Report on SPECTER: Document-level Representation Learning using Citation-informed Transformers

Anonymous ACL submission

Abstract

The research of classification and recommendation on scientific documents is critical in natural language processing systems. In this paper, a new method and a new evaluation benchmark are proposed to generate document-level embedding of scientific documents based on pretraining a Transformer language model on a powerful signal of document-level relatedness and evaluate citation prediction, document classification and recommendation. With this method we are able to apply SPECTER to downstream applications without task-specific fine-tuning,which also encourage further research on document-level models.

1 Introduction

With the increasing scientific publication,it has become critical to help users to search, discover and understand the scientific literature by using Natural Language Processing (NLP) tools. This model SPECTER can help us address these problems.Pretrained neural language models (LMs)(Radford et al., 2018; Devlin et al., 2019; Yang et al., 2019) is used to represent individual words or sentences.However, the whole-document embeddings need relatively more extensions.Likewise,the method that can produce whole-document embeddings(Tu et al., 2017; Chen et al., 2019) by using inter-document signals is lack of pretraining.In this paper,the goal is to leverage the power of pretrained language models to learn embeddings for scientific documents. Rich semantic content about the paper can be provided by a paper's title and abstract.However, the recent SciBERT (Beltagy et al., 2019) does not result in accurate paper representations because it passes these textual fields to an "off-the-shelf" pretrained language model simply.It doesn't help in document-level tasks such as topic classification or recommendation.

In this paper,the authors propose a new method to learn general-purpose vector representations of scientific documents.They use the system SPECTER to incorporates inter-document context into the Transformer (Vaswani et al., 2017)language models (e.g., SciBERT (Beltagy et al., 2019)) to learn document representations that are effective across a wide-variety of downstream tasks, [without the need for any task-specific fine-tuning of the pretrained language model]C. So it is less costly and perform better.In addition,they specifically use citations as a naturally occurring, inter-document incidental supervision signal indicating which documents are most related and formulate the signal into a triplet-loss pretraining objective.Compared with the prior work, the improvement of their work is that their model does not require any citation information. Their results show that SPECTER can be a powerful Natural Language Processing (NLP) tool to work on a variety of document-level tasks, including topic classification, citation prediction, and recommendation.

In this paper they also introduce and release SCIDOCS as a collection of data sets and an evaluation suite for documentlevel embeddings in the scientific domain,which covers seven tasks and tens of thousands of examples of anonymized user signals of document relatedness.Their training set and trained embedding model and its associated code base are all released.

2 Related work

SCIBERT, this pretrained language model based on BERT (Devlin et al., 2019) can address the lack of high-quality, large-scale labeled scientific data. It is widely used to leverage unsupervised pretraining on a large multi-domain corpus of scientific publications to improve performance on downstream scientific NLP tasks. Many previous works such as GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) show that unsupervised pretraining of language models on large corpora significantly improves performance on many NLP tasks. These models return contextualized embeddings for each token which can be passed into minimal task-specific neural architectures. (Radford et al., 2018) demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. They also propose a large model, GPT-2, a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. (Devlin et al., 2019) further design a new language representation model called BERT to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial taskspecific architecture modifications. In this paper,we focus on downstreaming applications without task-specific fine-tuning.

3 Methods

3.1 Overview

In order to learn task-independent representations of academic papers,they use the Transformer model architecture as basis of encoding the input paper.In addition, they also use citations as an inter-document relatedness signal and formulate it as a triplet loss learning objective to learn high-quality document-level representations . Then they pretrain the model on a large corpus of citations using this objective, encouraging it to output representations that are more similar for papers that share a citation link than for those that do not.

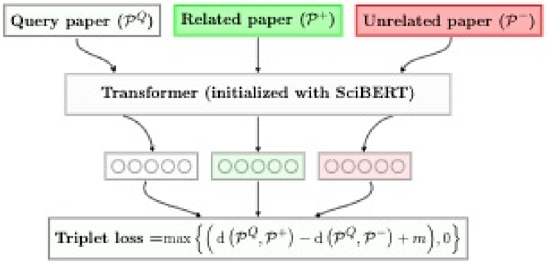

Query paper ($\mathcal{P}^Q$)
Related paper ($\mathcal{P}^+$)
Unrelated paper ($\mathcal{P}^-$)
Transformer (initialized with SciBERT)
Triplet loss =max$\left\{\left(\mathrm{d}\left(\mathcal{P}^Q,\mathcal{P}^+\right)-\mathrm{d}\left(\mathcal{P}^Q,\mathcal{P}^-\right)+m\right),0\right\}$

Figure 1: Overview of SPECTER

3.2 Background

Based on the models from pretrained Transformer networks they create SPECTER. Especially they use SciBERT(Beltagy et al., 2019) because it adapts to the original BERT (Devlin et al., 2019) architecture in the scientific domain. Bert is trained on a large unlabelled dataset and it achieve state-of-the-art results on 11 different NLP tasks. Bidirectional means that BERT learns information from both the left and the right side of a token's context during the training phase.The BERT is based on Transformer.There are 3 embeddings in every input embedding–Position Embeddings, Segment Embeddings, Token Embeddings. When a token is given, its input representation is constructed by summing the corresponding token, segment, and position embeddings.

Document Representation In order to represent a given paper P they use a dense vector v which can represent the paper best and also can be used in downstream tasks. The embeddings in SPECTER are from the title and abstract of a paper. Because these two fields can usually provide a succinct and comprehensive summary of the paper and therefore they can also produce accurate embeddings. Then they use a Transformer LM (e.g., SciBERT) to encode the concatenated title and abstract and the output is the final representation of the [CLS] token.

$$v = Transformer(input)_{[CLS]},$$

Transformer's forward function is Transformer and input is the concatenation of the [CLS] token and WordPieces (Wu et al., 2016) of the title and abstract of a paper, separated by the [SEP] token. They use general formulation and use SciBERT as model initialization .But this model does not take global inter-document information into account.Because SciBERT only predict words or sentences given their in-document, nearby textual context.

3.3 Citation-Based Pretraining Objective

The documents can be related by a citation from one document to another.Therefore they design a loss function to train the Transformer model so the model can learn closer representations for papers when one cites the other, and more distant representations otherwise. They make a triplet of papers: a query paper P^Q, a positive paper P^+ and a negative paper P in each training instance.The positive paper is a paper that the query paper cites, and the negative paper is a paper that is not cited by the query paper (but that may be cited by P^+). They then train the model using the following triplet margin loss function:

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

where d is a distance function and m is the loss margin hyperparameter (they empirically choose m = 1). Here, they use the L2 norm distance:

$$d(P^A, P^B) = \|v_A v_B\|_2$$

where v_A is the vector corresponding to the pooled output of the Transformer run on paper A (Equation 1).Starting from the trained SciBERT model, they pretrain the Transformer parameters on the citation objective to learn paper representations that capture document relatedness.

3.4 Selecting Negative Distractors

It is important to choose negative example when training the model. There are two different sets of negative examples. One is selected papers from the corpus,the other is [“hard negatives”]C.the papers that are not cited by the query paper, but are cited by a paper cited by the query paper.

4 SCIDOCS Evaluation Framework

In order to evaluate scientific document they introduce a new comprehensive evaluation framework called SCIDOCS. This framework consists of citation prediction, prediction of user activity, document classification and paper recommendation. [Note that SPECTER will not be further fine-tuned on any of the tasks; we simply plug in the embeddings as features for each task.]C.Retrieving relevant documents from a corpus is typically based on the semantic similarity between the document content and query text.(Raman et al., 2022)

4.1 Document Classification

There are two main scientific tasks in document classification which are MeSH Classification and Paper Topic Classification. The key point of these two tasks is if it can predict the class of the documents.

MeSH Classification This task is classifying scientific papers using their Medical Subject Headings (MeSH) (Lipscomb, 2000). The dataset is about 23K academic medical papers.For each of those, the topic of the paper is about one of 11 top-level disease such as cardiovascular diseases, diabetes, digestive diseases derived from the MeSH vocabulary. The largest number of disease is Neoplasms (cancer) with 5.4K instances,which is more than one fifth of the total dataset. On the contrary, the minimum number of the disease is Hepatitis (1.7% of the total dataset).

Paper Topic Classification In this task ,the goal is to predict the topic of the paper according to the predefined topic categories of the Microsoft Academic Graph (MAG) (Sinha et al., 2015). MAG is a database full of the papers which are tagged with a list of topics.The dataset is about 25K papers. For the whole dataset, they are evenly assigned to 19 different classes of level 1 categories in MAG. While in MAG,the level 1 is the most general and the level 5 is the most specific.

4.2 Citation Prediction

In this section, they talk about how citations can effect on the relatedness between papers. They focus on predicting direct citations and predicting co-citations.

Direct citations In this task,there is a given set of candidate papers which include the query paper for the model to predict.The dataset fot evaluating is about 30K and this includes 1K query papers and 25 uncited papers which are also random selected. The goal of this task is to rank the cited papers higher than the uncited papers.

Co-Citations The goal of this task is to predict a highly co-cited paper with a given paper,which is quite similar to the last task. Paper A and B will be highly related,if they always cited together by some papers.The dataset for this task is also 30K papers and is built similar to the last task.

4.3 User Activity

If the similarity of the two papers is close enough so the embeddings of these papers shouble be close too.Therefore they use co-views and co-reads as user activity to identify similar papers and test if the model is able to recover this information.

Co-Views The dataset for this task is about 30K papers.They select 1K random papers which are not involved in train or development set and 25 random papers. Furthermore,the embedding model can rank co-viewed papers higher than the random papers.

Co-Reads They put the activity from the reader such as clicking to access the PDF of a paper as a strong sign of interest in the paper. It is regarded as a “read” action. The dataset for this task is about 30K papers.

4.4 Recommendation

In this task ,they build a system to recommend “similar papers” for a given paper.They make 20K examples for training, 1K examples for validation and 1K examples for testing. The recommendation system includes features like title,author similarity, reference and citation overlap.

5 Results

Before this section,they compare their model results with below methods.[We compare with several strong textual models: SIF(Arora et al., 2017), a method for learning document representations by removing the first principal component of aggregated word-level embeddings which we pretrain on scientific text; SciBERT(Beltagy et al., 2019) a state-of-the-art pretrained Transformer LM for scientific text; and Sent-BERT(Reimers and Gurevych, 2019), a model that uses negative sampling to tune BERT for producing optimal sentence embeddings. We also compare with Citeomatic(Bhagavatula et al., 2018) , a closely related paper representation model for citation prediction which trains content-based representations with citation graph information via dynamically sampled triplets, and SGC (Wu et al., 2019), astate-of-the-art graph-convolutional approach]C.There are many baseline details in appendix.

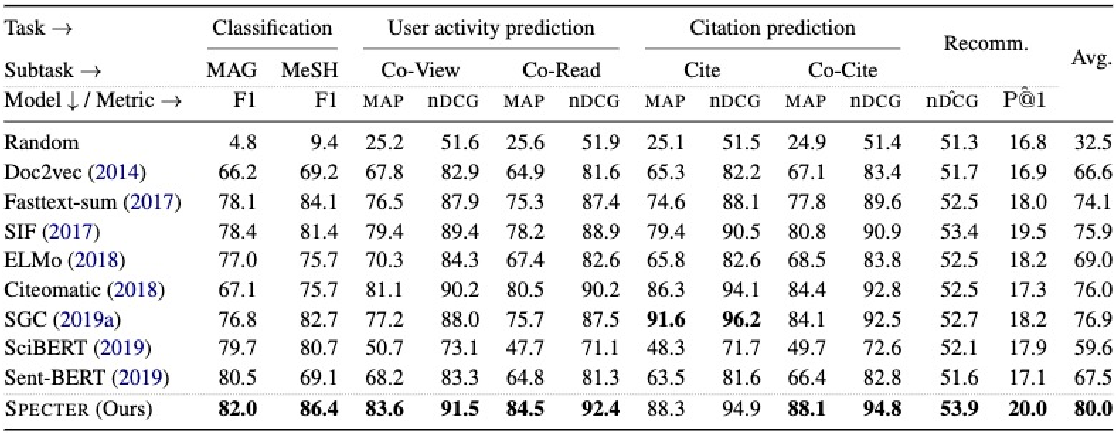
In this section, the results of the paper are presented.First the results of document classification and then recommendation task. Both of the results get substantial improvements. As for document classification,Especially the model outperform on the MeSH dataset.They obtain an 86.4 F1 score which is absolutely increase over the best baseline.Furthermore,they make the SPECTER model reach the 83.8 and 84.5 score on the co-view and co-read task,which improve by 2.7 and 4.0 points than the best baseline. However,in “citation” and “co-citation” tasks, SGC model outperform the SPECTER with 91.6 and 96.2 score. But SGC model can not be used in reality to embed new papers that are not cited yet. Finally.for recommendation task, SPECTER outperform all other models with nDCG of 53.9. But the differences are much smaller.Because embedding variants have less chance to impact on overall performances.

6 Analysis

In this section the work is critically assessed and the relation towards social good in the fields of positive outcomes and public policy is discussed. This is then followed by possible negative consequences of such work and concluded by possible future directions of research.

6.1 Shortcomings

This paper only used MeSH as dataset to classify scientific papers.Additionally, they can also use SNOMED CT(Donnelly et al., 2006) and UMLS(Bodenreider, 2004) as dataset.SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) is a standardized, multilingual vocabulary of clinical terminology that is used by physicians and other health care providers for the electronic exchange of clinical health information.SNOMED CT includes more than 300,000 medical concepts which can be divided into body structure, clinical findings, geographic location and pharmaceutical/biological product. The Unified Medical Language System(UMLS) integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records.The reporting F1 score would have become more informative if they use data from SNOMED CT and UMLS since most of the input scientific papers from medical domain.Another shortcoming is,if they put the author names as input,this will hurt the performance of the model.The reason is that the model cannot relate the author with the document because the author names are sparsed. [We observe that removing the abstract from the textual input and relying only on the title results in a substantial decrease in performance.]C.The abstract provides usually the main topic of the paper and during reading the abstract, the readers can catch the keywords and understand the meaning.There is no doubt that removing the abstract from the input will decrease the performance.



Task →	Classification		User activity prediction				Citation prediction				Recomm.		Avg.
Subtask →	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model ↓ / Metric →	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nDĈG	P@̂1	
Random	4.8	9.4	25.2	51.6	25.6	51.9	25.1	51.5	24.9	51.4	51.3	16.8	32.5
Doc2vec (2014)	66.2	69.2	67.8	82.9	64.9	81.6	65.3	82.2	67.1	83.4	51.7	16.9	66.6
Fasttext-sum (2017)	78.1	84.1	76.5	87.9	75.3	87.4	74.6	88.1	77.8	89.6	52.5	18.0	74.1
SIF (2017)	78.4	81.4	79.4	89.4	78.2	88.9	79.4	90.5	80.8	90.9	53.4	19.5	75.9
ELMo (2018)	77.0	75.7	70.3	84.3	67.4	82.6	65.8	82.6	68.5	83.8	52.5	18.2	69.0
Citeomatic (2018)	67.1	75.7	81.1	90.2	80.5	90.2	86.3	94.1	84.4	92.8	52.5	17.3	76.0
SGC (2019a)	76.8	82.7	77.2	88.0	75.7	87.5	**91.6**	**96.2**	84.1	92.5	52.7	18.2	76.9
SciBERT (2019)	79.7	80.7	50.7	73.1	47.7	71.1	48.3	71.7	49.7	72.6	52.1	17.9	59.6
Sent-BERT (2019)	80.5	69.1	68.2	83.3	64.8	81.3	63.5	81.6	66.4	82.8	51.6	17.1	67.5
SPECTER (Ours)	**82.0**	**86.4**	**83.6**	**91.5**	**84.5**	**92.4**	88.3	94.9	**88.1**	**94.8**	**53.9**	**20.0**	**80.0**

Table 1: Results on the SCIDOCS evaluation suite consisting of 7 tasks

6.2 Relation to positive outcomes

SciBERT can help to contribute to other scientific topics. For example Sharma and Roychowdhury (2019) proposed a QSpider system based on SciBERT which can capture both question types and semantic relations . They used Scibert-scivocab-uncased as the vocabulary for QSpider model,stacked the vectors of 1's and 0's of length 13(number of question types) for question1 and question2 horizontally and use these features to train Gradient Boosting Classifier.The goal of this model is to predict whether two questions are an entailment or not. With the help of SciBERT and Hinge Loss,these got the best performance during Validation phase.

Additionally,it can motivate the use of SciBERT model for various subtasks in Gangwar et al. (2021),such as pre processing,quantity extraction,measured entity and has quantity extraction.SciBERT help their SciBERT + CRF model to get significant improvement in performance over the baseline model and works well in all scientific subtasks.

Furthermore, there is an essay(Anteghini et al., 2020) based on SciBERT about automatically semantify, thereby structure, unstructured bioassay text.In the context of the current Covid-19 pandemic, bioassays are critical, for example, for vaccine development,which reveal the functional and biologically relevant immunological responses that correlate with vaccine efficacy. However, massive volumes of bioassays are being produced and researchers are inundated with this information. They present a solution as a step in the easier knowledge acquisition of bioassays for researchers: the neural-based automated structuring of unstructured, non-standardized bioassays based on the standardized BioAssay Ontology (BAO).

6.3 Relation to Public Policy

Through a series of analysis,the authors show the value and potential to use its paper embeddings,which often perform better and less costly because without fine-tuning.This also give other scientific paper groups more opportunity and more convenience to develop their own system and model.Through their experienments and evaluations, with the foundation of SciBERT,they can promise a new baseline approach.

6.4 Future work

In the future,we can explore more bert based models such as BioBERT(Lee et al., 2020), BlueBERT, ClinicalBERT(Alsentzer et al., 2019), BioMedRoBERT.Additionally,[Another item of future work is to develop better multitask approaches to leverage multiple signals of relatedness information during training.] C. it is also critical to get supported from other metrics from bibliometrics literature(Klavans and Boyack, 2006) to create relatedness graphs.

7 Conclusion

In this paper,a new model based on a Transformer language model is developed for learning representations of scientific papers,which the Transformer language model is pretrained on citations. It also shows how to leverage the power of pretrained language models to learn embeddings for scientific documents. Additionally, a new evaluation method including seven document-levels called SCIDOCS is also introduced and the corresponding datesets are also released. These works exhibit that SciBERT can be used as an efficient and powerful tool to study the relatedness information in scientific papers.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Marco Anteghini, Jennifer D'Souza, Vitor AP Santos, and Sören Auer. 2020. Scibert-based semantification of bioassays in the open research knowledge graph. *arXiv preprint arXiv:2009.08801*.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. *arXiv preprint arXiv:1802.08301*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Liqun Chen, Guoyin Wang, Chenyang Tao, Dinghan Shen, Pengyu Cheng, Xinyuan Zhang, Wenlin Wang, Yizhe Zhang, and Lawrence Carin. 2019. Improving textual network embedding with global attention via optimal transport. *arXiv preprint arXiv:1906.01840*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. arxiv. *arXiv preprint arXiv:1810.04805*.

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Akash Gangwar, Sabhay Jain, Shubham Sourav, and Ashutosh Modi. 2021. Counts@ iitk at semeval-2021 task 8: Scibert based entity and semantic relation extraction for scientific data. *arXiv preprint arXiv:2104.01364*.

Richard Klavans and Kevin W Boyack. 2006. Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2):251–263.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Natraj Raman, Sameena Shah, and Manuela Veloso. 2022. Structure with semantics: Exploiting document relations for retrieval. *arXiv preprint arXiv:2201.03720*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Prakhar Sharma and Sumegh Roychowdhury. 2019. Iit-kgp at mediqa 2019: Recognizing question entailment using sci-bert stacked with a gradient boosting classifier. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 471–477.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Cunchao Tu, Han Liu, Zhiyuan Liu, and Maosong Sun. 2017. Cane: Context-aware network embedding for relation modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1722–1731.

Ashish Vaswani, Noam Shazeer, and Niki Parmar. 2017. Jakob, uszkoreit, llion jones, aidan n. *Gomez, Lukasz, Kaiser, and Illia Polosukhin, "Attention is all you need,", in, NIPS*.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.