On the Principles of Evaluation for Natural Language Generation

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

Dissertation

zur Erlangung des akademischen Grades Dr.-Ing.

vorgelegt von
Wei Zhao
geboren in Shanghai, China

Tag der Einreichung: 11. November. 2022
Tag der Disputation: 13. December. 2022

Referenten: Prof. Dr. Iryna Gurevych, Darmstadt, Germany
Prof. Dr. Steffen Eger, Bielefeld, Germany
Prof. Dr. Goran Glavaš, Würzburg, Germany

Darmstadt 2022
D17

Wei Zhao: On the Principles of Evaluation for Natural Language Generation
Darmstadt, Technische Universität Darmstadt
Year thesis published in TUprints: 2023
Day of the viva voce: 13. December 2022
Please cite this document as
URN: urn:nbn:de:tuda-tuprints-232959
URL: https://tuprints.ulb.tu-darmstadt.de/id/eprint/23295

This document is provided by tuprints,
E-Publishing-Service of the TU Darmstadt
http://tuprints.ulb.tu-darmstadt.de
mailto:tuprints@ulb.tu-darmstadt.de

Ehrenwörtliche Erklärung[1]

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades "Dr.-Ing." mit dem Titel "On the Principles of Evaluation for Natural Language Generation" selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 11. November 2022

Wei Zhao

[1] Gemäß §9 Abs. 1 der Promotionsordnung der TU Darmstadt

Abstract

Natural language processing is concerned with the ability of computers to understand natural language texts, which is, arguably, one of the major bottlenecks in the course of chasing the holy grail of general Artificial Intelligence. Given the unprecedented success of deep learning technology, the natural language processing community has been almost entirely in favor of practical applications with state-of-the-art systems emerging and competing for human-parity performance at an ever-increasing pace. For that reason, fair and adequate evaluation and comparison, responsible for ensuring trustworthy, reproducible and unbiased results, have fascinated the scientific community for long, not only in natural language but also in other fields. A popular example is the ISO-9126 evaluation standard for software products, which outlines a wide range of evaluation concerns, such as cost, reliability, scalability, security, and so forth. The European project EAGLES-1996, being the acclaimed extension to ISO-9126, depicted the fundamental principles specifically for evaluating natural language technologies, which underpins succeeding methodologies in the evaluation of natural language.

Natural language processing encompasses an enormous range of applications, each with its own evaluation concerns, criteria and measures. This thesis cannot hope to be comprehensive but particularly addresses the evaluation in natural language generation (NLG), which touches on, arguably, one of the most human-like natural language applications. In this context, research on quantifying day-to-day progress with evaluation metrics lays the foundation of the fast-growing NLG community. However, previous works have failed to address high-quality metrics in multiple scenarios such as evaluating long texts and when human references are not available, and, more prominently, these studies are limited in scope, given the lack of a holistic view sketched for principled NLG evaluation.

In this thesis, we aim for a holistic view of NLG evaluation from three complementary perspectives, driven by the evaluation principles in EAGLES-1996: (i) high-quality evaluation metrics, (ii) rigorous comparison of NLG systems for properly tracking the progress, and (iii) understanding evaluation metrics. To this end, we identify the current state of challenges derived from the inherent characteristics of these perspectives, and then present novel metrics, rigorous comparison approaches, and explainability techniques for metrics to address the identified issues.

We hope that our work on evaluation metrics, system comparison and explainability for metrics inspires more research towards principled NLG evaluation, and contributes to the fair and adequate evaluation and comparison in natural language processing.

Acknowledgments

I would like to express my greatest thanks to all people who supported me during this journey as advisors and as friends. This thesis would not have been possible without many of you.

First, I would like to thank my primary supervisor Prof. Steffen Eger. You provided me with tremendous support, advice, and inspiration. You are always patient with me when I had a hard time in language and culture integration. You are a great person showing an extreme devotion of science and scientific research with critical and ethical thinking.

Second, I am very thankful to my secondary supervisor Prof. Iryna Gurevych for creating an excellent environment at the graduate school on Adaptive Preparation of Information from Heterogeneous Sources (AIPHES). AIPHES provides a multidisciplinary research curriculum for Ph.D students, which enlarges my vision in scientific research, and gives me the opportunity to pursue my research interests collaboratively.

Third, I would like to thank Prof. Goran Glavaš and Prof. Isabelle Augenstein. You supported me all the way during my visit: providing the warmest welcome, and the greatest support to my research.

Further, I am very thankful to my collaborators for your guidance during my Ph.D studies: Prof. Johannes Bjerva, Prof. Michael Strube, Prof. Fei Liu, Dr. Yang Gao, Dr. Maxime Peyrard, Prof. Robert West and Prof. Erik Cambria. I learned very much from working with you!

Now, I would like to thank my colleagues: Prasetya Ajie Utama, Leonardo Filipe Rodrigues Ribeiro, P.V.S. Avinesh, Andreas Hanselowski, Debjit Paul, Fabrizio Ventola and many others. I enjoyed fun chats very much with you in-person and on zoom during the pandemic.

After all, I would like to thank my parents, my wife and many friends for always being there and for growing up together. I am grateful to have you in my life!

Sincerely,
Wei Zhao

Contents

Part I

Synopsis

Publications and Author Contributions

This thesis is based on eight scientific papers completed within a three-year doctoral program. Most of the papers have been published in top-tier international conferences from the major NLP events such as ACL and EMNLP. All papers have joint authorship. In the following, I list these papers in the order of their appearance in the thesis, and provide the statements of author contributions in each paper.

Subpart A corresponds to the following papers:

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer and Steffen Eger. 2019b. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China. Association for Computational Linguistics.

In this project, Maxime proposed the idea of combining Earth Mover Distance (EMD) with better embeddings such as ELMO and BERT. I proposed further ideas regarding how best to combine them: (a) n-gram EMD on the scale of word mover up to sentence mover; (b) routing and power means for aggregating word embeddings and (c) fine-tuning BERT. I also proposed a Proposition for the theoretical comparison between BERTScore and MoverScore. I did all the implementations (based on the scripts provided by Maxime), the experiments and analyses. I drafted the paper. Yang, Christian joined the project at a later time. Steffen was the advisor for this work. My advisor and all co-authors provided thoughtful feedback that greatly improved the final texts. Fei wrote the introduction and helped with related work (along with Maxime). All co-authors helped revise the introduction, corrected grammar mistakes, rewrote stylistically odd sentences, and helped shape the story. All authors agree with the use of this paper as part of Wei's cumulative doctoral thesis.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia. Association for Computational Linguistics.

This project encompassed my own spirits in ideas, writing, programming, experiments, analyses, etc, which was undertaken in a way to assess my research independence. Steffen was the advisor for this work. Michael and my advisor provided several rounds of thoughtful feedback that greatly improved the final texts. All authors agree with the use of this paper as part of Wei's cumulative doctoral thesis.

Wei Zhao, Haiyun Peng, Steffen Eger, Erik Cambria and Min Yang. 2019a. Towards Scalable and Reliable Capsule Networks for Challenging NLP Applications. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy. Association for Computational Linguistics.

This project was partially done when I did my internship at Nanyang Technological University, and completed at Darmstadt. Steffen and Erik were the advisors for this work. I proposed the ideas, and did most of the implementations and all the analyses. Haiyun did a substantial part of data preprocessing, and took most of the work on the literature review. Min did early experiments for the QA task. I wrote the paper, and did the major revisions based on thoughtful feedback from my advisors and all co-authors. All authors agree with the use of this paper as part of Wei's cumulative doctoral thesis.

Subpart B corresponds to the following papers:

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West and Steffen Eger. 2020. On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics.

This project was partially done when I visited the University of Mannheim, and completed at Darmstadt. Steffen and Goran were the advisors for this joint work. I proposed the ideas, and completed all the experiments and analyses. Steffen proposed two additional ideas regarding (a) combining metrics with a target-side language model and (b) the analysis of metric preference. I implemented these ideas, with the second idea completed in the camera ready version. Maxime, Yang and Robert joined the project at a later time. I drafted the paper, and did the revisions based on the iterations of thoughtful feedback from my advisors and all co-authors. My advisors did the annotation work for human judgments, rewrote the introduction and conclusion, restructured and polished the paper. Steffen wrote the section of metric preference in the final version. All authors agree with the use of this paper as part of Wei's cumulative doctoral thesis.

Wei Zhao, Steffen Eger, Johannes Bjerva, Isabelle Augenstein. Inducing Language-Agnostic Multilingual Representations. 2021. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Se-*

mantics. Bangkok, Thailand (online). Association for Computational Linguistics.

This project was partially done when I did my internship at the University of Copenhagen, and completed at Darmstadt. Steffen and Isabelle were the advisors for this joint work. I proposed the ideas for analyzing the source of cross-lingual ability. Johannes provided me with many insights based on a crucial linguistic typology resource (WALS). Building upon these insights, I proposed linguistic ideas in the form of text normalization. I did all the experiments and preliminary analyses. Johannes provided linguistic analysis to the results, which led to the iterations of the ideas. I proposed two additional ideas, and did all the experiments and analyses. I drafted the paper, and did the major revisions in a way to address the iterations of thoughtful feedback from my advisors and Johannes—who also helped finalize some texts and shape the story. Johannes took the work for the literature review of linguistic typology. My advisors rewrote the introduction, corrected grammatical mistakes, and shortened the text. All authors agree with the use of this paper as part of Wei's cumulative doctoral thesis.

Wei Zhao and Steffen Eger. 2022. Constrained Density Matching and Modeling for Cross-lingual Alignment of Contextualized Representations. In *Proceedings of The 14th Asian Conference on Machine Learning*. Hyderabad, India. Proceedings of Machine Learning Research (PMLR).

This project was undertaken for assessing my research independence. Steffen was the advisor for this work. I supervised a Hiwi student, Dan Liu, who ran several baselines for the word alignment task. This project involved extensive work in machine learning and statistics, to which my advisor provided in-depth feedback. My advisor provided thoughtful feedback that greatly improved the final texts. All authors agree with the use of this paper as part of Wei's cumulative doctoral thesis.

Subpart C corresponds to the following paper:

Maxime Peyrard, **Wei Zhao**, Steffen Eger and Robert West. 2021. Better than Average: Paired Evaluation of NLP systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*. Online. Association for Computational Linguistics.

This project was based on Maxime's idea. Robert was the advisor for this work. Maxime and I together did the implementations and extensive experiments on 296 NLP evaluation setups across 4 tasks and 18 evaluation metrics. Maxime did the writing and revisions based on several iterations of the feedback from other co-authors and me. I did the addition to this work: I applied the idea of paired comparison to the evaluation of Eval4NLP shared tasks, and identified a crucial limitation of this idea. I did all the experiments, analyses and writing

in this matter. The results and texts have not been published yet. I add these novel content to the appendix of Chapter 11 in the thesis for the interest of completeness. All authors agree with the use of this paper as part of Wei's cumulative doctoral thesis.

Subpart D corresponds to the following paper:

Marvin Kaster, **Wei Zhao** and Steffen Eger. (2021). Global Explainability of BERT-Based Evaluation Metrics by Disentangling along Linguistic Factors. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online. Association for Computational Linguistics.

This project was initialized by Steffen, who was the advisor for this work. Micha Dippel did early experiments for his Bachelor work, but showed no interest in writing a joint paper. Marvin was a Hiwi student, taking over the experiments after Micha. Marvin and my advisor took most of the writing. I joined this project at a later time, contributing to the related work, experiments, data processing, and feedback. I took the work for a crucial analysis regarding the root causes of low R^2 scores, but completed the work (including ideas, experiments and writing) in the camera ready version. I corrected some mistakes in Marvin's implementation and the results in the final version. All authors agree with the use of this paper as part of Wei's cumulative doctoral thesis.

The following publications are the outcomes of my contributions to the workshops on Evaluation and Comparison of NLP Systems, co-located at EMNLP. Given my contributions are mostly in the form of organization, these publications are not included in the text of this thesis.

Marina Fomicheva, Piyawat Lertvittayakumjorn, **Wei Zhao**, Steffen Eger and Yang Gao. The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results. 2021a. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics.

Gao Yang, Steffen Eger, **Wei Zhao**, Piyawat Lertvittayakumjorn and Marina Fomicheva. 2021. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics.

Steffen Eger, Yang Gao, Maxime Peyrard, **Wei Zhao** and Eduard Hovy. 2020. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics.

The following papers are published during the course of my doctoral studies. Some of the papers (for which my contributions are not substantial) are strongly related to this thesis, whereas others are not. All these papers do not meet the

formal criteria for inclusion to Kumulative Dissertation at Fachbereich Informatik, TU Darmstadt. I list them here to demonstrate my enormous research efforts and interests in the topic of evaluation, as well as the broadness of my research.

- Yang Gao, **Wei Zhao**, Steffen Eger. (2020). SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

- Gregor Geigle, Jonas Elias Stadtmüller, **Wei Zhao**, Jonas Pfeiffer and Steffen Eger. TUDa at WMT21: Sentence-Level Direct Assessment with Adapters. (2021). In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

- Haixia Chai, **Wei Zhao**, Steffen Eger and Michael Strube. (2020). Evaluation of Coreference Resolution Systems Under Adversarial Attacks, In *Proceedings of the First Workshop on Computational Approaches to Discourse*. Association for Computational Linguistics.

- Yang Li, **Wei Zhao**, Erik Cambria, Suhang Wang and Steffen Eger. (2021). Graph Routing Between Capsules. In *Journal of Neural Networks*. Elsevier.

The source code of the papers related to this thesis is publicly available at https://github.com/AIPHES.

Chapter 1

Introduction

1.1 Motivation

Natural language processing (NLP), as an engineering branch of computational linguistics, is concerned with designing computational systems for accomplishing a variety of practical tasks on natural language texts. Accordingly, system evaluation and comparison come in pairs, aiming to evaluate how much a system's behavior deviates from anticipation and to affirm a system superior to the other. To date, within deep learning revolution, reporting trustworthy and unbiased results has become progressively challenging, as the majority of systems have trouble in reproducibility and transparency. However, these issues have to be addressed in order to properly track the advances in the fast-growing NLP community.

Fair and adequate evaluation and comparison have fascinated the scientific community for long, and the genesis of the field can be traced back to Woods (1977), which sketched the blueprint for NLP evaluation from the following two perspectives:

- A system with a smaller size wins under the condition that two systems capture the same amount of linguistic phenomena.
- A good NLP system should run faster than a bad system implemented cleverly.

However, given the shortage of computational power back then, the sketched blueprint cannot carry out. Therefore, NLP researchers resorted to case studies on chosen examples and counterexamples for evaluating systems. Up until the late 90's, with the increased amount of computing power according to Moore's law, Guida and Mauri (1986) and Hollnagel (1986) introduced *quality assurance* concerning the assessment of system performance in quantitative and diagnostic manners as listed below:

- Quantify the extent to which one system surpasses the other.
- Troubleshoot malfunctions of a system through case studies.

Jones and Galliers (1995) sketched evaluation concerns from two perspectives, i.e., intrinsic and extrinsic evaluation. While the former evaluates the desired functionalities of the system itself, the latter evaluates the impact of a system on an external task.

The European project EAGLES-1996[1], being the acclaimed extension to the ISO-9126 standard for software quality evaluation, is, arguably, one of the most comprehensive studies on the fundamental principles pertaining to evaluation concerns, criteria and measures for evaluating natural language technologies. This lays the foundations of succeeding methodologies in the evaluation of natural language. In particular, EAGLES-1996 outlines evaluation principles from three complementary perspectives as follows:

- *Adequacy Evaluation* quantifies the extent to which a system is adequate for intended use.
- *Progress Evaluation* investigates whether or not the current state of a system achieves the state-of-the-art.
- *Diagnostic Evaluation* concerns the rationales on why system output deviates from anticipation.

NLP encompasses an abundance of applications, each with its own evaluation concerns, criteria and measures. In this thesis, we do not aim for a comprehensive study but particularly focus on the evaluation in natural language generation (NLG), which touches on, arguably, the most human-like NLP applications. We hope that our research contributes to the evaluation of general language technologies. In the following, we sketch EAGLES-1996 evaluation principles in the form of NLG evaluation.

Adequacy Evaluation → Evaluation Metrics. NLG systems aim for producing natural language outputs corresponding to the given inputs in various forms, be it text, image or table. Jones (2001) acknowledged the importance of human evaluation, for which human experts are asked to judge *the adequacy of system outputs* according to predetermined criteria. For instance, given a system translation and a human reference, translators are asked to rate the translation in how adequately it reflects the meaning of the reference. However, a well-designed human evaluation requires a large investment of time and resources, thus not favored today by most. In stark contrast, cost-efficient NLG evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), evaluating system performance in seconds or minutes, have received considerable attention, and they quantify day-to-day progress and form the basis for ever-increasing NLG community. However, previous works cannot address high-quality metrics in multiple scenarios, such as long texts and in the absence of lexical overlap and human references.

[1] https://www.issco.unige.ch/en/research/projects/ewg96/ewg96.html

Progress Evaluation → System Comparison. As for all tasks in NLP, a state-of-the-art system, being a direct reflection of the most recent advances in the field, has to be affirmed with no doubt. Leaderboards, arguably, are the standard evaluation hubs that benchmark competing systems with evaluation metrics, publish their results and affirm a winner of the systems (Tague-Sutcliffe, 1992; Voorhees, 2003; Hu et al., 2020). However, high-quality metrics could misjudge the state-of-the-art when systems are not compared in a rigorous manner, such as employing inappropriate significance tests (Simpson, 2021) and reporting single-point estimate of performance (Reimers and Gurevych, 2017).

Diagnostic Evaluation → Explainability. Explainability has been researched for long in artificial intelligence, which not only concerns the understanding of model process and model outputs, but also lays the foundations of identifying model's limitations, and as such is more comprehensive than the scope of diagnostic evaluation. In NLG evaluation, though recent metrics based on blackbox language models exhibit high quality levels, few work has touched on the understanding of these non-transparent metrics. Therefore, the judgments from these metrics can hardly be justified.

1.2 Research Objectives

Given the complementarity of evaluation metrics, system comparison and explainability, it is evident that a holistic view of principled NLG evaluation from these perspectives is required in order to ensure trustworthy, reproducible and unbiased results. In the following, we outline the research questions driven by the current state of challenges pertaining to the inherent characteristics of these perspectives.

Evaluation Metrics. As the cost-efficient alternative to human evaluation, automatic evaluation with its own criteria and metrics has received attention for two decades. However, previous metrics correlate poorly with human judgment of text quality in multiple scenarios. For instance, research showed that BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which count the n-gram matches between system output and human reference, are inadequate in evaluating NLG systems in the absence of lexical overlap between system and human pair (Reiter, 2018). As a second example, though recent metrics building on contextualized representations are shown to be high quality in reference-based evaluation, they could not show advantages in the absence of human references and in evaluating long texts.

Recently, aiming to fuel more research on this topic, we organized the workshops on Evaluation and Comparison of NLP Systems (Eger et al., 2020; Gao et al., 2021), outlining the anticipated properties of evaluation metrics: (i) high correlation with human assessments in both reference-based and reference-free evaluations and (ii) robustness across lengths of input and output sequences. We now dissect these properties into three research questions.

- **RQ1. What are essential elements for reference-based metrics to be**

high quality in the absence of lexical overlap?

Recognizing lexical similarity is crucial in order for reference-based metrics to properly judge system outputs, especially in the case where system output and human reference lack lexical overlap. Research showed that text representations, grounded in the distributional hypothesis (Harris, 1954), can address lexical similarity with broader semantic relatedness of words in vector space (Hill et al., 2015). However, little is known how best to exploit the powers of these representations for achieving high-quality metrics in *reference-based scenarios*.

- **RQ2. How to design reference-based metrics targeting the assessment of text coherence for evaluating long texts?**

  Text coherence plays a vital role in the assessment of long texts. Research showed that recent metrics based on contextualized representations cannot recognize text coherence (Fabbri et al., 2020; Yuan et al., 2021; Sai et al., 2021). This is not surprising as language models responsible for producing contextualized embeddings mostly do not take account of the interdependence between sentences, such as discourse phenomena in the inter-sentence context.

- **RQ3: What are essential elements for reference-free metrics to be high quality in the absence of human references?**

  In *reference-free scenarios*, evaluation metrics aim for unlimited evaluation by means of removing the need for human references. However, the proposals of reference-free metrics are in need of human ratings as supervision (Specia et al., 2010) and language-specific preprocessing (Lo et al., 2014; Lo, 2019), which hinder the broader use of these metrics. Further, research showed that the qualities of these metrics are far below reference-based BLEU, invented two decades ago (Ma et al., 2019).

System Comparisons. Evaluation metrics and system comparisons are interrelated, and both affect evaluation results. For instance, high-quality metrics can misjudge a state-of-the-art system on leaderboards in which systems are ranked according to the average of instance-level evaluation scores. We show that global statistics such as average and median cannot carry out rigorous comparison, as they ignore the fact that systems are evaluated on the same test instances (see Chapter 11). We now present the research question related to rigorous system comparison.

- **RQ4: What is the rigorous comparison approach in order for leaderboards to report correct system rankings?**

Explainability for Evaluation Metrics. Unlike the advances in explainable artificial intelligence, few work has touched on the understanding of evaluation metrics along the dimensions, such as (i) visualizing the process of evaluation metrics; (ii) understanding what linguistic factors these metrics capture; (iii) providing rationales as to why one metric is superior to another and (iv) providing justifications to the judgments from these metrics. We now present the research question corresponding

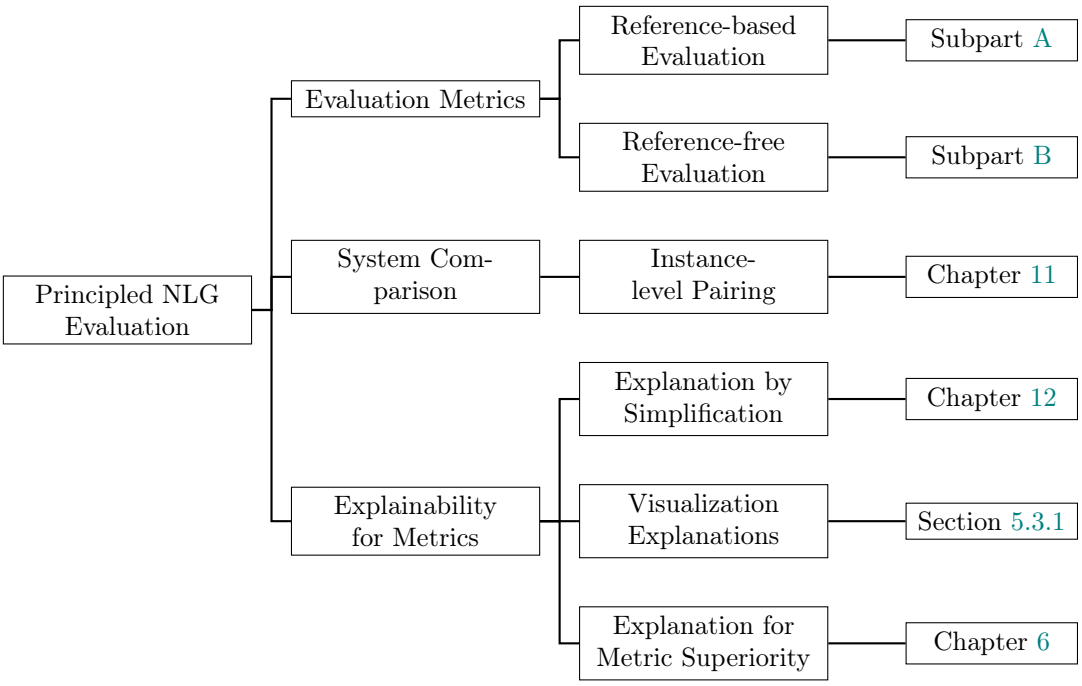


Figure 1.1: Classification of the contributions presented in this thesis [2].

to the explainability of evaluation metrics.

- **RQ5. What insights can be drawn from explainable artificial intelligence in order to understand non-transparent evaluation metrics?**

1.3 Thesis Organization

Figure 1.1 provides the overview structure of the contributions presented in this thesis. In Chapter 2, we provide an in-depth background on principled evaluation in natural language generation from three complementary perspectives: evaluation metrics, system comparison and explainability, and we summarize our contributions that address the five research questions above.

In Subpart A, we introduce two evaluation metrics in the presence of human references. In Chapter 5, we introduce MoverScore to address the absence of lexical overlap with contextualized word embeddings, while we introduce DiscoScore to address long-text evaluation based on word embeddings and discourse coherence in Chapter 6. In Chapter 7, we propose the KDE routing for aggregating capsules (another form of embeddings)—which we use to address the aggregation of word embeddings across the layers of contextualized encoders for our evaluation metrics.

In Subpart B, we introduce XMoverScore, a parameterized metric in the absence of human references, particularly to address reference-free machine translation evaluation, i.e., comparing system translations with source language texts. XMoverScore

[2] We adopt the standard classification of explainable artificial intelligence to distinguish the explainability techniques for understanding evaluation metrics (see more details in Section 4).

can be parameterized with the choice of solutions to rectifying the vector space of different languages—which results in high-quality cross-lingual embeddings. In particular, we reduce language bias and rotate the vector space in Chapter 8. Chapter 9 addresses the removal of language identity signals from the vector space, while Chapter 10 addresses the vector space alignment for low-resource languages.

In Chapter 11, we propose pairwise comparison approaches for reporting correct system rankings on leaderboards. Our contributions to the explainability for metrics are manifold. We visualize the process of MoverScore in Section 5.3.1, and provide rationales as to why one metric is superior to another in Chapter 6. Chapter 12 addresses understanding on what linguistic factors recently proposed metrics capture. In Chapter 13, we conclude this thesis and present an outlook for future work.

Chapter 2

Text Quality Evaluation

Natural language generation (NLG) systems aim for producing system outputs faithful to the given source inputs in various forms, such as source language texts in machine translation, documents in text summarization, and utterances in conversational dialog. Evaluating these systems has a long transition course from human-centralized to automatic evaluation, both concerned with assessing the text quality of system outputs. In this chapter, we start by outlining the progression of human evaluation, and then introduce automatic evaluation that performs assessments with low-cost evaluation metrics. Lastly, we summarize our contributions presented in this thesis.

2.1 Human Evaluation

As for all tasks in NLG, system outputs are deemed open-ended, i.e., that multiple outputs can be correct to a given input. For instance, a source language text allows for numerous paraphrased system translations. In this context, the judgments of human experts, e.g., professional translators, appear to be the gold standard for evaluating system outputs, even though human experts often produce inconsistent assessments after time-consuming work. We now outline the evaluation concerns, criteria and guidelines responsible for ensuring high-quality human evaluation.

2.1.1 Evaluation Concerns and Criteria

NLG encompasses a range of tasks, each with its own course of human evaluation, driven by the fundamental nature of the task. This section particularly discusses the concerns and criteria of human evaluation in summarization and machine translation.

Text Summarization. Figure 2.1 shows the timeline and progression of human evaluation criteria in summarization. In the following, we outline these criteria and the major concerns of them. Table 2.1 shows evaluation corpora adopting these criteria to carry out human evaluation.

2018 - Recent Edition of Readability (Grusky et al., 2018)
2005 - Responsiveness (Dang, 2005)
2004 - Content Coverage of Summary Content Units (Nenkova and Passonneau, 2004)
2004 - Content Coverage of Factoids (Teufel and Halteren, 2004)
2003 - Content Coverage of Discourse Units (Lin, 2001)
2003 - Readability (Lin, 2001)

Figure 2.1: Timeline and progression of human evaluation criteria in summarization.

Corpora	DiscoUnit	Factoid	SCU	Resp	Info	Flu	Rel	Coh
REALSUM-2020			✓					
SummEval-2020		✓			✓	✓		✓
NewsRoom-2018					✓	✓	✓	✓
Rank2019		✓						
QAGS-2020		✓						
DUC-[01,02,03,04]	✓							
DUC-2005	✓			✓				
DUC-2006			✓	✓				
DUC-2007			✓	✓				
TAC-[08,09,10,11]			✓	✓				

Table 2.1: Relationships between human evaluation criteria and corpora in summarization. SCU, Resp, Info, Flu, Rel, Coh mean Summary Content Unit (Pyramid), Responsiveness, Informativeness, Fluency, Relevance, and Coherence. These corpora consist of source language texts, system and human translations, as well as human ratings according to the selected human evaluation criteria.

- *Content Coverage of Discourse Units*: Summary Evaluation Environment (SEE) pioneered the human evaluation criterion in summarization. SEE starts by extracting the elementary discourse units from human reference with SPADE[1], a discourse annotation tool, and then recruits human raters to underline the sentences within system summary that are relevant to the extracted discourse units, and finally denotes *content coverage* by the number of the marked sentences (Lin, 2001) (see Figure 2.2). This criteria was adopted in the early iterations of the evaluation campaigns in the Document Understanding Conferences (DUC).

  > **Limitation**: Lin and Hovy (2002) acknowledged the inherent limitations in SEE: (i) given a source document, SEE only offered a single human reference, which falsely assumes a unique solution for system summary, and (ii) given discourse tools are prone to mistakes, this criterion cannot reach an acceptable agreement level between human raters.

- *Content Coverage of Factoids*: to be able to overcome the issues in SEE, Teufel and Halteren (2004) introduced a factoid-based, content coverage criterion,

[1] https://www.isi.edu/publications/licensed-sw/spade/

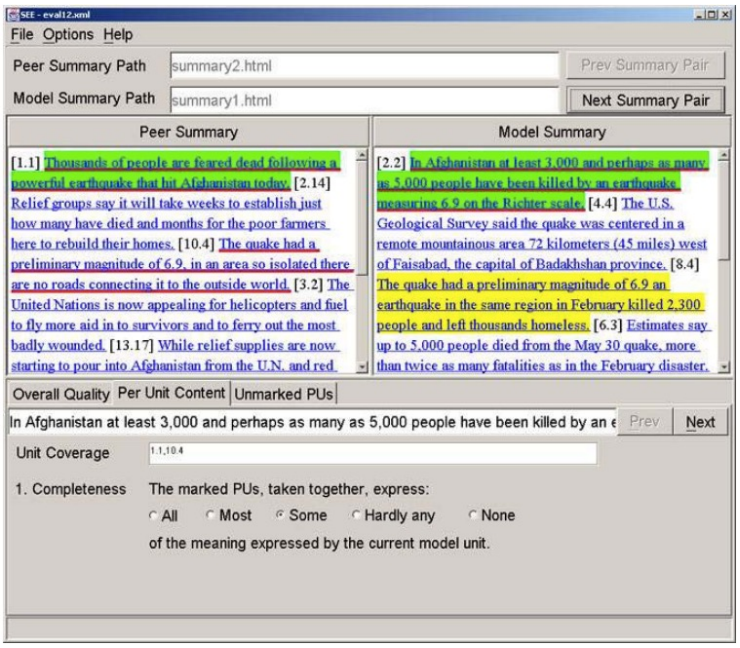


Figure 2.2: Summary Evaluation Environment (Lin, 2001).

which includes multiple human references for evaluating a system summary, and asks humans to produce the factoids from the references and the system summary. Consider the following example:

> **Example**
>
> Given the sentence "*A man was accused of murdering his wife Jeniffer.*", the corresponding factoids could be "*The man is a killer.*", and "*The man is a husband of Jeniffer.*".

As such, *content coverage* denotes the number of factoid overlaps between system and reference summaries. This criteria has been adjusted and adopted in recent evaluation corpora such as SUMMEval (Fabbri et al., 2020), Rank2019 (Falke et al., 2019), and QAGS (Wang et al., 2020).

- *Content Coverage of Summary Content Units*: Nenkova and Passonneau (2004) introduced another content coverage criterion, the so-called Pyramid, which evaluates system summary in multi-reference scenarios, and asks humans to annotate Summary Content Units (SCUs) in system and human summaries. Each SCU can be weighted and placed in a certain level of a pyramid according to its occurrence in the reference summaries. Thus, *content coverage* denotes the number of SCU overlaps between system and human summaries. Fuentes et al. (2005) extended upon the idea of Pyramid by exploring approaches for automatic SCU annotation. Pyramid has been adopted in DUC-2006 and DUC-2007 and in the evaluation campaigns in Text Analysis Conferences (TAC)[1], and the recent evaluation corpus— REALSUM (Bhandari et al., 2020).

[1] http://www.nist.gov/tac/

Quizzes
1. About how many gross capitalization errors are there?
2. About how many sentences have incorrect word order?
3. About how many times does the subject fail to agree in number with the verb?
4. About how many of the sentences are missing important components (e.g. the subject, main verb, direct object, modifier) causing the sentence to be ungrammatical, unclear, or misleading?
5. About how many times are unrelated fragments joined into one sentence?
6. About how many times are articles (a, an, the) missing or used incorrectly?
7. About how many pronouns are there whose antecedents are incorrect, unclear, missing, or come only later?
8. For about how many nouns is it impossible to determine clearly who or what they refer to?
9. About how times should a noun or noun phrase have been replaced with a pronoun?
10. About how many dangling conjunctions are there, such as *and* and *however* ?
11. About many instances of unnecessarily repeated information are there?
12. About how many sentences strike you as being in the wrong place because they indicate a strange time sequence, suggest a wrong cause-effect relationship, or just do not fit in topically with neighboring sentences?

Table 2.2: Linguistic Quizzes for evaluating readability (Over and Liggett, 2002).

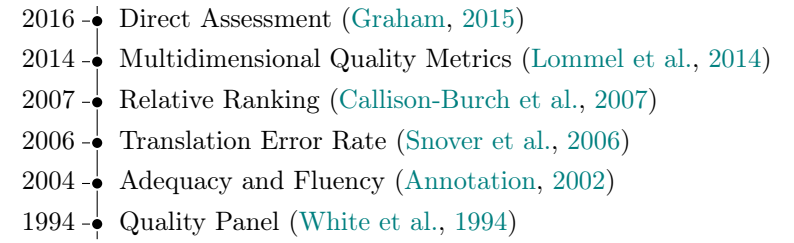


Figure 2.3: Timeline and progression of human evaluation criteria in MT.

- *Responsiveness*: much unlike previous criteria, Dang (2005) introduced an reference-free criterion, which removes the dependence on human summaries, and compares system summary with the topic statement of source document, aiming to measure the amount of information in the summary meeting the information need expressed in the topic statement. Humans are asked to read the topic statement and the associated summaries, and then rate the summary on 5-point scale from worst to best. This criteria was adopted in the editions of DUC from 2005 to 2007, and the TAC evaluation campaigns.

- *Readability and Recent Edition*: further, aiming to know whether system outputs conform to linguistic rules, the DUC-2002 evaluation campaign (Over and Liggett, 2002) adopted the SEE environment to measure the *readability* of system summaries by filling out twelve linguistic quizzes (see Table 2.2), which is time-consuming, however. Accordingly, recent evaluation corpora, such as SUMMEval (Fabbri et al., 2020) and RealSUM (Bhandari et al., 2020), adopted a new edition of *readability* to judge system summaries, which focuses on four aspects: informativeness, coherence, relevance and fluency, and as such requires lower human effort.

Machine Translation. Figure 2.3 shows the timeline and progression of human evaluation criteria in machine translation. In the following, we outline these criteria

Corpora	QP	Adequacy	Fluency	RR	DA	HTER	MQM
ARPA-1994	✓						
MTC-P4		✓	✓				
OpenMT-2015		✓	✓				
WMT-[14,15]				✓			
WMT-[16,17,18,19,20]				✓	✓		
WMT-21					✓		✓
WMT-QE-[19,20]					✓	✓	✓
MLQE-2020						✓	
MLQE-PE-2020						✓	
Eval4NLP-2021						✓	

Table 2.3: Links between human evaluation criteria and corpora in machine translation. QP, RR, DA, HTER and MQM mean Quality Panel, Relative Ranking, Direct Assessment, Human Translation Error Rate and Multidimensional Quality Metrics. These corpora consist of source language texts, system and human translations, as well as human ratings according to the selected human evaluation criteria.

and the major concerns of them. Table 2.3 shows evaluation corpora adopting these criteria to carry out human evaluation.

- *Quality Panel*: professional translators are asked to judge system translations through multi-round panel discussions along predefined linguistic criteria, such as lexical choice, grammaticality, semantics, stylistics, fluency, and so forth. ARPA[2] recruited translators to perform human evaluation with these criteria (White et al., 1994). While reliable, multi-round panel discussions are extremely laborious, and thus do not exist for long.

- *Adequacy and Fluency*: aiming for low-cost human evaluation, Linguistics Data Consortium introduced two criteria, namely *Adequacy and Fluency* on a 5-point Likert scale, which do not require internal discussion, but rather ask translators to complete judgments independently (Annotation, 2002). The annual Open Machine Translation campaigns[3] adopted these criteria to carry out human evaluation. Table 2.4 shows the 5-point scales of these criteria.

  > Adequacy: to what extent the information of professional human translation is expressed in system translation?
  >
  > Fluency: to what extent system translation is well-formed according to the grammar of target language?

  > **Limitation**: given a 5-point scale, these criteria are inadequate to discriminate system translations of varying qualities.

- *Relative Ranking*: in order to carry out nuanced, fine-grained human evalu-

[2] The ARPA MT Initiative is part of the Human Language Technologies Program of the Advanced Research Projects Agency Software and Intelligent Systems Technology Office.

[3] https://www.nist.gov/itl/iad/mig/open-machine-translation-evaluation

Adequacy	Fluency
5: All	5: Flawless
4: Most	4: Good
3: Much	3: Non-native
2: Little	2: Disfluent
1: None	1: Incomprehensible

Table 2.4: Adequacy and Fluency on 5-point scales in MT human evaluation.

ation, Callison-Burch et al. (2007) recruited translators to rank translations produced by multiple systems in pairs. For instance, given a source language text and two translations at a time, translators are asked to judge which one is superior to the other. Ties are allowed when two translations are of similar qualities. This criterion has become popular in human evaluation, adopted in recent iterations of the WMT metrics shared tasks (Bojar et al., 2017; Ma et al., 2018, 2019; Mathur et al., 2020b).

> **Limitation**: ranking translations in pairs cannot distinguish the magnitude of the differences in translations quality (Bojar et al., 2011).

- *Direct Assessment*: further, to be able to quantify the differences in translation quality, Graham (2015) extended upon the idea of *Adequacy* by taking a continuous scale from 0 to 100, showing that recruiting 15 professional translators can achieve an acceptable agreement between translators. This criterion has been applied in recent editions of the WMT metrics shared tasks (Bojar et al., 2017; Ma et al., 2018, 2019; Mathur et al., 2020b), and the WMT quality estimation shared task (Fonseca et al., 2019; Specia et al., 2020), and the Eval4NLP shared task (Fomicheva et al., 2021b).

- *Human Translation Error Rate*: Snover et al. (2006) introduced a post-editing criterion, which asks translators to underline edits, such as insertions, deletions and replacements, needed for correcting a system translation, and computes the ratio between the number of edits and the reference translation length. This criterion has been adopted in the WMT shared tasks for quality estimation (Fonseca et al., 2019; Specia et al., 2020), the MLQE (Fomicheva et al., 2020b) and MLQE-PE (Fomicheva et al., 2020a) corpora.

- *Multidimensional Quality Metrics*: aiming for mastering the judgment of translation quality, Lommel et al. (2014) assembled over 100 translation errors in the literature, and provided the taxonomy of these errors, termed the Multidimensional Quality Metrics (MQM) framework. The WMT shared tasks for quality estimation (Fonseca et al., 2019; Specia et al., 2020) selected a subset of translation errors from the MQM framework to perform fine-grained human evaluation, such as assigning error types to mistranslated words. For that reason, research showed that MQM is superior to coarse-grained *Direct Assessment* in discriminating human and system translations (Freitag et al., 2021). Figure 2.4 shows the overview structure of the MQM framework.

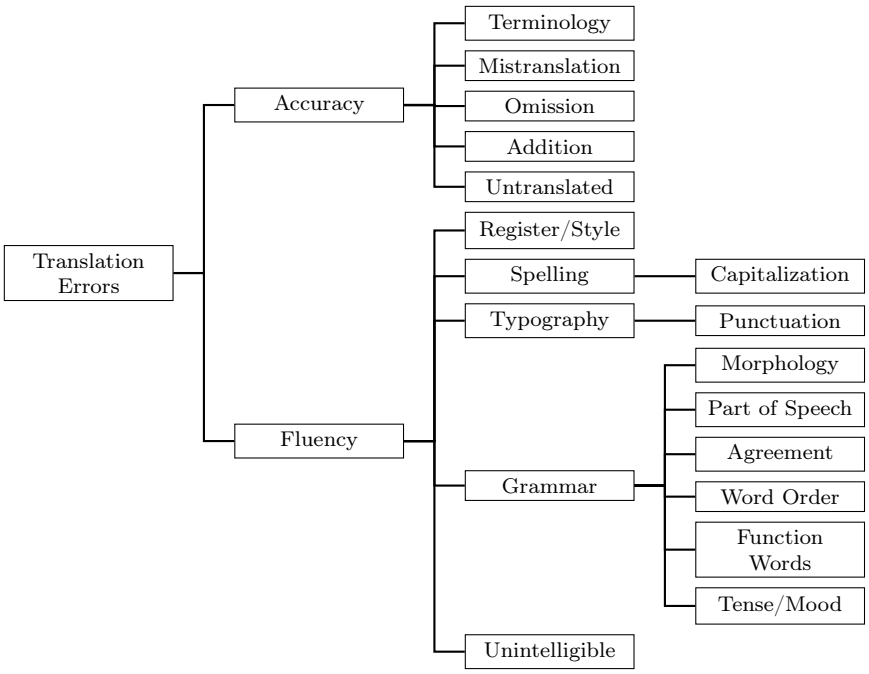


Figure 2.4: High-level taxonomy of translation errors outlined in MQM.

2.1.2 Evaluation Guideline

Human judgments are notoriously subjective, inconsistent, not reproducible. As a consequence, a well-designed guideline is required for ensuring high-quality human evaluation. We discuss the major elements of an evaluation guideline as follows:

- *The number of human raters recruited*: in machine translation, research showed that a number of 15 professional translators is at minimum in order to reach an acceptable agreement level between translators when adopting Direct Assessment to judge system translations on segment level (Graham et al., 2015). However, the minimum number has to increase to 100 when making judgments on document level (Mathur et al., 2020b).

- *Pre-qualification tests to filter out disqualified raters*: in order to ensure high-quality judgments, human raters are asked to participate in qualification tests. Only raters who pass the tests are allowed to continue the evaluation process. Research showed that disqualified raters can be recognized with demographic characteristics, such as nationality and age (Downs et al., 2010), but the qualification of these raters can be improved after training (Mitra et al., 2015).

- *The number of assessment rounds*: negotiation plays a vital role in the quality of human judgments. Research showed that, though human raters often disagree with each other in the first round, they negotiate on the details of evaluation criteria in follow-up meetings, and finally can result in an acceptable agreement (Iskender et al., 2020).

- *Experts vs non-experts*: recruiting human experts or non-experts to carry out human evaluation seems a controversy in the literature. For instance, research

Scenarios	Source	Hypothesis	Reference	Human Rating
Evaluation without Supervision	(✓)	✓	(✓)	
Evaluation with Supervision	(✓)	✓	(✓)	✓

Table 2.5: Classification of evaluation scenarios. Hypothesis means system output. (✓) means the dependence on source text and human reference is optional.

showed that non-experts, such as crowdsourcing workers and laboratory students, are indifferent from experts in quality when carrying out human evaluation in text classification (Snow et al., 2008) and machine translation (Callison-Burch, 2009). However, non-experts exhibit quality levels well below experts in summarization (Gillick and Liu, 2010; Lloret et al., 2013). Iskender et al. (2020) recently showed that the overall quality of non-experts can be considerably improved in summarization by increasing the number of crowdsourcing workers, but the study is limited in scope to German as the only language.

Inter-rater Agreement As discussed above, the consensus between human raters determines the quality of human evaluation. We briefly outline the four common measures of inter-rater agreement as follows:

- *Percent Agreement* refers to the percent of instances for which two human raters agree with one another, but considers all instances equally.
- *Cohen's k* (Cohen, 1960) builds on top of Percent Agreement, which considers the odds of that two human raters arrive at the same judgments.
- *Fleiss's k* (Fleiss, 1971) extends Cohen's k to operate in multi-raters settings, which computes the extent of observed agreements over the agreements expected by chance.
- Krippendorff's α (Krippendorff, 1970) is a measure of inter-rater disagreement, which computes the extent of observed disagreements over the disagreements expected by chance.

2.2 Automatic Evaluation

Human evaluation carried out by following the instructions of a well-designed evaluation guideline requires a large investment of time and resources. For instance, research showed that a cost of 3,000 hours is required for human raters to complete evaluation in multi-documents summarization (Lin, 2004). Accordingly, automatic evaluation emerged, which assesses system outputs in seconds or minutes with evaluation metrics, and as such has been favored by most today.

Table 2.5 shows two evaluation scenarios corresponding to two classes of evaluation metrics: unsupervised and supervised metrics, which are distinguished by whether or not they use human rating as supervision. Both classes of metrics can be deemed *reference-based* or *reference-free*, which depends on the input arguments

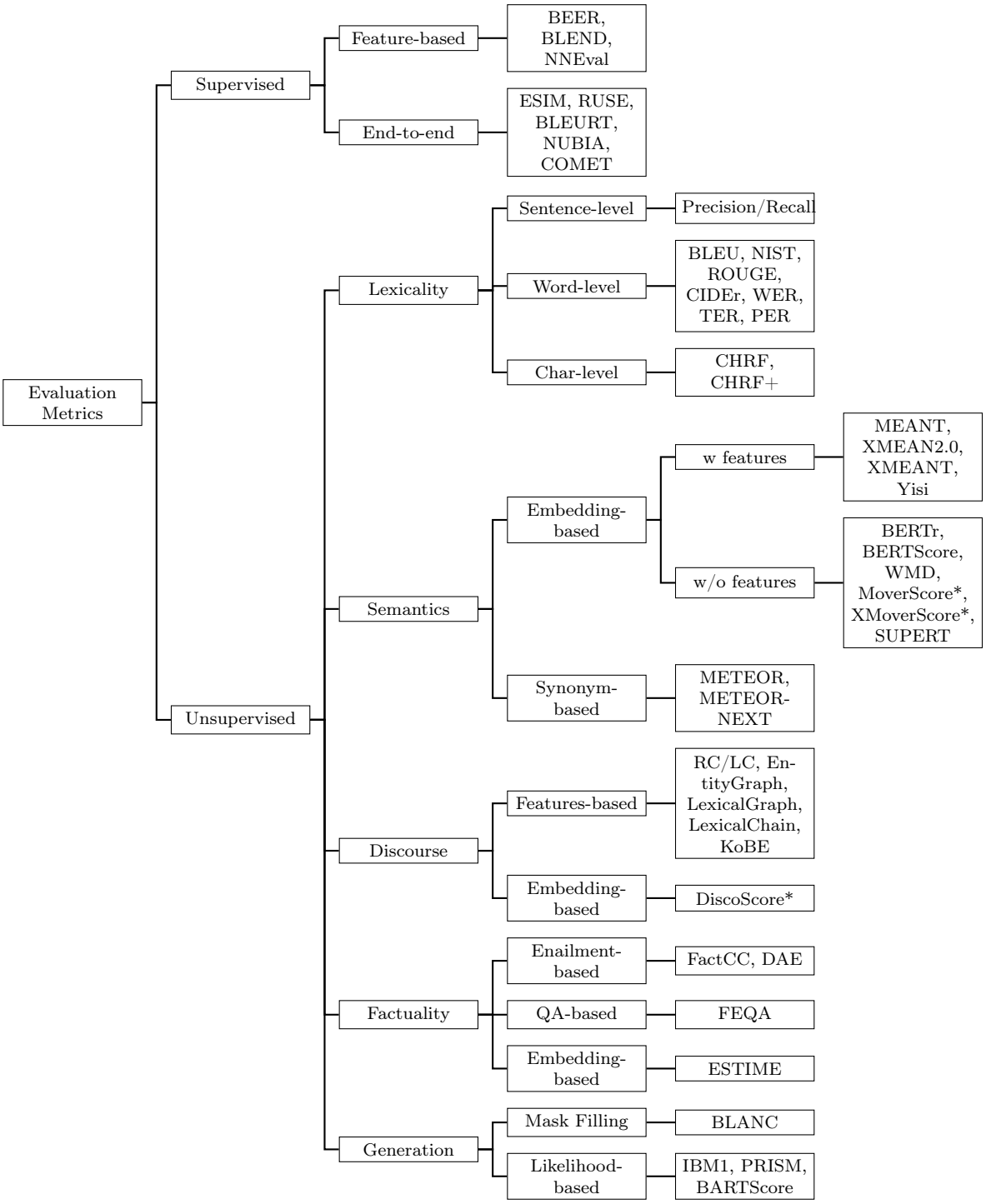


Figure 2.5: Taxonomy of NLG evaluation metrics. (*) marks the metrics proposed in this thesis.

they take, i.e., either they compare system output with human reference or with source text. In this section, we will outline unsupervised and supervised metrics. Figure 2.5 shows the taxonomy of these metrics.

2.2.1 Evaluation without Supervision

As discussed above, unsupervised evaluation metrics judge the text quality of system outputs without the access to human ratings. Often, recent surveys on NLG evaluation classify unsupervised metrics into two classes: (i) lexical-based and (ii) semantic-based (Celikyilmaz et al., 2020; Sai et al., 2022). We complement this classification of unsupervised metrics with three extra classes: (iii) discourse-based; (iv) factual-based and (v) generation-based. We now discuss these metrics under each class as follows:

Lexical-based Metrics. Lexical-based evaluation metrics, comparing system output with human reference or with source text on lexical level, are often termed retrieval-based metrics, which treat text as a bag of words or sentences. We now outline these metrics as follows:

- *Precision* and *Recall* (Kupiec et al., 1995) are sentence-level, reference-based evaluation metrics, pioneered in summarization, which compares system summary with human reference with Recall and Precision. Recall computes the fraction of the sentences in human reference that occur in system summary, while Precision computes the faction of the sentences in system summary that occur in human reference. Later on, Jing et al. (1998) extended Recall and Precision to operate in multi-reference settings.

  > **Limitation**: these metrics counting sentence-level matches cannot recognize the extract word matches in system and reference pairs. Consider the following example in which we underline the exact matches on word level.

  > **Example**
  >
  > - **System output**: The German Johannes Gutenberg introduced printing to Europe, whose invention allowed for the production of printed books.
  > - **Human reference**: Printing was introduced by the German Johannes Gutenberg, to Europe, which allowed the production of printed books and open circulation of information.
  >
  > Given the lack of sentence-level matches, *Precision* and *Recall* falsely assign a score of 0 to the above system summary, despite the fact that system and human summaries express similar meanings.

- *BLEU* (Papineni et al., 2002) is a precision-based valuation metric, which operates *Precision* on word and phrase levels in reference-based scenarios, and can be parametrized with the window size of words, termed n-gram. In particular, BLEU replaces the sentence-level matches with the n-gram matches

between system output and human reference. Recent research extended upon the idea of BLEU from multiple perspectives. Post (2018) proposed a standardized metric, the so-called sacreBLEU, which does not allow users to adjust the metric configuration, such as the window size, the scheme of tokenization, and so forth, as recommended in the annual Conference on Machine Translation (WMT). Galley et al. (2015) introduced △-BLEU, which rewards n-gram matches between system output and high-quality reference, and penalizes the matches with low-quality reference.

- *NIST* (Doddington, 2002) is a precision- and reference-based evaluation metric, which counts the n-gram matches in system output and human reference, much like BLEU. However, NIST weights each n-gram according to its information gain on human reference, which rewards n-gram matches for which the n-gram is deemed rare, and as such decreases the chance of manipulating the metric with unimportant n-gram.

- *ROUGE* (Lin, 2004) is a recall-based evaluation metric, which operates *Recall* on word and phrase levels in reference-based scenarios, and can be parametrized with the window size of words and the weighting scheme. For instance, ROUGE-N counts the n-gram matches between system output and human reference, much like BLEU-N. ROUGE-L measures the longest common subsequence (LCS) between system and reference pairs. ROUGE-W weights LCS matches according to the consecutiveness of these matches.

- *CHRF* (Popović, 2015) is a n-gram based evaluation metric operating on character level, which does not require tokenization, but directly counts character n-gram matches in system output and human reference, particularly useful when system outputs are in morphologically rich languages. Popović (2017) proposed CHRF+ considering n-gram matches on both word and character levels.

- *CIDEr* (Vedantam et al., 2015) is a weighted n-gram based evaluation metric for image captioning evaluation, which employs "term-frequency and inverse-document-frequency" (TF-IDF) to weight each n-gram matches according to its frequency not only in human reference, but also in the entire corpus consisting of all references over images. In particular, CIDEr is in favor of rare n-grams relevant to an image, and penalizes n-grams with high occurrences in the corpus.

- *Word Error Rate (WER)* (Su et al., 1992) is reference-based evaluation metric, which concerns the edits, such as the number of insertions, deletions, substitutions, and so forth, required for correcting a system output. Research showed that *WER* excessively punishes word order displacement, i.e., that correcting a misplaced word requires a two-step correction: a deletion followed by an insertion. As such, *WER* does not allow for high-quality system output with different word ordering from human reference. To this end, Translation Edit Rate (TER) (Snover et al., 2006) introduced one-step correction to address word order difference. Position-independent Edit Rate (PER) (Tillmann

et al., 1997) proposed to base the edits on the alignment of word pairs in system output and human reference.

Limitation: as Lavie and Agarwal (2007) state, lexical metrics, operating on surface level, have failed to recognize semantic similarity between system output and human reference in the absence of exact n-gram matches. Consider the following example in which the exact matches are underlined:

Example

- **System output**: This makes an increase in immigration unavoidable.
- **Human reference**: As a result, immigration will increase.

Given only two exact matches seen in the above system and human translations, lexical metrics cannot properly judge translation quality, notwithstanding the identical meanings expressed in these two translations.

Semantic-based Metrics. Later on, evaluation metrics focusing on soft lexical matching emerged, which address lexical similarity with (i) synonym matching and (ii) broader word relatedness in vector space. We outline these metrics as follows:

- *METEOR* (Lavie and Agarwal, 2007) is a reference-based evaluation metric, which allows for synonym matching in the absence of lexical overlap. In particular, METEOR, considering both precision and recall, carry outs a series of matching operations, including exact word matches, stemmed word matches, and synonym matches, over system and human pairs. Unlike n-gram based metrics, METEOR only considers unigram matches, but it rewards longer contiguous matches with a penalty term called "fragmentation penalty". Denkowski and Lavie (2010) proposed METEOR-NEXT, which computes weighted precision and recall by assigning different weights to different matching operations.

- *Word Mover Distance (WMD)* (Kusner et al., 2015) is a distance-based evaluation metric, which computes the semantic distance between system output and human reference by solving an optimization problem: given the Euclidean distances between word embeddings pertaining to the system and reference words, what is the minimum cost in order to transform the system words into the reference words. WMD allows for one-to-many word transformation, i.e., a word on one side can transform into multiple words on the other. Chow et al. (2019) extended WMD with a penalty term, which controls the punishment of word order difference.

- *ROUGE-WE* (Ng and Abrecht, 2015) is a reference-based metric, which extends upon the idea of ROUGE by leveraging pre-trained word2vec embeddings (Mikolov et al., 2013). ROUGE-WE computes the word relatedness in vector space, thereby addressing the challenge of operating ROUGE in the absence of lexical overlap between system and reference texts.

- *MEANT* (Lo and Wu, 2013) attributes the difference between system output and human reference into two aspects: structure and semantics. In particular, MEANT employs a semantic parser to predict the role of each word in system and reference texts, and then align role labels and role fillers with bipartite graph matching based on word embeddings. Lastly, MEANT computes F-score based on the matches of these role frames. XMEANT (Lo et al., 2014) extends MEAT to operate in reference-free scenarios. MEANT2.0 (Lo, 2017) weights each word with TF-IDF, aiming to reward content words and penalize function words.

- *Yisi* (Lo, 2019) is a parametrized evaluation metric, with parameters on the dependence of human reference and semantic parsers. YiSi-1 extends MEANT2.0 by replacing static word embeddings by recently proposed contextualized embeddings and making language-specific semantic parsers optional. In particular, YiSi-1 is a reference-based metric, which computes cosine similarity between system and reference texts on n-gram level based on contextualized embeddings, and optionally combines an additional semantic score based on the matches of role labels and role fillers. YiSi-2 operates YiSi-1 in reference-free scenarios, which compares system and source texts.

- *SIMILE* (Wieting et al., 2019) is a reference-based metric, which trains a sequence-to-sequence based language model on paraphrase text pairs extracted from ParaNMT, and then computes the cosine similarity of two sentence embeddings pertaining to system and reference texts. As in BLEU, SIMILE uses a length penalty (LP) term to penalize needless word repetition.

- *BERTr* (Mathur et al., 2020a) is a recall-based evaluation metric, which starts by assembling maximum cosine similarity scores between each word in human reference and each word in system output based on contextualized word embeddings, and then take the average of these similarity scores as the metric score. BERTr is based on the greedy one-to-one alignment (Rus and Lintean, 2012), which matches each word in human reference to the closest word in system output based on the cosine similarity between the embeddings of these words.

- *BERTScore* (Zhang* et al., 2020) is a reference-based evaluation metric, which operates BERTr in forward and reverse directions to obtain precision and recall scores from BERTr, aiming to ensure the symmetry of the metric.

- *MoverScore* (see Chapter 5) is a set-based evaluation metric as the extension of WMD, which computes the similarity between two sets of word embeddings corresponding to system output and human reference by solving an optimization problem, as it is the case for WMD. Much unlike BERTScore, MoverScore allows for one-to-many word matching, termed soft word alignment. XMoverScore extends MoverScore to operate in reference-free scenarios (see Chapter 8).

- *SUPERT* (Gao et al., 2020) is a reference-free evaluation metric, which em-

ploys graph-based approaches to extract pseudo reference summary from source document, and compares system summary with the extracted pseudo reference using MoverScore and contextualized representations.

> **Limitation**: research showed that recent semantic-based metrics, such as BERTScore and MoverScore, building on contextualized representations cannot recognize text coherence, and fail to punish the incoherent elements in system outputs (Fabbri et al., 2020). This is not surprising as language models responsible for producing contextualized embeddings mostly do not consider the interdependence between sentences, such as discourse phenomena in the inter-sentence context.

Discourse-based Metrics. We now outline early discourse metrics, as well as popular coherence models treated as metrics, which were initially proposed to judge text coherence in discourse tasks.

- *RC* and *LC* (Wong and Kit, 2012) are reference-free, discourse metrics, which require neither source texts nor references and use lexical cohesion devices (e.g., repetition) within hypothesis to predict text coherence. LC computes the proportion of words within hypothesis that are lexical cohesion devices, while RC computes the proportion of times that lexical cohesion devices appear in hypothesis.

- *Entity Graph* (Guinaudeau and Strube, 2013) and Lexical Graph (Mesgar and Strube, 2016) are popular coherence models known to perform discourse tasks such as essay scoring, both of which introduce a graph with nodes as sentences and adjacency matrices as the connectivity between sentences. Here, we use the average of adjacency matrices from the hypothesis as the proxy of hypothesis coherence. While Entity Graph draws an edge between two sentences if both sentences have at least one noun in common, Lexical Graph draws an edge if two sentences have a pair of similar words in common, i.e., the cosine similarity between their embeddings greater than a threshold.

- *Lexical Chain* (Gong et al., 2015) is a reference-based discourse metric, which extracts multiple lexical chains from hypothesis and reference. Each word is associated to a lexical chain if a word appears in more than one sentence. A lexical chain contains a set of sentence positions in which a word appears. Finally, the metric performs soft matching to measure lexical chain overlap between hypothesis and reference.

- *DiscoScore* (see Chapter 6) is a reference-based, parametrized discourse metric, which builds upon contextualized representations, and models discourse coherence through the lens of readers' focus, driven by Centering theory. DiscoScore can be parameterized with the choices of focus modeling: (i) modeling the frequency and semantics of foci, and compare the difference of foci in hypothesis and reference and (ii) employing focus transitions over sentences to model the interdependence between sentences.

- *KoBE* (Gekhman et al., 2020) is a reference-free, discourse metric, which counts the exact matches on entity level between system translations and source language text. KoBE is inspired by the fact that a coreferent entity is linked to a set of referring, language-agnostic expressions, and it addresses entity matches in two steps: (i) resorting to a large-scale multilingual knowledge base to extract entity mentions from system translation and source language text, and (ii) employing Recall to measure the ratio of entity matches and the number of entities in source text.

Factual-based Metrics. Apart from the inability to recognize discourse coherence, semantic-based metrics also fail to recognize factuality, and as such cannot punish the factual errors in system output (Maynez et al., 2020; Fabbri et al., 2020). Recently factual-based evaluation metrics have been proposed to address factual consistency evaluation. We now outline these metrics as follows:

- *ESTIME* (Vasilyev and Bohannon, 2021) is a reference-free, factual-based evaluation metric, which judges the factuality of system output by comparing it with source text. In particular, ESTIME extracts word pairs in system and source texts according to the word relatedness in vector space, and then counts the number of word pairs for which the corresponding words are not identical, which reflects factual inconsistency.

- *DAE* (Goyal and Durrett, 2020) is a reference-free evaluation metric, which judges the text quality of system output in the form of textual entailment. In particular, DAE begins with employing Stanford CoreNLP (Manning et al., 2014) to extract the dependence tree from system output. Each arc in the tree describes a relationship over two words in text. DAE predicts the probability that the relationships are the logically necessary consequences of source text.

- *FactCC* (Kryscinski et al., 2020) is a reference-free, entailment-based evaluation metric, which finetunes RoBERTa (Liu et al., 2019) on synthetic, textual entailment data, and then employs the finetuned RoBERTa to produce the probability that individual sentences in system output are factual consistent to source text, and finally extracts a span of source text as justification to the model prediction.

- *FEQA* (Durmus et al., 2020) is a reference-free evaluation metric, which judges the factuality of system output in the form of question answering. FEQA starts by producing the questions from system output, and then employs QA models to extract the corresponding answers from source text and system output. Non-matched answers on both sides reflect the factual inconsistency of system output.

Generation-based Metrics. Much unlike previous metrics, generation-based metrics judge the text quality of system output in the form of text generation, based on pre-trained language models. We discuss these metrics as follows:

- *BLANC* (Vasilyev et al., 2020) is a reference-free metric concerning to what

degree system output can assist language models such as BERT in performing cloze-test tasks. In particular, given the access to system output, BLANC asks BERT to fill in the masked-out tokens in source text. The more correct words fill in the masked places, the more system output is deemed informative.

- *IBM1* (Popović et al., 2011) is a reference-based metric judging translation quality, which trains a bag-of-word translation model on machine translation corpora, and then computes the likelihood that system output is the translation of source language text.

- *PRISM* (Thompson and Post, 2020) is a reference-based metric, which trains a sequence-to-sequence language model on paraphrased text pairs, and then employs the language model to compute the likelihood that system output and human reference are paraphrases.

- *BARTScore* (Yuan et al., 2021) is a parameterized evaluation metric, with parameters regarding the dependence on human reference or on source text. In reference-based scenarios, BARTScore finetunes BART on CNN/Daily-Mail (Hermann et al., 2015), and then employs the finetuned BART to measure how likely system output and reference are paraphrased according to the probability of one given the other, as it is the case for PRISM. In reference-free scenarios, BARTScore employs the finetuned BART to measure the likelihood that system output and source text are relevant.

2.2.2 Evaluation with Supervision

As evaluation metrics aim for imitating the behavior of human raters, supervised metrics have been studied, which builds on regression models trained with human ratings as supervision. Often, these metrics can be distinguished by the inputs arguments they take: (i) feature-based: heuristic features, such as the evaluation scores obtained from unsupervised metrics and linguistic features extracted from text, and (ii) end-to-end: system output and reference (or source text) without the need for any features. We outline these metrics as follows:

- *BEER* (Stanojević and Sima'an, 2014) is a feature-based evaluation metric trained on machine translation corpora with human rated translation quality as supervision, which linearly combines statistical measures, such as precision and recall scores on both character and word levels, and the features derived from permutation trees (Zhang and Gildea, 2007) to consider word ordering in system output and human reference.

- *BLEND* (Ma et al., 2019) is a feature-based metric trained on machine translation corpora, which uses an SVM regression model to combine metric scores from a total of 57 evaluation metrics, falling under three categories: lexical metrics such as BLEU and ROUGE, syntactic metrics from the Asiya toolkit (Giménez and Marquez, 2010), and discourse metrics focusing on name entities, semantic roles and discourse representation.

- *NNEval* (Sharif et al., 2018) is a feature-based metric trained on image captioning corpora, which uses a multi-layer neural network to combine lexical and semantic based metrics, and produces the probability that system caption exhibits quality levels similar to human.

- *ESIM* (Chen et al., 2017) is a sequence-to-sequence BiLSTM model initially proposed to perform natural language inference, which has been recently adapted to predicting translation quality. Mathur et al. (2019) introduced an end-to-end, supervised metric based on ESIM, which uses ESIM to encode sentence embeddings of system and reference texts, and then trains a neural model on these embeddings to rate system translations.

- *RUSE* (Shimanaka et al., 2018) is an end-to-end evaluation metric trained on machine translation corpora, which removes the need for evaluation metrics as input features, and uses a neural regression model to predict translation quality based on a combination of three pre-trained sentence embeddings: InferSent (Conneau et al., 2017), Quick-Thought (Logeswaran and Lee, 2018), and Universal Sentence Encoder (Cer et al., 2018).

- *BLEURT* (Sellam et al., 2020) is an end-to-end, supervised evaluation metric predicting translation quality based on finetuned BERT embeddings. BLEURT finetunes BERT on synthetic sentence pairs and machine translation corpora in succession, in which the synthetic pairs are produced by perturbing Wikipedia sentences via (i) mask-filling; (ii) back-translation and (iii) randomly dropping words. Pu et al. (2021) extended BLEURT with RemBERT to operate in reference-free scenarios.

- *NUBIA* (Kane et al., 2020) is an end-to-end, supervised evaluation metric trained on machine translation (or image captioning) corpora, which dissects the assessment of text quality into three sub-tasks: (i) NUBIA finetunes RoBERTa (Liu et al., 2019) embeddings on STS corpora (Cer et al., 2017) to predict the sentence similarity between system and reference texts; (i) NUBIA finetunes RoBERTa embeddings on MNLI corpora (Wang et al., 2018) to infer the relationship between system and reference texts; and (iii) NUBIA uses GPT-2 (Radford et al., 2018) to rate system output, which produces perplexity score reflecting text fluency. Lastly, NUBIA uses a neural regression model to aggregate these three scores.

- *COMET* (Rei et al., 2020) is an end-to-end, reference-free evaluation metric trained on machine translation corpora, which builds on XLM-RoBERTa and can be parametrized with the form of human ratings as supervision: (i) COMET-DA uses Direct Assessment as supervision, (ii) COME-HTER uses Human Translation Edit Rate (HTER) and (iii) COME-HTER uses Multi-dimensional Quality Metrics and (iv) COMET-Rank uses Relative Ranking. Bhosale et al. (2020) extended COMET with a triplet of input arguments considering source, system and reference texts.

> **Limitation**: given the poor generalization being the fundamental issue in supervised training, supervised evaluation metrics trained towards one NLG task generalize poorly towards another, and thus are often termed task-specific metrics.

2.2.3 Meta-Evaluation

As Lin and Hovy (2002) state, evaluation metrics aim for correlating highly, positively and consistently with human judgment of text quality. As a consequence, performance gains by metrics can truly reflect improved text quality of system outputs. We list the following common correlation measures, which concern a meta-level evaluation for evaluation metrics as opposed to evaluating system outputs.

- *Spearman's* r quantifies the monotonic relationship between metric and human scores assigned to system outputs.
- *Pearson's* ρ, on the other hand, quantifies the linear relationship between metric and human scores.
- *Kendall's* τ quantifies the ordinal association, i.e., the relationship between the rankings pertaining to metric and human scores.

In NLG, there appears to be a tradition carrying out meta-evaluation for evaluation metrics on two levels: system and instance levels as follows:

- *System Level*: as for all tasks in NLG, evaluation metrics are responsible for comparing systems, i.e., affirming the ranking of these systems. In this context, system-level correlation between judgments by humans and by metrics aims for quantifying how much the ranking of the systems is trustworthy. This setup has to assemble system-level metric and human scores—both are the average of instance-level scores weighted by the number of references.
- *Instance Level*: in order to complement meta-evaluation on system level, instance-level correlation between evaluation metrics and human judgments is introduced, which aims for investigating metric behaviors on individual cases, such as the extent to which evaluation metrics discriminate system outputs of varying text qualities, and in which cases metrics misjudge text quality by assigning high scores to low-quality texts and vice versa.

2.3 Our Contributions

2.3.1 Reference-based Evaluation

Evaluation metrics, as low-cost alternatives to human evaluation, has become the standard for evaluating the performance of text generation systems. The major, if not utmost, concern in evaluation metrics is the magnitude that these metrics correlate with human ratings, as the higher the correlations are, the more performance

gains by metrics can genuinely reflect improved text quality of system outputs. However, previous n-gram based metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which have been extensively employed in recent NLG evaluation (Ma et al., 2018, 2019), cannot recognize semantic similarity when system output and human reference have no lexical overlap. Building upon these insights, we present the research question as follows:

> **RQ1. What are essential elements for reference-based metrics to be high quality in the absence of lexical overlap?**

In Chapter 5, aiming to design high-quality metrics in the absence of lexical overlap, we propose MoverScore for system output assessment, which addresses lexical overlap with semantic relatedness of words in vector space derived from recently proposed contextualized representations. MoverScore consists of four essential elements: (i) leveraging strong contextualized encoders such as BERT (Devlin et al., 2018); (ii) employing Earth Mover Distance to compute the similarity between system output and reference, based on BERT embeddings; (iii) aggregating the word embeddings across the layers of BERT and (iv) finetuning the embeddings on Natural Language Inference and Paraphrase corpora. We show that MoverScore strongly correlates with human assessment in machine translation, summarization and image captioning, surpassing BLEU by up to 25 correlation points.

However, MoverScore and other recently proposed metrics based on BERT, known as sentence-level metrics, cannot recognize coherence and fail to punish incoherent elements in system outputs, and as such are inadequate to evaluate long texts. We present the following research question.

> **RQ2. How to design reference-based metrics targeting the assessment of text coherence for evaluating long texts?**

In Chapter 6, we begin with investigating the extent to which recent BERT-based evaluation metrics can recognize text coherence. We show that these metrics correlate much worse with human rated coherence than early discourse metrics such as RC and LC (Wong and Kit, 2012), invented a decade ago. To this end, we propose DiscoScore targeting text coherence assessment, which uses BERT to model discourse coherence through the lens of readers' focus of attention, driven by Centering theory (Grosz et al., 1995). We show that DiscoScore achieves strong system-level correlation with human ratings, not only in coherence but also in factual consistency and other aspects.

Aggregation of embeddings. Contextualized word embeddings have been shown to be responsible for the successful stories of recent BERT-based metrics. However, the selection of word embeddings can be challenging, as contextualized encoders produce considerably different embeddings across the layers of the encoders. Research addressed this issue by either selecting the word embeddings from an intermediate layer or linearly combining word embeddings across layers with task-dependent supervisions (Tenney et al., 2019; Liu et al., 2018, 2019)

In Section 5.3.2, we apply the KDE routing to the aggregation of word embeddings across the layers of contextualized encoders for our evaluation metrics. In Chapter 7, we elaborate on the KDE routing, a kernel-based density estimator, used to aggregate capsules (another form of embeddings) across layers without supervision. The routing process uses an adaptive optimizer for adjusting the number of times to perform aggregation.

2.3.2 Reference-free Evaluation

While recent BERT-based evaluation metrics achieve strong correlation with human judgment of text quality, the majority of these metrics require costly human references, and thus are limited to language pairs with available parallel data (source and reference pairs), especially in machine translation evaluation. In contrast, reference-free metrics removes the dependence of human references by directly comparing system translations with source language texts, aiming for unlimited, web-scale evaluation of natural language generation (NLG) systems. In this context, we present the following research question:

> **RQ3. What are essential elements for reference-free metrics to be high quality in the absence of human references?**

Indeed, each NLG task has its own challenges driven by the inherent characteristics of the task. In this thesis, we particularly address reference-free evaluation in machine translation (MT). In Chapter 8, we outline the challenges stemming from the absence of human references in MT:

Resource and typology disparities across languages. We extend MoverScore to operate in reference-free setups, which compares system translations with source language texts based on their BERT-based, multilingual embeddings. However, these embeddings (responsible for high-quality reference-based metrics) could not show advantages in a cross-lingual setup. This is because these embeddings exhibit a strong language bias with the qualities considerably different across languages. This issue results in the poor qualities of reference-free metrics with the translations to be assessed in low-resource languages and in languages dissimilar to the source. Besides, we show that these metrics have the inability to punish "translationese", i.e., low-quality word-by-word translations.

To address this issue, we introduce XMoverScore, a parameterized metric to address reference-free MT evaluation. XMoverScore can be parameterized with the choice of solutions to rectifying the vector space of different languages—which results in high-quality cross-lingual embeddings. In the following, we briefly outline these solutions:

In Chapter 8, we propose two supervised approaches with the one rotating the vector space, with another reducing "language bias" by subtracting embeddings from a language-dependent bias vector. Besides, we couple XMoverScore with a target-side language modeling to penalize unnatural word-by-word translations. XMover-

Score surpasses reference-based BLEU, constituting the new frontier in reference-free machine translation evaluation.

In Chapter 9, we propose three approaches to rectify the vector space: (i) re-mapping the vector space of target languages (all together) to a pivot source language with large parallel data as supervision; (ii) normalizing vector space by removing language-specific means and variances, which yields better discriminativeness of embeddings as a by-product and (iii) normalizing input texts by removing morphological contractions and sentence reordering, which aims to increase typological similarity across languages. Our findings are manifold: First, normalizing vector spaces is surprisingly effective, rivals much more resource-intensive approaches such as re-mapping, and leads to more consistent gains. Second, text normalization yields benefits in the setups where language-dependent, linguistic phenomena do exist in input texts. Lastly, the three approaches are orthogonal and their gains often stack.

In Chapter 10, we propose weakly supervised and unsupervised approaches to re-map the vector space of different languages, which remove the need for large parallel data, thereby addressing the data scarcity issue for low-resource languages. We expose the two limitations of previous resource-intensive approaches: (i) the inability to sufficiently leverage data and (ii) these approaches are not trained properly. To address these issues, we introduce weakly supervised and unsupervised density-based approaches, which dissect the re-mapping of vector space into density matching and density modeling. Besides, we propose two validation criteria to guide both supervised and unsupervised training. Our experiments demonstrate the effectiveness of our approaches in the scenarios of limited and no parallel data. First, our supervised approach trained on 20k parallel data mostly surpasses previous resource-intensive approaches trained on much larger parallel data. Second, parallel data can be removed without sacrificing performance when integrating our unsupervised approach in our bootstrapping procedure, which is theoretically motivated to enforce equality of multilingual subspaces.

Chapter 3

System Comparison

System comparison, aiming for rigorously affirming one system over another, plays an essential role, as evaluation metrics do, in NLG evaluation. For instance, high-quality evaluation metrics could misjudge the state-of-the-art system when systems are not compared in a rigorous manner, such as employing inappropriate significance tests. Few work has surveyed the methodologies responsible for ensuring unbiased and trustworthy comparison results. Here, we dissect methodologies into multiple classes (see Figure 3.1). We now discuss them under each class.

- Statistics for the trustworthiness of results, via appropriate significance tests.
- Comparing score distributions instead of single-point estimates.
- Comparing systems under a given computational budget.
- Reporting consistent evaluation results with parameterized evaluation metrics.

3.1 Significance Testing

Significance testing is a long-lasting topic in statistics, which concerns the possibility that experimental results are coincidental. As for all tasks in NLP, when the performances of two systems only differ by a small amount according to evaluation metrics, researchers resort to report significance results in order to justify the superiority of one system over the other. However, research showed that results can be misleading when employing inappropriate significance tests or carrying out them incorrectly (Simpson, 2021). In statistics, significance tests fall under two categories: parametric and non-parametric tests.

For (i) parametric tests, evaluation scores have to follow a well-known distribution with predefined parameters. A popular example was given by Student's t-test (Fisher, 1935), which assumes that evaluation scores follow a Gaussian distribution, and computes the difference between two population means of evaluation scores pertaining to two systems. For (ii) non-parametric tests, no assumption is

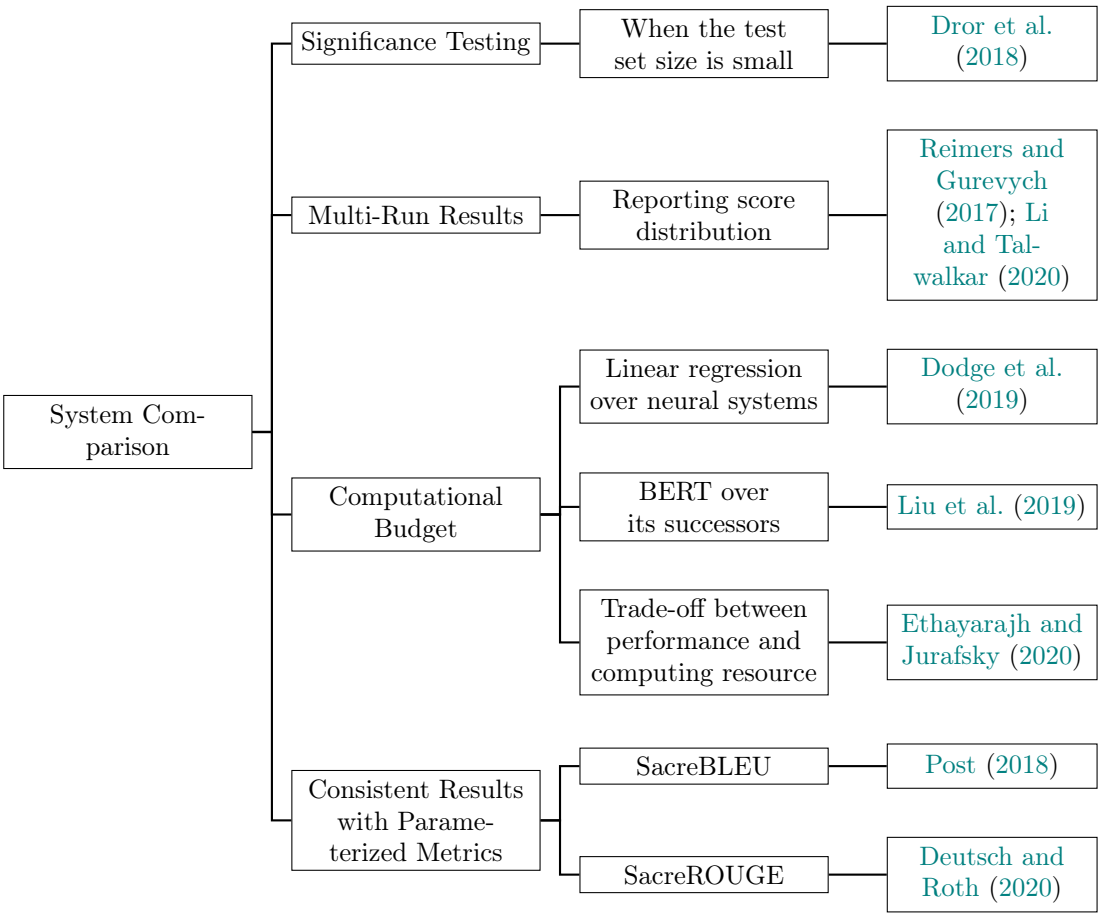


Figure 3.1: Classification of methodologies for rigorous system comparison.

required. For instance, Sign Test computes the percent of instances for which one system surpasses the other. Wilcoxon Signed Rank Test (Wilcoxon, 1945) assembles the score differences of two pairs of evaluation scores, and justifies whether the distribution of the score differences is symmetric around zero.

While parametric tests have much stronger statistical power than non-parametric counterparts, the data assumption of parametric tests cannot be justified. Accordingly, Dror et al. (2018) suggested to use non-parametric tests for reporting significance results when the test set size is small, and provided examples of performing these tests in a proper manner (Collins et al., 2005; Chan et al., 2007; Rush et al., 2012).

3.2 Reporting Multi-Run Results

The NLP community has been entirely in favor of neural systems for long. However, given the randomness in weight initialization, data shuffling and dropout techniques, the results of neural systems are not reproducible (Zhuang et al., 2021). As an example, after carrying out a randomization study by testing 86 seed values, Reimers and Gurevych (2017) showed that the choice of random seed value can result in the significant difference ($p < 10^{-4}$) between the best and worse performance of

a state-of-the-art NLP system. To this end, they recommended comparing score distributions assembled from the results of multiple runs. Li and Talwalkar (2020) pointed out the issue that, given the randomness of neural networks, it is challenging for a state-of-the-art neural system to reproduce the results in different computing environments. Thus, they recommended reporting score distribution instead of single-run performance.

3.3 Not Forgetting Computational Budget

Research showed that a large computational budget, such as the training time, computing resources, the parameter size and data size, sometimes outranks advanced techniques in system performance. For instance, an inferior linear regression system can outperform a superior neural system when restricting the parameter sizes of the two systems to be equal (Dodge et al., 2019). In a second study, Liu et al. (2019) showed that the BERT (Devlin et al., 2018) language model endowed with larger computational budget (e.g., involving more training data) can surpass the subsequent, more advanced language models such as XLNet (Yang et al., 2019). Moreover, as Ethayarajh and Jurafsky (2020) state, much unlike performance-centralized leaderboards, NLP practitioners concern the trade-off between system performance and computing resources. Given these three examples, it is evident that comparing NLG systems under a given computational budget is crucial for fair and energy-efficient comparison.

3.4 Evaluation Metrics are Parameterized

Recent research acknowledged that evaluation metrics can be parametrized by the choice of hyperparameters and the text preprocessing scheme. Accordingly, evaluation scores pertaining to a metric vary by a large amount under different parameter configurations. For instance, as Post (2018) states, BLEU is a parameterized metric, consisting of five parameters: (i) the number of references used, (ii) the length penalty, (iii) the maximum n-gram length, (iv) smoothing for zero n-gram overlaps, and (v) the text preprocessing scheme, which concerns tokenization, splitting compound words, removing stopwords and special characters, and so forth. All these factors play vital roles in metric scores. For that reason, parameterized metrics, if not handled properly, could yield biased evaluation results of NLG systems. To this end, SacreBLEU (Post, 2018) and SacreROUGE (Deutsch and Roth, 2020) emerged, which propose to use official evaluation scripts to compute BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores, i.e., preventing users from adjusting the configuration of parameters.

3.5 Our Contributions

Evaluation metrics and system comparison are interrelated, and both affect evaluation results. For instance, if text generation systems on leaderboards are not

rigorously compared, then evaluation metrics can misjudge the state-of-the-art system. Oftentimes, NLG systems are compared on a common set of test instances, and ranked in a way according to the *average* of instance-level evaluation scores. However, we argue *average* is notoriously sensitive to outliers, which undermines the reliability of comparison results. We now present the research question concerning rigorous system comparison:

> **RQ4. What is the rigorous comparison approach in order for leaderboards to report correct system rankings?**

In Chapter 11, we show global statistics such as *average* and *median* cannot carry out rigorous comparison, as they ignore the fact that systems are evaluated on the same test set. We propose pairwise comparison approaches based on the Bradley Terry (BT) model (Bradley and Terry, 1952) to compare systems in pairs on instance level, which base the prediction of system strengths on the probability of one system over the other on the same test set. We comparably evaluate our approaches, *average* and *median* across four text generation tasks and 18 evaluation metrics, and show that they yield different conclusions as to which systems are state of the art in about 30% of the setups. Lastly, we employ our approaches to perform paired evaluation of Eval4NLP shared tasks, and expose the limitation of such paired evaluation: it could fail to show a clear winner, which is in line with the Arrow's impossibility theorem (Arrow, 1950).

Chapter 4

Explainability

4.1 From Artificial Intelligence to Evaluation Metrics

Given the extraordinary success of deep learning technologies, artificial intelligence has engaged steadily in the human decision making and aims to make decisions on the behalf of human experts in the long run. As Ridgeway et al. (1998) state, human decisions are interpretable and logically defensible, which underpins the importance of explainable artificial intelligence providing justifications to ensure legitimate model decisions. Further, when in cooperation with artificial intelligence, humans cannot make critical decisions without the understanding of model predictions (Lipton, 2018). For instance, to carry out a medical diagnosis, doctors demand more information from a model than binary predictions (Tjoa and Guan, 2020).

Arrieta et al. (2020) dissected the goal of explainable artificial intelligence into multiple essential characteristics, such as Trustworthiness, Causality, Transferability, Ethical Concerns, and so forth. In the following, we briefly outline two of these characteristics, and discuss them in the context of NLG evaluation metrics.

Trustworthiness. According to (Lipton, 2018), trustworthiness is not a matter of model performance, but rather concerns the confidence that a model makes a correct prediction, which results in research towards model selection in artificial intelligence. For instance, when making crucial decisions, humans are shown to be in favor of the trade-off between model performance and model trustworthiness, instead of the state-of-the-art model (Došilović et al., 2018).

> **NLG Evaluation**: the confidence of text quality judgment is not yet a property of off-the-shelf evaluation metrics. This poses a problem, as knowing the confidence of judgments made on individual cases with evaluation metrics can be crucial for ensuring unbiased evaluation, such as excluding low-confidence judgments in the comparison of NLG systems.

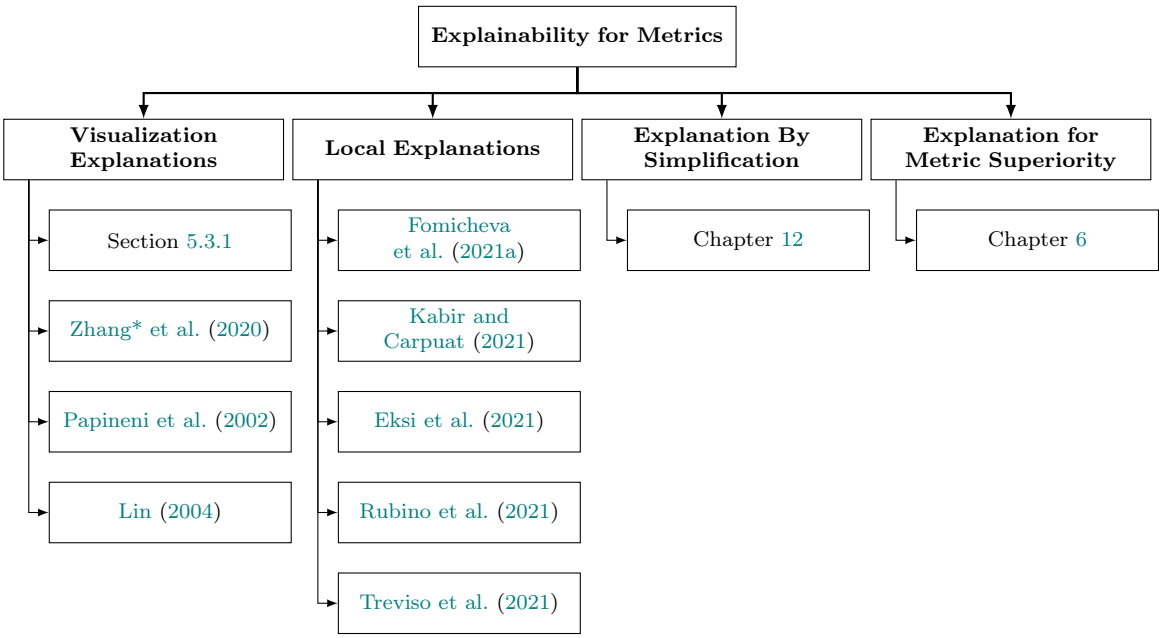


Figure 4.1: Taxonomy of explainability techniques for evaluation metrics.

Ethical Consideration. Humans with different cultures and traditions exhibit cognitive prejudices for individuals in minoritized groups, such as gender bias in career choice, and religion bias in violence. As a consequence, these prejudices are reflected in human-produced data, and then influence the model decisions given by data-driven artificial intelligence (Lauscher, 2021). For that reason, the European Parliament released the Data Protection Regularization, aiming to make artificial intelligence conform to ethical standards, and foster explainable artificial intelligence providing explanations to individuals who affected by model decisions (Goodman and Flaxman, 2017).

> **NLG Evaluation**: much like artificial intelligence, ethical concerns also occur in NLG evaluation. For instance, when reference-free metrics are used to judge system translations, they aim to carry out unbiased judgments for translations in *any* language, and as such underpin the inclusion and democratization of NLG technologies. However, given resource and typology disparities of languages, recent metrics based on multilingual representations exhibit a strong language bias, resulting in the different qualities of these metrics across languages (see Chapter 9).

4.2 Post-hoc Explainability Techniques

Transparent and Non-transparent Models Modern artificial intelligence began with statistical machine learning, which studies statistics-driven algorithms improved through experience, i.e., the number of training examples that algorithms have seen (Mitchell, 1997). Popular examples are Linear Regression, Decision Trees, and Bayesian Models, all of which can directly explain the model decisions with

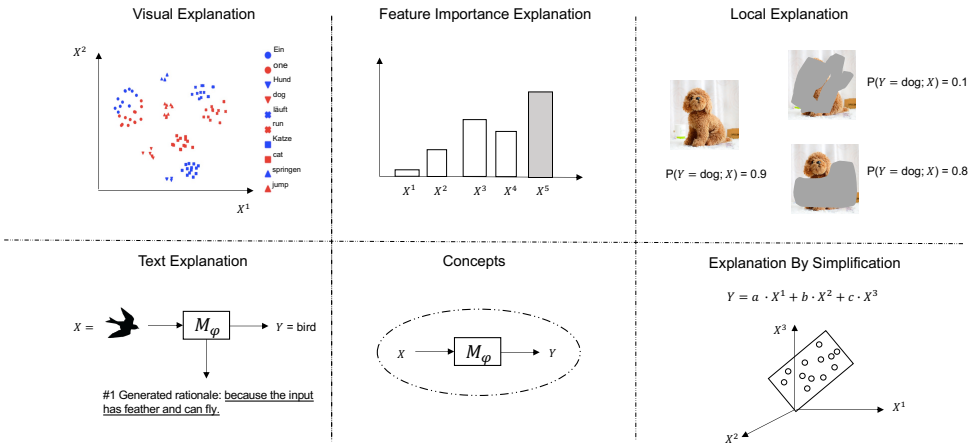


Figure 4.2: Classification of explainability techniques for artificial intelligence.

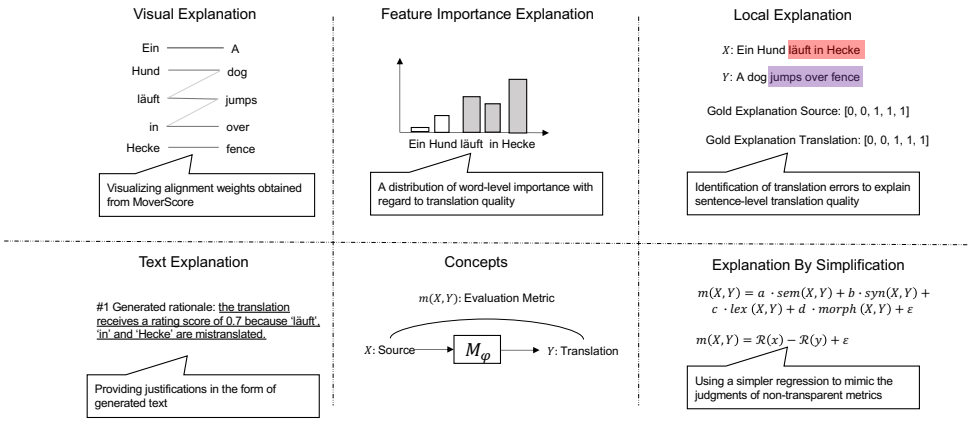


Figure 4.3: Classification of explainability techniques for evaluation metrics.

simple ideas such as feature importance, hierarchical structure and conditional probability, thereby placed under the umbrella of transparent models. By the time large amount of data and computing resources came to be accessible, artificial intelligence entered into the new frontier of machine learning, i.e., data-driven deep learning, which trains neural networks on given inputs to perform a specific task. Hornik et al. (1990) proved that any task being a function of inputs can be well-performed by neural networks. However, this comes with the cost of complexity due to huge parametric space, such as dozens of layers and millions of parameters. As such, neural networks are overcomplicated for humans to comprehend, thereby termed blackbox models. For that reason, explainability techniques emerged.

Arrieta et al. (2020) dissected explainability techniques for artificial intelligence into multiple classes. We follow Arrieta et al. (2020) and adopt this classification (see Figure 4.2), as the authors managed to distinguish over 160 techniques under this umbrella. In the following, we start by briefly outlining each of the classes, and discuss them in the context of evaluation metrics. Figure 4.1 shows the taxonomy of explainability techniques for evaluation metrics.

- *Explanation by Simplification* is achieved by training simpler models, such as Linear Regression and Decision Tree, aiming to mimic the outputs of complex

models. Given the transparency nature of simpler models, they can be employed to explain the process of non-transparent models. Zhang et al. (2019) presented an approach to interpret convolutional neural networks, which employs decision tree to decompose feature representations within convolutional layers, and serves as rule-based rationales providing information on model predictions.

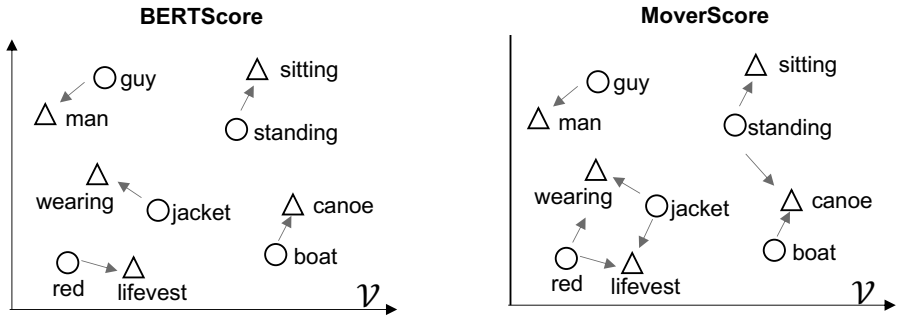
> **NLG Evaluation**: in order to comprehend and mimic the judgments obtained from recent non-transparent metrics based on blackbox language models, we employ linear regression to dissect metric scores into four linguistic factors, including semantics, syntax, morphology, and lexical overlap (see Chapter 12).

- *Visual Explainations* provide understanding on the level of model parameters, or on the level of input representations encoded by models—shown to be effective for identifying model limitations. For instance, research demonstrated the misalignment of multilingual subspaces by visualizing multilingual embeddings in a shared vector space with dimensionality reduction techniques (Cao et al., 2020). Other work showed that the visualization of self-attention weights of language models can interpret word relations, such as a pronoun and its antecedent.

> **NLG Evaluation**: there are several transparent and semi-transparent evaluation metrics, which we can visualize their process of text quality judgment. For instance, BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) counting word overlaps between system output and human reference are transparent metrics. MoverScore (see Chapter 5) and BERTScore (Zhang* et al., 2020) based on blackbox language models are semi-transparent, as these metrics can align word pairs in system output and human reference—which are indicative of their judgments. Figure 4.4 shows word overlap and word alignment.

- *Feature Importance Explanations* are rationales in the form of importance distribution, i.e., how important each input feature is for model output. For instance, given the input text *"I like your website"* and the output *{positive}* as a gold label, the rationale is said to be a distribution of word-level importance with *'like'* at peak.

> **NLG Evaluation**: we organized a shared task at Eval4NLP21, which concerns the probability distribution of words being mistranslated in system translation used to explain human judgment of translation quality (Fomicheva et al., 2021b). For instance, given a system translation *"this is not a dog"* and a source text *"Das ist ein Hund"*, the desired rationale is a probability distribution with *'not'* at peak on system translation, accounting for translation errors.



System Output ○	Reference △	BERT	Mover	BLEU	ROUGE
A guy with a red jacket is standing on a boat	A man wearing a lifevest is sitting in a canoe	80.3	53.6	5.30	4.00

Figure 4.4: Visualization of word-level alignments marked by arrows in vector space and word overlaps underlined in table. BERT and Mover denote BERTScore and MoverScore.

- *Local Explainations* are concerned with the identification of certain features highly relevant to model output, which can be derived from a distribution of feature importance (see *Feature Importance Explanations*). Popular examples are LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), which identify prominent features through input perturbation. Concretely, when minimal changes to a feature results in a noticeable change in model output, the feature is deemed important.

> **NLG Evaluation**: the Eval4NLP21 shared task provided two baselines, one with SHAP and one with LIME, both of which identify prominent words termed translation errors through masking out words in system translation (Fomicheva et al., 2021a). Rubino et al. (2021) and Treviso et al. (2021) identified translation errors by placing attention mechanism over text embeddings pertaining to source and translation pairs. Eksi et al. (2021) investigated a class of gradient-based methods, which assigns gradients obtained from backpropagation to words in system translation, and then considers the words associated to high gradients as translation errors. Kabir and Carpuat (2021) presented an approach employing Divergence-mBERT (Briakou and Carpuat, 2020) to underline the words in system translation that are likely mistranslated.

- *Text Explanations* are rationales, either in the form of generated texts providing justifications to model decisions (Bennetot et al., 2019), or in the form of generated rules providing understanding on the process of decision making (Arrieta et al., 2020). Consider the following example:

Example

- **Generated text**: the news article classified as sports entertainment is because the keywords detected in the article are football teams and players.
- **Generated rule**: given the input x, the output is y because $x_1 > \alpha$ and $x_2 < \beta$.

NLG Evaluation: such rationales could be "*the translation receives a rating score of 0.7 because word A, word B and word C are mistranslated.*"

4.3 Our Contributions

Recent years have seen rapid advances in BERT-based evaluation metrics with much better qualities than traditional metrics such as BLEU for text quality evaluation. However, these metrics build upon black-box language models, which makes their judgments of text quality hardly understandable. Based on the insights of the explainable artificial intelligence literature, we present the following research question:

RQ5. What insights can be drawn from explainable artificial intelligence in order to understand non-transparent evaluation metrics?

In Section 5.3.1, we visualize the process of MoverScore by picturing the alignment of word pairs in system output and human reference. In Chapter 12, we propose a simple regression based explainability technique to dissect metric scores into semantics, syntax, morphology, and lexical overlap. We show that recent BERT-based metrics are similar to BLEU and ROUGE in a way that they all are substantially sensitive to lexical overlap. Accordingly, these metrics cannot discriminate semantically non-sensical word-by-word translations and paraphrases, which we show in an adversarial test scenario. In Chapter 6, we address the understanding of metric superiority as to why one metric outperforms another. In particular, we derive simple features from the inherent characteristics of non-transparent metrics, and show that these features are responsible for the performance gaps between the metrics. We find that the more discriminative the features are in separating system output from human reference, the better the metrics perform. This attributes the superiority of a metric to the fact that the feature can better separate hypothesis and reference.

Part II

Publications

Subpart A

Reference-based Evaluation

Chapter 5

MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance

MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance

Wei Zhao[†], **Maxime Peyrard**[†], **Fei Liu**[‡], **Yang Gao**[†], **Christian M. Meyer**[†], **Steffen Eger**[†]
[†] Computer Science Department, Technische Universität Darmstadt, Germany
[‡] Computer Science Department, University of Central Florida, US
zhao@aiphes.tu-darmstadt.de, maxime.peyrard@epfl.ch
feiliu@cs.ucf.edu, yang.gao@rhul.ac.uk
meyer@ukp.informatik.tu-darmstadt.de
eger@aiphes.tu-darmstadt.de

Abstract

A robust evaluation metric has a profound impact on the development of text generation systems. A desirable metric compares system output against references based on their semantics rather than surface forms. In this paper we investigate strategies to encode system and reference texts to devise a metric that shows a high correlation with human judgment of text quality. We validate our new metric, namely MoverScore, on a number of text generation tasks including summarization, machine translation, image captioning, and data-to-text generation, where the outputs are produced by a variety of neural and non-neural systems. Our findings suggest that metrics combining contextualized representations with a distance measure perform the best. Such metrics also demonstrate strong generalization capability across tasks. For ease-of-use we make our metrics available as web service.[1]

1 Introduction

The choice of evaluation metric has a significant impact on the assessed quality of natural language outputs generated by a system. A desirable metric assigns a single, real-valued score to the system output by comparing it with one or more reference texts for content matching. Many natural language generation (NLG) tasks can benefit from robust and unbiased evaluation, including text-to-text (*machine translation* and *summarization*), data-to-text (*response generation*), and image-to-text (*captioning*) (Gatt and Krahmer, 2018). Without proper evaluation, it can be difficult to judge on system competitiveness, hindering the development of advanced algorithms for text generation.

It is an increasingly pressing priority to develop better evaluation metrics given the recent advances in neural text generation. Neural models provide the flexibility to copy content from source text as well as generating unseen words (See et al., 2017). This aspect is hardly covered by existing metrics. With greater flexibility comes increased demand for unbiased evaluation. Diversity-promoting objectives make it possible to generate diverse natural language descriptions (Li et al., 2016; Wiseman et al., 2018). But standard evaluation metrics including BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) compute the scores based primarily on n-gram co-occurrence statistics, which are originally proposed for diagnostic evaluation of systems but not capable of evaluating text quality (Reiter, 2018), as they are not designed to measure if, and to what extent, the system and reference texts with distinct surface forms have conveyed the same meaning. Recent effort on the applicability of these metrics reveals that while compelling text generation system ascend on standard metrics, the text quality of system output is still hard to be improved (Böhm et al., 2019).

Our goal in this paper is to devise an automated evaluation metric assigning a single holistic score to any system-generated text by comparing it against human references for content matching. We posit that it is crucial to provide a holistic measure attaining high correlation with human judgments so that various neural and non-neural text generation systems can be compared directly. Intuitively, the metric assigns a perfect score to the system text if it conveys the same meaning as the reference text. Any deviation from the reference content can then lead to a reduced score, e.g., the system text contains more (or less) content than the reference, or the system produces ill-formed text that fails to deliver the intended meaning.

We investigate the effectiveness of a spectrum of distributional semantic representations to encode system and reference texts, allowing them to be compared for semantic similarity across

[1] Our code is publicly available at http://tiny.cc/vsqtbz

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 563–578,
Hong Kong, China, November 3–7, 2019.

multiple natural language generation tasks. Our new metric quantifies the semantic distance between system and reference texts by harnessing the power of contextualized representations (Peters et al., 2018; Devlin et al., 2018) and a powerful distance metric (Rubner et al., 2000) for better content matching. Our contributions can be summarized as follows:

- We formulate the problem of evaluating generation systems as measuring the semantic distance between system and reference texts, assuming powerful continuous representations can encode any type of semantic and syntactic deviations.
- We investigate the effectiveness of existing contextualized representations and Earth Mover's Distance (Rubner et al., 2000) for comparing system predictions and reference texts, leading to our new automated evaluation metric that achieves high correlation with human judgments of text quality.
- Our metric outperforms or performs comparably to strong baselines on four text generation tasks including summarization, machine translation, image captioning, and data-to-text generation, suggesting this is a promising direction moving forward.

2 Related Work

It is of fundamental importance to design evaluation metrics that can be applied to natural language generation tasks of similar nature, including summarization, machine translation, data-to-text generation, image captioning, and many others. All these tasks involve generating texts of sentence or paragraph length. The system texts are then compared with one or more reference texts of similar length for semantic matching, whose scores indicate how well the systems perform on each task. In the past decades, however, evaluation of these natural language generation tasks has largely been carried out independently within each area.

Summarization A dominant metric for summarization evaluation is ROUGE (Lin, 2004), which measures the degree of lexical overlap between a system summary and a set of reference summaries. Its variants consider overlap of unigrams (-1), bigrams (-2), unigrams and skip bigrams with a maximum gap of 4 words (-SU4), longest common subsequences (-L) and its weighted version (-W-1.2), among others. Metrics such as Pyramid (Nenkova and Passonneau, 2004) and BE (Hovy et al., 2006; Tratz and Hovy, 2008) further compute matches of content units, e.g., (head-word, modifier) tuples, that often need to be manually extracted from reference summaries. These metrics achieve good correlations with human judgments in the past. However, they are not general enough to account for the relatedness between abstractive summaries and their references, as a system abstract can convey the same meaning using different surface forms. Furthermore, large-scale summarization datasets such as CNN/Daily Mail (Hermann et al., 2015) and Newsroom (Grusky et al., 2018) use a *single reference* summary, making it harder to obtain unbiased results when only lexical overlap is considered during summary evaluation.

Machine Translation A number of metrics are commonly used in MT evaluation. Most of these metrics compare system and reference translations based on surface forms such as word/character n-gram overlaps and edit distance, but not the meanings they convey. BLEU (Papineni et al., 2002) is a precision metric measuring how well a system translation overlaps with human reference translations using n-gram co-occurrence statistics. Other metrics include SentBLEU, NIST, chrF, TER, WER, PER, CDER, and METEOR (Lavie and Agarwal, 2007) that are used and described in the WMT metrics shared task (Bojar et al., 2017; Ma et al., 2018). RUSE (Shimanaka et al., 2018) is a recent effort to improve MT evaluation by training sentence embeddings on large-scale data obtained in other tasks. Additionally, preprocessing reference texts is crucial in MT evaluation, e.g., normalization, tokenization, compound splitting, etc. If not handled properly, different preprocessing strategies can lead to inconsistent results using word-based metrics (Post, 2018).

Data-to-text Generation BLEU can be poorly suited to evaluating data-to-text systems such as dialogue response generation and image captioning. These systems are designed to generate texts with lexical and syntactic variation, communicating the same information in many different ways. BLEU and similar metrics tend to reward systems that use the same wording as reference texts, causing repetitive word usage that is deemed undesirable to humans (Liu et al., 2016). In a similar vein, evaluating the quality of image captions can be challenging. CIDEr (Vedantam et al., 2015) uses tf-idf weighted n-grams for similarity estimation; and SPICE (Anderson et al., 2016) incorporates

synonym matching over scene graphs. Novikova et al. (2017) examine a large number of word- and grammar-based metrics and demonstrate that they only weakly reflect human judgments of system outputs generated by data-driven, end-to-end natural language generation systems.

Metrics based on Continuous Representations Moving beyond traditional metrics, we envision a new generation of automated evaluation metrics comparing system and reference texts based on semantics rather than surface forms to achieve better correlation with human judgments. A number of previous studies exploit static word embeddings (Ng and Abrecht, 2015; Lo, 2017) and trained classifers (Peyrard et al., 2017; Shimanaka et al., 2018) to improve semantic similarity estimation, replacing lexical overlaps.

In contemporaneous work, Zhang et al. (2019) describe a method comparing system and reference texts for semantic similarity leveraging the BERT representations (Devlin et al., 2018), which can be viewed as a special case of our metrics and will be discussed in more depth later. More recently, Clark et al. (2019) present a semantic metric relying on sentence mover's similarity and the ELMo representations (Peters et al., 2018) and apply them to summarization and essay scoring. Mathur et al. (2019) introduce unsupervised and supervised metrics based on the BERT representations to improve MT evaluation, while Peyrard (2019a) provides a composite score combining redundancy, relevance and informativeness to improve summary evaluation.

In this paper, we seek to accurately measure the (dis)similarity between system and reference texts drawing inspiration from contextualized representations and Word Mover's Distance (WMD; Kusner et al., 2015). WMD finds the "traveling distance" of moving from the word frequency distribution of the system text to that of the reference, which is essential to capture the (dis)similarity between two texts. Our metrics differ from the contemporaneous work in several facets: (i) we explore the granularity of embeddings, leading to two variants of our metric, word mover and sentence mover; (ii) we investigate the effectiveness of diverse pretrained embeddings and finetuning tasks; (iii) we study the approach to consolidate layer-wise information within contextualized embeddings; (iii) our metrics demonstrate strong generalization capability across four tasks, oftentimes outperforming the supervised ones. We now describe our method in detail.

3 Our MoverScore Meric

We have motivated the need for better metrics capable of evaluating disparate NLG tasks. We now describe our metric, namely MoverScore, built upon a combination of (i) contextualized representations of system and reference texts and (ii) a distance between these representations measuring the semantic distance between system outputs and references. It is particularly important for a metric to not only capture the amount of *shared content* between two texts, i.e., intersect(A,B), as is the case with many semantic textual similarity measures (Peters et al., 2018; Devlin et al., 2018); but also to accurately reflect to what extent the system text has *deviated* from the reference, i.e., union(A,B) - intersect(A,B), which is the intuition behind using a distance metric.

3.1 Measuring Semantic Distance

Let $\boldsymbol{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_m)$ be a sentence viewed as a sequence of words. We denote by $\boldsymbol{x}^n$ the sequence of n-grams of $\boldsymbol{x}$ (i.e., $\boldsymbol{x}^1 = \boldsymbol{x}$ is the sequence of words and $\boldsymbol{x}^2$ is the sequence of bigrams). Furthermore, let $\boldsymbol{f}_{\boldsymbol{x}^n} \in \mathbb{R}_+^{|\boldsymbol{x}^n|}$ be a vector of weights, one weight for each n-gram of $\boldsymbol{x}^n$. We can assume $\boldsymbol{f}_{\boldsymbol{x}^n}^\intercal \mathbf{1} = 1$, making $\boldsymbol{f}_{\boldsymbol{x}^n}$ a distribution over n-grams. Intuitively, the effect of some n-grams like those including function words can be downplayed by giving them lower weights, e.g., using Inverse Document Frequency (IDF).

Word Mover's Distance (WMD) (Kusner et al., 2015), a special case of Earth Mover's Distance (Rubner et al., 2000), measures semantic distance between texts by aligning semantically similar words and finding the amount of flow traveling between these words. It was shown useful for text classification and textual similarity tasks (Kusner et al., 2015). Here, we formulate a generalization operating on n-grams. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two sentences viewed as sequences of n-grams: $\boldsymbol{x}^n$ and $\boldsymbol{y}^n$. If we have a distance metric d between n-grams, then we can define the *transportation cost* matrix $\boldsymbol{C}$ such that $\boldsymbol{C}_{ij} = d(\mathrm{x}_i^n, \mathrm{y}_j^n)$ is the distance between the i-th n-gram of $\boldsymbol{x}$ and the j-th n-gram of $\boldsymbol{y}$. The WMD between the two sequences of n-grams $\boldsymbol{x}^n$ and $\boldsymbol{y}^n$ with associated n-gram weights $\boldsymbol{f}_{\boldsymbol{x}^n}$ and $\boldsymbol{f}_{\boldsymbol{y}^n}$ is then given by:

$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$
$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^\intercal \mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

where $\boldsymbol{F}$ is the *transportation flow* matrix with $\boldsymbol{F}_{ij}$ denoting the amount of flow traveling from the i-th n-gram x_i^n in $\boldsymbol{x}^n$ to the j-th n-gram y_j^n in $\boldsymbol{y}^n$. Here, $\langle \boldsymbol{C}, \boldsymbol{F} \rangle$ denotes the sum of all matrix entries of the matrix $\boldsymbol{C} \odot \boldsymbol{F}$, where $\odot$ denotes element-wise multiplication. Then $\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n)$ is the minimal transportation cost between $\boldsymbol{x}^n$ and $\boldsymbol{y}^n$ where n-grams are weighted by $\boldsymbol{f}_{\boldsymbol{x}^n}$ and $\boldsymbol{f}_{\boldsymbol{y}^n}$.

In practice, we compute the Euclidean distance between the embedding representations of n-grams: $d(\mathrm{x}_i^n, \mathrm{y}_j^n) = ||E(\mathrm{x}_i^n) - E(\mathrm{y}_j^n)||_2$ where E is the embedding function which maps an n-gram to its vector representation. Usually, *static* word embeddings like word2vec are used to compute E but these cannot capture word order or compositionality. Alternatively, we investigate contextualized embeddings like ELMo and BERT because they encode information about the whole sentence into each word vector.

We compute the n-gram embeddings as the weighted sum over its word embeddings. Formally, if $\boldsymbol{x}_i^n = (\mathrm{x}_i, \ldots, \mathrm{x}_{i+n-1})$ is the i-th n-gram from sentence $\boldsymbol{x}$, its embedding is given by:

$$E(\boldsymbol{x}_i^n) = \sum_{k=i}^{i+n-1} \mathrm{idf}(\mathrm{x}_k) \cdot E(\mathrm{x}_k) \qquad (1)$$

where $\mathrm{idf}(\mathrm{x}_k)$ is the IDF of word x_k computed from all sentences in the corpus and $E(\mathrm{x}_k)$ is its word vector. Furthermore, the weight associated to the n-gram $\boldsymbol{x}_i^n$ is given by:

$$\boldsymbol{f}_{\boldsymbol{x}_i^n} = \frac{1}{Z} \sum_{k=i}^{i+n-1} \mathrm{idf}(\mathrm{x}_k) \qquad (2)$$

where Z is a normalizing constant s.t. $\boldsymbol{f}_{\boldsymbol{x}^n}^\top \mathbf{1} = 1$,

In the limiting case where n is larger than the sentence's size, $\boldsymbol{x}^n$ contains only one n-gram: the whole sentence. Then $\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n)$ reduces to computing the distance between the two sentence embeddings, namely Sentence Mover's Distance (SMD), denoted as:

$$\mathrm{SMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := ||E(\boldsymbol{x}_1^{l_x}) - E(\boldsymbol{y}_1^{l_y})||$$

where l_x and l_y are the size of sentences.

Hard and Soft Alignments In contemporaneous work, BERTScore (Zhang et al., 2019) also models the semantic distance between system and reference texts for evaluating text generation systems. As shown in Figure 1, BERTScore (precision/recall) can be intuitively viewed as hard alignments (one-to-one) for words in a sentence pair, where each word in one sequence travels to the most semantically similar word in the other sequence. In contrast, MoverScore goes beyond BERTScore as it relies on soft alignments (many-to-one) and allows to map semantically related words in one sequence to the respective word in the other sequence by solving a constrained optimization problem: finding the minimum effort to transform between two texts.

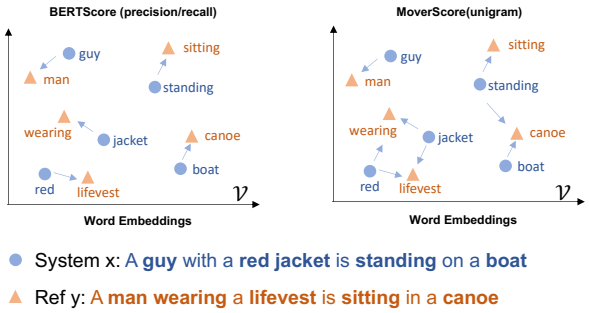


Figure 1: An illustration of MoverScore and BERTScore.

The formulation of Word Mover's Distance provides an important possibility to bias the metric towards precision or recall by using an asymmetric transportation cost matrix, which bridges a gap between MoverScore and BERTScore:

Proposition 1 *BERTScore (precision/recall) can be represented as a (non-optimized) Mover Distance $\langle \boldsymbol{C}, \boldsymbol{F} \rangle$, where $\boldsymbol{C}$ is a transportation cost matrix based on BERT and $\boldsymbol{F}$ is a uniform transportation flow matrix.*[2]

3.2 Contextualized Representations

The task formulation naturally lends itself to deep contextualized representations for inducing word vectors $E(\mathrm{x}_i)$. Despite the recent success of multi-layer attentive neural architectures (Devlin et al., 2018; Peters et al., 2018), consolidating layer-wise information remains an open problem as different layers capture information at disparate scales and task-specific layer selection methods may be limited (Liu et al., 2018, 2019). Tenney et al. (2019) found that a scalar mix of output layers trained from task-dependent supervisions would be effective in a deep transformer-based model. Instead, we investigate aggregation functions to consolidate layer-wise information, forming stationary representations of words without supervision.

Consider a sentence $\boldsymbol{x}$ passed through contextualized encoders such as ELMo and BERT with L layers. Each layer of the encoders produces a vec-

[2] See the proof in the appendix.

tor representation for each word x_i in $\boldsymbol{x}$. We denote by $\boldsymbol{z}_{i,l} \in \mathbb{R}^d$ the representation given by layer l, a d-dimensional vector. Overall, x_i receives L different vectors $(\boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,L})$. An aggregation ϕ maps these L vectors to one final vector:

$$E(x_i) = \phi(\boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,L}) \quad (3)$$

where $E(x_i)$ is the aggregated representation of the word x_i.

We study two alternatives for ϕ: (i) the concatenation of power means (Rücklé et al., 2018) as a generalized pooling mechanism, and (ii) a routing mechanism for aggregation (Zhao et al., 2018, 2019). We relegate the routing method to appendix, as it does not yield better results than power means.

Power Means Power means is an effective generalization of pooling techniques for aggregating information. It computes a non-linear average of a set of values with an exponent p (Eq. (4)). Following Rücklé et al. (2018), we exploit power means to aggregate vector representations $(\boldsymbol{z}_{i,l})_{l=1}^{L}$ pertaining to the i-th word from all layers of a deep neural architecture. Let $p \in \mathbb{R} \cup \{\pm\infty\}$, the p-mean of $(\boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,L})$ is:

$$\boldsymbol{h}_i^{(p)} = \left(\frac{\boldsymbol{z}_{i,1}^p + \cdots + \boldsymbol{z}_{i,L}^p}{L} \right)^{1/p} \in \mathbb{R}^d \quad (4)$$

where exponentiation is applied elementwise. This generalized form can induce common named means such as arithmetic mean ($p = 1$) and geometric mean ($p = 0$). In extreme cases, a power mean reduces to the minimum value of the set when $p = -\infty$, and the maximum value when $p = +\infty$. The concatenation of p-mean vectors we use in this paper is denoted by:

$$E(x_i) = \mathbf{h}_i^{(p_1)} \oplus \cdots \oplus \mathbf{h}_i^{(p_K)} \quad (5)$$

where $\oplus$ is vector concatenation; $\{p_1, \ldots, p_K\}$ are exponent values, and we use $K = 3$ with $p = 1, \pm\infty$ in this work.

3.3 Summary of MoverScore Variations

We investigate our MoverScore along four dimensions: (i) the granularity of embeddings, i.e., the size of n for n-grams, (ii) the choice of pretrained embedding mechanism, (iii) the fine-tuning task used for BERT[3] (iv) the aggregation technique (p-means or routing) when applicable.

[3]ELMo usually requires heavy layers on the top, which restricts the power of fine-tuning tasks for ELMo.

Granularity We used $n = 1$ and $n = 2$ as well as full sentences ($n =$ size of the sentence).

Embedding Mechanism We obtained word embeddings from three different methods: *static* embedding with word2vec as well as contextualized embedding with ELMo and BERT. If $n > 1$, n-gram embeddings are calculated by Eq. (1). Note that they represent sentence embeddings when $n =$ size of the sentence.

Fine-tuning Tasks Natural Language Inference (NLI) and paraphrasing pose high demands in understanding sentence meaning. This motivated us to fine-tune BERT representations on two NLI datasets, MultiNLI and QANLI, and one Paraphrase dataset, QQP—the largest datasets in GLUE (Wang et al., 2018). We fine-tune BERT on each of these, yielding different contextualized embeddings for our general evaluation metrics.

Aggregation For ELMo, we aggregate word representations given by all three ELMo layers, using p-means or routing (see the appendix). Word representations in BERT are aggregated from the last five layers, using p-means or routing since the representations in the initial layers are less suited for use in downstream tasks (Liu et al., 2019).

4 Empirical Evaluation

In this section, we measure the quality of different metrics on four tasks: *machine translation*, *text summarization*, *image captioning* and *dialogue generation*. Our major focus is to study the correlation between different metrics and human judgment. We employ two text encoders to embed n-grams: $\text{BERT}_{\text{base}}$, which uses a 12-layer transformer, and $\text{ELMO}_{\text{original}}$, which uses a 3-layer BiLSTM. We use Pearson's r and Spearman's ρ to measure the correlation. We consider two variants of MoverScore: *word mover* and *sentence mover*, described below.

Word Mover We denote our word mover notation containing four ingredients as: *WMD-Granularity+Embedding+Finetune+Aggregation*. For example, WMD-1+BERT+MNLI+PMEANS represents the semantic metric using word mover distance where unigram-based word embeddings fine-tuned on MNLI are aggregated by p-means.

Sentence Mover We denote our sentence mover notation with three ingredients as: *SMD+Embedding+Finetune+Aggregation*. For example, SMD+W2V represents the semantic

metric using sentence mover distance where two sentence embeddings are computed as the weighted sum over their word2vec embeddings by Eq. (1).

Baselines We select multiple strong baselines for each task for comparison: SentBLEU, METEOR++ (Guo et al., 2018), and a supervised metric RUSE for machine translation; ROUGE-1 and ROUGE-2 and a supervised metric S^3_{best} (Peyrard et al., 2017) for text summarization; BLEU and METEOR for dialogue response generation, CIDEr, SPICE, METEOR and a supervised metric LEIC (Cui et al., 2018) for image captioning. We also report BERTScore (Zhang et al., 2019) for all tasks (see §2). Due to the page limit, we only compare with the strongest baselines, the rest can be found in the appendix.

4.1 Machine Translation

Data We obtain the source language sentences, their system and reference translations from the WMT 2017 news translation shared task (Bojar et al., 2017). We consider 7 language pairs: from German (de), Chinese (zh), Czech (cs), Latvian (lv), Finnish (fi), Russian (ru), and Turkish (tr), resp. to English. Each language pair has approximately 3,000 sentences, and each sentence has one reference translation and multiple system translations generated by participating systems. For each system translation, at least 15 human assessments are independently rated for quality.

Results Table 1: In all language pairs, the best correlation is achieved by our word mover metrics that use a BERT pretrained on MNLI as the embedding generator and PMeans to aggregate the embeddings from different BERT layers, i.e., WMD-1/2+BERT+MNLI+PMeans. Note that our unsupervised word mover metrics even outperforms RUSE, a supervised metric. We also find that our word mover metrics outperforms the sentence mover. We conjecture that important information is lost in such a sentence representation while transforming the whole sequence of word vectors into one sentence embedding by Eq. (1).

4.2 Text Summarization

We use two summarization datasets from the Text Analysis Conference (TAC)[4]: TAC-2008 and TAC-2009, which contain 48 and 44 *clusters*, respectively. Each cluster includes 10 news articles (on the same topic), four reference summaries, and 57 (in TAC-2008) or 55 (in TAC-2009) system summaries generated by the participating systems. Each summary (either reference or system) has fewer than 100 words, and receives two human judgment scores: the *Pyramid* score (Nenkova and Passonneau, 2004) and the *Responsiveness* score. Pyramid measures how many important semantic content units in the reference summaries are covered by the system summary, while Responsiveness measures how well a summary responds to the overall quality combining both content and linguistic quality.

Results Tables 2: We observe that lexical metrics like ROUGE correlate above-moderate on TAC 2008 and 2009 datasets. In contrast, these metrics perform poorly on other tasks like Dialogue Generation (Novikova et al., 2017) and Image Captioning (Anderson et al., 2016). Apparently, strict matches on surface forms seems reasonable for *extractive* summarization datasets. However, we still see that our word mover metrics, i.e., WMD-1+BERT+MNLI+PMeans, perform better than or come close to even the supervised metric S^3_{best}.

4.3 Data-to-text Generation

We use two task-oriented dialogue datasets: BAGEL (Mairesse et al., 2010) and SFHOTEL (Wen et al., 2015), which contains 202 and 398 instances of Meaning Representation (MR). Each MR instance includes multiple references, and roughly two system utterances generated by different neural systems. Each system utterance receives three human judgment scores: *informativeness*, *naturalness* and *quality* score (Novikova et al., 2017). Informativeness measures how much information a system utterance provides with respect to an MR. Naturalness measures how likely a system utterance is generated by native speakers. Quality measures how well a system utterance captures fluency and grammar.

Results Tables 3: Interestingly, no metric produces an even moderate correlation with human judgments, including our own. We speculate that current contextualizers are poor at representing named entities like hotels and place names as well as numbers appearing in system and reference texts. However, best correlation is still achieved by our word mover metrics combining contextualized representations.

[4] http://tac.nist.gov

Setting	Metrics	Direct Assessment cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
BASELINES	METEOR++	0.552	0.538	0.720	0.563	0.627	0.626	0.646	0.610
	RUSE(*)	0.624	0.644	0.750	0.697	0.673	0.716	0.691	0.685
	BERTSCORE-F1	0.670	0.686	0.820	0.710	0.729	0.714	0.704	0.719
SENT-MOVER	SMD + W2V	0.438	0.505	0.540	0.442	0.514	0.456	0.494	0.484
	SMD + ELMO + PMEANS	0.569	0.558	0.732	0.525	0.581	0.620	0.584	0.595
	SMD + BERT + PMEANS	0.607	0.623	0.770	0.639	0.667	0.641	0.619	0.652
	SMD + BERT + MNLI + PMEANS	0.616	0.643	0.785	0.660	0.664	0.668	0.633	0.667
WORD-MOVER	WMD-1 + W2V	0.392	0.463	0.558	0.463	0.456	0.485	0.481	0.471
	WMD-1 + ELMO + PMEANS	0.579	0.588	0.753	0.559	0.617	0.679	0.645	0.631
	WMD-1 + BERT + PMEANS	0.662	0.687	0.823	0.714	0.735	0.734	0.719	0.725
	WMD-1 + BERT + MNLI + PMEANS	0.670	0.708	**0.835**	**0.746**	**0.738**	0.762	**0.744**	**0.743**
	WMD-2 + BERT + MNLI + PMEANS	**0.679**	**0.710**	0.832	0.745	0.736	**0.763**	0.740	**0.743**

Table 1: Absolute Pearson correlations with segment-level human judgments in 7 language pairs on WMT17 dataset.

Setting	Metrics	TAC-2008 Responsiveness r	Responsiveness ρ	Pyramid r	Pyramid ρ	TAC-2009 Responsiveness r	Responsiveness ρ	Pyramid r	Pyramid ρ
BASELINES	S^3_{best} (*)	0.715	0.595	0.754	0.652	0.738	**0.595**	**0.842**	**0.731**
	ROUGE-1	0.703	0.578	0.747	0.632	0.704	0.565	0.808	0.692
	ROUGE-2	0.695	0.572	0.718	0.635	0.727	0.583	0.803	0.694
	BERTSCORE-F1	0.724	0.594	0.750	0.649	0.739	0.580	0.823	0.703
SENT-MOVER	SMD + W2V	0.583	0.469	0.603	0.488	0.577	0.465	0.670	0.560
	SMD + ELMO + PMEANS	0.631	0.472	0.631	0.499	0.663	0.498	0.726	0.568
	SMD + BERT + PMEANS	0.658	0.530	0.664	0.550	0.670	0.518	0.731	0.580
	SMD + BERT + MNLI + PMEANS	0.662	0.525	0.666	0.552	0.667	0.506	0.723	0.563
WORD-MOVER	WMD-1 + W2V	0.669	0.549	0.665	0.588	0.698	0.520	0.740	0.647
	WMD-1 + ELMO + PMEANS	0.707	0.554	0.726	0.601	0.736	0.553	0.813	0.672
	WMD-1 + BERT + PMEANS	0.729	0.595	0.755	0.660	0.742	0.581	0.825	0.690
	WMD-1 + BERT + MNLI + PMEANS	**0.736**	**0.604**	**0.760**	**0.672**	**0.754**	0.594	0.831	0.701
	WMD-2 + BERT + MNLI + PMEANS	0.734	0.601	0.752	0.663	0.753	0.586	0.825	0.694

Table 2: Pearson r and Spearman ρ correlations with summary-level human judgments on TAC 2008 and 2009.

4.4 Image Captioning

We use a popular image captioning dataset: MS-COCO (Lin et al., 2014), which contains 5,000 images. Each image includes roughly five reference captions, and 12 system captions generated by the participating systems from 2015 COCO Captioning Challenge. For the system-level human correlation, each system receives five human judgment scores: M1, M2, M3, M4, M5 (Anderson et al., 2016). The M1 and M2 scores measure overall quality of the captions while M3, M4 and M5 scores measure correctness, detailedness and saliency of the captions. Following Cui et al. (2018), we compare the Pearson correlation with two system-level scores: M1 and M2, since we focus on studying metrics for the overall quality of the captions, leaving metrics understanding captions in different aspects (correctness, detailedness and saliency) to future work.

Results Table 4: Word mover metrics outperform all baselines except for the supervised metric LEIC, which uses more information by considering *both* images and texts.

4.5 Further Analysis

Hard and Soft Alignments BERTScore is the harmonic mean of BERTScore-Precision and BERTScore-Recall, where both two can be decomposed as a combination of "Hard Mover Distance" (HMD) and BERT (see Prop. 1).

We use the representations in the 9-th BERT layer for fair comparison of BERTScore and MoverScore and show results on the machine translation task in Table 5. MoverScore outperforms both asymmetric HMD factors, while if they are combined via harmonic mean, BERTScore is on par with MoverScore. We conjecture that BERT softens hard alignments of BERTScore as contextualized embeddings encode information about the whole sentence into each word vector. We also observe that WMD-BIGRAMS slightly outperforms WMD-UNIGRAMS on 3 out of 4 language pairs.

Setting	Metrics	BAGEL Inf	BAGEL Nat	BAGEL Qual	SFHOTEL Inf	SFHOTEL Nat	SFHOTEL Qual
BASELINES	BLEU-1	0.225	0.141	0.113	0.107	0.175	0.069
	BLEU-2	0.211	0.152	0.115	0.097	0.174	0.071
	METEOR	0.251	0.127	0.116	0.111	0.148	0.082
	BERTSCORE-F1	0.267	0.210	**0.178**	0.163	0.193	0.118
SENT-MOVER	SMD + W2V	0.024	0.074	0.078	0.022	0.025	0.011
	SMD + ELMO + PMEANS	0.251	0.171	0.147	0.130	0.176	0.096
	SMD + BERT + PMEANS	0.290	0.163	0.121	0.192	0.223	0.134
	SMD + BERT + MNLI + PMEANS	0.280	0.149	0.120	0.205	0.239	0.147
WORD-MOVER	WMD-1 + W2V	0.222	0.079	0.123	0.074	0.095	0.021
	WMD-1 + ELMO + PMEANS	0.261	0.163	0.148	0.147	0.215	0.136
	WMD-1 + BERT + PMEANS	**0.298**	**0.212**	0.163	0.203	0.261	0.182
	WMD-1 + BERT + MNLI + PMEANS	0.285	0.195	0.158	**0.207**	**0.270**	**0.183**
	WMD-2 + BERT + MNLI + PMEANS	0.284	0.194	0.156	0.204	0.270	0.182

Table 3: Spearman correlation with utterance-level human judgments for BAGEL and SFHOTEL datasets.

Setting	Metric	M1	M2
BASELINES	LEIC(*)	**0.939**	**0.949**
	METEOR	0.606	0.594
	SPICE	0.759	0.750
	BERTSCORE-RECALL	0.809	0.749
SENT-MOVER	SMD + W2V	0.683	0.668
	SMD + ELMO + P	0.709	0.712
	SMD + BERT + P	0.723	0.747
	SMD + BERT + M + P	0.789	0.784
WORD-MOVER	WMD-1 + W2V	0.728	0.764
	WMD-1 + ELMO + P	0.753	0.775
	WMD-1 + BERT + P	0.780	0.790
	WMD-1 + BERT + M + P	**0.813**	**0.810**
	WMD-2 + BERT + M + P	0.812	0.808

Table 4: Pearson correlation with system-level human judgments on MSCOCO dataset. 'M' and 'P' are short names.

Metrics	cs-en	de-en	fi-en	lv-en
RUSE	0.624	0.644	0.750	0.697
HMD-F1 + BERT	0.655	0.681	0.821	0.712
HMD-RECALL + BERT	0.651	0.658	0.788	0.681
HMD-PREC + BERT	0.624	0.669	0.817	0.707
WMD-UNIGRAM + BERT	0.651	0.686	**0.823**	0.710
WMD-BIGRAM + BERT	**0.665**	**0.688**	0.821	**0.712**

Table 5: Comparison on hard and soft alignments.

Distribution of Scores In Figure 2, we take a closer look at sentence-level correlation in MT. Results reveal that the lexical metric SENTBLEU can correctly assign lower scores to system translations of low quality, while it struggles in judging system translations of high quality by assigning them lower scores. Our finding agrees with the observations found in Chaganty et al. (2018); Novikova et al. (2017): lexical metrics correlate better with human judgments on texts of low quality than high quality. Peyrard (2019b) further show that lexical metrics cannot be trusted because

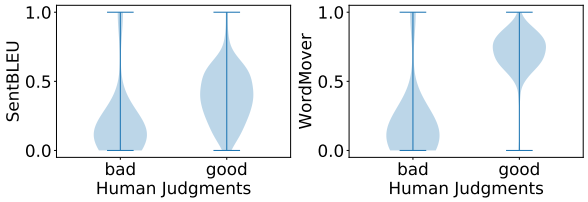


Figure 2: Score distribution in German-to-English pair.

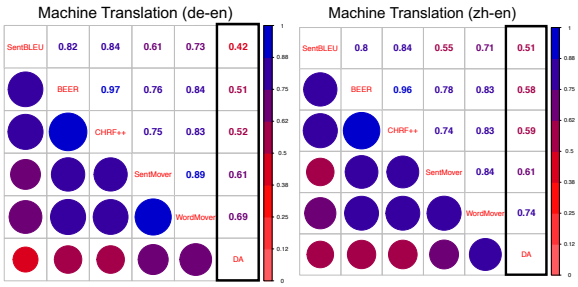


Figure 3: Correlation in similar language (de-en) and distant language (zh-en) pair, where bordered area shows correlations between human assessment and metrics, the rest shows inter-correlations across metrics and DA is direct assessment rated by language experts.

they strongly disagree on high-scoring system outputs. Importantly, we observe that our word mover metric combining BERT can clearly distinguish texts of two polar qualities.

Correlation Analysis In Figure 3, we observe existing metrics for MT evaluation attaining medium correlations (0.4-0.5) with human judgments but high inter-correlations between themselves. In contrast, our metrics can attain high correlations (0.6-0.7) with human judgments, performing robust across different language pairs. We believe that our improvements come from clearly distinguishing translations that fall on two extremes.

Impact of Fine-tuning Tasks Figure 4 com-

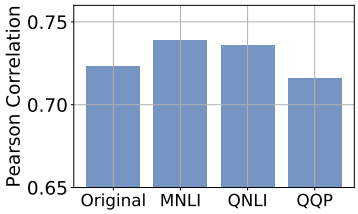


Figure 4: Correlation is averaged over 7 language pairs.

pares Pearson correlations with our word mover metrics combining BERT fine-tuned on three different tasks. We observe that fine-tuning on closely related tasks improves correlations, especially fine-tuning on MNLI leads to an impressive improvement by 1.8 points on average.

4.6 Discussions

We showed that our metric combining contextualized embeddings and Earth Mover's Distance outperforms strong unsupervised metrics on 3 out of 4 tasks, i.e., METEOR++ on machine translation by 5.7 points, SPICE on image captioning by 3.0 points, and METEOR on dialogue response generation by 2.2 points. The best correlation we achieved is combining contextualized word embeddings and WMD, which even rivals or exceeds SOTA task-dependent *supervised* metrics across different tasks. Especially in machine translation, our word mover metric pushes correlations in machine translation to 74.3 on average (5.8 points over the SOTA supervised metric and 2.4 points over contemporaneous BERTScore). The major improvements come from contextualized BERT embeddings rather than word2vec and ELMo, and from fine-tuning BERT on large NLI datasets. However, we also observed that soft alignments (MoverScore) marginally outperforms hard alignments (BERTScore). Regarding the effect of n-grams in word mover metrics, unigrams slightly outperforms bigrams on average. For the effect of aggregation functions, we suggested effective techniques for layer-wise consolidations, namely p-means and routing, both of which are close to the performance of the best layer and on par with each other (see the appendix).

5 Conclusion

We investigated new unsupervised evaluation metrics for text generation systems combining contextualized embeddings with Earth Mover's Distance. We experimented with two variants of our metric, sentence mover and word mover. The latter has demonstrated strong generalization ability across four text generation tasks, oftentimes even outperforming supervised metrics. Our metric provides a promising direction towards a holistic metric for text generation and a direction towards more 'human-like' (Eger et al., 2019) evaluation of text generation systems.

In future work, we plan to avoid the need for costly human references in the evaluation of text generation systems, and instead base evaluation scores on source texts and system predictions only, which would allow for 'next-level', unsupervised (in a double sense) and unlimited evaluation (Louis and Nenkova, 2013; Böhm et al., 2019).

Acknowledgments

We thank the anonymous reviewers for their comments, which greatly improved the final version of the paper. This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1. Fei Liu is supported in part by NSF grant IIS-1909603.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 382–398.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China.

Ondrej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Conference on Machine Translation (WMT)*.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of*

the 57th Annual Meeting of the Association for Computational Linguistics, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Dorin Comaniciu and Peter Meer. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):603–619.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5804–5812.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research (JAIR)*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. NEWSROOM: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Yinuo Guo, Chong Ruan, and Junfeng Hu. 2018. Meteor++: Incorporating copy knowledge into machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 740–745.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of Neural Information Processing Systems (NIPS)*.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006)*, pages 604–611.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of ACL workshop on Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Liyuan Liu, Xiang Ren, Jingbo Shang, Xiaotao Gu, Jian Peng, and Jiawei Han. 2018. Efficient contextualized representation: Language model pruning for sequence labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 589–597.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Qingsong Ma, Ondrej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation

using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 145–152. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Maxime Peyrard. 2019a. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard. 2019b. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *CoRR*, abs/1803.01400.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Stephen Tratz and Eduard H Hovy. 2008. Summarization Evaluation Using Transformed Basic Elements. In *Proceedings of the text analysing conference, (TAC 2008)*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575.

Matt P Wand and M Chris Jones. 1994. *Kernel smoothing*. Chapman and Hall/CRC.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2018. Learning neural templates for text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Suofei Zhang, Wei Zhao, Xiaofu Wu, and Quan Zhou. 2018. Fast dynamic routing based on weighted kernel density estimation. *CoRR*, abs/1805.10807.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Wei Zhao, Haiyun Peng, Steffen Eger, Erik Cambria, and Min Yang. 2019. Towards scalable and reliable capsule networks for challenging NLP applications. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1549–1559, Florence, Italy. Association for Computational Linguistics.

Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Suofei Zhang, and Zhou Zhao. 2018. Investigating capsule networks with dynamic routing for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

A Supplemental Material

A.1 Proof of Prop. 1

In this section, we prove Prop. 1 in the paper about viewing BERTScore (precision/recall) as a (non-optimized) Mover Distance.

As a reminder, the WMD formulation is:

$$\text{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \sum_{i,j} \boldsymbol{C}_{ij} \cdot \boldsymbol{F}_{ij}$$
$$\text{s.t. } \mathbf{1}^\intercal \boldsymbol{F}^\intercal \mathbf{1} = 1, \quad \mathbf{1}^\intercal \boldsymbol{F} \mathbf{1} = 1.$$

where $\boldsymbol{F}^\intercal \mathbf{1} = \boldsymbol{f}_x^n$ and $\boldsymbol{F}\mathbf{1} = \boldsymbol{f}_y^n$. Here, $\boldsymbol{f}_x^n$ and $\boldsymbol{f}_y^n$ denote vectors of weights for each n-gram of $\boldsymbol{x}^n$ and $\boldsymbol{y}^n$.

BERTScore is defined as:

$$R_{\text{BERT}} = \frac{\sum_{y_i^1 \in \boldsymbol{y}^1} \text{idf}(y_i^1) \max_{x_j^1 \in \boldsymbol{x}^1} E(x_j^1)^\intercal E(y_i^1)}{\sum_{y_i^1 \in \boldsymbol{y}^1} \text{idf}(y_i^1)}$$
$$P_{\text{BERT}} = \frac{\sum_{x_j^1 \in \boldsymbol{x}^1} \text{idf}(x_j^1) \max_{y_i^1 \in \boldsymbol{y}^1} E(y_i^1)^\intercal E(x_j^1)}{\sum_{x_j^1 \in \boldsymbol{x}^1} \text{idf}(x_j^1)}$$
$$F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad .$$

Then, R_{BERT} can be formulated in a "quasi" WMD form:

$$R_{\text{BERT}}(\boldsymbol{x}^1, \boldsymbol{y}^1) := \sum_{i,j} \boldsymbol{C}_{ij} \cdot \boldsymbol{F}_{ij}$$
$$\boldsymbol{F}_{ij} = \begin{cases} \frac{1}{M} & \text{if } x_j = \arg\max_{\hat{x}_j^1 \in \boldsymbol{x}^1} E(y_i^1)^\intercal E(\hat{x}_j^1) \\ 0 & \text{otherwise} \end{cases}$$
$$\boldsymbol{C}_{ij} = \begin{cases} \frac{M}{Z} \text{idf}(y_i^1) E(x_j^1)^\intercal E(y_i^1) & \text{if } x_j = \arg\max_{\hat{x}_j^1 \in \boldsymbol{x}^1} E(y_i^1)^\intercal E(\hat{x}_j^1) \\ 0 & \text{otherwise} \end{cases}$$

where $Z = \sum_{y_i^1 \in \boldsymbol{y}^1} \text{idf}(y_i^1)$ and M is the size of n-grams in $\boldsymbol{x}^1$. Similarly, we can have P_{BERT} in a quasi WMD form (omitted). Then, F_{BERT} can be formulated as harmonic-mean of two WMD forms of P_{BERT} and R_{BERT}.

A.2 Routing

In this section, we study the aggregation function ϕ with a routing scheme, which has achieved good results in other NLP tasks (Zhao et al., 2018, 2019). Specifically, we introduce a nonparametric clustering with Kernel Density Estimation (KDE) for routing since KDE bridges a family of kernel functions with underlying empirical distributions, which often leads to computational efficiency (Zhang et al., 2018), defined as:

$$\min_{\boldsymbol{v}, \gamma} f(\boldsymbol{z}) = \sum_{i=1}^{L} \sum_{j=1}^{T} \gamma_{ij} k(d(\boldsymbol{v}_j - \boldsymbol{z}_{i,j}))$$
$$\textit{s.t.} \quad \forall i, j : \gamma_{ij} > 0, \sum_{j=1}^{L} \gamma_{ij} = 1.$$

where $d(\cdot)$ is a distance function, γ_{ij} denotes the underlying closeness between the aggregated vector $\boldsymbol{v}_j$ and vector $\boldsymbol{z}_i$ in the i-th layer, and k is a kernel function. Some instantiations of $k(\cdot)$ (Wand and Jones, 1994) are:

$$Gaussian : k(x) \triangleq \exp\left(-\frac{x}{2}\right), \quad Epanechnikov : k(x) \triangleq \begin{cases} 1-x & x \in [0,1) \\ 0 & x \geq 1. \end{cases}$$

One typical solution for KDE clustering to minimize $f(\boldsymbol{z})$ is taking Mean Shift (Comaniciu and Meer, 2002), defined as:

$$\nabla f(\boldsymbol{z}) = \sum_{i,j} \gamma_{ij} k'(d(\boldsymbol{v}_j, \boldsymbol{z}_{i,j})) \frac{\partial d(\boldsymbol{v}_j, \boldsymbol{z}_{i,j})}{\partial \boldsymbol{v}}$$

Firstly, $\boldsymbol{v}_j^{\tau+1}$ can be updated while $\gamma_{ij}^{\tau+1}$ is fixed:

$$\boldsymbol{v}_j^{\tau+1} = \frac{\sum_i \gamma_{ij}^{\tau} k'(d(\boldsymbol{v}_j^{\tau}, \boldsymbol{z}_{i,j}))\boldsymbol{z}_{i,j}}{\sum_{i,j} k'(d(\boldsymbol{v}_j^{\tau}, \boldsymbol{z}_{i,j}))}$$

Intuitively, $\boldsymbol{v}_j$ can be explained as a final aggregated vector from L contextualized layers. Then, we adopt SGD to update $\gamma_{ij}^{\tau+1}$:

$$\gamma_{ij}^{\tau+1} = \gamma_{ij}^{\tau} + \alpha \cdot k(d(\boldsymbol{v}_j^{\tau}, \boldsymbol{z}_{i,j}))$$

where α is a hyperparameter to control step size. The routing process is summarized in Algorithm 1.

Algorithm 1 Aggregation by Routing

procedure ROUTING($\boldsymbol{z}_{ij}, \ell$)
Initialize $\forall i, j : \gamma_{ij} = 0$
while true **do**
  foreach representation i and j in layer ℓ and $\ell+1$ **do** $\gamma_{ij} \leftarrow softmax(\gamma_{ij})$
  foreach representation j in layer $\ell+1$ **do**
    $\boldsymbol{v}_j \leftarrow \sum_i \gamma_{ij} k'(\boldsymbol{v}_j, \boldsymbol{z}_i)\boldsymbol{z}_i / \sum_i k'(\boldsymbol{v}_i, \boldsymbol{z}_i)$
  foreach representation i and j in layer ℓ and $\ell+1$ **do** $\gamma_{ij} \leftarrow \gamma_{ij} + \alpha \cdot k(\boldsymbol{v}_j, \boldsymbol{z}_i)$
  loss $\leftarrow \log(\sum_{\mathrm{i,j}} \gamma_{\mathrm{ij}} \mathrm{k}(\boldsymbol{v}_{\mathrm{j}}, \boldsymbol{z}_{\mathrm{i}}))$
  if $|\text{loss} - \text{preloss}| < \epsilon$ **then**
    break
  else
    preloss $\leftarrow$ loss
return v_j

Best Layer and Layer-wise Consolidation Table 6 compares our word mover based metric combining BERT representations on different layers with stronger BERT representations consolidated from these layers (using p-means and routing). We often see that which layer has best performance is task-dependent, and our word mover based metrics (WMD) with p-means or routing schema come close to the oracle performance obtained from the best layers.

Experiments Table 7, 8 and 9 show correlations between metrics (all baseline metrics and word mover based metrics) and human judgments on machine translation, text summarization and dialogue response generation, respectively. We find that word mover based metrics combining BERT fine-tuned on MNLI have highest correlations with humans, outperforming all of the unsupervised metrics and even supervised metrics like RUSE and S^3_{full}. Routing and p-means perform roughly equally well.

Metrics	Direct Assessment cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en
WMD-1 + BERT + LAYER 8	.6361	.6755	.8134	.7033	.7273	.7233	.7175
WMD-1 + BERT + LAYER 9	.6510	.6865	.8240	.7107	.7291	**.7357**	**.7195**
WMD-1 + BERT + LAYER 10	.6605	**.6948**	**.8231**	.7158	**.7363**	.7317	.7168
WMD-1 + BERT + LAYER 11	**.6695**	.6845	.8192	.7048	.7315	.7276	.7058
WMD-1 + BERT + LAYER 12	.6677	.6825	.8194	**.7188**	.7326	.7291	.7064
WMD-1 + BERT + ROUTING	.6618	.6897	.8225	.7122	.7334	.7301	.7182
WMD-1 + BERT + PMEANS	.6623	.6873	.8234	.7139	.7350	.7339	.7192

Table 6: Absolute Pearson correlations with segment-level human judgments on WMT17 to-English translations.

Setting	Metrics	Direct Assessment cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
BASELINES	BLEND	0.594	0.571	0.733	0.594	0.622	0.671	0.661	0.635
	RUSE	0.624	0.644	0.750	0.697	0.673	0.716	0.691	0.685
	SENTBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.481
	CHRF++	0.523	0.534	0.678	0.520	0.588	0.614	0.593	0.579
	METEOR++	0.552	0.538	0.720	0.563	0.627	0.626	0.646	0.610
	BERTSCORE-F1	0.670	0.686	0.820	0.710	0.729	0.714	0.704	0.719
WORD-MOVER	WMD-1 + W2V	0.392	0.463	0.558	0.463	0.456	0.485	0.481	0.471
	WMD-1 + BERT + ROUTING	0.658	0.689	0.823	0.712	0.733	0.730	0.718	0.723
	WMD-1 + BERT + MNLI + ROUTING	0.665	0.705	0.834	0.744	0.735	0.752	0.736	0.739
	WMD-2 + BERT + MNLI + ROUTING	0.676	0.706	0.831	0.743	0.734	0.755	0.732	0.740
	WMD-1 + BERT + PMEANS	0.662	0.687	0.823	0.714	0.735	0.734	0.719	0.725
	WMD-1 + BERT + MNLI + PMEANS	0.670	0.708	**0.835**	**0.746**	**0.738**	0.762	**0.744**	**0.743**
	WMD-2 + BERT + MNLI + PMEANS	**0.679**	**0.710**	0.832	0.745	0.736	**0.763**	0.740	**0.743**

Table 7: Absolute Pearson correlations with segment-level human judgments on WMT17 to-English translations.

Setting	Metrics	TAC-2008 Responsiveness r	Responsiveness ρ	Pyramid r	Pyramid ρ	TAC-2009 Responsiveness r	Responsiveness ρ	Pyramid r	Pyramid ρ
BASELINES	S^3_{full}	0.696	0.558	0.753	0.652	0.731	0.552	0.838	0.724
	S^3_{best}	0.715	0.595	0.754	0.652	0.738	**0.595**	**0.842**	**0.731**
	TF∗IDF-1	0.176	0.224	0.183	0.237	0.187	0.222	0.242	0.284
	TF∗IDF-2	0.047	0.154	0.049	0.182	0.047	0.167	0.097	0.233
	ROUGE-1	0.703	0.578	0.747	0.632	0.704	0.565	0.808	0.692
	ROUGE-2	0.695	0.572	0.718	0.635	0.727	0.583	0.803	0.694
	ROUGE-1-WE	0.571	0.450	0.579	0.458	0.586	0.437	0.653	0.516
	ROUGE-2-WE	0.566	0.397	0.556	0.388	0.607	0.413	0.671	0.481
	ROUGE-L	0.681	0.520	0.702	0.568	0.730	0.563	0.779	0.652
	FRAME-1	0.658	0.508	0.686	0.529	0.678	0.527	0.762	0.628
	FRAME-2	0.676	0.519	0.691	0.556	0.715	0.555	0.781	0.648
	BERTSCORE-F1	0.724	0.594	0.750	0.649	0.739	0.580	0.823	0.703
WORD-MOVER	WMD-1 + W2V	0.669	0.559	0.665	0.611	0.698	0.520	0.740	0.647
	WMD-1 + BERT + ROUTING	0.729	0.601	0.763	0.675	0.740	0.580	0.831	0.700
	WMD-1 + BERT + MNLI + ROUTING	0.734	**0.609**	**0.768**	**0.686**	0.747	0.589	0.837	0.711
	WMD-2 + BERT + MNLI + ROUTING	0.731	0.593	0.755	0.666	0.753	0.583	0.827	0.698
	WMD-1 + BERT + PMEANS	0.729	0.595	0.755	0.660	0.742	0.581	0.825	0.690
	WMD-1 + BERT + MNLI + PMEANS	**0.736**	0.604	0.760	0.672	**0.754**	0.594	0.831	0.701
	WMD-2 + BERT + MNLI + PMEANS	0.734	0.601	0.752	0.663	0.753	0.586	0.825	0.694

Table 8: Correlation of automatic metrics with summary-level human judgments for TAC-2008 and TAC-2009.

Setting	Metrics	BAGEL			SFHOTEL		
		Inf	**Nat**	**Qual**	**Inf**	**Nat**	**Qual**
BASELINES	BLEU-1	0.225	0.141	0.113	0.107	0.175	0.069
	BLEU-2	0.211	0.152	0.115	0.097	0.174	0.071
	BLEU-3	0.191	0.150	0.109	0.089	0.161	0.070
	BLEU-4	0.175	0.141	0.101	0.084	0.104	0.056
	ROUGE-L	0.202	0.134	0.111	0.092	0.147	0.062
	NIST	0.207	0.089	0.056	0.072	0.125	0.061
	CIDER	0.205	0.162	0.119	0.095	0.155	0.052
	METEOR	0.251	0.127	0.116	0.111	0.148	0.082
	BERTSCORE-F1	0.267	0.210	**0.178**	0.163	0.193	0.118
WORD-MOVER	WMD-1 + W2V	0.222	0.079	0.123	0.074	0.095	0.021
	WMD-1 + BERT + ROUTING	0.294	0.209	0.156	0.208	0.256	0.178
	WMD-1 + BERT + MNLI + ROUTING	0.278	0.180	0.144	**0.211**	0.252	0.175
	WMD-2 + BERT + MNLI + ROUTING	0.279	0.182	0.147	0.204	0.252	0.172
	WMD-1 + BERT + PMEANS	**0.298**	**0.212**	0.163	0.203	0.261	0.182
	WMD-1 + BERT + MNLI + PMEANS	0.285	0.195	0.158	0.207	**0.270**	**0.183**
	WMD-2 + BERT + MNLI + PMEANS	0.284	0.194	0.156	0.204	**0.270**	0.182

Table 9: Spearman correlation with utterance-level human judgments for BAGEL and SFHOTEL datasets.

Chapter 6

DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence

DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence

Wei Zhao[12] **Michael Strube**[1] **Steffen Eger**[3]
[1]Heidelberg Institute for Theoretical Studies [2]Technische Universität Darmstadt
www.h-its.org/research/nlp/
[3]NLLG, Faculty of Technology, Bielefeld University
{wei.zhao, michael.strube}@h-its.org
steffen.eger@uni-bielefeld.de
nl2g.github.io

Abstract

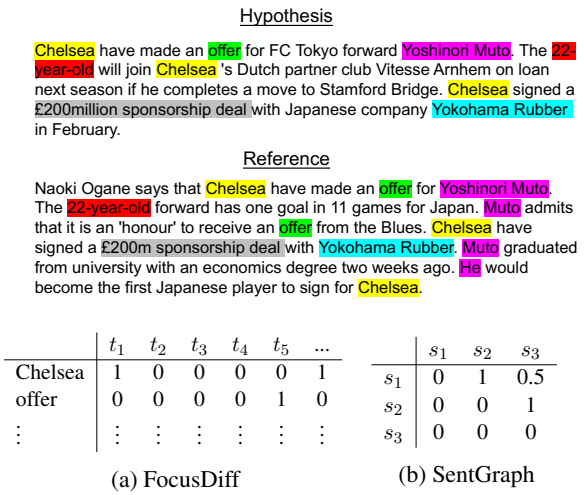
Recently, there has been a growing interest in designing text generation systems from a discourse coherence perspective, e.g., modeling the interdependence between sentences. Still, recent BERT-based evaluation metrics are weak in recognizing coherence, and thus are not reliable in a way to spot the discourse-level improvements of those text generation systems. In this work, we introduce DiscoScore, a parametrized discourse metric, which uses BERT to model discourse coherence from different perspectives, driven by Centering theory. Our experiments encompass 16 non-discourse and discourse metrics, including DiscoScore and popular coherence models, evaluated on summarization and document-level machine translation (MT). We find that (i) the majority of BERT-based metrics correlate much worse with human rated coherence than early discourse metrics, invented a decade ago; (ii) the recent state-of-the-art BARTScore is weak when operated at system level—which is particularly problematic as systems are typically compared in this manner. DiscoScore, in contrast, achieves strong system-level correlation with human ratings, not only in coherence but also in factual consistency and other aspects, and surpasses BARTScore by over 10 correlation points on average. Further, aiming to understand DiscoScore, we provide justifications to the importance of discourse coherence for evaluation metrics, and explain the superiority of one variant over another. Our code is available at https://github.com/AIPHES/DiscoScore.

1 Introduction

In discourse, coherence refers to the continuity of semantics in text. Often, discourse relations and lexical cohesion devices, such as repetition and coreference, are employed to connect text spans, aiming to ensure text coherence. Popular theories in the linguistics community on discourse were provided by Grosz et al. (1995) and Mann and Thompson (1988). They formulate coherence through the lens of readers' focus of attention, and rhetorical discourse structures over sentences. Later on, coherence models as computational approaches of these theories emerged to judge text coherence in discourse tasks such as sentence ordering and essay scoring (Barzilay and Lapata, 2008; Lin et al., 2011; Guinaudeau and Strube, 2013).

While humans also often use text planning at discourse level prior to writing and speaking, up until recently, the majority of natural language generation (NLG) systems, be it text summarization or document-level MT, has performed sequential word prediction without considering text coherence. For instance, MT systems mostly do not model the interdependence between sentences and translate a document at sentence level, and thus produce many incoherent elements such as coreference mistakes in system outputs (Maruf et al., 2021). Only more recently has there been a surge of interest towards discourse based summarization and MT systems, aiming to model inter-sentence context, with a focus on pronominal anaphora (Voita et al., 2018; Liu et al., 2021) and discouse relations (Miculicich et al., 2018; Xu et al., 2020).

However, there appears a mismatch between discourse based NLG systems and non-discourse NLG evaluation metrics such as MoverScore (Zhao et al., 2019) and BERTScore (Zhang et al., 2020) which have recently become popular for MT and summarization evaluation. As these metrics base their judgment on semantic similarity (and lexical overlap (Kaster et al., 2021)) between hypotheses and references—which by design does not target text coherence—it is not surprising that they do not correlate well with human rated coherence (Fabbri et al., 2021; Yuan et al., 2021; Sai et al., 2021). Recently, BARTScore (Yuan et al., 2021) receives increasingly attention, which uses sequence-to-sequence language models to measure the likeli-



	t_1	t_2	t_3	t_4	t_5	...
Chelsea	1	0	0	0	0	1
offer	0	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

(a) FocusDiff

	s_1	s_2	s_3
s_1	0	1	0.5
s_2	0	0	1
s_3	0	0	0

(b) SentGraph

Figure 1: Sample hypothesis and reference from SUMMEval. Each focus[1] is marked in a different color, corresponding to multiple tokens as instances of a focus. Foci shared in Hypothesis and Reference are marked in the same color. (a)+(b) are adjacency matrices used to model focus-based coherence for Hypothesis; for simplicity, adjacency matrices for Reference are omitted. FocusDiff and SentGraph are the variants of DiscoScore. For FocusDiff, we use (a) to depict the relations between foci and tokens, reflecting focus frequency. For SentGraph, we use (b) to depict the interdependence between sentences according to the number of foci shared between sentences and the distance between sentences.

hood that hypothesis and reference are paraphrases, and that cannot contrast text pairs at discourse level.

In this work, we fill the gap of missing discourse metrics in MT and summarization evaluation, particularly in reference-based evaluation scenarios. We introduce DiscoScore, a parametrized discourse metric, which uses BERT to model discourse coherence through the lens of readers' focus, driven by Centering theory (Grosz et al., 1995). The DiscoScore variants can be distinguished in how we use *focus*—see Figure 1: (i) we model focus frequency and semantics, and compare their difference between hypothesis and reference and (ii) we use focus transitions to model the interdependence between sentences. Building upon this, we present a simple graph-based approach to compare hypothesis with reference.

We compare DiscoScore with a range of baselines, including discourse and non-discourse metrics, and coherence models on summarization and document-level MT datasets. Our contributions and findings are summarized as follows:

- Recent BERT-based metrics and the state-of-the-art BARTScore (Yuan et al., 2021) are all weak in system-level correlation with human ratings, not only in coherence but also in other aspects such as factual consistency. Most of them are even worse than very early discourse metrics, RC and LC (Wong and Kit, 2012)—which require neither source texts nor references and use discourse features to predict hypothesis coherence.
- DiscoScore strongly correlates with human rated coherence and many other aspects, over 10 points (on average across aspects) better than BARTScore and two strong baselines RC and LC in the single and multi-references settings. This indicates that either leveraging contextualized encoders or finding discourse features is not sufficient, suggesting to combine both as DiscoScore does.
- We demonstrate the importance of including discourse signals in the assessment of system outputs, as the discourse features derived from DiscoScore can strongly separate hypothesis from reference. Further, we show that the more discriminative these features are, the better the metrics perform, which allows for interpreting the performance gaps between the variants of DisoScore.
- We investigate two focus choices popular in the discourse community, i.e., noun (Elsner and Charniak, 2011) and semantic entity (Mesgar and Strube, 2016). Our results show that entity as focus is not always helpful, but when it helps, the gain is big.

2 Related work

Evaluation Metrics. Traditional metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) measure lexical n-gram overlap between a hypothesis and a human reference. As they fail to measure semantic similarity in the absence of lexical overlap, several metrics have been proposed to overcome this issue, which carry out soft lexical matching with static word embeddings (Ng and Abrecht, 2015) and synonym matching (Lavie and Agarwal, 2007). However, none of those metrics

[1]The formal definition of focusing in discourse is given on two levels (Grosz et al., 1977): (i) readers are said to be *globally* focusing on a set of entities relevant to the overall discourse, and (ii) readers focus on a particular entity that an utterance *locally* concerns most. Section 3 elaborates on focus as a key ingredient of DiscoScore.

can properly judge text coherence (Kryscinski et al., 2019; Zhu and Bhat, 2020).

Recently, a class of novel metrics based on BERT (Devlin et al., 2019) has received a surge of attention, as they correlate strongly with human judgment of text quality in both reference-based and reference-free scenarios (Zhao et al., 2019; Zhang et al., 2020; Sellam et al., 2020; Rei et al., 2020; Gao et al., 2020; Thompson and Post, 2020; Zhao et al., 2020; Pu et al., 2021; Chen et al., 2021). While strong at sentence-level, these metrics are weak in recognizing coherence in inter-sentence contexts (just like BLEU and ROUGE), as BERT and the majority of BERT variants[2] that these metrics build on only capture discourse phenomena to a certain extent (Koto et al., 2021; Laban et al., 2021; Beyer et al., 2021). Thus, they are not suitable for evaluating long texts as in document-level MT evaluation. Works that either (i) average sentence-level evaluation scores as document score or (ii) assign a score to the concatenation of sentences within a document (Xiong et al., 2019; Liu et al., 2020; Saunders et al., 2020) do not factor interdependence between sentences into a document score, e.g., do not explicitly punish incoherent elements, thus are also inadequate.

Several attempts have been made towards discourse metrics in MT evaluation. Wong and Kit (2012); Gong et al. (2015); Cartoni et al. (2018) use the frequency of lexical cohesion devices (e.g., word repetition) over sentences to predict coherence of hypothesis translations, while Guzmán et al. (2014) and Joty et al. (2017) suggest to compare the difference of rhetorical structures between hypothesis and reference translations. Recently, Jiang et al. (2021) measure the inconsistency between hypothesis and reference translations in several aspects such as verb tense and named entities. However, these metrics do not leverage strong contextualized encoders, as has been shown to be a key ingredient for recent success of BERT-based metrics. Most recently, BARTScore (Yuan et al., 2021) uses sequence-to-sequence pretrained language models such as BART (Lewis et al., 2020) to measure how likely hypothesis and reference are paraphrased according to the probability of one given the other. While BARTScore constitutes the recent state-of-the-art in sentence-level correlation with human ratings in several aspects (incl. discourse), we find that (i) it performs still poorly at system level—which is particularly problematic as systems are typically compared in this manner. (ii) As based on a 'blackbox' language model, it cannot offer insights towards how it models coherence and what discourse phenomena it does (not) capture.

Coherence Models. In discourse, there have been many computational models (Barzilay and Lapata, 2008; Guinaudeau and Strube, 2013; Pitler and Nenkova, 2008; Lin et al., 2011) for text coherence assessment, the majority of which differ in *regularities* that they use to distinguish coherent from incoherent text, driven by different linguistic theories, *v.i.z.*, a pattern of (i) focus transitions in adjacent sentences (Grosz et al., 1995) and (ii) text organization regarding discourse relations over sentences (Mann and Thompson, 1988). For instance, Barzilay and Lapata (2008) and Guinaudeau and Strube (2013) use the distribution of entity transitions over sentences to predict text coherence, while Pitler and Nenkova (2008) and Lin et al. (2011) suggest to produce discourse relations over sentences with a discourse parser, showing that the relations are indicative of text coherence. In the last few years, neural coherence models have been explored. Popular examples are Tien Nguyen and Joty (2017), Mesgar and Strube (2018) and Moon et al. (2019). As they and the recent state-of-the-art (Mesgar et al., 2021) all have been trained on text readability datasets, with readability labels as supervision, they may suffer issues of domain shift when applied to MT and summarization evaluation. More importantly, they judge hypothesis coherence in the absence of reference, thus are not sufficient for reference-based evaluation. Our experiments involve two popular, unsupervised coherence models, entity graph (Guinaudeau and Strube, 2013) and lexical graph (Mesgar and Strube, 2016) treated as discourse metrics with the advantages on robustness (Lai and Tetreault, 2018).

Discourse Test Sets. Apart from evaluation metrics, there have been several discourse-focused test sets proposed to compare NLG systems, most of which have been studied in MT evaluation. For instance, the DiscoMT15 shared task (Hardmeier et al., 2015) compares MT systems, not based on translation adequacy but on the accuracy of pronoun translation for English-to-French, i.e., counting the number of correctly translated pronouns, given the annotated ones in reference. Bawden

[2]Recently, several discourse BERT variants such as Conpono (Iter et al., 2020) have been proposed, but they are not always helpful for evaluation metrics—see Table 2 (appendix).

et al. (2018) extend this by labeling both anaphoric pronouns and lexical cohesion devices on test sets, while Voita et al. (2018) construct English-to-Russian test sets focusing on deixis, ellipsis and lexical cohesion. Guillou et al. (2018); Lopes et al. (2020) construct English-to-German and English-to-French test sets targeting pronouns. While reliable, these test sets involve costly manual annotation, thus are limited to few language pairs.

In this work, we introduce DiscoScore to judge system outputs, which uses BERT to model readers' focus within hypothesis and reference, and thus clearly outlines the discourse phenomena being captured, serving as low-cost alternatives to discourse test sets for comparing discourse based NLG systems. More prominently, we derive discourse features from DiscoScore, which we use to understand the importance of discourse for evaluation metrics, and explain why one metric is superior to another. This parallels recent effort towards explainability for non-discourse evaluation metrics (Kaster et al., 2021; Fomicheva et al., 2021). Finally, we show that simple features can be indicative of the superiority of a metric, which fosters research towards finding insightful features with domain expertise and building upon these insights to design high-quality metrics.

3 Our Approach

In the following, we elaborate on the two variants of DiscoScore, FocusDiff and SentGraph, which we refer to as DS-FOCUS and DS-SENT.

Focus Difference. In discourse, there have been many corpus-based studies towards modeling focus transitions over sentences, showing that focus transition patterns are indicative of text coherence (Barzilay and Lapata, 2008; Guinaudeau and Strube, 2013). When reading a document, readers may have multiple *focus of attention*,

each associated to a group of expressions: (i) referring expressions such as pronouns and (ii) semantically related elements such as [*Berlin, capital*].

Here, we assume two focus based conditions that a coherent hypothesis should meet in reference-based evaluation scenarios:

- A large number of focus overlaps between a hypothesis and a reference.
- Each focus overlap is nearly identical in terms of semantics and frequency, where frequency shows how often a focus is mentioned in a hypothesis or in a reference.

In the following, we present focus modeling towards semantics and frequency, according to which we compare hypothesis with reference.

For a hypothesis, we introduce a bipartite graph $\mathcal{G}^{\text{hyp}} = (\mathcal{V}, \mathcal{S}, \mathbf{A}^{\text{hyp}})$, where $\mathcal{V}$ and $\mathcal{S}$ are two sets of vertices corresponding to a set of foci and all tokens (per occurrence a word is a separate token) within a hypothesis. Let $\mathbf{A} = \{0,1\}^{n \times m}$ be an adjacency matrix where n and m are the number of foci and tokens respectively, and A_{ij} equals 1 if and only if the i-th focus associates to the j-th token. Let $\mathbf{F}^{\text{hyp}} \in \mathbb{R}^{n \times d}$ be a matrix of focus embeddings and $\mathbf{Z}^{\text{hyp}} \in \mathbb{R}^{m \times d}$ be a matrix of contextualized token embeddings with d as the embedding size. Similarly, we use notation $\mathcal{G}^{\text{ref}}$, $\mathbf{F}^{\text{ref}}$ and $\mathbf{Z}^{\text{ref}}$ for a human reference.

We use contextualized encoders such as BERT to produce token embeddings $\mathbf{Z}^{\text{hyp}}$ and $\mathbf{Z}^{\text{ref}}$. We use a simple approach to model both semantics and frequency of a focus. That is, we assign per focus v an embedding by summing token embeddings that a focus is associated to:

$$\mathbf{F}_v^{\text{hyp}} = \sum_{u \in \mathcal{N}(v)} \mathbf{Z}_u^{\text{hyp}}, \; \mathbf{F}_v^{\text{ref}} = \sum_{u \in \mathcal{N}(v)} \mathbf{Z}_u^{\text{ref}} \quad (1)$$

where $\mathcal{N}(v)$ is a set of tokens (e.g., a group of semantically related expressions) associated with a focus v. In matrix notation, we rewrite Eq. (1) to $\mathbf{F}^{\text{hyp}} = \mathbf{A}^{\text{hyp}}\mathbf{Z}^{\text{hyp}}$, similarly for $\mathbf{F}^{\text{ref}}$.

Next, we measure the distance between a common set of foci Ω in a hypothesis and reference pair based on their embeddings:

$$\text{DS-FOCUS}(\text{hyp}, \text{ref}) = \frac{1}{\text{N}} \sum_{\text{u} \in \Omega} \|\mathbf{F}_{\text{u}}^{\text{hyp}} - \mathbf{F}_{\text{u}}^{\text{ref}}\| \quad (2)$$

where DS-FOCUS is scaled down by the factor of N, the number of foci in hypothesis.

Sentence Graph. Few contextualized encoders can produce high-quality sentence embeddings in the document context, as they do not model inter-dependence between sentences. According to Centering theory (Grosz et al., 1995), two sentences are marked continuous in meaning when they share at least one focus, on the one hand; one marks a meaning shift for two sentences when no focus appears in common, on the other hand. From this, one can aggregate sentence embeddings for which

corresponding sentences are considered continuous. In the following, we present a graph-based approach to do so.

For a hypothesis[3], let $\mathbf{S}^{\text{hyp}} \in \mathbb{R}^{n \times d}$ be a matrix of sentence embeddings with n and d as the number of sentences and the embedding size. We introduce a graph $\mathcal{G}^{\text{hyp}} = (\mathcal{V}, \mathbf{A}^{\text{hyp}})$ where $\mathcal{V}$ is a set of sentences and $\mathbf{A}^{\text{hyp}}$ is an adjacency matrix weighted according to the number of foci shared between sentences and the distance between sentences as listed below to depict two variants of $\mathbf{A}^{\text{hyp}}$:

- unweighted: $\mathbf{A}^{\text{hyp}}_{ij} = 1/(j-i)$ if the i-th and the j-th sentences have at least one focus in common (otherwise 0), where $j-i$ denotes the distance between two sentences and $\mathbf{A}^{\text{hyp}}_{ij} = 0$ when $j \leq i$.
- weighted: $\mathbf{A}^{\text{hyp}}_{ij} = a/(j-i)$, where a is the number of foci shared in the i-th and the j-th sentences, with the same constraints on j and i as above.

Analyses by Guinaudeau and Strube (2013) indicate that global statistics (e.g., average) over such adjacency matrices can distinguish incoherent from coherent text to some degree. Here we depict adjacency matrices as a form of sentence connectivity derived from focus transitions over sentences. We use them to aggregate sentence embeddings from hypothesis and from reference:

$$\hat{\mathbf{S}}^{\text{hyp}} = (\mathbf{A}^{\text{hyp}} + \mathbf{I})\mathbf{S}^{\text{hyp}}, \; \hat{\mathbf{S}}^{\text{ref}} = (\mathbf{A}^{\text{ref}} + \mathbf{I})\mathbf{S}^{\text{ref}}$$

where $\mathbf{I}$ is an identity matrix that adds a self-loop to a graph so as to include self-embeddings when updating them.

Next, we derive per graph an embedding with simple statistics from $\hat{\mathbf{S}}^{\text{hyp}}$ and $\hat{\mathbf{S}}^{\text{ref}}$, i.e., the concatenation of mean-max-min-sum embeddings. Finally, we compute the cosine similarity between two graph-level embeddings:

$$\text{DS-SENT}(\text{hyp}, \text{ref}) = \text{cosine}(\mathcal{G}^{\text{hyp}}, \mathcal{G}^{\text{ref}}) \quad (3)$$

Choice of Focus. In discourse, often four popular choices are used to describe a focus: (i) a noun that heads a NP (Barzilay and Lapata, 2008), (ii) a noun (Elsner and Charniak, 2011), (iii) a coreferent entity associated with a set of referring expressions (Guinaudeau and Strube, 2013) and (iv) a semantic entity associated with a set of lexical related words (Mesgar and Strube, 2016).

In this work, we investigate two focus choices: noun (NN) and semantic entity (Entity). Linguistically speaking, the latter is a lexical cohesion device in the form of repetition. From this, NN as focus yields few useful coherence signals but a lot of noise, while Entity as focus uses 'signal compression' by means of aggregation to produce better signals. To produce entities, we first extract all nouns in hypothesis (or reference), and aggregate them into different semantic entities if their cosine similarities based on Dep2Vec word embeddings (Levy and Goldberg, 2014) is greater than a threshold—assuming that nouns with high similarity refer to the same semantic entity.

4 Experiments

4.1 Evaluation Metrics

In the following, we list all of the evaluation metrics, and elaborate on them in Appendix A.1.

Non-discourse Metrics. We consider BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), SBERT (Reimers and Gurevych, 2019), S^3-pyr (Peyrard et al., 2017), BLEURT (Sellam et al., 2020), BARTScore (Yuan et al., 2021), PRISM (Thompson and Post, 2020).

Discourse Metrics. We consider RC and LC (Wong and Kit, 2012) and Lexical Chain (Gong et al., 2015). We consider two coherence models, EntityGraph (Guinaudeau and Strube, 2013) and LexicalGraph (Mesgar and Strube, 2016), and treat them as discourse metrics.

DiscoScore. DS-FOCUS can be parameterized with two focus choices: noun (NN) or semantic entity (Entity). DS-SENT can be parameterized not only with focus, but also with the choices of *unweighted* (-U) and *weighted* (-W). For DS-FOCUS, we use Conpono (Iter et al., 2020) that finetuned BERT with a novel discourse-level objective regarding sentence ordering. For DS-SENT, we use BERT-NLI. This is because we find this configuration performs best after initial trials—see Table 2 (appendix). Figure 5 (appendix) shows all variants of DiscoScore. Concerning the threshold of Dep2Vec to produce entities, after experimenting with several alternatives we set it to 0.8 for DS-FOCUS (Entity) in all setups, and to 0.8 in summarization and to 0.5 in MT for DS-SENT (Entity).

[3]For simplicity, we omit the notation $\mathbf{S}^{\text{ref}}$ and $\mathcal{G}^{\text{ref}}$ for a reference.

4.2 Datasets

We consider two datasets in summarization: SummEval (Fabbri et al., 2021) and NeR18 (Grusky et al., 2018), and one dataset in document-level MT: WMT20 (Mathur et al., 2020). Note that these datasets consist of hypotheses paired with human-written references, where hypotheses are machine-generated texts of varying qualities given by neural and non-neural, extractive and abstractive language models. We outline these datasets in Appendix A.2, and provide data statistics in Table 9 (appendix).

5 Results

We first examine the importance of discourse for evaluation metrics—which underpins the usefulness of discourse metrics, and then benchmark DiscoScore on summarization and MT datasets.

Importance of Discourse. DS-FOCUS and DS-SENT concern the modeling of discourse coherence on two different levels: (i) the occurrences of foci, and (ii) the interdependence between sentences driven by focus transitions, both reflecting the discourse characteristics of a text. In the following, we describe these discourse features, and examine their importance for assessing system outputs by contrasting the discourse patterns of hypothesis and reference.

- **Focus Frequency**, denoted by FREQ(x), equals the ratio between the total frequencies of foci and the number of foci in a text x, where x is hypothesis or reference. We exclude foci occurring only once.
- **Sentence Connectivity**, denoted by CONN(x), equals the average of all elements in adjacency matrix representing the interdependence between sentences in a text x (hypothesis/reference).
- As in DiscoScore, we consider two focus choices (NN and Entity) and the choices of *unweighted* (-U) and *weighted* (-W) for these discourse features. Figure 5 (appendix) shows the links between DiscoScore and the features.

Figure 2 shows that the scales on FREQ(ref) and FREQ(hyp) in summarization differ by a large amount, i.e., from 0.5 to 2.5 on y-axis and up to 6 on x-axis. This means that hypothesis and reference can be strongly distinguished by FREQ(x), which underpins the usefulness of including such

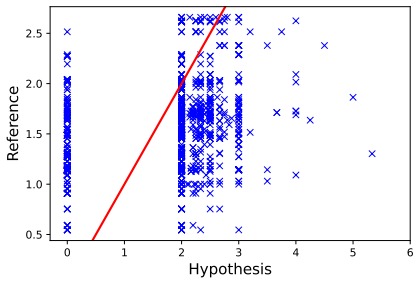


Figure 2: Scatter plot to display FREQ(hyp) (based on NN) on x-axis and FREQ(ref) on y-axis on SUMMEval. Each point contains two frequencies from a pair of hypothesis and reference. The points below the auxiliary line are the ones for which FREQ(hyp) > FREQ(ref).

discourse signals in the assessment of system outputs when references are available. Further, the larger scale on FREQ(hyp) indicates that foci in hypothesis are more repetitive than in reference, as a result of needless repetition in poor summaries—in line with previous studies on incoherent machine translations (Guillou, 2013; Voita et al., 2019). The results for other discourse features are similar, we provide them in Figure 6 (appendix).

Overall, these results show discourse features can separate hypothesis from reference.

5.1 Text Summarization

Correlation Results. Table 1 compares metrics on SUMMEval on system level. Most of non-discourse metrics have a lowest correlation with human rated coherence among four quality aspects. Even worse, ROUGE-L and SBERT do not correlate with coherence whatsoever. BARTScore, the recent state-of-the-art metric, is very weak when operated on system level, notwithstanding that it has been fine-tuned on "document-to-summary" parallel data from CNN/DailyMail—which SUMMEval is constructed from. We note that SUMMEval uses multiple references. BARTScore by default compares a hypothesis with one reference at a time, then takes the average of multiple evaluation scores as a final score. Table 8 (appendix) shows that we can improve system-level BARTScore to some degree by replacing 'average' with 'max' (i.e., taking the maximum score), but DS-FOCUS is still much better overall, i.e., surpassing BARTScore by ca. 10 points on average.

Table 7 (appendix) reports correlation results on NeR18 that uses single reference. We find that half of hypotheses do not contain 'good foci', and as such the foci-based discourse features outlined

Settings	Metrics	Coherence	Consistency	Fluency	Relevance	Average
	Non-discourse metrics					
$m(\text{hyp}, \text{ref})$	ROUGE-1	9.09	27.27	18.18	9.09	15.91
	ROUGE-L	0.00	36.36	21.21	18.18	18.94
	BERTScore	30.30	30.30	51.52	54.55	41.67
	MoverScore	36.36	42.42	63.64	60.61	50.76
	SBERT	3.03	33.33	30.30	27.27	23.48
	BLEURT	45.45	**51.52**	**72.73**	63.64	58.33
	BARTScore	60.61	36.36	45.45	48.48	47.73
	PRISM	51.52	39.39	**72.73**	69.70	58.33
	S^3-pyr	18.18	24.24	9.09	6.06	14.39
	Discourse metrics					
$m(\text{hyp})$	RC	45.45	**51.52**	**54.55**	57.58	**52.27**
	LC	**51.52**	45.45	48.48	57.58	50.76
	Entity Graph	42.42	12.12	15.15	18.18	21.97
	Lexical Graph	48.48	6.06	15.15	18.18	21.97
$m(\text{hyp}, \text{ref})$	Lexical Chain	42.42	6.06	9.09	18.18	18.94
	DS-FOCUS (NN)	**75.76**	**63.64**	**78.79**	**81.82**	**75.00**
	DS-FOCUS (Entity)	69.70	57.58	72.73	75.76	68.94
	DS-SENT-U (NN)	48.48	54.55	63.64	60.61	56.82
	DS-SENT-U (Entity)	54.55	60.61	75.76	66.67	64.39
	DS-SENT-W (NN)	51.52	51.52	66.67	63.64	58.33
	DS-SENT-W (Entity)	51.52	57.58	66.67	63.64	59.85

Table 1: System-level Kendall correlations between metrics and human ratings of summary quality on SUMMEval. We bold numbers that significantly outperform others according to paired t-test (Fisher et al., 1937). m is a metric.

previously are less discriminative on NeR18 than on SUMMEval—see Table 9 (appendix). However, DS-FOCUS is still strong, ca. 20 points better than BARTScore in all aspects, despite that DS-FOCUS uses a much smaller contextualized encoder[4]. We note that the 'F-score' version of DS-FOCUS seems extremely strong on NeR18, but it is not robust across datasets, e.g., much worse than the original, precision-based DS-FOCUS on SUMMEval.

On a side note, coherence (mostly) strongly correlates with the other rating aspects on both SUMMEval and NeR18—see Figure 3. Thus, it is not surprising that both DS-FOCUS and DS-SENT correlate well with these aspects, despite that we have not targeted them. While strong on system level, DiscoScore could not show advantages on summary level—see Table 5 (appendix), but we argue that system-level correlation deserves the highest priority as systems are compared in this manner.

Overall, these results show that BERT-based non-discourse metrics correlate weakly with human ratings on system level. BARTScore also does so, though we improve it to some degree in multi-references settings. DiscoScore, particularly DS-FOCUS, performs consistently best in both single- and multi-references settings, and it is equally strong in all aspects.

As for discourse metrics, RC and LC that use discourse features are strong baselines as they outperform most of non-discourse metrics and coherence models (i.e., Entity and Lexical Graph) without the access to source texts and references. However, they are worse than both DS-FOCUS and DS-SENT. This confirms the inadequacy of RC and LC in that they do not leverage strong contextualized encoders and judge hypothesis in the absence of references. Moreover, we compare DiscoScore to a combination of two strong, complementary baselines, BARTScore and RC—a simple solution to address text coherence of non-discourse metrics. To combine them, we simply average their scores. We see the gains are additive in all aspects but coherence. DS-FOCUS wins all the time by a large margin—see Table 10 (appendix).

Taken together, these results show that any of the three—(i) leveraging contextualized encoders as in BERT-based metrics and BARTScore; (ii) leveraging discourse features as in RC and (iii) the ensemble of (i) and (ii) by averaging—is not sufficient, suggesting to combine (i) and (ii) as DiscoScore does.

Understanding DiscoScore. As for all variants of DiscoScore, we provide understanding on why

[4]DS-FOCUS uses Conpono on the same size of BERTBase. BARTScore uses BARTLarge finetuned on CNN/DailyMail.

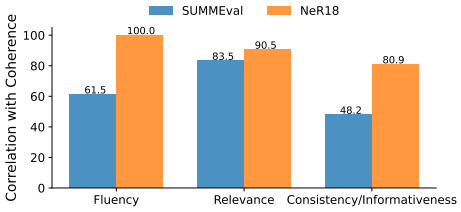


Figure 3: Pearson Correlation between coherence and other aspects on system level. SUMMEval and NeR18 use Consistency and Informativeness respectively.

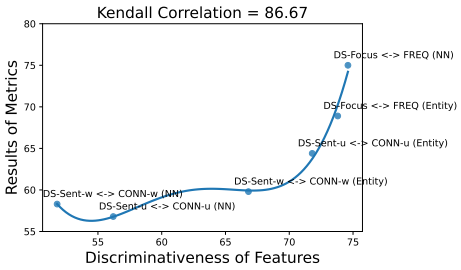


Figure 4: Correlations between the results of metrics and the discriminativeness of features on SUMMEval. Metric results are averaged across four rating aspects.

one variant is superior to another with the discourse features outlined in Figure 5 (appendix). To this end, we begin with defining the *discriminativeness* of these features as the magnitude of separating hypothesis from reference:

$$\mathcal{D}_{\mathcal{R}}(\text{hyp}, \text{ref}) := \frac{|\{(\text{hyp}, \text{ref}) | \mathcal{R}(\text{ref}) < \mathcal{R}(\text{hyp})\}|}{N} \quad (4)$$

where N is a normalization term, $\mathcal{R}$ is any one of the discourse features in Figure 5 (appendix).

Figure 4 shows that the discriminativeness of these features strongly correlate with the results of the DiscoScore variants, i.e., that the more discriminative the features are, the better the metrics perform. This attributes the superiority of a metric to the fact that the discourse feature can better separate hypothesis and reference.

From this, we can interpret the performance gaps between the DiscoScore variants, namely (i) DS-FOCUS over DS-SENT: given *Focus Frequency* is more discriminative than *Sentence Connectivity*, it is not surprising that DS-FOCUS modeling discourse coherence with the former outperforms DS-SENT modeling with the latter, and (ii) DS-Focus (NN) outperforms DS-Focus (Entity) because *Frequency (NN)* can better separate hypothesis from reference than *Frequency (Entity)*.

Analyses. We provide analyses on the configuration of DiscoScore from three perspectives—see Appendix A.3: (i) the choice of BERT variants towards discourse- versus non-discourse BERT; (ii) the impact of adjacency matrices accounting for the interdependence between sentences and (iii) that we compare statistics- and alignment-based approaches to examine the best configuration for DS-SENT. Our results show the advantages of adjacency matrices and statistics based approach, and that discourse BERT only helps for DS-FOCUS.

5.2 Document-level Machine Translation

Correlation Results. Table 12 (appendix) compares metrics on WMT20. We see that non-discourse metrics seem much better, but these results are not consistent to the discriminativeness of the discourse features—see Table 11 (appendix). For instance, in cs-en, the discourse features (Frequency and Connectivity) corresponding to DS-FOCUS and DS-SENT clearly separate hypothesis from reference due to the probability of $\mathcal{D} > 0$ being over 70%. However, both DS-FOCUS and DS-SENT correlate weakly with human rated adequacy. Recently, Freitag et al. (2021a) provide justification to the inadequacy of the 'adequacy' ratings, as 'adequacy' sometimes cannot distinguish human from system translations and correlates weakly with multiple aspects (e.g., fluency and accuracy). Thus, they re-annotate WMT20 with the MQM and pSQM rating schemes, which has been subsumed into the annotation guideline of the most recent WMT evaluation campaign (Freitag et al., 2021b). Here, we perform an extra study on these ratings on both document- and system-levels. Note that system-level ratings are said to be the average of document-level ones in our setting. Table 6 (appendix) shows that DS-SENT is much better than BARTScore on system level, surpassing it by 25 points in terms of MQM and 14 points in pSQM.

Overall, these results in MT are consistent with those in summarization, i.e., DiscoScore is strong on system levels for both tasks, but it cannot show gains on fine-grained levels. Section A.4 (appendix) show inter-correlations between metrics.

6 Conclusions

Given the recent growth in discourse based NLG systems, evaluation metrics targeting the assessment of text coherence are essential next steps for properly tracking the progress of these systems.

Although there have been several attempts made towards discourse metrics, they all do not leverage strong contextualized encoders which have been held responsible for the recent success story of NLP. In this work, we introduced DiscoScore that uses BERT to model discourse coherence from two perspectives of readers' focus: (i) frequencies and semantics of foci and (ii) focus transitions over sentences used to predict interdependence between sentences. We find that BERT-based non-discourse metrics cannot address text coherence, even much worse than early feature-based discourse metrics invented a decade ago. We also find that the recent state-of-the-art BARTScore correlates weakly with human ratings on system level. DiscoScore, on the other hand, performs consistently best in both single- and multi-reference settings, equally strong in coherence and several other aspects such as factual consistency, despite that we have not targeted them. More prominently, we provide understanding on the importance of discourse for evaluation metrics, and explain the superiority of one metric over another with simple features, in line with recent work on explainability for evaluation metrics (Kaster et al., 2021; Fomicheva et al., 2021).

Scope for future research is huge, e.g., developing reference-free discourse metrics comparing source text to hypothesis, improving discourse metrics on fine-grained levels[5], and ranking NLG systems via discourse metrics and rigorous approaches (Peyrard et al., 2021; Kocmi et al., 2021).

7 Impact and Limitation

To our knowledge, we, for the first time, combine the elements of discourse and BERT representations to design an evaluation metric (DiscoScore) for text quality assessment in summarization and MT. While our experiments are conducted on English datasets, DiscoScore could adapt to many other languages in which references and foci are available. We believe that this work fosters future research on text generation systems endowed with the ability to produce well-formed texts in discourse.

However, we acknowledge several limitations of this work, which require further investigation in future. We now discuss them in the following:

Entity as Focus. We follow the idea of Mesgar and Strube (2016) in the discourse community, which clusters nouns into entities based on their static word embeddings. Although simple, it sometimes helps for DiscoScore. However, alternatives aiming to produce better entities have not been explored in this work, e.g., replacing static with contextualized embeddings, and weighting entities by their occurrences in hypothesis/reference.

Weakness on Fine-Grained Assessment. In summarization and MT, we show that our novel DiscoScore largely outperforms the current state-of-the-art BARTScore on system levels for both tasks, while it cannot show advantages on finer-grained levels such as document- and summary-levels. This might be because modeling focus alone is insufficient to perform much more challenging, finer-grained assessment of text quality. Future work could also factor other discourse phenomena (e.g., discourse connectives and coreference) into the assessment of text coherence.

[5]Recently, Steen and Markert (2022) introduce a fine-grained evaluation setup to compute summary-level correlation, which performs computing over summaries not produced by multiple systems, but rather by a single system. This is because systems sometimes substantially differ in quality, which implies that involving multiple systems could result in inaccurate evaluation outcomes in the presence of system-level confounders.

Acknowledgments

We thank the anonymous reviewers for their thoughtful comments that greatly improved the texts. This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1 and the Klaus Tschira Foundation, Heidelberg, Germany. Steffen Eger is funded by DFG Heisenberg grant EG 375/5-1.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bruno Cartoni, Jindřich Libovický, and Thomas Brovelli (Meyer), editors. 2018. *Machine Translation Evaluation beyond the Sentence Level*. Alicante, Spain.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 649–662, Hong Kong, China. Association for Computational Linguistics.

Wang Chen, Piji Li, and Irwin King. 2021. A training-free and reference-free summarization evaluation metric via centrality-weighted relevance and self-referenced redundancy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 404–414, Online. Association for Computational Linguistics.

Elisabet Comelles, Jesús Giménez, Lluís Màrquez, Irene Castellón, and Victoria Arranz. 2010. Document-level automatic MT evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 333–338, Uppsala, Sweden. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Ronald Aylmer Fisher et al. 1937. The design of experiments. *The design of experiments.*, (2nd Ed).

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2015. Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Barbara J Grosz et al. 1977. The representation and use of focus in a system for understanding dialogs. In *IJCAI*, volume 67, page 76. Citeseer.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the*

2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Liane Guillou. 2013. Analysing lexical consistency in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 10–18, Sofia, Bulgaria. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4859–4870, Online. Association for Computational Linguistics.

Yuchen Jiang, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, and Ming Zhou. 2021. Blond: An automatic evaluation metric for document-level machinetranslation. *CoRR*, abs/2103.11878.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2017. Discourse structure in machine translation evaluation. *Computational Linguistics*, 43(4):683–722.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of BERT-based evaluation metrics by disentangling along linguistic factors. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8912–8925, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *CoRR*, abs/2107.10821.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. Can transformer models measure coherence in text: Re-thinking the shuffle test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd*

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 302–308.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of ACL workshop on Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online. Association for Computational Linguistics.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Thomas Mcdonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online.

Mohsen Mesgar, Leonardo F. R. Ribeiro, and Iryna Gurevych. 2021. A neural graph-based local coherence model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2316–2321, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1414–1423, San Diego, California. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Han Cheol Moon, Tasnim Mohiuddin, Shafiq Joty, and Chi Xu. 2019. A unified neural coherence model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2262–2272, Hong Kong, China. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*,

pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of NLP systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2301–2315, Online. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2020. Using context in neural machine translation training objectives. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7764–7770, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2022. How to find strong summary coherence measures? a toolbox and a comparative study for summary coherence measure evaluation. *arXiv preprint arXiv:2209.06517*.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Dat Tien Nguyen and Shafiq Joty. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, Vancouver, Canada. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020,*

Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

A Appendix

A.1 Evaluation Metrics

Non-discourse Metrics. We consider the following non-discourse metrics.

- BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are precision- and recall-oriented metrics respectively, both of which measure n-gram overlap between a hypothesis and a reference.
- BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019) are set-based metrics used to measure the semantic similarity between hypothesis and reference. BERTScore uses greedy alignment to compute the similarity between two sets of BERT-based word embeddings from hypothesis and from reference, while MoverScore uses optimal alignments based on Word Mover's Distance (Kusner et al., 2015) to do so.
- SBERT (Reimers and Gurevych, 2019) fine-tunes BERT on the NLI datasets and uses pooling operations to produce sentence embeddings. We compute the cosine similarity between two sentence representations from hypothesis and from reference.
- S^3-pyr and S^3-resp (Peyrard et al., 2017) are supervised metrics that linearly combine ROUGE, JS-divergence and ROUGE-WE scores, trained on the TAC datasets with human annotated pyramid and responsiveness scores as supervision.
- BLEURT (Sellam et al., 2020) is another supervised metric that fine-tunes BERT on the concatenation of WMT datasets and synthetic data in the MT domain, with human judgment of translation quality as supervision.
- BARTScore (Yuan et al., 2021) and PRISM (Thompson and Post, 2020) depict sequence-to-sequence language models as metrics to compare hypothesis with reference. In reference-based settings, they both measure the likelihood that hypothesis and reference are paraphrases, but differ in the language models they rely on. PRISM has been based on a neural MT system trained from scratch on parallel data in MT, while BARTScore uses BART (Yuan et al., 2021) that has been fine-tuned on CNN/DailyMail (Hermann et al., 2015)—which is parallel data in summarization. We use the 'F-score' version of BARTScore as recommended in Yuan et al. (2021).

Discourse Metrics. We consider the following discourse metrics (including ours and coherence models).

- RC and LC (Wong and Kit, 2012) require neither source texts nor references and use lexical cohesion devices (e.g., repetition) within a hypothesis to predict text coherence. LC computes the proportion of words within hypothesis that are lexical cohesion devices, while RC computes the proportion of times that lexical cohesion devices appear in hypothesis.
- Entity Graph (Guinaudeau and Strube, 2013) and Lexical Graph (Mesgar and Strube, 2016) are popular coherence models used to perform discourse tasks such as essay scoring, both of which introduce a graph with nodes as sentences and adjacency matrices as the connectivity between sentences. Here, we use the average of adjacency matrices from the hypothesis as the proxy of hypothesis coherence. While Entity Graph draws an edge between two sentences if both sentences have at least one noun in common, Lexical Graph draws an edge if two sentences have a pair of similar words in common, i.e., the cosine similarity between their embeddings greater than a threshold.
- Lexical Chain (Gong et al., 2015) extracts multiple lexical chains from hypothesis and from reference. Each word is associated to a lexical chain if a word appears in more than one sentence. A lexical chain contains a set of sentence positions in which a word appears. Finally, the metric performs soft matching to measure lexical chain overlap between hypothesis and reference.
- FocusDiff and SentGraph are the two variants of DiscoScore, which use BERT to model semantics and coherence of readers' focus in hypothesis and reference. In particular, FocusDiff measures the difference between a common set of foci in hypothesis and reference in

terms of semantics and frequency, while SentGraph measures the semantic similarity between two sets of sentence embeddings from hypothesis and reference—which are aggregated according to the number of foci shared across sentences and the distance between sentences.

A.2 Datasets

We outline two datasets in summarization, and one in document-level MT.

Text Summarization. While DUC[6] and TAC[7] datasets with human rated summaries, constructed one decade ago, were the standard benchmarks for comparing evaluation metrics in summmarization, they collect summaries only from extractive summarization systems. In the last few years, abstractive systems have become popular; however, little is known how well metrics judge them. Recently, several datasets based on CNN/DailyMail have been constructed to address this. For instance, SummEval (Fabbri et al., 2021), REALSumm (Bhandari et al., 2020), XSum (Maynez et al., 2020) and FEQA (Durmus et al., 2020) all collect summaries from both extractive and abstractive systems, but differ in the aspects human experts rate summaries. In this work, we consider the following two complementary summarization datasets.

- SummEval has been constructed in multiple-references settings, i.e., that each hypothesis is associated to multiple references. It contains human judgments of summary coherence, factual consistency, fluency and relevance. We only consider abstractive summaries as they have little lexical overlap with references.
- NeR18 (Grusky et al., 2018), in contrast, has been constructed in single-reference settings. It contains human judgments of summary coherence, fluency, informativeness and relevance. As majority of summaries are extractive, we include both extractive and abstractive for the inclusive picture.

Document-level Machine Translation. As document-level human ratings in MT are particularly laborious, hardly ever have there been MT datasets directly addressing them. First attempts suggested to use the average of much cheaper sentence-level ratings as a document score for comparing document-level metrics (Comelles et al., 2010; Wong and Kit, 2012; Gong et al., 2015). However, human experts were asked to rate sentences in isolation within a document. Thus, human ratings at both sentence and document levels cannot reflect inter-sentence coherence. Recently, the WMT20 workshop (Mathur et al., 2020) asks humans to rate each sentence translation in the document context, and follows the previous idea of 'average' to yield document scores.

In this work, we use the WMT20 dataset with 'artificial' document-level ratings. Note that WMT20 comes with two issues: (i) though sentences are rated in the document context, averaging sentence-level ratings may zero out negative effects of incoherent elements on document level and (ii) unlike SummEval and NeR18, WMT20 only contains human judgment of translation *adequacy* (which may subsume multiple aspects), not *coherence*.

For simplicity, we exclude system and reference translations with lengths greater than 512—the number of tokens at maximum allowed by BERT, as only a small portion of instances is over the token limit. Note that it is effortless to replace BERT with Longformer (Beltagy et al., 2020) to deal with longer documents for DiscoScore.

[6] https://duc.nist.gov/data.html
[7] https://tac.nist.gov/data/

Metrics	Encoders	Average
DS-FOCUS (NN)	+ BERT	71.97
	+ BERT-NLI	70.45
	+ Conpono	75.00
DS-SENT-U (NN)	+ BERT	35.61
	+ BERT-NLI	56.82
	+ Conpono	23.48

Table 2: Results of three contextualized encoders on SUMMEval. Results are averaged across four aspects.

Metrics	Average
DS-SENT-U (NN)	56.82
w/o sentence aggregation	46.21

Table 3: Ablation study on the use of adjacency matrix to aggregate sentence embeddings on SUMMEval.

A.3 Analyses on Text Summarization

Choice of BERT Variants. Table 2 compares the impact of three BERT variants on DiscoScore. Conpono, referred to as a discourse BERT, has fine-tuned BERT with a novel discourse-level objective regarding sentence ordering. While strong on discourse evaluation benchmarks (Chen et al., 2019),

Metrics	Mechanisms	Average
DS-SENT-U (NN)	+ greedy align	21.97
	+ optimal align	26.52
	+ mean-max-min-sum	56.82

Table 4: Averaged results of SentGraph variants based on three mechanisms on SUMMEval.

Metrics	SUMMEval	NeR18
BARTScore	14.13	24.78
PRISM	14.92	18.89
DS-FOCUS (NN)	10.81	10.42
DS-SENT-U (NN)	15.71	3.81

Table 5: Summary-level averaged Kendall correlations across all rating aspects.

Conpono is not always helpful, e.g., BERT-NLI is better for DS-SENT. These results suggest the best configuration for DiscoScore.

Impact of Sentence Connectivity. Table 3 shows an ablation study on the use of sentence connectivity. Aggregating sentence embeddings with our adjacency matrices (see Eq.3) helps considerably. This confirms the usefulness of aggregation from which we include coherence signals in sentence embeddings.

SentGraph Variants. Table 4 compares three DS-SENT variants as to how we measure the distance between two sets of sentence embeddings from hypothesis and reference. In particular, we refer to BERTScore (Zhang et al., 2020) as a 'greedy align' mechanism used to compute the similarity between two sets of sentence embeddings. As for 'optimal align', we use MoverScore (Zhao et al., 2019) to do so. While the two alignments directly measure the distance between the two sets, the simple statistics, i.e., mean-max-min-sum, derives a graph embedding from each set and computes the cosine similarity between two graph embeddings. We see that the 'statistics' wins by a big margin, and thus adopt this DS-SENT variant in all setups.

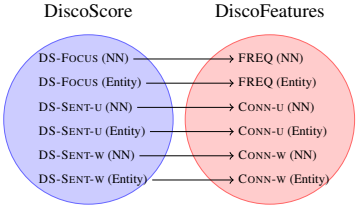


Figure 5: Links between the DiscoScore variants and discourse features.

Metrics	Sys-level MQM	Sys-level pSQM	Doc-level MQM	Doc-level pSQM
BARTScore	45.57	55.50	**34.90**	**28.96**
*DS-FOCUS (NN)	42.12	40.89	19.10	9.98
DS-SENT-U (NN)	**70.77**	**69.74**	19.98	14.49

Table 6: Document-level Kendall and system-level Pearson correlations between metrics and MQM/pSQM ratings on WMT20 in Chinese-to-English—which is the only language pair with such ratings in reference-based settings. *DS-FOCUS (NN) excludes focus that occurs only once in hypothesis/reference.

A.4 Analyses on MT

Correlation between Metrics. Figure 7 shows inter-correlations between metrics on WMT20 across languages. Overall, correlations are mostly high between non-discourse metrics, much weaker between discourse and non-discourse metrics—which confirms the orthogonality of them in that they rate translations in different aspects. We note that DS-FOCUS has the lowest correlations with all other metrics. For instance, DS-FOCUS is almost orthogonal to BERTScore and MoverScore. We investigated whether combining them receives additive gains. We find that a combination of DS-FOCUS and BERTScore (or MoverScore) provides little help in correlation with adequacy.

Settings	Metrics	Coherence	Fluency	Informative	Relevance	Average
$m(\text{hyp}, \text{ref})$	BARTScore	42.58	42.58	23.80	33.33	35.57
	PRISM	51.52	42.58	42.86	52.38	47.33
	DS-FOCUS (NN)	61.90	61.90	42.86	52.38	54.76
	DS-FOCUS* (NN)	**80.95**	**80.95**	**100.00**	**90.47**	**88.09**
	DS-SENT-U (NN)	14.29	14.29	14.29	23.81	16.67

Table 7: System-level Kendall correlations between metrics and human ratings on NeR18. DS-FOCUS* is the 'F-score' version of DS-FOCUS.

Settings	Metrics	Coherence	Consistency	Fluency	Relevance	Average
$m(\text{hyp}, \text{ref})$	BARTScore (max)	**78.79**	48.48	63.64	72.73	65.91
	BARTScore (original)	60.61	36.36	45.45	48.48	47.73
	FocusDiff (NN)	**75.76**	**63.64**	**78.79**	**81.82**	**75.00**
	FocusDiff (Entity)	69.70	57.58	72.73	75.76	68.94
	SentGraph-u (NN)	48.48	54.55	63.64	60.61	56.82
	SentGraph-u (Entity)	54.55	60.61	75.76	66.67	64.39

Table 8: System-level Kendall correlations between metrics and human ratings on SUMMEval in multi-reference settings. BARTScore (original) compares a hypothesis with one reference at a time, and takes the average of evaluation scores as a final score, while BARTScore (max) takes the maximum score.

			WMT20			
	SUMMEval	NeR18	cs-en	de-en	ja-en	ru-en
Number of references	11	1	1	1	1	1
Number of systems	12	7	13	14	11	13
Number of hypothesis per system	100	60	102	118	80	91
Number of sentences per hypothesis	3.13	1.90	15.21	13.84	11.29	9.46
Average number of foci in hypothesis	15.18	12.85	62.01	56.68	57.09	44.99
Average number of 'good foci' in hypothesis	2.47	2.56	13.16	13.37	15.07	9.95
Percent of hypotheses with 'good foci'	80.50%	43.80%	100%	98.60%	100%	100%

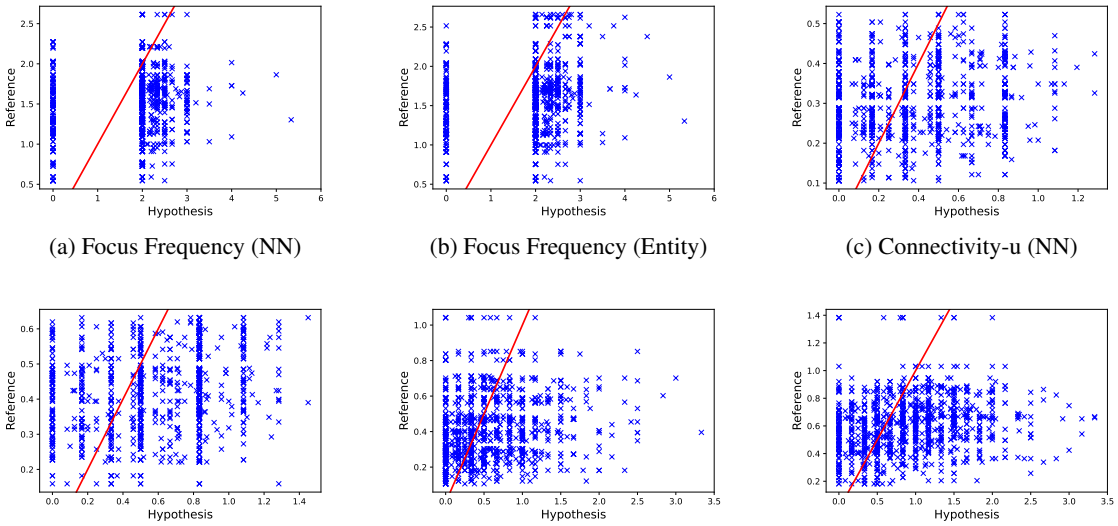
Table 9: Characteristics of summarization and MT datasets. 'good foci' denotes a focus appearing more than once in hypothesis. The more often a focus appears, the stronger the discourse signals are.

Metrics	Coherence	Consistency	Fluency	Relevance	Average
RC	45.45	51.52	54.55	57.58	52.27
BARTScore (max)	**78.79**	48.48	63.64	72.73	65.91
BARTScore (max) + RC	66.67	54.55	69.70	78.79	67.42
DS-FOCUS (NN)	75.76	**63.64**	**78.79**	**81.82**	**75.00**

Table 10: Ensemble of non-discourse and discourse metrics (BARTScore + RC) vs DiscoScore.

	cs-en			de-en			ja-en			ru-en		
DiscoFeatures	$\mathcal{D}>0$	$\mathcal{D}=0$	$\mathcal{D}<0$	$\mathcal{D}>0$	$\mathcal{D}=0$	$\mathcal{D}<0$	$\mathcal{D}>0$	$\mathcal{D}=0$	$\mathcal{D}<0$	$\mathcal{D}>0$	$\mathcal{D}=0$	$\mathcal{D}<0$
Frequency (NN)	74.18	2.00	23.82	57.38	9.65	32.97	53.04	2.63	44.33	52.77	7.31	39.92
Frequency (Entity)	76.17	1.76	22.07	59.74	8.38	31.88	52.38	1.48	46.14	53.61	7.31	39.08
Connectivity-u (NN)	78.05	0.35	21.60	63.11	8.29	28.60	59.61	5.25	35.14	52.04	10.03	37.93
Connectivity-u (Entity)	79.46	0.35	20.19	62.02	8.20	29.78	59.44	5.09	35.47	52.87	9.40	37.72
Connectivity-w (NN)	77.93	0.24	21.83	64.85	4.64	30.51	59.12	0.49	40.39	59.98	5.12	34.90
Connectivity-w (Entity)	80.40	0.23	19.37	63.48	4.73	31.79	60.76	0.33	38.91	60.82	4.60	34.58

Table 11: Statistics of discourse features on WMT20. $\mathcal{D}>0$ denotes the percent of 'reference-hypothesis' pairs for which $\mathcal{R}(\text{ref}) > \mathcal{R}(\text{hyp})$ with $\mathcal{R}$ as any one of these features, similarly for the definitions of $\mathcal{D}=0$ and $\mathcal{D}<0$. We exclude the pairs for which hypothesis and reference are the exact same.



(a) Focus Frequency (NN) (b) Focus Frequency (Entity) (c) Connectivity-u (NN)

(d) Connectivity-u (Entity) (e) Connectivity-w (NN) (f) Connectivity-w (Entity)

Figure 6: Distribution of discourse features over hypothesis and reference on SUMMEval.

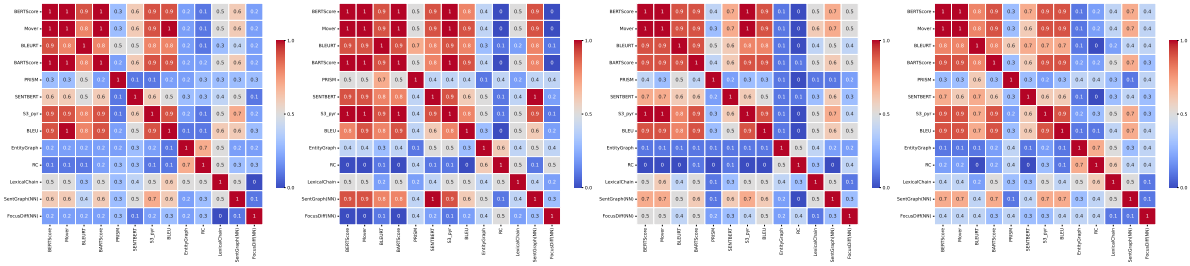


Figure 7: Pearson Correlations between metrics on WMT20 in cs-en, de-en, ja-en and ru-en (from left to right).

Settings	Metrics	Direct Assessment (Adequacy) cs-en	de-en	ja-en	ru-en	Average
	Non-discourse metrics					
$m(\text{hyp}, \text{ref})$	BLEU	7.44	57.52	41.48	10.74	29.30
	BERTScore	10.82	60.38	**46.95**	13.08	32.81
	MoverScore	**15.40**	**61.69**	42.12	13.78	**33.25**
	BARTScore	10.82	60.26	46.30	14.95	33.09
	PRISM	8.64	58.83	32.48	15.42	28.84
	SBERT	13.20	55.26	33.44	10.04	27.99
	BLEURT	12.01	58.83	37.94	**18.22**	31.75
	S^3-pyr	6.25	58.83	42.44	13.78	30.33
	S^3-resp	5.85	58.59	**47.26**	14.71	31.61
	Discourse metrics					
$m(\text{hyp})$	RC	5.85	7.19	**8.68**	9.34	7.77
	LC	**9.23**	1.72	3.53	6.07	5.14
	Entity Graph	5.06	**43.24**	3.53	10.51	15.59
	Lexical Graph	2.28	**43.60**	5.14	**13.55**	**16.15**
	Discourse metrics					
$m(\text{hyp}, \text{ref})$	Lexical Chain	21.54	35.15	15.11	16.12	21.99
	FocusDiff (NN)	7.64	33.13	19.29	2.57	15.66
	FocusDiff (Entity)	6.45	33.73	19.94	1.64	15.44
	SentGraph-u (NN)	7.64	57.16	39.22	18.22	**30.56**
	SentGraph-u (Entity)	7.65	57.17	39.23	18.22	**30.57**
	SentGraph-w (NN)	7.65	57.18	39.22	18.21	**30.57**
	SentGraph-w (Entity)	7.65	57.17	39.23	18.22	**30.57**

Table 12: Document-level Kendall correlations between metrics and human rated translation quality on WMT20.

Chapter 7

Towards Scalable and Reliable Capsule Networks for Challenging NLP Applications

Towards Scalable and Reliable Capsule Networks for Challenging NLP Applications

Wei Zhao[†], **Haiyun Peng**[‡], **Steffen Eger**[†], **Erik Cambria**[‡] **and Min Yang**[Φ]
[†] Computer Science Department, Technische Universität Darmstadt, Germany
[‡] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[Φ] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
www.aiphes.tu-darmstadt.de

Abstract

Obstacles hindering the development of capsule networks for challenging NLP applications include poor scalability to large output spaces and less reliable routing processes. In this paper, we introduce (i) an agreement score to evaluate the performance of routing processes at instance level; (ii) an adaptive optimizer to enhance the reliability of routing; (iii) capsule compression and partial routing to improve the scalability of capsule networks. We validate our approach on two NLP tasks, namely: multi-label text classification and question answering. Experimental results show that our approach considerably improves over strong competitors on both tasks. In addition, we gain the best results in low-resource settings with few training instances.[1]

1 Introduction

In recent years, deep neural networks have achieved outstanding success in natural language processing (NLP), computer vision and speech recognition. However, these deep models are data-hungry and generalize poorly from small datasets, very much unlike humans (Lake et al., 2015).

This is an important issue in NLP since sentences with different surface forms can convey the same meaning (paraphrases) and not all of them can be enumerated in the training set. For example, *Peter did not accept the offer* and *Peter turned down the offer* are semantically equivalent, but use different surface realizations.

In image classification, progress on the generalization ability of deep networks has been made by capsule networks (Sabour et al., 2017; Hinton et al., 2018). They are capable of generalizing to the same object in different 3D images with various viewpoints.

[1] Our code is publicly available at http://bit.ly/311Dcod

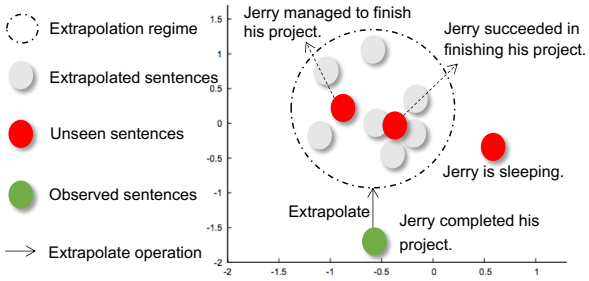


Figure 1: The extrapolation regime for an observed sentence can be found during training. Then, the unseen sentences in this regime may be generalized successfully.

Such generalization capability can be learned from examples with few viewpoints by extrapolation (Hinton et al., 2011). This suggests that capsule networks can similarly abstract away from different surface realizations in NLP applications.

Figure 1 illustrates this idea of how observed sentences in the training set are generalized to unseen sentences by extrapolation. In contrast, traditional neural networks require massive amounts of training samples for generalization. This is especially true in the case of convolutional neural networks (CNNs), where pooling operations wrongly discard positional information and do not consider hierarchical relationships between local features (Sabour et al., 2017).

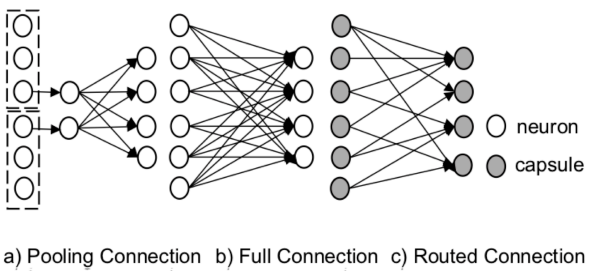


Figure 2: Outputs attend to a) active neurons found by pooling operations b) all neurons c) relevant capsules found in routing processes.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1549–1559
Florence, Italy, July 28 - August 2, 2019.

Capsule networks, instead, have the potential for learning hierarchical relationships between consecutive layers by using routing processes without parameters, which are clustering-like methods (Sabour et al., 2017) and additionally improve the generalization capability. We contrast such routing processes with pooling and fully connected layers in Figure 2.

Despite some recent success in NLP tasks (Wang et al., 2018; Xia et al., 2018; Xiao et al., 2018; Zhang et al., 2018a; Zhao et al., 2018), a few important obstacles still hinder the development of capsule networks for mature NLP applications.

For example, selecting the number of iterations is crucial for routing processes, because they iteratively route low-level capsules to high-level capsules in order to learn hierarchical relationships between layers. However, existing routing algorithms use the same number of iterations for all examples, which is not reliable to judge the convergence of routing. As shown in Figure 3, a routing process with five iterations on all examples converges to a lower training loss at system level, but on instance level for one example, convergence has still not obtained.

Additionally, training capsule networks is more difficult than traditional neural networks like CNN and long short-term memory (LSTM) due to the large number of capsules and potentially large output spaces, which requires extensive computational resources in the routing process.

In this work, we address these issues via the following contributions:

- We formulate routing processes as a proxy problem minimizing a total negative agreement score in order to evaluate how routing processes perform at instance level, which will be discussed more in depth later.
- We introduce an adaptive optimizer to self-adjust the number of iterations for each example in order to improve instance-level convergence and enhance the reliability of routing processes.
- We present capsule compression and partial routing to achieve better scalability of capsule networks on datasets with large output spaces.
- Our framework outperforms strong baselines on multi-label text classification and question answering. We also demonstrate its superior generalization capability in low-resource settings.

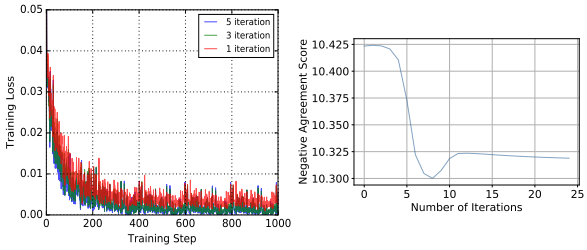


Figure 3: left) System-level routing evaluation on all examples; right) Instance-level routing evaluation on one example.

2 NLP-Capsule Framework

We have motivated the need for better capsule networks being capable of scaling to large output spaces and higher reliability for routing processes at instance level. We now build a unified capsule framework, which we call NLP-Capsule. It is shown in Figure 4 and described below.

2.1 Convolutional Layer

We use a convolutional operation to extract features from documents by taking a sliding window over document embeddings.

Let $\boldsymbol{X} \in \mathbb{R}^{l \times v}$ be a matrix of stacked v-dimensional word embeddings for an input document with l tokens. Furthermore, let $\boldsymbol{W}^a \in \mathbb{R}^{l \times k}$ be a convolutional filter with a width k. We apply this filter to a local region $\boldsymbol{X}^{\intercal}_{i:i+k-1} \in \mathbb{R}^{k \times l}$ to generate one feature:

$$m_i = f(\boldsymbol{W}^a \circ \boldsymbol{X}^{\intercal}_{i:i+k-1})$$

where $\circ$ denotes element-wise multiplication, and f is a nonlinear activation function (i.e., ReLU). For ease of exposition, we omit all bias terms.

Then, we can collect all m_i into one feature map $(m_1, \ldots, m_{(v-k+1)/2})$ after sliding the filter over the current document. To increase the diversity of features extraction, we concatenate multiple feature maps extracted by three filters with different window sizes (2,4,8) and pass them to the primary capsule layer.

2.2 Primary Capsule Layer

In this layer, we use a group-convolution operation to transform feature maps into primary capsules. As opposed to using a scalar for each element in the feature maps, capsules use a group of neurons to represent each element in the current layer, which has the potential for preserving more information.

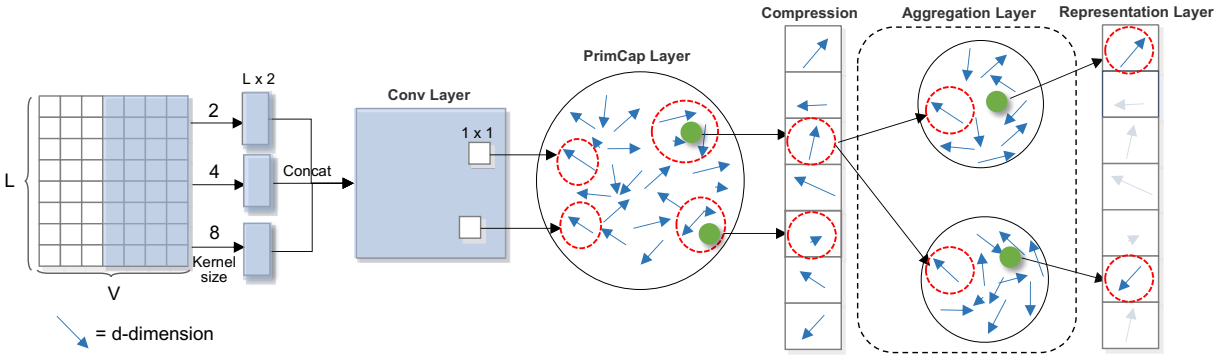


Figure 4: An illustration of NLP-Capsule framework.

Using 1×1 filters $\boldsymbol{W}^b = \{w_1, ..., w_d\} \in \mathbb{R}^d$, in total d groups are used to transform each scalar m_i in feature maps to one capsule $\boldsymbol{p}_i$, a d- dimensional vector, denoted as:

$$\boldsymbol{p}_i = g(p_{i1} \oplus p_{i2} \oplus \cdots \oplus p_{id}) \in \mathbb{R}^d$$

where $p_{ij} = m_i \cdot w_j \in \mathbb{R}$ and $\oplus$ is the concatenation operator. Furthermore, g is a non-linear function (i.e., squashing function). The length $||\boldsymbol{p}_i||$ of each capsule $\boldsymbol{p}_i$ indicates the probability of it being useful for the task at hand. Hence, a capsule's length has to be constrained into the unit interval $[0, 1]$ by the squashing function g:

$$g(\boldsymbol{x}) = \frac{||\boldsymbol{x}||^2}{1 + ||\boldsymbol{x}||^2} \frac{\boldsymbol{x}}{||\boldsymbol{x}||}$$

Capsule Compression One major issue in this layer is that the number of primary capsules becomes large in proportion to the size of the input documents, which requires extensive computational resources in routing processes (see Section 2.3). To mitigate this issue, we condense the large number of primary capsules into a smaller amount. In this way, we can merge similar capsules and remove outliers. Each condensed capsule $\boldsymbol{u}_i$ is calculated by using a weighted sum over all primary capsules, denoted as:

$$\hat{\boldsymbol{u}}_i = \sum_j b_j \boldsymbol{p}_j \in \mathbb{R}^d$$

where the parameter b_j is learned by supervision.

2.3 Aggregation Layer

Pooling is the simplest aggregation function routing condensed capsules into the subsequent layer, but it loses almost all information during aggregation. Alternatively, routing processes are introduced to iteratively route condensed capsules into the next layer for learning hierarchical relationships between two consecutive layers. We now describe this iterative routing algorithm. Let $\{\boldsymbol{u}_1, \ldots, \hat{\boldsymbol{u}}_m\}$ and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a set of condensed capsules in layer ℓ and a set of high-level capsules in layer $\ell+1$, respectively. The basic idea of routing is two-fold.

First, we transform the condensed capsules into a collection of candidates $\{\hat{\boldsymbol{u}}_{j|1}, \ldots, \hat{\boldsymbol{u}}_{j|m}\}$ for the j-th high-level capsule in layer $\ell + 1$. Following Sabour et al. (2017), each element $\hat{\boldsymbol{u}}_{j|i}$ is calculated by:

$$\hat{\boldsymbol{u}}_{j|i} = \boldsymbol{W}^c \boldsymbol{u}_i \in \mathbb{R}^d$$

where $\boldsymbol{W}^c$ is a linear transformation matrix.

Then, we represent a high-level capsule $\boldsymbol{v}_j$ by a weighted sum over those candidates, denoted as:

$$\boldsymbol{v}_j = \sum_{i=1}^{m} c_{ij} \hat{\boldsymbol{u}}_{j|i}$$

where c_{ij} is a coupling coefficient iteratively updated by a clustering-like method.

Our Routing As discussed earlier, routing algorithms like dynamic routing (Sabour et al., 2017) and EM routing (Hinton et al., 2018), which use the same number of iterations for all samples, perform well according to training loss at system level, but on instance level for individual examples, convergence has still not been reached. This increases the risk of unreliability for routing processes (see Figure 3).

To evaluate the performance of routing processes at instance level, we formulate them as a proxy problem minimizing the negative agreement score (NAS) function:

$$\min_{c,\boldsymbol{v}} f(\boldsymbol{u}) = -\sum_{i,j} c_{ij} \langle \boldsymbol{v}_j, \boldsymbol{u}_{j|i} \rangle$$

$$\text{s.t.} \quad \forall i, j : c_{ij} > 0, \quad \sum_j c_{ij} = 1.$$

The basic intuition behind this is to assign higher weights c_{ij} to one agreeable pair $\langle \boldsymbol{v}_j, \boldsymbol{u}_{j|i} \rangle$ if the capsule $\boldsymbol{v}_j$ and $\boldsymbol{u}_{j|i}$ are close to each other such that the total agreement score $\sum_{i,j} c_{ij} \langle \boldsymbol{v}_j, \boldsymbol{u}_{j|i} \rangle$ is maximized. However, the choice of NAS functions remains an open problem. Hinton et al. (2018) hypothesize that the agreeable pairs in NAS functions are from Gaussian distributions. Instead, we study NAS functions by introducing Kernel Density Estimation (KDE) since this yields a non-parametric density estimator requiring no assumptions that the agreeable pairs are drawn from parametric distributions. Here, we formulate the NAS function in a KDE form.

$$\min_{c,\boldsymbol{v}} f(\boldsymbol{u}) = -\sum_{i,j} c_{ij} k(d(\boldsymbol{v}_j, \boldsymbol{u}_{j|i})) \qquad (1)$$

where d is a distance metric with ℓ_2 norm, and k is a Epanechnikov kernel function (Wand and Jones, 1994) with:

$$k(x) = \begin{cases} 1 - x & x \in [0, 1) \\ 0 & x \geq 1 \end{cases}$$

The solution we used for KDE is taking Mean Shift (Comaniciu and Meer, 2002) to minimize the NAS function $f(\boldsymbol{u})$:

$$\nabla f(\boldsymbol{u}) = \sum_{i,j} c_{ij} k'(d(\boldsymbol{v}_j, \boldsymbol{u}_{j|i})) \frac{\partial d(\boldsymbol{v}_j, \boldsymbol{u}_{j|i})}{\partial \boldsymbol{v}}$$

First, $\boldsymbol{v}_j^{\tau+1}$ can be updated while $c_{ij}^{\tau+1}$ is fixed:

$$\boldsymbol{v}_j^{\tau+1} = \frac{\sum_{i,j} c_{ij}^{\tau} k'(d(\boldsymbol{v}_j^{\tau}, \hat{\boldsymbol{u}}_{j|i})) \boldsymbol{u}_{j|i}}{\sum_{i,j} k'(d(\boldsymbol{v}_j^{\tau}, \boldsymbol{u}_{j|i}))}$$

Then, $c_{ij}^{\tau+1}$ can be updated using standard gradient descent:

$$c_{ij}^{\tau+1} = c_{ij}^{\tau} + \alpha \cdot k(d(\boldsymbol{v}_j^{\tau}, \boldsymbol{u}_{j|i}))$$

where α is the hyper-parameter to control step size.

To address the issue of convergence not being reached at instance level, we present an adaptive optimizer to self-adjust the number of iterations for individual examples according to their negative agreement scores (see Algorithm 1). Following Zhao et al. (2018), we replace standard softmax with leaky-softmax, which decreases the strength of noisy capsules.

Algorithm 1 Our Adaptive KDE Routing

procedure ROUTING($\boldsymbol{u}_{j|i}$, ℓ)
Initialize $\forall i, j : c_{ij} = 1/n_{\ell+1}$
while true **do**
  foreach capsule i, j in layer ℓ, $\ell + 1$ **do**
    $c_{ij} \leftarrow$ leaky-softmax(c_{ij})
  foreach capsule j in layer $\ell + 1$ **do**
    $\boldsymbol{v}_j \leftarrow \frac{\sum_i c_{ij} k'(d(\boldsymbol{v}_j, \boldsymbol{u}_{j|i})) \hat{\boldsymbol{u}}_{j|i}}{\sum_{i=1}^{n} k'(d(\boldsymbol{v}_i, \boldsymbol{u}_{j|i}))}$
  foreach capsule i, j in layer ℓ, $\ell + 1$ **do**
    $c_{ij} \leftarrow c_{ij} + \alpha \cdot k(d(\boldsymbol{v}_j, \boldsymbol{u}_{j|i}))$
  foreach capsule j in layer $\ell + 1$ **do**
    $a_j \leftarrow |v_j|$
  NAS $= \log(\sum_{\mathrm{i,j}} \mathrm{c_{ij}} \mathrm{k}(\mathrm{d}(\boldsymbol{v}_{\mathrm{j}}, \boldsymbol{u}_{\mathrm{j|i}})))$
  if $|\mathrm{NAS} - \mathrm{Last\_NAS}| < \epsilon$ **then**
    break
  else
    Last_NAS $\leftarrow$ NAS
return v_j, a_j

2.4 Representation Layer

This is the top-level layer containing final capsules calculated by iteratively minimizing the NAS function (See Eq. 1), where the number of final capsules corresponds to the entire output space. Therefore, as long as the size of an output space goes to a large scale (thousands of labels), the computation of this function would become extremely expensive, which yields the bottleneck of scalability of capsule networks.

Partial Routing As opposed to the entire output space on data sets, the sub-output space corresponding to individual examples is rather small, i.e., only few labels are assigned to one document in text classification, for example. As a consequence, it is redundant to route low-level capsules to the entire output space for each example in the training stage, which motivated us to present a partial routing algorithm with constrained output spaces, such that our NAS function is described as:

$$\min_{c,\boldsymbol{v}} -\sum_i \Big(\sum_{j \in D^+} c_{ij} \langle \boldsymbol{v}_j, \boldsymbol{u}_{j|i} \rangle + \lambda \cdot \sum_{k \in D^-} c_{ik} \langle \boldsymbol{v}_k, \boldsymbol{u}_{k|i} \rangle \Big)$$

where D^+ and D^- denote the sets of real (positive) and randomly selected (negative) outputs for each example, respectively. Both sets are

far smaller than the entire output space. λ is the hyper-parameter to control aggregation scores from negative outputs.

3 Experiments

The major focus of this work is to investigate the scalability of our approach on datasets with a large output space, and generalizability in low-resource settings with few training examples. Therefore, we validate our capsule-based approach on two specific NLP tasks: (i) multi-label text classification with a large label scale; (ii) question answering with a data imbalance issue.

3.1 Multi-label Text Classification

Multi-label text classification task refers to assigning multiple relevant labels to each input document, while the entire label set might be extremely large. We use our approach to encode an input document and generate the final capsules corresponding to the number of labels in the representation layer. The length of final capsule for each label indicates the probability whether the document has this label.

Dataset	#Train/Test/Labels	Avg-docs
RCV1	23.1K/781.2K/103	729.67
EUR-Lex	15.4K/3.8K/3.9K	15.59

Table 1: Characteristics of the datasets. Each label of RCV1 has about 729.67 training examples, while each label of EUR-Lex has merely about 15.59 examples.

Experimental Setup We conduct our experiments on two datasets selected from the extreme classification repository:[2] a regular label scale dataset (RCV1), with 103 labels (Lewis et al., 2004), and a large label scale dataset (EUR-Lex), with 3,956 labels (Mencia and Fürnkranz, 2008), described in Table 1. The intuition behind our datasets selection is that EUR-Lex, with 3,956 labels and 15.59 examples per label, fits well with our goal of investigating the scalability and generalizability of our approach. We contrast EUR-Lex with RCV1, a dataset with a regular label scale, and leave the study of datasets with extremely large labels, e.g., Wikipedia-500K with 501,069 labels, to future work.

Baselines We compare our approach to the following baselines: non-deep learning approaches using TF-IDF features of documents as inputs: FastXML (Prabhu and Varma, 2014), and PD-Sparse (Yen et al., 2016), deep learning approaches using raw text of documents as inputs: FastText (Joulin et al., 2016), Bow-CNN (Johnson and Zhang, 2014), CNN-Kim (Kim, 2014), XML-CNN (Liu et al., 2017)), and a capsule-based approach Cap-Zhao (Zhao et al., 2018). For evaluation, we use standard rank-based measures (Liu et al., 2017) such as Precision@k, and Normalized Discounted Cumulative Gain (NDCG@k).

Implementation Details The word embeddings are initialized as 300-dimensional GloVe vectors (Pennington et al., 2014). In the convolutional layer, we use a convolution operation with three different window sizes (2,4,8) to extract features from input documents. Each feature is transformed into a primary capsule with 16 dimensions by a group-convolution operation. All capsules in the primary capsule layer are condensed into 256 capsules for RCV1 and 128 capsules for EUR-Lex by a capsule compression operation.

To avoid routing low-level capsules to the entire label space in the inference stage, we use a CNN baseline (Kim, 2014) trained on the same dataset with our approach, to generate 200 candidate labels and take these labels as a constrained output space for each example.

Experimental Results In Table 2, we can see a noticeable margin brought by our capsule-based approach over the strong baselines on EUR-Lex, and competitive results on RCV1. These results appear to indicate that our approach has superior generalization ability on datasets with fewer training examples, i.e., RCV1 has 729.67 examples per label while EUR-Lex has 15.59 examples.

In contrast to the strongest baseline XML-CNN with 22.52M parameters and 0.08 seconds per batch, our approach has 14.06M parameters, and takes 0.25 seconds in an acceleration setting with capsule compression and partial routing, and 1.7 seconds without acceleration. This demonstrates that our approach provides competitive computational speed with fewer parameters compared to the competitors.

Discussion on Generalization To further study the generalization capability of our approach, we vary the percentage of training examples from 100% to 50% on the entire training set, leading to the number of training examples per label de-

[2] https://manikvarma.github.io

Datasets	Metrics	FastXML	PD-Sparse	FastText	Bow-CNN	CNN-Kim	XML-CNN	Cap-Zhao	NLP-Cap	Impv
RCV1	PREC@1	94.62	95.16	95.40	96.40	93.54	96.86	96.63	**97.05**	+0.20%
	PREC@3	78.40	79.46	79.96	81.17	76.15	81.11	81.02	**81.27**	+0.20%
	PREC@5	54.82	55.61	55.64	**56.74**	52.94	56.07	56.12	56.33	-0.72%
	NDCG@1	94.62	95.16	95.40	96.40	93.54	96.88	96.63	**97.05**	+0.20%
	NDCG@3	89.21	90.29	90.95	92.04	87.26	92.22	92.31	**92.47**	+0.17%
	NDCG@5	90.27	91.29	91.68	92.89	88.20	92.63	92.75	**93.11**	+0.52%
EUR-Lex	PREC@1	68.12	72.10	71.51	64.99	68.35	75.65	-	**80.20**	+6.01%
	PREC@3	57.93	57.74	60.37	51.68	54.45	61.81	-	**65.48**	+5.93%
	PREC@5	48.97	47.48	50.41	42.32	44.07	50.90	-	**52.83**	+3.79%
	NDCG@1	68.12	72.10	71.51	64.99	68.35	75.65	-	**80.20**	+6.01%
	NDCG@3	60.66	61.33	63.32	55.03	59.81	66.71	-	**71.11**	+6.59%
	NDCG@5	56.42	55.93	58.56	49.92	57.99	64.45	-	**68.80**	+6.75%

Table 2: Comparisons of our NLP-Cap approach and baselines on two text classication benchmarks, where '-' denotes methods that failed to scale due to memory issues.

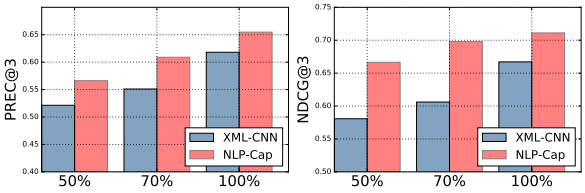


Figure 5: Performance on EUR-Lex by varying the percentage of training examples (X-axis).

Method	#Training	PREC@1	PREC@3	PREC@5
XML-CNN	100% examples	75.65	61.81	50.90
NLP-Capsule	50% examples	73.69	56.62	44.36
	60% examples	74.83	58.48	46.33
	70% examples	77.26	60.90	47.73
	80% examples	77.68	61.06	48.28
	90% examples	79.45	63.95	50.90
	100% examples	**80.20**	**65.48**	**52.83**
Method	#Training	NDCG@1	NDCG@3	NDCG@5
XML-CNN	100% examples	75.65	66.71	64.45
NLP-Capsule	50% examples	73.69	66.65	67.36
	60% examples	74.83	67.87	68.62
	70% examples	77.26	69.79	69.65
	80% examples	77.67	69.43	69.27
	90% examples	79.45	**71.64**	**71.06**
	100% examples	**80.21**	71.11	68.80

Table 3: Experimental results on different fractions of training examples from 50% to 100% on EUR-Lex.

creasing from 15.59 to 7.77. Figure 5 shows that our approach outperforms the strongest baseline XML-CNN with different fractions of the training examples.

This finding agrees with our speculation on generalization: the distance between our approach and XML-CNN increases as fewer training data samples are available. In Table 3, we also find that our approach with 70% of training examples achieves about 5% improvement over XML-CNN with 100% of examples on 4 out of 6 metrics.

Routing Comparison We compare our routing with (Sabour et al., 2017) and (Zhang et al., 2018b) on EUR-Lex dataset and observe that it performs best on all metrics (Table 4). We speculate that the improvement comes from enhanced reliability of routing processes at instance level.

3.2 Question Answering

Question-Answering (QA) selection task refers to selecting the best answer from candidates to each question. For a question-answer pair (q, a), we use our capsule-based approach to generate two final capsules $\boldsymbol{v}_q$ and $\boldsymbol{v}_a$ corresponding to the respective question and answer. The relevance score of question-answer pair can be defined as their cosine similarity:

$$s(q,a) = \cos(\boldsymbol{v}_{\mathrm{q}}, \boldsymbol{v}_{\mathrm{a}}) = \frac{\boldsymbol{v}_{\mathrm{q}}^{\intercal}\boldsymbol{v}_{\mathrm{a}}}{||\boldsymbol{v}_{\mathrm{q}}|| \cdot ||\boldsymbol{v}_{\mathrm{a}}||}$$

Experiment Setup In Table 5, we conduct our experiments on the TREC QA dataset collected from TREC QA track 8-13 data (Wang et al., 2007). The intuition behind this dataset selection is that the cost of hiring human annotators to collect positive answers for individual questions can be prohibitive since positive answers can be conveyed in multiple different surface forms. Such issue arises particularly in TREC QA with only 12%

Method	PREC@1	PREC@3	PREC@5
XML-CNN	75.65	61.81	50.90
NLP-Capsule + Sabour's Routing	79.14	64.33	51.85
NLP-Capsule + Zhang's Routing	80.20	65.48	52.83
NLP-Capsule + Our Routing	**80.62**	**65.61**	**53.66**
Method	**NDCG@1**	**NDCG@3**	**NDCG@5**
XML-CNN	75.65	66.71	64.45
NLP-Capsule + Sabour's Routing	79.14	70.13	67.02
NLP-Capsule + Zhang's Routing	80.20	71.11	68.80
NLP-Capsule + Our Routing	**80.62**	**71.34**	**69.57**

Table 4: Performance on EUR-Lex dataset with different routing process.

Dataset	#Questions	#QA Pairs	%Positive
Train/Dev/Test	1229/82/100	53417/1148/1517	12%

Table 5: Characteristic of TREC QA dataset. %Positive denotes the percentage of positive answers.

positive answers. Therefore, we use this dataset to investigate the generalizability of our approach.

Baselines We compare our approach to the following baselines: CNN + LR (Yu et al., 2014b) using unigrams and bigrams, CNN (Severyn and Moschitti, 2015) using additional bilinear similarity features, CNTN (Qiu and Huang, 2015) using neural tensor network, LSTM (Tay et al., 2017) using single and multi-layer, MV-LSTM (Wan et al., 2016), NTN-LSTM and HD-LSTM (Tay et al., 2017) using holographic dual LSTM and Capsule-Zhao (Zhao et al., 2018) using capsule networks. For evaluation, we use standard measures (Wang et al., 2007) such as Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR).

Implementation Details The word embeddings used for question answering pairs are initialized as 300-dimensional GloVe vectors. In the convolutional layer, we use a convolution operation with three different window sizes (3,4,5). All 16-dimensional capsules in the primary capsule layer are condensed into 256 capsules by the capsule compression operation.

Experimental Results and Discussions In Table 6, the best performance on MAP metric is achieved by our approach, which verifies the effectiveness of our model. We also observe that our approach exceeds traditional neural models like CNN, LSTM and NTN-LSTM by a noticeable margin.

This finding also agrees with the observation we found in multi-label classification: our approach has superior generalization capability in low-resource setting with few training examples. In contrast to the strongest baseline HD-LSTM with 34.51M and 0.03 seconds for one batch, our approach has 17.84M parameters and takes 0.06 seconds in an acceleration setting, and 0.12 seconds without acceleration.

Method	MAP	MRR
CNN + LR (unigram)	54.70	63.29
CNN + LR (bigram)	56.93	66.13
CNN	66.91	68.80
CNTN	65.80	69.78
LSTM (1 layer)	62.04	66.85
LSTM	59.75	65.33
MV-LSTM	64.88	68.24
NTN-LSTM	63.40	67.72
HD-LSTM	67.44	**75.11**
Capsule-Zhao	73.63	70.12
NLP-Capsule	**77.73**	74.16

Table 6: Experimental results on TREC QA dataset.

4 Related Work

4.1 Multi-label Text Classification

Multi-label text classification aims at assigning a document to a subset of labels whose label set might be extremely large. With increasing numbers of labels, issues of data sparsity and scalability arise. Several methods have been proposed for the large multi-label classification case.

Tree-based models (Agrawal et al., 2013; Weston et al., 2013) induce a tree structure that recursively partitions the feature space with non-leaf nodes. Then, the restricted label spaces at leaf nodes are used for classification. Such a solution entails higher robustness because of a dynamic hyper-plane design and its computational efficiency. FastXML (Prabhu and Varma, 2014) is one such tree-based model, which learns a hierarchy of training instances and optimizes an NDCG-based objective function for nodes in the tree structure.

Label embedding models (Balasubramanian and Lebanon, 2012; Chen and Lin, 2012; Cisse et al., 2013; Bi and Kwok, 2013; Ferng and Lin, 2011; Hsu et al., 2009; Ji et al., 2008; Kapoor et al., 2012; Lewis et al., 2004; Yu et al., 2014a) address the data sparsity issue with two steps: compression and decompression. The compression step learns a low-dimensional label embedding that is projected from original and high-dimensional label space. When data instances are classified to these label embeddings, they will be projected back to the high-dimensional label space, which is the decompression step. Recent works came up with different compression or decompression techniques, e.g., SLEEC (Bhatia et al., 2015).

Deep learning models: FastText (Joulin et al., 2016) uses averaged word embeddings to classify documents, which is computationally efficient but ignores word order. Various CNNs inspired by Kim (2014) explored MTC with dynamic pooling, such as Bow-CNN (Johnson and

Zhang, 2014) and XML-CNN (Liu et al., 2017).

Linear classifiers: PD-Sparse (Yen et al., 2016) introduces a Fully-Corrective Block-Coordinate Frank-Wolfe algorithm to address data sparsity.

4.2 Question and Answering

State-of-the-art approaches to QA fall into two categories: IR-based and knowledge-based QA.

IR-based QA firstly preprocesses the question and employ information retrieval techniques to retrieve a list of relevant passages to questions. Next, reading comprehension techniques are adopted to extract answers within the span of retrieved text. For answer extraction, early methods manually designed patterns to get them (Pasca). A recent popular trend is neural answer extraction. Various neural network models are employed to represent questions (Severyn and Moschitti, 2015; Wang and Nyberg, 2015). Since the attention mechanism naturally explores relevancy, it has been widely used in QA models to relate the question to candidate answers (Tan et al., 2016; Santos et al., 2016; Sha et al., 2018). Moreover, some researchers leveraged external large-scale knowledge bases to assist answer selection (Savenkov and Agichtein, 2017; Shen et al., 2018; Deng et al., 2018).

Knowledge-based QA conducts semantic parsing on questions and transforms parsing results into logical forms. Those forms are adopted to match answers from structured knowledge bases (Yao and Van Durme, 2014; Yih et al., 2015; Bordes et al., 2015; Yin et al., 2016; Hao et al., 2017). Recent developments focused on modeling the interaction between question and answer pairs: Tensor layers (Qiu and Huang, 2015; Wan et al., 2016) and holographic composition (Tay et al., 2017) have pushed the state-of-the-art.

4.3 Capsule Networks

Capsule networks were initially proposed by Hinton (Hinton et al., 2011) to improve representations learned by neural networks against vanilla CNNs. Subsequently, Sabour et al. (2017) replaced the scalar-output feature detectors of CNNs with vector-output capsules and max-pooling with routing-by-agreement.

Hinton et al. (2018) then proposed a new iterative routing procedure between capsule layers based on the EM algorithm, which achieves better accuracy on the smallNORB dataset. Zhang et al. (2018a) applied capsule networks to relation extraction in a multi-instance multi-label learning framework. Xiao et al. (2018) explored capsule networks for multi-task learning.

Xia et al. (2018) studied the zero-shot intent detection problem with capsule networks, which aims to detect emerging user intents in an unsupervised manner. Zhao et al. (2018) investigated capsule networks with dynamic routing for text classification, and transferred knowledge from the single-label to multi-label cases. Cho et al. (2019) studied capsule networks with determinantal point processes for extractive multi-document summarization.

Our work is different from our predecessors in the following aspects: (i) we evaluate the performance of routing processes at instance level, and introduce an adaptive optimizer to enhance the reliability of routing processes; (ii) we present capsule compression and partial routing to achieve better scalability of capsule networks on datasets with a large output space.

5 Conclusion

Making computers perform more like humans is a major issue in NLP and machine learning. This not only includes making them perform on similar levels (Hassan et al., 2018), but also requests them to be robust to adversarial examples (Eger et al., 2019) and generalize from few data points (Rücklé et al., 2019). In this work, we have addressed the latter issue.

In particular, we extended existing capsule networks into a new framework with advantages concerning scalability, reliability and generalizability. Our experimental results have demonstrated its effectiveness on two NLP tasks: multi-label text classification and question answering.

Through our modifications and enhancements, we hope to have made capsule networks more suitable to large-scale problems and, hence, more mature for real-world applications. In the future, we plan to apply capsule networks to even more challenging NLP problems such as language modeling and text generation.

6 Acknowledgments

We thank the anonymous reviewers for their comments, which greatly improved the final version of the paper. This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of In-

formation from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1.

References

Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. 2013. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the 22nd international conference on World Wide Web*, pages 13–24. ACM.

Krishnakumar Balasubramanian and Guy Lebanon. 2012. The landmark selection method for multiple output prediction. *arXiv preprint arXiv:1206.6479*.

Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse local embeddings for extreme multi-label classification. In *Advances in Neural Information Processing Systems*, pages 730–738.

Wei Bi and James Kwok. 2013. Efficient multi-label classification with many labels. In *International Conference on Machine Learning*, pages 405–413.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Yao-Nan Chen and Hsuan-Tien Lin. 2012. Feature-aware label space dimension reduction for multi-label classification. In *Advances in Neural Information Processing Systems*, pages 1529–1537.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Moustapha M Cisse, Nicolas Usunier, Thierry Artieres, and Patrick Gallinari. 2013. Robust bloom filters for large multilabel classification tasks. In *Advances in Neural Information Processing Systems*, pages 1851–1859.

Dorin Comaniciu and Peter Meer. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.

Yang Deng, Ying Shen, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. 2018. Knowledge as a bridge: Improving cross-domain answer selection with external knowledge. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3295–3305.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding nlp systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

C-S Ferng and H-T Lin. 2011. Multi-label classification with error-correcting codes. In *Asian Conference on Machine Learning*, pages 281–295.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 221–231.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. 2011. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer.

Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. 2018. Matrix capsules with em routing.

Daniel J Hsu, Sham M Kakade, John Langford, and Tong Zhang. 2009. Multi-label prediction via compressed sensing. In *Advances in neural information processing systems*, pages 772–780.

Shuiwang Ji, Lei Tang, Shipeng Yu, and Jieping Ye. 2008. Extracting shared subspace for multi-label classification. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 381–389. ACM.

Rie Johnson and Tong Zhang. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Ashish Kapoor, Raajay Viswanathan, and Prateek Jain. 2012. Multilabel classification using bayesian compressed sensing. In *Advances in Neural Information Processing Systems*, pages 2645–2653.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124. ACM.

Eneldo Loza Mencia and Johannes Fürnkranz. 2008. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 50–65. Springer.

Marius Pasca. *Open-Domain Question Answering from Large Text Collections*, volume 29.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Yashoteja Prabhu and Manik Varma. 2014. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 263–272. ACM.

Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Andreas Rücklé, Nafise Sadat Moosavi, and Iryna Gurevych. 2019. Coala: A neural coverage-based approach for long answer selection with small data. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866.

Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.

Denis Savenkov and Eugene Agichtein. 2017. Evinets: Neural networks for combining evidence signals for factoid question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 299–304.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 373–382. ACM.

Lei Sha, Xiaodong Zhang, Feng Qian, Baobao Chang, and Zhifang Sui. 2018. A multi-view fusion neural network for answer selection. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ying Shen, Yang Deng, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. 2018. Knowledge-aware attentive neural network for ranking question answer pairs. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 901–904. ACM.

Ming Tan, Cicero Dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 464–473.

Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. 2017. Learning to rank question answer pairs with holographic dual lstm architecture. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 695–704. ACM.

Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. 2016. A deep architecture for semantic matching with multiple positional sentence representations. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Matt P Wand and M Chris Jones. 1994. *Kernel smoothing*. Chapman and Hall/CRC.

Di Wang and Eric Nyberg. 2015. A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 707–712.

Mengqiu Wang, Noah A Smith, and Teruko Mitamura. 2007. What is the jeopardy model? a quasi-synchronous grammar for qa. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Mingxuan Wang, Jun Xie, Zhixing Tan, Jinsong Su, et al. 2018. Towards linear time neural machine translation with capsule networks. *arXiv preprint arXiv:1811.00287*.

Jason Weston, Ameesh Makadia, and Hector Yee. 2013. Label partitioning for sublinear ranking. In *International Conference on Machine Learning*, pages 181–189.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S Yu. 2018. Zero-shot user intent detection via capsule neural networks. *arXiv preprint arXiv:1809.00385*.

Liqiang Xiao, Honglun Zhang, Wenqing Chen, Yongkun Wang, and Yaohui Jin. 2018. Mcapsnet: Capsule network for text with multi-task learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4565–4574.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 956–966.

Ian En-Hsu Yen, Xiangru Huang, Pradeep Ravikumar, Kai Zhong, and Inderjit Dhillon. 2016. Pd-sparse: A primal and dual sparse approach to extreme multiclass and multilabel classification. In *International Conference on Machine Learning*, pages 3069–3077.

Scott Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base.

Wenpeng Yin, Mo Yu, Bing Xiang, Bowen Zhou, and Hinrich Schütze. 2016. Simple question answering by attentive convolutional neural network. *arXiv preprint arXiv:1606.03391*.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. 2014a. Large-scale multi-label learning with missing labels. In *International conference on machine learning*, pages 593–601.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014b. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*.

Ningyu Zhang, Shumin Deng, Zhanling Sun, Xi Chen, Wei Zhang, and Huajun Chen. 2018a. Attention-based capsule network with dynamic routing for relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 986–992.

Suofei Zhang, Wei Zhao, Xiaofu Wu, and Quan Zhou. 2018b. Fast dynamic routing based on weighted kernel density estimation. *arXiv preprint arXiv:1805.10807*.

Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Suofei Zhang, and Zhou Zhao. 2018. Investigating capsule networks with dynamic routing for text classification. In *Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP)*, pages 3110–3119.

Subpart B

Reference-free Evaluation

Chapter 8

On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation

On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation

Wei Zhao[†], **Goran Glavaš**[‡], **Maxime Peyrard**[Φ], **Yang Gao**[⋆], **Robert West**[Φ], **Steffen Eger**[†]
[†] Technische Universität Darmstadt [‡] University of Mannheim, Germany
[Φ] EPFL, Switerland [⋆] Royal Holloway University of London, UK
{zhao,eger}@aiphes.tu-darmstadt.de
goran@informatik.uni-mannheim.de, yang.gao@rhul.ac.uk
{maxime.peyrard,robert.west}@epfl.ch

Abstract

Evaluation of cross-lingual encoders is usually performed either via zero-shot cross-lingual transfer in supervised downstream tasks or via unsupervised cross-lingual textual similarity. In this paper, we concern ourselves with reference-free machine translation (MT) evaluation where we directly compare source texts to (sometimes low-quality) system translations, which represents a natural adversarial setup for multilingual encoders. Reference-free evaluation holds the promise of web-scale comparison of MT systems. We systematically investigate a range of metrics based on state-of-the-art cross-lingual semantic representations obtained with pretrained M-BERT and LASER. We find that they perform poorly as semantic encoders for reference-free MT evaluation and identify their two key limitations, namely, (a) a semantic mismatch between representations of mutual translations and, more prominently, (b) the inability to punish "translationese", i.e., low-quality literal translations. We propose two partial remedies: (1) post-hoc re-alignment of the vector spaces and (2) coupling of semantic-similarity based metrics with target-side language modeling. In segment-level MT evaluation, our best metric surpasses reference-based BLEU by 5.7 correlation points. We make our MT evaluation code available.[1]

1 Introduction

A standard evaluation setup for supervised machine learning (ML) tasks assumes an evaluation metric which compares a gold label to a classifier prediction. This setup assumes that the task has clearly defined and unambiguous labels and, in most cases, that an instance can be assigned few labels. These assumptions, however, do not hold for natural language generation (NLG) tasks like machine translation (MT) (Bahdanau et al., 2015; Johnson et al., 2017) and text summarization (Rush et al., 2015; Tan et al., 2017), where we do not predict a single discrete label but generate natural language text. Thus, the set of labels for NLG is neither clearly defined nor finite. Yet, the standard evaluation protocols for NLG still predominantly follow the described default paradigm: (1) evaluation datasets come with human-created reference texts and (2) evaluation metrics, e.g., BLEU (Papineni et al., 2002) or METEOR (Lavie and Agarwal, 2007) for MT and ROUGE (Lin and Hovy, 2003) for summarization, count the exact "label" (i.e., n-gram) matches between reference and system-generated text. In other words, established NLG evaluation compares semantically ambiguous labels from an unbounded set (i.e., natural language texts) via hard symbolic matching (i.e., string overlap).

The first remedy is to replace the hard symbolic comparison of natural language "labels" with a soft comparison of texts' meaning, using semantic vector space representations. Recently, a number of MT evaluation methods appeared focusing on semantic comparison of reference and system translations (Shimanaka et al., 2018; Clark et al., 2019; Zhao et al., 2019). While these correlate better than n-gram overlap metrics with human assessments, they do not address inherent limitations stemming from the need for reference translations, namely: (1) references are expensive to obtain; (2) they assume a single correct solution and bias the evaluation, both automatic and human (Dreyer and Marcu, 2012; Fomicheva and Specia, 2016), and (3) limitation of MT evaluation to language pairs with available parallel data.

Reliable *reference-free* evaluation metrics, directly measuring the (semantic) correspondence between the source language text and system translation, would remove the need for human references and allow for unlimited MT evaluations: *any*

[1] https://github.com/AIPHES/ACL20-Reference-Free-MT-Evaluation

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1656–1671
July 5 - 10, 2020.

monolingual corpus could be used for evaluating MT systems. However, the proposals of reference-free MT evaluation metrics have been few and far apart and have required either non-negligible supervision (i.e., human translation quality labels) (Specia et al., 2010) or language-specific preprocessing like semantic parsing (Lo et al., 2014; Lo, 2019), both hindering the wide applicability of the proposed metrics. Moreover, they have also typically exhibited performance levels well below those of standard reference-based metrics (Ma et al., 2019).

In this work, we comparatively evaluate a number of reference-free MT evaluation metrics that build on the most recent developments in multilingual representation learning, namely cross-lingual contextualized embeddings (Devlin et al., 2019) and cross-lingual sentence encoders (Artetxe and Schwenk, 2019). We investigate two types of cross-lingual reference-free metrics: **(1)** *Soft token-level alignment* metrics find the optimal soft alignment between source sentence and system translation using Word Mover's Distance (WMD) (Kusner et al., 2015). Zhao et al. (2019) recently demonstrated that WMD operating on BERT representations (Devlin et al., 2019) substantially outperforms baseline MT evaluation metrics in the reference-based setting. In this work, we investigate whether WMD can yield comparable success in the reference-free (i.e., cross-lingual) setup; **(2)** *Sentence-level similarity* metrics measure the similarity between sentence representations of the source sentence and system translation using cosine similarity.

Our analysis yields several interesting findings. **(i)** We show that, unlike in the monolingual reference-based setup, metrics that operate on contextualized representations generally do not outperform symbolic matching metrics like BLEU, which operate in the reference-based environment. **(ii)** We identify two reasons for this failure: (a) firstly, cross-lingual semantic mismatch, especially for multi-lingual BERT (M-BERT), which construes a shared multilingual space in an unsupervised fashion, without any direct bilingual signal; (b) secondly, the inability of the state-of-the-art cross-lingual metrics based on multilingual encoders to adequately capture and punish "translationese", i.e., literal word-by-word translations of the source sentence—as translationese is an especially persistent property of MT systems, this problem is particularly troubling in our context of reference-free MT evaluation. **(iii)** We show that by executing an additional weakly-supervised cross-lingual re-mapping step, we can to some extent alleviate both previous issues. **(iv)** Finally, we show that the combination of cross-lingual reference-free metrics and language modeling on the target side (which is able to detect "translationese"), surpasses the performance of reference-based baselines.

Beyond designating a viable prospect of web-scale domain-agnostic MT evaluation, our findings indicate that the challenging task of reference-free MT evaluation is able to expose an important limitation of current state-of-the-art multilingual encoders, i.e., the failure to properly represent corrupt input, that may go unnoticed in simpler evaluation setups such as zero-shot cross-lingual text classification or measuring cross-lingual text similarity not involving "adversarial" conditions. We believe this is a promising direction for nuanced, fine-grained evaluation of cross-lingual representations, extending the recent benchmarks which focus on zero-shot transfer scenarios (Hu et al., 2020).

2 Related Work

Manual human evaluations of MT systems undoubtedly yield the most reliable results, but are expensive, tedious, and generally do not scale to a multitude of domains. A significant body of research is thus dedicated to the study of automatic evaluation metrics for machine translation. Here, we provide an overview of both reference-based MT evaluation metrics and recent research efforts towards reference-free MT evaluation, which leverage cross-lingual semantic representations and unsupervised MT techniques.

Reference-based MT evaluation. Most of the commonly used evaluation metrics in MT compare system and reference translations. They are often based on surface forms such as n-gram overlaps like BLEU (Papineni et al., 2002), SentBLEU, NIST (Doddington, 2002), chrF++ (Popović, 2017) or METEOR++(Guo and Hu, 2019). They have been extensively tested and compared in recent WMT metrics shared tasks (Bojar et al., 2017a; Ma et al., 2018a, 2019).

These metrics, however, operate at the surface level, and by design fail to recognize semantic equivalence lacking lexical overlap. To overcome these limitations, some research efforts exploited static word embeddings (Mikolov et al., 2013b) and trained embedding-based supervised metrics on sufficiently large datasets with available human judgments of translation quality (Shimanaka

et al., 2018). With the development of contextual word embeddings (Peters et al., 2018; Devlin et al., 2019), we have witnessed proposals of semantic metrics that account for word order. For example, Clark et al. (2019) introduce a semantic metric relying on sentence mover's similarity and the contextualized ELMo embeddings (Peters et al., 2018). Similarly, Zhang et al. (2019) describe a reference-based semantic similarity metric based on contextualized BERT representations (Devlin et al., 2019). Zhao et al. (2019) generalize this line of work with their MoverScore metric, which computes the mover's distance, i.e., the optimal soft alignment between tokens of the two sentences, based on the similarities between their contextualized embeddings. Mathur et al. (2019) train a supervised BERT-based regressor for reference-based MT evaluation.

Reference-free MT evaluation. Recently, there has been a growing interest in reference-free MT evaluation (Ma et al., 2019), also referred to as "quality estimation" (QE) in the MT community. In this setup, evaluation metrics semantically compare system translations directly to the source sentences. The attractiveness of automatic reference-free MT evaluation is obvious: it does not require any human effort or parallel data. To approach this task, Popović et al. (2011) exploit a bag-of-word translation model to estimate translation quality, which sums over the likelihoods of aligned word-pairs between source and translation texts. Specia et al. (2013) estimate translation quality using language-agnostic linguistic features extracted from source lanuage texts and system translations. Lo et al. (2014) introduce XMEANT as a cross-lingual reference-free variant of MEANT, a metric based on semantic frames. Lo (2019) extended this idea by leveraging M-BERT embeddings. The resulting metric, YiSi-2, evaluates system translations by summing similarity scores over words pairs that are best-aligned mutual translations. YiSi-2-SRL optionally combines an additional similarity score based on the alignment over the semantic structures (e.g., semantic roles and frames). Both metrics are reference-free, but YiSi-2-SRL is not resource-lean as it requires a semantic parser for both languages. Moreover, in contrast to our proposed metrics, they do not mitigate the misalignment of cross-lingual embedding spaces and do not integrate a target-side language model, which we identify to be crucial components.

Recent progress in cross-lingual semantic similarity (Agirre et al., 2016; Cer et al., 2017) and unsupervised MT (Artetxe and Schwenk, 2019) has also led to novel reference-free metrics. For instance, Yankovskaya et al. (2019) propose to train a metric combining multilingual embeddings extracted from M-BERT and LASER (Artetxe and Schwenk, 2019) together with the log-probability scores from neural machine translation. Our work differs from that of Yankovskaya et al. (2019) in one crucial aspect: the cross-lingual reference-free metrics that we investigate and benchmark do not require any human supervision.

Cross-lingual Representations. Cross-lingual text representations offer a prospect of modeling meaning across languages and support cross-lingual transfer for downstream tasks (Klementiev et al., 2012; Rücklé et al., 2018; Glavaš et al., 2019; Josifoski et al., 2019; Conneau et al., 2020). Most recently, the (massively) multilingual encoders, such as multilingual M-BERT (Devlin et al., 2019), XLM-on-RoBERTa (Conneau et al., 2020), and (sentence-based) LASER, have profiled themselves as state-of-the-art solutions for (massively) multilingual semantic encoding of text. While LASER has been jointly trained on parallel data of 93 languages, M-BERT has been trained on the concatenation of monolingual data in more than 100 languages, without any cross-lingual mapping signal. There has been a recent vivid discussion on the cross-lingual abilities of M-BERT (Pires et al., 2019; K et al., 2020; Cao et al., 2020). In particular, Cao et al. (2020) show that M-BERT often yields disparate vector space representations for mutual translations and propose a multilingual re-mapping based on parallel corpora, to remedy for this issue. In this work, we introduce re-mapping solutions that are resource-leaner and require easy-to-obtain limited-size word translation dictionaries rather than large parallel corpora.

3 Reference-Free MT Evaluation Metrics

In the following, we use $\mathbf{x}$ to denote a source sentence (i.e., a sequence of tokens in the source language), $\mathbf{y}$ to denote a system translation of $\mathbf{x}$ in the target language, and $\mathbf{y}^\star$ to denote the human reference translation for $\mathbf{x}$.

3.1 Soft Token-Level Alignment

We start from the MoverScore (Zhao et al., 2019), a recently proposed reference-based MT evaluation

metric designed to measure the semantic similarity between system outputs ($\mathbf{y}$) and human references ($\mathbf{y}^\star$). It finds an optimal soft semantic alignments between tokens from $\mathbf{y}$ and $\mathbf{y}^\star$ by minimizing the Word Mover's Distance (Kusner et al., 2015). In this work, we extend the MoverScore metric to operate in the cross-lingual setup, i.e., to measure the semantic similarity between n-grams (unigram or bigrams) of the source text $\mathbf{x}$ and the system translation $\mathbf{y}$, represented with embeddings originating from a cross-lingual semantic space.

First, we decompose the source text $\mathbf{x}$ into a sequence of n-grams, denoted by $\boldsymbol{x}_n = (\mathrm{x}_1^n, \ldots, \mathrm{x}_m^n)$ and then do the same operation for the system translation $\mathbf{y}$, denoting the resulting sequence of n-grams with $\boldsymbol{y}_n$. Given $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$, we can then define a distance matrix $\boldsymbol{C}$ such that $\boldsymbol{C}_{ij} = \|E(\mathrm{x}_i^n) - E(\mathrm{y}_j^n)\|_2$ is the distance between the i-th n-gram of $\boldsymbol{x}$ and the j-th n-gram of $\boldsymbol{y}$, where E is a *cross-lingual* embedding function that maps text in different languages to a shared embedding space. With respect to the function E, we experimented with cross-lingual representations induced (a) from static word embeddings with RCSLS (Joulin et al., 2018)) (b) with M-BERT (Devlin et al., 2019) as the multilingual encoder; with a focus on the latter. For M-BERT, we take the representations of the last transformer layer as the text representations.

WMD between the two sequences of n-grams $\boldsymbol{x}^n$ and $\boldsymbol{y}^n$ with associated n-gram weights [2] to $\boldsymbol{f}_{\boldsymbol{x}^n} \in \mathbb{R}^{|\boldsymbol{x}^n|}$ and $\boldsymbol{f}_{\boldsymbol{y}^n} \in \mathbb{R}^{|\boldsymbol{y}^n|}$ is defined as:

$$m(\boldsymbol{x}, \boldsymbol{y}) := \mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) = \min_{\boldsymbol{F}} \sum_{ij} \boldsymbol{C}_{ij} \cdot \boldsymbol{F}_{ij},$$

$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^\intercal \mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n},$$

where $\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}$ is a transportation matrix with $\boldsymbol{F}_{ij}$ denoting the amount of flow traveling from x_i^n to y_j^n.

3.2 Sentence-Level Semantic Similarity

In addition to measuring semantic distance between $\mathbf{x}$ and $\mathbf{y}$ at word-level, one can also encode them into sentence representations with multilingual sentence encoders like LASER (Artetxe and Schwenk, 2019), and then measure their cosine distance

$$m(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{E(\boldsymbol{x})^\intercal E(\boldsymbol{y})}{\|E(\boldsymbol{x})\| \cdot \|E(\boldsymbol{y})\|}.$$

[2] We follow Zhao et al. (2019) in obtaining n-gram embeddings and their associated weights based on IDF.

3.3 Improving Cross-Lingual Alignments

Initial analysis indicated that, despite the multilingual pretraining of M-BERT (Devlin et al., 2019) and LASER (Artetxe and Schwenk, 2019), the monolingual subspaces of the multilingual spaces they induce are far from being semantically well-aligned, i.e., we obtain fairly distant vectors for mutual word or sentence translations.[3] To this end, we apply two simple, weakly-supervised linear projection methods for post-hoc improvement of the cross-lingual alignments in these multilingual representation spaces.

Notation. Let $\boldsymbol{D} = \{(w_\ell^1, w_k^1), \ldots, (w_\ell^n, w_k^n)\}$ be a set of matched word or sentence pairs from two different languages ℓ and k. We define a re-mapping function f such that any $f(E(w_\ell))$ and $E(w_k)$ are better aligned in the resulting shared vector space. We investigate two resource-lean choices for the re-mapping function f.

Linear Cross-lingual Projection (CLP). Following related work (Schuster et al., 2019), we re-map contextualized embedding spaces using linear projection. Given ℓ and k, we stack all vectors of the source language words and target language words for pairs $\boldsymbol{D}$, respectively, to form matrices $\boldsymbol{X}_\ell$ and $\boldsymbol{X}_k \in \mathbb{R}^{n \times d}$, with d as the embedding dimension and n as the number of word or sentence alignments. The word pairs we use to calibrate M-BERT are extracted from EuroParl (Koehn, 2005) using FastAlign (Dyer et al., 2013), and the sentence pairs to calibrate LASER are sampled directly from EuroParl.[4] Mikolov et al. (2013a) propose to learn a projection matrix $W \in \mathbb{R}^{d \times d}$ by minimizing the Euclidean distance beetween the projected source language vectors and their corresponding target language vectors:

$$\min_{\boldsymbol{W}} \|\boldsymbol{W}\boldsymbol{X}_\ell - \boldsymbol{X}_k\|_2.$$

Xing et al. (2015) achieve further improvement on the task of bilingual lexicon induction (BLI) by constraining $\boldsymbol{W}$ to an orthogonal matrix, i.e., such that $\boldsymbol{W}^\intercal\boldsymbol{W} = \mathbf{I}$. This turns the optimization into the well-known Procrustes problem (Schönemann, 1966) with the following closed-form solution:

$$\hat{\boldsymbol{W}} = \boldsymbol{U}\boldsymbol{V}^\intercal, \boldsymbol{U}\Sigma\boldsymbol{V}^\intercal = \mathrm{SVD}(\boldsymbol{X}_\ell\boldsymbol{X}_k^\intercal)$$

[3] LASER is jointly trained on parallel corpora of different languages, but in resource-lean language pairs, the induced embeddings from mutual translations may be far apart.

[4] While LASER requires large parallel corpora in pretraining, we believe that fine-tuning/calibrating the embeddings post-hoc requires fewer data points.

We note that the above CLP re-mapping is known to have deficits, i.e., it requires the embedding spaces of the involved languages to be approximately isomorphic (Søgaard et al., 2018; Vulić et al., 2019). Recently, some re-mapping methods that reportedly remedy for this issue have been suggested (Glavaš and Vulić, 2020; Mohiuddin and Joty, 2020). We leave the investigation of these novel techniques for our future work.

Universal Language Mismatch-Direction (UMD) Our second post-hoc linear alignment method is inspired by the recent work on removing biases in distributional word vectors (Dev and Phillips, 2019; Lauscher et al., 2019). We adopt the same approaches in order to quantify and remedy for the "language bias", i.e., representation mismatches between mutual translations in the initial multilingual space. Formally, given ℓ and k, we create individual misalignment vectors $E(w_\ell^i) - E(w_k^i)$ for each bilingual pair in $\boldsymbol{D}$. Then we stack these individual vectors to form a matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times d}$. We then obtain the global misalignment vector $\boldsymbol{v}_B$ as the top left singular vector of $\boldsymbol{Q}$. The global misalignment vector presumably captures the direction of the representational misalignment between the languages better than the individual (noisy) misalignment vectors $E(w_\ell^i) - E(w_k^i)$. Finally, we modify all vectors $E(w_\ell)$ and $E(w_k)$, by subtracting their projections onto the global misalignment direction vector v_B:

$$f(E(w_\ell)) = E(w_\ell) - \cos(E(w_\ell), v_B)v_B.$$

Language Model BLEU scores often fail to reflect the fluency level of translated texts (Edunov et al., 2019). Hence, we use the language model (LM) of the target language to regularize the cross-lingual semantic similarity metrics, by coupling our cross-lingual similarity scores with a GPT language model of the target language (Radford et al., 2018). We expect the language model to penalize translationese, i.e., unnatural word-by-word translations and boost the performance of our metrics.[5]

4 Experiments

In this section, we evaluate the quality of our MT reference-free metrics by correlating them with human judgments of translation quality. These quality judgments are based on comparing human references and system predictions. We will discuss this discrepancy in §5.3.

Word-level metrics. We denote our word-level alignment metrics based on WMD as MoverScore-ngram + Align(Embedding), where Align is one of our two post-hoc cross-lingual alignment methods (CLP or UMD). For example, Mover-2 + UMD(M-BERT) denotes the metric combining MoverScore based on bigram alignments, with M-BERT embeddings and UMD as the post-hoc alignment method.

Sentence-level metric. We denote our sentence-level metrics as: Cosine + Align(Embedding). For example, Cosine + CLP(LASER) measures the cosine distance between the sentence embeddings obtained with LASER, post-hoc aligned with CLP.

4.1 Datasets

We collect the source language sentences, their system and reference translations from the WMT17-19 news translation shared task (Bojar et al., 2017b; Ma et al., 2018b, 2019), which contains predictions of 166 translation systems across 16 language pairs in WMT17, 149 translation systems across 14 language pairs in WMT18 and 233 translation systems across 18 language pairs in WMT19. We evaluate for X-en language pairs, selecting X from a set of 12 diverse languages: German (de), Chinese (zh), Czech (cs), Latvian (lv), Finnish (fi), Russian (ru), and Turkish (tr), Gujarati (gu), Kazakh (kk), Lithuanian (lt) and Estonian (et). Each language pair in WMT17-19 has approximately 3,000 source sentences, each associated to one reference translation and to the automatic translations generated by participating systems.

4.2 Baselines

We compare with a range of reference-free metrics: ibm1-morpheme and ibm1-pos4gram (Popović, 2012), LASIM (Yankovskaya et al., 2019), LP (Yankovskaya et al., 2019), YiSi-2 and YiSi-2-srl (Lo, 2019), and reference-based baselines BLEU (Papineni et al., 2002), SentBLEU (Koehn et al., 2007) and ChrF++ (Popović, 2017) for MT evaluation (see §2).[6] The main results are reported on WMT17. We report the results obtained on WMT18 and WMT19 in the Appendix.

[5] We linearly combine the cross-lingual metrics with the LM scores using a coefficient of 0.1 for all setups. We choose this value based on initial experiments on one language pair.

[6] The code of these unsupervised metrics is not released, thus we compare to their official results on WMT19 only.

Setting	Metrics	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
$m(\mathbf{y}^*, \mathbf{y})$	SENTBLEU	43.5	43.2	57.1	39.3	48.4	53.8	51.2	48.1
	CHRF++	**52.3**	**53.4**	**67.8**	**52.0**	**58.8**	**61.4**	**59.3**	**57.9**
$m(\mathbf{x}, \mathbf{y})$	*Baseline with Original Embeddings*								
	MOVER-1 + M-BERT	22.7	37.1	34.8	26.0	26.7	**42.5**	**48.2**	34.0
	COSINE + LASER	**32.6**	**40.2**	**41.4**	**48.3**	**36.3**	42.3	46.7	**41.1**
	Cross-lingual Alignment for Sentence Embedding								
	COSINE + CLP(LASER)	33.4	**40.5**	42.0	**48.6**	**36.0**	44.7	42.2	**41.1**
	COSINE + UMD(LASER)	**36.6**	28.1	**45.5**	48.5	31.3	**46.2**	**49.4**	40.8
	Cross-lingual Alignment for Word Embedding								
	MOVER-1 + RCSLS	18.9	26.4	31.9	33.1	25.7	31.1	34.3	28.8
	MOVER-1 + CLP(M-BERT)	33.4	38.6	50.8	48.0	33.9	**51.6**	53.2	44.2
	MOVER-2 + CLP(M-BERT)	**33.7**	38.8	**52.2**	**50.3**	**35.4**	51.0	**53.3**	**45.0**
	MOVER-1 + UMD(M-BERT)	22.3	38.1	34.5	30.5	31.2	43.5	48.6	35.5
	MOVER-2 + UMD(M-BERT)	23.1	38.9	37.1	34.7	33.0	44.8	48.9	37.2
	Combining Language Model								
	COSINE + CLP(LASER) ⊕ LM	48.8	46.7	63.2	66.2	51.0	54.6	48.6	54.2
	COSINE + UMD(LASER) ⊕ LM	**49.4**	46.2	**64.7**	66.4	**51.1**	**56.0**	52.8	**55.2**
	MOVER-2 + CLP(M-BERT) ⊕ LM	46.5	46.4	63.3	63.8	47.6	55.5	**53.5**	53.8
	MOVER-2 + UMD(M-BERT) ⊕ LM	41.8	**46.8**	60.4	59.8	46.1	53.8	52.4	51.6

Table 1: Pearson correlations with segment-level human judgments on the WMT17 dataset.

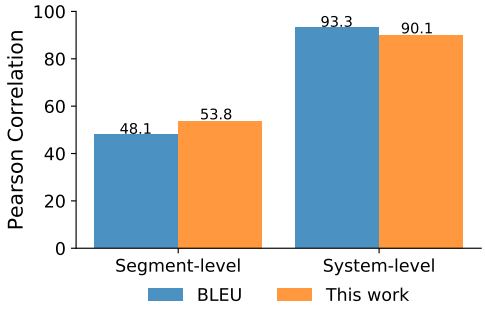


Figure 1: Average results of our best-performing metric, together with reference-based BLEU on WMT17.

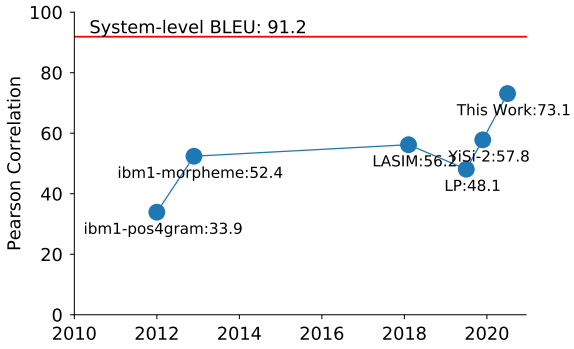


Figure 2: Average results of our metric best-performing metric, together with the official results of reference-free metrics, and reference-based BLEU on system-level WMT19.

4.3 Results

Figure 1 shows that our metric MOVER-2 + CLP(M-BERT) ⊕ LM, operating on modified M-BERT with the post-hoc re-mapping and combining a target-side LM, outperforms BLEU by 5.7 points in segment-level evaluation and achieves comparable performance in the system-level evaluation. Figure 2 shows that the same metric obtains 15.3 points gains (73.1 vs. 57.8), averaged over 7 languages, on WMT19 (system-level) compared to the the state-of-the-art reference-free metric YiSi-2. Except for one language pair, gu-en, our metric performs on a par with the reference-based BLEU (see Table 8 in the Appendix) on system-level.

In Table 1, we exhaustively compare results for several of our metric variants, based either on M-BERT or LASER. We note that re-mapping has considerable effect for M-BERT (up to 10 points improvements), but much less so for LASER. We believe that this is because the underlying embedding space of LASER is less 'misaligned' since it has been (pre-)trained on parallel data.[7] While the re-mapping is thus effective for metrics based on M-BERT, we still require the target-side LM to outperform BLEU. We assume the LM can address challenges that the re-mapping apparently is not able to handle properly; see our discussion in §5.1.

Overall, we remark that none of our metric com-

[7]However, in the appendix, we find that re-mapping LASER using 2k parallel sentences achieves considerable improvements on low-resource languages, e.g., kk-en (from -61.1 to 49.8) and lt-en (from 68.3 to 75.9); see Table 8.

binations performs consistently best. The reason may be that LASER and M-BERT are pretrained over hundreds of languages with substantial differences in corpora sizes in addition to the different effects of the re-mapping. However, we observe that MOVER-2 + CLP(M-BERT) performs best on average over all language pairs when the LM is not added. When the LM is added, MOVER-2 + CLP(M-BERT) $\oplus$ LM and COSINE + UMD (LASER) $\oplus$ LM perform comparably. This indicates that there may be a saturation effect when it comes to the LM or that the LM coefficients should be tuned individually for each semantic similarity metric based on cross-lingual representations.

5 Analysis

We first analyze preferences of our metrics based on M-BERT and LASER (§5.1) and then examine how much parallel data we need for re-mapping our vector spaces (§5.2). Finally, we discuss whether it is legitimate to correlate our metric scores, which evaluate the similarity of system predictions and source texts, to human judgments based on system predictions and references (§5.3).

5.1 Metric preferences

To analyze why our metrics based on M-BERT and LASER perform so badly for the task of reference-free MT evaluation, we query them for their preferences. In particular, for a fixed source sentence $\mathbf{x}$, we consider two target sentences $\tilde{\mathbf{y}}$ and $\hat{\mathbf{y}}$ and evaluate the following score difference:

$$d(\tilde{\mathbf{y}}, \hat{\mathbf{y}}; \mathbf{x}) := m(\mathbf{x}, \tilde{\mathbf{y}}) - m(\mathbf{x}, \hat{\mathbf{y}}) \qquad (1)$$

When $d > 0$, then metric m prefers $\tilde{\mathbf{y}}$ over $\hat{\mathbf{y}}$, given $\mathbf{x}$, and when $d < 0$, this relationship is reversed. In the following, we compare preferences of our metrics for specifically modified target sentences $\tilde{\mathbf{y}}$ over the human references $\mathbf{y}^\star$. We choose $\tilde{\mathbf{y}}$ to be (i) a random reordering of $\mathbf{y}^\star$, to ensure that our metrics do not have the BOW (bag-of-words) property, (ii) a word-order preserving translation of $\mathbf{x}$, i.e., (ii-a) an expert reordering of the human $\mathbf{y}^\star$ to have the same word order as $\mathbf{x}$ as well as (ii-b) a word-by-word translation, obtained either using experts or automatically. Especially condition (ii-b) tests for preferences for literal translations, a common MT-system property.

Expert word-by-word translations. We had an expert (one of the co-authors) translate 50 German sentences word-by-word into English. Table 2 illustrates this scenario. We note how bad the word-by-word translations sometimes are even for closely related language pairs such as German-English. For example, the word-by-word translations in English retain the original German verb final positions, leading to quite ungrammatical English translations.

Figure 3 shows histograms for the d statistic for the 50 selected sentences. We first check condition (i) for the 50 sentences. We observe that both MOVER + M-BERT and COSINE+LASER prefer the original human references over random reorderings, indicating that they are not BOW models, a reassuring finding. Concerning (ii-a), they are largely indifferent between correct English word order and the situation where the word order of the human reference is the same as the German. Finally, they strongly prefer the expert word-by-word translations over the human references (ii-b).

Condition (ii-a) in part explains why our metrics prefer expert word-by-word translations the most: for a given source text, these have higher lexical overlap than human references and, by (ii-a), they have a favorable target language syntax, *viz.*, where the source and target language word order are equal. Preference for translationese, (ii-b), in turn is apparently a main reason why our metrics do not perform well, by themselves and without a language model, as reference-free MT evaluation metrics. More worryingly, it indicates that cross-lingual M-BERT and LASER are not robust to the 'adversarial inputs' given by MT systems.

Automatic word-by-word translations. For a large-scale analysis of condition (ii-b) across different language pairs, we resort to automatic word-by-word translations obtained from Google Translate (GT). To do so, we go over each word in the source sentence $\mathbf{x}$ from left to right, look up its translation in GT *independently of context* and replace the word by the obtained translation. When a word has several translations, we keep the first one offered by GT. Due to context-independence, the GT word-by-word translations are of much lower quality than the expert word-by-word translations since they often pick the wrong word senses—e.g., the German word *sein* may either be a personal pronoun (*his*) or the infinitive *to be*, which would be selected correctly only by chance; cf. Table 2.

Instead of reporting histograms of d, we define a "W2W" statistic that counts the relative number of

$\mathbf{x}$	Dieser von Langsamkeit geprägte Lebensstil scheint aber ein Patentrezept für ein hohes Alter zu sein.
$\mathbf{y}^\star$	However, this slow pace of life seems to be the key to a long life.
$\mathbf{y}^\star$-random	To pace slow seems be the this life. life to a key however, of long
$\mathbf{y}^\star$-reordered	This slow pace of life seems however the key to a long life to be.
$\mathbf{x}'$-GT	This from slowness embossed lifestyle seems but on nostrum for on high older to his.
$\mathbf{x}'$-expert	This of slow pace characterized life style seems however a patent recipe for a high age to be.
$\mathbf{x}$	Putin teilte aus und beschuldigte Ankara, Russland in den Rücken gefallen zu sein.
$\mathbf{y}^\star$	Mr Putin lashed out, accusing Ankara of stabbing Moscow in the back.
$\mathbf{y}^\star$-random	Moscow accusing lashed Putin the in Ankara out, Mr of back. stabbing
$\mathbf{y}^\star$-reordered	Mr Putin lashed out, accusing Ankara of Moscow in the back stabbing.
$\mathbf{x}'$-GT	Putin divided out and accused Ankara Russia in the move like to his.
$\mathbf{x}'$-expert	Putin lashed out and accused Ankara, Russia in the back fallen to be.

Table 2: Original German input sentence $\mathbf{x}$, together with the human reference $\mathbf{y}^\star$, in English, and a randomly ($\mathbf{y}^\star$-random) and expertly reordered ($\mathbf{y}^\star$-reordered) English sentence as well as expert word-by-word translation ($\mathbf{x}'$) of the German source sentence. The latter is either obtained by the human expert or by Google Translate (GT).

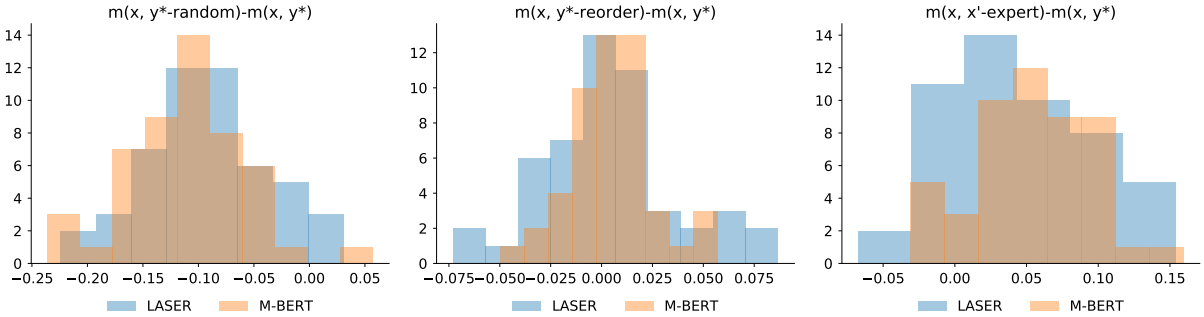


Figure 3: Histograms of d scores defined in Eq. (1). Left: Metrics based on LASER and M-BERT favor gold over randomly-shuffled human references. Middle: Metrics are roughly indifferent between gold and reordered human references. Right: Metrics favor expert word-by-word translations over gold human references.

times that $d(\mathbf{x}', \mathbf{y}^\star)$ is positive, where $\mathbf{x}'$ denotes the described literal translation of $\mathbf{x}$ into the target language:

$$\text{W2W} := \frac{1}{N} \sum_{(\mathbf{x}', \mathbf{y}^\star)} I(\, d(\mathbf{x}', \mathbf{y}^\star) > 0\,) \qquad (2)$$

Here N normalizes W2W to lie in $[0, 1]$ and a high W2W score indicates the metric prefers translationese over human-written references. Table 3 shows that reference-free metrics with original embeddings (LASER and M-BERT) either still prefer literal over human translations (e.g., W2W score of 70.2% for cs-en) or struggle in distinguishing them. Re-mapping helps to a small degree. Only when combined with the LM scores do we get adequate scores for the W2W statistic. Indeed, the LM is expected to capture unnatural word order in the target language and penalize word-by-word translations by recognizing them as much less likely to appear in a language.

Note that for expert word-by-word translations, we would expect the metrics to perform even worse.

Metrics	cs-en	de-en	fi-en
COSINE + LASER	70.2	65.7	53.9
COSINE + CLP(LASER)	70.7	64.8	53.7
COSINE + UMD(LASER)	67.5	59.5	52.9
COSINE + UMD(LASER) $\oplus$ LM	**7.0**	**7.1**	**6.4**
MOVER-2 + M-BERT	61.8	50.2	45.9
MOVER-2 + CLP(M-BERT)	44.6	44.5	32.0
MOVER-2 + UMD(M-BERT)	54.5	44.3	39.6
MOVER-2 + CLP(M-BERT) $\oplus$ LM	**7.3**	**10.2**	**6.4**

Table 3: W2W statistics for selected language pairs. Numbers are in percent.

5.2 Size of Parallel Corpora

Figure 4 compares sentence- and word-level re-mapping trained with a varying number of parallel sentences. Metrics based on M-BERT result in the highest correlations after re-mapping, even with a small amount of training data (1k). We observe that COSINE + CLP(LASER) and MOVER-2 + CLP(M-BERT) show very similar trends with a sharp increase with increasing amounts of parallel data and then level off quickly. However, the M-BERT based Mover-2 reaches its peak and outperforms the original baseline with only 1k data, while LASER needs 2k before beating the corre-

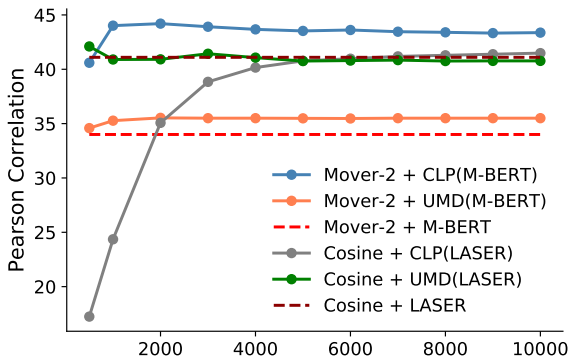


Figure 4: Average results of our metrics based on sentence- and word-based re-mappings of vector spaces as a function of different sizes of parallel corpus (x-axis).

sponding original baseline.

5.3 Human Judgments

The WMT datasets contain segment- and system-level human judgments that we use for evaluating the quality of our reference-free metrics. The segment-level judgments assign one direct assessment (DA) score to each pair of system and human translation, while system-level judgments associate each system with a single DA score averaged across all pairs in the dataset. We initially suspected the DA scores to be biased for our setup—which compares $\mathbf{x}$ with $\mathbf{y}$—as they are based on comparing $\mathbf{y}^\star$ and $\mathbf{y}$. Indeed, it is known that (especially) human professional translators "improve" $\mathbf{y}^\star$, e.g., by making it more readable, relative to the original $\mathbf{x}$ (Rabinovich et al., 2017). We investigated the validity of DA scores by collecting human assessments in the cross-lingual settings (CLDA), where annotators directly compare source and translation pairs $(\mathbf{x}, \mathbf{y})$ from the WMT17 dataset. This small-scale manual analysis hints that DA scores are a valid proxy for CLDA. Therefore, we decided to treat them as reliable scores for our setup and evaluate our proposed metrics by comparing their correlation with DA scores.

6 Conclusion

Existing semantically-motivated metrics for reference-free evaluation of MT systems have so far displayed rather poor correlation with human estimates of translation quality. In this work, we investigate a range of reference-free metrics based on cutting-edge models for inducing cross-lingual semantic representations: cross-lingual (contextualized) word embeddings and cross-lingual sentence embeddings. We have identified some scenarios in which these metrics fail, prominently their inability to punish literal word-by-word translations (the so-called "translationese"). We have investigated two different mechanisms for mitigating this undesired phenomenon: (1) an additional (weakly-supervised) cross-lingual alignment step, reducing the mismatch between representations of mutual translations, and (2) language modeling (LM) on the target side, which is inherently equipped to punish "unnatural" sentences in the target language. We show that the reference-free coupling of cross-lingual similarity scores with the target-side language model surpasses the reference-based BLEU in segment-level MT evaluation.

We believe our results have two relevant implications. First, they portray the viability of reference-free MT evaluation and warrant wider research efforts in this direction. Second, they indicate that reference-free MT evaluation may be the most challenging ("adversarial") evaluation task for multilingual text encoders as it uncovers some of their shortcomings—prominently, the inability to capture semantically non-sensical word-by-word translations or paraphrases—which remain hidden in their common evaluation scenarios.

We release our metrics under the name *XMoverScore* publicly: https://github.com/AIPHES/ACL20-Reference-Free-MT-Evaluation.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions, which greatly improved the final version of the paper. This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1. The contribution of Goran Glavaš is supported by the Eliteprogramm of the Baden-Württemberg-Stiftung, within the scope of the grant AGREE.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego,

California. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Ondrej Bojar, Yvette Graham, and Amir Kamran. 2017a. Results of the WMT17 metrics shared task. In *Proceedings of the Conference on Machine Translation (WMT)*.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017b. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 879–887.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Markus Dreyer and Daniel Marcu. 2012. Hyter: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2019. On the evaluation of machine translation systems trained with back-translation. *CoRR*, abs/1908.05204.

Marina Fomicheva and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016-Short Papers*, pages 77–82. ACL Home Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of ACL*.

Yinuo Guo and Junfeng Hu. 2019. Meteor++ 2.0: Adopt Syntactic Level Paraphrase Knowledge into Machine Translation Evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 501–506, Florence, Italy. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a ViÃ©gas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Martin Josifoski, Ivan S Paskov, Hristo S Paskov, Martin Jaggi, and Robert West. 2019. Crosslingual document embedding as reduced-rank ridge regression. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 744–752.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2019. A general framework for implicit and explicit debiasing of distributional word vector spaces. *arXiv preprint arXiv:1909.06092*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Chi-kiu Lo. 2019. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. XMEANT: Better semantic MT evaluation without reference translations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 765–771, Baltimore, Maryland. Association for Computational Linguistics.

Qingsong Ma, Ondrej Bojar, and Yvette Graham. 2018a. Results of the WMT18 metrics shared task. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018b. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy". Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting Evaluation in Context: Contextual Embeddings Improve Machine Translation Evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on*

Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States., pages 3111–3119.

Bari Saiful M Mohiuddin, Tasnim and Shafiq Joty. 2020. Lnmap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. *CoRR, abs/1309.4168*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2012. Morpheme- and POS-based IBM1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 133–137, Montréal, Canada. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: Words Helping Character N-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark.

Maja Popović, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: IBM1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 99–103, Edinburgh, Scotland. Association for Computational Linguistics.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1181. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4398–4409.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality Estimation and Translation Metrics via Pre-trained Word and Sentence Embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 101–105, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Zero-shot Transfer to Resource-lean Language

Our metric allows for estimating translation quality on new domains. However, the evaluation is limited to those languages covered by multilingual embeddings. This is a major drawback for low-resource languages—e.g., Gujarati is not included in LASER. To this end, we take multilingual USE (Yang et al., 2019) as an illustrating example which covers only 16 languages (in our sample Czech, Latvian and Finish are not included in USE). We re-align the corresponding embedding spaces with our re-mapping functions to induce evaluation metrics even for these languages, using only 2k translation pairs. Table 4 shows that our metric with a composition of re-mapping functions can raise correlation from zero to 0.10 for cs-en and to 0.18 for lv-en. However, for one language pair, fi-en, we see correlation goes from negative to zero, indicating that this approach does not always work. This observation warrants further investigation.

Metrics	cs-en	fi-en	lv-en
BLEU	0.849	0.834	0.946
COSINE + LAS	-0.001	-0.149	0.019
COSINE + CLP(USE)	0.072	-0.068	0.109
COSINE + UMD(USE)	0.056	-0.061	0.113
COSINE + CLP ∘ UMD(USE)	0.089	-0.030	0.162
COSINE + UMD ∘ CLP(USE)	**0.102**	-0.007	**0.180**

Table 4: The Pearson correlation of merics on segment-level WMT17. '∘' marks the composition of two re-mapping functions.

Setting	Metrics	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
$m(\mathbf{y}^*, \mathbf{y})$	BLEU	**0.971**	0.923	0.903	**0.979**	0.912	**0.976**	0.864	0.933
	CHRF++	0.940	**0.965**	**0.927**	0.973	**0.945**	0.960	**0.880**	**0.941**
$m(\mathbf{x}, \mathbf{y})$	*Baseline with Original Embeddings*								
	MOVER-1 + M-BERT	0.408	**0.905**	0.570	0.571	0.855	0.576	**0.816**	0.672
	COSINE + LASER	**0.821**	0.821	**0.744**	**0.754**	**0.895**	**0.890**	0.676	**0.800**
	Cross-lingual Alignment for Sentence Embedding								
	COSINE + CLP(LASER)	0.824	0.830	**0.760**	**0.766**	0.900	**0.942**	**0.757**	**0.826**
	COSINE + UMD(LASER)	**0.833**	**0.858**	0.735	0.754	**0.909**	0.870	0.630	0.798
	Cross-lingual Alignment for Word Embedding								
	MOVER-1 + RCSLS	-0.693	-0.053	0.738	0.251	0.538	0.380	0.439	0.229
	MOVER-1 + CLP(M-BERT)	0.796	0.960	0.879	0.874	0.894	0.864	0.898	0.881
	MOVER-2 + CLP(M-BERT)	**0.818**	0.971	**0.885**	**0.887**	0.878	**0.893**	0.896	**0.890**
	MOVER-1 + UMD(M-BERT)	0.610	0.956	0.526	0.599	**0.906**	0.538	0.898	0.719
	MOVER-2 + UMD(M-BERT)	0.650	**0.973**	0.574	0.649	0.888	0.634	**0.901**	0.753
	Combining Language Model								
	COSINE + CLP(LASER) $\oplus$ LM	**0.986**	0.909	0.868	0.968	0.858	0.910	0.800	0.900
	COSINE + UMD(LASER) $\oplus$ LM	0.984	0.904	0.861	**0.968**	0.850	**0.922**	**0.817**	**0.901**
	MOVER-2 + CLP(M-BERT) $\oplus$ LM	0.977	0.923	**0.873**	0.944	0.863	0.880	0.803	0.895
	MOVER-2 + UMD(M-BERT) $\oplus$ LM	0.968	**0.934**	0.832	0.951	**0.871**	0.862	0.821	0.891

Table 5: Pearson correlations with system-level human judgments on the WMT17 dataset.

Setting	Metrics	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	Average
$m(\mathbf{y}^*, \mathbf{y})$	SENTBLEU	0.233	0.415	0.285	0.154	0.228	0.145	0.178	0.234
	YISI-1	**0.319**	**0.488**	**0.351**	**0.231**	**0.300**	**0.234**	**0.211**	**0.305**
$m(\mathbf{x}, \mathbf{y})$	*Baseline with Original Embeddings*								
	MOVER-1 + M-BERT	0.005	0.229	0.179	0.115	0.100	0.039	**0.082**	0.107
	COSINE + LASER	**0.072**	**0.317**	**0.254**	**0.155**	**0.102**	**0.086**	0.064	**0.150**
	Cross-lingual Alignment for Word Embedding								
	COSINE + CLP(LASER)	0.093	0.323	0.254	**0.151**	0.112	0.086	0.074	0.156
	COSINE + UMD(LASER)	0.077	0.317	0.252	**0.145**	0.136	0.083	0.053	0.152
	COSINE + UMD $\circ$ CLP(LASER)	0.090	**0.337**	**0.255**	0.139	**0.145**	**0.090**	**0.088**	**0.163**
	COSINE + CLP $\circ$ UMD(LASER)	**0.096**	0.331	0.254	0.153	0.122	0.084	0.076	0.159
	Cross-lingual Alignment for Sentence Embedding								
	MOVER-1 + CLP(M-BERT)	**0.084**	0.279	0.207	0.147	**0.145**	**0.089**	**0.122**	**0.153**
	MOVER-2 + CLP(M-BERT)	0.063	**0.283**	0.193	0.149	0.136	0.069	0.115	0.144
	MOVER-1 + UMD(M-BERT)	0.043	0.264	0.193	0.136	0.138	0.051	0.113	0.134
	MOVER-2 + UMD(M-BERT)	0.040	0.268	0.188	0.143	0.141	0.055	0.111	0.135
	MOVER-1 + UMD $\circ$ CLP(M-BERT)	0.024	0.282	0.192	0.144	0.133	0.085	0.089	0.136
	MOVER-1 + CLP $\circ$ UMD(M-BERT)	0.073	0.277	**0.208**	0.148	0.142	0.086	0.121	0.151
	MOVER-2 + CLP $\circ$ UMD(M-BERT)	0.057	0.283	0.194	**0.149**	0.137	0.069	0.114	0.143
	Combining Language Model								
	COSINE + UMD $\circ$ CLP(LASER) $\oplus$ LM	0.288	0.455	0.226	0.321	0.263	0.159	0.192	0.272
	COSINE + CLP $\circ$ UMD(LASER) $\oplus$ LM	0.283	0.457	0.228	0.321	0.265	0.150	0.198	0.272
	MOVER-1 + CLP $\circ$ UMD(M-BERT) $\oplus$ LM	0.268	0.428	0.292	0.213	0.261	0.152	0.192	0.258
	MOVER-2 + CLP $\circ$ UMD(M-BERT) $\oplus$ LM	0.254	0.426	0.285	0.203	0.251	0.146	0.193	0.251

Table 6: Kendall correlations with segment-level human judgments on the WMT18 dataset.

Setting	Metrics	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	Average
$m(\mathbf{y}^*, \mathbf{y})$	BLEU	**0.970**	0.971	**0.986**	**0.973**	0.979	0.657	**0.978**	0.931
	METEOR++	0.945	**0.991**	0.978	0.971	**0.995**	**0.864**	0.962	**0.958**
$m(\mathbf{x}, \mathbf{y})$	*Baseline with Original Embeddings*								
	MOVER-1 + M-BERT	-0.629	0.915	0.880	0.804	0.847	0.731	**0.677**	0.604
	COSINE + LASER	**-0.348**	**0.932**	**0.930**	**0.906**	**0.902**	**0.832**	0.471	**0.661**
	Cross-lingual Alignment for Sentence Embedding								
	COSINE + CLP(LASER)	-0.305	0.934	0.937	0.908	0.904	0.801	0.634	0.688
	COSINE + UMD(LASER)	-0.241	0.944	0.933	0.906	0.902	0.842	0.359	0.664
	COSINE + UMD ∘ CLP(LASER)	**0.195**	**0.955**	**0.958**	**0.913**	0.896	**0.899**	**0.784**	**0.800**
	COSINE + CLP ∘ UMD(LASER)	-0.252	0.942	0.941	0.908	**0.919**	0.811	0.642	0.702
	Cross-lingual Alignment for Word Embedding								
	MOVER-1 + CLP(M-BERT)	**-0.163**	0.943	**0.918**	**0.941**	0.915	0.628	**0.875**	**0.722**
	MOVER-2 + CLP(M-BERT)	-0.517	0.944	0.909	0.938	0.913	0.526	0.868	0.654
	MOVER-1 + UMD(M-BERT)	-0.380	0.927	0.897	0.886	0.919	**0.679**	0.855	0.683
	MOVER-2 + UMD(M-BERT)	-0.679	0.929	0.891	0.896	**0.920**	0.616	0.858	0.633
	MOVER-1 + UMD ∘ CLP(M-BERT)	-0.348	**0.949**	0.905	0.890	0.905	0.636	0.776	0.673
	MOVER-1 + CLP ∘ UMD(M-BERT)	-0.205	0.943	0.916	0.938	0.913	0.641	0.871	0.717
	MOVER-2 + CLP ∘ UMD(M-BERT)	-0.555	0.944	0.908	0.935	0.911	0.551	0.863	0.651
	Combining Language Model								
	COSINE + UMD ∘ CLP(LASER) ⊕ LM	0.979	0.967	0.979	0.947	0.942	0.673	0.954	0.919
	COSINE + CLP ∘ UMD(LASER) ⊕ LM	0.974	0.966	0.983	0.951	0.951	0.255	0.961	0.863
	MOVER-1 + CLP ∘ UMD(M-BERT) ⊕ LM	0.956	0.960	0.949	0.973	0.951	0.097	0.954	0.834
	MOVER-2 + CLP ∘ UMD(M-BERT) ⊕ LM	0.959	0.961	0.947	0.979	0.951	-0.036	0.952	0.815

Table 7: Pearson correlations with system-level human judgments on the WMT18 dataset.

		Direct Assessment							
Setting	Metrics	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	Average
$m(\mathbf{y}^*, \mathbf{y})$	BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899	0.907
$m(\mathbf{x}, \mathbf{y})$	*Existing Reference-free Metrics*								
	IBM1-MORPHEME(Popović, 2012)	0.345	0.740	-	-	0.487	-	-	-
	IBM1-POS4GRAM(Popović, 2012)	0.339	-	-	-	-	-	-	-
	LASIM(Yankovskaya et al., 2019)	0.247	-	-	-	-	0.310	-	-
	LP(Yankovskaya et al., 2019)	0.474	-	-	-	-	0.488	-	-
	YISI-2(Lo, 2019)	0.796	0.642	0.566	0.324	0.442	0.339	0.940	0.578
	YISI-2-SRL(Lo, 2019)	0.804	-	-	-	-	-	0.947	-
	Baseline with Original Embeddings								
	MOVER-1 + M-BERT	**0.358**	0.611	**-0.396**	**0.335**	**0.559**	**0.261**	**0.880**	**0.373**
	COSINE + LASER	0.217	**0.891**	-0.745	-0.611	0.683	-0.303	0.842	0.139
	Our Cross-lingual based Metrics								
	MOVER-2 + CLP(M-BERT)	**0.625**	0.890	-0.060	**0.993**	**0.851**	**0.928**	**0.968**	**0.742**
	COSINE + CLP(LASER)	0.225	**0.894**	**0.041**	0.150	0.696	-0.184	0.845	0.381
	COSINE + UMD ∘ CLP(LASER)	0.074	0.835	-0.633	**0.498**	**0.759**	-0.201	0.610	0.277
	Our Cross-lingual based Metrics ⊕ LM								
	COSINE + CLP(LASER) ⊕ LM	0.813	0.910	-0.070	-0.735	0.931	0.630	0.711	0.456
	COSINE + UMD(LASER) ⊕ LM	0.817	0.908	-0.383	-0.902	0.929	0.573	0.781	0.389
	MOVER-2 + CLP(M-BERT) ⊕ LM	0.848	0.907	**-0.068**	**0.775**	0.963	**0.866**	0.827	**0.731**
	MOVER-2 + UMD(M-BERT) ⊕ LM	**0.859**	**0.914**	-0.181	**-0.391**	**0.970**	0.702	**0.874**	0.535

Table 8: Pearson correlations with system-level human judgments on the WMT19 dataset. '-' marks the numbers not officially reported in (Ma et al., 2019).

Chapter 9

Inducing Language-Agnostic Multilingual Representations

Inducing Language-Agnostic Multilingual Representations

Wei Zhao[†] **Steffen Eger**[†] **Johannes Bjerva**[Ψ,Φ] **Isabelle Augenstein**[Φ]
[†]Technische Universität Darmstadt [Φ]University of Copenhagen [Ψ]Aalborg University
{zhao,eger}@aiphes.tu-darmstadt.de
jbjerva@cs.aau.dk
augenstein@di.ku.dk

Abstract

Cross-lingual representations have the potential to make NLP techniques available to the vast majority of languages in the world. However, they currently require large pretraining corpora or access to typologically similar languages. In this work, we address these obstacles by removing language identity signals from multilingual embeddings. We examine three approaches for this: (i) re-aligning the vector spaces of target languages (all together) to a pivot source language; (ii) removing language-specific means and variances, which yields better discriminativeness of embeddings as a by-product; and (iii) increasing input similarity across languages by removing morphological contractions and sentence reordering. We evaluate on XNLI and reference-free MT across 19 typologically diverse languages. Our findings expose the limitations of these approaches—unlike vector normalization, vector space re-alignment and text normalization do not achieve consistent gains across encoders and languages. Due to the approaches' additive effects, their combination decreases the cross-lingual transfer gap by 8.9 points (m-BERT) and 18.2 points (XLM-R) on average across all tasks and languages, however. Our code and models are publicly available.[1]

1 Introduction

Cross-lingual text representations (Devlin et al., 2019; Conneau et al., 2019) ideally allow for transfer between *any* language pair, and thus hold the promise to alleviate the data sparsity problem for low-resource languages. However, until now, cross-lingual systems trained on English appear to transfer poorly to target languages dissimilar to English (Wu and Dredze, 2019; Pires et al., 2019) and for which only small monolingual corpora are available (Conneau et al., 2019; Hu et al., 2020; Lauscher et al., 2020), as illustrated in Fig. 1.[2]

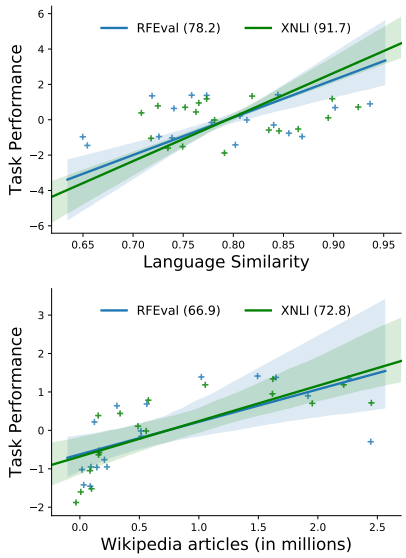


Figure 1: Zero-shot performance on XNLI and RFEval vs. language similarity to English (top), and data sizes in Wikipedia (bottom). Each point is a language; brackets give the Pearson correlation of points on the x- and y-axis. Zero-shot performance is based on the last layer of m-BERT and is standardized (zero mean, unit standard deviation) for better comparison.

As a remedy, recent work has suggested to train representations on larger multilingual corpora (Conneau et al., 2019) and, more importantly, to re-align them post-hoc so as to address the deficits of state-of-the-art contextualized encoders which have not seen any parallel data during training (Schuster et al., 2019; Wu and Dredze, 2019; Cao et al., 2020). However, re-mapping (i) can be costly, (ii) requires parallel data on word or sentence level, which may not be available abundantly in low-resource set-

[1] https://github.com/AIPHES/Language-Agnostic-Contextualized-Encoders

[2] We consider language similarity as the cosine similarity between the average representations of two languages over monolingual corpora from Wikipedia.

Proceedings of the 10th Conference on Lexical and Computational Semantics, pages 229–240
August 5–6, 2021, Bangkok, Thailand (online)

tings, and (iii) its positive effect has not yet been studied systematically.

Here, we explore *normalization* as an alternative to re-mapping. To decrease the distance between languages and thus allow for better cross-lingual transfer, we normalize (i) text inputs to encoders before vectorization to increase cross-lingual similarity, e.g., by reordering sentences according to typological features, and (ii) the representations themselves by removing their means and standard deviations, a common operation in machine and deep learning (LeCun et al., 1998; Rücklé et al., 2018). We evaluate vector normalization and post-hoc re-mapping across a typologically diverse set of 19 languages from five language families with varying sizes of monolingual corpora. However, input normalization is examined on a smaller sample of languages, as it is not feasible for languages whose linguistic features cannot be obtained automatically. We investigate two NLP tasks, and two state-of-the-art contextualized cross-lingual encoders—multilingual BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019). Further, we provide a thorough analysis to investigate the effects of these techniques: (1) across layers; (2) to decrease the cross-lingual transfer gap, especially for low-resource and dissimilar languages; and (3) to eliminate language identity signals from multilingual representations and thus induce language-agnostic representations.

We evaluate on two cross-lingual tasks of varying difficulty: (1) zero-shot cross-lingual natural language inference (XNLI) measures the transfer ability of inference from source to target languages, where only the source language is annotated;and (2) reference-free machine translation evaluation (RFEval) measures the ability of multilingual embeddings to assign adequate cross-lingual semantic similarity scores to text from two languages, where one is frequently a corrupt automatic translation.

Our contributions: We show that: (i) input normalization leads to performance gains of up to 4.7 points on two challenging tasks; (ii) normalizing vector spaces is surprisingly effective, rivals much more resource-intensive methods such as re-mapping, and leads to more consistent gains; (iii) all three techniques—vector space normalization, re-mapping and input normalization—are orthogonal and their gains often stack. This is a very important finding as it allows for improvements on a much larger scale, especially for typologically dissimilar and low-resource languages.

2 Related Work

Cross-lingual Transfer Static cross-lingual representations have long been used for effective cross-lingual transfer and can even be induced without parallel data (Artetxe et al., 2017; Lample et al., 2018). In the monolingual case, static cross-lingual embeddings have recently been succeeded by contextualized ones, which yield considerably better results. The capabilities and limitations of the contextualized multilingual BERT (m-BERT) representations is a topic of vivid discourse. Pires et al. (2019) show surprisingly good transfer performance for m-BERT despite it being trained without parallel data, and that transfer is better for typologically similar languages. Wu et al. (2019) show that language representations are not correctly aligned in m-BERT, but can be linearly re-mapped. Extending this, Cao et al. (2020) find that jointly aligning language representations to be more useful than language-independent rotations. However, we show that the discriminativeness of the resulting embeddings is still poor, i.e., random word pairs are often assigned very high cosine similarity scores by the upper layers of original encoders, especially for XLM-R.

Libovický et al. (2019) further observe that m-BERT representations of related languages are seemingly close to one another in the cross-lingual embedding space. They show that removing language-specific means from m-BERT can eliminate language identity signals. In contrast, we remove both language-specific means and variances as well as morphological contractions, and reorder sentences to reduce linguistic gaps between languages. In addition, our analysis covers more languages from a typologically broader sample, and shows that vector space normalization is as effective as other recently proposed fixes for m-BERT's limitations (especially re-mapping), but is much cheaper and orthogonal to other solutions (e.g., input normalization) in that gains are almost additive.

Linguistic Typology in NLP. Structural properties of many of the world's languages can be queried via databases such as WALS (Dryer and Haspelmath, 2013). O'Horan et al. (2016); Ponti et al. (2019) suggest to inject typological information into models to bridge the performance gap between high- and low-resource languages. Bjerva and Augenstein (2018); de Lhoneux et al. (2018); Bjerva and Augenstein (2021) show that cross-

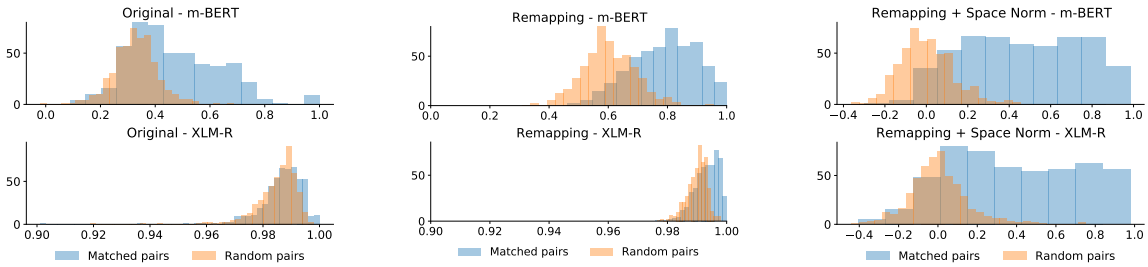


Figure 2: Histograms of cosine similarity scores of word pairs.

lingual transfer can be more successful between languages which share, e.g., morphological properties. We draw inspiration from Wang and Eisner (2016), who use dependency statistics to generate a large collection of synthetic languages to augment training data for low-resource languages. This intuition of modifying languages based on syntactic features can also be used in order to decrease syntactic and morphological differences between languages. We go further than using syntactic features, and remove word contractions and reorder sentences based on typological information from WALS.

3 Language-Agnostic Representations

Analyses by Ethayarajh (2019) indicate that random words are often assigned high cosine similarities in the upper layers of monolingual BERT. We examine this in a cross-lingual setting, by randomly selecting 500 German-English mutual word translations and random word pairs within parallel sentences from Europarl (Koehn, 2005). Fig. 2 (left) shows histograms based on the last layers of m-BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019), respectively, which show that XLM-R wrongly assigns nearly perfect cosine similarity scores (+1) to both mutual word translations (matched word pairs) and random word pairs, whereas m-BERT sometimes assigns low scores to mutual translations. This reaffirms that both m-BERT and XLM-R have difficulty in distinguishing matched from random word pairs. Surprisingly, vector space re-mapping does not seem to help for XLM-R, but better separates random from matched pairs for m-BERT (Fig. 2 (middle)). In contrast, the joint effect of normalization and re-mapping leads to adequate separation of the two distributions for both m-BERT and XLM-R, increasing the discriminative ability of both encoders.

3.1 Vector space re-alignment

m-BERT and XLM-R induce cross-lingual vector spaces in an unsupervised way—no parallel data is involved at training time. To improve upon these representations, recent work has suggested to re-map them, i.e., to use small amounts of parallel data to restructure the cross-lingual vector spaces. We follow the joint re-mapping approach of Cao et al. (2020), which has shown better results than rotation-based re-mapping.

Notation. Suppose we have k parallel corpora $C^1, \ldots, C^k$, i.e., $C^\nu = \{(\mathbf{s}^1, \mathbf{t}^1), \ldots, (\mathbf{s}^n, \mathbf{t}^n)\}$ is a set of corresponding sentence pairs from source and target languages, for $\nu = 1, \ldots, k$. We denote the alignments of words in a sentence pair $(\mathbf{s}, \mathbf{t})$ as $a(\mathbf{s}, \mathbf{t}) = \{(i_1, j_1), \ldots, (i_m, j_m)\}$, where (i, j) denotes that $\mathbf{s}_i$ and $\mathbf{s}_j$ are mutual translations. Let $f(i, \mathbf{u})$ be the contextual embedding for the i-th word in a sentence $\mathbf{u}$.

Joint Alignment via Fine-tuning. We align the monolingual sub-spaces of a source and target language by minimizing the distances of embeddings for matched word pairs in the corpus C^ν:

$$L(C^\nu, f_\Theta) = \sum_{(\mathbf{s},\mathbf{t}) \in C^\nu} \sum_{(i,j) \in a(\mathbf{s},\mathbf{t})} \|f_\Theta(i, \mathbf{s}) - f_\Theta(j, \mathbf{t}))\|_2^2 \quad (1)$$

where Θ are the parameters of the encoder f. As in Cao et al. (2020), we use a regularization term to avoid for the resulting (re-aligned) embeddings to drift too far away from the initial encoder state f_0:

$$R(C^\nu, f_\Theta) = \sum_{\mathbf{t} \in C^\nu} \sum_{i=1}^{\text{len}(\mathbf{t})} \|f_\Theta(i, \mathbf{t}) - f_0(i, \mathbf{t})\|_2^2 \quad (2)$$

Like for the multilingual pre-training of m-BERT and XLM-R, we fine-tune the encoder f on the concatenation of k parallel corpora to handle resource-

lean languages, which is in contrast to offline alignment with language-independent rotations (Aldarmaki and Diab, 2019; Schuster et al., 2019). Assume that English is a common pivot (source language) in all our k parallel corpora. Then the following objective function orients all non-English embeddings toward English:

$$\min_{\Theta} \sum_{\nu=1}^{k} L(C^{\nu}, f_{\Theta}) + R(C^{\nu}, f_{\Theta}) \qquad (3)$$

In §5, we refer to the above described re-alignment step as JOINT-ALIGN.

3.2 Vector space normalization

We add a batch normalization layer that constrains all embeddings of different languages into a distribution with zero mean and unit variance:

$$\bar{f}(i, \mathbf{s}) = \frac{f(i, \mathbf{s}) - \mu_{\beta}}{\sqrt{\sigma_{\beta}^2 + \epsilon}} \qquad (4)$$

where ϵ is a constant value for numerical stability, μ_{β} and σ_{β} are mean and variance, serving as per batch statistics for each time step in a sequence. In addition to a common effect during training, i.e., reducing covariate shift of input spaces, this additional layer in the cross-lingual setup may allow for 1) removing language identity signals, e.g. language-specific means and variances, from multilingual embeddings; and 2) increasing the discriminativeness of embeddings so that they can distinguish word pairs with different senses, as shown in Fig. 2 (right). We apply batch normalization to the last layer representations of m-BERT and XLM-R, and use a batch size of 8 across all setups. In §5, we refer to the above batch normalization step as NORM and contrast this with layer normalization. The latter yields batch-independent statistics, which are computed across all time steps for individual input sequences in a batch. This is predominantly used to stabilize the training process of RNN (Ba et al., 2016) and Transformer-based models (Vaswani et al., 2017).

3.3 Input normalization

In addition to joint alignment and vector space normalization, we investigate decreasing cross-linguistic differences between languages via the following surface form manipulation of input texts.

Removing Morphological Contractions. In many languages, e.g. Italian, prepositions and definite articles are often contracted. For instance, *de il* (*'of the'*) is usually contracted to *del*. This leads to a mismatch between, e.g., English and Italian in terms of token alignments, and increases the cross-lingual difference between the two. We segment an orthographic token (e.g. *del*) into several (syntactic) tokens (e.g. *de il*).[3] This yields a new sentence which no longer corresponds to typical standard Italian grammar, but which we hypothesise reduces the linguistic gap between Italian and English, thus increasing cross-lingual performance.

Sentence Reordering. Another typological feature which differs between languages, is the ordering of nouns and adjectives. For instance, WALS shows that Romance languages such as French and Italians often use noun-adjective ordering, e.g., *pomme rouge* in French, whereas the converse is used in English. Additionally, languages differ in their ordering of subjects, objects, and verbs. For instance, according to WALS, English firmly follows the subject-verb-object (SVO) structure, whereas there is no dominant order in German. We apply this reordering in order to decrease the linguistic gap between languages. For instance, when considering English and French, we reverse all noun-adjective pairings from French to match English. This alignment is done while considering a dependency tree. We re-align according to the typological features from WALS. Since such feature annotations are available for a large amount of languages, and can be obtained automatically with high accuracy (Bjerva et al., 2019a), we expect this method to scale to languages for which basic dependencies (such as noun-adjective attachment) can be obtained automatically. In §5, we refer to the above re-alignment step as TEXT.

4 Experiments

4.1 Transfer tasks

Cross-lingual embeddings are usually evaluated via zero-shot cross-lingual transfer for supervised text classification tasks, or via unsupervised cross-lingual textual similarity. For zero-shot transfer, fine-tuning of cross-lingual embeddings is done based on source language performance, and evaluation is performed on a held-out target language.

[3] We use UDPipe (Straka et al., 2016), which is a pipeline trained on UD treebank 2.5 (Nivre et al., 2020).

Language	Lang. family	Distance (EN-X)	Wiki-articles (in millions)	Sim level	Res level
Tagalog	α	29.3	0.08	low	low
Javanese	α	26.5	0.06	low	low
Bengali	γ	24.8	0.08	low	low
Marathi	γ	24.0	0.06	low	low
Estonian	η	23.8	0.20	low	middle
Hindi	γ	22.2	0.13	middle	low
Urdu	γ	21.7	0.15	middle	middle
Finnish	η	20.1	0.47	middle	middle
Hungarian	η	19.8	0.46	middle	middle
Afrikaans	β	19.6	0.09	middle	low
Malay	α	19.2	0.33	middle	middle
Spanish	δ	18.5	1.56	high	high
French	δ	18.2	2.16	high	high
Italian	δ	18.0	1.57	high	high
Indonesian	α	17.7	0.51	high	middle
Dutch	β	16.3	1.99	high	high
Portuguese	δ	16.2	1.02	high	high
German	β	15.6	2.37	high	high
English	β	0.0	5.98	high	high

Table 1: Languages used, with their language families: Austronesian (α), Germanic (β), Indo-Aryan (γ), Romance (δ), and Uralic (η). The cosine distances between target languages and English are measured using m-BERT.

This is, however, not likely to result in high quality target language embeddings and gives a false impression of cross-lingual abilities (Libovický et al., 2020). Zhao et al. (2020) use the more difficult task of reference-free machine translation evaluation (RFEval) to expose limitations of cross-lingual encoders, i.e., a failure to properly represent fine-grained language aspects, which may be exploited by natural adversarial inputs such as word-by-word translations.

XNLI. The goal of natural language inference (NLI) is to infer whether a premise sentence entails, contradicts, or is neutral towards a hypothesis sentence. Conneau et al. (2018) release a multilingual NLI corpus, where the English dev and test sets of the MultiNLI corpus (Williams et al., 2018) are translated to 15 languages by crowd-workers.

RFEval. This task evaluates the translation quality, i.e. similarity of a target language translation and a source language sentence. Following Zhao et al. (2020), we collect source language sentences with their system and reference translations, as well as human judgments from the WMT17 metrics shared task (Bojar et al., 2017), which contains predictions of 166 translation systems across 12 language pairs in WMT17. Each language pair has approximately 3k source sentences, each associated with one human reference translation and with the automatic translations of participating systems. As in Zhao et al. (2019, 2020), we use the Earth Mover Distance to compute the distances between source sentence and target language translations, based on the semantic similarities of their contextualized cross-lingual embeddings. We refer to this score as XMoverScore (Zhao et al., 2020) and report its Pearson correlation with human judgments in our experiments.

4.2 A Typologically Varied Language Sample

We evaluate multilingual representations on two sets of languages: (1) a default language set with 4 languages from the official XNLI test sets and 2 languages from the WMT17 test sets; (2) a diagnostic language set which contains 19 languages with different levels of data resources from a typologically diverse sample[4] covering five language families (each with at least three languages): Austronesian (α), Germanic (β), Indo-Aryan (γ), Romance (δ), and Uralic (η). For RFEval, we resort to pairs of translated source sentences and system translations. The former ones are translated from English human reference translations into 18 languages, obtained from Google Translate. For XNLI, we use translated test sets of all these languages from (Hu et al., 2020). Tab. 1 shows the overview of 19 languages which are labeled with 1) Similarity Level, i.e., the degree of similarity between target languages and English; and 2) Resource Level, i.e., the amount of data resources available in Wikipedia.

4.3 Cross-lingual Encoders

Our goal is to improve the cross-lingual abilities of established contextualized cross-lingual embeddings. These support around 100 languages and are pre-trained using monolingual language modeling.

m-BERT (Devlin et al., 2019) is pre-trained on 104 monolingual corpora from Wikipedia, with: 1) a vocabulary size of 110k; 2) language-specific tokenization tools for data pre-processing; and 3) two monolingual pre-training tasks: masked language modeling and next sentence prediction.

XLM-R (Conneau et al., 2019) is pre-trained on the CommonCrawl corpora of 100 languages, which contain more monolingual data than Wikipedia corpora, with 1) a vocabulary size of 250k; 2) a language-agnostic tokenization tool,

[4]This sample was chosen as it yields a large typological variety, with representatives from several language families across the world.

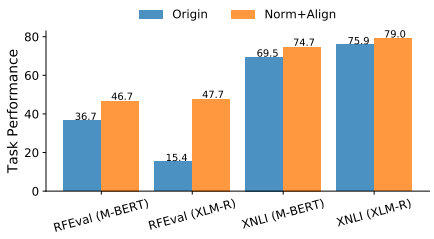


Figure 3: Results on RFEval are averaged over two language pairs (de-en and fi-en) from the WMT17 human translated test sets. Likewise, results on XNLI are averaged over four selected language pairs (en-fr, en-de, en-hi and en-es) from XNLI human translated test sets.

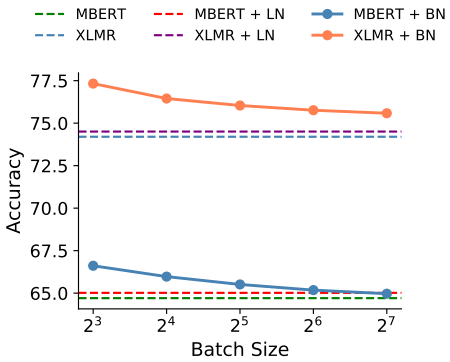


Figure 4: Results on XNLI on average across all language pairs. BN and LN denote batch and layer normalizations, respectively.

Sentence Piece (Kudo and Richardson, 2018) for data pre-processing; and 3) masked language modeling as the only monolingual pre-training task. We apply NORM, TEXT, JOINT-ALIGN and the combinations of these to the last layer of m-BERT and XLM-R, and report their performances on XNLI and RFEval in §5. To investigate the layer-wise effect of these modifications, we apply the modifications to individual layers and report the performances in §6. See the appendix for implementation details.

5 Results

Unlike re-mapping and vector space normalization, scaling input normalization to a large language sample is more difficult, as typological features differ across languages. Thus, we report the results of re-mapping and vector space normalization across 19 languages, while text normalization is evaluated on a smaller sample of languages.

Re-mapping and Vector Space Normalization. In Tab. 2, we show results on machine translated test sets. The m-BERT space modified by JOINT-ALIGN ⊕ NORM achieves consistent improvements on RFEval (+10.1 points) and XNLI (+7.6 points) on average. However, effects are different for XLM-R. The modified XLM-R outperforms the baseline XLM-R on RFEval by the largest margin (+33.5 points), but the improvement is much smaller (+2.8 points) on XNLI. These gains are not an artefact of machine-translated test sets: we observe similar gains on human-translated data (see Fig. 3).

In Tab. 3, we tease apart the sources of improvements. Overall, the impacts of NORM and JOINT-ALIGN are substantial, and their effect is additive and sometimes even superadditive (e.g., m-BERT improves by 10.1 points on RFEval when both NORM and JOINT-ALIGN are applied but only by 1.7 and 7.6 points individually). We note that the improvement from NORM is more consistent across tasks and encoders, despite its simplicity and negligible cost. In contrast, JOINT-ALIGN has a positive effect for m-BERT but it does not help for XLM-R on the XNLI task, notwithstanding the minor difference of two encoders, e.g., much larger training data and a different tokenizer used in XLM-R. We believe the poor discriminative ability of XLM-R, *viz.*, that it cannot distinguish word translations from random word pairs, leads to the inconsistent behavior of JOINT-ALIGN. As a remedy, negative examples such as random pairs could be included in Eq. (3) during training so as to decrease the discriminative gap between m-BERT and XLM-R. This suggests that future research efforts should focus on the robustness of cross-lingual alignments.

Batch vs. Layer Normalization. Unsurprisingly, the choice of batch size greatly influences XNLI performance when applying batch normalization for m-BERT and XLM-R (Fig. 4). We find that (i) the larger the batch size is, the smaller the impacts on XNLI, and (ii) a batch size of 8 performs best. Interestingly, layer normalization does not help for XNLI, even though it yields batch-independent statistics and is effective in stabilizing the training process (Vaswani et al., 2017). We note that per batch sequences with varying time steps (i.e., sentence length) are often padded with zero vectors in practice. This leads to inaccurate batch-independent statistics, as they are computed across all time steps, unlike batch normalization with per batch statistics for individual time steps. In addition to batch and layer normalizations, other nor-

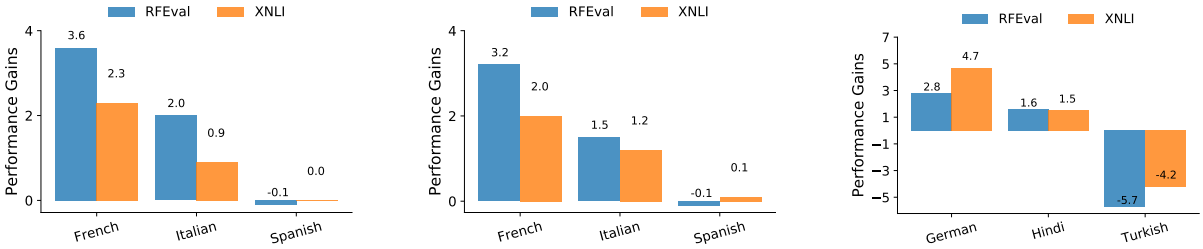
Model	Avg	△	Language Families α(4)	△	β(3)	△	γ(4)	△	δ(4)	△	η(3)	△
Original cross-lingual embeddings												
M-BERT	38.0	-	36.6	-	40.4	-	28.2	-	49.8	-	34.8	-
XLM-R	12.9	-	13.5	-	17.4	-	2.9	-	25.9	-	11.6	-
Modified cross-lingual embeddings												
M-BERT ⊕ JOINT-ALIGN ⊕ NORM	48.1	+10.1	45.9	+9.3	47.5	+7.1	32.4	+4.2	53.4	+3.6	46.0	+11.2
XLM-R ⊕ JOINT-ALIGN ⊕ NORM	46.4	+33.5	46.5	+33.0	48.2	+30.8	37.0	+34.1	53.8	+27.9	47.2	+35.6

(a) Cross-lingual Semantic Text Similarity on the RFEval task

Model	Avg	△	Language Families α(4)	△	β(3)	△	γ(4)	△	δ(4)	△	η(3)	△
Original cross-lingual embeddings												
M-BERT	64.7	-	60.8	-	69.1	-	57.9	-	73.1	-	63.4	-
XLM-R	74.8	-	72.4	-	76.3	-	70.9	-	78.4	-	76.1	-
Modified cross-lingual embeddings												
M-BERT ⊕ JOINT-ALIGN ⊕ NORM	72.3	+7.6	72.3	+11.5	75.8	+6.7	65.2	+7.3	77.4	+4.3	72.0	+8.6
XLM-R ⊕ JOINT-ALIGN ⊕ NORM	77.6	+2.8	74.8	+2.4	79.6	+3.3	73.7	+2.8	80.9	+2.5	78.8	+2.7

(b) Cross-lingual Zero-shot transfer on the XNLI task

Table 2: Overall results of established cross-lingual baselines and our modifications, for RFEval and XNLI. Brackets denote the number of languages per group. Results are averaged per group. △ is the difference between the performance of the original and the modified encoders.



(a) Removing contractions (b) Reversing adjective-noun order (c) Reversing object-verb order

Figure 5: Performance gains on RFEval and XNLI obtained by three types of TEXT operations .

Model	XNLI	RFEval
M-BERT ⊕ NORM	+1.9	+1.7
M-BERT ⊕ JOINT-ALIGN	+5.2	+7.6
M-BERT ⊕ JOINT-ALIGN ⊕ NORM	+7.6	+10.1
XLM-R ⊕ NORM	+2.5	+27.1
XLM-R ⊕ JOINT-ALIGN	−0.2	+11.6
XLM-R ⊕ JOINT-ALIGN ⊕ NORM	+2.8	+33.5

Table 3: Ablation tests of our modified encoders. Performance gains are averaged over all languages.

Model	XNLI	RFEval	Avg
M-BERT	17.4	24.5	21.0
XLM-R	11.1	37.8	24.5
M-BERT ⊕ JOINT-ALIGN ⊕ NORM	9.8	14.4	12.1
XLM-R ⊕ JOINT-ALIGN ⊕ NORM	8.4	4.3	6.3

Table 4: Performance gap (lower is better) for cross-lingual classification transfer, and reference-based and reference-free MT.

malizers such as GroupNorm (Wu and He, 2018) and PowerNorm (Shen et al., 2020) also receive attention in many communities. This raises another concern towards a systematic investigation of normalizers for future work.

Linguistic Manipulation. We apply input modifications to language pairs that contrast in either of three typological features: word contractions, noun-adjective and object-verb orderings. Fig. 5 shows that reducing the linguistic gap between languages by TEXT can sometimes lead to improvements (exemplified by m-BERT). Both French and Italian benefit considerably from both removing contractions (a) and reversing the order of adjectives and nouns (b), with no changes observed for Spanish.

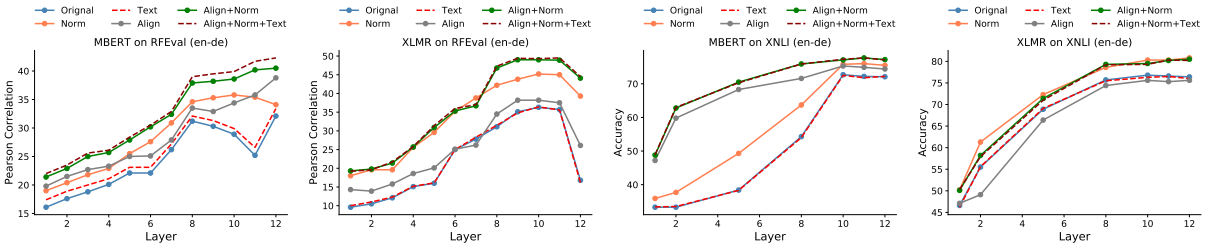


Figure 6: Results of m-BERT and XLM-R and our modifications across layers on the RFEval and XNLI tasks.

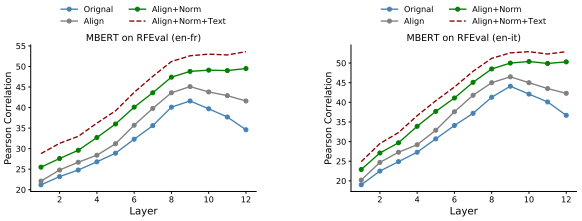


Figure 7: Results of m-BERT across layers on RFEval.

As for reversing object-verb order (c), we again see improvements for 2 out of 3 languages. We hypothesize that the few cases without gains are due to the differing frequencies of occurrences of linguistic phenomena in XNLI and RFEval. Another error source is the automatic analysis from Straka et al. (2016), and improving this pre-processing step may further increase the performance of TEXT.

6 Analysis

(Q1) *How sensitive are normalization and post-hoc re-mapping across layers?*

In Fig. 6, rather than checking results for the last layer only, we investigate improvements of our three modifications on RFEval across all layers of and XLM-R for one high-resource language pair (de-en) and one low-resource pair (jv-en) (see appendix). This reveals that, (1) for XNLI, applying JOINT-ALIGN, NORM and TEXT to the last layer of m-BERT and XLM-R consistently results in the best performance. This indicates that the modifications to the last layer could be sufficient for *supervised* cross-lingual transfer tasks. (2) However, the best results on RFEval are oftentimes obtained from an intermediate layer. Further, (3) we observe that JOINT-ALIGN is not always effective, especially for XLM-R. E.g., it leads to the worst performance across all layers on XNLI for XLM-R, even below the baseline performance. (4) Reporting improvements on only the last layer may sometimes give a false and inflated impression, especially for RFEval. E.g., the improvement (on RFEval) of the three modifications over the original embeddings is almost 30 points for the last layer of XLMR, but it is less than 15 points for the penultimate layer. (5) Normalization and re-mapping typically stabilize layer-wise variances. (6) The gains of the three modifications are largely complementary across layers. (see also Fig. 7).

(Q2) *To what extent can our modifications decrease the cross-lingual transfer gap, especially in low-resource scenarios and dissimilar languages?*

Tab. 4 shows that applying re-mapping and vector space normalization[5] to the last layer of m-BERT and XLM-R considerably reduces performance gaps *viz.*: a) zero-shot transfer performance on XNLI between the English test set and the average performance on the other 18 languages; b) the difference between mono- and cross-lingual textual similarity on RFEval, i.e., the difference between the average correlations of XMoverScore and human judgments on 19 languages obtained from *reference-based*[6] and *reference-free* MT evaluation setups. Although smaller, the remaining gaps indicates further potential for improvement. Fig. 9 shows the largest gains are on (1) low-resource languages and (2) languages most distant to English.

(Q3) *Are our modifications to contextualized cross-lingual encoders language-agnostic?*

Fig. 8 (a) shows that the centroid vectors[7] of languages within the same language family lie closely in the vector space, further showing that language

[5] We do not apply text normalization in this setup because not all languages are covered in UDPipe.

[6] Reference-based evaluation assigns semantic similarity scores to pairs of system and reference translations in English.

[7] Language centroids are representative (sentence) embeddings of languages averaged over monolingual Wikipedia data, as in Libovický et al. (2019). Although they use language families as a proxy, recent work shows that *structural similarities* of languages are a more likely candidate (Bjerva et al., 2019b).

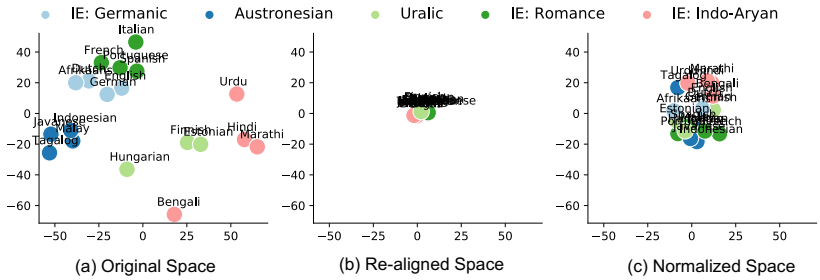


Figure 8: t-SNE distributions of language centroids based on the last m-BERT layer.

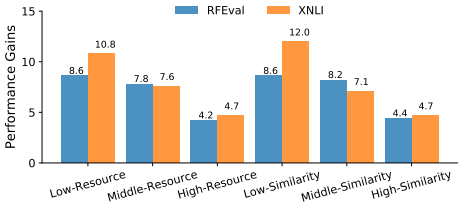


Figure 9: Performance gains across language groups for M-BERT ⊕ JOINT-ALIGN ⊕ NORM.

Model	τ	r	ρ
M-BERT	53.2	74.7	71.8
XLM-R	54.4	70.1	73.5
M-BERT ⊕ JOINT-ALIGN ⊕ NORM	17.5	57.3	21.2
XLM-R ⊕ JOINT-ALIGN ⊕ NORM	15.9	57.7	26.0

Table 5: Correlations (Kendall τ, Pearson r and Spearman ρ) between language similarities induced by m-BERT/XLM-R and WALS for 19 languages.

identity signals are stored in the m-BERT embeddings. Fig. 8 (b)+(c) shows that these signals are diminished in both re-aligned and normalized vector spaces, suggesting that the resulting embeddings in them are more language-agnostic.

(Q4) *To what extent do the typological relations learned from contextualized cross-lingual encoders deviate from those set out by expert typologists?*

Tab. 5 shows that language similarities, between English and other 18 languages, obtained from m-BERT and XLM-R have high correlations with structural language similarities[8] obtained from WALS[9] via the syntactic features listed, indicating that language identifiers stored in the original embeddings are a good proxy for the annotated linguistic features. In contrast, this correlation is smaller in the modified embedding spaces, which we believe is because language identity is a much less prominent signal in them.

7 Conclusion

Cross-lingual systems show striking performance for transfer, but their success crucially relies on two constraints: the similarity between source and target languages and the size of pre-training corpora. We comparatively evaluate three approaches to address these challenges, removing language-specific information from multilingual representations, thus learning language-agnostic representations. Our extensive experiments, based on a typologically broad sample of 19 languages, show that (vector space and input) normalization and re-mapping are oftentimes complementary approaches to improve cross-lingual performance, and that the popular approach of re-mapping leads to less consistent improvements than the much simpler and less costly normalization of vector representations. Input normalization yields benefits across a small sample of languages; further work is required for it to achieve consistent gains across a larger language sample.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions, which greatly improved the final version of the paper. This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1, as well as by the Swedish Research Council under grant agreement No 2019-04129.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the*

[8]The language similarity induced by WALS is the fraction of structural properties that have the same value in two languages among all 192 properties.

[9]WALS covers approximately 200 linguistic features over 2500 languages, annotated by expert typologists.

Association for Computational Linguistics, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3906–3911, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Johannes Bjerva and Isabelle Augenstein. 2021. Does typological blinding impede cross-lingual sharing? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 480–486, Online. Association for Computational Linguistics.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019a. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540, Minneapolis, Minnesota. Association for Computational Linguistics.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019b. What Do Language Representations Really Represent? *Computational Linguistics*, 45(2):381–389.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *CoRR*, abs/2002.03518.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *ICLR*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulic, and Goran Glavas. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *CoRR*, abs/2005.00633.

Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. 1998. Efficient backprop. In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, page 9–50, Berlin, Heidelberg. Springer-Verlag.

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4992–4997.

Jindrich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *CoRR*, abs/1911.03310.

Jindrich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. *CoRR*, abs/2004.05160.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis M. Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. *CoRR*, abs/2004.10643.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. *arXiv preprint arXiv:1610.03349*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv*.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Sheng Shen, Zhewei Yao, Amir Gholami, Michael Mahoney, and Kurt Keutzer. 2020. PowerNorm: Rethinking batch normalization in transformers. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8741–8751. PMLR.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-u files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *CoRR*, abs/1911.01464.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Yuxin Wu and Kaiming He. 2018. Group normalization. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, volume 11217 of *Lecture Notes in Computer Science*, pages 3–19. Springer.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by

reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Chapter 10

Constrained Density Matching and Modeling for Cross-lingual Alignment of Contextualized Representations

Constrained Density Matching and Modeling for Cross-lingual Alignment of Contextualized Representations

Wei Zhao WEI.ZHAO@H-ITS.ORG
Heidelberg Institute for Theoretical Studies
Fachbereich Informatik, Technische Universität Darmstadt

Steffen Eger STEFFEN.EGER@UNI-BIELEFELD.DE
Technische Fakultät, Universität Bielefeld
Fachbereich Informatik, Technische Universität Darmstadt

Editors: Emtiyaz Khan and Mehmet Gonen

Abstract

Multilingual representations pre-trained with monolingual data exhibit considerably unequal task performances across languages. Previous studies address this challenge with resource-intensive contextualized alignment, which assumes the availability of large parallel data, thereby leaving under-represented language communities behind. In this work, we attribute the data hungriness of previous alignment techniques to two limitations: (i) the inability to sufficiently leverage data and (ii) these techniques are not trained properly. To address these issues, we introduce supervised and unsupervised density-based approaches named Real-NVP and GAN-Real-NVP, driven by Normalizing Flow, to perform alignment, both dissecting the alignment of multilingual subspaces into density matching and density modeling. We complement these approaches with our validation criteria in order to guide the training process. Our experiments encompass 16 alignments, including our approaches, evaluated across 6 language pairs, synthetic data and 5 NLP tasks. We demonstrate the effectiveness of our approaches in the scenarios of limited and no parallel data. First, our supervised approach trained on 20k parallel data (sentences) mostly surpasses Joint-Align and InfoXLM trained on over 100k parallel sentences. Second, parallel data can be removed without sacrificing performance when integrating our unsupervised approach in our bootstrapping procedure, which is theoretically motivated to enforce equality of multilingual subspaces. Moreover, we demonstrate the advantages of validation criteria over validation data for guiding supervised training[1].

Keywords: Multilingual Embeddings; Cross-lingual Alignment

1. Introduction

Multilingual text encoders such as m-BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have been profiled as the de facto solutions to modeling languages at scale. However, research showed that such encoders pre-trained with monolingual data have failed to align multilingual subspaces, and exhibit strong language bias, i.e., the quality of these encoders largely differs across languages, particularly for dissimilar and low-resource languages (Pires et al., 2019; Zhao et al., 2021).

For that reason, supervised alignment techniques emerged, aiming to rectify multilingual representations post-hoc with cross-lingual supervision (Cao et al., 2020; Zhao et al., 2020; Chi et al., 2021), but previous studies are limited in scope to high-resource languages requiring large-scale par-

1. Our code and models are available at https://github.com/AIPHES/DensityAlign

allel data. In contrast, unsupervised alignment removing the dependence on parallel data allows for unlimited use in all languages (Artetxe et al., 2017, 2018). However, previous studies of unsupervised alignment focusing on static embeddings have not touched on contextualized representations.

In this work, we address cross-lingual alignment for contextualized representations, termed *contextualized alignment*, particularly in the scenarios of limited and no parallel data. In **supervised** settings, we identify two limitations responsible for the ineffectiveness of previous resource-intensive contextualized alignments: (i) the inability to sufficiently leverage data, i.e., that these techniques do not target the modeling of data density, and (ii) that they are not properly trained due to a lack of validation criteria—recent techniques, such as Wu and Dredze (2020) and Cao et al. (2020), have been trained for several epochs without access to any criteria for model selection, coming at the risk of being mistrained. To this end, we start by introducing a density-based, contextualized alignment, which dissects the alignment of multilingual subspaces into two sub-problems with one solution: density modeling and density matching, addressed by Normalizing Flows (Dinh et al., 2017). Second, in order to guide the training process, we present two validation criteria for model selection during training, and demonstrate the superiority of these criteria over validation data.

In **unsupervised settings**, aiming for unsupervised, contextualized alignment, we carry out density modeling and density matching in the form of adversarial learning (Goodfellow et al., 2014), and complement this learning process with the validation criteria mentioned previously to guide unsupervised training. Further, we identify a statistical issue of density matching in the unsupervised case: density matching only leads to a weak notion of equality of multilingual subspaces, *viz.*, equality in distribution. Accordingly, we present a bootstrapping procedure enhancing unsupervised alignment by promoting equality of multilingual subspaces. We evaluate our approaches across 6 language pairs, synthetic data and 5 NLP tasks. Our major findings are summarized as follows:

- With 20k parallel data we provided, our supervised alignment mostly surpasses Joint-Align (Cao et al., 2020) and InfoXLM (Chi et al., 2021) trained on much larger parallel data. This confirms the effectiveness of the conflation of density matching and density modeling as our alignment does. Second, our unsupervised alignment integrated in bootstrapping procedure rivals supervised counterparts, showing that parallel data can be removed without sacrificing performance. But we admit that these alignments, be it supervised or not, are poor in generalization (see §4.3), calling for an improvement in future work.

- Not only are validation criteria crucial for guiding unsupervised training, but also for supervised training. Given the performance on validation data and external tasks often correlates weakly, validation data is inappropriate for guiding supervised training. Above all, guiding contextualized alignment with validation criteria is challenging, as the model performances across tasks exhibit negative correlations in about 30% setups in our experiments, i.e., the better the alignment performs in one task, the worse it performs in the other. Thus, we base the evaluation of validation criteria on the model performances in all tasks. We find that validation criteria correlate much better than validation data (treated as criterion) with model performance on average across tasks for guiding supervised training.

2. Related Work

Recent advances in multilingual representations, such as m-BERT and XLM-R, boost the performance of cross-lingual NLP systems. However, such systems exhibit weak(er) performance for

dissimilar languages (Pires et al., 2019) and low-resource languages (Zhao et al., 2021). Accordingly, contextualized alignment emerged. Aldarmaki and Diab (2019) show that language-dependent rotation can linearly rectify m-BERT representations. Cao et al. (2020) find that jointly aligning multiple languages performs better. Zhao et al. (2020) show that removing language bias in multilingual representations mitigates the vector space misalignment between languages. More recently, Mengzhou et al. (2021) show that gradient-based alignment is effective for the languages not covered during pre-training in XLM-R. Alqahtani et al. (2021) use optimal transport to finetune multilingual representations, while Chi et al. (2021) finetune them with translation language modeling as the learning objective. However, these studies focused on supervised, resource-intensive alignment techniques and required from 250k to ca. 2M parallel sentences (or a large-scale analogy corpus) in each language pair for substantial improvement. As early attempts to remove the use for parallel data, Libovický et al. (2020) and Zhao et al. (2021) find applying vector space normalization is helpful to yield language-neural representations. However, there lacks a thorough study on unsupervised, contextualized alignment for multilingual representations.

As for unsupervised alignment, previous studies have predominantly focused on static embeddings, which mostly rely on iterative procedures in two steps, aiming to derive bilingual lexicons as cross-lingual supervision: (i) inducing seed dictionaries with different approaches, such as adversarial learning (Lample et al., 2018), similarity based heuristics (Artetxe et al., 2018) and identical strings (Artetxe et al., 2017), and (ii) applying Procrustes to augment induced lexicons (Lample et al., 2018) in an iterative fashion.

In this work, we present a principled, iterative procedure to enhance our unsupervised alignment on contextualized representations, which employs density-based approaches to induce bilingual lexicons, and then applies our bootstrapping procedure, theoretically grounded in statistics for equality of multilingual subspaces, to iteratively augment lexicons. Lastly, we complement the iterative procedure with validation criteria to guide unsupervised training. We contrast our approaches with other domain adaptation techniques in Section 6.

3. Contextualized Alignment

Let two random variables X and Y with densities P_X and P_Y describe two populations of contextual word embeddings pertaining to two languages ℓ_1 and ℓ_2, with Ω_{ℓ_1} and Ω_{ℓ_2} as two lexicons. Each occurrence of a word is associated to a separate entry in the lexicons. X maps all entries in Ω_{ℓ_1} to real-valued m-dimensional embedding vectors, denoted by $X : \Omega_{\ell_1} \to \mathbb{R}^m$, and similarly for Y. A bilingual lexicon Ω describes a set of translations between Ω_{ℓ_1} and Ω_{ℓ_2}.

Empirical inference. Assume a function $f : \mathbb{R}^m \to \mathbb{R}^m$ perfectly maps m-dimensional embedding vectors from X to Y. As standard in machine learning, a mapping function f_θ can be empirically inferred from data, with θ as model parameters. To this end, we assume data $\mathbf{M}_{\ell_1} \in \mathbb{R}^{n \times m}$ and $\mathbf{M}_{\ell_2} \in \mathbb{R}^{n \times m}$ are given, corresponding to two sets of contextual word embeddings with a common size of n for simplicity. Let a permutation matrix $\mathbf{P} \in \{0,1\}^{n \times n}$ ($\mathbf{P}\mathbf{1}_n = \mathbf{1}_n$ and $\mathbf{P}^\top \mathbf{1}_n = \mathbf{1}_n$) be a realization of Ω, serving as cross-lingual supervision when available. A random variable $\tilde{Y}$ with density $P_{\tilde{Y}}$ is a prediction of Y given X, i.e., $\tilde{Y} = f_{X \to Y}(X)$ where the subscript denotes mapping direction.

3.1. Supervised Alignment

When parallel data is available, a permutation matrix $\mathbf{P}$ can be effortless induced from parallel data with word alignment tools (Dyer et al., 2013; Jalili Sabet et al., 2020). We introduce a density-based mapping function focusing on two components: density matching and density modeling. To do so, we start by depicting the alignment of multilingual subspaces in the form of density matching between $P_{\tilde{Y}}$ and P_Y:

$$\begin{aligned} \mathrm{KL}(P_{\tilde{Y}}, P_Y) &= \mathrm{CE}(P_{\tilde{Y}}, P_Y) - \mathbb{E}_{y\sim P_Y}[\log P_Y(y)] \\ &= \|f_{X\to Y}(\mathbf{M}_{\ell_1}) - \mathbf{P}\mathbf{M}_{\ell_2}\|^2 - \mathbb{E}_{y\sim P_Y}[\log P_X(f^{-1}_{X\to Y}(y))|\det(\nabla_\theta f^{-1}_{X\to Y}(y))|] \end{aligned} \tag{1}$$

where $f_{X\to Y}$ is the trainable mapping function from X to Y. Given P_Y intractable to compute, previous supervised alignments always minimize the cross-entropy term alone by solving the least squares problem. Note that the density P_Y can be rewritten to $P_X(f^{-1}_{X\to\tilde{Y}}(y))|\det(\nabla_\theta f^{-1}_{X\to\tilde{Y}}(y))|$ by using the change-of-variable rule (assuming f is an invertible function). However, the density P_Y is still intractable given the unknown density P_X. We overcome this by introducing a generative model named Real-NVP (Dinh et al., 2017) as use case of Normalizing Flows (Rezende and Mohamed, 2015). Real-NVP is a popular example of invertible neural networks, which can be thought of as a bijective function between two domains of data points (e.g., random noise and real data). Here, we use Real-NVP to address density estimation, i.e., inferring the unknown distribution of word embeddings X and Y from a normal distribution of random noise via the change-of-variable rule.

To do so, we introduce a latent variable $Z \sim \mathcal{N}(0, \mathbf{I})$ with the normal density P_Z to describe random noise. We then use Real-NVP to infer P_Y from P_Z, denoted by $P_Y(y) = P_Z(f^{-1}_{Z\to Y}(y))|\det(\nabla_\theta f^{-1}_{Z\to Y}(y))|$ with $f_{Z\to Y}$ as a trainable mapping function from Z to Y. Lastly, we rewrite the entropy term in Eq. 1 to:

$$\mathbb{E}_{y\sim P_Y}[\log P_Y(y)] = \mathbb{E}_{y\sim P_Y}[\log \mathcal{N}(f^{-1}_{Z\to Y}(y), 0, \mathbf{I})] + \mathbb{E}_{y\sim P_Y}[\log|\det(\nabla_\theta f^{-1}_{Z\to Y}(y))|] \tag{2}$$

To consider density estimation (modeling) on both P_X and P_Y, we perform a dual form of density matching based on JS divergence. We omit the definition for simplicity. In §4, we refer to the above described approach as Real-NVP.

3.2. Unsupervised Alignment

When $\mathbf{P}$ is not given due to a lack of parallel data, we apply adversarial learning to align the two densities $P_{\tilde{Y}}$ and P_Y. As standard in adversarial training, we involve a min-max game between two components to perform density matching: (a) a discriminator distinguishing source and target word embeddings after mapping them and (b) a mapping function aligning source and target word embeddings in order to fool the discriminator. We use a popular adversarial approach, the Wasserstein GAN (Arjovsky et al., 2017), which aligns the densities $P_{\tilde{Y}}$ and P_Y by minimizing the Earth Mover distance (EMD) between these densities. To better leverage data, we include density estimation (modeling) based on Real-NVP in the procedure of adversarial training, which maximizes the data likelihood of X and Y. Taken together, we denote our density-based learning objective by:

$$\mathrm{EMD}(P_{\tilde{Y}}, P_Y) = \min_{f_{X\to Y}} \max_{h_\phi} \mathbb{E}_{y\sim P_Y}[h_\phi(y)] - \mathbb{E}_{\tilde{y}\sim P_{\tilde{Y}}}[h_\phi(\tilde{y})] + \mathbb{E}_{y\sim P_Y}[\log P_Y(y)] \tag{3}$$

where h_ϕ is a 1-Lipschitz constrained discriminator, $\tilde{y} = f_{X\to Y}(x)$ mapping X to Y. Note that $f_{X\to Y}$ is the composition of $f_{X\to Z}$ and $f_{Z\to Y}$, and the last entropy term aims to maximize the data log-likelihood of Y. As in the supervised case, we use a dual form of Eq. 3. In §4, we refer to the above described approach as GAN-Real-NVP.

Bootstrapping procedure. After adversarial learning Y and $\tilde{Y}$ are ideally equal in distribution, denoted by $Y \overset{\text{dist}}{=} \tilde{Y}$. However, this is not sufficient. For instance, let $Y \sim \text{Uniform}(-1,1)$ and $\tilde{Y} = -Y$. Clearly, Y and Y are equal in distribution, but they are identical only at the origin. Here, we derive the two following conditions that promote the equality of Y and $\tilde{Y}$ and enhance unsupervised alignment.

Proposition 1 *Given $Y \overset{\text{dist}}{=} \tilde{Y}$, $\tilde{Y}$ and Y are equal if one of the following conditions is met:*

(i) $Y = \mathbf{U}\tilde{Y}$, *where* $\mathbf{U}$ *is invertible and* $\mathbf{U}_{ij} \geq 0\ \forall i,j$.

(ii) $\text{cor}(\tilde{Y}_i, Y_i) = 1$ *for* $\forall i$, *where* $\tilde{Y}_i$ *and* Y_i *represent the i-th component in* $\tilde{Y}$ *and* Y.

Proof

(i) $P(Y \leq y) = P(\tilde{Y} \leq y)$ for all y, due to $Y \overset{\text{dist}}{=} \tilde{Y}$. If $Y = \mathbf{U}\tilde{Y}$, then $P(\tilde{Y} \leq y) = P(Y \leq y) = P(\mathbf{U}\tilde{Y} \leq y)$. If $\mathbf{U} \geq 0$, then $P(\mathbf{U}\tilde{Y} \leq y) = P(\tilde{Y} \leq \mathbf{U}^{-1}y)$. Thus, $P(\tilde{Y} \leq \mathbf{U}^{-1}y) = P(\tilde{Y} \leq y)$ for all y. This implies that $\mathbf{U} = \mathbf{I}$. Thus, $\tilde{Y} = Y$.

(ii) If $\text{cor}(\tilde{Y}_i, Y_i) = 1$ for $\forall i$, then $\text{Var}[(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})] = 0$, thus $\mathbb{E}[(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})^2] - \mathbb{E}[(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})]^2 = 0$. However, the second term equals to 0 by using $\mathbb{E}[(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})] = \frac{\mathbb{E}[\tilde{Y}_i]}{\sigma_{\tilde{y}_i}} - \frac{\mathbb{E}[Y_i]}{\sigma_{y_i}} = 0$ due to $\tilde{Y}_i \overset{\text{dist}}{=} Y_i$. Thus, $\mathbb{E}[(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})^2] = 0$, and this implies $\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} = \frac{Y_i}{\sigma_{y_i}}$ since the non-negative $(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})^2$ must be zero if its expectation is 0. Note that $\sigma_{\tilde{y}_i} = \sigma_{y_i}$ due to $\tilde{Y}_i \overset{\text{dist}}{=} Y_i$. This implies that $\tilde{Y}_i = Y_i$ for $\forall i$. ■

To design computational approaches meeting the above conditions, we introduce additional notation and the following lemma. Let $\mathbf{M}_X$, $\mathbf{M}_Y$ be embeddings from X and Y, and $\mathbf{M}_{\tilde{Y}}$=$f_\theta(\mathbf{M}_X)$.

Lemma 2 *If* $\mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^\intercal = \mathbf{M}_Y\mathbf{M}_Y^\intercal$ *and* $\mathbf{M}_{\tilde{Y}}$ *is invertible, then* $Y = \mathbf{U}\tilde{Y}$.

Proof If $\mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^\intercal = \mathbf{M}_Y\mathbf{M}_Y^\intercal$ and $\mathbf{M}_{\tilde{Y}}$ is invertible, then $\mathbf{M}_Y = \mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^{-1}\mathbf{M}_Y$. Let $\mathbf{U} = \mathbf{M}_{\tilde{Y}}^{-1}\mathbf{M}_Y$. Then, $\mathbf{M}_Y = \mathbf{M}_{\tilde{Y}}\mathbf{U}$. If this holds for all $\mathbf{M}_{\tilde{Y}}$ and $\mathbf{M}_Y$, then $Y = \mathbf{U}\tilde{Y}$. ■

In the following, we describe our computational approaches, and then include them as **constraints** in the adversarial training in order to promote the equality of Y and $\tilde{Y}$. Lastly, we discuss the connection of these constraints with canonical correlation and language isomorphism.

Graph structure. We depict $\mathbf{M}_{\tilde{Y}}$ and $\mathbf{M}_Y$ as m-dimensional vertices in two graphs, with $\mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^\intercal$ and $\mathbf{M}_Y\mathbf{M}_Y^\intercal$ as the weighted adjacency matrices on these graphs. As Lemma 2 states, minimizing the difference between these adjacency matrices allows to meet Prop.1(i). Thus, the objective becomes:

$$\text{EMD}(P_{\tilde{Y}}, P_Y) + \|\mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^\intercal - \mathbf{M}_Y\mathbf{M}_Y^\intercal\|^2 \tag{4}$$

However, we admit that Prop.1(i) cannot be strictly met, as guaranteeing $\mathbf{U} \geq 0$, i.e., $\mathbf{M}_{\tilde{Y}}^{-1}\mathbf{M}_Y \geq 0$ is not trivial. This might explain why *graph structure* is worse than *cross-correlation* in our experiments.

Cross-correlation. We maximize Pearson cross-correlation between de-meaned $\mathbf{M}_{\tilde{Y}}$ and $\mathbf{M}_Y$ in order to realize Prop.1(ii). The objective becomes:

$$\text{EMD}(P_{\tilde{Y}}, P_Y) + \| \frac{\text{diag}(\mathbf{M}_{\tilde{Y}}^\intercal \mathbf{M}_Y)}{\text{diag}(\mathbf{M}_{\tilde{Y}}^\intercal \mathbf{M}_{\tilde{Y}}) \text{diag}(\mathbf{M}_Y^\intercal \mathbf{M}_Y)} - \vec{1} \|^2 \tag{5}$$

Concerning the construction of $\mathbf{M}_Y$ and $\mathbf{M}_{\tilde{Y}}$, we use CSLS (Lample et al., 2018) to induce them from monolingual data, and then update them in an iterative fashion with Algorithm 1.

Algorithm 1: Bootstrapping Procedure

Input: $\mathbf{M}_X, \mathbf{M}_Y \leftarrow$ population word embeddings of X and Y
Input: $n \leftarrow$ number of bootstrapping iterations ▷ simulation: $n = 10$; real data: $n = 3$
Input: $f_{X \to Y} \leftarrow$ an identity matrix as initial mapping function
for $i \leftarrow 1$ **to** n **do**
  $\mathbf{M}_{\tilde{Y}} \leftarrow f_{X \to Y}(\mathbf{M}_X)$
  $\mathbf{P} \leftarrow \text{CSLS}(\mathbf{M}_Y, \mathbf{M}_{\tilde{Y}})$ ▷ induce permutation matrix
  $f_{X \to Y} \leftarrow \text{EMD}(P_{\tilde{Y}}, P_Y) + g(\mathbf{M}_Y, \mathbf{P}\mathbf{M}_{\tilde{Y}})$ ▷ g is a function of our constraints (see Eq. 4+5)
end
Output: $f_{X \to Y}$

Connection with canonical correlation. Often, cross-correlation between random vectors are computed using Canonical Correlation Analysis (CCA). Research showed that CCA is useful to improve static embeddings, but it requires finding k primary canonical variables (Faruqui and Dyer, 2014). In contrast, our solution is much cheaper to compute cross-correlation without the need for canonical variables (see Eq. 5).

Connection with language isomorphism. In graph theory, two graphs are called isomorphic when the two corresponding adjacency matrices are permutation similar. According to Eq. 4, our solution aims to minimize the difference between adjacency matrices, and as such lays the foundation of graph isomorphism—which is termed language isomorphism in the multilingual community. Taken together, our solution allows for yielding isomorphic multilingual subspaces for non-isomorphic languages such as typologically dissimilar languages.

4. Experiments

4.1. Baselines and Our Approach

Supervised alignments. (a) Rotation (Aldarmaki and Diab, 2019; Zhao et al., 2021): a linear orthogonal-constrained transformation; (b) GBDD (Zhao et al., 2020): subtracting a global language bias vector from multilingual representations; (c) FCNN: an architecture that contains three fully-connected layers followed by a tanh activation function each; (d) Joint-Align (Cao et al., 2020): jointly aligning many languages via fine-tuning; (e) InfoXLM (Chi et al., 2021): finetuning multilingual representations with translation language modeling and contrastive learning; (f) our Real-NVP.

Unsupervised alignments. (a) MUSE: the unsupervised variant of Rotation (Lample et al., 2018); (b) VecMap: a heuristic unsupervised approach based on the assumption that word translations have

similar distributions on word similarities (Artetxe et al., 2018); (c) vector space normalization (Zhao et al., 2021): removing language-specific means and variances of multilingual representations. MUSE and VecMap are popular unsupervised alignments on static embeddings; (d) our GAN-Real-NVP. For bootstrapping procedure, we use the notation: [Method]+[Constraint], where [Method] is MUSE or GAN-Real-NVP, and [Constraint] is Cross-Correlation or Graph-Structure or Procrustes—known to enhance unsupervised alignment on static embeddings.

Except for Joint-Align, InfoXLM, and Normalization, the others are trained individually across language pairs.

4.2. Validation Criterion

We present two validation criteria, and compare them with no-criteria (i.e., training for several epochs) in both supervised and unsupervised settings. In particular, we induce the 30k most confident word translations from monolingual data with CSLS, and then compute the two following criteria on these word translations.

- *Semantic criterion* was proposed for guiding the training of unsupervised alignment on static embeddings. Lample et al. (2018) assemble the 10k most frequent source words and generate target translations of these words. Next, they average cosine similarities on these translation pairs treated as validation criterion.
- *Structural Criterion*: we compute the difference between two ordered lists of singular values obtained from source and target word embeddings pertaining to the 30k most confident word translations. This criterion was initially proposed to measure language isomorphism (Dubossarsky et al., 2020).

4.3. Simulation

Bilingual Lexicon Induction (BLI) is a popular internal task known to evaluate alignment on static embeddings, as it covers ca. 100 language pairs and focuses on the understanding of the alignment itself other than its impact on external tasks. In particular, BLI bases the induction of bilingual lexicons on static word embeddings, and compares the induced lexicons with gold lexicons.

However, contextual embeddings lack such evaluation tasks. As Artetxe et al. (2020) state, when not evaluated under similar conditions, the lessons learned from static embeddings cannot transfer to contextual ones. To this end, we perform simulation to construct synthetic data as the contextual extension of BLI (CBLI), which focuses on evaluating the alignment of multilingual subspaces of contextualized embeddings.

We split CBLI data to train, validation and test sets, and report Precision@K, as in BLI evaluation. Our creation procedure is two-fold: First, we sample source embeddings from a two-dimensional Gaussian (normal) mixture distribution, and then perform different transformations on them to produce target embeddings. By doing so, we mimic typologically dissimilar languages—see Figure 1.

As for the construction of simulation setups, we adjust three parameters: (a) the occurrence for a word k, $k \in \{5, 10, \ldots, 100\}$—we use 20 words in all setups; (b) the degree of language isomorphism $t \in \{1, \ldots, 10\}$—which mimics different language pairs and (c) the distance ϵ between embeddings in train and test sets, $\epsilon \in \{0, 0.2, \ldots, 5\}$—which reflects different similarities between train and test domains. For the i-th word, in order to reflect word occurrence, we sample contextualized embeddings from a normal distribution $\mathcal{N}(\mu_i, \mathbf{I})$ for train sets, and from $\mathcal{N}(\mu_i + \epsilon, \mathbf{I})$ for validation

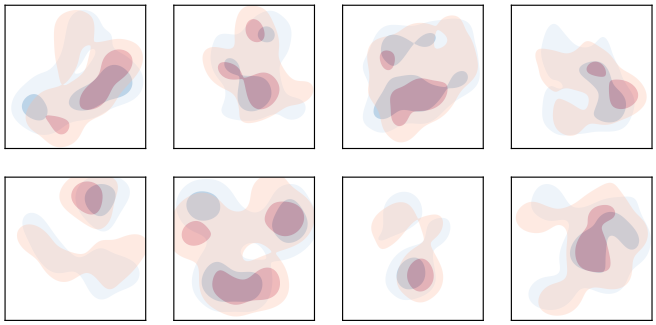
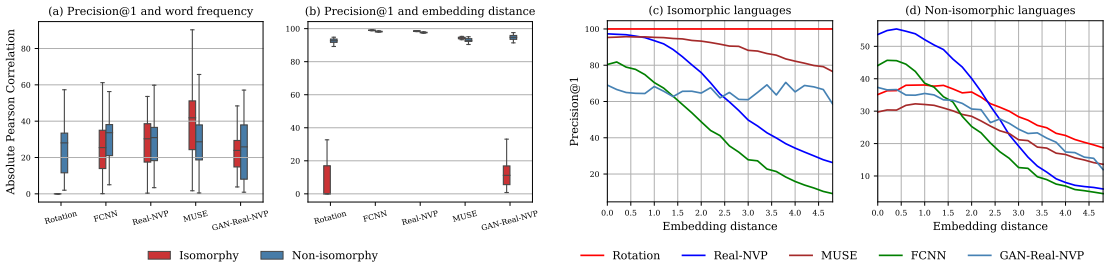


Figure 1: Eight figures are constructed in simulation. Each depicts two languages pertaining to two subspaces, colored in blue and red. Each subspace consists of up to 3 densities with each representing a word. Each density contains a number of data points sampled from a two-dimensional Gaussian distribution, as a reflection of word occurrence.

and test sets, based on the insights from the visualized m-BERT space: different instances of a word appear to follow a normal distribution (Cao et al., 2020). μ_i denotes a mean vector sampled uniformly from $[-5,5]$ for each component. For isomorphic languages ($t = 1$), we transform source into target embeddings with a rotation matrix. For non-isomorphic languages ($t > 1$), we alternate rotation with translation t times, assuming the more often we alternate, the more dissimilar two languages become.



(*I*) (*II*)

Figure 2: Absolute Pearson correlation between task performance and (a) word frequency (occurrence) and (b) similarity between train and test domains (the distance between embeddings on train and test sets). We set frequency bins $k \in \{5, 10, \ldots, 100\}$, and similarity bins $\epsilon \in \{0, 0.2, \ldots, 5\}$. We set t to 1 in all isomorphic settings, and t to 5 in non-isomorphic settings. (c)+(d) compares the generalization of approaches. Results are averaged across 10 runs.

Generalization to unseen words. Research showed that word frequency has a big impact on task performance for static embeddings (Czarnowska et al., 2019). However, Figure 2 (a)+(b) show that, in the contextual case, task performance often does not correlate with word frequency but strongly correlates with domain similarities between train and test sets. On a side note, Figure 2(b) shows that Rotation correlates poorly with embedding distance in "isomorphy", but rather highly in "non-isomorphy". This is because isomorphic spaces can be perfectly aligned via Rotation, independent of

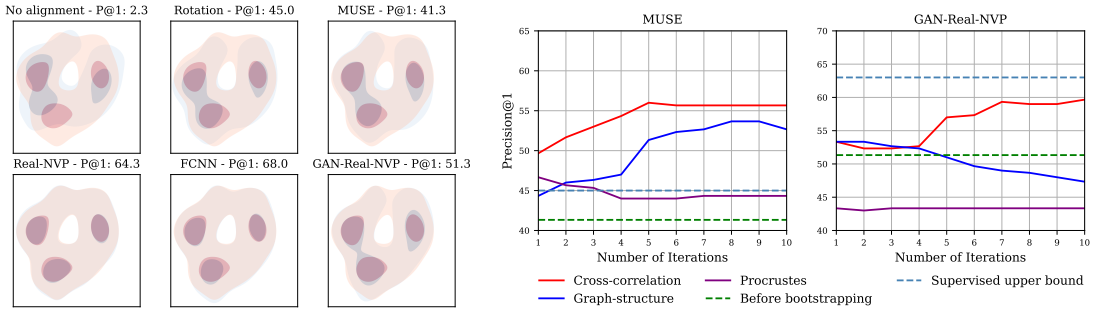
the degree of embedding distance. For non-isomorphic spaces, the bigger the embedding distance is, the worse Rotation performs, which results in a high absolute correlation.

Analyses by Glavaš et al. (2019) showed that linear alignments are much better than non-linear counterparts on static embeddings. In the following, we contrast linear with non-linear alignments on contextualized embeddings, aiming to understand in which cases one is superior to another.

In isomorphic settings, Figure 2 (c) shows that linear alignments, Rotation and MUSE, clearly win in both supervised and unsupervised settings. This means a simple, linear transformation is sufficient to align vector spaces for isomorphic languages. We mark this as a sanity test, as languages mostly are non-isomorphic (Søgaard et al., 2018).

In non-isomorphic settings, Figure 2 (d) shows that non-linear alignments, Real-NVP and GAN-Real-NVP, win by a large margin when train and test domains are similar. However, when train and test domains are dissimilar, linear alignments are indeed better. As such, non-linear alignments suffer from the issue of generalization.

Overall, we show that alignment on static and contextual embeddings yield different conclusions on *word frequency* and *the superiority of linear over non-linear alignments*. By contrasting them, we hope to provide better understanding on each.



(*I*) (*II*)

Figure 3: (I) shows how well two languages are aligned according to a visual introspection (subspace overlaps) and Precision@1; (II) compares unsupervised approaches (MUSE and GAN-Real-NVP) with the supervised counterparts (Rotation and Real-NVP) in non-isomorphic settings ($t = 5$). We set the occurrence per word k to 100.

Importance of bootstrapping procedure for unsupervised alignment. Figure 3 (I) shows how well two languages are aligned when train and test domains are similar. In this context, Real-NVP and GAN-Real-NVP win, and the resulting vector spaces are better overlapped (aligned) than others. This confirms the effectiveness of our density-based approaches. However, we still see a big performance gap between supervised and unsupervised approaches, especially for Real-NVP (64.3) vs. GAN-Real-NVP (51.3), notwithstanding large overlap in subspaces and small differences in model architectures. This confirms that density matching alone is not sufficient. Figure 3 (II) shows that after bootstrapping GAN-Real-NVP rivals Real-NVP. We also see similar results by contrasting Rotation with MUSE. Cross-correlation helps best in all cases, while graph-structure and Procrustes yield less consistent gains across approaches.

Overall, these results show that bootstrapping procedure plays a vital role in order for unsupervised alignments to rival supervised counterparts.

4.4. Experiments on Real Data

XTREME (Hu et al., 2020) has recently become popular for evaluating multilingual representations. However, it does not address word-level alignment as CBLI and BLI do, but rather focus on how multilingual representations impact cross-lingual systems. In this work, we evaluate both internal and external strengths of alignment, i.e., the internal alignment results on CBLI, and the impact of alignment on external tasks: (i) Align, RFEval and Tatoeba that require no supervised classifiers, and (ii) XNLI that requires a supervised classifier. We outline these tasks in the following:

- CBLI is the contextualized extension of BLI. Both contain a bilingual lexicon per language pair, but CBLI marks each occurrence of a word as an entry in lexicon. For each language pair, we extract 10k word translations from parallel sentences using FastAlign (Dyer et al., 2013). We report Precision@1. Note that we provide two complementary CBLI data: one is gold but simulated, while the other is real but contains noises.
- Alignment (Align) is a bilingual word retrieval task. Each language pair contains gold standard 2.5k word translations annotated by human experts. We use SimAlign (Jalili Sabet et al., 2020) to retrieve word translations from parallel sentences based on contextualized word embeddings. We report F-score that combines precision and recall.
- Reference-free evaluation (RFEval) measures the Pearson correlation between human and automatic judgments of translation quality. We use XMover (Zhao et al., 2019, 2020) to yield automatic judgment, which compares system translation with source sentence based on contextualized word embeddings. We exclude the target-side language model from XMover. Each language pair contains 3k source sentences.
- Tatoeba is a bilingual sentence retrieval task taken from XTREME. Each language pair contains 1k sentence translations. Given a source sentence, we retrieve the nearest translation from a pool of candidates based on cosine similarities between sentence embeddings. We report Precision@1.
- XNLI is a cross-lingual transfer task taken from XTREME, which aims to infer the relationship between a sentence pair of premise and hypothesis. Often, XNLI is evaluated in a zero-shot transfer setup, which measures the transfer ability from source to target languages, with cross-lingual systems trained on source language only. We report accuracy.

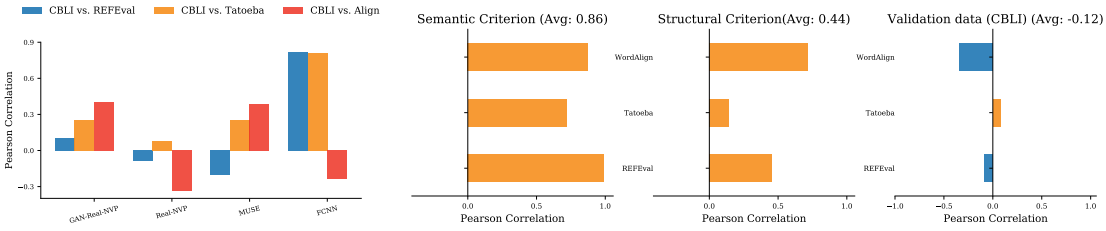
Tatoeba, CBLI and RFEval consist of six languages: German, Czech, Latvian, Finnish, Russian and Turkish, paired to English. In this work, we train alignments for these language pairs. Align considers two language pairs: German/Czech-to-English, and XNLI considers three: English-to-German/Russian/Turkish, as the other languages are not available. We consider two choices of multilingual representations: m-BERT and XLM-R.

Setup. To contrast supervised with unsupervised approaches, we consider two data scenarios: (i) limited parallel data and (ii) no parallel data. In case (i), we sample 20k (compared to ca. 250k often used in previous studies) parallel sentences from News-Commentary (Tiedemann, 2012) for Russian/Turkish-to-English, and from EuroParl (Koehn, 2005) for other languages. We use FastAlign to induce word translations from parallel sentences. Building upon these translations, we construct a permutation matrix $\mathbf{P}$ as cross-lingual supervision. In case (ii), we unpair the word translations

obtained from (i) by removing the use for the permutation matrix. As such, we compare supervised and unsupervised approaches under similar conditions, *viz.*, with similar scale of data.

How to select the best model. We compare two choices of model selection: (i) CBLI as validation data and (ii) our validation criteria.

Figure 4 (I) shows that the results on CBLI and on other tasks correlate poorly (even negatively) in both supervised and unsupervised settings. This means validation data is inappropriate for guiding both supervised and unsupervised training.



(*I*) (*II*)

Figure 4: Pearson correlation between task performances (I) and between validation criteria and task performance given by Real-NVP (II). Results are averaged across languages and encoders. For each task, we collect model performances and criteria scores over 20 epochs.

Settings	Alignments	$[-1,0]$	$(0,0.4]$	$(0.4,1]$
Supervised	FCNN-20k	50%	0%	50%
	Real-NVP-20k	33%	17%	50%
Unsupervised	MUSE-20k	17%	66%	17%
	GAN-Real-NVP-20k	17%	50%	33%

Table 1: Correlation statistics: the last three columns split the Pearson's ρ range into three intervals. Each entry denotes the percent of task pairs in which the correlation between model performances is in one of the intervals. For instance, the performances across tasks exhibit negative correlations in 17%-50% task pairs. For each task, we collect the model performances over 20 epochs. Results are averaged across language pairs and encoders.

Table 1 reports correlation statistics across approaches, showing that the model performances across tasks exhibit negative correlations in about 30% setups, i.e., the better the alignment performs in one task, the worse it performs in the other. This means that (i) guiding contextualized alignment with validation criteria is challenging, and (ii) the evaluation of validation criteria should consider the performances in all tasks. As a running example, we evaluate the two validation criteria and validation data (CBLI) treated as criterion, based on Real-NVP. Figure 4 (II) shows that both validation criteria correlate much better than validation data with the model performances across tasks. Further, *semantic criterion* wins by a large margin (0.86 versus 0.44 and -0.12). This means *semantic criterion* is the best option for guiding Real-NVP. We also see similar results on other approaches. Thus, we adopt *semantic criterion* to perform model selection in all setups.

Overall, these results show that (i) not only are validation criteria important for guiding unsupervised training, but also for guiding supervised training, and (ii) the evaluation of validation criteria should be based on the model performances in all tasks.

4.5. Results on Real Data

Table 2 contrasts unsupervised with supervised approaches. For ease of reading, we provide the average results across languages, and break them down into individual languages in Table 4 (appendix).

	m-BERT				XLM-R			
Alignments	RFEval	Tatoeba	CBLI	Align	RFEval	Tatoeba	CBLI	Align
Original	27.23	49.35	50.9	61.54	26.42	63.40	48.58	59.77
Supervised mapping functions								
Rotation-20k	38.73	55.28	58.45	**62.83**	34.67	68.60	53.67	60.85
FCNN-20k (semantic criterion)	**42.72**	61.18	55.30	61.50	36.67	**80.20**	50.88	59.87
FCNN-20k (5 epochs)	38.40	58.97	54.02	61.02	33.50	77.98	50.48	59.50
Real-NVP-20k (semantic criterion)	**42.32**	**62.87**	57.62	**62.59**	**44.17**	**80.08**	**61.63**	**62.84**
Real-NVP-20k (5 epochs)	40.12	60.70	58.52	61.80	42.24	78.75	61.76	61.20
GBDD-20k	28.77	52.28	51.42	61.71	27.13	68.85	48.42	59.81
Joint-Align-100k (3 epochs)	41.23	59.13	**64.67**	**62.30**	-	-	-	-
InfoXLM-42GB (150K training steps)	-	-	-	-	37.60	76.10	60.80	**62.94**
Unsupervised mapping functions								
Normalization	30.08	61.28	54.88	**62.54**	39.52	79.75	59.03	62.55
VecMap-20k (∼500 epochs)	30.77	55.00	**64.42**	**62.50**	-	-	-	-
MUSE-20k (5 epochs)	29.20	50.20	52.30	61.56	25.21	63.42	50.20	60.20
MUSE-20k (semantic criterion)	31.23	51.42	52.48	61.64	27.55	65.72	50.00	60.01
+ Cross-Correlation	35.25	52.90	52.87	**62.63**	32.05	69.23	49.80	60.49
+ Graph Structure	33.22	51.65	53.10	62.17	29.48	68.33	50.18	60.46
+ Procrustes	36.85	54.13	55.22	**62.71**	33.37	68.82	50.37	60.59
GAN-Real-NVP-20k (5 epochs)	32.24	59.10	56.79	61.80	39.61	77.77	60.83	60.90
GAN-Real-NVP-20k (semantic criterion)	33.90	61.20	57.03	**62.33**	41.72	79.67	**61.00**	**62.81**
+ Cross-Correlation	35.33	**62.32**	58.00	**62.70**	**42.60**	**80.50**	**61.23**	**63.15**
+ Graph Structure	34.32	61.82	56.65	**62.52**	41.55	80.02	60.83	**62.99**
+ Procrustes	**36.93**	53.95	56.05	**62.79**	33.78	67.95	51.60	60.51

Table 2: Results are averaged across language pairs. We bold numbers that significantly outperform others according to paired t-test. Joint-Align uses 100k parallel data per language pair; others only use 20k data. InfoXLM uses 42GB parallel data in total. Rotation, GBDD and Normalization with closed-form solutions do not require validation criteria for model selection.

Supervised settings. FCNN and Real-NVP training for 5 epochs are worse than those training with *semantic criterion* in almost all setups. This demonstrates the importance of validation criteria. We find that, though Joint-Align training with 100k parallel data wins on internal CBLI, it is worse than Real-NVP training with (i) merely 20k data and (ii) *semantic criterion* in the external tasks. This means Joint-Align overfits CBLI. We see similar results by contrasting VecMap with GAN-Real-NVP.

We see that the gains from all alignments on Align are much smaller than on others. This might be because SimAlign (used to induce word alignments) or the dataset cannot recognize the improved contextualized embeddings after alignment.

Real-NVP seems the strongest approach, which helps considerably for both m-BERT and XLM-R and surpasses recent InfoXLM by a large margin. InfoXLM training with much larger parallel data for 150k training steps cannot show advantages in the absence of validation criteria. Note that we do not apply validation criteria to Joint-Align and InfoXLM for further improvements, as these resource-intensive approaches have not been designed for low-resource languages, e.g., that the improvements by Joint-Align appear to vanish in the setup of limited parallel data Cao et al. (2020).

Unsupervised settings. Validation criteria are crucial: MUSE and GAN-Real-NVP with *semantic criterion* largely outperform those training for 5 epochs.

Much unlike the results in simulation, we see small performance gaps between supervised and unsupervised approaches, such as the gap between Real-NVP and GAN-Real-NVP (2 points vs 13 points in simulation). Thus, it is not surprising that the gains from the bootstrapping procedure are small in these tasks. Overall, we see *cross-correlation* is better than *graph structure* on MUSE and GAN-Real-NVP. The results for Procrustes are similar as in simulation—it improves MUSE but harms GAN-Real-NVP. GAN-Real-NVP training with *semantic criterion* and *cross-correlation* always wins, rivaling the best supervised approach Real-NVP.

Overall, these supervised and unsupervised results show that (i) validation criterion plays an essential role; (ii) density-based approaches targeting density matching and density modeling are effective in both supervised and unsupervised settings, and (iii) after bootstrapping unsupervised approaches are able to rival supervised counterparts.

	m-BERT		
Alignments	DE	RU	TR
Original	70.3	68.2	60.0
Rotation-20k	70.6	68.1	60.5
FCNN-20k (semantic criterion)	70.8	68.1	59.9
Real-NVP-20k (semantic criterion)	**73.6**	**72.7**	**62.9**
Joint-Align-100k (3 epochs)	**72.9**	**72.1**	**62.4**
GAN-R-NVP-20k (semantic criterion)	**73.5**	**72.2**	61.5
+ Cross-Correlation	**73.4**	**72.1**	61.3
+ Graph Structure	**73.4**	**72.3**	61.2
+ Procrustes	70.4	68.3	60.2

Table 3: Results on XNLI in a zero-shot cross-lingual setup from English to German/Russian/Turkish. After rectifying m-BERT with alignments we finetune m-BERT (coupled with a supervised classifier) on XNLI English train data. We restrict the evaluation to 3 languages, as the other languages on XNLI are not covered in our alignments.

Downstream Task. Table 3 shows the impacts of our approaches coupled with a supervised classifier to perform zero-shot text classification on the downstream task XNLI. While Rotation and FCNN yield better results in Table 2, their impacts vanish on XNLI. This might be because (i) rectifying m-BERT with 20k parallel data is not adequate to reflect improvements on downstream tasks, or (ii) alignment results may be orthogonal to downstream zero-shot performance. However, Real-NVP and GAN-Real-NVP trained on the same scale of data with Rotation and FCNN exhibit strong impacts on XNLI, on par with Joint-Align trained with 100k parallel data. Thus, 20k data is sufficient for our approaches to yield improvements on XNLI.

5. Conclusion

Given resource and typology disparities across languages, multilingual representations exhibit unequal capabilities between languages. Research showed that contextualized alignments overcome this challenge by producing language-agnostic representations. However, these techniques demand large parallel data, and thus cannot address the data scarcity issue in low-resource languages. Our contributions in this work are manifold. We start by introducing supervised and unsupervised density-based approaches, Real-NVP and GAN-Real-NVP, both dissecting the alignment of multilingual subspaces into density matching and density modeling in order to sufficiently leverage data. Second, we investigate the usefulness of validation criteria for guiding the training process of our approaches. Further, we present a bootstrapping procedure to enhance our unsupervised approach, which is theoretically grounded for promoting equality of multilingual subspaces. We demonstrated the effectiveness of our alignments in the scenarios of limited and no parallel data. With 20k parallel data we provided, our supervised approach mostly outperforms Joint-Align and InfoXLM trained on much larger parallel data. Next, we showed that validation criteria are imperative for guiding both supervised and unsupervised training. Finally, we demonstrated that parallel data could be removed without the loss of model performances after integrating our unsupervised approach in the bootstrapping procedure.

6. Broader Impact

As a class of domain adaptation techniques, density-based approaches have been shown useful in a range of cross-domain applications, such as image-captioning (Mahajan et al., 2020), image-to-image translation (Grover et al., 2020), alignment on static embeddings (Zhou et al., 2019) and machine translation (Setiawan et al., 2020). In this work, we showed that (i) density-based approaches could overfit validation data in the absence of validation criteria, and are weak in generalization (see §4.3), but (ii) bootstrapping procedures can improve these density-based approaches. While our analyses are limited in scope to contextualized alignment as the only cross-domain application, we hope that our results fuel future research towards effective domain adaptation techniques in other applications.

Acknowledgments

We thank the anonymous reviewers for their thoughtful comments that greatly improved the final texts. We also thank Dan Liu for her early experiments in the word alignment task. This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1 and the Klaus Tschira Foundation, Heidelberg, Germany.

References

Hanan Aldarmaki and Mona Diab. Context-aware cross-lingual mapping. In *NAACL*, 2019.

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *EMNLP*, 2021.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*, Vancouver, Canada, July 2017.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, Melbourne, Australia, July 2018.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. A call for more rigor in unsupervised cross-lingual learning. In *ACL*, pages 7375–7388, Online, July 2020.

Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual alignment of contextual word representations. In *ICLR*, 2020.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *NAACL*, 2021.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, Online, July 2020.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. In *EMNLP*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017.

Haim Dubossarsky, Ivan Vulić, Roi Reichart, and Anna Korhonen. The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures. In *EMNLP*, 2020.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In *NAACL*, 2013.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *EACL*, 2014.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *ACL*, Florence, Italy, July 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

Aditya Grover, Christopher Chute, Rui Shu, Zhangjie Cao, and Stefano Ermon. Alignflow: Cycle consistent learning from multiple domains via normalizing flows. In *AAAI*, 2020.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *ICML*, 2020.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of EMNLP*, 2020.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5. Citeseer, 2005.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *ICLR*, 2018.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. On the language neutrality of pre-trained multilingual representations. In *Findings of EMNLP*, Online, November 2020.

Shweta Mahajan, Iryna Gurevych, and Stefan Roth. Latent normalizing flows for many-to-many cross-domain mappings. In *ICLR*, 2020.

Xia Mengzhou, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Newbig, and Ahmed Hassan Awadallah. Metaxl: Meta representation transformation for low-resource cross-lingual learning. 2021.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *ACL*, Florence, Italy, July 2019.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015.

Hendra Setiawan, Matthias Sperber, Udhyakumar Nallasamy, and Matthias Paulik. Variational neural machine translation with normalizing flows. In *ACL*, July 2020.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *ACL*, 2018.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *LREC*, 2012.

Shijie Wu and Mark Dredze. Do explicit alignments robustly improve multilingual encoders? In *EMNLP*, Online, November 2020.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *EMNLP*, Hong Kong, China, November 2019.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *ACL*, 2020.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. Inducing language-agnostic multilingual representations. In **SEM*, Online, August 2021.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. Density matching for bilingual word embedding. In *NAACL*, Minneapolis, Minnesota, June 2019.

Lang	Original	Rotation	FCNN	NVP	GBDD	Joint-Align	NORM	MUSE	MUSE+C	MUSE+G	MUSE+P	GNVP	GNVP+C	GNVP+G	GNVP+P	infoXLM	VecMap
Reference-free Evaluation (m-BERT)																	
cs-en	23.6	32.5	36.2	30.2	24.4	32.5	22.5	24.8	29.7	26.7	30.4	24.9	25.3	25.2	30.5	-	25.0
de-en	29.8	31.6	33.1	34.2	30.0	39.9	31.4	29.9	31.3	31.1	31.2	32.8	33.0	32.8	31.3	-	31.8
fi-en	30.9	48.5	51.3	51.8	32.5	48.6	32.3	37.9	44.0	40.4	46.0	37.2	39.5	37.9	46.1	-	39.1
lv-en	23.0	44.0	50.2	52.1	25.7	46.1	31.3	33.4	35.9	33.5	40.9	38.0	40.7	38.8	41.0	-	24.9
ru-en	19.4	23.4	30.7	33.4	21.4	31.7	23.9	19.8	22.0	21.3	22.9	26.5	27.0	26.3	22.9	-	18.7
tr-en	36.7	52.4	54.8	52.2	38.6	48.6	39.1	41.6	48.6	46.3	49.7	44.0	46.5	44.9	49.8	-	44.5
Reference-free Evaluation (XLM-R)																	
cs-en	20.6	26.3	26.1	32.9	22.4	-	28.0	20.4	24.2	22.3	24.6	29.2	30.1	29.1	25.1	26.6	
de-en	25.1	27.4	28.2	40.5	25.4	-	39.5	25.2	26.5	26.0	26.6	40.4	40.4	40.5	27.2	37.3	
fi-en	26.6	39.9	44.6	47.7	26.0	-	41.6	29.0	35.7	32.3	37.8	44.3	47.2	44.5	38.2	36.1	
lv-en	28.3	42.3	48.6	51.2	29.6	-	44.1	30.8	37.9	32.7	40.7	47.3	48.7	47.4	40.9	38.3	
ru-en	21.4	25.6	27.7	38.0	21.5	-	34.3	21.5	23.3	22.4	24.6	35.9	36.3	36.1	25.1	34.9	
tr-en	36.5	46.5	44.8	54.7	37.9	-	49.6	38.4	44.7	41.2	45.9	53.2	52.9	51.7	46.2	52.6	
Tatoeba (m-BERT)																	
cs-en	47.5	53.1	61.4	61.3	48.3	51.1	60.6	48.2	50.8	48.7	51.6	60.6	61.3	60.7	51.5	-	47.5
de-en	76.6	79.5	84.5	87.3	78.6	89.6	86.0	77.7	79.3	78.7	78.9	86.2	86.5	86.0	78.8	-	81.2
fi-en	41.5	50.6	57.2	54.3	46.4	70.4	53.5	44.7	46.5	45.2	47.8	53.9	56.1	54.6	47.5	-	53.5
lv-en	31.7	39.8	47.4	51.2	35.0	31.1	44.0	35.3	35.0	33.8	37.1	44.1	44.7	43.8	37.0	-	33.5
ru-en	63.0	65.7	70.3	72.6	64.8	74.2	75.4	63.5	64.9	64.7	66.3	75.1	76.0	75.1	66.3	-	66.5
tr-en	35.8	43.0	46.3	50.5	40.6	38.4	48.2	39.1	40.9	38.8	43.1	47.3	49.3	48.1	42.6	-	47.6
Tatoeba (XLM-R)																	
cs-en	49.8	57.5	74.6	72.9	59.2	-	72.5	53.8	58.9	60.6	59.1	72.8	74.0	73.3	57.6	69.6	
de-en	89.0	91.0	94.5	94.9	90.1	-	95.2	88.7	91.0	89.6	90.8	94.6	95.2	94.8	90.5	95.5	
fi-en	63.8	68.6	81.3	80.8	69.3	-	79.6	65.6	69.8	67.2	68.8	80.6	81.2	80.6	67.7	68.4	
lv-en	52.2	61.3	72.6	72.4	58.9	-	71.9	56.4	63.1	59.2	61.0	71.5	71.5	71.7	59.7	51.2	
ru-en	70.2	71.5	83.1	83.9	73.2	-	83.9	71.2	70.6	70.2	72.2	83.4	84.2	84.0	70.9	85.7	
tr-en	55.4	61.7	75.1	75.6	62.4	-	75.4	58.6	62.0	63.2	61.0	75.1	76.9	75.7	61.3	86.1	
Contextual BLI (m-BERT)																	
cs-en	47.3	52.6	49.2	52.3	47.7	57.0	50.1	47.7	50.0	48.4	50.8	51.8	52.0	51.5	51.0	-	55.1
de-en	61.0	64.6	59.0	63.8	60.9	79.4	63.3	61.8	61.8	62.1	62.5	64.8	64.8	64.7	63.0	-	71.7
fi-en	36.8	48.9	45.8	45.8	37.9	74.2	43.0	40.7	40.6	40.7	43.4	45.0	47.6	44.7	44.6	-	56.7
lv-en	42.7	51.8	49.5	50.6	43.5	49.8	47.9	45.5	44.5	45.9	48.7	49.9	50.5	49.7	49.5	-	53.9
ru-en	66.5	70.9	67.6	71.4	67.0	67.4	69.2	67.2	67.5	67.8	69.5	70.5	71.0	70.6	69.9	-	76.9
tr-en	51.1	61.9	60.7	61.8	51.5	60.2	55.8	52.0	52.8	53.7	56.4	60.2	62.1	58.7	58.3	-	70.4
Contextual BLI (XLM-R)																	
cs-en	40.2	45.1	43.6	51.1	39.3	-	48.2	41.6	42.0	41.5	42.3	49.8	50.0	49.6	43.3	50.7	
de-en	54.1	57.9	53.2	64.6	54.3	-	62.7	55.2	55.1	55.3	55.4	65.1	65.3	64.9	56.3	67.3	
fi-en	44.2	51.0	46.1	60.9	44.0	-	58.1	46.8	45.7	47.2	46.6	60.2	60.5	59.9	48.3	57.1	
lv-en	44.2	50.0	48.3	57.6	44.2	-	54.8	46.4	45.7	47.3	46.5	56.9	57	56.7	47.8	53.5	
ru-en	60.9	64.4	60.3	71.7	60.7	-	70.3	61.4	61.8	61.2	62.2	71.5	71.5	71.4	63.0	72.4	
tr-en	47.9	53.6	53.8	63.9	48.0	-	60.1	48.6	48.5	48.6	49.2	62.5	63.1	62.5	50.9	64.0	
Word Alignment (m-BERT)																	
cs-en	44.0	45.0	43.4	45.7	44.1	44.2	44.6	44.0	45.0	44.5	45.0	44.6	45.1	44.9	45.1	-	45.0
de-en	79.1	80.6	79.6	79.5	79.3	80.2	80.5	79.3	80.2	79.9	80.4	80.0	80.3	80.2	80.5	-	80.0
Word Alignment (XLM-R)																	
cs-en	41.1	42.1	41.8	45.2	40.9	-	44.4	41.2	42.1	41.8	42.1	44.6	44.9	44.8	42.0	43.3	
de-en	78.5	79.6	77.9	80.5	78.7	-	80.8	78.8	78.9	79.2	79.1	81.0	81.4	81.2	79.0	82.6	

Table 4: Full results of baselines and our approaches. Real-NVP and GAN-Real-NVP are denoted by NVP and GNVP. [Method]+P/C/G denotes the integration in Procrustes refinement (P), in our bootstrapping procedure constrained with cross-correlation (C) and with graph structure (G).

Subpart C

System Comparison

Chapter 11

Better than Average: Paired Evaluation of NLP Systems

Better than Average: Paired Evaluation of NLP Systems

Maxime Peyrard[*], **Wei Zhao**[†], **Steffen Eger**[†], **Robert West**[*]
[*]EPFL, Switzerland
[†]Technische Universität Darmstadt, Germany
{maxime.peyrard,robert.west}@epfl.ch
{zhao,eger}@aiphes.tu-darmstadt.de

Abstract

Evaluation in NLP is usually done by comparing the scores of competing systems independently averaged over a common set of test instances. In this work, we question the use of averages for aggregating evaluation scores into a final number used to decide which system is best, since the average, as well as alternatives such as the median, ignores the pairing arising from the fact that systems are evaluated on the same test instances. We illustrate the importance of taking the instance-level pairing of evaluation scores into account and demonstrate, both theoretically and empirically, the advantages of aggregation methods based on pairwise comparisons, such as the Bradley–Terry (BT) model, a mechanism based on the estimated probability that a given system scores better than another on the test set. By re-evaluating 296 real NLP evaluation setups across four tasks and 18 evaluation metrics, we show that the choice of aggregation mechanism matters and yields different conclusions as to which systems are state of the art in about 30% of the setups. To facilitate the adoption of pairwise evaluation, we release a practical tool for performing the full analysis of evaluation scores with the mean, median, BT, and two variants of BT (Elo and TrueSkill), alongside functionality for appropriate statistical testing.

1 Introduction

Research is driven by evaluation results, with attention and resources being focused on methods identified as state of the art (SotA). The proper design of evaluation methodology is thus crucial to ensure progress in the field. In NLP, evaluation usually consists in comparing the averaged scores of competing systems over a common set of test instances. Indeed, averaging scores independently for each system and declaring the one with the highest average to be best is particularly simple, well understood, and mirrors the expected risk minimization paradigm used to train systems.

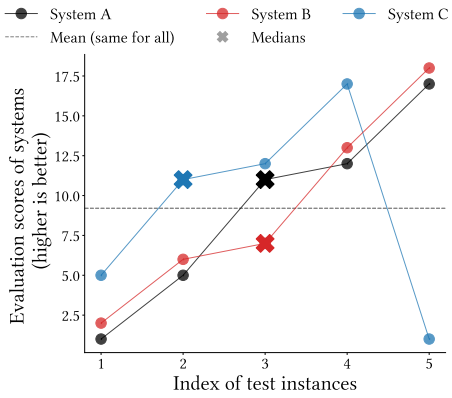


Figure 1: **Motivating example** (synthetic data). Evaluation scores of systems A, B, and C for five test instances. All systems have the same mean. C is better than A on all instances but one, so BT declares $C > A$ Also, B is better than A on all instances but one, so BT declares $B > A$, whereas the median of A is greater, and the means are the same. Overall, mean and median fail to capture the complex instance-level pairing.

Here, we critically assess the specific choice of the average to aggregate evaluation scores. In particular, we emphasize that there is a natural *instance-level pairing* between the evaluation scores of systems, which aggregation mechanisms such as the mean or median fail to take into account: as they produce a score for each system independently, systems that have the same set of scores (but potentially in different order) cannot be distinguished.

Consider the three systems A, B, and C compared on five test instances in Fig. 1. Despite a complex pairing structure, they all have the same mean score across test instances. Moreover, even though B is better than A on all test instances but one, the median of A is greater than the median of B.

In this work, we discuss an alternative aggregation mechanism: the Bradley–Terry (BT) model

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2301–2315
August 1–6, 2021.

(Bradley and Terry, 1952). BT compares systems for each test instance and estimates the latent strength of systems based on how frequently one system scores higher than another. Such paired mechanisms have already been successfully used to aggregate human judgments (Novikova et al., 2018; Sedoc and Ungar, 2020); for example, WMT evaluation protocols regularly employ TrueSkill (Herbrich et al., 2007), a Bayesian variant of BT (Sakaguchi et al., 2014).

Contributions. We contribute the first comprehensive analysis of the BT model (especially vis-à-vis mean and median) as an aggregation mechanism for comparing system scores in NLP.

(i) We illustrate the importance of accounting for instance-level pairing and discuss the conditions under which the mean, median, and BT disagree about the ordering of systems. In Sec. 3, we draw parallels with the field of statistical testing, where *paired statistical tests* are recommended when comparing paired variables. Thus, we argue that paired aggregation mechanisms such as BT are more robust alternatives to the mean and median. We support this argument with simulations in Sec. 4.

(ii) We show that the differences between mean, median, and BT matter in practice. By re-evaluating 296 real NLP evaluation setups across four tasks and 18 evaluation metrics, different aggregation mechanisms yield different conclusions as to which systems are SotA in about 30% of the setups (Sec. 5). These results hold when replacing BT by the Elo (Elo, 1978) and TrueSkill variants.

(iii) We discuss further advantages and potential limitations of BT, alongside possible resolutions, in Sec. 7.

(iv) We recommend replacing the mean by BT in future evaluations of NLP systems. To ease the adoption of more robust aggregation mechanisms, we release *Pairformance*,[1] a practical tool for performing full analyses of evaluation scores with mean, median, BT, and two variants of BT (Elo and TrueSkill). The tool reports paired evaluation results alongside appropriate statistical testing for all five aggregation mechanisms and various visualization functionalities to elucidate the pairing structure between system scores.

Code and data for replicating our analyses and experiments is available online.[2]

[1] https://github.com/epfl-dlab/pairformance

[2] https://github.com/epfl-dlab/BT-eval

2 Aggregation of evaluation results

In this section, we briefly present the three aggregation mechanisms we consider.

2.1 Terminology

A standard evaluation setup typically consists of four elements:

1. At least two **systems,** A and B, to compare, with latent strengths λ_A and λ_B that we aim to estimate.
2. A **test set** $T = \{(x_l, y_l) : l = 1, \ldots, n\}$ consisting of n test instances, where x_l is the input and y_l is the ground-truth target output.
3. An **evaluation metric** M for scoring system outputs based on target outputs y_l, resulting in the sequence of **evaluation scores** $\mathscr{M}_A = \langle M(A(x_l), y_l) : l = 1, \ldots, n \rangle$ for system A.
4. An **aggregation mechanism** Θ that decides whether system A is better than B based on the evaluation scores of the two systems. We use $\Theta_{T,M}(A,B) = \Theta(\mathscr{M}_A, \mathscr{M}_B)$ to denote the comparison mechanism between A and B on the test set T with evaluation metric M. Here, Θ outputs its guess about which system is the best (or declares the comparison inconclusive if the difference is not statistically significant). For simplicity, we drop the dependency on T and M in the notation, simply writing $\Theta(A,B)$.

For example in text summarization, x_l is a source document from the test set, y_l its corresponding reference summary, and M might be ROUGE (Lin, 2004). The decision mechanism Θ usually compares the individual systems' mean evaluation scores, where the system with the highest mean score (here mean ROUGE score) is declared better.

Consistent evaluation result. We say that the outcome of such an evaluation is *consistent* if it recovers the ordering of systems implied by the inherent strengths of systems: $\Theta(A,B) = A \iff \lambda_A > \lambda_B$.

Probabilistic model. As commonly done in the literature on statistical testing, we view the evaluation scores of a system A as n indexed random variables: $X_A^{(l)}$, $l = 1, \ldots, n$, where n is the size of the test set. Note that this sequence of random variables is not necessarily i.i.d. Furthermore, even though systems A and B are independent, their evaluation scores are not, since there is an instance-level **pairing.** Intuitively, knowing the score of A on an instance (x_l, y_l) can provide information about the expected

performance of B. For example, if A scores highly because (x_l, y_l) is an easy instance, one might expect B to also score highly.

2.2 Aggregation mechanisms

We now introduce three aggregation mechanisms Θ. We investigate their properties in subsequent sections.

Mean. This is the current standard: the system with the highest average score is declared the strongest. We denote this aggregation mechanism as MEAN. The average score of system A is computed as $E_A = \frac{1}{n}\sum_{l=1}^{n} X_A^{(l)}$.

Median. The median is an interesting alternative to the mean because it is robust to outliers. Here, the system with the highest median score is declared to be the strongest. The median score M_A of a system A is the central value in the sorted list of evaluation scores of A. We denote this aggregation mechanism as MEDIAN.

Bradley-Terry. The third option examined here is the Bradley–Terry (BT) model (Bradley and Terry, 1952). While MEAN and MEDIAN compute scores for systems A and B independently, BT is a function of the joint random variable $\left(X_A^{(l)}, X_B^{(l)}\right)$. BT estimates the relative strengths $\hat{\lambda}_A$ and $\hat{\lambda}_B$ of the two systems A and B, by comparing the evaluation scores for each test instance:

$$\mathbb{P}(A > B) = \frac{\hat{\lambda}_A}{\hat{\lambda}_A + \hat{\lambda}_B}. \quad (1)$$

Intuitively, $\mathbb{P}(A > B)$ is the probability that, for any given test instance, A scores higher than B. The BT model chooses $\hat{\lambda}_A$ and $\hat{\lambda}_B$ in order to best explain the observations. The system with the highest $\hat{\lambda}$ is declared strongest.

When considering only two systems, the latent strength $\hat{\lambda}_A$ is the number of instances for which A scores better than B (and similarly for $\hat{\lambda}_B$). When the number of systems is greater than two, BT solves an iterative optimization algorithm that is guaranteed to converge to a unique solution (Bradley and Terry, 1952). We give details about BT and its computation in the general case in Appendix E.

We denote as BT the decision mechanism based on the BT model. While it is much less common than MEAN and MEDIAN, we will see below that BT satisfies interesting properties making it a more robust alternative.

3 Comparison of assumptions

Since the roles played by A and B are symmetrical, we now assume without loss of generality that system A is better, i.e., $\lambda_A > \lambda_B$.

Proposition 1. *If $\lambda_A > \lambda_B$ then*

- MEAN *consistent* $\iff E_A - E_B > 0$,
- MEDIAN *consistent* $\iff M_A - M_B > 0$,
- BT *consistent* $\iff M_{A-B} > 0$,

where E_S and M_S are the mean and median of the evaluation scores of system S, and M_{A-B} is the median of the differences between the evaluation scores of A and B. Note that E_S, M_S, and M_{A-B} are all random variables.

The proof is given in Appendix B. Note that, whereas the expectation is linear ($E_A - E_B = E_{A-B}$), the median is not (in general, $M_A - M_B \neq M_{A-B}$).

Robustness to ouliers. The mean is not robust to outliers: E_{A-B} can be swayed above or below the threshold of 0 by a small number of test instances for which the difference between system scores is large. On the contrary, the median is a robust statistic that cannot be easily influenced by outliers. Similarly, BT is robust to outliers because its decision is based on the median of differences M_{A-B}.

Importance of pairing. The critical difference between BT, MEAN, and MEDIAN, is that only BT preserves the pairing information. Both MEAN and MEDIAN compute a statistic from the (unordered) set of scores $X_A^{(l)}$ and $X_B^{(l)}$ independently and then compare the aggregate statistics, losing the pairing structure. If the pairing actually does not matter, any permutation of the indices of system scores leaves the distribution of paired evaluation scores unchanged. This happens, for example, when both $X_A^{(l)}$ and $X_B^{(l)}$ are i.i.d.[3]

However, in the general case, the pairing matters. One particular example is when there exist different types of test instances and systems behave differently for different types, e.g., when there are *easy* instances on which all systems have higher scores. For example, consider the three systems and their evaluation scores on five test instances in Fig. 1. System A is worse than C on all instances but one, so $C > A$ according to BT, yet the median of A is greater than the median of C (10 vs. 7). At the same time, B outperforms C on all instances

[3] More generally, when the two sequences of random variables are exchangeable.

but one, so $B > C$ according to BT. For MEDIAN and MEAN, which ignore the pairing, A and B are completely equivalent, even though there is a clear difference regarding which system is more likely to be the best. This difference is revealed in the pairing structure. In general, any mechanism ignoring the pairing cannot capture the difference between A and B.

Choosing an aggregation mechanism. In Prop. 1, we stated the conditions for each mechanism to be *consistent*. Choosing an aggregation mechanism for a specific evaluation setup boils down to deciding what condition is more likely to hold in the setup. Note that none of the conditions implies any other condition in Prop. 1.

When comparing BT against MEAN (or MEDIAN), there are three possible scenarios: (i) BT agrees with MEAN (or MEDIAN), (ii) BT is consistent but MEAN (or MEDIAN) is not, and (iii) MEAN (or MEDIAN) is consistent but BT is not.

In case (i), it does not matter whether we use BT or MEAN (or MEDIAN).

In case (ii), for most instances, the better system has a higher score than the worse system, but MEAN (or MEDIAN) fails. For example, MEAN may be swayed by outliers, and MEDIAN may be swayed by jumps in score lists as in the example above.

In case (iii), for most instances, the better system has a lower score than the worse system, yet particular variations in the marginals make the MEAN or MEDIAN get the ordering correct. This is a very peculiar scenario: for MEAN, it implies that on the few instances on which the better system did better, the difference between evaluation scores was large enough to lift the mean of the better system above the other. We argue that if one really believes that the evaluation setup is likely to be in case (iii), then one does not trust the evaluation setup in the first place. It corresponds to assuming that the observed scores are inconsistent for the majority of test instances. If this is the case, one should rather improve the evaluation setup (e.g., metric, test set) in order to be more representative of the phenomena that one desires to capture.

Overall, the condition making BT consistent appears to be the most natural one. Trusting MEAN or MEDIAN more than BT implies holding an unintuitive belief about the evaluation setup, namely that the better system does worse than the worse system on a majority of test instances.

From another perspective, the random variables $E_A - E_B$ (MEAN) and $M_A - M_B$ (MEDIAN) are less likely to be (correctly) greater than zero in the presence of (i) complex pairing structures or (ii) outliers. The variable M_{A-B} (BT), on the contrary, is not affected by complex pairings or outliers.

3.1 Graphical criterion

Fig. 2 summarizes the problem of ignoring the pairing and offers a graphical criterion to understand the decisions made by MEAN, MEDIAN, and BT. In each plot, the densities are estimated by placing test instances at coordinates given by the evaluation scores of the two systems. The evaluation scores of A (green) are on the x-axis, and the evaluation scores of B (blue) on the y-axis. We also plot the *marginal distributions* of evaluation scores, from which we can read off means and medians. When the mean of $X_B^{(l)}$ is greater than that of $X_A^{(l)}$, the two extended lines representing the means meet in the upper triangle (above the line $X_A = X_B$), and analogously for the median. But mean and median being only functions of the marginals, they completely ignore the pairing. Fig. 2 illustrates this by depicting three completely different pairing structures where the marginals (and thus the means and medians) of A and B remain unchanged. (In Appendix A.1, we explain how to generate infinitely many such examples.) On the contrary, BT, being a property of the pairing (the 2D density), predicts that B is better than A when there is more mass in the upper triangle, i.e., more instances for which B scores higher than A. In the middle figure, the pairing indicates that A is better than B, in disagreement with the decisions of MEAN and MEDIAN.

3.2 Connection with statistical testing

The above discussion about the differences between MEAN, MEDIAN, and BT has interesting parallels with statistical testing.

When comparing the means of two systems over the same test set, the recommended statistical test is the *paired* t-test (Fisher, 1935). When comparing medians instead of means, the appropriate test is the sign test, which measures whether the median of the difference is significantly differerent from zero. Interestingly, the statistic of the sign test is precisely the one in the condition for BT to be consistent (see Prop. 1). Wilcoxon's signed-rank test (Wilcoxon, 1945) is often used as an alternative to the sign test because it has more statistical power (at the cost of making more assumptions). However,

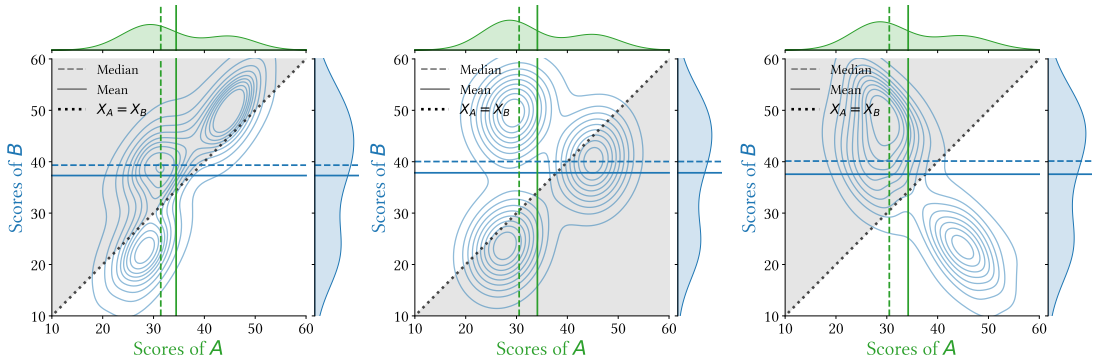


Figure 2: These 2D plots represent the distribution of test instances with coordinates given by the scores of the two systems being compared, i.e., the x-axis is the score $X_A^{(l)}$ of system A on some test instance (x_l, y_l), and the y-axis is the score $X_B^{(l)}$ of system B on the same instance. While the 3 plots represent different instance-level performances of A and B, the marginal (unpaired) distribution of scores of A and B remain unchanged. From such 2D plots, not only do we see the global structure of the pairing between the scores of A and B, we can also read off the decision of MEAN, MEDIAN and BT based on simple **geometrical criteria:** (i) if the prolongation of the means intersect above the $X_A = X_B$ line, then MEAN predicts that A is better, (ii) if the prolongation of the medians intersect above the $X_A = X_B$ line, then MEDIAN predicts that A is better, (iii) if there is more mass in the upper-left triangle, then BT predicts that system A is better. The latter case corresponds to most of the test instances being located in the upper-left triangle ($A > B$). The half-space with more mass is shaded.

Divine et al. (2018) showed that Wilcoxon's signed-rank test does not always properly account for the pairing of data, unlike the sign test.

When performing statistical testing, it seems obvious that we should use the paired version of tests when the data is naturally paired (Rankel et al., 2011). Even works discussing statistical testing in NLP recommend Wilcoxon's signed-rank test (Graham, 2015; Owczarzak et al., 2012; Dror et al., 2018). Yet, to obtain aggregated scores for systems, the community still mostly uses aggregation mechanisms that ignore the pairing, such as MEAN. MEDIAN is the outlier-resistant version of MEAN, and BT is the paired variant of MEDIAN. Whenever one recommends a paired test of medians, such as the sign test or Wilcoxon's signed-rank test, to obtain p-values, one should use BT to compare system scores.

4 Simulations with synthetic data

Next, we perform simulations to extend the analysis of the previous section to (i) $N > 2$ systems, (ii) finitely many test samples, (iii) a practical implementation of BT (for $N > 2$ systems, BT is an iterative optimization algorithm, as discussed in Appendix E).

We synthesize evaluation scores with various properties starting with systems of predefined implicit strengths λ_i. To create situations where the pairing of evaluation scores matters, we introduce multiple test instance types. For each type, systems perform differently but still have the same relative strength ($\mathbb{P}(A > B)$), differing only by an added offset. For example, the evaluation scores obtained by A and B could be sampled from $\mathcal{N}(\lambda_A, \sigma)$ and $\mathcal{N}(\lambda_B, \sigma)$ for one test instance type, and by $\mathcal{N}(\lambda_A + \epsilon, \sigma)$ and $\mathcal{N}(\lambda_B + \epsilon, \sigma)$ for another type, with ϵ being the offset. We sample evaluation setups by varying the following properties: the number of systems, the number of test instances, the percentage of outliers, the numbers of test instance types, and the level of noise. This results in 2,880 simulated evaluation setups. In Appendix A.2, we give the detailed algorithm and parameters used to generate the data.

In Fig. 3, we report Kendall's τ between the latent scores λ_i and the aggregated scores estimated by MEAN, MEDIAN, and BT. When the evaluation setup does not present any difficulty (Fig. 3(a, b)), all aggregation mechanisms work equally well (within each other's 95% error bounds), improving with more samples (Fig. 3(b)) and deteriorating with more systems (Fig. 3(a)). Unsurprisingly, MEAN fails in the presence of outliers, whereas MEDIAN and BT are unaffected (Fig. 3(c, e, f)). When several types of test instances are considered, MEDIAN begins to fail (Fig. 3(d)), which is made worse when outliers are also present (Fig. 3(f)). Overall, BT is more robust and does not fail when the pairing matters Fig. 3(g, h).

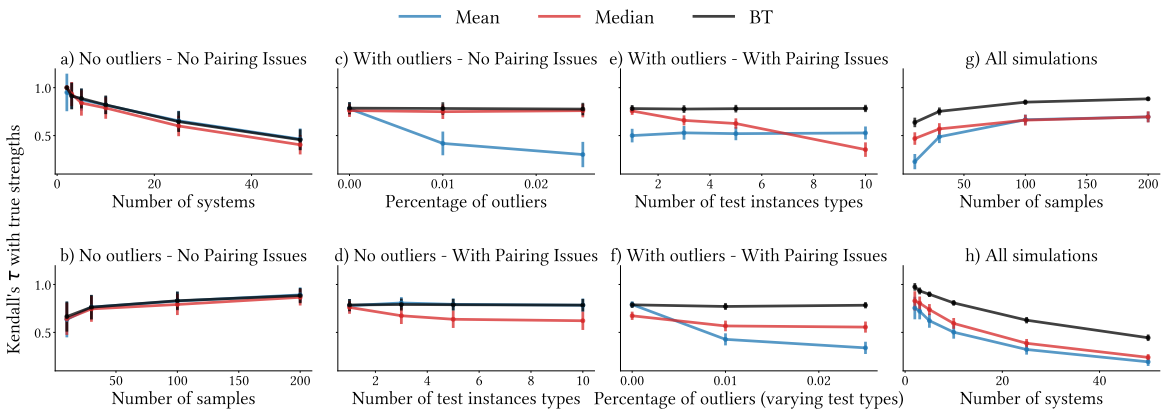


Figure 3: The y-axis is the Kendall's τ correlation between latent scores λ_i of systems and the scores obtained after aggregating simulated evaluation scores with MEAN, MEDIAN, or BT. Fig. 3(a) and Fig. 3(b) corresponds to the intuitive case where no problem occurs (no outliers, no pairing issues). Fig. 3(c) adds outlier problems only, and Fig. 3(d) adds pairing issues only by increasing the number of types of test instances. Fig. 3(e) and (f) show the combined effect of outliers and pairing issues. Finally, Fig. 3(g) and Fig. 3(h) collect all the simulations. The error bars represent 95% confidence intervals obtained with bootstrap resampling.

5 Analysis of empirical data

In this section, we perform large-scale experiments using real evaluation scores from four NLG tasks. For summarization, we use the TAC-08, TAC-09, TAC-11 and CNN/DM (Hermann et al., 2015) datasets. For machine translation, we use the shared tasks of WMT-17 (Bojar et al., 2017), WMT-18 (Ma et al., 2018), and WMT-19 (Ma et al., 2019). For image captioning, we use the MSCOCO (Lin et al., 2014) dataset, and for dialogue, we use the PersonaChat and TopicalChat (Mehri and Eskenazi, 2020) datasets. The evaluation scores are obtained with a total of 18 different evaluation metrics: BLEU-[1,2,3,4] (Papineni et al., 2002), ROUGE-[1,2,L] (Lin, 2004), ROUGE-WE-[1,2] (Ng and Abrecht, 2015), JS-[1,2] (Lin et al., 2006), S3-[pyr, resp] (Peyrard et al., 2017), CIDEr (Vedantam et al., 2015), Chrfpp (Popovic, 2017), METEOR (Lavie and Agarwal, 2007), MoverScore (Zhao et al., 2019), and BERTScore (Zhang et al., 2020). Some metrics are only available for some task; e.g., CIDEr, METEOR are only available for the image captioning task. We provide details about datasets, metrics, and their statistics in Appendix A.3.

Overall, across datasets and metrics we have 296 evaluation setups, 73,471 pairs of systems, and 91,197 test instances. We also experiment with sub-sampling different sizes of test sets (see Appendix A.3) to simulate varying train/dev/test splits or cross-validation.

5.1 Comparison of BT, MEAN, and MEDIAN

In Table 1, we report the disagreement between aggregation mechanisms over all the data with three measures: the percentage of pairs ranked in a different order (rescaled version of Kendall's τ), the percentage of setups where the state-of-the-art (SotA) systems are different, and the percentage of setups where the top 3 systems are different (compared as sets). A significant fraction of pairs of systems (about 10%) are ranked differently by different mechanisms. More importantly, top systems are often different (in about 40% of setups for top 1 and 50% for top 3). We can conclude that the choice of aggregation mechanism has a real impact on evaluation outcome. The observed disagreement between the three aggregation metrics implies that we are not in the case depicted by Fig. 3(a) and Fig. 3(b), i.e., the pairing matters and there are outliers in real data. In the next paragraphs, we break down the disagreement per evaluation metric, task, and test set size. Detailed results are provided in Appendix C.

Which metrics are impacted most? We report in Fig. 4(a) the percentage of disagreement between aggregation mechanisms per metric averaged over datasets, when subsampling test sets of different sizes uniformly (see Appendix A.3 for details). While most metrics are available for all four tasks, METEOR and CIDEr are only available for the captioning task. Therefore, the observed disagreements for these metrics may be a feature of the task instead of the metrics. Interestingly, recent metrics

	Disagree	≠ SotA	≠ Top-3
MEAN VS.MEDIAN	4%	18%	30%
MEAN VS. BT	9%	40%	49%
MEDIAN VS. BT	9%	41%	55%

Table 1: Disagreement between aggregation mechanisms. The first column shows the percentage of system pairs ordered differently by two aggregation mechanisms. The second column shows the percentage of setups where two aggregation mechanisms find different SotA, and the third column shows the percentage of setups where the top-3 systems are different (compared as sets).

such as BERTScore and MOVERScore seem less affected. On the other hand, BLEU variants are the most impacted, particularly when comparing MEAN or MEDIAN against BT. The disagreement between MEAN and MEDIAN is stable across metrics. In general, MEAN and MEDIAN are more in agreement with one another than they are with BT, which indicates that pairing issues have a stronger effect than outliers.

Which tasks are impacted most? Fig. 4(b) summarizes an analysis as above, but across tasks instead of metrics. Again, to control for the fact that some tasks may have larger datasets, we subsample uniformly from various test set sizes. The results are averaged over evaluation metrics. Machine translation and summarization suffer the least while dialogue and image captioning display larger disagreement between aggregation mechanisms. This suggests important future research directions to improve the evaluation setups in these tasks.

Importance of dataset size. In Fig. 4(c), we report disagreement across test set sizes, while averaging over datasets and evaluation metrics. It is reassuring to observe that with larger test sets, the different mechanisms tend to agree more, such that it matters less which one is actually chosen. However, for MEAN vs. BT and MEDIAN vs. BT, the disagreement does not continue to decrease below 10% with more test instances. For MEAN and BT the disagreement is lower but exhibits the same behavior, never falling below a certain threshold.

Different perspectives on uncertainty. In standard evaluation setups, not only system scores are reported but also whether the differences are statistically significant (Dror et al., 2018). Therefore, we ask how often differences that are statistically significant for one test are also statistically significant for another. The details of this experiments are presented in Appendix D and show, perhaps unsurprisingly, different behavior for different tests. In particular, the paired t-test is the one that most often finds differences to be significant (for 41% of pairs); Mood's test, an unpaired test to compare medians, finds significance for only 21% of pairs; and the sign test and Wilcoxon's sign-rank test (related to BT) are in between (for 35% and 40% of the pairs, respectively).

Sources of disagreement. Based on the analysis of Sec. 3, we know that the difference between MEAN and MEDIAN is due to the presence of statistical outliers, while the difference between MEDIAN and BT is due to the presence of different test instance types (Fig. 3). With real NLP datasets, in Fig. 4, we observe some discrepancy between MEAN and MEDIAN, indicating the presence of outliers. There is even more disagreement between MEDIAN and BT, indicating the presence of different types of test instances, as illustrated in Fig. 3.

6 Related work

Several studies have made a critical assessment of the standard evaluation methodologies. For example, Freitag et al. (2020) demonstrate the advantages of carefully choosing which references to use for NLG evaluation. Mathur et al. (2020) show that outliers matter in practice. Recently, Graham et al. (2020) draws attention on test set size. Several works have emphasized the importance of careful statistical testing (Rankel et al., 2011; Owczarzak et al., 2012; Graham, 2015; Dror et al., 2018). They recommend *paired* statistical tests. Finally, Novikova et al. (2018) report that "relative rankings yield more discriminative results than absolute assessments", which further motivates aggregation mechanisms like BT.

Aggregations. Pairwise comparison mechanisms date back to Thurstone (1927). Subsequently, the Bradley-Terry (BT) model has become a standard pairwise comparison model (Bradley and Terry, 1952). In NLP, BT-inspired mechanisms have sometimes been used to aggregate human assessments. For instance, Deriu et al. (2020) ranked chatbots regarding their ability to mimic conversational behavior of humans. Item response theory (IRT) has a similar formulation as BT, but also estimates the difficulty of each test instances using a latent-variable Bayesian model (Dras, 2015).

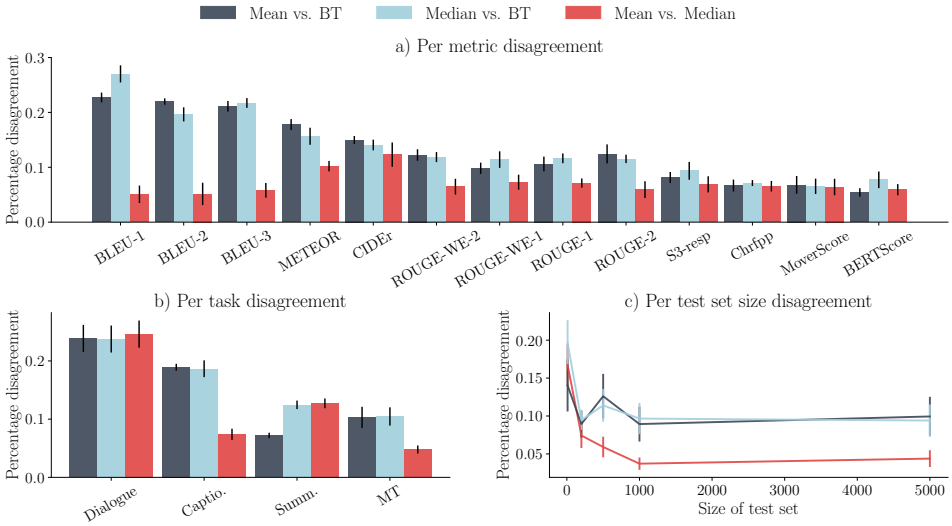


Figure 4: This figure measures the percentage of disagreement between each pair of aggregation mechanisms across different dimensions with real evaluation setups. Fig. 4(a) shows the disagreement per evaluation metric averaged over tasks and uniformly subsampled test set sizes, Fig. 4(b) shows the disagreement per task averaged over evaluation metrics and uniformly subsampled test set sizes, and Fig. 4(c) shows the disagreement across test set sizes averaged over tasks and evaluation metrics.

IRT has been applied to perform dataset filtering (Lalor et al., 2016, 2019), evaluate chatbots from human assessments (Sedoc and Ungar, 2020), and aggregate human assessments in machine translation (Dras, 2015). Elo (Elo, 1978) and TrueSkill (Herbrich et al., 2007) are famous extensions of the BT model commonly used to rate players in the context of gaming or sports events. Elo views player strengths as normally distributed random variables. TrueSkill is a Bayesian variant of Elo. Since 2015, the Workshop on Machine Translation (WMT) has been using TrueSkill to rank models based on human assessments following the methodology of Sakaguchi et al. (2014). We provide a detailed presentation and comparison of BT, Elo, and TrueSkill in Appendix G, and make both Elo and TrueSkill available as alternatives to BT in the released tool. The arguments in favor of BT made in this work transfer to its variants, including IRT, Elo, and TrueSkill, and the conclusions drawn from the experiments of Sec. 5 still hold when replacing BT by Elo or TrueSkill (Appendix G). Our work extends previous works that has considered BT variants by analyzing the potential causes for disagreement with MEAN and MEDIAN and by measuring the disagreement in real NLP evaluation setups.

7 Discussion

We briefly discuss some possible questions raised by the use of BT-like metrics, with more details provided in Appendix E, F, G, and H.

Extension to other evaluation setups. The experiments of Sec. 5 focus on reference-based NLG evaluation metrics. However, the arguments laid out throughout the paper apply beyond this setup. Any comparison of systems based on score aggregation is susceptible to suffer from outliers and complex pairing structures (e.g., Fig. 2). Future work should replicate our experimental setup for reference-free NLG (Zhao et al., 2020), classification, or regression tasks.

Type imbalance. Imagine a test set with a majority of easy instances and few hard ones. A system *A* could perform slightly worse than *B* on easy instances but much better on hard ones and will be declared worse by BT. If one views this decision as problematic then one should probably acknowledge that the test set is not representative of what should be measured. If hard instances matter more there should be a majority of them in the test set. Hoping that MEAN will be swayed to output the *intuitive* ordering of systems from a minority of test instances is a peculiar expectation to have about the evaluation setup. To diagnose such pathological cases, our tool, *Pairformance,* offers the possibility to view pairwise plots (as in Fig. 2) and histograms

of score differences. More generally, better aggregation mechanisms such as BT do not solve all potential problems of evaluation methodologies. Other aspects (such as choosing evaluation metrics or meaningful, representative, and large test sets) are all independent of the choice of aggregation mechanism, but also critical to the quality of the evaluation.

Transitivity. BT is not computed independently for each system, and it can happen that adding or removing a baseline impacts the scores of other systems. We explain this phenomenon in Appendix F and show that it is rarely a problem in real data. More generally, we discuss the connection with Arrow's impossibility theorem in the context of the aggregation of social preferences (Arrow, 1950). The *Pairformance* tool gets around this difficulty by offering the possibility of analyzing each pair of systems independently.

Relaxing assumptions. BT assumes that the relative strengths of systems remain constant across test instances. This might not always be true, especially when some systems are crafted for some specific kind of instances but perform badly on others. In such cases, BT still produces meaningful and easily interpretable results but fails to capture the latent structure of system strengths. Several refinements of BT are possible; e.g., item response theory extends BT by modeling instance difficulty, and Elo and TrueSkill allow system strengths to be stochastic and vary across instances. These refinements come at the cost of introducing new parameters, and it remains unclear how to choose these parameters in practice. Future work should investigate systematic ways to choose these parameters.

Tool description. We release *Pairformance,* a tool for performing full diagnostic analyses based on an evaluation dataframe made of the evaluation scores of systems and baselines. It can perform the analysis based on MEAN, MEDIAN, BT, Elo, and TrueSkill. For each aggregation technique, it outputs a full pairwise analysis of all pairs of systems. For MEAN and MEDIAN it compares score differences for pairs of systems. For BT, Elo, and TrueSkill, it estimates the probability that one system is better than another. All analysis is accompanied by appropriate statistical testing. See Fig. 5 for an example based on the BT mechanism. Furthermore, the tool can plot the histogram of paired differences $X_A^{(l)} - X_B^{(l)}$, allowing for the direct identification of pathological patterns such as those discussed above.

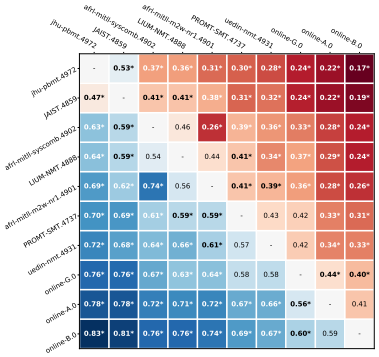


Figure 5: Pairwise system comparison with BT for machine translation with ROUGE-1, as output by the *Pairformance* tool released as part of this work.

8 Conclusion

We performed a critical assessment of the standard NLP evaluation methodology based on averaged scores, which ignores the natural instance-level pairing of evaluation scores when comparing systems. We showed the importance of the pairing and demonstrated the advantages of paired mechanisms such as Bradley–Terry (BT) over more standard aggregation schemes such as the mean or median. The choice of aggregation mechanism matters in real evaluation setups, and we therefore recommend BT as a robust aggregation mechanism. To facilitate adoption, we release *Pairformance,* a new tool to perform full analyses of system scores using BT and two of its variants, Elo and TrueSkill.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions, which greatly improved the final version of the paper. With support from Swiss National Science Foundation (grant 200021_185043), European Union (TAILOR, grant 952215), and gifts from Google, Facebook, Microsoft.

References

Kenneth J. Arrow. 1950. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine*

Translation, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3971–3984, Online. Association for Computational Linguistics.

George W. Divine, H. James Norton, Anna E. Barón, and Elizabeth Juarez-Colunga. 2018. The wilcoxon–mann–whitney procedure fails as a test of medians. *The American Statistician*, 72(3):278–286.

Mark Dras. 2015. Squibs: Evaluating human pairwise preference judgments. *Computational Linguistics*, 41(2):309–317.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Arpad E. Elo. 1978. *The rating of chessplayers, past and present*. Arco Publishing.

Ronald A. Fisher. 1935. *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill™: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19, pages 569–576. MIT Press.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, pages 1693–1701. Curran Associates, Inc.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259, Hong Kong, China. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 463–470, New York City, USA. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume*

2: Shared Task Papers, Day 1), pages 62–90, Florence, Italy". Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*.

Maja Popovic. 2017. chrF++: Words Helping Character n-grams. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 612–618.

Peter Rankel, John Conroy, Eric Slud, and Dianne O'Leary. 2011. Ranking human and machine summarization systems. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 467–473, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

João Sedoc and Lyle Ungar. 2020. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, Online. Association for Computational Linguistics.

Louis Leon Thurstone. 1927. A law of comparative judgement. *Psychological Review*, 34:278–286.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulleting*, 6:80–83.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

	type 1	type 2	type 3	type 4	type 5
S	23	50	40	70	60
B	28	45	30	65	50

Table 2: Example of two systems S and B with their strengths $\lambda_{t_i,S}$ and $\lambda_{t_i,B}, i \in [1,5]$ associated to each type of test instances. types.

A Reproducibility

In this section, we give additional details to ensure the reproducibility of our experiments. Furthermore, the code and data to reproduce each figure and table of the main paper is available at: https://github.com/epfl-dlab/BT-eval.

A.1 Pairing examples

It is straightforward to generate examples where the marginal distribution of the evaluation scores of two systems remain unchanged even when the pairing varies.

To do so, one can define k types of test instances. For each type t_i, each system has a probability distribution of scores for this type: $\mathcal{N}(\lambda_{t_i,S}, 1)$. So for instances of type t_i, the system S has score $\lambda_{t_i,S}$ in expectation with a variance of $\sigma^2 = 1$. Similarly, another system B can have different $\lambda_{t_i,B}$ parameters. An example is given in Table 2.

Now, observe that permuting the columns of S without changing the row B leaves the marginal distribution of S and B unchanged but changes the pairing. Then, one can simply iterate over all permutations of the row S to obtain many different pairings with fixed marginal distributions.

A.2 Simulation

We discuss the synthetic data and experiments depicted in Fig. 3.

To introduce pairing issues, we create a variable number of test instance types: N_{types}. For each test type, each system has a different distribution of scores. On test type t_i, the system s_j has a normal distribution of scores: $\mathcal{N}(\lambda_{i,j}, \sigma^2)$, where we fix $\sigma^2 = 1$ throughout our experiments. For each system, the $\lambda_{i,j}$ are sampled uniformly from $[0,1]$. Depending on the values of $\lambda_{i,j}$, the score distribution of system s_j can become multimodal. When, there is only one test type, the score of each system s_j is a normal $\mathcal{N}(\lambda_j, \sigma^2)$. In that case, the pairing can be ignored and MEAN and MEDIAN are expected to work well.

For outliers, we define f as the fraction of test instances on which systems' scores are not drawn from their distribution scores. For such instances, we first draw the scores for each systems according to their distribution and then perform a random permutation, so that each system receives a score that is not sampled from its score distribution.

Then, we vary the number of systems present in the evaluation N_{sys} and the number of test instances M. Each choice of N_{types}, f, N_{sys}, and M gives a dataframe corresponding to an evaluation setup on which we can compare MEAN, MEDIAN, and BT against the *true* latent strengths of systems $\lambda_{i,j}$. The evaluation and the y-axis in Fig. 3 is then the Kendall's τ between the ordering resulting from MEAN, MEDIAN, or BT against the ordering resulting from the $\lambda_{i,j}$.

We consider the following variations for the parameters of the experiments:

- $N_{types} \in \{1,3,5,10\}$,
- $f \in \{0.,0.01,0.025\}$,
- $N_{sys} \in \{2,3,5,10,25,50\}$,
- $M \in \{10,30,100,200\}$.

In total, we have: $4 \cdot 3 \cdot 6 \cdot 4 = 288$ parameter choices. For each we sample 10 datasets resulting in $2,880$ synthetic evaluation setups.

A.3 Real data

Each of the dataset we use contains the evaluation results of a varying number of systems for a varying number of evaluation metrics:

Summarization: CNN/DM (Hermann et al., 2015): 11,432 test instances, 12 summarization systems, and 13 evaluation metrics. TAC-08: 48 test instances, 58 summarization systems, and 13 evaluation metrics. TAC-09: 44 test instances, 55 summarization systems, and 13 evaluation metrics. TAC-11: 44 test instances, 50 summarization systems, and 13 evaluation metrics. **Captioning**: MSCOCO (Lin et al., 2014): 40,504 test instances, 12 systems, and 7 evaluation metrics. **Dialogue**: Topical-Chat (Mehri and Eskenazi, 2020): 60 test instances, 5 systems, and 13 evaluation metrics. Persona-Chat (Mehri and Eskenazi, 2020): 60 test instances, 4 systems, and 13 evaluation metrics. **MT**: WMT-17 (Bojar et al., 2017): evaluated with 11 evaluation metrics, we have the following pairs: *lv-en* (2,001 instances, 9 systems), *de-en* (3,004 instances, 11 systems), *ru-en* (3,001 instances, 9 systems), *tr-en* (3,007 instances, 10 systems), and

zh-en (2,001 instances, 16 systems). WMT-18 (Ma et al., 2018): evaluated with 13 evaluation metrics we have the following pairs: *de-en* (2,998 instances, 16 systems), *et-en* (2,000 instances, 14 systems), *fi-en* (3,000 instances, 9 systems), *ru-en* (3,000 instances, 8 systems), and *zh-en* (3,981 instances, 14 systems). WMT-19 (Ma et al., 2019): evaluated with 13 evaluation metrics we have the following pairs: *de-en* (2,000 instances, 16 systems), *fi-en* (1,996 instances, 12 systems), *gu-en* (1,016 instances, 12 systems), *kk-en* (1,000 instances, 11 systems), *lt-en* (1,000 instances, 11 systems), *ru-en* (2,000 instances, 14 systems), and *zh-en* (2,000 instances, 15 systems).

The evaluation metrics considered are: BLEU-[1,2,3,4] (Papineni et al., 2002), ROUGE-[1,2,L] (Lin, 2004), ROUGE-WE-[1,2] (Ng and Abrecht, 2015), JS-[1,2] (Lin et al., 2006), S3-[pyr, resp] (Peyrard et al., 2017), CIDEr (Vedantam et al., 2015), Chrfpp (Popovic, 2017), METEOR (Lavie and Agarwal, 2007), MoverScore (Zhao et al., 2019), and BERTScore (Zhang et al., 2020). This is a total of 18 metrics.

Sub-sampling test set sizes. In experiments reported by Fig. 4 the results are averaged after resampling test sets of different sizes. The test set sizes used are: $[10, 50, 100, 500, 1000, 5000]$. Results broken down per dataset and per metric that does not need resampling of test set sizes is proposed in Appendix C.

A.4 Implementations

We implement BT with `scipy.org` and numpy. For the statistical tests, we use the default implementation from `scipy.org`. For Elo, we implement a wrapper around existing code: `https://github.com/ddm7018/Elo`. Similarly, for TrueSkill, we implement a wrapper around existing code: `https://pypi.org/project/trueskill/`.

B Proof of Proposition 1

Proof. We observe that the case of the MEAN and the MEDIAN are direct by definition.

$M_{A-B} > 0$ is equivalent to saying that for more than 50% of instances, $X_A^{(l)} > X_B^{(l)}$, i.e., A is better than B on more than 50% of instances. On the other hand, BT correctly gives A better than $B \iff \mathbb{P}(A > B) > \mathbb{P}(B > A) \iff \mathbb{P}(A > B) > \frac{1}{2}$, i.e., A is better than B on more than 50% of instances. So, BT is consistent $\iff$ A is better than B on more than 50% of instances $\iff$ $M_{A-B} > 0$. □

C Disagreement breakdown

Compared to experiments in the main paper, we provide a more detailed breakdown of the disagreement in Table 3.

D Different view on uncertainty

As argued in the main paper (Sec. 3.2), the choice of aggregation mechanism bears strong similarities with the choice of statistical test. Thus, we measure in how many setups difference between systems that are statistically significant according to one test are also significant according to another.

We compare: paired t-test (usually to compare means), the Mood's median test, and the sign test (consistent with BT). We also add the Wilcoxon sign-rank test as it was often recommended by previous work (Owczarzak et al., 2012; Dror et al., 2018).

In Fig. 6, we plot the frequency with which test j yields a significant difference among the pairs of systems for which the test i has already yielded a significant difference. The diagonal depicts the overall percentage of pairs of systems for which the test finds a significant difference. Note that the matrix is not symmetric.

Interestingly, when the Mood's median test says the difference between two system is significant, 98% of the times it is also the case for the paired t-test and 89% of the times it is also the case for the Sign test. So the Mood's median is the most restrictive, finding less often significant difference than the other two. In comparison, the Sign test and the Wilcoxon's sign-rank test find significant differences between systems much more frequently. In general, the paired t-test is the one finding differences the most frequently.

E Details about the Bradley–Terry model

Given a pair of systems S_i and S_j, the Bradley–Terry model estimates the probability $p_{i,j}$ that the system S_i is better than the system S_j based on their relative strengths: $\frac{\lambda_i}{\lambda_i+\lambda_j}$.

BT estimates these parameters λ_i for each of the n systems from the observed results of evaluation. We denote as $\omega_{i,j}$ the number of instances for which S_i scores higher than S_j. Note that, in our setup, there is one comparison per test instance. In the main paper, we said that the solutions for $\hat{\lambda}$ are found in closed-form for $n = 2$. When the number of systems is greater than 2, the parameters are

		BLEU			ROUGE			ROUGE-WE			MoverScore			BERTScore		
		Mean/BT	Med/BT	Mean/Med	Mean/BT	Med/BT	Mean/Med	Mean/BT	Med/BT	Mean/Med	Mean/BT	Med/BT	Mean/Med	Mean/BT	Med/BT	Mean/Med
TAC08	Disagree.	.09	.13	.15	.07	.13	.14	.12	.06	.13	.05	.11	.12	.05	.11	.12
	≠ SotA	.43	.73	.47	.33	.52	.47	.58	.20	.47	.10	.50	.47	.13	.17	.27
	≠ Top3	.73	.77	.77	.61	.80	.81	.87	.65	.80	.43	.73	.70	.60	.93	.87
TAC09	Disagree.	.08	.13	.13	.08	.16	.16	.07	.15	.16	.06	.14	.13	.06	.12	.12
	≠ SotA	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	≠ Top3	.70	.70	.70	.63	.87	.82	.48	.73	.75	.33	.70	.70	.43	.73	.67
TAC11	Disagree.	.07	.12	.12	.06	.13	.12	.05	.13	.12	.04	.11	.10	.04	.11	.10
	≠ SotA	.37	.67	.50	.42	.64	.61	.33	.67	.65	.40	.63	.63	.27	.73	.63
	≠ Top3	.73	.87	.83	.58	.88	.87	.60	.93	.92	.57	.87	.80	.43	.87	.83
CNN/DM	Disagree.	.14	.17	.12	.08	.07	.02	.06	.05	.02	.07	.06	.08	.08	.08	.04
	≠ SotA	.53	.80	.83	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	≠ Top3	.97	.97	.90	.73	.49	.24	.90	.42	.48	.00	.00	.00	.90	.90	.06
WMT17	Disagree.	.07	.08	.05	.07	.07	.04	.07	.08	.04	.03	.04	.03	.03	.04	.03
	≠ SotA	.17	.19	.14	.28	.42	.23	.35	.40	.19	.22	.15	.24	.15	.22	.24
	≠ Top3	.43	.57	.40	.56	.63	.29	.57	.67	.40	.26	.37	.37	.23	.27	.33
WMT18	Disagree.	.09	.09	.03	.11	.11	.04	.12	.12	.04	.06	.06	.04	.06	.06	.03
	≠ SotA	.67	.63	.24	.55	.65	.26	.61	.67	.66	.47	.49	.18	.43	.47	.31
	≠ Top3	.77	.74	.25	.56	.69	.39	.66	77	.40	.57	.58	.33	.57	.58	.19
WMT19	Disagree.	.07	.08	.04	.10	.11	.04	.11	.11	.05	.05	.04	.05	.04	.04	.05
	≠ SotA	.32	.36	.25	.44	.45	.18	.46	.48	.16	.32	.25	.33	.31	.17	.35
	≠ Top3	.54	.42	.30	.48	.54	.30	.51	.54	.33	.54	.41	.46	.39	.26	.39
TC	Disagree.	.26	.22	.34	.24	.19	.24	.27	.28	.22	.28	.19	.29	.18	.24	.20
	≠ SotA	.53	.43	.66	.52	.46	.40	.53	.63	.45	.63	.33	.53	.30	.40	.27
	≠ Top3	.57	.60	.63	.57	.56	.60	.62	.55	.47	.63	.60	.60	.53	.57	.57
PC	Disagree.	.28	.24	.32	.25	.23	.22	.21	.22	.22	.12	.20	.19	.13	.12	.13
	≠ SotA	.50	.50	.63	.42	.53	.43	.28	.33	.30	.33	.47	.50	.30	.37	.43
	≠ Top3	.33	.33	.43	.42	.60	.55	.37	.72	.63	.23	.30	.27	.27	.20	.07
MSCOCO	Disagree.	.20	.18	.12	.18	.14	.03	-	-	-	-	-	-	-	-	-
	≠ SotA	1.0	1.0	.00	.03	.03	.00	-	-	-	-	-	-	-	-	-
	≠ Top3	1.0	1.0	.17	1.0	1.0	.47	-	-	-	-	-	-	-	-	-

Table 3: Disagreement between aggregation mechanisms per dataset and per metric.

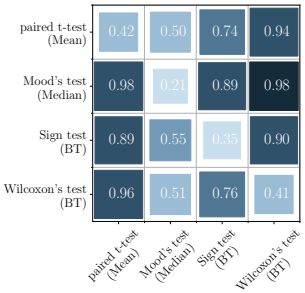


Figure 6: In this matrix, the cell in row i and column j indicates the frequency with which the test j finds a difference significant among the pairs of systems for which the test i has found the difference significant. For example, when the Mood's median test finds a significant difference between a pair, 98% of the times, the paired t-test also finds the difference significant.

found by an iterative optimization algorithm that maximizes the following log-likelihood:

$$\mathscr{L}(\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_{i,j}\log(\lambda_i) - \omega_{i,j}\log(\lambda_i+\lambda_j), \tag{2}$$

where $\lambda = [\lambda_1, \ldots, \lambda_n]$.

Denote W_i as the number of comparison in which system i is better: $W_i = \sum_j \omega_{i,j}$. Then, the algorithm iteratively performs the following two updates (at step t):

$$\hat{\lambda}_i = W_i \left(\sum_{i \neq j} \frac{\omega_{i,j} + \omega_{j,i}}{\lambda_i^{(t)} + \lambda_j^{(t)}} \right)^{-1}, \forall i, \tag{3}$$

$$\lambda_i^{(t+1)} = \frac{\hat{\lambda}_i}{\sum_k \hat{\lambda}_k}, \forall i. \tag{4}$$

It can be shown that starting from a random λ this algorithm improves the log-likelihood at every iteration and converges to a unique maximum.

For the practical implementation, only a threshold ϵ defining when to stop has to be decided. We choose to stop iterating when at step t, if the new vector of parameter λ remains close to the previous one: $\|\lambda^{(t+1)} - \lambda^{(t)}\|^2 < \epsilon$. Throughout our experiments, we always set $\epsilon = 1 \cdot 10^{-9}$.

F Transitivity with BT and Arrow's theorem

One possibly counter-intuitive behaviour of BT is that adding or removing a baseline can impact the scores and ordering of other systems. For example, consider two systems A and B with the following scores: $\mathscr{M}_A = [1,2,3]$ and $\mathscr{M}_B = [2,3,1]$. Then, BT identifies system B has better with a relative strengths of $\frac{2}{3}$. Now suppose another system C is added with scores $\mathscr{M}_C = [3,2,1]$, running BT on these 3 systems together gives the result that all systems have an equal strength, so now B is not seen as better than A.

We search for triple of systems which exhibit this pattern in our data and couldn't find any as long as we use more than 10 test instance.

Can we hope to fix this weakness? Arrow's impossibility theorem says no (Arrow, 1950). Our setup matches very well the problem of aggregating social preferences from voters. In this context, Arrow (1950) proved that no aggregation mechanism with more than 2 voters and 3 possibilities can simulataneously meet the 3 following criterion: (i) monotonicity: if every voter prefers X over Y, then the aggregation ranks X above Y, (ii) (IAA) the aggregated preference between X and Y should remain unchanged if voter preferences between other pairs change, and (iii) no dictators: the outcome is not decided by a single voter. In our framework, voters are test instance and preferences are given by the evaluation metrics. BT can fail on the second criteria, and MEAN and MEDIAN can be dictatorial (as seen in the paper). A way around this problem is to remain with pairwise comparisons of systems $n < 3$ and use BT. In that case, there is no possibility for BT to fail on IIA.

G Variants of BT: Elo and TrueSkill

BT has been extended in various ways. We discuss here two important variants that we incorporate in our analysis tool: Elo and TrueSkill.

G.1 Elo ratings

The Elo rating (Elo, 1978) is variant of the BT with an online update rule, i.e., the rating of systems (players) is updated as new test instances (new games) arrive. As BT, Elo computes the probability that systems S_i beats system S_j. Now, the t-th test instance arrives and system S_i receives the score s_i and system S_j receives the score s_j. We update the rating R based on this observed difference $\delta_{i,j}$:

$$R_k^{(t+1)} = R^{(t)} + K\left(\delta_{i,j} - \frac{Q_i}{Q_i + Q_j}\right), \qquad (5)$$

where K is parameter that has to be chosen, R the rating of some system, and Q plays a role analogous to λ_k in BT. K controls how much each new instance can change the ratings. It can be shown that, implicitly, Elo corresponds to a version of BT where the strength of systems is represented by a normal distribution: $\lambda_i + \epsilon_i,\ \ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$, with a variance σ^2 shared by all players (Elo, 1978). In our implementation, we provide the user with the ability to choose K and set it to 20 by default.

	Disagree.	≠ SotA	≠ Top-3
MEAN vs. MEDIAN	4%	18%	30%
MEAN vs. BT	9%	40%	49%
MEDIAN vs. BT	9%	41%	55%
MEAN vs. Elo	20%	55%	84%
MEDIAN vs. Elo	19%	56%	84%
MEAN vs. TrueSkill	18%	44%	76%
MEDIAN vs. TrueSkill	17%	46%	79%
BT vs. Elo	16%	38%	75%
BT vs. TrueSkill	18%	53%	72%
Elo vs. TrueSkill	18%	45%	71%

Table 4: Global disagreement (as in Table 1) between aggregation mechanisms repeated with Elo and TrueSkill.

G.2 TrueSkill

TrueSkill (Herbrich et al., 2007) is Bayesian variant of the Elo rating system. It also updates the ratings of systems online, i.e., ratings change as new test instances arrive. Now, the strength of a system S_i is represented by a normal distribution, $\mathcal{N}(\lambda_i, \sigma_i^2)$. In contrast to Elo, each player has its own variance. The update follows Bayes rule, but is intractable in general, so message passing approximation are often employed.

H Comparison of Elo, TrueSkill, and BT

We repeat the experiments of Table 1 from the main paper by replacing BT with Elo and TrueSkill with their default parameters. The results are shown in Table 4. With Elo and TrueSkill, the same conclusions from the main paper hold, i.e., paired aggregation mechanisms exhibit significant disagreement with MEAN and MEDIAN. Some discrepancies between BT, Elo, and TrueSkill remain which calls for further investigations about which one to choose.

	de-zh	ro-en	et-en	ru-de
source.ap-scores	9%	14%	7%	3%
source.auc-scores	3%	7%	9%	0%
source.rec-topk-scores	38%	20%	23%	14%
target.ap-scores	4%	10%	2%	2%
target.auc-scores	2%	8%	4%	0%
target.rec-topk-scores	31%	19%	13%	13%

Table 11.1: Disagreement of system rankings between *mean* and BT across six evaluation metrics and four language pairs. Each cell shows the percent of system pairs ordered differently by *mean* and BT according to the recalled version of Kendall's τ supported on $[0, 1]$. Higher scores indicate higher disagreement.

11.8 Appendix: Application to Eval4NLP Shared Task

In the Eval4NLP shared task (Fomicheva et al., 2021a), systems are ranked according to their global independent statistics, e.g., *mean* AUC scores of different systems over a common set of test instances. However, aggregation mechanisms such as the *mean* ignores which system beats others over individual instances, and thus may lead to false conclusions. Here, we adopt BT to conduct rigorous comparison for competing systems. Recall that BT leverages instance-level pairing of metric scores from different systems, and assumes that a winning system should beat others over the majority of instances. In the concrete case – the shared task – this would mean that a system could have very high AUC scores on few instances, which inflate its *mean* AUC, but otherwise performs worse in the majority of instances.

We analyze whether *mean* and BT yield similar results on the shared task. First, we quantify the disagreement between *mean* and BT. Table 11.1 shows that *mean* and BT often disagree on the ranking of systems, especially for "source.rec-topk-scores" and "target.rec-topk-scores". This might undermine the reliability of these recall-based metrics, as they are very sensitive to the aggregation scheme (BT vs. *mean*), unlike 'ap-scores' and 'auc-scores' that consider both precision and recall.

We then provide justifications to understand the judgments of BT and *mean* on German-Chinese as use case (see Figure 11.1). We find that BT and *mean* both may yield wrong judgments as to which system is the state-of-the-art. We illustrate this below:

- *mean* might be wrong (Fig. 11.1, top): Considering plausibility of explanations on source sentences, *mean* declares Kyoto-1 as the best system; however, it significantly outperforms merely 3 out of 9 systems according to pairwise comparison. This indicates that *MEAN* results are very likely wrong. In contrast, BT chooses Unbabel-18 according to that it wins in 8 out of 9 cases.

- BT might be wrong (Fig. 11.1, bottom): Considering plausibility of explana-

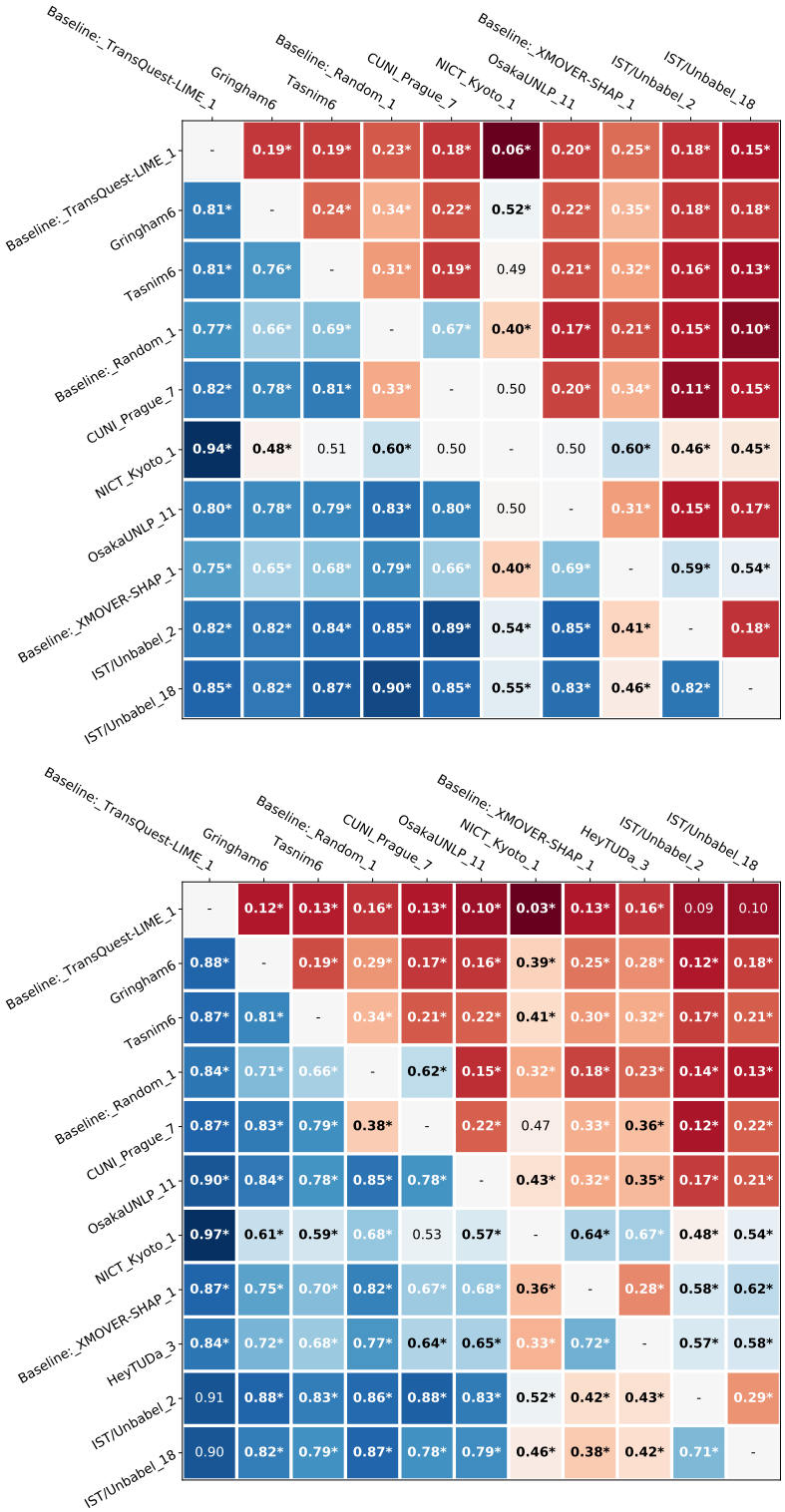


Figure 11.1: Results of pairwise comparison according to (top) "source.rec-topk-scores" and (bottom) "target.rec-topk-scores" over system pairs for German-Chinese. Each cell denotes the percent of the instances in which one system (in rows) beats another (in columns). We mark cells for which system pairs have significant differences according to Sign test with '*'. Systems have been ranked reversely by *BT*, e.g., systems in final rows are declared the best. *mean* declares Kyoto-1 as the best in both (top) and (bottom) settings.

tions on target sentences, BT declares Unbabel-18 as the best, as it beats 7 out of 10 systems with clear wins. On the other hand, Kyoto-1 (ranked 5th according to BT) wins in 9 out of 10 systems, and it also beats Unbabel-18. This means Kyoto-1 might be the winner, but that BT nevertheless favors Unbabel-18 most as BT considers the number of instances of one system superior to another. Concretely, though Kyoto-1 beats the greatest number of systems, it outperforms these systems on slightly over half of instances, which reflects the weak strength from a BT perspective. In contrast, Unbabel-18 wins globally on the greatest number of instance-level pairs assembled across all systems. We depict this issue of BT as the inconsistency between global and local judgments, i.e., that one locally beats another in the case of two systems, but the judgment of system superiority may change in the global view when involving more systems in the comparison. Indeed, the 'inconsistency' can hardly be addressed according to the Arrow's impossibility theorem (Arrow, 1950).

Our analysis shows how subtle the evaluation of systems in the case of the shared task can be and that there is no clear winner, as none of the systems beats all other systems according to pairwise comparison.

Subpart D

Explainability for Evaluation Metrics

Chapter 12

Global Explainability of BERT-Based Evaluation Metrics by Disentangling along Linguistic Factors

Global Explainability of BERT-Based Evaluation Metrics by Disentangling along Linguistic Factors

Marvin Kaster, Wei Zhao, Steffen Eger
Natural Language Learning Group (NLLG)
Technische Universität Darmstadt, Germany
marvin.kaster@stud.tu-darmstadt.de
{zhao,eger}@aiphes.tu-darmstadt.de

Abstract

Evaluation metrics are a key ingredient for progress of text generation systems. In recent years, several BERT-based evaluation metrics have been proposed (including BERTScore, MoverScore, BLEURT, etc.) which correlate much better with human assessment of text generation quality than BLEU or ROUGE, invented two decades ago. However, little is known what these metrics, which are based on black-box language model representations, actually capture (it is typically assumed they model semantic similarity). In this work, we use a simple regression based global explainability technique to disentangle metric scores along linguistic factors, including semantics, syntax, morphology, and lexical overlap. We show that the different metrics capture all aspects to some degree, but that they are all substantially sensitive to lexical overlap, just like BLEU and ROUGE. This exposes limitations of these novelly proposed metrics, which we also highlight in an adversarial test scenario.

1 Introduction

Evaluation metrics are a key ingredient in assessing the quality of text generation systems, be it machine translation, summarization, or conversational AI models. Traditional evaluation metrics in machine translation and summarization, BLEU and ROUGE (Papineni et al., 2002a; Lin, 2004), have measured lexical n-gram overlap between system prediction and a human reference. While simple and easy to understand, early on, limitations of such lexical overlap metrics have been recognized (Callison-Burch et al., 2006), e.g., in that they can only measure surface level similarity, and they are especially inadequate when it comes to assessing current high-quality text generation systems (Rei et al., 2020; Mathur et al., 2020; Marie et al., 2021).

Recently, a class of novel evaluation metrics based on BERT and its variants has been explored that correlates much better with human assessments of translation quality. For example, BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), BLEURT (Sellam et al., 2020), XMoverScore (Zhao et al., 2020), and COMET (Rei et al., 2020) all use large-scale pretrained language models, but differ in whether they compare hypotheses to references, to source texts, to both, on the one hand, and whether they use human scores for supervision or not, on the other. Since these models all leverage large-scale language models which have pushed the state-of-the-art in many areas of NLP, their success comes with little surprise.

To better understand these novel metrics based on black-box language representations is a prerequisite for identifying their limitations, e.g., to adversarial inputs. For example, if an evaluation metric is sensitive to lexical overlap, it can be fooled by using the same words but in different order.

In this work, we fill the existing 'explainability gap' and introspect linguistic properties encoded in BERT-based evaluation metrics. Although there is already considerable work on introspecting and understanding BERT (see Rogers et al. (2020) for an overview), e.g., via probing (Tenney et al., 2019), analyzes by Hewitt and Liang (2019); Eger et al. (2020); Ravichander et al. (2021) indicate that probing results (based on supervision) are not always trustworthy. More importantly, the modern evaluation metrics sketched above rely on at least two factors: BERT (or its variants) and different aggregation schemes, such as Earth Mover Distance (Kusner et al., 2015; Zhao et al., 2019) or greedy alignment (Zhang et al., 2020), on top of BERT. Understanding BERT alone is thus not sufficient for explaining BERT-based evaluation metrics.

Here, we present a simple global explanation technique of BERT-based evaluation metrics which disentangles them on prominent linguistic factors, *viz.*, syntax, semantics, morphology and lexical overlap. We find that all metrics capture these linguistic aspects to certain (but differing) degrees,

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8912–8925
November 7–11, 2021.

and they are particularly sensitive to lexical overlap, which makes them prone to similar adversarial fooling (cf. Li et al., 2020; Eger et al., 2019; Keller et al., 2021) as BLEU-based lexical overlap metrics. Overall, our contributions are:

- We disentangle a multitude of current BERT-based evaluation metrics on four linguistic factors using linear regression.
- We show that all metrics are sensitive to all factors and especially to lexical overlap, as confirmed both by the linear regression and an adversarial experiment.
- Based on the insight that different metrics capture different linguistic factors to varying degrees, we ensemble metrics and identify an average improvement of between 8 and 13% for the most heterogeneous metrics.

2 Related work

Our work concerns reference-based and reference-free metrics, on the one hand, and model introspection (or 'explainability'), on the other.

Evaluation metrics for Natural Language Generation In the last few years, several strong performing evaluation metrics have been proposed, the majority of which is based on BERT and similar high-quality text representations. They can be differentiated along two dimensions: (i) the input arguments they take, and (ii) whether they are supervised or unsupervised. *Reference-based* metrics compare human references to system predictions. Popular metrics are BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), and BLEURT (Sellam et al., 2020). In contrast, *reference-free* metrics directly compare source texts to system predictions, thus they are more resource-lean. Popular examples are XMoverScore (Zhao et al., 2020), Yisi-2 (Lo, 2019), KoBE (Gekhman et al., 2020), and SentSim (Song et al., 2021). Rei et al. (2020) use all three information signals: source text, hypotheses and human references. There are also reference-free metrics outside the field of machine translation; for example, SUPERT (Gao et al., 2020) for summarization. *Supervised* metrics train on human sentence-level scores, e.g., Direct Assessment (DA) scores or post-editing effort (HTER) for MT. These metrics include BLEURT and COMET (Rei et al., 2020). In MT, most metrics from the so-called 'Quality Estimation' (QE) tasks are also supervised reference-free metrics, e.g., TransQuest (Ranasinghe et al., 2020) and BERGAMOT-LATTE (Fomicheva et al., 2020b). *Unsupervised* metrics require no such supervisory signal (e.g., MoverScore, BERTScore, XMoverScore, SentSim).

Model introspection There has been a recent surge in interest in explaining deep learning models. The techniques for explainability differ in whether they provide justification or information for model outputs on individual instances (*local explainability*) or focus on a model as a whole and disclose its internal structure (*global explainability*) (Danilevsky et al., 2020). Popular examples for local explainability are LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) that find features from the input (such as particular words) relevant to model outputs.

Concerning (global) interpretability of text representations, previous works (Adi et al., 2017; Conneau et al., 2018) introspect the properties encoded in vector representations through probing classifiers—trained on external data to perform a certain linguistic task, such as inducing the dependency tree depth from a text representation (of a sentence). Tenney et al. (2019) extend this idea by inspecting BERT representations layer-by-layer, and find that BERT captures more semantic information in its higher layers and more syntactic and morphological information in its lower layers. However, probing results are not always trustworthy due to the sensitivity to probing design choices, e.g., data size and classifier choices (Eger et al., 2020), and data artefacts (Ravichander et al., 2021). More importantly, evaluation metrics use BERT differently: some are supervised and others are unsupervised, some fine-tune BERT on semantic similarity datasets, and they generally differ on how they aggregate and compare BERT representations. This means to understand these metrics it does not suffice to understand BERT alone.

In our work, we disentangle BERT based evaluation metrics along linguistic factors as a form of global explainability of those metrics. This yields insights into which linguistic information signals specific metrics use, in general, and may expose their limitations.

3 Our approach

In our scenario, we consider different metrics $\mathfrak{m}$ taking two arguments and assigning them a real-

valued score

$$\mathfrak{m} : (x, y) \mapsto s_{\mathfrak{m}} \in \mathbb{R}$$

where x and y are source text and hypothesis text, respectively (for so-called *reference-free* metrics) or alternatively x and y are reference and hypothesis text, respectively (for so-called *reference-based* metrics). The scores s that metrics assign to x and y can be considered the *similarity* between x and y or *adequacy* of y given x. In our experiments in Section 4, we will focus on machine translation (MT) as use case; it is arguably the most popular and prominent text generation task. Thus, x is either a (sentence-level) source text in one language and y the corresponding MT output, or x is the human reference for the original source text.

To better understand evaluation metrics, we decompose their scores $s_{\mathfrak{m}}$ along multiple linguistic factors. An example is outlined in Table 1.

We follow a long line of research in the applied sciences, and use a *linear model* to explain a *target variable* (in our case, the metric score), also called *response variable*, in terms of multiple *regressors*, also called *explanatory variables*. That is, we estimate the linear regression

$$\begin{aligned}\mathfrak{m}(x, y) = \alpha \cdot \mathit{sem}(x, y) + \beta \cdot \mathit{syn}(x, y) \\ + \gamma \cdot \mathit{lex}(x, y) + \delta \cdot \mathit{morph}(x, y) + \epsilon\end{aligned} \quad (1)$$

Here, $\text{sem}(x, y)$, $\text{syn}(x, y)$, $\text{morph}(x, y)$ and $\text{lex}(x, y)$ are scores which describe the semantic, syntactic, morphological, and lexical similarity between the two argument sentences. The real coefficients α, β, γ and δ are the regressors' weights, estimated from data. Finally, ϵ is an error term.

Linear regression assumes a linear relationship between the target variable and the regressors. It may fail when the relationship is non-linear, but its simple model structure provides interpretability: a larger positive coefficient means the respective regressor has higher positive impact on the target variable (fixing all other variables), a coefficient close to zero means no linear relationship, and a larger negative coefficient means an inverse linear relationship between regressor and target variable.

The coefficient of determination R^2 describes how well the regression model reflects the data. It is defined as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

where *SSE* denotes the sum of squared errors and *SST* denotes the sum of squared totals. They are defined as

$$\text{SSE} = \sum_i (y_i - \hat{y}_i)^2, \quad \text{SST} = \sum_i (y_i - \overline{y}_i)^2$$

respectively, where $\hat{y}_i$ is the prediction of the model, y_i is the true score, and $\overline{y}_i$ is the mean, $\overline{y}_i = \frac{1}{N}\sum_i^N y_i$. R^2 is 1 for a perfect fit, 0 if it always predicts the mean and negative if the model is worse than this baseline.

To ensure comparability of the different regressions, we normalize the scores of our regressors and the response variable with the z-normalization, i.e., subtracting the mean and dividing by the standard deviation, per variable. In the following, we define our regressors.

Semantic score (SEM) The semantic scores are provided by the datasets and were annotated by humans who rated e.g. the translation quality. See Section 4.2 for details.

Syntactic score (SYN) To measure the syntactic similarity of the argument sentences x and y, we compare their dependency trees. Both sentences are parsed by the Stanford dependency parser (Chen and Manning, 2014). Then, the tree edit distance (TED) (Bille, 2005) between the resulting trees is calculated. As an extension of string edit distance, TED measures how many operations are necessary to transform one tree into the other. Only the structures of the trees are considered in the calculation. We ignore the actual words.

We normalized the TED to ensure comparability between sentences with different lengths (Zhang and Shasha, 1989). The final score is calculated as

$$\mathit{syn}(x, y) = 1 - \frac{\mathit{TED}}{l_x + l_y}$$

where l_x and l_y are the lengths of the sentences. Figure 1 shows an example of the TED calculation for sentences in the same language (monolingual) and Figure 2 shows a cross-lingual example.

Lexical overlap score (LEX) We measure the lexical overlap between x and y by the BLEU score (Papineni et al., 2002b): BLEU_n calculates the precision based on how many n-grams of one sentence can be found in the other sentence. In the experiments below, we use BLEU_1. Using unigrams assures that word order is ignored. The simple precision count is modified so that identical words are only counted once.

Hypothesis (y)	**Reference/Source** (x)	**SEM**	**SYN**	**LEX**	**MOR**
It is a boy , likes to sport , but it cannot do it because of their very.	He is a boy, he likes sports but he can't take part because of his knee.	-1.57	0.98	-0.59	-0.87
Zwei Besatzungsmitglieder galten als vermisst.	Two crew members were regarded as missing.	0.83	0.99	0.46	-2.40

Table 1: Example setups with normalized semantic, syntactic, lexical overlap and morphological scores.

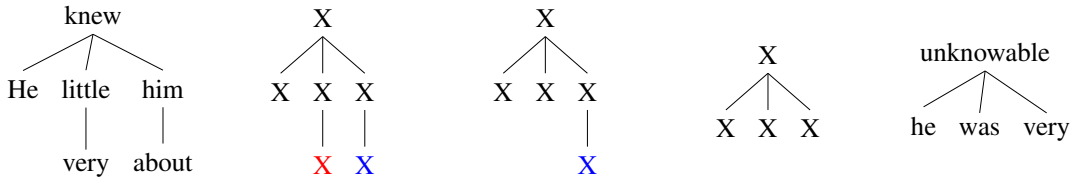


Figure 1: Monolingual tree edit distance example. The left-most tree is the first sentence of the sentence pair and the right-most tree is the second. To transform the left-most sentence into the right-most sentence, two leaves are removed. The unnormalized tree edit distance is thus 2. The normalized score is $1 - \frac{2}{6+4} = 0.8$.

For monolingual reference-based metrics, the BLEU score is calculated directly on the sentence pairs. To use BLEU for cross-lingual reference-free metrics, we choose to translate the non-English sentences into English via Google Translate, as it remains unclear how else to define lexical overlap between sentences from different languages. We compute BLEU scores on original English and translated sentence pairs.

Word 1	**Word 2**	**UD Tags**
reached	combined	Tense=Past VerbForm=Part
stay	sein	VerbForm=Inf

Table 2: Example of a morphological lexicon that contains word pairs with identical morphological UD tags, on which we finetune FastText word embeddings.

Morphological score (MOR) We introduce a morphological score $morph(x, y)$. To do so, we use static FastText word embeddings (Bojanowski et al., 2017) and increase morphological information in the original embeddings: (a) first, we produce two morphological lexicons based on words from WMT and STS, each containing word pairs with identical UD morphological tags (Nivre et al., 2020) (see Table 2). (b) Then, we finetune/retrofit the embeddings on the morphological lexicons using the method described in Faruqui et al. (2015), so that words with the same morphological tags have more similar representations.

The final morphological score for a sentence pair is the cosine similarity between two averaged sentence embeddings over refined word vectors of each sentence. Note that, while we refer to these embeddings as morphological, they actually capture multiple linguistic factors and can only be considered *more* morphological than standard static vector spaces.

If the overlap of morphological features between a language pair is very low, the morphological score will not be meaningful. We exclude the morphological score for such language pairs.

4 Experiments

We analyze different evaluation metrics by calculating their score for sentence pairs. We use both reference-based metrics, which operate in a monolingual space, and reference-free metrics, which operate in a cross-lingual space.

4.1 Metrics

Reference-based metrics We consider the following reference-based metrics.

- BERTScore (Zhang et al., 2020) aggregates and compares BERT embeddings by determining a greedy alignment between words in two sentences and summing up the cosine similarities of representations of aligned words.
- MoverScore (Zhao et al., 2019) computes an optimal alignment between words in the two sentences using word mover distance (Kusner et al., 2015) on top of BERT representations.
- Sentence BERT (SBERT) (Reimers and Gurevych, 2019) fine-tunes Siamese BERT networks on NLI data, and produces sentence embeddings by using pooling on top of BERT

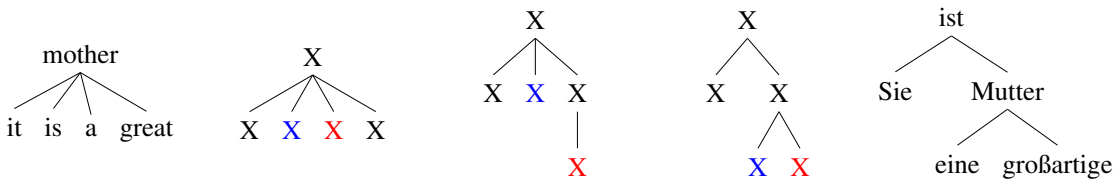


Figure 2: Cross-lingual tree edit distance example. To transform the left-most into the right-most sentence, two leaves (shown in red and blue) are moved to other locations. This takes 2 operations. The normalized score is $1 - \frac{2}{5+5} = 0.8$.

representations. We compute the cosine similarity between SBERT representations.

- SBERT-WK (Wang and Kuo, 2020) is a variant of SBERT which weighs different layers of BERT.
- In contrast to the others, BLEURT (Sellam et al., 2020) is a supervised metric and fine-tunes BERT on the WMT datasets with available human assessment of translation quality.

Reference-free metrics We consider the following reference-free metrics:

- Multilingual Sentence BERT (mSBERT) (Reimers and Gurevych, 2020) is a multilingual version of Sentence BERT trained on parallel data with multilingual knowledge distillation. A teacher model trained on STS and NLI provides English sentence embeddings. mSBERT (student model) is trained to produce embeddings for the English sentence and its translation which are close to the embeddings of the teacher model.
- Multilingual Universal Sentence Encoder (MUSE) (Yang et al., 2019) is a multilingual sentence embedding. It is a dual-encoder model which was trained on multiple tasks such as NLI and translation ranking. MUSE was trained on mono- and multilingual data.
- LaBSE (Feng et al., 2020) is a dual-encoder framework. The encoders re-use pre-trained BERT and finetune it with Masked Language Modeling and Translation Language modeling on monolingual and parallel data.
- LASER (Artetxe and Schwenk, 2019) is a BiLSTM encoder trained on parallel corpora. It produces language-agnostic representations. The encoder-decoder architecture is trained jointly on different languages.
- XMoverScore (Zhao et al., 2020) extends MoverScore to operate in the cross-lingual setup, and relies on re-aligned multilingual BERT representations. Note that we exclude a target-side language model integrated in XMoverScore to have a similar setup as for MoverScore.

Except for XMoverScore, all metrics are based on calculation of cosine similarity between the source-language and target-language sentence embeddings. Except for MUSE and LASER, all metrics are based on BERT representations. Note that some multilingual reference-free metrics can also be used in the monolingual reference-based case, especially those based on calculating cosine similarity on top of sentence embeddings, thus we will include them in both settings.

4.2 Datasets

We use the datasets of the WMT shared task and the Semantic Text Similarity Benchmark (STSB) in our experiments. In the appendix, Table 8 shows the statistics in each dataset, and Table 9 shows examples for the sentences of the dataets.

WMT The WMT datasets contain an input sentence in the source language, the hypothesis translation of an MT system and a human reference sentence in the target language. Humans have rated the similarity between human reference and MT hypothesis using so-called 'direct assessment' (DA) scores which are framed in terms of one sentence 'adequately expressing the meaning' of another. We use these ratings as semantic scores in our setup.

For the reference-based case, we use the hypotheses and the references as sentence pairs. This data is collected over multiple language pairs which have English as target language (so both the human reference and the hypothesis are in English) of WMT15-WMT17. For the reference-free scenario, we pair the source texts with MT hypotheses and use the corresponding reference-to-hypothesis DA scores for similarity score. For German, we take the data from WMT15 (Bojar et al., 2015), WMT16 (Bojar et al., 2016) and WMT17 (Bojar et al., 2017). Chinese is only available in WMT17.

STSB The Semantic Text Similarity Benchmark (Cer et al., 2017) consists of English sentence pairs and a semantic similarity score for each pair. The scores were annotated by humans. The score is used as semantic score in the regression. In contrast to WMT, some sentence pairs in STSB are designed to have a different structure but a similar meaning.

While we use the sentence pairs directly for the monolingual case, we translate the sentences in the cross-lingual case using Google Translate, following Chidambaram et al. (2018).

5 Results

Reference-based Metrics Table 3 shows the results for the reference-based metrics. The R^2 values range from 0.43 to 0.76 on WMT, and from 0.58 to 0.91 on STS. This means we can reasonably well explain the metrics from our four explanatory variables and using a linear model. mSBERT can best be explained with an R^2 value of 0.91 on STS; however, since it has been trained on STS, this merely indicates overfitting. All metrics have positive coefficients for SEM, indicating that they all reflect semantic similarity and (semantic) 'adequacy' (as measured by DA), respectively: the weights range from 0.19 to 0.48 on WMT and from 0.12 to 0.76 on STS (ignoring mSBERT). The SYN coefficients are much lower and range from -0.05 to 0.16 for WMT and from -0.03 to 0.24 on STS. MoverScore and BERTScore are most affected by syntactic similarity (0.11 to 0.24), while the sentence embedding based techniques have coefficients around zero. The coefficients for MOR are low on STS, except for SBERT-WK, and moderate for WMT.

All metrics have comparatively large coefficients for lexical overlap, especially on WMT: the coefficient values range from 0.24 to 0.64 on WMT and from 0.14 to 0.67 on STS. Especially LEX dominates for MoverScore and BERTScore, indicating that these two metrics are most sensitive to lexical adversaries, potentially making them vulnerable to inputs such as 'man bites dog' vs. 'dog bites man'.

Reference-free Metrics Table 4 shows the results for ZH-EN in the reference-free setup (omitting the score for MOR as indicated earlier). Many SYN coefficients are now zero or negative, meaning that a larger syntactic difference between the input arguments leads to a higher metric score, indicating that metrics are sensitive to syntactic language differences. LEX is still significant in all cases. SEM has higher coefficient values than LEX in 6 out of 10 cases, and when it 'wins', it wins by a large margin. However, we note that the R^2 are low: they range from 0.3-0.39 on WMT and from 0.24-0.59 on STS. This means we can either not (well) explain the metrics given our current regressors or the relationship is not well explained by a linear model. The results for DE-EN are similar; we provide them in Table 10 (appendix).

To explore why R^2 scores are now lower, we note that reference-free metrics based on BERT might contain a form of *cross-lingual bias* (CLB) in that they do not properly score mutual translations, as Cao et al. (2020) and many others have shown that the multilingual subspaces induced by BERT are mis-aligned. We thus include a factor CLB as a regressor to measure how significant this bias is in different metrics. Note that the metrics use either different BERT variants or other representations such as LASER, which points to different sources of CLB. Therefore, we realize CLB differently across metrics. For each metric regression, we use the same metric but take a parallel sentence, i.e., source text and Google translation (as we assume that Google Translate has very high quality in general), as input arguments, and take the metric score as a proxy of the CLB factor. If a metric does not contain cross-lingual bias, it should assign almost full scores to parallel sentences; this constant would then be meaningless in the regression.

In Table 5, we show that including the CLB factor in the regression improves the R^2. We substantially improve the R^2 for XMoverScore (from 0.39 to 0.68), but observe little improvements for the remaining metrics (especially on STS). This is because XMoverScore is more problematic than the other metrics in terms of properly scoring mutual translations, given that the other metrics use BERT variants (or LASER) that have been finetuned (or trained) on parallel sentences. Apart from CLB, both SEM and LEX are the dominating factors in the regression. The DE-EN results are similar—see Table 11 (appendix).

Limitations The R^2 scores for the WMT dataset are almost always lower than the corresponding STS scores. One may not forget that STS sentences are in a sense artificial sentences of the form 'a girl is playing a guitar' while WMT contains more realistic sentences as well as their (possibly faulty, non-grammatical) translations. The WMT datasets are furthermore inhomogeneous in that we used different years from 2015 to 2017, which has dif-

Metric	WMT SEM	SYN	LEX	MOR	R^2	Metric	STS SEM	SYN	LEX	MOR	R^2
MoverScore	0.28	0.15	0.64	-0.06	0.76	MoverScore	0.30	0.24	0.45	0.04	0.61
BERTScore	0.27	0.16	0.61	-0.01*	0.74	BERTScore	0.12	0.11	0.67	0.06	0.69
LASER	0.19	0.04	0.33	0.32	0.53	LASER	0.55	0.05	0.28	0.09	0.63
SBERT	0.37	-0.02*	0.22	0.22	0.42	SBERT	0.73	-0.03	0.14	-0.06	0.60
SBERT-WK	0.30	-0.02	0.32	0.33	0.58	SBERT-WK	0.26	0.04	0.31	0.30	0.55
LaBSE	0.31	0.00*	0.36	0.22	0.53	LaBSE	0.60	0.07	0.27	0.06	0.67
mSBERT	0.41	-0.05	0.24	0.18	0.43	mSBERT	0.93	-0.02	0.04	0.00*	0.91
BLEURT	0.48	0.07	0.30	0.05	0.57	BLEURT	0.62	0.08	0.19	0.03*	0.58
mUSE	0.27	-0.03	0.39	0.19	0.49	mUSE	0.76	0.00*	0.22	-0.11	0.68

Table 3: Regression results for reference-based metrics. Coefficient values for linguistic regressors and R^2 values. * denotes $p \geq 0.05$ (non-significance). Intercept coefficients are small values and we omit them for clarity.

Metric	WMT SEM	SYN	LEX	R^2	Metric	STS SEM	SYN	LEX	R^2
XMoverScore	0.37	-0.08*	0.40	0.39	XMoverScore	0.09	-0.12	0.46	0.24
LASER	0.29	-0.03*	0.38	0.30	LASER	0.59	0.02*	0.27	0.51
mUSE	0.34	0.01*	0.35	0.33	mUSE	0.72	0.02*	0.12	0.59
LaBSE	0.47	-0.05*	0.17	0.30	LaBSE	0.63	0.08	0.26	0.56
mSBERT	0.42	-0.04*	0.25	0.31	mSBERT	0.87	-0.01*	0.05	0.80

Table 4: Regression results for Chinese-English reference-free metrics.

Metric	WMT SEM	SYN	LEX	CLB	R^2	Metric	STS SEM	SYN	LEX	CLB	R^2
XMoverScore	0.18	0.10	0.18	0.59	0.68	XMoverScore	0.08	0.16	0.36	0.48	0.53
LASER	0.18	-0.02*	0.32	0.34	0.43	LASER	0.56	0.03*	0.28	0.17	0.55
mUSE	0.26	0.00*	0.25	0.30	0.40	mUSE	0.72	0.02*	0.12	0.14	0.61
LaBSE	0.37	-0.05*	0.14	0.30	0.40	LaBSE	0.63	0.09	0.25	0.10	0.58
mSBERT	0.35	-0.04*	0.16	0.31	0.40	mSBERT	0.87	-0.01*	0.05*	0.07	0.81

Table 5: Regression results for Chinese-English reference-free metrics. We add the CLB factor in the regression.

ferent participating MT systems as well as slightly different task definitions, corresponding to an aggregation of different domains. The STS dataset is monolingual and was translated by Google Translate for the cross-lingual scenario. The latter may lower the quality of the data, but WMT data also contains translations.

The WMT scores measure the similarity between the reference and the hypothesis but we compare the source with the hypothesis in the cross-lingual scenario, which reflects a mismatch. Fomicheva et al. (2020a) provide a dataset which gives human DA scores between source and hypothesis. We repeated the experiments with this dataset. The full results are shown in Table 12 in the appendix (omitting the CLB factor). The R^2 scores of 3 out of 5 metrics improve (slightly) compared to the WMT dataset for German-English but all R^2 scores are lower for Chinese-English. This means that mismatched DA scores are apparently not the main reason for our low regression fits. With the new DA scores, all coefficients for SEM are considerably lower. They are in the range of range 0.06 to 0.14 compared to the maximum of 0.47 for WMT. In contrast, all SYN (0-0.12) scores are higher especially vof German. All MOR (0.16-0.33) and most LEX scores are higher but they are still in a similar range as for the original DA scores.

6 Analysis

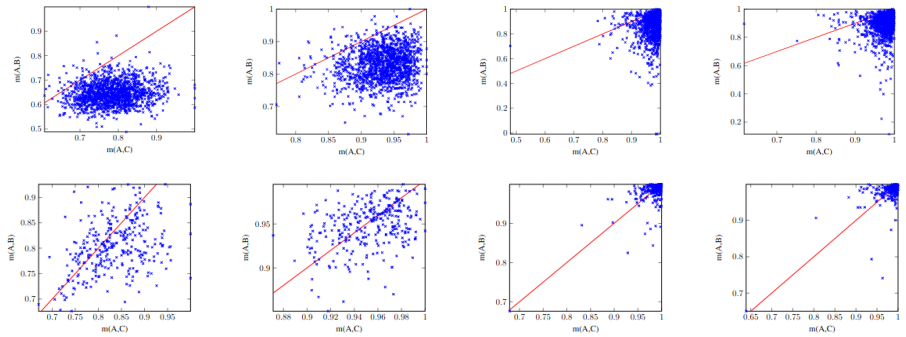
	Lex(A,B)	Lex(A,C)	Size
Freitag et al. (2020)	0.49	0.99	1452
PAWS	0.84	0.94	100

Table 6: Averaged lexical overlap and size of the datasets for the adversarial experiments.

In the following, we analyze two observations from our previous experiments in more depth: (i) the sensitivity of metrics to lexical overlap; (ii) the orthogonality of metrics in that they capture different linguistic signals.

PAWS		
Sentence A	**Sentence B**	**Sentence C**
Later in 2014 , Dassault Systèmes was bought by Quintiq.	Dassault Systèmes was bought in 2014 by Quintiq.	In 2014 , Quintiq was bought by Dassault Systèmes.
They are high , built of concrete faced with small blocks of stone.	They are high built of concrete with small stone blocks.	They are small , built of concrete with high stone blocks.
Freitag et al. (2020) translated		
Sentence A	**Sentence B**	**Sentence C**
Shark injures 13-year-old on lobster dive in California	A 13-year-old is injured by a shark while diving for lobsters in California	Shark injures 13-year-old on dive lobster in California
Kovacic did a quick give-and-go at midfield.	Kovacic managed a quick one-two in midfield.	Kovacic did a quick midfield at give-and-go.

Table 7: Selected sentences for the adversarial experiments.



(a) MoverScore (b) BERTScore (c) mSBERT (d) SBERT

Figure 3: Distribution of $\mathfrak{m}(A, B)$ vs. $\mathfrak{m}(A, C)$. Top: Dataset from Freitag et al. (2020). Bottom: PAWS dataset.

6.1 Adversarial experiments

According to our results, all metrics rely on lexical overlap, which indicates that they may not be robust to adversarial examples. We check this by an additional experiment, for reference-based metrics, in which we query their pairwise preferences over three sentences: Sentence A is the anchor sentence; sentence B is a paraphrase of sentence A with little lexical overlap; sentence C is a non-paraphrase with high lexical overlap. A good metric $\mathfrak{m}$ would have $\mathfrak{m}(A, B) > \mathfrak{m}(A, C)$ but the high lexical overlap between A and C makes this task difficult.

Freitag et al. Sentences A are taken as source sentences from WMT19. Freitag et al. (2020) provided alternative references for WMT19; they instructed human professional translators to paraphrase the references as much as possible in terms of lexical choice and sentence structure but keep the same semantics. We take these as sentences B. We produce sentences C from sentences A: for each sentence A, we detect the nouns within the sentence using the NLTK POS tagger, and then we randomly permute them to produce a sentence C. Since Freitag et al. (2020) provided sentences in German, we translate all sentences into English using Google translate. We note that by inspection the translations are generally of high quality and satisfy our constraints of inducing paraphrases with low lexical overlap and non-paraphrases with high lexical overlap. Examples are shown in Table 7 and statistics in Table 6.

PAWS We complement the analysis with the native English PAWS dataset (Zhang et al., 2019) which consists of paraphrase and non-paraphrase pairs that have high lexical overlap. For each sentence in the dataset, there are multiple paraphrases and non-paraphrases. For a given sentence A, we use the paraphrase with the smallest amount of lexical overlap as sentence B and sentence C is the non-paraphrase with the highest amount of lexical overlap with A. We note that PAWS is more problematic as even the paraphrases B with minimum amount of lexical overlap do have considerable lexical overlap. Therefore, we select the 100 sentence pairs with the smallest amount of lexical overlap between sentence A and B. Table 6 shows the lex-

ical overlap and the size of the datasets. Indeed, for PAWS, sentences B have only a little less lexical overlap with A than sentences C, while the dataset of Freitag et al. (2020) has a much clearer separation between B and C in this respect.

Figure 3 shows the distribution of $\mathfrak{m}(A, B)$ and $\mathfrak{m}(A, C)$ for selected metrics $\mathfrak{m}$. Overall, the adversarial results on translated and non-translated datasets point in a similar direction. We see that metrics clearly prefer the high lexical overlap sentences which are non-paraphrases (sentences C) in the translated dataset of Freitag et al. (2020). For non-translated PAWS, metrics are at least to some degree indifferent, but tend to prefer B on average, with MoverScore and BERTScore having higher preference for C than mSBERT and SBERT, which confirms our linear regression results.

Overall, these experiment show that metrics are indeed not robust to lexical adversarial examples.

6.2 Ensemble of Models

Our experiments in Section 4 show that the different metrics use different information signals, even when they use the same underlying BERT representations. For example, BERTScore relies more on lexical overlap and mSBERT relies more on semantics; BERTScore and MoverScore both capture syntax, while the other metrics are less sensitive to it. This means that the metrics are to some degree orthogonal. Thus, we suspect that a combination of metrics yields a considerably better metric. We check this hypothesis through an extra experiment. We combine especially BERTScore and MoverScore with SBERT and mSBERT based metrics.

We evaluate the metrics on segment-level. In segment-level evaluation, each sentence pair gets a score from $\mathfrak{m}$. The Pearson correlation coefficient is then calculated between these scores and the human judgement, disregarding the systems which generated the translations. To combine metrics, we simply average their scores. In the evaluation, we use the best performance of the single metrics as baseline and compare it to our combined metrics.

Table 13 (appendix) shows the improvements for different language pairs. In the reference-free case, the two best ensembles combine XMoverScore with mSBERT and LaBSE—the latter two rely less on lexical overlap than the first—leading to big improvements of 11-13% over the best individual metric. Combining metrics relying on similar factors shows less improvement, and often even leads to worse results. In the reference-based case, we combine BERTScore with mSBERT and observe an improvement of 8%, more than for any other combination we tested. These results show that combing metrics relying on different factors can largely improve their performance.[1]

7 Conclusions

We disentangled BERT-based evaluation metrics along four linguistic factors: semantics, syntax, morphology, and lexical overlap. The results indicate that (i) the different metrics capture these different aspects to different degrees but (ii) they all rely on semantics and lexical overlap. The first observation indicates that combining metrics may be helpful, which we confirmed: simple parameter-free averaging of hetereogenous metric scores can improve correlations with humans by up to more than 13% in our experiments. The second observation shows that these metrics may be prone to adversarial fooling, just like BLEU and ROUGE, which we confirmed in an additional experiment in which we queried metric preferences over paraphrases with little lexical overlap and non-paraphrases with high lexical overlap. Future metrics should especially take this last aspect into account, and improve their robustness to adversarial conditions.

There is much scope for future research, e.g., in developing better global explanations for reference-free metrics (as we cannot yet well explain these metrics), better linguistic factors (e.g., a clearer conceptualization of morphological similarity of two sentences) and in developing local explainability techniques for evaluation metrics (Fomicheva et al., 2021a,b).[2]

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions, which greatly improved the final version of the paper. We also thank Micha Dippell for his early experiments contributing to this work. This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1.

[1] In work independent from us, Song et al. (2021) also ensemble BERT based reference-free evaluation metrics.

[2] Code and data for our experiments are available from https://github.com/SteffenEger/global-explainability-metrics.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond.

Philip Bille. 2005. A survey on tree edit distance and related problems. *Theoretical computer science*, 337(1-3):217–239.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2020. How to probe sentence embeddings in low-resource languages: On structural design choices for probing task evaluation. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 108–118, Online. Association for Computational Linguistics.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021a. The eval4nlp shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2021b. Translation error detection as rationale extraction.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2020a. MLQE-PE: A multilingual quality estimation and post-editing dataset. *arXiv preprint arXiv:2010.04480*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020b. BERGAMOT-LATTE submissions for the WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1010–1017, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Zorik Gekhman, Roee Aharoni, Genady Beryozkin, Markus Freitag, and Wolfgang Macherey. 2020. KoBE: Knowledge-based machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3200–3207, Online. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Yannik Keller, Jan Mackensen, and Steffen Eger. 2021. BERT-defense: A probabilistic model based on BERT to combat cognitively inspired orthographic adversarial attacks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1616–1629, Online. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of*

the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest at WMT2020: Sentence-level direct assessment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1049–1055, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. SentSim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156, Online. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Bin Wang and C. C. Jay Kuo. 2020. Sbert-wk: A sentence embedding method by dissecting bert-based word models.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual universal sentence encoder for semantic retrieval.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

8 Appendix

The following tables contain remaining experimental results.

Dataset	English	Chinese	German
WMT	8595	560	1620
STSB	5749	5749	5749

Table 8: Number of sentence pairs in each dataset and language pair.

Sentence 1	Sentence 2	SEM
WMT		
Why is it such a difference?	Why such a difference?	0.94
There are coincidences.	Zufälle gibt es.	0.57
STS		
Some men are fighting.	Two men are fighting.	4.25
A woman is writing.	Eine Frau schwimmt.	0.1

Table 9: Example of WMT and STS datasets.

Metric	WMT SEM	SYN	LEX	MOR	R^2	Metric	STS SEM	SYN	LEX	MOR	R^2
XMoverScore	0.37	-0.07*	0.33	-0.01*	0.31	XMoverScore	0.08	-0.02*	0.51	0.03	0.30
LASER	0.21	-0.08	0.31	0.31	0.29	LASER	0.54	0.00*	0.28	0.15	0.54
mUSE	0.19	-0.00*	0.40	0.08	0.25	mUSE	0.71	0.00*	0.09	0.07	0.6
LaBSE	0.30	-0.03*	0.26	0.07	0.21	LaBSE	0.60	0.05	0.29	0.04	0.60
mSBERT	0.39	-0.06*	0.22	0.05*	0.26	mSBERT	0.89	-0.02	0.05	0.02	0.84

Table 10: Regression results for German-English reference-free metrics. Coefficient values for linguistic regressors and R^2 values. * denotes $p \geq 0.05$ (non-significance).

Metric	WMT SEM	SYN	LEX	MOR	CLB	R^2	Metric	STS SEM	SYN	LEX	MOR	CLB	R^2
XMoverScore	0.14	0.08	0.16	-0.03*	0.63	0.61	XMoverScore	0.09	0.25	0.43	-0.09	0.40	0.54
LASER	0.11	-0.03*	0.28	0.19	0.40	0.46	LASER	0.51	0.01*	0.32	0.12	0.11	0.58
mUSE	0.19	0.02*	0.28	0.03*	0.35	0.34	mUSE	0.70	0.02*	0.12	0.05*	0.11	0.61
LaBSE	0.29	0.00*	0.14	0.00*	0.45	0.38	LaBSE	0.59	0.06	0.30	0.02*	0.12	0.62
mSBERT	0.36	-0.02*	0.09	0.02*	0.36	0.36	mSBERT	0.88	-0.02*	0.05*	0.02*	0.03*	0.85

Table 11: Regression results for German-English reference-free metrics. Coefficient values for linguistic regressors and R^2 values. * denotes $p \geq 0.05$ (non-significance). We add the CLB factor in the regression.

Metric	German-English SEM	SYN	LEX	MOR	R^2	Metric	Chinese-English SEM	SYN	LEX	R^2
XMoverScore	0.16	0.05*	0.33	0.16	0.20	XMoverScore	0.14	0.04*	0.35	0.15
LASER	0.12	0.08	0.30	0.46	0.41	LASER	0.06*	0.03*	0.35	0.14
mUSE	0.12	0.12	0.25	0.43	0.35	mUSE	0.12	0.05*	0.42	0.21
LaBSE	0.10	0.08	0.23	0.51	0.41	LaBSE	0.16	0.04*	0.43	0.23
mSBERT	0.14	0.06*	0.21	0.33	0.22	mSBERT	0.24	0.00*	0.31	0.17

Table 12: Regression results for reference-free metrics on the Fomicheva et al. (2020a) dataset which contains DA scores comparing sources and hypotheses (rather than references and hypotheses). Coefficient values for linguistic regressors and R^2 values. * denotes $p \geq 0.05$ (non-significance).

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Avg
Reference-based								
BERTScore + mSBERT	8%	15%	7%	7%	11%	2%	4%	8%
SBERT + MoverScore	3%	10%	1%	11%	6%	-1%	1%	4%
SBERT + LASER	6%	6%	0%	2%	5%	3%	10%	5%
SBERT + mUSE	5%	10%	4%	2%	8%	4%	2%	5%
Reference-free								
XMoverScore + mSBERT	9%	19%	11%	12%	18%	13%	12%	13%
XMoverScore + LaBSE	8%	14%	5%	10%	17%	9%	17%	11%
mSBERT + LASER	8%	9%	-2%	1%	3%	0%	12%	4%
LaBSE + LASER	7%	15%	-4%	2%	1%	-2%	4%	3%
XMoverScore + LASER	0%	7%	-5%	2%	4%	1%	13%	3%
mSBERT + LaBSE	3%	7%	-11%	-6%	-7%	-6%	2%	-3%

Table 13: Performance gains from the ensemble metrics over single best metrics.

Part III

Epilogue

Chapter 13

Conclusion

Given the unprecedented success of deep learning technology, state-of-the-art natural language generation (NLG) systems compete for human-parity performance at an ever-increasing pace, and as such fair and adequate evaluation plays a vital role for properly tracking the progress of NLG systems. In this work, we sketch a holistic view of principled NLG evaluation from three complementary perspectives: (i) evaluating the *adequacy* of NLG systems with high-quality evaluation metrics; (ii) comparing NLG systems for properly tracking the *progress* and (iii) providing the understanding of evaluation metrics, all of which are driven by the evaluation principles, pertaining to *adequacy*, *progress* and *diagnostic*, outlined in the European project EAGLES-1996 for evaluating natural language technologies. In particular, we identify a series of challenges derived from the inherent characteristics of these perspectives, and address the identified issues by proposing novel evaluation metrics, rigorous comparison approaches and explainability techniques for understanding non-transparent metrics.

Reference-based Evaluation. As the low-cost alternatives to human evaluation, traditional evaluation metrics such as BLEU and ROUGE, evaluating the performance of NLG systems in seconds or minutes, have been extensively adopted in recent NLG evaluation campaigns. However, these metrics have failed to judge the text quality of system outputs when system output and human reference have no lexical overlap. Thus, the challenge of addressing lexical similarity in the absence of lexical overlap has become the major bottleneck in reference-based evaluation (**RQ1**). In Chapter 5, we proposed MoverScore, a reference-based evaluation metric, to overcome this challenge, which addresses the lexical similarity by using broader semantic relatedness of words in vector space. We demonstrated that MoverScore strongly correlates with human ratings in machine translation, summarization and image captioning, surpassing BLEU by up to 25 correlation points.

In relation to **RQ2**, we investigated the extent to which recent BERT-based metrics can recognize text coherence for evaluating long texts. In Chapter 6, we showed that these metrics cannot penalize incoherent elements in system outputs, and correlate poorly with human rated coherence. To address this issue, we proposed

DiscoScore, a reference-based metric, targeting the assessment of text coherence, driven by Centering theory. We showed that DiscoScore achieves strong system-level correlation with human ratings, not only in coherence but also in factual consistency and other aspects.

In Section 5.3.2, we showed that the selection of word embeddings across the layers of BERT is of importance for the performance of BERT-based evaluation metrics. We proposed to aggregate word embeddings across layers by using the KDE routing. In Chapter 7, we elaborate on the KDE routing, a kernel-based density estimator, used to aggregate capsules (another form of embeddings) across layers without supervision.

Reference-free Evaluation. Much unlike reference-based metrics, reference-free metrics remove the need for human reference by directly comparing system output with source text, allowing for unlimited evaluation of NLG systems. Therefore, reference-free metrics have been researched for long. However, the proposal of previous metrics required human ratings as supervision and language-dependent preprocessing, hindering the wide applicability of these metrics.

In Subpart B, we particularly addressed reference-free evaluation in machine translation, and showed that contextualized encoders such as BERT (responsible for the success of reference-based metrics) cannot show advantages in the absence of human reference. In relation to **RQ2**, we proposed XMoverScore, a parameterized, reference-free evaluation metric, which operates MoverScore in a cross-lingual setup. XMoverScore can be parameterized by the choice of solutions to address high-quality cross-lingual embeddings, which we identify to be crucial for XMoverScore surpassing reference-based BLEU.

Regarding the choice of solutions, we proposed to rotate the vector space and to reduce "language bias" in Chapter 8. Chapter 9 addressed the removal of language identity signals from the vector space, while Chapter 10 addressed the vector space alignment with density-based approaches for low-resource languages.

System Comparison. Proper evaluation concerns not only the designing of evaluation metrics, but also the comparison of NLG systems. In relation to **RQ4**, we questioned the use of *average* to report system rankings—see Chapter 11. In particular, we showed that global statistics such as *average* and *median* of instance-level evaluation scores cannot carry out rigorous comparison, as they ignore the fact that systems are evaluated on the same test instances. To address this issue, we introduced pairwise comparison approaches to compare text generation systems on instance level, which bases the prediction of system strengths on the probability of one system over the other.

Explainability for Evaluation Metrics. Recent evaluation metrics building on contextualized encoders exhibit much better quality levels than traditional metrics such as BLEU and ROUGE. However, given the complexity and non-transparency of the encoders, understanding these metrics is challenging (**RQ5**). In this thesis,

we provided several explainability techniques to address this issue, inspired by the recent advances in explainable artificial intelligence.

In Section 5.3.1, we visualized the process of MoverScore by picturing the alignment of word pairs in system output and human reference. In Chapter 6, we provided justifications to the superiority of one metric over another, showing that the more discriminative the features (derived from the metrics) are in separating system output from human reference, the better the metrics perform. In Chapter 12, we provided understanding on what linguistic factors evaluation metrics capture, and showed that both reference-based and reference-free metrics based on BERT are sensitive to lexical overlap, much like BLEU and ROUGE.

To summarize, in this thesis, we have acknowledged the importance of three complementary perspectives constituting the holistic view of principled evaluation in NLG: (i) evaluation metrics, (ii) system comparison and (iii) explainability for metrics. To this end, we have outlined the current state of challenges pertaining to the inherent characteristics of these perspectives.

We have addressed these challenges through individual studies and the proposals of novel metrics, rigorous comparison approaches, explainability techniques for evaluation metrics. Given text generation encompassing an enormous range of tasks, we acknowledge that our research cannot hope to be a comprehensive treatment for solving evaluation in all tasks, but we have made significant contributions to machine translation and text summarization evaluation, paving the path towards fair and adequate evaluation in other tasks, such as dialogue generation, story generation, and so forth.

Building upon our work, the scope for future work is huge. We now outline only a few possible directions:

As long-text generation keeps growing continuously, recognizing text coherence has become crucial in the assessment of system outputs, not only for reference-based evaluation (see Chapter 6). In relation to (**RQ2**), we intend to design discourse metrics by freeing themselves from the need for human reference. As an example in machine translation, we will start by understanding discourse coherence in a cross-lingual setup (Menzel et al., 2017), such as how coherence is realized across languages. We then design reference-free discourse metrics by modeling such language-dependent coherence phenomena in source language texts and system translations.

In relation to (**RQ4**), we intend to study how to rigorously compare evaluation metrics. Evaluation metrics are often compared according to the correlation between metric and human scores. For system-level correlation, one has to assemble system-level metric and human scores by averaging instance-level scores. As *average* cannot properly compare NLG systems (see Chapter 11), we suspect that comparing evaluation metrics in this manner is also problematic. To this end, we will investigate a range of aggregation mechanisms such as *median* and *BT*, as it is the case for our studies in comparing NLG systems. Further, we will pair evaluation metrics with aggregation mechanisms, such as {ROUGE, average} and {BLEU, BT}, and

then compare such pairs for ensuring rigorous comparison of evaluation metrics.

We hope that the contributions presented in this thesis fuel and inspire more research towards principled NLG evaluation responsible for ensuring trustworthy, reproducible and unbiased results.

Bibliography

Linguistic Data Annotation. 2002. Assessment of fluency and adequacy in arabic-english and chinese-english translations.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Kenneth J. Arrow. 1950. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346.

Adrien Bennetot, Jean-Luc Laurent, Raja Chatila, and Natalia Díaz-Rodríguez. 2019. Towards explainable neural-symbolic visual reasoning. *arXiv preprint arXiv:1909.09065*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shruti Bhosale, Kyra Yee, Sergey Edunov, and Michael Auli. 2020. Language models not just for pre-training: Fast online neural noisy channel modeling. In *Proceedings of the Fifth Conference on Machine Translation*, pages 584–593, Online. Association for Computational Linguistics.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A grain of salt for the wmt manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 1563–1580, Online. Association for Computational Linguistics.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*, volume abs/2002.03518.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Haixia Chai, Wei Zhao, Steffen Eger, and Michael Strube. 2020. Evaluation of coreference resolution systems under adversarial attacks. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 154–159, Online. Association for Computational Linguistics.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang 0001, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1657–1668. Association for Computational Linguistics.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. WMDO: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 494–500, Florence, Italy. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference (DUC 2005)*, volume 2005, pages 1–12.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, Uppsala, Sweden. Association for Computational Linguistics.

Daniel Deutsch and Dan Roth. 2020. Sacrerouge: An open-source library for using and developing summarization evaluation metrics. *CoRR*, abs/2007.05374.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, pages 4171–4186.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE.

Julie S Downs, Mandy B Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system? screening mechanical turk workers. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2399–2402.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy, editors. 2020. *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics, Online.

Melda Eksi, Erik Gelbing, Jonathan Stieber, and Chi Viet Vu. 2021. Explaining errors in machine translation with absolute gradient ensembles. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 238–249, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*, 9:391–409.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Ronald A. Fisher. 1935. *The Design of Experiments*. 2nd Ed. Oliver and Boyd, Edinburgh.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021a. The eval4nlp shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021b. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2020a. MLQE-PE: A multilingual quality estimation and post-editing dataset. *arXiv preprint arXiv:2010.04480*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020b. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Maria Fuentes, Edgar González, Daniel Ferrés, and Horacio RODRiguez. 2005. Qasum-talp at duc 2005 automatically evaluated with a pyramid based metric. In *HLT-EMNLP Workshop (DUC 2005), Vancouver, Canada*.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 445–450, Beijing, China. Association for Computational Linguistics.

Yang Gao, Steffen Eger, Wei Zhao, Piyawat Lertvittayakumjorn, and Marina Fomicheva, editors. 2021. *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics, Punta Cana, Dominican Republic.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Gregor Geigle, Jonas Stadtmüller, Wei Zhao, Jonas Pfeiffer, and Steffen Eger. 2021. TUDa at WMT21: Sentence-level direct assessment with adapters. In *Proceedings of the Sixth Conference on Machine Translation*, pages 911–919, Online. Association for Computational Linguistics.

Zorik Gekhman, Roee Aharoni, Genady Beryozkin, Markus Freitag, and Wolfgang Macherey. 2020. KoBE: Knowledge-based machine translation evaluation. In

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3200–3207, Online. Association for Computational Linguistics.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Jesús Giménez and L Asiya Marquez. 2010. An open toolkit for automatic machine translation (meta-) evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94:77–86.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2015. Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40.

Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. NEWSROOM: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 708–719, New Orleans, Louisiana, USA.

Giovanni Guida and Giancarlo Mauri. 1986. Evaluation of natural language processing systems: Issues and approaches. *Proceedings of the IEEE*, 74(7):1026–1035.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 1693–1701, Montreal, Quebec, Canada.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Erik Hollnagel. 1986. Cognitive system performance analysis. In *Intelligent decision support in process environments*, pages 211–226. Springer.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural networks*, 3(5):551–560.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020. Best practices for crowd-based evaluation of German summarization: Comparing crowd, expert and automatic evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online. Association for Computational Linguistics.

Hongyan Jing, Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *AAAI symposium on intelligent summarization*, pages 51–59.

Karen Sparck Jones. 2001. Automatic language and information processing: rethinking evaluation. *Natural Language Engineering*, 7(1):29–46.

Karen Sparck Jones and Julia R Galliers. 1995. *Evaluating natural language processing systems: An analysis and review*, volume 1083. Springer Science & Business Media.

Tasnim Kabir and Marine Carpuat. 2021. The UMD submission to the explainable MT quality estimation shared task: Combining explanation models with sequence labeling. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 230–237, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of bert-

based evaluation metrics by disentangling along linguistic factors. In *EMNLP 2021*, pages 8912–8925, Online. Association for Computational Linguistics.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, Washington, USA. Association for Computing Machinery.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 957–966.

Anne Lauscher. 2021. Language representations for computational argumentation.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics, Association for Computational Linguistics.

Liam Li and Ameet Talwalkar. 2020. Random search and reproducibility for neural architecture search. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 367–377. PMLR.

Yang Li, Wei Zhao, Erik Cambria, Suhang Wang, and Steffen Eger. 2021. Graph routing between capsules. *Neural Networks*, 143:345–354.

Chin-Yew Lin. 2001. See-summary evaluation environment. *WWW site, URL: http://www. isi. edu/cyl/SEE*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and Automatic Evaluation of Summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization - Volume 4*, pages 45–51.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Liyuan Liu, Xiang Ren, Jingbo Shang, Xiaotao Gu, Jian Peng, and Jiawei Han. 2018. Efficient contextualized representation: Language model pruning for se-

quence labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, page arXiv:1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2013. Analyzing the Capabilities of Crowdsourcing Services for Text Summarization. *Language Resources and Evaluation*, 47(2):337–369.

Chi-kiu Lo. 2017. Meant 2.0: Accurate semantic mt evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. XMEANT: Better semantic MT evaluation without reference translations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 765–771, Baltimore, Maryland. Association for Computational Linguistics.

Chi-kiu Lo and Dekai Wu. 2013. MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based MT evaluation metric. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 422–428, Sofia, Bulgaria. Association for Computational Linguistics.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vish-

wanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy". Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Katrin Menzel, Ekaterina Lapshinova-Koltunski, and Kerstin Kunz. 2017. *New perspectives on cohesion and coherence: Implications for translation*, volume 6. Language Science Press.

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1414–1423, San Diego, California. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In

Advances in neural information processing systems, pages 3111–3119, Lake Tahoe, Nevada, USA. Curran Associates, Inc.

Tom Mitchell. 1997. Machine learning.

Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1345–1354.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Paul Over and Walter Liggett. 2002. Introduction to duc: An intrinsic evaluation of generic news text summarization systems. *Proc. DUC. http://wwwnlpir. nist. gov/projects/duc/guidelines/2002. html.*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of NLP systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2301–2315, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisboa, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: IBM1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 99–103, Edinburgh, Scotland. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Greg Ridgeway, David Madigan, Thomas Richardson, and John O'Kane. 1998. Interpretable boosted naïve bayes classification. In *KDD*, pages 101–104.

Raphael Rubino, Atsushi Fujita, and Benjamin Marie. 2021. Error identification for machine translation with metric embedding and attention. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 146–156, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montréal, Canada. Association for Computational Linguistics.

Alexander Rush, Roi Reichart, Michael Collins, and Amir Globerson. 2012. Improved parsing and POS tagging using inter-sentence consistency constraints. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1434–1444, Jeju Island, Korea. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M.

Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. 2018. Learning-based composite metrics for improved caption evaluation. In *Proceedings of ACL 2018, Student Research Workshop*, pages 14–20, Melbourne, Australia. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Edwin D. Simpson. 2021. Statistical Significance Testing for Natural Language Processing. *Computational Linguistics*, 46(4):905–908.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, volume 200, pages 223–231, Cambridge, Massachusetts, USA. Cambridge, MA, Association for Machine Translation in the Americas.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha, Qatar. Association for Computational Linguistics.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation systems. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Jean Tague-Sutcliffe. 1992. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4):467–490.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Simone Teufel and H Van Halteren. 2004. Evaluating information content by factoid analysis: human annotation and stability.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Christoph Tillmann, Stephan Vogel, Hermann Ney, A. Zubiaga, and Hassan Sawaf. 1997. Accelerated DP Based Search for Statistical Translation. In *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*.

Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*.

Marcos Treviso, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2021. IST-unbabel 2021 submission for the explainable quality estimation shared task. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 133–145, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oleg Vasilyev and John Bohannon. 2021. ESTIME: Estimation of summary-to-text inconsistency by mismatched embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 94–103, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575.

Ellen M Voorhees. 2003. Evaluating the evaluation: a case study using the trec 2002 question answering track. In *Proceedings of the 2003 Human Language Technology*

Conference of the North American Chapter of the Association for Computational Linguistics, pages 260–267.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

John S White, Theresa A O'Connell, and Francis E O'Mara. 1994. The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulleting*, 6:80–83.

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

William A Woods. 1977. A personal view of natural language understanding. *ACM SIGART Bulletin*, (61):17–20.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Hao Zhang and Daniel Gildea. 2007. Factorization of synchronous context-free grammars in linear time. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 25–32, Rochester, New York. Association for Computational Linguistics.

Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. 2019. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6261–6270.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, volume abs/1904.09675. OpenReview.net.

Wei Zhao and Steffen Eger. 2022. Constrained density matching and modeling for cross-lingual alignment of contextualized representations. In *Proceedings of The 14th Asian Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Wei Zhao, Haiyun Peng, Steffen Eger, Erik Cambria, and Min Yang. 2019a. Towards scalable and reliable capsule networks for challenging NLP applications. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1549–1559, Florence, Italy. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019b. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. Discoscore: Evaluating text generation with bert and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics.

Donglin Zhuang, Xingyao Zhang, Shuaiwen Leon Song, and Sara Hooker. 2021. Randomness in neural network training: Characterizing the impact of tooling. *CoRR*, abs/2106.11872.