On the Use of Language Models for Function Identification of Citations in Scholarly Papers

Tomoki Ikoma[1] and **Shigeki Matsubara**[1,2]
[1]Graduate School of Informatics, Nagoya University, Japan
[2]Information Technology Center, Nagoya University, Japan
ikoma.tomoki.d0@s.mail.nagoya-u.ac.jp

Abstract

Citation graphs represent the citation relations between papers, and they are commonly used by researchers to identify relevant papers. However, citation graphs do not always represent how papers are related to each other. To make more effective use of citation graphs to discover relevant papers, we can consider identifying the functions of citations and label each edge in the citation graphs with its function. This paper proposes a method to identify the functions of citations automatically. The proposed model utilizes language models, e.g., SciBERT, to identify the description of citation functions. However, the language models are limited in terms of the number of input tokens; thus, the entire citing paragraph cannot be processed at once. To overcome this problem, we analyzed the distribution of the descriptions of citation functions in the citing paragraphs and determined the focusing part in identifying the citation functions. Experiments conducted on scientific paper data demonstrated the effectiveness of the proposed method.

1 Introduction

Scientific papers cite publications for various reasons, and the connections between papers are established through citations. In addition, citation graphs[1] represent citations in a graph structure, and they are commonly used by researchers to identify relevant papers. However, the edges in citation graphs only represent the citations between papers; thus, citation graphs do not always represent how papers are related to each other. To make more effective use of citation graphs in order to discover relevant papers, we can consider identifying the functions of citations and label each edge in the citation graphs with its function[2].

Thus, in this paper, we propose a method to identify the functions of citations automatically based on the text of citing paragraphs. The proposed model utilizes language models, e.g., SciBERT (Beltagy et al., 2019), to identify the citation functions. However, the language models are limited in terms of the number of input tokens; thus, the entire citing paragraph cannot be processed at once. To overcome this problem, we analyzed the distribution of the descriptions of citation functions in the citing paragraphs and determined the focusing part in identifying citation functions. Experiments conducted on scientific paper data demonstrated the effectiveness of the proposed method.

2 Datasets for Citation Function Identification

A previous study (Teufel et al., 2006) published the first dataset for the citation function identification task. They manually annotated 548 citation instances extracted from 161 papers in the computational linguistics domain as one of the 12 classes of citation functions. However, their dataset suffered from several limitations, e.g., the small data size and the coverage of only one research domain. Despite these issues, no new datasets were created for years due to various difficulties, including the definition of labeling schema and the annotation of gold labels (Kunnath et al., 2022b).

Recently, several new datasets for the citation classification task, e.g., ACL-ARC (Jurgens et al., 2018) and SciCite (Cohan et al., 2019), have been created and made available to the public. The ACL-ARC dataset comprises approximately 2,000 citation instances extracted from papers in the ACL Anthology, where each instance is labeled as either one of the six classes of citation functions, i.e., `background`, `compares_contrasts`, `extension`, `future`, `motivation` and `uses`. The SciCite dataset contains approximately 11,000 citation instances sampled from papers in the computer science and medical domains with class labels of either `background`, `method`, or `result`.

[1] https://citationgraph.org/

[2] For citations via URLs, (Tsunokake and Matsubara, 2022) proposed a method to identify the function of citations.

Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications, pages 130–135

In addition, a large and diverse dataset has been created for the Citation Context Classification Shared Task (Kunnath et al., 2020, 2021). The shared task provided a dataset of 3,000 citation instances sampled from papers in various domains. Here, each citation instance was labeled by the authors of the citing papers under the same schema as ACL-ARC. However, the classification labels were annotated at each author's own discretion; thus, the consistency of the labels over the entire dataset was not guaranteed.

3 Task Definition and Data Analysis

3.1 Task Definition

We propose a method to automatically identify the citation functions based on the text containing citations. Specifically, from a given paragraph containing citations, we propose a method to extract the part that describes why the target paper was cited, and classify the described citation function into one of the eight categories[3]: `background`, `motivation`, `uses`, `extends`, `similarities`, `differences`, `compare/contrast`, and `future work`.

3.2 FOCAL Dataset

In this study, we used the dataset from the Function Of Citation in Astrophysics Literature (FOCAL) shared task (Grezes et al., 2023). The FOCAL dataset comprises of 2,421 training examples, 606 validation examples, and 821 test examples extracted from papers in the astrophysics domain, and each example contains the paragraph text and the single or multiple positional information of the target citation. Training examples also include the positional information and the class label of the descriptions of the citation functions for data analysis and model training. Note that some examples have multiple spans that describe the citation function, and the class label is annotated on each span in such cases.

3.3 Data Analysis

We analyzed the class label distribution of the citation functions and the positional relations between citation tags and citation function descriptions in the training set.

[3]A detailed explanation of each category is available at `https://ui.adsabs.harvard.edu/WIESP/2023/LabelDefinitions`.

Table 1: Number of examples with each class of citation function. Note that the sum of each row does not match the number of training examples, because some examples are labeled with more than one citation function class.

Function class	Number of examples	
Background	1,098	45.35%
Motivation	161	6.65%
Uses	605	24.99%
Extends	7	0.29%
Similarities	202	8.34%
Differences	87	3.59%
Compare/Contrast	400	16.52%
Future work	27	1.12%

Table 2: Percentage of sentences containing descriptions of citation functions. For preceding sentences, cases with no sentences before the citing sentence are excluded. For following sentences, cases with no sentences after the citing sentence are excluded.

	Inclusion percentage	
Preceding sentences	285/2,419	11.78%
Citing sentences	2,436/2,464	98.86%
Following sentences	258/2,419	10.66%

3.3.1 Distribution of Citation Function Label

Table 1 shows the number of examples labeled for each class of citation function. As can be seen, the most frequent class is background representing approximately 45% of the analyzed examples. In contrast, other classes, e.g., extends and future work, include low number of examples.

3.3.2 Positional Relation with Citation Tags

We analyzed the positional relations between the citation tags and the citation function descriptions. Here, we initially split each paragraph into sentences using the NLTK sentence tokenizer (Bird et al., 2009), and then we extracted the citing sentences and their preceding and following sentences. Next, we computed the percentage of sentences containing the descriptions of the citation functions for the citing, preceding and following sentences.

Table 2 shows the percentage of sentences containing the descriptions of the citation functions. As shown, the preceding and following sentences contained descriptions of citation functions approximately 10% of cases. In contrast, only 28 citing sentences without citation function descriptions were found. These results indicate that the citing sentences almost always contain descriptions of

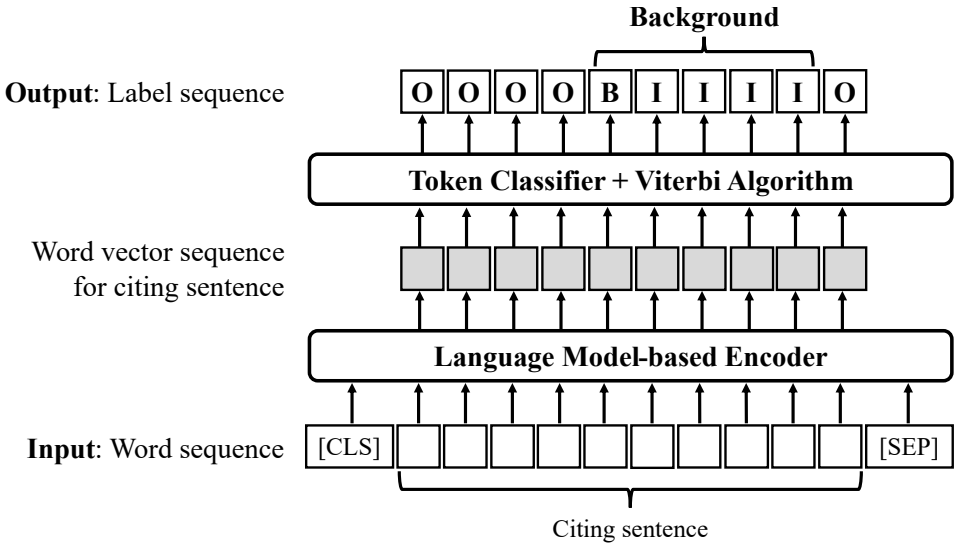

Figure 1: Structure of the proposed model

the citation functions; however, it is rare for such descriptions to extend to nearby sentences.

4 Method

4.1 Model Structure

The proposed method utilizes language models, e.g., SciBERT (Beltagy et al., 2019) to identify the citation functions as shown in Figure 1. The model comprises an encoder and a token classifier, and it identifies the citation functions as follows:

1. Convert the input text to a sequence of words and add [CLS] and [SEP] at the start and end of the sequence.

2. Transform the words in the citing sentence to feature vectors using the encoder.

3. Output a BIO tag sequence that indicates whether each word is the beginning, inside or outside of the span describing the citation function with the token classifier. Here, the Viterbi algorithm (Forney, 1973) is employed to avoid generating invalid sequences, e.g., sequences where I follows O.

4. Generate a class label of citation functions for each subsequence starting with B.

4.2 Range of Input Text

Note that the citing paragraph cannot be processed at once by language models, e.g., SciBERT, due to the limitation in the number of input tokens; thus, we must determine which part of a paragraph to focus on prior to inputting the text to the model. When training the model, the focusing part can be determined as the sentence containing the annotated span of the citation function descriptions and the m preceding and n following sentences. However, such annotations of the span of the citation function descriptions are not given at the time of prediction. Thus, based on the results of the analysis described in Section 3.3.2, we determine the focusing part for prediction as the citing sentence and the m preceding and n following sentences.

Table 3: Citation function identification performance of different language models

Model	Word accuracy	Exact match
SciBERT	**68.13**	34.14
RoBERTa	67.33	33.70
ALBERT-v2	67.01	32.06
DeBERTa-v3	67.76	**35.02**

5 Experiment

5.1 Selection of Language Models

We compared the performance of several language models on identifying the citation functions. Here, we trained the SciBERT (Beltagy et al., 2019), RoBERTa (Liu et al., 2019), ALBERT-v2 (Lan et al., 2019), and DeBERTa-v3 (He et al., 2023) models on 85% of the FOCAL training data as the training subset, and we evaluated each model on the remaining 15% of the data as the development subset. During the training process, we fine-tuned

Table 4: Citation function identification performance with different input text window sizes

Input text window		Evaluation metrics		
prev(m)	next(n)	Full	Generic	Labels
0	0	51.11	78.03	64.82
0	1	51.26	75.49	66.23
0	2	49.23	74.16	64.61
0	3	49.60	74.83	65.15
1	0	50.87	79.82	64.57
1	1	50.93	76.61	64.75
1	2	49.08	76.45	64.72
2	0	**51.97**	**79.99**	64.23
2	1	49.73	74.52	65.96
3	0	51.90	79.13	**67.10**

each language model on 2,677 sentences containing citation function descriptions in the training subset over 30 epochs. At the end of each epoch, we evaluated the trained models by the word-based labeling accuracy on 454 sentences in the development subset and saved the best model.

Table 3 shows the performance of each model evaluated by the word-based accuracy and sentence-based exact match rate on the development subset. As can be seen, the best word-based accuracy was achieved by the SciBERT, and the best sentence-based exact match rate was obtained by the DeBERTa-v3 model.

5.2 Selection of Input Text Window Size

We searched for the best setting for the focusing part in the citing paragraphs by training the SciBERT with different settings. Following the experimental setup presented in the literature (Kunnath et al., 2022a), we set the number of m preceding and n following sentences. Here, for each m and n value, we fine-tuned the SciBERT model on the sentence containing citation function descriptions, m preceding sentences, and n following sentences in the training subset. We then saved the best model over 30 epochs and evaluated this model on the development subset in terms of the following metrics.

Full F1 score that considers the predictions to be correct if both of the predicted placement and class labels are correct.

Generic F1 score that considers the predictions to be correct if the predicted placement is correct.

Table 5: Experimental result on validation data

	Full	Generic	Labels
Baseline	23.68	59.86	42.87
Proposed model	**54.08**	**79.92**	**65.94**

Labels F1 score that considers the predictions to be correct if the predicted class label is correct.

The evaluation results are shown in Table 4. As can be seen, model performance was improved by extending the focusing part to the preceding sentences; however, extending the focusing part to the following sentences did not contribute performance improvement.

5.3 Final Evaluaiton

Based on the results of the experiments discussed in Sections 5.1 and 5.2, we fine-tuned the SciBERT model over 30 epochs on the sentences containing the citation function descriptions and 3 preceding sentences in the training subset. At the end of each epoch, we evaluated the performance of the model according to the word-based accuracy on the development subset and saved the best model. Then, on the FOCAL validation and test data, we identified the citation functions using the trained model. On the validation data, we compared the performance to a baseline that always predicts the description of the citation function as the citing sentence and labels as background.

Table 5 shows the experimental results obtained on the validation data. As shown, the proposed model exhibited better results for all three evaluation metrics compared to the baseline, which indicates the effectiveness of the proposed model.

On the testing data, the proposed model achieved the scores of 51.97 Full, 73.00 Generic and 69.44 Labels.

6 Conclusion

This paper has proposed a method to identify the functions of citations automatically based on the text of citing paragraphs. The proposed method utilizes the SciBERT model to identify the citation function based on the citing sentences and nearby sentences under the assumption that citation functions are described near the citation. Experiments conducted on scientific paper data demonstrated the effectiveness of the proposed method.

Table 6: Performance of identifying sentences containing citation function descriptions

Citing sentences (365 examples)			
	Precision	Recall	F1 score
Baseline	99.18	100.00	99.59
SciBERT	99.39	90.06	94.49
Non-citing sentences (4,153 examples)			
	Precision	Recall	F1 score
Baseline	0.00	0.00	0.00
SciBERT	0.73	10.87	1.36
Overall (4,518 examples)			
	Precision	Recall	F1 score
Baseline	99.18	79.74	88.40
SciBERT	19.74	74.01	31.17

Limitations

The proposed method assumes that the function of the citation is always described in the citing sentence and its surrounding sentences, while sentences distant from the citing sentence do not contain descriptions of citation functions. Thus, the proposed method cannot extract descriptions of citation functions for cases where the citation function is described in text distant from the citing sentence. Although we uniformly determined the part of the citing paragraph to focus on experimentally, the part of the citing paragraph to focus on should be dynamically determined.

To decide the focusing part of the citing paragraph, we can consider using language models, e.g., SciBERT, to identify sentences that are likely to contain descriptions of the citation function. To evaluate the effectiveness of this strategy, we trained SciBERT to predict whether a given sentence is likely to contain description of the citation function and evaluated the performance of the trained model. Here, for training, we split the citing paragraphs in the training subset into sentences using the NLTK sentence tokenizer and used sentences containing descriptions of the citation functions as positive examples, and sentences without description of citation functions were used as negative examples. Then, the model was trained using all positive examples and 10% of randomly sampled negative examples and evaluated in terms of precision, recall and F1 score on the development subset. To better understand of the model's performance, we computed the evaluation metrics for citing and non-citing sentences separately, and we compared this model to a baseline that always predicts citing sentences to contain descriptions of the citation functions.

Table 6 shows evaluation results. As can be seen, the performance of the trained model for non-citing sentences was very poor. In addition, the overall performance was considerably worse than that of the baseline. These results indicate that the predictions are influenced greatly by whether each given sentence is a citing sentence; thus, classifying sentences with language models is not an effective method to identify the sentences containing the descriptions of citation functions.

Acknowledgments

This work was partially supported by the Grant-in-Aid for Challenging Research (Exploratory) (No. 23K18506) of JSPS and by JST SPRING, Grant Number JPMJSP2125. The computation was carried out on supercomputer "Flow" at Information Technology Center, Nagoya University.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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 3586–3596. Association for Computational Linguistics.

G.D. Forney. 1973. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Felix Grezes, Thomas Allen, Tirthankar Ghosal, and Sergi Blanco-Cuaresma. 2023. Function of citation in astrophysics literature (FOCAL): Findings of the shared task. In *Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications*. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The 11th International Conference on Learning Representations*.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

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

Suchetha Nambanoor Kunnath, David Pride, Drahomira Herrmannova, and Petr Knoth. 2021. Overview of the 2021 SDP 3C citation context classification shared task. In *Proceedings of 2nd Workshop on Scholarly Document Processing*, pages 150–158. Association for Computational Linguistics.

Suchetha Nambanoor Kunnath, David Pride, and Petr Knoth. 2022a. Dynamic context extraction for citation classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 539–549. Association for Computational Linguistics.

Suchetha Nambanoor Kunnath, Valentin Stauber, Ronin Wu, David Pride, Viktor Botev, and Petr Knoth. 2022b. ACT2: A multi-disciplinary semi-structured dataset for importance and purpose classification of citations. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3398–3406. European Language Resources Association.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv:1909.11942[cs.CL]*.

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:1907.11692[cs.CL]*.

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*, pages 103–110. Association for Computational Linguistics.

Masaya Tsunokake and Shigeki Matsubara. 2022. Classification of URL citations in scholarly papers for promoting utilization of research artifacts. In *Proceedings of the 1st Workshop on Information Extraction from Scientific Publications*, pages 8–19. Association for Computational Linguistics.