Digital Transformation of Science: AI-Assisted Collaborative Reading and Evaluation

Jan Buchmann, Nils Dycke, Dennis Zyska, Iryna Gurevych
Ubiquitous Knowledge Processing Lab, TU Darmstadt
January 23, 2023

1 Assisting Peer Review

Peer review is a common instrument of quality control in the academic world. New scientific knowledge is only accepted and, in many cases, only published when it has passed this barrier. However, in its current state, peer review has two major shortcomings. First, it is work-intensive and slow. Paired with the continuous increase in submissions (Publons, 2018), this means that more and more time of researchers is spent with the peer review of others' submissions, and publications are delayed. Second, there have been several investigations showing that acceptance decisions based on peer review are in part arbitrary (Cortes & Lawrence, 2021; Lee et al., 2013). Our research aims to address some of these shortcomings by assisting researchers in reading and evaluation of scientific publications. To help make peer reviewing faster and more reliable, and to give guidance in this process to young researchers, our lab is developing CARE (Collaborative Augmented Reading Environment, formerly PEER), an Artificial Intelligence (AI) assisted software to support researchers in the annotation phase of reading and evaluation of scientific publications. To provide the maximum benefit to researchers from any field, we design CARE to adapt to (i) the user, (ii) the domain of research, and (iii) the document at hand. For this, we need to know how researchers from various fields interact with scientific publications, and what their needs in such a tool are. Our research at the Center for Advanced Internet Studies (CAIS) was aimed to answer these questions, and brought valuable insights that will guide the future development of CARE.

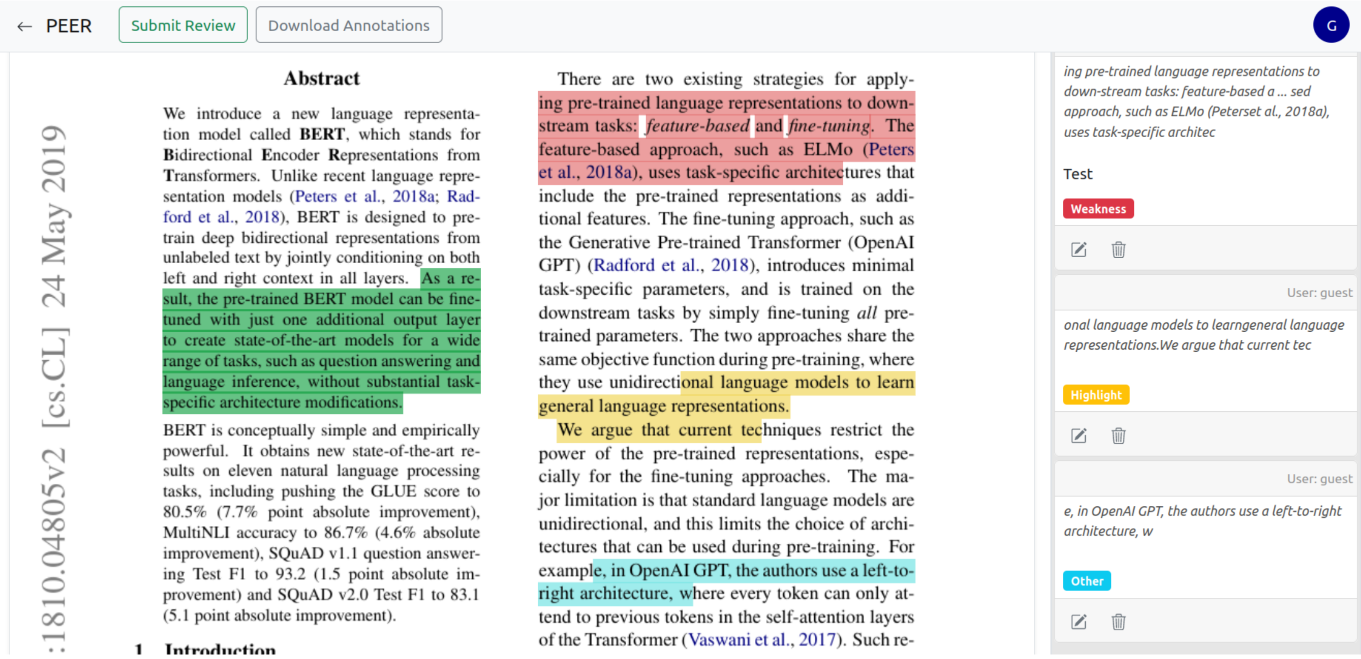
This report first introduces the CARE software tool in greater detail. It then presents the setup and results of two studies performed at CAIS. In one, a survey was distributed among the CAIS members, in which we wanted to know more about the practices of publishing, reading and reviewing in the scientific communities of the CAIS researchers, and their personal scientific work. In the other study, the CAIS members tried out CARE first hand in a simulated peer reviewing scenario, and were asked for feedback on the software. Finally, the report summarizes the main findings and provides an outlook on the future of CARE.

2 The CARE Tool

CARE is meant to support researchers in reading and evaluation of scientific articles. The process from the first read of a scientific manuscript to a (written) evaluation usually consists of several steps such as an initial fast read, in-depth read of specific sections, writing comments or annotations, notes in bullet point style and composing the final evaluation text. CARE focuses on the annotation phase. First, because we assume (and here verify) that annotation is a common and natural mode of interaction with unknown texts. Second, because it is an under-researched aspect of the reviewing process, which means that the potential for assistance is not fully explored.

Our software is accessible from the web browser with an easy-to-use interface to load, read, highlight and comment on scientific articles (or any other document in PDF format). Comments can be made on the full article or any part (a text span, a table, ...) of it. We are developing two modes of AI assistance the tool can provide.

In the first one, the user is guided through the evaluation of an article along a sequence of aspects, which serve as a "scaffold" known from educational research (Sandoval & Reiser, 2004). An example for such an aspect would be the evaluation of the appropriateness of the method for the given research question. For each of the aspects in a scaffold, the AI agent retrieves and highlights relevant text spans from the article and adds a comment with the aspect (e.g. "Is the method appropriate for the research question?"). The user can then reply to the comment with their evaluation of the respective aspect. The guidance by CARE should make peer reviewing more accessible to junior researchers (Fok et al., 2022).



arXiv:1810.04805v2 [cs.CL] 24 May 2019

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pretrained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such re-

ing pre-trained language representations to down-stream tasks: feature-based a ... sed approach, such as ELMo (Peterset al., 2018a), uses task-specific architec

Test

Weakness

User: guest

onal language models to learngeneral language representations.We argue that current tec

Highlight

User: guest

e, in OpenAI GPT, the authors use a left-to-right architecture, w

Other

Figure 1: The CARE interface (review view). The colors show highlights of different categories. The comments are on the right.

The second AI functionality of CARE is meant to facilitate the generation of a full review report from the comments made in the annotation phase, as well as collaborative reviewing. A user can leave comments on any part of an article. CARE's AI agent recognizes the category of the comment (e.g. "strength" or "weakness") and marks the comment with the category. The classification itself, but not the implementation in a user-oriented software, has been done in other studies for several sets of categories (Hua et al., 2019; Kennard et al., 2022). CARE can then assemble an overview of the comments, structured according to the categories. This structured overview can help the user to formulate the full review text. It can also be exchanged between researchers to give an easily accessible overview of a peer's opinions expressed in the comments.

CARE was initially developed in the context of computer science and the life sciences. As we want it to be useful for researchers from many disciplines, we need to know their specific requirements for the functionalities of the software. Furthermore, we would like to receive input on the usability of the tool in its current form. With its fellowships and members, the CAIS gathers excellent researchers interested in the digital transformation of society. Therefore, the CAIS was an ideal platform to learn about their needs when working with scientific articles, to test CARE, and to promote its interdisciplinary use.

3 Practices of Peer Reviewing: Survey

As the first step of our research agenda, we elucidated the practices of publishing, reading and reviewing in the scientific communities of the CAIS researchers, and their personal scientific work. Our questions could be summarized as

- How do people review, and where is potential for improvement by assistance?
- How should a software tool for reading and reviewing be designed (e.g. which file formats should be supported)?

To answer these questions, we designed a survey with 31 content items, 6 demographic items and a feedback section. The demographic items asked for information such as the field of study and academic position. They helped us to assess the representivity of our sample. The content items were grouped into four topics: (1) Scientific publishing, (2) properties of scientific articles and personal reading practices, (3) peer review in the field of study of the participant, and (4) personal peer review practices.

The survey was distributed via E-Mail among the CAIS members and their affiliates. In total, the survey was completed by 17 researchers. 9 participants identified as female, 8 as male. All participants worked in academia, most being PhD students or post-doctoral researchers, and 5 being professors. The participants came from diverse fields such as philosophy, digital studies and media psychology. All came from the humanities or social sciences. Except 2 participants, all had served as a reviewer at least once in the year before the survey. While being a rather small group of participants, the diversity of fields of study represented and the shared academic background make for an interesting snapshot of reading and reviewing in humanities and social sciences. In the following, we will present a selection of results from the survey.

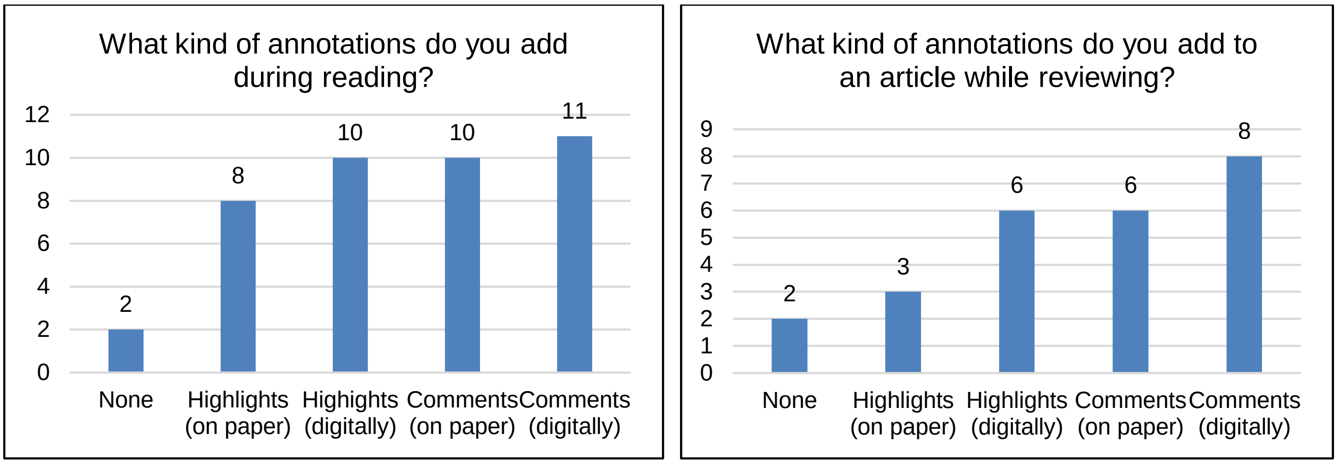


Figure 2: The importance of annotations during review. n=17, multiple answers possible.

As our software CARE is focused on the annotation phase, one of our biggest priorities was to see how common the practice of annotation actually is. The results can be seen in Fig. 2. It is clear that the large majority of participants add some sort of annotations to articles they read and review, whether it be on paper or in digital form. Annotation indeed is an important step in the process of analyzing scientific articles, which validates our approach in CARE. When asking for the problems in the current peer review system, subjective reviews without sufficient standardization and time constraints were mentioned multiple times. With CARE we try to help with both points, by assisting in the education of reviewers and streamlining the process from annotations to the full review report.

Two questions in the survey asked which parts of a paper the participants would read if they only had a short time to assess whether an article is worthy for reading or publishing, respectively. For the hypothetical review under time constraints, many participants mentioned that they would start by reading the methods section. When deciding whether to fully read a paper, several participants would read the abstract and the conclusions first and some would check the bibliography. In a related question, we asked whether the participants follow a typical reading order during reviewing (irrespective of time constraints). The majority read sequentially from start to finish, but some have special orders such as jumping to the conclusions after reading the abstract. The answers to these questions inform the assistance workflows to be implemented in CARE.

A distinguishing feature of CARE is the addition of categories to comments made by the users. The categories can be added either by the users themselves or the integrated AI agent. In the survey, we asked the participants for typical categories they see in structured review reports. Here, a range of possibilities was mentioned such as originality or readability. However, we could not identify a common set of categories, showing that in CARE, the set of categories should be flexible, and adapted to the field of study of the user.

4 Annotation-based Peer Reviewing: User Study

The survey provided theoretical insights to guide the development of CARE. As our second research intervention at CAIS, we carried out a user study of CARE. With this user study, we wanted to get practical feedback on the usability of our software and the focus on annotations as a main part of the reviewing process.

We implemented a simulated peer reviewing scenario. The participants were split into two groups, each of which received a different scientific article from the open science platform F1000Research. Both articles were short, previously unknown to the participants, and from the field of meta-science. They were selected such that they could be understood by most practicing researchers. The participants were asked to individually produce an annotation-based review of the article they received using our software. For this, they were given 40 minutes. They could add highlights, attach comments to highlights and comment on the full article. All comments had to be given a category, which could be strength, weakness, or other (for all comments that did not fit into the first two categories). After the 40 minutes ended, the reviews (in the form of an article, highlights and comments) were exchanged between the groups, such that each participant received a review and an article he or she had not seen before. The participants had to make an "acceptance decision" based on the article and the review under a time constraint of 15 minutes. They were given less time in this phase to have them use the review as much as possible. To make the use of the reviews easier, the participants could switch to a structured report view, in which the review comments were sorted by category. After this practical part of the user study, the participants were given a questionnaire. This questionnaire contained demographic items similar to those in the survey, as well as questions on the user experience and the potential of the use of annotations in peer review.

https://f1000research.com

11 persons participated in the study, 9 of them having reviewed at least one scientific article in the last 5 years. All of them at least sometimes add annotations to an article during reviewing. They were again from a diverse set of fields from the humanities and social sciences.

The participants wrote around 150 categorized comments during reviewing. This data can be used to train and evaluate our AI assistance models, and give insights into the commenting behavior of the participants.

The majority of the participants were satisfied with the ease of use of CARE, and reported that they were able to quickly complete the tasks given in the user study with the help of the software. Most participants also agreed that CARE could help to speed up the reviewing process. However, the majority also thought that the software does not yet have all the functions and capabilities they would expect it to have. This feedback shows that we are on the right track with our software, but that some further development is needed.

We were interested in the potential of using annotations as a form of communication in peer review. Most participants agreed that it was easy for them to put all their thoughts into the comment format. We asked the participants whether annotation-based reviews could replace traditional peer review reports, and obtained mixed feedback. Nevertheless, this is a promising sign towards the usefulness of comments themselves. During an informal feedback round after the user study, several participants mentioned that they liked the structured report function of CARE. However, they found that it would be of better use for the reviewer him- or herself rather than an editor. This is because the editor usually does not know the full text of the reviewed article, and the comments can be hard to understand in isolation. The reviewer can use the structured report as the basis for a full-text review.

5 The Potential of Annotation-based Peer Reviewing

Our research at CAIS showed significant potential in the use of CARE as a software tool for annotation-based reviewing. Our software was well-received in the user study, as the participants found it both easy-to-use and were able to complete the tasks effectively. We could verify our central assumption that annotations are a natural way of interacting with text, and a common step during peer review. Several participants in the survey mentioned problems of current peer reviewing systems regarding time constraints and subjective reviewing. CARE can help make reviewing more streamlined for experienced researchers, and can be used as an educational tool, tackling both problems mentioned. The potential of CARE will fully unfold when the AI-enhanced functionalities of CARE, such as automatic categorization of comments, have come to shape. We are one step closer towards this goal with the comment data collected during the user study, as it will help us to train our AI models. In our upcoming work on CARE, we will add functionalities that help transform the structured report into a full-text review.

Our software can also be applied in scenarios other than academic peer reviewing, such as in high school and undergraduate education, collaborative writing or the collection of annotation data.

6 Acknowledgements

This research was conducted in the context of a fellowship at the Center for Advanced Internet Studies (CAIS). The work has been funded by the German Research Foundation (DFG) as part of the PEER project (grant GU 798/28-1) and by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

Cortes, C., & Lawrence, N. D. (2021). “Inconsistency in conference peer review: Revisiting the 2014 neurips experiment”. *arXiv: 2109.09774*.

Fok, R., Head, A., Bragg, J., Lo, K., Hearst, M. A., & Weld, D. S. (2022). “Scim: Intelligent faceted highlights for interactive, multi-pass skimming of scientific papers”. *arXiv:2205.04561*.

Hua, X., Nikolov, M., Badugu, N., & Wang, L. (2019). “Argument mining for understanding peer reviews”. *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)*, 2131–2137.

Kennard, N., O’Gorman, T., Das, R., Sharma, A., Bagchi, C., Clinton, M., Yelugam, P. K., Zamani, H., & McCallum, A. (2022). “Disapere: A dataset for discourse structure in peer review discussions”. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1234–1249.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). “Bias in peer review”. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17.

Publons. (2018). “Global state of peer review”.

Sandoval, W. A., & Reiser, B. J. (2004). “Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry”. *Science education*, *88*(3), 345–372.