SciDocFind - Faceted Ranked Retrieval of Scientific Research Papers

Soni Aditya Bharatbhai, Likhith Reddy Morreddigari, Rishi Raj, Shreyas Jena

Abstract

This paper proposes a novel approach for faceted retrieval of scientific research papers using the CSFCube dataset for evaluation. Given a query paper Q, a facet f $\in$ {*background, method, result*} and a set of candidate papers $\{C_1, C_2, ..., C_n\}$, the task is to rank candidate papers similar to Q with respect to the facet f. In particular we fine-tuned the SentBERT-PP model using the PARADE dataset and observed improvement in the performance metrics when the facet being searched for is *result*. Additionally we also fine-tuned the SPECTER and the Sci-NCL models using a subset of the HIC dataset. For the fine-tuned SPECTER model, we observed improvement in performance metrics when the facet being searched for is *method*.

1 Introduction

The subject of faceted scientific paper retrieval has recently gained attention within the Information Retrieval community. This task has significant potential benefits for the research community, as it enables researchers to retrieve research papers that are more relevant to their specific interests, thereby facilitating the dissemination of scientific knowledge.

Recent research in this domain has highlighted the importance of pre-trained language models (Beltagy et al., 2019) in acquiring effective document representations, allowing retrieval models to yield more relevant search results. To further advance this line of inquiry, current state-of-the-art research (Cohan et al., 2020) (Ostendorff et al., 2022) has explored the use of citation network data to fine-tune language models pre-trained on scientific text. This approach enables the models to capture inter-document relationships, resulting in even more significant improvements in learning document representations.

Despite their demonstrated effectiveness in other contexts, these methods have been found to be less effective when evaluated on faceted query retrieval datasets such as CSFCube (Mysore et al., 2021). Specifically, when tested on faceted scientific retrieval tasks involving three distinct facets - Background, Method, and Result, the models typically exhibit suboptimal overall performance, particularly concerning the Method and Result facets. This outcome underscores the need for a more robust inter-document similarity criterion in the learning of faceted relevance, one that surpasses the limitations of the citation network data currently available.

To address this issue, we leveraged the SciRepEval dataset (Singh et al., 2022), a large-scale benchmark that includes multiple tasks designed to evaluate generic scientific paper comprehension. Our proposed solution is an end-to-end approach that builds on a pre-trained language model trained on scientific paper citation network data, which is further fine-tuned using Contrastive Loss, using triples generated from the Highly Influential Citations dataset. Our experimental results reveal that our proposed method achieves significant improvement in the Method facet of the CSFCube dataset. Furthermore, even when trained using limited computational resources, our approach exhibits competitive performance in other facets, thereby highlighting its effectiveness.

1.1 Related Work

Large language models have been widely used to create high-quality document representations for evaluating scientific paper similarity. SciBERT (Beltagy et al., 2019) involves a BERT (Devlin et al., 2018) model which is pretrained on a corpus of scientific papers. Since SciBERT is found to underperform on scientific paper retrieval tasks, recent works like SPECTER and SciNCL (Cohan et al., 2020) (Ostendorff et al., 2022) make use of citation network data to help the model learn inter-document similarity effectively. But the un-

derwhelming performance of such state-of-the-art models on faceted scientific paper retrieval datasets like CSFCube across various facets indicates the need for a stronger method of representing inter-document similarity than citation data.

Another work of significance is the SciRepEval dataset, a large-scale benchmark dataset for evaluating the performance of scientific retrieval models on a diverse array of tasks. Two datasets of the benchmark are of interest here - Field of Study (FoS) and Highly Influential Citations (HIC). The FoS dataset scientific papers labelled with their corresponding domain/field of study. The HIC dataset consists of tuples of query and candidate papers, where a positive candidate is said to be highly influential only if it is cited by the query paper atleast 4 times, otherwise the candidate is marked as negative (cited less than 4 times by the query paper). We make use of the stringent similarity criterion mentioned by the HIC dataset to make the model learn stronger notions of document similarity.

2 Method

2.1 Highly Influential Citations (HIC)

Given the promising results of pretrained language models that are fine-tuned on citation network data in scientific query retrieval tasks, we employ two such models - SPECTER and SciNCL - as baselines for producing scientific document representations. To further enhance the models' understanding of inter-document similarity, we employ the Highly Influential Citations (HIC) dataset, which is part of the SciRepEval benchmark. The HIC dataset comprises papers from 23 different scientific domains and contains tuples of query and candidate papers of the form:

$$\langle q, \langle c_1, s_1 \rangle, \langle c_2, s_2 \rangle, \ldots, \langle c_n, s_n \rangle \rangle$$

Here, q denotes the query paper, while $\langle c_i, s_i \rangle$ refers to the set of candidate papers, with $s_i \in \{0, 1\}$ indicating their corresponding relevance scores. The paper c_i is deemed a highly influential citation for q_i if q_i cites c_i at least four times in its paper body. We select this dataset for fine-tuning because the stringent conditions set by it impose a much stronger constraint on a paper's high relevance. Additionally, the corresponding negative examples allow for the creation of a large number of hard negatives, enabling the model to acquire a more specific notion of inter-document similarity.

2.2 Multi-Label Scientific Paper Classifier

To account for the lack of domain labels in the Highly Influential Citations dataset, we employ the Field of Study (FoS) dataset of the SciRepEval benchmark, which contains around 676k papers labeled across 23 different domains. Out of these, around 6% of the papers belong to the Computer Science domain. We employ a pre-trained SciBERT model and finetune it for multi-label classification. The classification metrics of the resultant multi-label classifier are given in Table 1.

CS-Precision	CS-Recall
0.6304	0.7023

Table 1: Classification metrics for the finetuned multi-label SciBERT classifier. CS-Precision computes the ratio of correctly predicted CS papers to the total CS papers predicted, while CS-Recall denotes the ratio of correctly predicted CS papers to the total CS papers in the dataset.

As shown in Table 1, the finetuned classifier shows strong performance in correctly classifying Computer Science papers and classifies a relatively small fraction of non-CS papers incorrectly, despite the large fraction of non-CS papers in the dataset. We leverage the effectiveness of the classifier to classify and filter out those query papers from the HIC dataset which are classified as CS papers.

2.3 Triplet Loss-based Finetuning

Considering the relatively stringent criterion imposed by the HIC dataset, we argue that the resultant triples formed from the filtered dataset can serve as sufficiently hard negatives for triplet loss-based finetuning of the base models. We therefore create a balanced dataset of triples from the resultant dataset and perform triplet loss-based finetuning on the SPECTER and SciNCL models. The resultant approach is shown to produce overall improvements in baseline model performance across specific facets for the finetuned SPECTER model, as explained in the Analysis section.

2.4 PARADE Dataset

PARADE (He et al., 2020) is a paraphrase-identification dataset comprising of sentences belonging to the Computer Science domain. It consists of pairs of sentences $\langle \langle p_i, q_i \rangle, s_i \rangle$, where $s_i \in \{0, 1\}$ and $s_i = 1$ if q_i is a paraphrase of p_i. We perform finetuning of the SentBERT-PP (Reimers

and Gurevych, 2019) model on this paraphrase-identification dataset and the relevant details are mentioned in the Experiments section.

3 Experiments

For each paper in the CSF-Cube dataset, the sentences of the paper's abstract are assigned a label ∈{*background, method, result, other*}. Similarly, for faceted search, the facets ∈{*background, method, result*}.

3.1 Techniques for Ranked Retrieval

3.1.1 Sentence-level models:

For sentence-level models, the sentences in the abstract of the query paper and the abstract of the candidate paper are encoded using the model and then ranking is done based on the maximum pairwise cosine similarity between the query sentences and the candidate sentences. We employ three different methods to choose the sentences from the abstract of the query and the candidate papers:

1. **Query Abstract - Candidate Abstract (QA-CA)**: For both the query and the candidate paper, all the sentences in the abstract are considered.

2. **Query Facet - Candidate Abstract (QF-CA)**: For the query paper, only those sentences are considered from the abstract for which the label is the same as the facet being searched for. All sentences from the abstract are considered for the candidate.

3. **Query Facet - Candidate Facet (QF-CF)**: For both the query and the candidate paper, only those sentences are considered from the abstract for which the label is the same as the facet being searched for.

3.1.2 Abstract-level models:

For the abstract-level models, SPECTER and Sci-NCL the query and candidate papers are represented using the sequence by title + [tokenizer.SEP_token] + abstract. For SciBERT, the query and candidate papers are represented using the abstract. The CLS embedding is considered the dense vector representation of the paper. The ranking is done by increasing the order of L2 distance between the query and the candidate paper.

For all three models, all sentences of the abstract may not be considered as shown below. There are three different methods by which we construct the abstract:

1. **Query Abstract - Candidate Abstract (QA-CA)**: For both the query and the candidate paper, the entire abstract is considered.

2. **Query Facet - Candidate Abstract (QF-CA)**: For the query paper, the abstract is constructed by concatenating those sentences of abstract which have the same label as the facet being searched for. For candidate paper, the entire abstract is considered.

3. **Query Facet - Candidate Facet (QF-CF)**: For both the query and the candidate paper, the abstract is constructed by concatenating those sentences from the abstract which have the same label as the facet being searched for.

Concatenation of sentences is done in the same order as the sentences that appear in the abstract.

3.2 Fine-tuning SPECTER and Sci-NCL using the HIC dataset

As described in section 2.2, first the HIC dataset is pre-processed using the multi-label classifier. As a result, 6165 query papers were obtained. For each query paper, if there are n1 positive and n2 negative papers, then $\min(n1, n2)$ number of triples {*query, positive paper, negative paper*} is constructed by randomly choosing $\min(n1, n2)$ pairs of (positive paper, negative paper) where no positive or negative paper is repeated. As a result, 16907 triples are obtained. Each training instance is a triplet of papers: a query paper P^Q, a positive paper P^+, and a negative paper P^-. Triplet loss function is used for fine-tuning as shown:

$$L = \max\left\{\left(d(P^Q, P^+) - d(P^Q, P^-) + m\right), 0\right\}$$

Where d is the L2 norm distance and m is empirically chosen as 1.

3.3 Fine-tuning SentBERT-PP using the PARADE dataset

The PARADE dataset is composed of a collection of paired sentences, denoted by p_i and q_i, and associated binary class labels. Notably, the dataset is well-balanced, which ensures that the model is exposed to a diverse set of examples during training. To optimize the model's performance, a contrastive loss function is employed, which seeks to minimize the embedding distance between similar sentence

pairs and maximize the distance between dissimilar sentence pairs. Contrastive loss is defined as

$$L = (1-Y)*||x_i-x_j||^2+Y*max(0, m-||x_i-x_j||^2)$$

where m is a hyperparameter, defining the lower bound distance between dissimilar samples. To further enhance the model's accuracy, it is fine-tuned over the course of 3 epochs, with a batch size of 32 and learning rate 2e^{-5}. m is empirically chosen to be 1.

4 Analysis

4.1 Baselines

For the abstract level models, SPECTER and Sci-NCL (both fine-tuned using SciBERT as base model) significantly outperform Sci-BERT indicating that fine-tuning Sci-BERT using citation data has improved performance for SPECTER and Sci-NCL [1]. Additionally, we observe that SPECTER works the best for the QA-CA retrieval method as compared to QF-CF and QF-CA. However, for Sci-NCL the trend is slightly different. Both QA-CA and QF-CF retrieval methods yield comparable results. This indicates that SPECTER is impacted by the loss of context in the QF-CF method which is not the case for Sci-NCL.

For sentence-level models, all the models perform similarly but SentBERT-PP is the best-performing model. A general trend is observed that the QF-CA and QF-CF retrieval methods perform better than QA-CA. However, SPECTER and SciNCL significantly outperform SentBERT-PP which means that the retrieval is improved when the context is provided.

4.2 Fine-tuning the models

For comparison of the fine-tuned and the baseline models, we have considered SPECTER and SentBERT-PP. [2]

For SPECTER, when the facet is "method" and the retrieval approach is QA-CA, Recall@20 increases by 4% and the remaining metrics have similar values for the 2 models. However, the performance of the fine-tuned model on the other two facets is worse than the baseline model. This may be caused by the fact that a paper cited less than 4 times by a query paper may have a similar background/result but if a paper is cited 4 or more times by a query paper then there is a high chance that these two papers have similar methods.

For SentBERT-PP, when the facet is "result" and the retrieval approach is QF-CA, Recall@20 increases by 5% and the remaining metrics have similar values for the 2 models. For other facets, the fine-tuned model performs better or it is similar to the baseline model. Thus, fine-tuning SentBERT-PP on PARADE helps the model to better capture the similarity between two sentences of the computer science domain.

This means that the fine-tuned model retrieves more relevant candidates in the first 20 retrieved papers for both abstract and sentence-level baselines.

4.3 Future work

We have obtained convincing results after fine-tuning the SentBERT-PP model on the PARADE dataset. We tried to use a similar dataset called ParaSci (Dong et al., 2021) which consists of only pairs of sentences in the CS domain which are paraphrases of each other. However we could not figure out a proper method to obtain pairs of sentences which are not paraphrases of each other using this dataset. Additionally we observed that the CSF-Cube dataset has a lot of latex snippets which are not handled in our present experiments. Due to computational constraints we have fine-tuned SPECTER and Sci-NCL only on those query papers from HIC dataset which belong to CS domain. The effect of using the entire HIC dataset remains unexplored.

Type	R-Prec	Prec@20	Recall@20	NDCG
Baseline	9.9	14.89	43.8	65.97
Finetuned	11.53	13.89	**47.81**	64.39

Table 2: Analysis of specter baseline on QA-CA facet

Type	R-Prec	Prec@20	Recall@20	NDCG
Baseline	13.6	19.83	41.73	71.9
Finetuned	12.98	20.42	**47.06**	71.76

Table 3: Analysis of Sentbert-PP for QF-CA facet

[1]The results for the baseline models are available here.
[2]The results for the fine-tuned models are available here.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 424–434.

Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020. PARADE: A new dataset for paraphrase identification requiring computer science domain knowledge. *CoRR*, abs/2010.03725.

Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. CSFCube - a test collection of computer science research articles for faceted query by example. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In *The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, Abu Dhabi. Association for Computational Linguistics. 7-11 December 2022. Accepted for publication.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *ArXiv*, abs/2211.13308.

Appendix

A. Comparison of Baseline and Fine tuned Models

Candidate Title	Gold Label
Adaptation of context-dependent deep neural networks for automatic speech recognition:	0
Mixture-Model Adaptation for SMT:	0
Efficient training of large neural networks for language modeling:	0
Training Continuous Space Language Models: Some Practical Issues:	0
Strategies for Training Large Vocabulary Neural Language Models:	0

Table 4: Query Title - Incremental Adaptation Strategies for Neural Network Language Models

Table 4 displays the top 5 candidates retrieved with their Gold Labels(as annotated by CSF-Cube Authors) using the baseline **Specter** for the shown Query. QA-CA Retrieval approach is used and the facet being searched is *method*.

Keeping the Query, retrieval approach and the facet the same the top 5 candidates retrieved by **Finetuned-Specter** are shown in Table 5. Clearly the fine-tuned version provides a better retrieval.

Candidate Title	Gold Label
On the dynamic adaptation of stochastic language models:	1
Training Continuous Space Language Models: Some Practical Issues:	0
Incremental Learning Through Deep Adaptation:	2
A Comparison of Adaptation Techniques and Recurrent Neural Network Architectures:	0
Learning Simpler Language Models with the Delta Recurrent Neural Network Framework:	0

Table 5: Query Title - Incremental Adaptation Strategies for Neural Network Language Models

Candidate Title	Gold Label
Human-Level Performance on Word Analogy Questions by Latent Relational Analysis:	3
Measuring Semantic Similarity by Latent Relational Analysis:	3
Correlation Coefficients and Semantic Textual Similarity:	0
Comparing measures of semantic similarity:	1
Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses:	1

Table 6: Query Title - Similarity of Semantic Relations

Table 6 displays the top 5 candidates retrieved with their Gold Labels using the baseline **SentBERT-PP** for the shown Query. QF-CA Retrieval approach is used and the facet being searched is *result*.

Candidate Title	Gold Label
Human-Level Performance on Word Analogy Questions by Latent Relational Analysis:	3
Measuring Semantic Similarity by Latent Relational Analysis:	3
Alternative measures of word relatedness in distributional semantics:	0
Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses:	1
Improved Sentence Similarity Algorithm based on VSM and its application in Question Answering System:	2

Table 7: Query Title - Similarity of Semantic Relations

Keeping the Query, retrieval approach and the facet the same the top 5 candidates retrieved by

Finetuned-SentBERT-PP are shown in Table 7. Clearly the fine-tuned version provides a better retrieval.

B. Work Distribution

Name	Experiments, Ideation, Comments
Soni Aditya Bharatbhai	Literature survey, Coding baselines, Fine tuning SPECTER and SciNCL, Design Decision about which datasets to choose for Fine tuning SPECTER, SCINCL and SentBERT-PP, Preparation of Presentation Slides and Report
Morreddigari Likhith Reddy	Literature survey, Coding baselines, fine tuning SentBERT-PP, Preparation of Demo Video and Demo Code, Preparation of Slides
Rishi Raj	Literature survey, preparation of CSF-Cube, PARADE and HIC dataset, Analysis of results, Preparation of Report
Shreyas Jena	Literature survey, Coding the Multi-label classifier, preparation of HIC dataset, Preparation of Report