SciBERT Sentence Representation for Citation Context Classification

Himanshu Maheshwari[1] **, Bhavyajeet Singh**[1] **, Vasudeva Varma**[2]
IIIT Hyderabad
[1]{himanshu.maheshwari,bhavyajeet.singh}@research.iiit.ac.in
[2]vv@iiit.ac.in

Abstract

This paper describes our system (IREL) for 3C-Citation Context Classification shared task of the Scholarly Document Processing Workshop at NAACL 2021. We participated in both subtask A and subtask B. Our best system achieved a Macro F1 score of 0.26973 on the private leaderboard for subtask A and was ranked one. For subtask B our best system achieved a Macro F1 score of 0.59071 on the private leaderboard and was ranked two. We used similar models for both the subtasks with some minor changes, as discussed in this paper. Our best performing model for both the subtask was a finetuned SciBert model followed by a linear layer. We provide a detailed description of all the approaches we tried and their results. The code can be found here

1 Introduction

Recent years have witnessed a massive increase in the amount of scientific literature and research data being published online, providing revelation about the advancements in different domains. The introduction of aggregator services like CORE has enabled unprecedented levels of open access to scholarly publications. The availability of the full text of the research documents facilitates the possibility of extending the bibliometric studies by identifying the context of the citations (Pride and Knoth, 2017). The shared task organized as part of the SDP 2021 focuses on classifying citation context in research publications based on their influence and purpose. Subtask A aims at identifying the purpose of the citation. Subtask A involves a multiclass classification of citations into one of six classes: Background, Uses, Compare and Contrast, Motivation, Extension, and Future. Subtask B aims at identifying the importance of the citation. It is a binary classification of citations into one of two classes: Incidental and Influential.

2 Related Work

In this section, we briefly describe other work in this and closely related areas. Several machine learning-based frameworks have been developed in the past which attempt to classify scientific citation text based on its context.

Before the machine learning era, Garfield (1965) introduced the idea of identifying the reasons for a particular citation and presented a list of 15 reasons why an author would include a citation. There have been other works and discussions about the classification of citations and its importance in this classification in developing models of referencing (Moravcsik and Murugesan, 1975; Chubin and Moitra, 1975).

However, despite these works, the literature on automating the task of classifying citations has been limited. Garzone and Mercer (2000) treated the citation classification as a task of sentence categorization. They extracted a sentence that incorporated citations and then applied manually curated lexical and grammar rules to assign categories to the citations. Teufel et al. (2006) used supervised machine learning techniques to classify citations into 12 categories. The authors annotated 2,829 citation contexts from 116 articles, using linguistic features, including the cue phrases. Agarwal et al. (2010) used algorithms like LDA and SVM to develop an automated citation classifier specific to the domain of biomedical text classification.

Cohan et al. (2019) proposed structural scaffolds, a multitask model to incorporate structural information of scientific papers into citations for effective classification of citation intents. Their model did not rely on external linguistic resources or hand-engineered features as done in previous methods. They also introduced a new dataset of citation intents (Sci-Cite) which covers multiple scientific domains. A pretrained language model based on BERT(Devlin et al., 2018), SciBERT (Beltagy

et al., 2019) was introduced, which leverages unsupervised pretraining on a large multi-domain corpus of scientific publications and can be used for multiple downstream tasks like this one.

de Andrade and Gonçalves (2020) tackled the same problem of classifying the citations based on purpose and influence. Their solution relies on combining different, potentially complementary, text representations to enhance the final obtained results. A combination of TF-IDF (capturing statistical information), LDA (capturing topical information), and Glove word embeddings (capturing contextual information) was used for the task of classifying the context of the citation. (Kunnath et al., 2020) presented an overview of all approaches used for the previous edition of this shared task, highlighting the data distribution and the results achieved, and has been used as a baseline for our further work.

3 Methodology

BERT-based models have shown significant performance in different NLU tasks. This paper uses sentence representation from the BERT-based model and experiments with different machine learning and deep learning models.

The training data contains nine different fields viz. unique identifier, COREID of citing paper, citing paper title, citing paper author, cited paper title, cited paper author, citation context, citation class label, citation influence label. For a machine learning or deep learning solution, we identify that only the citation context is significant. We also experiment by adding cited paper title information with citation context and observe a drop in Macro F1 score. Thus, for all our experiments, unless specified otherwise, we use citation context only.

3.1 Transfer learning with different BERT-based model

BERT model has shown significant performance in different NLU tasks like sentiment analysis (Xie et al., 2019), hate speech detection (Mathew et al., 2020), etc. The RoBERTa (Liu et al., 2019) model showed an improvement over BERT for different NLU tasks. Different works (Beltagy et al., 2019; Chalkidis et al., 2020; Nguyen et al., 2020) pretrain a BERT-like model from scratch using large-scale in-domain data to learn the domain-specific language patterns. Beltagy et al. (2019) trained the BERT model from scratch using scientific documents sand saw an improvement in different NLU tasks involving scientific documents.

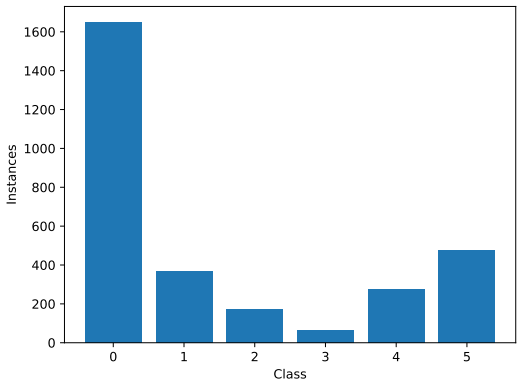

Figure 1: Class distribution for Subtask A.

We finetune BERT, RoBERTa, and SciBERT on our training data. We use a linear layer for the classification of sentence embeddings that we got from these models. Experiment results show that SciBERT-uncased performed the best, and thus for our further experiments, we used SciBERT to get our sentence representations. As already noted, we use citation context information.

3.2 SciBERT with Bi-LSTMs

LSTMs have shown good performance for various text classification task (Badjatiya et al., 2017; Tang et al., 2015). So, we finetune, SciBERT on our training data. We use a Bi-LSTM model for the classification of sentence embeddings that we got from the SciBERT model. The entire model is finetuned end-to-end.

3.3 Random Forest

The training data size is small, so we tried the Random Forest model with a smaller number of parameters to train. A max tree depth of 40 and forest size of 1200 gave decent results. The sentence embeddings that we got from the SciBERT model were used.

3.4 Cited Title with Citation Context

We have argued that citation context is the only relevant field for our deep learning models. To verify that, we include cited paper title information in the model's input. We concatenate the sentence representation of the cited paper title and the citation context and present it as input to our model. We finetune SciBERT with a linear layer as described in section 3.1.

Subtask A		Subtask B	
Model	**Macro F1**	**Model**	**Macro F1**
BERT-uncased + Linear	**0.4350 ± 0.00439**	BERT-uncased + Linear	0.6611 ± 0.0065
RoBERTa + Linear	0.4258 ± 0.0264	RoBERTa + Linear	0.6636 ± 0.0038
SciBERT-cased + Linear	0.4232 ± 0.0157	SciBERT-cased + Linear	0.6720 ± 0.0041
SciBERT-uncased + Linear	**0.4333 ± 0.0094**	SciBERT-uncased + Linear	**0.6778 ± 0.0098**
SciBERT-uncased + Bi-LSTM	0.4246 ± 0.01267	SciBERT-uncased + Bi-LSTM	0.6741 ± 0.0101
Random Forest	0.2742	Random Forest	0.6559
Title + Citation Context	0.4232 ± 0.01486	Title + Citation Context	**0.6781 ± 0.0077**

Table 1: Results of subtask A and subtask B.

Weighted Loss Function	Macro F1
True	0.4333 ± 0.0094
False	0.4146 ± 0.0133

Table 2: The weighted loss function performed better than weighted loss function for Subtask A.

4 Dataset

The labeled training dataset contains 3000 instances. The training data includes nine different fields viz. The unique identifier, COREID of citing paper, citing paper title, citing paper author, cited paper title, cited paper author, citation context, citation class label, citation influence label. We randomly divided 500 instances into validation data and used the remaining 2500 instances for training the model. It was verified that the validation and training data shared similar class distribution. During inference, we trained the model by combining training and validation data.

Fig 1 shows the class distribution for sub-task A and sub-task B. The data is imbalanced, so we use the weighted loss function. The balanced heuristic is inspired by (King and Zeng, 2001). We use sklearn implementation to compute class weight[1]. The weighted loss function outperformed the unweighted loss function for the model described in section 3.1, as shown in Table 2. So for all the tasks, we use the weighted loss function.

5 Results and Discussion

Table 1 shows the results for both the subtasks. As already noted, we use the weighted loss function for Subtask A. We see that for subtask A, BERT-uncased + Linear performed better than SciBERT-uncased + Linear, however on the actual test data, the latter outperformed the former. So for all our further experiments, we used SciBERT-uncased. Similarily for subtask B, for our validation set, using citation title gave the best results, however on the actual test data, this model performed very poorly, so we do not use citing title information.

Finetuning BERT or BERT like model is high variance (Phang et al., 2018). So, we experimented with the same model thrice and reported the mean macro-F1 score and standard deviation.

For subtask A, SciBERT-uncased + Linear performed the best on the private leaderboard and achieved a Macro F1 score of 0.26973 and was ranked one. For subtask B, SciBERT-uncased + Linear performed the best on the private leaderboard, achieved a Macro F1 score of 0.59071, and was ranked two.

We used grid search to find the optimal hyperparameter. For subtask A, we found the best learning rate to be 10^{-5}, batch size as four, and dropout to be 0.1. Optimizing hyperparameter offered significant performance gain. For subtask B, we found the best learning rate to be 10^{-5}, batch size as eight, and dropout to be 0.1. We used cross-entropy loss (weighted for subtask A and unweighted for subtask B), sigmoid activation function, and Adam optimizer.

6 Conclusion and Future Work

This work shows how SciBERT embeddings can capture nuances of scientific documents, and simple deep learning and machine learning models could give competitive results. Despite a small dataset, good results could be achieved.

There is much noise in the training data, like numbers, HTML elements, etc. A more detailed study of features of the training data could improve the results. Since the dataset is small, data augmentation techniques could be used as future work.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

References

S. Agarwal, L. Choubey, and H. Yu. 2010. Automatically classifying the role of citations in biomedical articles. *Annual Symposium proceedings. AMIA Symposium*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. *CoRR*, abs/1706.00188.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: the muppets straight out of law school. *CoRR*, abs/2010.02559.

Daryl E. Chubin and Soumyo D. Moitra. 1975. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, 5(4):423–441.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *CoRR*, abs/1904.01608.

Claudio Moisés Valiense de Andrade and Marcos André Gonçalves. 2020. Combining representations for effective citation classification. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 54–58, Wuhan, China. Association for Computational Linguistics.

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.

E. Garfield. 1965. Can citation indexing be automated. *Statistical association methods for mechanized documentation, symposium proceedings*, Vol. 269:189–19.

Mark Garzone and Robert E. Mercer. 2000. Towards an automated citation classifier. In *Advances in Artificial Intelligence*, pages 337–346, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163.

Suchetha Nambanoor Kunnath, David Pride, Bikash Gyawali, and Petr Knoth. 2020. Overview of the 2020 WOSP 3C citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 75–83, Wuhan, China. 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.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *CoRR*, abs/2012.10289.

Michael J. Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science*, 5(1):86–92.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

David Pride and Petr Knoth. 2017. Incidental or influential? ? a decade of using text-mining for citation function classification. In *16th International Society of Scientometrics and Informetrics Conference*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Target-dependent sentiment classification with long short term memory. *CoRR*, abs/1512.01100.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, page 103–110, USA. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation. *CoRR*, abs/1904.12848.