Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022, pages 1–7
Marseille, 24 June 2022

Automatic Approach for Building Dataset of Citation Functions for COVID-19 Academic Papers

Setio Basuki, Masatoshi Tsuchiya
Department of Computer Science and Engineering, Toyohashi University of Technology
1–1 Hibarigaoka, Tempaku-cho, Toyohashi 441-8580, Aichi, Japan
{setio,tsuchiya}@is.cs.tut.ac.jp

Abstract

This paper develops a new dataset of citation functions of COVID-19-related academic papers. Because the preparation of new labels of citation functions and building a new dataset requires much human effort and is time-consuming, this paper uses our previous citation functions that were built for the Computer Science (CS) domain, which consists of five coarse-grained labels and 21 fine-grained labels. This paper uses the COVID-19 Open Research Dataset (CORD-19) and extracts 99.6k random citing sentences from 10.1k papers. These citing sentences are categorized using the classification models built from the CS domain. The manually check on 475 random samples resulted accuracies of 76.6% and 70.2% on coarse-grained labels and fine-grained labels, respectively. The evaluation reveals three findings. First, two fine-grained labels experienced meaning shift while retaining the same idea. Second, the COVID-19 domain is dominated by statements highlighting the importance, cruciality, usefulness, benefit, consideration, etc. of certain topics for making sensible argumentation. Third, discussing State of The Arts (SOTA) in terms of their outperforming previous works in the COVID-19 domain is less popular compared to the CS domain. Our results will be used for further dataset development by classifying citing sentences in all papers from CORD-19.

Keywords: citation function, citing sentence, COVID-19, state of the art.

1. Introduction

Citation functions represent the reason why authors of academic papers cite previous works. Valenzuela et al. (2015) stated that the citations should not be treated equally. This is because citations indicate different roles, e.g., introducing the background, comparing and contrasting between studies, using or extending of existing methods, criticizing the previous works, etc. The existence of citations plays an important role in the preparation of a research manuscript since it helps the authors understand the big picture of a topic (Qayyum & Afzal, 2018), position their proposed research in the broad literature (Lin & Sui, 2020), and show their research novelty (Tahamtan & Bornmann, 2019). Moreover, citations can indicate the quality of proposed research (Casey et al., 2019; Raamkumar et al., 2016). Therefore, providing appropriate citations requires serious attention to support research dissemination.

There is a continuous development in designing labels for Rhetorical Structures (RS) and building datasets in the medical domain. Existing works have designed RS and developed the dataset (Alliheedi et al., 2019; Dayrell et al., 2012; Dernoncourt & Lee, 2017; Green, 2015; Jia, 2018; Kim et al., 2011; Liakata, 2010; Shatkay et al., 2008; Wilbur et al., 2006). However, several issues appear in these works. The first issue is that not all these RS were developed based on full text papers; several works built the RS using only papers' abstracts. The second issue is that most of the RS were not specifically designed for *citing sentences* (i.e., sentences which contain citation marks). Since the existing RS covers both *citing sentences* and non-*citing sentences*, the number of labels is considered small, which causes several potential missing *citation functions* being accommodated—the last issue. Moreover, due to the COVID-19 pandemic, the number of published papers covering this topic has significantly increased. Existing RS is not designed specifically for this purpose, and this has become an additional issue. Considering this, we aim to develop a new dataset of *citation functions* that contains more detailed labels, covers full text papers, and is specific for the COVID-19 domain.

Designing new labels of citation functions and building a new dataset is challenging. This is because we need to provide large, labeled training data, which is time-consuming, expensive, and requires much human effort. To obtain the labeled instance with less effort, this paper uses our previous labels of *citation functions* that have been built based on Computer Science (CS) papers. Note that the process to design the labels was accomplished prior to and has not become part of this paper. The developed labels consist of five *coarse-grained* labels and 21 *fine-grained* labels. By using these labels, we obtained classification models with accuracies of 83.6% and 90.1% for *coarse-grained* labels and fine-grained labels, respectively. This paper uses these models to categorize *citing sentences* on COVID-19-related papers obtained from the COVID-19 Open Research Dataset (CORD-19) (Lu Wang et al., 2020).

Through completing this research, we deliver several contributions:

- The automatic classification of *citation functions* on COVID-19 domain achieved accuracies of 76.6% for *coarse-grained* label and 70.2% for *fine-grained* labels.
- The experimental results show that several *fine-grained* labels experienced a meaning shift (the expansion of the labels' definition while remaining in the same idea).
- The COVID-19 domain is dominated by statements highlighting the importance, cruciality,

usefulness, benefit, consideration, etc. of certain topics for making sensible argumentation.
- We noticed that discussing State of The Arts (SOTA) in terms of outperforming previous studies in the COVID-19 domain is less popular compared to the CS domain.
- Lastly, we released a final dataset consisting of 99.6k labeled *citing sentences*[1].

This paper uses two main terms: *citing paper*, which is used to define the paper citing other papers, and *cited paper*, which is used to define the papers cited by the *citing paper*.

2. Dataset Development

This section describes how our proposed dataset of *citation functions* is developed. The dataset consists of several parts, the first of which concerns the obtainment of data sources of COVID-19-related papers. The second part describes the labels of *citation functions*, and the last part builds the dataset of *citation functions* on COVID-19 domains.

2.1 COVID-19-related Papers

This paper uses a collection of papers from the COVID-19 Open Research Dataset (CORD-19) (Lu Wang et al., 2020). Initially, this dataset provided 28k papers. The present number of papers has significantly increased during the continuous development. The CORD-19[2] collected papers from several sources (e.g., PubMed Central (PMC), PubMed, and the World Health Organization's COVID-19 Database). Moreover, it contains a collection from preprint servers such as bioRxiv, medRxiv, and ArXiv. This paper uses the latest version of the dataset **(version: 2021-12-20)** from JSON parsed from the full text of 314,391 (PDF) and 243,652 (PMC) papers. The distribution of CORD-19 is shown in Figure 1.

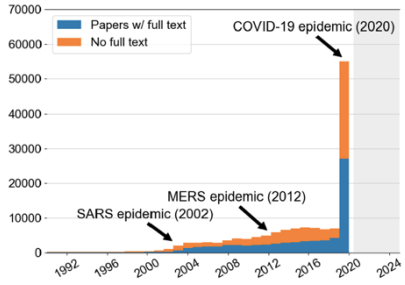

Figure 1: The paper distribution in CORD-19. The *x* axis depicts year, and the *y* axis depicts the number of papers. This figure is taken from Lu Wang et al.'s work (2020).

2.2 The Labels of Citation Functions

The labeling scheme of *citation functions* used in this paper is obtained from our previous research. The dataset used to develop the scheme is obtained from Färber et al. (2018), which provided 90,278 papers from ArXiv in the CS domain from January 1993 until December 31, 2017. Since this data source provides all parsed sentences from research papers, we perform *filtering* to separate the *citing sentences* and non-*citing sentences*. The *filtering* is performed using regular expressions based on certain citations tags. In this stage, we obtained around 1.6 million instances of *citing sentences*.

Our proposed labels of *citation functions* are developed through three steps: top-down analysis, bottom-up analysis, and annotation experiment. While the top-down analysis reviews the definitions of the labels from existing works, e.g., *background, usage,* and *comparison,* the bottom-up analysis is performed to identify the *citing sentence* patterns on the dataset. At this point, we obtained an initial dataset consisting of 5,669 samples. Next, we conducted the pre-annotations experiment to develop and finalize both labels of citation guidance and the annotation guidance. The final labels themselves consist of two categories: five *coarse-grained* labels and 21 *fine-grained*. The *coarse-grained* labels represent the generic idea of the *citation functions*, which are divided into *background* for stating certain topics, *citing paper work* for focusing on what is done by author, *cited paper work* to show what has been done by previous works, *compare and contrast* to discuss the similarity between the citing paper and the cited paper, and *other* for all categories that do not match the above criteria. To obtain more specific functions, these *coarse-grained* categories are broken down into *fine-grained* categories.

The annotation step was performed by two annotators who have master's degrees in Computer Science. They were provided with annotation guidance which covers an introduction to the task, labeling examples, and checking mechanism for the annotators' understanding. For the real experiment, the annotators are supplied with an excel spreadsheet that consisted of 421 random *unlabeled citing sentences*. Our experiment shows that the inter-annotator agreement shows 88.59% for *coarse-grained* labels and 72.44% for *fine-grained* labels. Moreover, Cohen's Kappa shows 0.85 for *coarse-grained* labels and 0.71 for fine-grained labels.

Coarse-grained Label: Background
Describes the *citing sentences* referring to the theory, principle, concept, topic, problem, etc. from cited papers.
Fine-grained Label: • **(atr0) definition**: explains the definition of general theory, principle, concept, topic, problem, etc. *example:* Gianna <citation> is a precursor visual environment for modeling CSP. • **(atr1) suggest**: provides the reader with suggestions to refer, see more details, and explore other cited papers. *example:* The interested reader may dig deeper on this subject by referring to <citation>. • **(atr2) judgment**: highlights the positivity/negativity, usefulness/non-usefulness, etc. of concepts, topics, problems, etc. *example*: The n-coalescent has some interesting statistical properties <citation>. • **(atr3) technical**: explains how a theory, principle, concept, topic, problem, etc. is applied. *example:* The WMF model <citation> learns the latent factors by preserving the personalized rankings. • **(atr4) trend**: explains the significance of the research topic, theory, principle, concept, topic, problem, etc. *example:* CNN has been gaining attention and is now used in many text classification tasks <citation>.
Coarse-grained Label: Citing Paper Work
What is proposed by the author?
Fine-grained Label:

[1] https://github.com/tutcsis/COVID-19

[2] https://www.semanticscholar.org/cord19/download

• **(atr5) corroboration**: while proposing a research topic, citing paper cites cited paper. *example:* To do this we build upon the concept of continuous regression <citation>. • **(atr6) based on**: states that the citing paper follows, considers, is built based on, or is inspired by the cited paper. *example:* Here we follow closely the definition of GPs given by <citation>. • **(atr7) use**: cites paper use, implements, employs, or adopts the concept, dataset, technique, etc. *example:* The proof systems we use were originally defined in <citation> which is the presentation we follow. • **(atr8) extend**: the citing paper extends, adapt, improves, adds, or modifies the cited paper's work. *example:* Our proposed method (multiCCA) extends the bilingual embeddings of <citation>. • **(atr9) dominant**: the citing paper outperforms the cited paper. *example:* Our PredNet model outperforms the model by <citation>. • **(atr10) future**: mentions the future plan of the citing paper. *example:* In fact, we plan in the future of reproducing all the algorithms in Common2 <citation>, in that spirit.
Coarse-grained Label: Cited Paper Work
What is done by the cited papers?
Fine-grained Label: • **(atr11) propose**: describes the proposed research by the cited paper. *example:* <citation> used CCA to learn bilingual lexicons from monolingual corpora. • **(atr12) success**: highlights the success of the cited paper. *example:* <citation> successfully extracts body appearance and topology from synthetic and real input. • **(atr13) weakness**: highlights the weakness of the cited paper. *example:* The limitation of <citation> is that the traffic is assumed always cross directional. • **(atr14) result**: describes the result of the cited paper (neutral). *example:* In 1994, Kosaraju <citation> reported another solution to this problem. • **(atr15) dominant**: states the superiority of the cited paper when compared to the citing paper. *example:* However, <citation> performs better than our method on class accuracy.
Coarse-grained Label: Compare and Contrast
The citing paper and the cited paper are compared and contrasted.
Fine-grained Label: • **(atr16) compare**: describes the similarity between citing and cited papers. *example:* Recent work by Xia <citation> is independent from, and closely related to, our work. • **(atr17) contrast:** describes the differences between citing and cited papers. *example:* However, unlike <citation> we did not observe an increased convergence speed.
Coarse-grained Label: Other
This label is prepared for *citing sentences* that do not match all criteria.
Fine-grained Label: • **(atr18) comparison**: comparison between cited papers (whether similarities or differences between them). *example:* This idea was first proposed by Google <citation> and was then further developed by <citation>. • **(atr18) multiple_intent**: *citing sentences* have two or more citation marks for different intents. *example:* It is noteworthy that while <citation> fared better than our system with the SemEval data, our system outperformed <citation> on the OEC dataset. • **(atr18) other**: this label is designed for *citing sentences* that do not meet all of the label categories described above. *example:* The first one is due to Valtr <citation>.

Table 1: The labels of *citation functions* on CS domains.

2.3 Dataset Development in COVID-19 Domains

The proposed dataset of COVID-19 domains is built using an automatic approach by following several steps. **The first step** is preparing the source of the papers. In this step, we do a simple data analysis to gather a deep understanding of the parsed JSON structures of CORD-19. Following this, the analysis is accompanied by *filtering* to select only *citing sentences*. **The second step** is classifying all extracted *citing sentences* using the best models obtained from the dataset of the CS domain. These models were obtained by experimenting with several machine learning (ML) approaches such as Logistic Regression and Deep Learning based on Long Short-Term Memory (LSTM) architecture. Considering the limitations of available instances, we consider using pre-trained word embedding that is both non-contextual, such as Glove (Pennington et al., 2014), word2vec (Mikolov et al., 2013), and fasttext (Bojanowski et al., 2017) and contextual such as the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and work by Beltagy et al. (2019). **The last step** is verifying the automatically labeled *citing sentences* by performing a random selection of 475 instances and checking the predicted labels manually.

3. Experiment Results

This section explains the results of the automatic classification used to build a dataset of *citation functions* of the COVID-19 domain. The results are divided into several parts: brief information about classification models developed using the CS domain, the automatic classification of *citing sentences* of the COVID-19 domain, and the evaluation of classification through a manual label check. Note that, the classification in the CS domain and the COVID-19 domain is done through two stages, namely the *filtering* stage and the *fine-grained* stage. While the *filtering* stage is used to classify the *citing sentences* into two categories, i.e., *Other (atr18)* and *No-Other (atr0-atr17)*, the *fine-grained* stage is applied to classify the *citing sentences* belonging to *No-Other* class into 18 *fine-grained* classes. Finally, the proportional distribution of labeled instances and a discussion of results are also presented.

3.1 Classification Results for CS Domain

Here, we demonstrate the best results for classifications in the CS domain. In the *filtering* stage, BERT and SciBERT showed identical accuracies of 90.12%, as shown in Table 2. To achieve these accuracies, both methods used different settings as shown in Table 3.

Methods	Accuracy	Macro avg precision	Macro avg recall	Macro avg f1
BERT	90.12	71.58	**85.15**	75.99
SciBERT	**90.12**	**74.53**	82.72	**77.73**

Table 2: The best results on filtering stage.

Techniques	Parameters
BERT	$2e^{-5}$; batch 64; imbalance
SciBERT	$3e^{-5}$; batch 32; balance

Table 3: Best parameters on the filtering stages.

In the *fine-grained* stage, the best result was obtained by using SciBERT by 83.64%, as shown in Table 4 and the hyperparameters is shown in Table 5.

Methods	Accuracy	Macro avg precision	Macro avg recall	Macro avg f1
BERT	80.95	80.98	82.40	81.06
SciBERT	**83.64**	**83.46**	**85.35**	**84.07**

Table 4: The best results on fine-grained classification.

Techniques	Parameters
BERT	$3e^{-5}$; batch 32; imbalance
SciBERT	$3e^{-5}$; batch 32; balance

Table 5: Best parameters on the fine-grained classification.

3.2 Dataset on the COVID-19 Domain

The classification experiment is conducted on 99.6k instances generated from 10.1k parsed paper files (JSON format). The automatic classification begins with the extraction of all the sentences in the JSON files. Next, all extracted sentences are filtered to keep only *citing sentences*. Similar to the dataset on the CS domain, the classification is then applied by following two classification stages, namely the *filtering* stage and the *fine-grained* stage. To measure the accuracy of labeled instances, we perform a manual label check on 25 random samples for each label, for a total of 475 samples (18 fine-grained labels + 1 other label).

After completing the manual label check, we obtained accuracies 76.63% and 70.20% for *coarse-grained* labels and *fine-grained* labels, respectively. The accuracy of *coarse-grained* labels is easily obtained by summing the proportion of correctly and wrongly *fine-grained* labels. Since each label in the *fine-grained* labels has the same number of instances, it is easy to use the confusion matrix to compare each label's accuracy, as shown in Figure 2. The highest number of correctly predicted labels is achieved by the label *technical,* with 24 correct predictions and only a single incorrect prediction. In contrast, the label *cited_paper_dominant* has the lowest number of correctly predicted labels with only nine correct and 16 incorrect predictions.

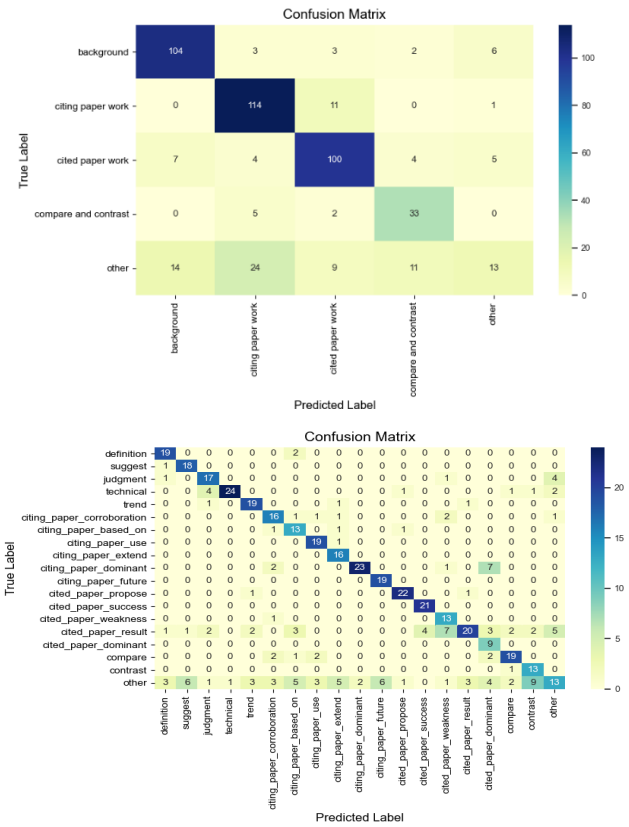

Figure 2: Confusion Matrix of manually label checking for (top) *coarse-grained* labels and (bottom) *fine-grained* labels.

Fine-grained Labels	Number of Instances		Label Proportion	
	CS Domain	COVID-19 Domain	CS Domain	COVID-19 Domain
definition	55,508	3,151	4.18%	3.77%
suggest	51,987	355	3.91%	0.42%
judgment	215,428	**37,885**	16.21%	**45.34%**
technical	85,374	5,557	6.42%	6.65%
trend	66,594	6,579	5.01%	7.87%
citing_paper_corroboration	113,488	2,571	8.54%	3.08%
citing_paper_based_on	55,878	531	4.20%	0.64%
citing_paper_use	115,215	1,114	8.67%	1.33%
citing_paper_extend	28,779	241	2.17%	0.29%
citing_paper_dominant	24,823	294	1.87%	0.35%
citing_paper_future	5,439	424	0.41%	0.51%
cited_paper_propose	**243,031**	5,442	**18.29%**	6.51%
cited_paper_success	34,505	2,128	2.60%	2.55%
cited_paper_weakness	15,054	1,072	1.13%	1.28%
cited_paper_result	154,394	15,063	11.62%	18.03%
cited_paper_dominant	**3,215**	**31**	**0.24%**	**0.04%**
compare	39,364	677	2.96%	0.81%
contrast	20,909	439	1.57%	0.53%
Total	**1,328,985**	**83,554**	**100%**	**100%**

Table 6: The distribution comparison of automatically labeled instances in CS domain and COVID-19 domain. The comparison consists of two parts: (a) the number of instances on each label and (b) the proportion of instance on each label to the total instances in the dataset.

Applying classification models built from CS papers to COVID-19 related papers results in two consequences. The first consequence is that there is a decrease of *fine-grained* label accuracy from 83.64% in CS domain to 70.2% in COVID-19 domain. The second consequence is that two *fine-grained* labels experienced a meaning shift: the label *citing_paper_dominant* and the label *citing_paper_future*. The definition of the label *citing_paper_dominant* changed

from expressing the *citing paper's* performance over *cited paper* to discussing the success of *citing paper*, with or without comparison. On the other hand, the definition of the label *citing_paper_future* changed from stating the future plan of the *citing paper* to a general recommendation without specifying whether it is done by *citing paper* or *cited paper*.

3.3 Citation Functions Distribution

To give more analysis on the current COVID-19 dataset, in Table 6 we show a comparison of the distribution datasets in the CS domain and COVID-19 domains. Note that, the distribution in this table represents the number of automatically labeled *citing sentences* in the datasets. The current dataset in this paper consists of 99,691 labeled instances, of which *No-Other* label has 83,554 instances and the *Other* label 16,137 instances. Since the labels of *citation functions* are designed for CS papers, it is worth determining whether the classification models are effective for domains related to COVID-19. Instead of using the number of instances to compare both datasets, this paper uses the proportion of labels as indicators due to the datasets having different sizes.

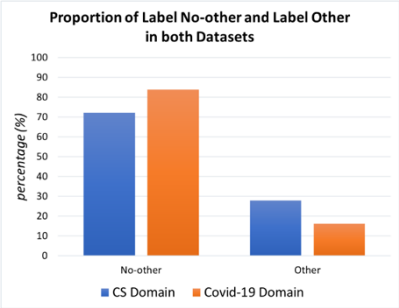

Figure 3: Proportion Comparison between *No-Other* and *Other* labels.

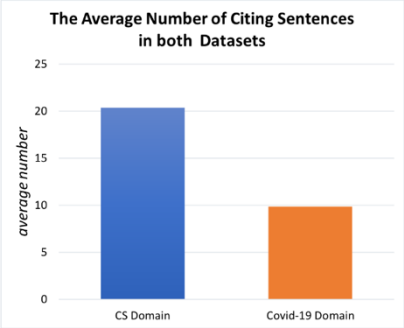

Figure 4: The average number of *citing sentences* in each paper

First, the comparison is done on the *filtering* stage to show the percentage of *No-Other* vs *Other* labels as depicted in Figure 3. In this figure, it is seen that both domains share the same trend in that the proportion of *No-Other* label much higher than *Other* label. Surprisingly, the label *judgment* in the COVID-19 domain has a proportion of almost half at 45.34%. In second place, the label *cited_paper_result* has 18.03% of proportion. The rest of labels constitute less than 10% of the proportion. Furthermore, there are eight labels have only under 1% of the proportion, with the lowest proportion obtained by the label *cited_paper_dominant* with 0.04%, which is equivalent with 31 instances. The CS domain faces a similar situation in that this label has the lowest proportion at 0.24%. However, this proportion is not as severe as in the COVID-19 domain. In the dataset of CS domain, the distribution trend is varied among labels, and no single label exceeds 20% of the proportion.

Another comparative indicator between both domains is the average number of *citing sentences* in each paper. Figure 4 demonstrates that the CS domain has a higher number of citing sentences than the COVID-19 domain. To be more specific, the dataset of CS domain consists of 1,840,815 *citing sentences* extracted from 90,278 papers, while the dataset of COVID-19 domain contains 99,691 *citing sentences* extracted from 10,102 papers.

3.4 Discussion

The experiments conducted in this paper reveal several notable findings. The first finding is a phenomenon of meaning shift in two *fine-grained* labels. This corroborates the assertion that even as this paper achieves acceptable accuracies, there still exists an issue regarding the labels' compatibility between two domains. Next, the large proportion of label *judgment* (constituting almost half of dataset) indicates that *citation functions* in the COVID-19 papers are dominated with statements highlighting the importance, cruciality, usefulness, benefit, consideration, etc. of certain topics for making sensible argumentation. Conversely, the smallest proportion, represented by the label *cited_paper_dominant,* which is followed by several labels with proportions less than 1% (e.g., *compare, citing_paper_extend, contrast, citing_paper_dominant,* and *citing_paper_based_on*) indicates that discussing State of the Arts (SOTA) in the COVID-19 domain is less popular compared to the CS Domain. This trend is supported by the average number of *citing sentences* in the CS domain being higher than in the COVID-19 domain, which emphasizes the fact that discussing the SOTA needs more *citing sentences* and *cited papers*.

4. Conclusion

This paper developed the dataset of *citation functions* using *citing sentences* extracted from COVID-19 related papers. Instead of designing new labels of *citation functions* from scratch and preparing training data, this paper uses our previously developed labels and applied the best ML models that have been built from the CS domain. The experiments show that the application of labels of the CS domain to the COVID-19 domain is promising. Furthermore, the evaluation for obtaining the automatic labeling accuracies uncovers several notable patterns such as label compatibility between two domains, the dominant citation roles on each domain, and the relation between a *citing paper* and the SOTA. For future work, we intend to apply the labels and the models to all papers in the CORD-19 dataset.

5. Acknowledgements

This research is supported by the Toyohashi University of Technology – Japan and Amano Institute of Technology Scholarship.

6. Bibliographical References

Alliheedi, M., Mercer, R. E., & Cohen, R. (2019).

Annotation of Rhetorical Moves in Biochemistry Articles. *Proceedings of the 6th Workshop on Argument Mining*, 113–123. https://doi.org/10.18653/v1/W19-4514

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. https://doi.org/10.18653/v1/D19-1371

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Casey, A., Webber, B., & Glowacka, D. (2019). A Framework for Annotating 'Related Works' to Support Feedback to Novice Writers. *Proceedings of the 13th Linguistic Annotation Workshop*, 90–99. https://doi.org/10.18653/v1/W19-4011

Dayrell, C., Candido, A., Lima, G., MacHado, D., Copestake, A., Feltrim, V. D., Tagnin, S., & Aluisio, S. (2012). Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 1604–1609.

Dernoncourt, F., & Lee, J. Y. (2017). *PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts*. 308–313. http://arxiv.org/abs/1710.06071

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Färber, M., Thiemann, A., & Jatowt, A. (2018). A high-quality gold standard for citation-based tasks. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1885–1889. https://www.aclweb.org/anthology/L18-1296

Green, N. (2015). Identifying Argumentation Schemes in Genetics Research Articles. *Proceedings of the 2nd Workshop on Argumentation Mining*, 12–21. https://doi.org/10.3115/v1/w15-0502

Jia, M. (2018). Citation Function and Polarity Classification in Biomedical Papers. *The University of Western Ontario*. https://ir.lib.uwo.ca/etd/5367/

Kim, S. N., Martinez, D., Cavedon, L., & Yencken, L. (2011). Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics*, *12*(SUPPL. 2). https://doi.org/10.1186/1471-2105-12-S1-S5

Liakata, M. (2010). Zones of conceptualisation in scientific papers: a window to negative and speculative statements. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 1–4. https://www.aclweb.org/anthology/W10-3101

Lin, K. L., & Sui, S. X. (2020). Citation Functions in the Opening Phase of Research Articles: A Corpus-based Comparative Study. *Corpus-Based Approaches to Grammar, Media and Health Discourses*, 233–250. https://doi.org/https://doi.org/10.1007/978-981-15-4771-3_10

Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., … Kohlmeier, S. (2020). CORD-19: The Covid-19 Open Research Dataset. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Conference on Neural Information Processing Systems*. https://papers.nips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Qayyum, F., & Afzal, M. T. (2018). Identification of important citations by exploiting research articles' metadata and cue-terms from content. *Scientometrics*, *118*, 21–43. https://doi.org/10.1007/s11192-018-2961-x

Raamkumar, A. S., Foo, S., & Pang, N. (2016). Survey on inadequate and omitted citations in manuscripts: A precursory study in identification of tasks for a literature review and manuscript writing assistive system. *Information Research*, *21*(4). http://informationr.net/ir/21-4/paper733.html

Shatkay, H., Pan, F., Rzhetsky, A., & Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, *24*(18), 2086–2093. https://doi.org/10.1093/bioinformatics/btn381

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. In *Scientometrics* (Vol. 121). Springer International Publishing. https://doi.org/10.1007/s11192-019-03243-4

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. *Association for the Advancement of Artificial Intelligence (AAAI)*.

Wilbur, W. J., Rzhetsky, A., & Shatkay, H. (2006). New directions in biomedical text annotation: Definitions, guidelines and corpus construction. *BMC Bioinformatics*, *7*, 1–10. https://doi.org/10.1186/1471-2105-7-356

7. Language Resource Reference

Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., … Kohlmeier, S. (2020). CORD-19: The Covid-19 Open Research Dataset. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.