Reference classification using BERT models to support scientific-document writing

Ryoma Hosokawa[1][0009−0004−0782−2522], Junji Yamato[1], Ryuichiro Higashinaka[2], Genichiro Kikui[1], and Hiroaki Sugiyama[2]

[1] Kogakuin University
[2] NTT Corporation

Abstract. We are presently developing a document clustering method to group reference papers based on semantic similarity to support the writing of scientific papers by organizing the citations more appropriately. Currently, no dataset can be used for clustering experiments and evaluations for this purpose. In this study, we created two datasets of papers and corresponding references from PMC, an online medical paper archive. Then we performed clustering of the reference papers based on their abstracts using BERT, BioBERT, and SciBERT. In this case, we input the number of clusters for clustering. Clustering by BERT-based models trained on the similarity of pairs of references was more accurate than clustering by embedding the abstracts of references in each BERT model. Moreover, the trained BERT-based models had a clustering accuracy better or comparable to that of human experts. In addition, we predicted the number of clusters that used information from the references. The prediction accuracy for the number of clusters was about 40%. Evaluation measures for the clustering results are also discussed.

Keywords: Classification· Clustering· Natural Language Processing (NLP)· Neural Networks· Text Classification· Transformer· Word Embedding

1 Introduction

The number of scientific papers continues to increase rapidly. As the number of authors increases, so does the number of people seeking support with paper writing. As a result, research to support the writing of scientific papers has become increasingly important [1]. Paper summarization [2], citation recommendation [3], and generation of citation text tasks [4] have been reported. Narimatsu et al. [5] were the first to present the big picture on scientific writing support and summarize the necessary tasks for each research stage.

This paper addresses the "citation categorization task" defined by Narimatsu et al. [5]. According to Narimatsu et al., tasks to support scientific paper writing include citation extraction tasks, citation worthiness tasks, citation allocation tasks, citation recommendation for sentence tasks, and citation categorization tasks. Of these tasks, the citation categorization task has the highest priority in terms of being researched because it is related to other tasks and has not yet been reported. This task facilitates writing related work sections by categorizing each reference into an appropriate cluster when a set of reference papers is given so that the citations can be more appropriately organized. This task will be combined with the two subtasks that follow. There are citation sentence

generation tasks and citation text generation tasks. This group of tasks aims to generate related work sections automatically. The citation categorization task assumes a situation in which a group of citations is given , and support is needed for beginners writing scientific papers.

To accomplish the citation categorization task, we need a dataset showing the correspondence between a paper and which references papers are cited in the paper's related work section. However, no dataset shows these correspondences. Therefore, we created two separate datasets for this study and evaluated two methods with three BERT models. We also compared the clustering results with those of human experts.

In this paper, we adopted three BERT-based models for the citation categorization task: BERT and two fine-tuned models for scientific vocabulary called BioBERT [12] and SciBERT [13]. We utilized them to vectorize the abstracts of reference papers and used their vector similarity for clustering. In addition, we also fine-tuned the BERT models to calculate the similarity between a pair of papers by learning whether each pair of reference papers appears in the same paragraph of the citing papers in a related work section. We then clustered the references based on their similarity by k-means or hierarchical clustering.

We collected the papers from Pub Med Central (PMC) and listed the reference papers cited as related work in the "background" sections of the citing papers. We utilized references cited in one paragraph as one cluster. Specifically, the clustering problem is defined as performing clustering on all the papers cited in the background section of one paper based on a similarity measure of the cited papers. We then evaluated the correctness of the clustering by comparing it with the actual groups corresponding to paragraphs of the background section. Note that to perform BERT-based clustering, we use only the abstracts of the reference papers as the max length of the tokens of the BERT models is restricted. In this study, the number of clusters was given when clustering the reference papers. As an additional experiment, we predicted the number of clusters from the information in reference papers.

Our research questions are as follows.

- RQ1: Does fine-tuning BERT-based models contribute to the clustering performance?
- RQ2: Do BERT models achieve better clustering than humans with abstract information?

To answer these, we constructed two datasets from the PubMed archive (Section Ⅲ) and conducted two experiments to evaluate the clustering performances (Section Ⅳ).

2 Related Work

Much has already been done to support writing for scientific documentation. Citation sentence generation, citation context classification, cited text identification, citation recommendation, and multi-label classification of scientific documentation have already been researched [16], [17], [18], [19].Arita's study proposed citation sentence generation based on sentences in the citing paper and the cited papers [16]. These studies have been particularly advanced in recent years.

For the citation categorization task, TF-IDF-based methods have been standard [6] for years, but Transformer-based models such as BERT [7] and XL-Net [8] have recently delivered outstanding accuracy in a variety of NLP tasks. BERT models have demonstrated excellent performance with document clustering tasks as well [9], [10], [11]. However, these prior studies have not targeted scientific documents. Therefore, we conducted a clustering of scientific papers. In the clustering of conventional methods for scientific papers, the clustering was performed based on citation networks [20]. The aim of this study is to perform clustering based on the semantic similarity of abstract in papers. Since datasets for the clustering of scientific papers have not been constructed, we built datasets composed of scientific papers. Moreover, we evaluated the clustering performance for references.

We made the BERT models vectorize the abstracts of the references. We also let the BERT models learn the abstracts of the references that appeared in pairs for each cluster. Then, based on the learning, we calculated the similarity of the pairs of references in papers used for testing. Using that information, we used k-means or hierarchical clustering to cluster the references. We then predicted the number of clusters and envisioned what actual scientific documentation support would look like. Comparisons were also made with the result of reference clustering done by human experts to demonstrate validity. We also discuss an evaluation measure for the clustering results. In future work, we will work with other tasks to support scientific documentation writing.

3 Creation of Datasets

We obtained paper data from PMC, an online medical article archive, to build datasets with different search queries for the clustering experiments. PMC has data from many papers. Moreover, we can get that data in XML format. Accordingly, we can quickly analyze the data. Also, in this archive, there are many past papers. Thus, we considered it likely that we could obtain all the references as well.

For this purpose, PMC was selected as the paper archive. First, we collected 194,787 papers retrieved using the query words "artificial intelligence," "deep learning," or "image recognition." The papers thus collected were set as Dataset 1.

Second, we collected 137,998 papers retrieved using the query word "echocardiography." The papers thus collected were set as Dataset 2. The reason for the multiple queries for the first set of collected papers was to ensure that the number of papers was sufficient. We chose echocardiography because we will use our laboratory as a model case. In this study, it was necessary to target papers in the medical field to serve as a model case for the research areas of related topics being pursued by the authors.

Among these obtained papers, we eliminated papers whose type was "case report" to extract research articles. Case reports are common among medical papers. Case reports and other documents were eliminated from the datasets because they were deemed to have a different structure than the scientific papers intended for this study. In addition, we extracted those that begin with the background section (which is supposed to include related work information). For this reason, papers beginning with an introduction section instead of a background one were eliminated. For medical papers, the related work section is often

described in the background section. In addition, we then selected only papers written in English. This process resulted in 12,178 papers for Dataset 1 and 7,599 for Dataset 2.

Figure 1 shows an example of a background section. We extracted each paragraph from the reference papers. Table 1 shows an example entry from Dataset 1, which includes three paragraphs in a background section citing 14 papers.

We obtained an abstract from PubMed for each reference in the papers we obtained. Unfortunately, not all of these papers could be used for our clustering experiments because some of the reference papers were missing in PubMed, and we could not obtain the abstracts. If we eliminated papers with at most one reference paper missing, the dataset would be too small; therefore, we decided to allow for two missing references as a compromise. We also eliminated papers with only one paragraph in the background section because the number of clusters has to be more than one to perform the clustering properly.

As a result, the final number of papers for Datasets 1 and 2 was 2,752 and 3,890, respectively. This selection process is summarized in Table 2. We generated 768-dimension vectors from the reference abstracts of these papers using the BERT, BioBERT, and SciBERT models. The tokenizer used during the pre-trained of each model was used to convert the abstract sentences into tokens. The BERT models have an upper limit on the length that can be converted into tokens. Therefore, we entered the longest possible abstract text into the BERT models. We vectorized the abstract for clustering references.

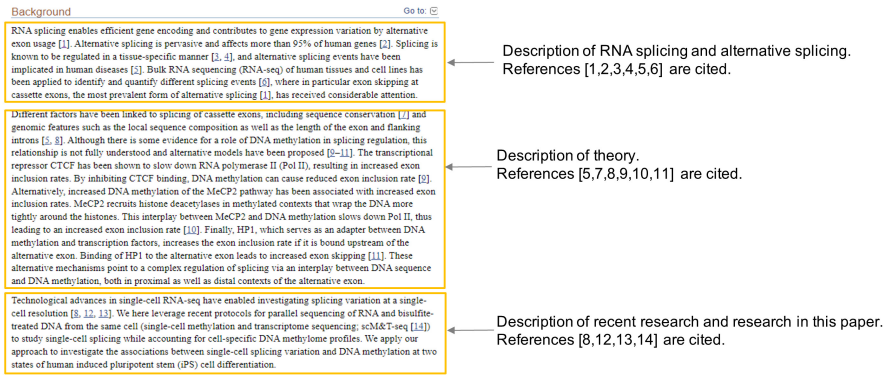

Fig. 1: Example of paragraph content and reference groups.

4 Methods

We conducted three experiments in this study. The first and second experiments used a given number of clusters. The first experiment regarded a clustering method that utilizes vectors of the abstracts of the references. The second used the similarity learned from whether each pair of reference papers appears in the same paragraph of the citing paper. The similarity was defined as the reciprocal of the distance between references. The third experiment predicted the number of clusters.

Table 1: Example of entry in Dataset 1.

Paper ID	PMC6371455
Paragraph 1	[1,2,3,4,5,6]
Paragraph 2	[5,7,8,9,10,11]
Paragraph 3	[8,12,13,14]
Reference 1–14	12626338, 23258890, 22909801, 20573213, 25525159, 18688268, 27717327, 28655331, 21964334, 23938295, 25704815, 28673540, 26740580, 26752769

The number in reference represents the PMID.

Table 2: Number of data per condition in dataset.

	Dataset 1	Dataset 2
Query	Artificial intelligence OR Deep learning OR Image recognition	Echocardiography
Number of retrieved papers	194,787	137,998
Articles that are [research-article] AND Include a [background] section	12,178	7,998
Cluster number ≥ 2 AND Missing references < 2	2,752	3,890

4.1 Clustering method

Four clustering methods were used: k-means, hierarchical clustering, x-means, and VBGMM (Variational Bayesian Gaussian Mixture Model). In this study, the k-means method and hierarchical clustering were given a number of clusters.

- *k-means method:*
  The k-means method is a basic clustering algorithm. First, the center of gravity of each cluster is set randomly. Then, for each dataset, the distance from the center of gravity of each cluster is calculated and assigned to the cluster with the closest distance. The above process is performed until the clusters to be assigned no longer change.
- *Hierarchical clustering:*
  Hierarchical clustering is a basic clustering algorithm. We used bottom-up hierarchical clustering. First, each piece of data is assigned to a cluster. Then, data that are close in distance to each other are merged into a single cluster. The above process is performed until the data are combined into a single cluster. Then, based on the generated dendrogram, we assign the data to an appropriate number of clusters.
- *X-means method:*
  The x-means method is an extension of the k-means method proposed by Dau [15]. This method allows for the prediction of clusters without giving the number of clusters. This is a significant difference from the k-means method. First, the k-means method is run with a small number of clusters to predict the number of clusters. Then, it is performed within the created clusters with k set to 2, and the clusters are divided. If the Bayesian Information Criterion (BIC) value increases, the clusters are split, and the number of clusters increases. The above process involves predicting the number of clusters by increasing the number of clusters from a small value.

– *VBGMM:*
The VBGMM (Variational Bayesian Gaussian Mixture) is a clustering method for cases where the number of clusters is unknown. This method is an application of Variational Bayes to the Gaussian Mixture Model. This method assumes multiple Gaussian distributions, and clustering is performed by determining which Gaussian distribution each element belongs to. The Gaussian distribution is updated until convergence based on the input data.

4.2 Evaluation measures

We used purity, pairwise F1, and pairwise accuracy to evaluate the clustering accuracy. Purity is known as a simple and transparent evaluation measure [14]. Equations (1) and (2) were used to calculate the purity, where N is the number of elements in a cluster, C_i are each generated cluster, and A_j is the cluster of correct answers.

In calculating pairwise F1 and pairwise accuracy, true positive is the number of correct reference pairs, not the number of elements, of reference pairs belonging to the same cluster.

$$\text{purity} = \sum_i \frac{|C_i|}{N} \text{max}_\text{j} \text{Precision}(C_i, A_j) \tag{1}$$

$$\text{Precision(C, A)} = \frac{|C \cap A|}{|C|} \tag{2}$$

Purity is used in this study as the evaluation measure for clustering. This is a measure whose value varies depending on the correspondence between the clusters predicted and the clusters of correct answers. Therefore, when evaluating the clustering results in this study, all patterns of correspondence between the clusters created and the clusters of correct answers were calculated. The correspondence of the clusters with the best purity was then used as the clustering result.

4.3 Experiment 1: Vector-based clustering for abstracts

In Experiment 1, we performed clustering on the dataset of papers created using the vectors of the abstracts of the references by the k-means method, with the number of clusters given. In this experiment, clustering was performed using three groups of vectors generated by BERT, BioBERT, and SciBERT, with each paper's abstract. The vectors of the abstracts of the references in this study came from the CLS token output of BERT.

BERT is a pre-training model with bidirectional transformers using large datasets [7]. SciBERT is a BERT model trained on 1.14 million scientific papers from Semantic Scholar [13]. BioBERT is a model trained using 4.5 billion words from paper abstracts in PubMed and 13.5 billion words from PMC papers.

We utilized random clustering as a baseline for this experiment. Random clustering in this experiment refers to assigning each element such that each cluster has at least one element.

4.4 Experiment 2: Pairwise learning based clustering

In Experiment 2, we used information on whether two references belong to the same cluster for training the three BERT-based models. We utilized the BERTForSequenceClassification[3] and prepared every combination of reference papers and a label, 1 or 0, meaning references belong to the same cluster (paragraph) or not the same cluster, respectively. BERTForSequenceClassification is a BERT model transformer with a sequence classification head on top. Figure 2 shows the sequence classification model. This allows the BERT model to classify sentences. We used the output of classification as a distance measure for clustering. In this experiment, we divided each dataset into training/validation/test sets at a ratio of 8:1:1.

For Dataset 1, 360,834 reference pairs generated from 2200 papers were used as training data, 45,754 references generated from 276 papers were used as validation data, and 276 papers were used as test data. For Dataset 2, we used 292,972 reference pairs from 3,112 papers for training data, 39,331 references from 389 papers for validation data, and 389 papers for test data. Among the training data, 38% in Dataset 1 and 42% in Dataset 2 were labeled 1. Since there was a slight bias, we introduced class weights in the training process. The percentage of class labels determined the class weights. This improved the low amount of data for label 1. Based on validation loss, a model trained for one epoch was used to predict clusters for the test data.

The value of logit output normalized by a softmax function was used as the distance between a pair of references. The distance value ranges from 0 to 1, decreasing if the two references are similar. A distance matrix was created by finding the distance between all references. The distance matrix was used to cluster the references. The k-means method and hierarchical clustering were used with the number of clusters given. We evaluated the clustering accuracy using purity and pairwise F1.

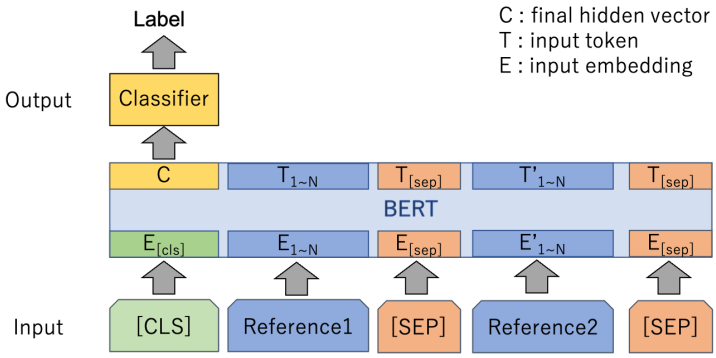

Fig. 2: Structure of BERT for sequence classification model.

[3] https://huggingface.co/docs/transformers/model_doc/bert

Human experts' performance We compared the accuracy of the clustering results to that of human experts. Collaborators with sufficient medical knowledge read the abstracts of the references and performed clustering, equivalent to the model in Experiment 2. Ten randomly selected sample papers were examined, and the average results of the BERT-based methods for the same ten samples were compared. We recruited three human experts for each dataset using a crowdsourcing service[4] under the condition of "licensed physicians or senior medical school students only."

The experimental collaborators were told the number of clusters for each paper. However, we did not give them the original reference numbers in the sample paper. This is because the referenced papers have a series of numbers, which we thought would give them a hint about clustering. Instead, we gave them a new number that we independently assigned to each referenced paper. They were asked to read the papers' abstracts and cluster them to reach a given number of clusters. In this experiment, the experts needed a long time to consider clustering. Therefore, we did not set a time limit for the clustering of each sample. They were allowed to do it whenever they wanted, with no time or order restrictions. These reasons were to confirm the optimal clustering results by human experts. The clustering results obtained in this way were evaluated in terms of purity. These samples all contained references, and the number of clusters was given.

4.5 Experiment 3: Predicting number of clusters

In Experiment 3, we predicted the number of clusters. In the previous experiments of this study, clustering was performed with the number of clusters given. Here, the number of clusters was predicted using the following two methods. In this experiment, if the number of clusters predicted and the number of clusters of correct answers were different, we added empty sets until we had the same number of clusters.

X-means algorithm For the first condition, the x-means algorithm was used. To predict the number of clusters in a paper, we used a vector of abstracts from the references used in Experiment 1, transformed by each BERT model X-means was used to determine how many clusters the vector was divided into. The x-means method was run with an initial number of clusters of 2. Hierarchical clustering was performed using the mode of the number of clusters obtained from ten runs of the x-means method and, as in Experiment 2, using the distance between references.

VBGMM algorithm For the second condition, we used the VBGMM algorithm to prediction the number of clusters. Hierarchical clustering was then performed with the number of clusters obtained using the distance between reference papers.

5 Results

The results table shows each dataset's average value of the evaluation results. The bold type in the table indicates the maximum value of the evaluation for

[4] https://crowdworks.co.jp/en/

the dataset for each method. In Tables 9 and 10, bold type indicates the number of correct cluster predictions.

Tables 3, 4 and 5 show the clustering accuracy in the experiments. We can see that the clustering evaluation value of any of the BERT models was increased by learning pairs of references. From Tables 3 and 4, we can see that the clustering results by BioBERT were evaluated highly. However, we confirmed that SciBERT was higher in terms of purity for the clustering results of Dataset 2. As we can see, all the BERT models had a higher purity and pairwise F1 than the random clustering result in Table 3 of Experiment 1. We can confirm that the pairwise accuracy of the BERT and SciBERT models is lower than the random clustering. In the results by random clustering, the number of elements in each cluster tends to take an average number. On the other hand, clustering by the BERT model results in some cases in a concentration of elements in one cluster. Pairwise evaluation tends to be lower when the cluster sizes are largely biased. As a result, the accuracy is considered to be lower than that of random clustering.

Table 6 shows the results of the cross-domain accuracy testing for each of the trained BERT models in different datasets. These results confirm that BERT and BioBERT had higher purity when trained on Dataset 1 than when trained on Dataset 2. They also confirm that SciBERT had higher purity for the training datasets than for the different test datasets. Still, even for the different test datasets, all BERT-based models had better accuracy for purity than the random clustering of Experiment 1. The clustering accuracy was improved even when trained on a different domain than the test dataset.

Tables 7 and 8 show the clustering accuracy in Experiment 3. Table 7 confirms that the clustering results obtained with SciBERT had the highest purity and pairwise evaluation for Dataset 2. Table 8 confirms that the clustering results obtained with BioBERT had the highest pairwise evaluation. We can also confirm that both methods' clustering results obtained with SciBERT had the highest purity evaluation. Moreover, the results of Experiment 3 had better purity accuracy than the random clustering of Experiment 1.

Tables 9, and 10 show the number of correct papers and the percentage of proper papers for the number of clusters.In Tables 9 and 10, C indicates the number of correct clusters, and P indicates the number of predicted clusters. It can be seen that the number of predictions was often smaller than the number of correct answers. The results in Table 9 and 10 also confirm that the VBGMM method was better at predicting the number of clusters. The results from the VBGMM method confirm that the prediction of the number of clusters correctly estimated by BioBERT was better. Still, the purity was lower than the other BERT models.

Table 11 shows a comparison with human clustering accuracy in terms of purity, where the human values show the average of the three human experts. The time required for one clustering task varied from person to person, ranging from 5 to 90 minutes. We found that there were individual differences in this task. As we can see, the results of the learned BERT models were better than the human results for Dataset 1. In contrast, those of the humans were slightly better than SciBERT and BioBERT for Dataset 2. In the result of Table 4, SciBERT and BioBERT also have a purity of 0.721 and 0.718, which is better than the human results. These results are based on a small number of samples, so it is difficult to say with certainty whether the performance of each domain is better or worse.

Table 3: Evaluation of clustering in Experiment 1.

Evaluation Measure	DS	BERT	SciBERT	BioBERT	Random
Purity	1	0.573	0.581	**0.652**	0.543
	2	0.605	**0.618**	0.600	0.578
Pairwise(F1)	1	0.367	0.364	**0.458**	0.319
	2	0.358	0.382	**0.390**	0.354
Pairwise(Accuracy)	1	0.504	0.506	**0.628**	0.516
	2	0.463	0.470	**0.542**	0.512

DS stands for dataset.

Table 4: Evaluation of clustering in Experiment 2 (purity).

Method	DS	BERT	SciBERT	BioBERT
k-means	1	0.693	0.699	**0.704**
	2	0.712	**0.721**	0.718
hierarchy	1	0.701	0.704	**0.710**
	2	0.714	0.721	**0.723**

Table 5: Evaluation of clustering in Experiment 2 (pairwise).

Evaluation Measure	DS	BERT	SciBERT	BioBERT
F1	1	0.597	0.607	**0.615**
	2	**0.601**	0.596	0.600
Accuracy	1	0.701	0.711	**0.722**
	2	0.725	0.728	**0.729**

Table 6: Trained BERT model results per dataset (purity).

BERT		Training		SciBERT	Training		BioBERT	Training	
	DS	1	2	DS	1	2	DS	1	2
Test	1	**0.693**	0.675	1	**0.699**	0.690	1	**0.704**	0.685
	2	**0.726**	0.712	2	0.711	**0.721**	2	**0.726**	0.712

Table 7: Results of clustering by x-means in Experiment 3

Evaluation Measure	DS	BERT	SciBERT	BioBERT
Purity	1	0.666	**0.685**	0.673
	2	0.683	**0.701**	0.693
Pairwise(F1)	1	0.572	0.603	**0.608**
	2	0.558	**0.599**	0.583
Pairwise(Accuracy)	1	**0.664**	0.657	0.654
	2	0.693	**0.696**	0.693

Table 8: Results of clustering by VBGMM in Experiment 3

Evaluation Measure	DS	BERT	SciBERT	BioBERT
Purity	1	0.666	**0.678**	0.676
	2	**0.703**	**0.703**	0.692
Pairwise(F1)	1	0.605	0.617	**0.623**
	2	0.607	0.604	**0.613**
Pairwise(Accuracy)	1	0.670	0.683	**0.692**
	2	0.700	0.701	**0.703**

Table 9: Number of papers for which number of clusters was correctly estimated (Dataset1, Experiment 3).

	Method	BERT	SciBERT	BioBERT
C < P	x-means	74	17	11
	VBGMM	45	42	39
C = P	x-means	74	85	83
	VBGMM	**93**	**93**	**103**
C > P	x-means	128	174	182
	VBGMM	138	141	134

C stands for number of correct clusters. P stands for number of predicted clusters.

Table 10: Number of papers for which number of clusters was correctly estimated (Dataset2, Experiment 3).

	Method	BERT	SciBERT	BioBERT
C < P	x-means	118	24	64
	VBGMM	60	68	44
C = P	x-means	148	152	141
	VBGMM	**160**	**154**	**169**
C > P	x-means	123	213	184
	VBGMM	169	167	176

C stands for number of correct clusters. P stands for number of predicted clusters.

6 Discussion

From the results of Experiment 1 in Table 3, we confirmed that the most accurate clustering for Dataset 1 was the one converted into a vector using BioBERT. This is presumably because BioBERT, a model specialized for medical vocabulary, could grasp the features of sentences better than the other two models since the paper data were medical papers from PMC. From the results of Experiment 2 in Table 4, we confirmed that the clustering accuracy was increased by learning pairs of references since the purity, accuracy, and F1 were higher than those of Experiment 1. From these results, we conclude that fine-tuning the vocabulary and domain knowledge contributed to the clustering performance to some extent (RQ1).

From the results in Table 6, we confirmed that the purity of the trained model for a dataset different from the training dataset was low. BERT had the lowest purity, indicating a difference in the dataset used for pre-training with the other BERT models. BERT, which had less pre-training for its area of expertise than the other BERT models, showed the most significant difference between the same training/test set result and the different training/test one. This might be because the pure BERT model without fine-tuning had the most significant margin to improve with domain knowledge.

From the results in Tables 4 and Table11, the BERT-based models showed better or comparable accuracy in terms of purity to the human experts (RQ2). This demonstrates that using BERT-based models for categorizing references has a good potential to be used in the writing support task defined by Narimatsu et al. [5]. It also indicates that the information contained in an abstract may be sufficient for such a clustering task. It is also likely that the higher purity value of the BERT models than in Experiment 2 was due to the papers selected. Unlike

Table 11: Human clustering result (purity).

DS	BERT	SciBERT	BioBERT	Human
1	0.762	**0.765**	0.759	0.735
2	**0.739**	0.674	0.688	0.699

the datasets for the other experiments, the papers selected for this experiment were those for which abstracts were obtained for all references. This allowed us to use data from each reference, and all relationships between references could be expressed. This is a merit that datasets with missing references do not have. There is merit in avoiding the need to take missing references into account when clustering. We believe this resulted in higher purity than the other data sets with missing references.

Regarding the prediction of the number of clusters by the x-means method, the results in Tables 9 and 10 confirm that even SciBERT, which is the most accurate, was correct only about 40% of the time. However, the results from SciBERT confirm that the most common case is that the number of clusters of correct answers was greater than the number of clusters of predictions. The errors in the number of predicted clusters and the number of correct clusters were generally 1. The number of clusters predicted by SciBERT was often two. The most common mistake was to predict the number of clusters to be 2 instead of 3. It is likely that the number of clusters predicted for papers with more than three clusters is wrong.

The results of BioBERT for Dataset 1 in Experiment 1 were superior to the other BERT models. However, BioBERT was also inaccurate in predicting the number of clusters with the x-means method in Experiment 3. This suggests that the method for predicting the number of clusters needs further improvement.

The VBGMM method was better at predicting the number of clusters, possibly because the features used differed from the x-means method. The x-means method used a vector of features from the abstracts of the references. However, VBGMM used the distance between references output by the trained model in Experiment 2. The output of the trained model showed better clustering results than the clustering using the output of the vectors by the BERT model. Therefore, the output was considered superior in predicting the number of clusters.

In this study, we used purity and pairwise evaluation to evaluate clustering. Pairwise evaluation can be performed regardless of the number of clusters. However, the evaluation decrease is significant if the wrong cluster belongs to a cluster with many elements. In contrast, the decrease is constant in purity evaluation no matter which element is mistaken. However, evaluation is difficult when the number of clusters is different. Therefore, we can evaluate the clustering method by checking the results of the two evaluation measures.

7 Conclusion and Future Work

In this paper, we investigated a citation categorization task in the context of support for scientific paper writing by creating two datasets for clustering references and proposed clustering methods. Our findings showed that the clustering accuracy is increased by having the BERT model learn labels that indicate whether two references belong to the same cluster. Moreover, the accuracy is better or comparable to human experts.

Future work will include clustering to allow common references for multiple clusters. In our experiments, a reference belongs to only one cluster because of the limitation of the clustering algorithm we used. However, a reference may sometimes belong to multiple clusters (paragraphs) in actual papers.

Future work will also include finding the adequate evaluation measure for clustering when the number of clusters is different from the correct number. In this study, we responded by adding an empty set when evaluating the basis of purity. The recall used in obtaining the F1 is characterized by a tendency to lead to a higher evaluation value when the number of clusters in the prediction is smaller than the number of clusters of correct answers. A problem with this property is that a lenient evaluation is obtained when the number of clusters is smaller than the true number.

The evaluation measure in this study should be able to correctly evaluate for common references and penalize an incorrect number of clusters in the prediction. Each of the clustering evaluation measures has its own merits. For this reason, we concluded to use multiple evaluation measures. In pairwise evaluation, it is possible to evaluate regardless of the number of clusters. On the other hand, purity evaluation is complex when the number of clusters differs. However, with purity evaluation, it is possible to evaluate even when multiple clusters do not have common references. In pairwise evaluation, clustering with no common references in the prediction results in a reduced evaluation. In addition, a disadvantage exists in that the range of decreases in the evaluation value is large when a cluster with a large number of elements mistakenly belongs to a cluster with a large number of elements.

The datasets for this experiment were created from the PMC archive. In the future, for non-medical papers, a dataset of papers and references in a particular field needs to be created. It will be important to construct datasets with a high acquisition rate of references.

In addition, clustering was performed by human experts with only ten samples out of the thousands contained in each dataset due to resource limitations. We plan to expand the sample size and the number of experts in the future.

In this study, clustering was performed using abstracts for the dataset, which is information that can be easily collected. For clustering, we used the abstract information from the references, so we need to consider other information that can be used for clustering. For example, we might use the amount of text per paragraph in the "background" section. Paragraphs with a large amount of text may have a large number of references. Thus, clustering accuracy may be improved by using information other than abstract information.

References

1. Bai, Xiaomei, Wang, Mengyang, Lee, Ivan, Yang, Zhuo, Kong, Xiangjie, and Xia, Feng, "Scientific Paper Recommendation: A Survey," in *IEEE Access*, vol. 7, 2019, pp. 9324–9339.
2. Vahed Qazvinian and Dragomir R. Radev, "Scientific Paper Summarization Using Citation Summary Networks," in *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, 2008, pp. 689–696.
3. Huang, Wenyi and Zhaohui Wu and Mitra, Prasenjit and Giles, C. Lee "RefSeer: A Citation Recommendation System," in *IEEE/ACM Joint Conference on Digital Libraries*, 2014, pp. 371–374.

4. Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan, "Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6181–6190.
5. Hiromi Narimatsu, Kohei Koyama, Kohji Dohsaka, Ryuichiro Higashinaka, Yasuhiro Minami, and Hirotoshi Taira,, "Task Definition and Integration for Scientific-document Writing Support," in *Proceedings of the Second Workshop on Scholarly Document Processing*, 2021, pp. 18–26.
6. Prafulla Bafna, Dhanya Pramod, and Anagha Vaidya, "Document Clustering: TF-IDF Approach," in *Proceedings of the 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016
7. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "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*, 2019, pp. 4171–4186.
8. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2020
9. Michal Marcińczu, Mateusz Gniewkowski, Tomasz Walkowiak, and Marcin Bedkowski, "Text Document Clustering: Wordnet vs. TF-IDF vs. Word Embeddings," in *Proceedings of the 11th Global Wordnet Conference*, 2021, pp. 207–214.
10. Yutong Li, Juanjuan Cai, and Jingling Wang, "A Text Document Clustering Method Based on Weighted BERT Model," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 2020
11. Haoxiang Shi and Cen Wang, "Self-supervised Document Clustering Based on BERT with Data Augment," in *ArXiv*, abs/2011.08523, 2020
12. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, "BioBERT: a Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," in *Bioinformatics, Volume 36, Issue 4*, 2020, pp.1234–1240.
13. Iz Beltagy, Kyle Lo, and Arman Cohan, "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 (EMNLP-IJCNLP)*, 2019, pp. 3615–3620.
14. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, "Introduction to Information Retrieval," in *Cambridge University Press*, 2008
15. Dau Pelleg and Andrew Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," in *Proceedings of the 17th International Conf. on Machine Learning*, 2000, pp. 727–734.
16. Akito Arita, Hiroaki Sugiyama, Kohji Dohsaka, Rikuto Tanaka, and Hirotoshi Taira, "Citation Sentence Generation Leveraging the Content of Cited Papers," in *Proceedings of the Third Workshop on Scholarly Document Processing*, 2022, pp. 170–174.
17. Sonita Te, Amira Barhoumi, Martin Lentschat, Frédérique Bordignon, Cyril Labbé, and François Portet, "Citation Context Classification: Critical vs Non-critical," in *Proceedings of the Third Workshop on Scholarly Document Processing*, 2022, pp. 49–53.
18. Autumn Toney and James Dunham, "Multi-label Classification of Scientific Research Documents Across Domains and Languages," in *Proceedings of the Third Workshop on Scholarly Document Processing*, 2022, pp. 105–114.
19. Zoran Medić and Jan Snajder, "Large-scale Evaluation of Transformer-based Article Encoders on the Task of Citation Recommendation," in *Proceedings of the Third Workshop on Scholarly Document Processing*, 2022, pp. 19–31.
20. Lovro Šubelj, Nees Jan van Eck, Ludo Waltman, "Clustering scientific publications based on citation relations: A systematic comparison of different methods," in *PLoS ONE 11(4)*, 2016.