Task Definition and Integration for Scientific-Document Writing Support

Hiromi Narimatsu[1], **Kohei Koyama**[2], **Kohji Dohsaka**[3], **Ryuichiro Higashinaka**[1]
Yasuhiro Minami[2], **Hirotoshi Taira**[4]
[1]NTT Communication Science Laboratories
[2]The University of Electro-Communication
[3]Akita Prefectural University, [4]Osaka Institute of Technology
{hiromi.narimatsu.eg, ryuichiro.higashinaka.tp}@hco.ntt.co.jp
k1710245@edu.cc.uec.ac.jp, dohsaka@akita-pu.ac.jp
minami.yasuhiro@is.uec.ac.jp, hirotoshi.taira@oit.ac.jp

Abstract

With the increase in the number of published academic papers, growing expectations have been placed on research related to supporting the writing process of scientific papers. Recently, research has been conducted on various tasks such as citation worthiness (judging whether a sentence requires citation), citation recommendation, and citation-text generation. However, since each task has been studied and evaluated using data that has been independently developed, it is currently impossible to verify whether such tasks can be successfully pipelined to effective use in scientific-document writing. In this paper, we first define a series of tasks related to scientific-document writing that can be pipelined. Then, we create a dataset of academic papers that can be used for the evaluation of each task as well as a series of these tasks. Finally, using the dataset, we evaluate the tasks of citation worthiness and citation recommendation as well as both of these tasks integrated. The results of our evaluations show that the proposed approach is promising.

1 Introduction

When writing a scientific paper, it is important to search for relevant papers and cite them appropriately. However, despite the importance of this requirement, the recent sharp increase in published scientific papers is making it difficult for researchers to comprehensively carry out this process. Consequently, much work has been devoted to developing systems that support the writing of scientific papers.

For example, some studies have attempted to summarize papers on a particular subject (Teufel and Moens, 2002; Qazvinian and Radev, 2008; Bai et al., 2019). The creation of knowledge graphs of scientific papers has also been proposed (Dessì et al., 2020), and Gábor et al. (2018) proposed an automatic content-analysis method by extracting the semantic relations of entities in abstracts.

Other studies have focused on citation recommendation (Huang et al., 2014; He et al., 2010) and generation of citation text (Xing et al., 2020; Luu et al., 2020). Using the database of PubMed[1] papers, Bhagavatula et al. (2018) proposed recommending citations on the basis of keywords as well as the contents of a paper. Mohammad et al. (2009) proposed the generation of citation text, and Färber et al. (2018) proposed a classification model for the task of judging whether a sentence requires citation (citation worthiness).

Although many reports have been presented and an abundance of effort has been expended on data creation (Färber and Jatowt, 2020; Kardas et al., 2020; Saier and Färber, 2020), each previous study has focused on a particular problem in scientific-writing support and has been performed independently using its own specific dataset. Therefore, we do not yet know whether these investigations can be successfully pipelined nor how to ascertain the overall performance of a system that can comprehensively recommend citations. Consequently, it is currently impossible to verify that the technologies centered around scientific-paper writing are actually helpful in comprehensively supporting real-world scientific-paper writing.

In this paper, we first define a series of tasks related to scientific-paper writing that can be pipelined. Then, we create a dataset [2] of academic papers that can be used for the evaluation of each task in scientific-paper writing as well as a series of these tasks. Finally, using the dataset, we evalu-

[1] https://pubmed.ncbi.nlm.nih.gov/

[2] Our dataset is available at https://github.com/citation-minami-lab/citation-dataset.

Proceedings of the Second Workshop on Scholarly Document Processing, pages 18–26
June 10, 2021. ©2021 Association for Computational Linguistics

ate the individual tasks of citation worthiness and citation recommendation as well as the integrated task composed of these two individual tasks. Experimental results show that our task setting and the dataset can be successfully used for scientific-paper writing support.

2 Handling "Related Work" Section

In a scientific paper, the section generally called "Related Work" is important for situating one's research in the field and clarifying the new contribution of the proposed work. However, the task of writing the Related Work section is time-consuming because one needs to read through many papers in related areas and carefully cite them. Due to this cost, much work has been directed to improving the efficiency of this process.

At the beginning stages of this line of research, we saw many studies aimed at helping authors understand the gist of a paper, that is, preparing a summary of the paper highlighting important points such as objective, problem, and methods (Teufel and Moens, 2002). There have also been studies that consider how a paper is cited in summarizing the paper in question (Qazvinian and Radev, 2008). The summarization of scientific papers continues to be an important research focus (Yasunaga et al., 2019). However, capturing the summarization of a particular paper in isolation would obviously not produce a universal solution when facing the abundance of papers that are available to readers.

Recent years have seen an increase in work related to citation recommendation, and this work has been greatly aided by the availability of large-scale article data in electronic form. Such studies have mainly focused on the papers that one should cite due to their authority and relevance based on keywords (Ren et al., 2014). Recently, some studies have focused on recommending papers that might be overlooked by limiting the scope to authority and relevance. Such methods utilize a citation network and more fine-grained content similarity, making it possible to identify specific papers that should be cited (Chakraborty et al., 2015; Bhagavatula et al., 2018). Moreover, Ali et al. (2020) proposed a method for citation recommendation by categorizing relevant papers on the basis of their data, methods, and problems. In our approach, we list tasks related to scientific-paper writing and include the task of citation recommendation. We show how this task can be combined with other tasks as well as how individual problems in citation recommendation can be combined.

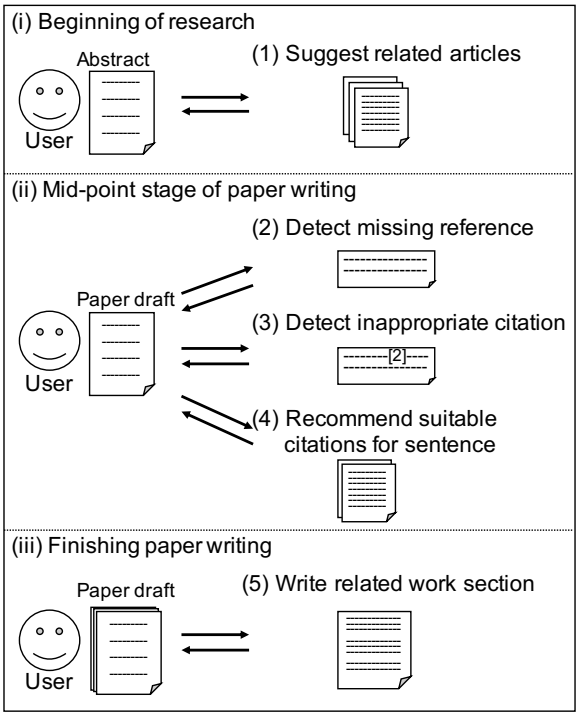

Figure 1: Scientific-paper writing support for each phase of research

In order to facilitate paper writing and peer review, the task of citation worthiness, that is, detecting whether a sentence requires citation, has been carried out (Färber et al., 2018). Färber et al. also released a dataset of scientific papers for this particular task. Other studies have generated citation texts given a portion of a Related Work section. Mohammad et al. (2009) used a rule-based technique to generate citation texts using, as templates, the sentences of the same author and the generic sentences that can be used for citation.

We believe today's high research activity related to handling citation has the potential to create technologies that can actually be useful for human support; however, we also believe that these studies need to be combined appropriately for them to be useful. This has motivated us to list up tasks related to citation and to create a dataset that enables us to evaluate combined tasks as well as individual ones.

3 Listing of tasks

We listed tasks related to citation that can cover the main phases in scientific-paper writing: (i) when

we conceive an idea, (ii) when we obtain research results and are ready to situate the work, and (iii) when we finalize the Related Work section.

Figure 1 shows how scientific-paper writing can be supported in each phase. At (i), which happens at the beginning of research, one is apprehensive that the conceived idea may not be original and thus feels the need to perform a survey of related work. In this situation, it is desired that the scientific-paper writing support system recommends relevant papers (Fig. 1 (1)) using as input the research problem and its approach, which are typically written in the paper's abstract. At (ii), which is done in the mid-point stage of paper writing, there may be cases when citations are not appropriate or missing. Therefore, it is necessary to provide support for the detection of missing references (Fig. 1 (2)), the detection of inappropriate citations (Fig. 1 (3)), and the recommendation of suitable citations (Fig. 1 (4)). At (iii), which is when the author applies finishing touches to the paper, support for tailoring the Related Work section would be appropriate (Fig. 1 (5)), such as how the references should be categorized and how they should be presented.

Tasks (1)–(5) in the figure can be broken down into more fine-grained tasks as follows:

(1)-1: Citation extraction Given an abstract, the task of citation extraction retrieves relevant papers from a large database of scientific documents.

(1)-2: Citation recommendation for draft paper Given a draft paper comprising an abstract plus some body text, this task presents the list of relevant papers retrieved from a large database of scientific documents.

(2): Citation worthiness Given a sentence in the Related Work section of a draft, this task detects whether the sentence needs citations.

(3)-1: Citation allocation Given sentences in the Related Work section of a draft and the body of relevant papers, this task allocates appropriate papers to the sentences.

(3)-2: Sentence-citation pair classification Given a sentence and its possible citation, this task classifies whether the allocation of the citation is appropriate for that paper. This is a sub-task of (3)-1.

(4): Citation recommendation for sentence In (2) and (3), there may be sentences with missing citations, that is, when the sentence requires citation but the allocation of citations has failed. In such a case, a citation needs to be retrieved from a large body of scientific papers. This task performs citation recommendation for a citation-missing sentence. Note that this task focuses only on the sentences suggested as citation-worthy by the citation worthiness task because these tasks form a pipeline.

(5)-1: Citation categorization Given sentences with citations, this task categorizes them based on their underlying themes so that the citations can be more appropriately organized.

(5)-2: Citation sentence generation Given sentences with citations, this task suggests alternative citation text for the sentences to achieve better clarity and fluency.

(5)-3: Citation text generation Given Related Work text, which includes multiple sentences with citations, this task suggests alternative citation text for the content. This task is different from (5)-2 in that the text of the entire Related Work section is generated instead of simply generating a sentence for a citation.

As can be seen in the above listing, the tasks follow the chronological order of how a paper is written in its research phases. They can be pipelined. These tasks have mostly been identified and tackled in previous studies, but they have been researched separately. The list of tasks includes (3)-2, which we newly conceived in this work; in pipelining the paper-writing process, we considered this a useful sub-task for citation allocation.

4 Data Creation

After having defined the tasks, we created a dataset for the evaluation of the individual tasks and, moreover, the integrated (pipelined) tasks. For this purpose, we use the same data as source.

4.1 Procedure

The process of data creation is depicted in Figure 2. We first extract key materials from a *target paper* in an archive of published papers. The target

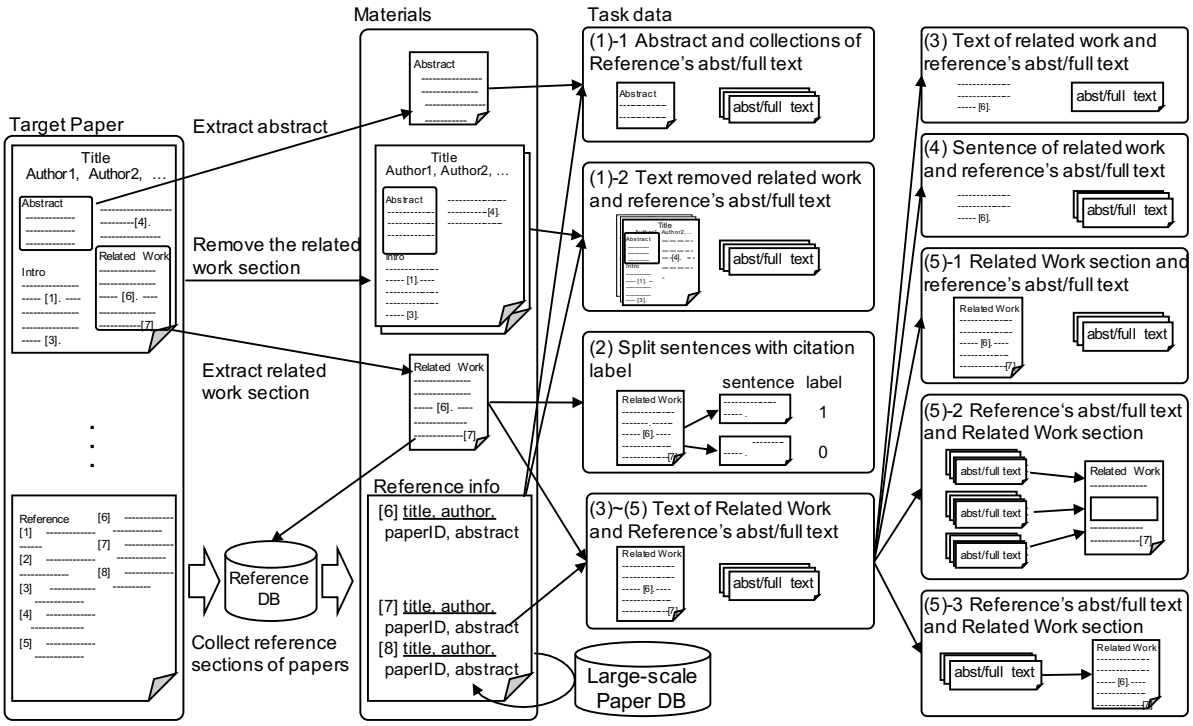

Figure 2: Data-creation process

paper, which is an arbitrary paper in the archive, is the starting point for creating the dataset. We extract elements from the target paper or remove certain elements from it so that we can simulate the incomplete versions of the paper as they appear during the research phases. First, from the target paper, we extract four key elements: abstract, paper without Related Work, Related Work, and references. These can be created directly by extracting certain parts of the target paper. As for the references, we refer to a large-scale reference database to extract their paper IDs (e.g., arXiv paper ID), abstracts, and full text (if retrievable) by using the title and author names. On the basis of these key materials, we create task data for the tasks listed in the previous section. In the following, we describe the detailed process for creating the task data for tasks (1)–(5).

(1)-1: Citation extraction We use the abstract and the list of references for the target paper. Since this is the initial phase of research, we remove from the abstract those sentences related to experiments and results with a rule-based extractor, using the resulting text as the input for this task. The references become the gold data to be retrieved from a large-scale paper database.

(1)-2: Citation recommendation for draft paper We use the paper without the Related Work section as input and use the references as gold output.

(2): Citation worthiness We use the sentences in the Related Work section as task data. We create task data by coupling each sentence with a label indicating whether that sentence has a citation.

(3)-1: Citation allocation We use the text of related work and the references' abstracts and full text (if available) for this task data. We extract sentences from the text of related work and retain only those sentences with citations. Then, we couple these sentences with their citations to create the task data. Although it is not done in this paper, surrounding sentences can also be included in the task data because such sentences may contain helpful information and can serve as context.

(3)-2: Sentence-citation pair classification For the sentences with citations, we create pairs of a sentence with the gold citation and also a pair containing a sentence with an incorrect citation taken from the references of the target paper.

(4): Citation recommendation for sentence We use the sentences of the Related Work section and its references as task data. As gold citations, we use those in the Related Work section. The task is to accurately retrieve the references from the large-scale paper database.

(5)-1: Citation categorization We use the Related Work section and the references as task data. We first extract paragraphs from the text and identify the clusters of citations by extracting citations from each paragraph. The task is to correctly allocate citations to each paragraph.

(5)-2: Citation sentence generation We use the references cited in the Related Work section and the sentences in the Related Work section. For each sentence with a citation, the task is to generate a sentence from the cited reference with its abstract/full text.

(5)-3: Citation text generation We use the references with abstract/full text and the entire text of Related Work. The task is to generate a complete Related Work section using the reference information.

4.2 Created dataset

In this work, we created the task data for (2), (3)-1 and (3)-2. We created a dataset for these tasks because we wanted to verify our approach within a minimal setting; these tasks can be tackled with only the related work sections and the abstracts of the papers cited in them, without requiring the large-scale paper DB or the papers' full texts. Using data covering multiple tasks, we can at least verify whether it is possible to evaluate the performance of individual tasks as well as the integrated task. Although we created the data for the subset of the listed tasks, as can be seen, the procedure for creating task data is mostly straightforward. Once we have verified our approach, as we do in this paper, we will be able to construct data covering all tasks.

To create the data, we first collected target papers from the AxCell dataset (Kardas et al., 2020), which has been made public for the purpose of leaderboard generation. AxCell contains approximately 100K papers.

Since we need papers having a Related Work section, we extracted papers with section titles such as "Related work" and "Related studies." As a result, we successfully obtained 34,416 papers. The sentences included in the Related Work sections of these papers become the task data for (2) Citation worthiness. Table 1 shows the statistics. The numbers of total, positive, and negative examples of the task data for (2) are shown in the first row. We first randomly split the papers into three sets having 22,416, 6,000, and 6,000 target papers. Then we made train/dev/test sets by extracting sentences from these sets. The test data are used for testing throughout the following tasks in order to guarantee a fair evaluation. The inclusion relationship among datasets is shown in Figure 3.

Next, from the target papers used for the task data of (2), we created the task data for (3)-2 Sentence-citation pair classification. Using the citations in the Related Work sections and matching them with the references in the paper in the bbl files, we obtained titles and authors. Then, we used the titles and authors to retrieve their paper IDs and abstracts through the arXiv API [3]. We also retrieved full text when available as tex source or a PDF file. We obtained 7,946 target papers that contain Related Work sections having citations with retrieved abstracts. The number has been reduced greatly due to the fact that many abstracts could not be retrieved via the arXiv API. These examples were split into three sets having 6,946, 500, and 500 target papers, maintaining the inclusion relationship shown in Figure 3. Then we made train/dev/test sets by extracting sentences with citations as positive examples and creating the same number of negative examples by randomly assigning a different citation. For the total number of examples in the task data of (3)-2, see the third row in Table 1.

From the target papers in the test data of (3)-2, we first extracted those that have Related Work sections with three or more citations. We found 600 such sentences. Then, from these, we extracted sentences with only one citation in order to create the task data for (3)-1 Citation allocation. We found 586 such sentences (see second row in Table 1). These sentences are used as test data for (3)-1. Note that, since citation allocation is performed by using the trained model of (3)-2, we have only test data for this task, although the model for this task can also be trained by creating

[3] https://arxiv.org/help/api/

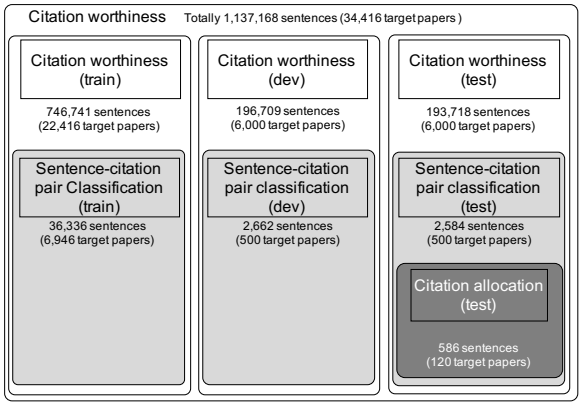

Figure 3: Inclusion relationship among datasets

its individual train/dev data.

5 Experiment

Using the task data, we evaluated baseline performance for these tasks. The aim of the experiment is to show the feasibility of our approach, that is, to test the performance of individual tasks and the integrated task using the same dataset. If this were successful, it would mean that our approach is effective for supporting various phases in scientific-paper writing.

5.1 Citation worthiness

Using the dataset for (2), we trained a BERT-based classifier (Devlin et al., 2019). We used `BertForSequenceClassification` from huggingface[4]. We used the bert-base-uncased model. For training, we used the train/dev data for this task as described in the previous section. The input format used for the classifier was "[CLS] sentence [SEP]." We used the Adam optimizer at a learning rate of $1.0e^{-5}$. We trained for 50 epochs and chose the model that achieved the highest accuracy for the development set. As evaluation metrics, in addition to accuracy, we used precision, recall, and F1 of positive labels (i.e., needs citation). Table 2 shows the results. As can be seen, the accuracy as well as F1 is quite high, much higher than previously reported (Bonab et al., 2018; Färber et al., 2018), which is probably due to the use of BERT.

5.2 Sentence-citation pair classification

Before citation allocation, we first describe the results of sentence-citation pair classification because it is a sub-task. This task determines whether a pair of a sentence and the abstract of its citation form a valid pair. Using the dataset for (3)-2, we trained a BERT-based classifier. In addition to a random baseline, we also prepared two other classifiers: a Doc2Vec-based classifier and an XLNet-based classifier. The methods used for comparison in this experiment are summarized below.

Random Randomly determines whether the citation is appropriate.

Doc2Vec This method utilizes Doc2Vec (Lau and Baldwin, 2016) to vectorize sentences and abstracts for similarity calculation. The Doc2Vec model was trained with the training data of this task. For all sentences with citations, we first concatenated a sentence and the abstract of the cited paper, then a Doc2Vec model was trained using the gensim[5] library. The trained model was used to convert a sentence and an abstract into vectors in order to calculate their cosine similarity. When the similarity increases above a predefined threshold (empirically set to 0.02 using the dev set), it is deemed appropriate.

BERT For training, each of the sentences and each abstract text in the references are paired to create training data while regarding the correct pair as a positive example or otherwise as a negative example. Then, the data are used for training a BERT-based classifier. Here, the input format is "[CLS] sentence [SEP] abstract text [SEP]." In the test phase, a pair consisting of a sentence and an abstract is fed to the trained classifier. We use the probability threshold of 0.5 to determine whether the pair is valid.

XLNet Instead of the BERT model, this method uses the XLNet (Yang et al., 2019) model (XLNet-base-model) for classification. This can be easily done using the huggingface library. The input format is the same as that for BERT.

For BERT and XLNet, we used the train/dev data for this task for fine-tuning. The training setting was the same as that used for citation worthiness.

[4] https://huggingface.co/

[5] https://radimrehurek.com/gensim/

Task	# Examples	# Positive examples	# Negative examples
(2) Citation worthiness	1,137,168	461,882	675,286
(3)-1 Citation allocation	586	N/A	N/A
(3)-2 Sentence-citation pair classification	41,582	20,791	20,791

Table 1: Statistics of task data

	Accuracy	P	R	F1
Random	0.499	0.408	0.500	0.449
BERT	0.911	0.925	0.852	0.887

Table 2: Accuracy for (2) Citation worthiness

	Accuracy	P	R	F1
Random	0.488	0.489	0.492	0.490
Doc2Vec	0.558	0.541	0.763	0.633
BERT	0.816	0.822	0.806	0.814
XLNet	0.844	0.846	0.841	0.843

Table 3: Accuracy for (3)-2 Sentence-citation pair classification

	Accuracy
Random	0.280
Doc2Vec	0.349
BERT	0.747
XLNet	0.795

Table 4: Accuracy for (3)-1 Citation allocation

	Accuracy
BERT	0.623

Table 5: Accuracy of integrated task composed of (2) Citation worthiness and (3)-1 Citation allocation, which includes (3)-2 Sentence-citation pair classification.

Table 3 shows the results for sentence-citation pair classification. As can be seen, the random baseline performs rather poorly, with an accuracy below 0.5. This is surpassed by the Doc2Vec method, which performed at an accuracy of 0.558. However, the two other models based on BERT and XLNet overwhelmed these with over 0.8 accuracy and F1. In this experiment, we can see that XLNet performs better than BERT.

5.3 Citation allocation

Using the dataset for (3)-1, we compared the four methods used in (3)-2, as shown below.

Random This method randomly chooses a citation from a list of possible references.

Doc2Vec This method uses the results of cosine similarity for sentence-citation pairs and chooses the highest-ranking one when it surpasses a predefined threshold of 0.02.

BERT This method uses the output of the BERT-based classifier for sentence-citation pairs. The highest-ranking pair is chosen as its citation when the output probability surpasses 0.5.

XLNet In place of the BERT model, this method uses the XLNet model for sentence-citation pairs.

The evaluation was carried out using test data containing 586 sentences.

Table 4 shows the results. The results clearly follow those of (3)-2, but accuracy is visibly lower. This is reasonable, since the results build on the sub-task. Reflecting the results obtained for sentence-citation pair classification, XLNet achieved the best performance at 0.795.

5.4 Integration of citation worthiness and citation allocation

We performed another experiment that spans two tasks: (2) Citation worthiness and (3)-1 Citation allocation. Note that task (3)-2 is included in (3)-1. Here, the input is a sentence that is first checked for citation worthiness. When it is determined that a citation is needed, the sentence is coupled with the abstracts of possible citations to check whether the pair is appropriate according to the sentence-citation pair classifier. Finally, the citation with the highest probability is chosen when it surpasses a predefined threshold. In this experiment, we used BERT-based methods for all tasks.

Table 5 shows the result of 0.623 for accuracy, indicating that cascading the tasks worsens per-

formance in comparison with the individual tasks. Although a reasonable accuracy can be achieved for a single task, this result shows that when they are combined, the performance may not be comparably high. The results would likely be even lower when more tasks are combined, which can give us clues on to how to improve overall performance and how to jointly train models. Using our design, it would thus be possible to evaluate the performance of a method to support scientific-paper writing at the various phases of research.

6 Summary and future work

In this paper, to achieve better support of scientific-paper writing, we first defined a series of tasks that can be pipelined. Then, focusing on the tasks of citation worthiness, citation allocation, and sentence-citation pair classification, we created a dataset of academic papers that could be used for the evaluation of each task as well as an integrated series of the tasks. We showed experimental results for citation worthiness, citation allocation, and sentence-citation pair classification for individual tasks as well as the case when these tasks are combined. Our series of experimental results shows the feasibility of our approach. We also showed the current performance using the same dataset.

Future work includes creating data for other tasks and performing experiments with them as well as their combinations in pipelined tasks. We will also consider the use of domain-specific pretrained language models, such as SciBERT (Beltagy et al., 2019), in order to improve performance. Furthermore, we plan to perform a human-in-the-loop evaluation in which a system supports researchers in their various writing phases. Finally, it would also be useful to improve the accuracy of the tasks we tackled in this paper.

Acknowledgments

We thank Hiroaki Sugiyama of NTT Communication Science Laboratories, Junji Yamato of Kogakuin University, and Genichiro Kikui of the National Agricultural Research Organization for their helpful comments and suggestions.

References

Zafar Ali, Guilin Qi, Pavlos Kefalas, Waheed Ahmad Abro, and Bahadar Ali. 2020. A graph-based taxonomy of citation recommendation models. *Artificial Intelligence Review*, 53(7):1573–7462.

Xiaomei Bai, Mengyang Wang, Ivan Lee, Zhuo Yang, Xiangjie Kong, and Feng Xia. 2019. Scientific paper recommendation: A survey. *IEEE Access*, 7:9324–9339.

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".

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 238–251.

Hamed Bonab, Hamed Zamani, Erik Learned-Miller, and James Allan. 2018. Citation worthiness of sentences in scientific reports. In *Proceedings of the 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, page 1061–1064.

Tanmoy Chakraborty, Natwar Modani, Ramasuri Narayanam, and Seema Nagar. 2015. DiSCern: A diversified citation recommendation system for scientific queries. In *Proceedings of the 2015 IEEE 31st International Conference on Data Engineering*, pages 555–566.

Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, Enrico Motta, and Harald Sack. 2020. AI-KG: An automatically generated knowledge graph of artificial intelligence. In *Proceedings of International Semantic Web Conference*, pages 127–143. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405.

Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018. To cite, or not to cite? Detecting citation contexts in text. In *Advances in Information Retrieval*, pages 598–603, Cham. Springer International Publishing.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task

7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 421–430.

Wenyi Huang, Zhaohui Wu, Prasenjit Mitra, and C. Lee Giles. 2014. Refseer: A citation recommendation system. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 371–374.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. Axcell: Automatic extraction of results from machine learning papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8580–8594.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany. Association for Computational Linguistics.

Kelvin Luu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2020. Citation text generation. *ArXiv*, abs/2002.00317.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 584–592.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 689–696.

Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. 2014. Cluscite: Effective citation recommendation by information network-based clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 821–830.

Tarek Saier and Michael Färber. 2020. unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *aScientometrics*, 125(3):3085–3108.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6181–6190.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of Advances in Neural Information Processing Systems 32*, pages 5753–5763.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7386–7393.