Towards Better Citation Intent Classification

Anonymous ACL submission

Abstract

Accurate classification of citation intents in a scientific article provides deeper contextual understanding of and better quantifies the contributions of cited articles. This improves scientific literature platform capabilities such as search relevance, ranking and more. To our knowledge, we present the most comprehensive survey of Transformer-based language models performance on the citation intent classification task using SciCite dataset. Here, we make three recommendations. Firstly, we propose to report model performance as a distribution in contrast to a single averaged performance value. This arises from our observation that model performance is sensitive to the random seed choice resulting in wide performance variations from multiple finetuning runs. Secondly, this provides practical insights for model selection, showing the model's best possible performance. Thus, we propose that practitioners perform multiple finetuning runs before selecting the best performing model. Thirdly, we propose a simple data augmentation to improve the distribution of model performance overall. Moving forward, we suggest exploring improvements to the finetuning and model selection process as promising future directions.

1 Introduction

Citations are a core part of scientific literature, providing an avenue to acknowledge the various contributions of various scientific articles. Citations are provided with specific intents, such as to provide background information, to present the use of methods or compare results from other works.

Citation intent classification is the task of identifying the intent of a specific citation. By quantifying the distribution of intents of various citations received by a particular paper, we can generate a better understanding of the nature of contributions provided by a paper beyond the citation count. For example, we can infer if a particular paper provides a useful method or result. This is useful in many applications, such as identifying and ranking scientific articles according to the nature of their contributions.

In our paper, we will survey and evaluate the effectiveness of various state-of-the-art language models on the task of citation intent classification. In addition, we propose a data augmentation approach that improves the performance across all the surveyed models. Lastly, we discuss the implications of the results we obtained through our experiments and provide some possible directions for future work.

1.1 Dataset

There are various datasets of scientific literature that contain citation information. Some, such as S2ORC (Lo et al., 2020), are large but do not provide annotations for the citation intents, while smaller datasets such as ACL-ARC (Bird et al., 2008) and SciCite (Cohan et al., 2019) do provide annotations for citation intents. SciCite is an order of magnitude larger than ACL-ARC, containing 6 times the number of citations gathered from about 35 times more scientific articles from the Computer Science and Medical domains.

In this work, we focus on the citation intent classification with **SciCite** as the benchmark dataset. More specifically, the task is to correctly classify citation intents into one of three classes: Background, Method, and Result Comparison.

1.2 Models

In recent years, Transformer-based large language models have been the dominant approach for achieving state of the art results on NLP tasks. One reason for the effectiveness of these models is the ability to learn good language representations by pre-training on a large corpus before undergoing fine-tuning on a task-specific dataset.

In our paper, we will survey the effectiveness of a total of nine Transformer-based language models on the SciCite benchmark dataset. The nine models are as follows:

1. **BERT** (Devlin et al., 2018) (Bidirectional Encoder Representations from Transformers) is a Transformer model that uses a bidirectional self-attention mechanism and pretrained on a large text corpus. It achieves high performance on many NLP tasks via transfer learning.

2. **RoBERTa** (Liu et al., 2019) is based on BERT, with an improved training regime using 10 times more data.

3. **DSP-RoBERTa** (Gururangan et al., 2020) is part of a family of RoBERTa models with further pretraining to adapt them for various target domains.

4. **Biomed-RoBERTa** (Lewis et al., 2020) is a RoBERTa model that is pre-trained on the full texts of about 2.7 million scientific papers from the Semantic Scholar corpus, improving the performance of the model on tasks in the biomedical domain.

5. **SciBERT** (Beltagy et al., 2019) is a BERT model pretrained on scientific text from 1.14M papers in order to handle language processing tasks in scientific field.

6. **PubMedBERT** (Gu et al., 2020) is pretrained using abstracts from PubMed and full text articles from PubMedCentral, achieving state-of-the-art results on various tasks in the biomedical domain.

7. **XLNet** (Yang et al., 2019) is a Transformer-based auto-regressive language model that retains the ability to learn bidirectional contexts.

8. **DeBERTa** (He et al., 2020) is a Transformer model with distangled attention mechanism and enhanced mask decoder replacing the final Softmax layer.

9. **ALBERT** (A Lite BERT) (Lan et al., 2020) is a variant of BERT that utilizes parameter sharing and embedding factorization to reduce the number of parameters compared to an equivalently sized BERT model, although at slightly lower performance.

We used the implementations and pretrained weights of these models through the Hugging Face Transformers library (HuggingFace, 2019). The precise implementations and checkpoints used are listed in the Appendix. Of these nine models, the some are of particular interest due to possible relevance in the corpus used for pretraining, in particular SciBERT, DSP-RoBERTa, and PubMedBERT, which is pretrained on scientific literature. We also test some variations of a subset of the models to investigate the impact of model size and vocabulary casing differences between some variants of the models.

2 Related Work

Previous work on the Sci-Cite dataset was presented in (Cohan et al., 2019), where a BiLSTM model with Attention and ELMo embeddings, along with structural scaffolds, was used to achieve a macro F1-score of 84.0. Following up, (Beltagy et al., 2019) presented fine-tuned BERT and SciBERT models that achieved with F1-score of 84.85 and 85.49 respectively.

3 Methods

3.1 Training Settings

To be consistent with the SciBERT paper (Beltagy et al., 2019) which first performed evaluation of BERT-base and SciBERT on SciCite dataset, we chose the batch size of 32 and fine-tuned all nine models separately with a learning rate of 1e-5. We run each training multiple times and report the mean performance.

3.2 Data Augmentation

We postulate that there is often a correlation between the intent of a citation and the section in which the citation is located. For example, the section "Related Work", a citation is likely to be used for describing the background and positioning, while citation in the "Results" section is likely cited with the intent of result comparison. By adding in the parent section of the citation as additional context, we hypothesize that we can improve the context available to the model to be used for classification. We term this data augmentation by adding in a **section hint**. The section hint takes the form of a sentence, which contains only section title, being added to the start of the original text. Some preprocessing is performed on the raw text of the section headers available in the SciCite dataset to normalize various formatting quirks of the raw data, such as inconsistent capitalization schemes, inclusion of section numbers etc.

Example of augmentation (marked in red):

> Discussion. More examples of contradictory results have been observed in bovines; some reports (Zakhartchenko et al., 2001; Bhuiyan et al., 2004) indicated a significant decrease in blastocyst.

We test this simple form of augmentation as an added experiment during our survey to see if it impacts the various models differently.

4 Results

Our experiment results are reported in **Table 1** below with macro F1-score on both original data and augmented data.

Model	RawData	Aug.Data	Δ
BERT-base	84.80	84.82	+0.02
RoBERTa-base	84.01	84.59	+0.58
DSP-RoBERTa	84.90	85.61	+0.71
Biomed-RoBERTa	86.32	86.63	+0.31
SciBERT	**86.74**	**87.08**	+0.34
PubMedBERT	85.55	86.14	+0.59
XLnet-base	84.59	85.49	+0.90
DeBERTa-base	84.75	85.90	**+1.15**
ALBERT-base	84.03	84.45	+0.42

We find that the 3 best performing models (SciBERT, Biomed-RoBERTa and PubMedBERT) are the ones pretrained on biomedical text, which is the text most relevant to the SciCite dataset. Hence, we have a clear indication on the benefit of pretraining on a relevant corpus.

DeBERTa shows the greatest improvement as a result of the augmented data, while also being, in theory, the most powerful model among those surveyed, however lacking in relevant pretraining. We hypothesize that DeBERTa pretrained on biomedical text would possibly be the best performing model.

SciBERT has the best F1-score for both the raw and augmented data. In fact, our result shows an improvement over the previous reported work, even without the augmented data. This is related to our next observation: In our experiments, we often observe a large variance in model performance, across all models due to different random seeds used in training. We postulate that this could be due to large variations in model performance due to different random seeds, and perform further analysis in Section 5.

We can also see that the data augmentation is effective across all the models surveyed.

5 Analysis and Discussion

In our experiments, we observe large variations in model performance due to different random seeds. Thus, we recorded the performance of our model across 9 training runs per model, and plotted them in a box-and-whisker plot (**Figure 1**) to visualize the variances in performance.

We observe that the original reported F1 score for SciBERT (85.49) falls well within the statistical inter-quartile range for our 9 SciBERT training runs, although our recorded mean and median values are both higher. This means that the original reported score is not unexpected given a different random seed, and in our training runs with more random seeds, we have trained models which on average perform better than what was previously recorded.

In addition, the absolute best performing model (F1-score of 88.5) recorded far exceeds the best average score as reported in **Table 1**. In fact, the best performing model is Biomed-RoBERTa with a particular random seed, and not SciBERT, although it has the best score on average.

From the visualization in **Figure 1**, we can also see that the data augmentation is effective in at least one of the two following ways for every model we tested, resulting in an improved distribution of performance across multiple training runs:

1. To improve the absolute best performance of the model
2. To reduce the variance in performance observed during the model training

We also performed additional analysis on several variations of the model to determine if they have an impact on the task performance.

1. **Cased vs Uncased Models** We studied the difference between cased and uncased versions of BERT-base, BERT-large and SciBERT, finding no clear correlation between the type of casing used and the performance of the model. The performance of the cased and uncased models are visualized in **Figure 2** in Appendix B.

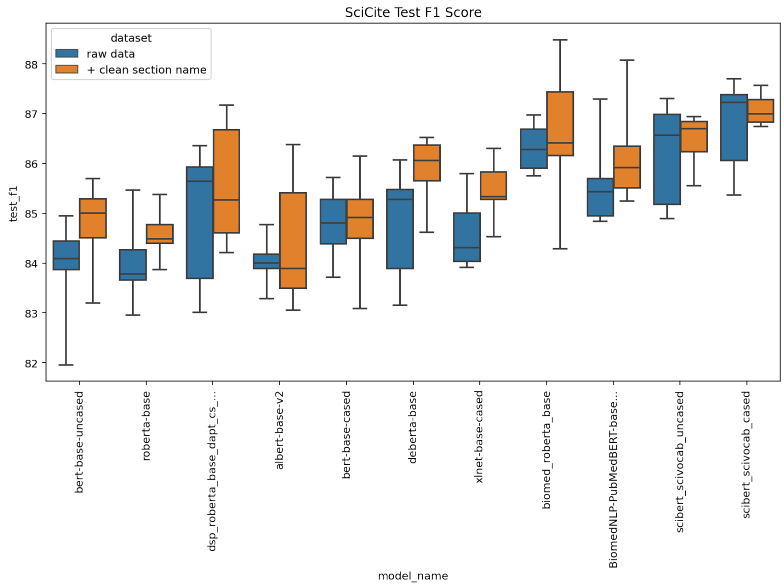

Figure 1: Spread of results.

2. **Base vs Large Models** We studied the difference between base (100M) and large (300M) variants of BERT, RoBERTa, DeBERTa, XLNet and the corresponding variants of ALBERT. We observe a clear relationship between a larger model and improved performance on the SciCite task. This also indicates to us that a larger model pretrained on biomedical tasks would likely perform better than the currently available models. The performance of the cased and uncased models are visualized in **Figure 3** in Appendix B.

6 Recommendations for Practitioners

In this paper, we have arrived at results that allow us to propose the following recommendations for practitioners to achieve better practical performance on the citation intent classification task:

1. Train models multiple times with different random seeds in order to find the best performing model

2. Utilize data augmentation, such as the simple strategy we demonstrated, as it shows measurable improvements across all the models surveyed

3. Report the performance score of models as a distribution, rather than a singular score, in order to provide a better overview of the average and best possible score that can result from a particular model

7 Conclusion and Future Work

In future work, we hope to improve the practical performance on the citation intent classification task. We have outlined a few directions below:

1. Explore more methods and different permutations of data augmentation techniques to add more context into the model input.

2. Introduce methods to perform training with random seeds found using search algorithms that allow us to train better performing models.

3. Create a better pretrained model. Based on our observations, a DeBERTa-large model pretrained on biomedical texts would be a prime candidate for the best performing model when applied to the citation intent classification task.

4. Perform our experiments on other citation intent classification datasets, such as ACL-ARC, to study the transferability of our proposed methods.

In this short paper, we have presented a preliminary data augmentation technique that demonstrates adding more context improves the performance of a model on the citation intent classification task. Our method improve the performance across all the models that we surveyed.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan Gibson, Mark T. Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proc. of the 6th International Conference on Language Resources and Evaluation Conference (LREC'08)*, pages 1755–1759.

Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

HuggingFace. 2019. Pytorch transformer repository. `https://huggingface.co/`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

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 preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

A Model Implementations Used

We used the PyTorch implementations of the following models from the HuggingFace Transformers library, as well as the following model checkpoints from the HuggingFace model hub:

1. **BERT** bert-base-cased and bert-base-uncased
2. **RoBERTa** roberta-base and roberta-large
3. **DSP-RoBERTa** dsp_roberta_base_dapt_cs_tapt_citation_intent
4. **Biomed-RoBERTa** allenai/biomed_roberta_base
5. **SciBERT** allenai/scibert_scivocab_cased and allenai/scibert_scivocab_uncased
6. **PubMedBERT** BiomedNLP-PubMedBERT-base
7. **XLNet** xlnet-base-cased and xlnet-large-cased
8. **DeBERTa** microsoft/deberta-base and microsoft/deberta-large
9. **ALBERT** albert-base-v2 and albert-large-v2

The training was performed on a single V100 32GB GPU with automatic mixed precision enabled.

B Additional diagrams

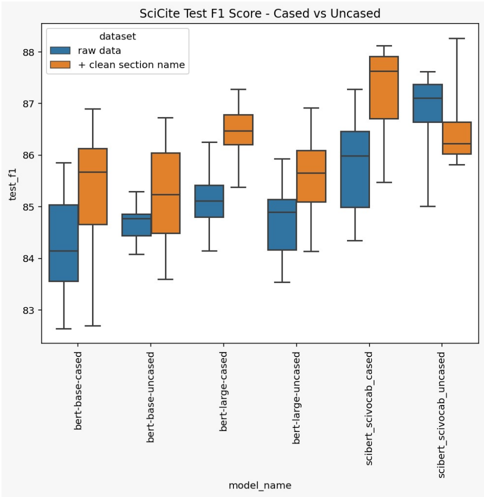

Figure 2: Spread of results.

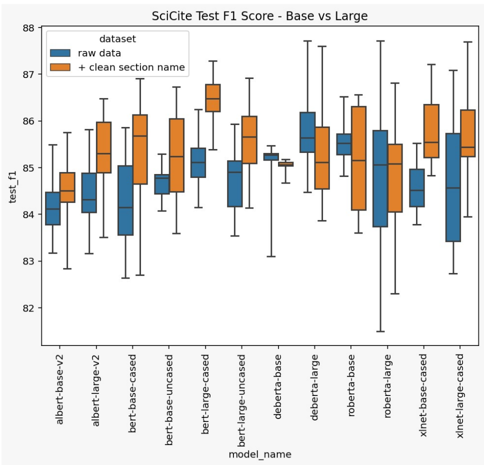

Figure 3: Spread of results.