Language Modeling of Physics and Computer Science Texts with BERT Models

HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Jeremy Mah Zhee Kein

Supervisors: Prof. Dr. Stefan Rapp, Prof. Dr. Antje Jahn

University of Applied Sciences Darmstadt

Motivation and Goal

With the introduction of the transformer network and its derivative Bidirectional Encoder Representations from Transformers (BERT), contextual embeddings of words can be used to model text data, that contain long-range information. At the beginning, research in this field was based on English text data. This work presents new BERT language models, that are pre-trained with text data from the physics and computer science domains obtained from the open-access repository arXiv [1]. Although **SciBERT** [4] was also pre-trained on computer science texts, it is interesting to investigate, whether further expanding the unlabeled text corpus for pre-training **SciBERT** on computer science texts will improve its performance on downstream tasks in these domains. Hence, the goal of this thesis is to model texts from the physics and computer science domains using a method known as masked language modeling with the help of BERT models initiated with weights from $\mathbf{BERT_{BASE}}$ [5] and **SciBERT**. The models pre-trained on texts from computer science and physics domains and initialized with $\mathbf{BERT_{BASE}}$ and **SciBERT** weights in are known as PCBERT and PCSciBERT, respectively. These models are then evaluated based on their performances in named entity recognition (NER) as a downstream task. Further studies from [7, 3] have also introduced the application of Conditional Random Fields (CRF) on language models for the improvement of their performance in sequence labeling tasks. It would therefore also be in this work's interest to investigate whether the addition of a CRF layer to the pre-trained language models would perform better than their counterparts without the CRF layer.

Methodology

Texts from 1,560,661 research articles were extracted from the computer science and physics categories of the open-access repository arXiv. The extracted texts were then used to build a computer science and physics text corpus. The cased and uncased versions of $\mathbf{BERT_{BASE}}$ and **SciBERT** were pre-trained using masked-language modeling on the built text corpus. To evaluate their language modeling capabilities, the pseudo-perplexities of the models were measured. Subsequently, the models were evaluated on their micro F1 scores as evaluation metrics on two NER datasets. The Workshop on Information Extraction from Scientific Publications (WIESP) [2] and Computer Science Named Entity Recognition in the Open Research Knowledge Graph (CS-NER) [6] datasets were chosen for the physics and computer science domains, respectively. A CRF layer was also added to the models pre-trained in this work to investigate whether their NER performances could be improved.

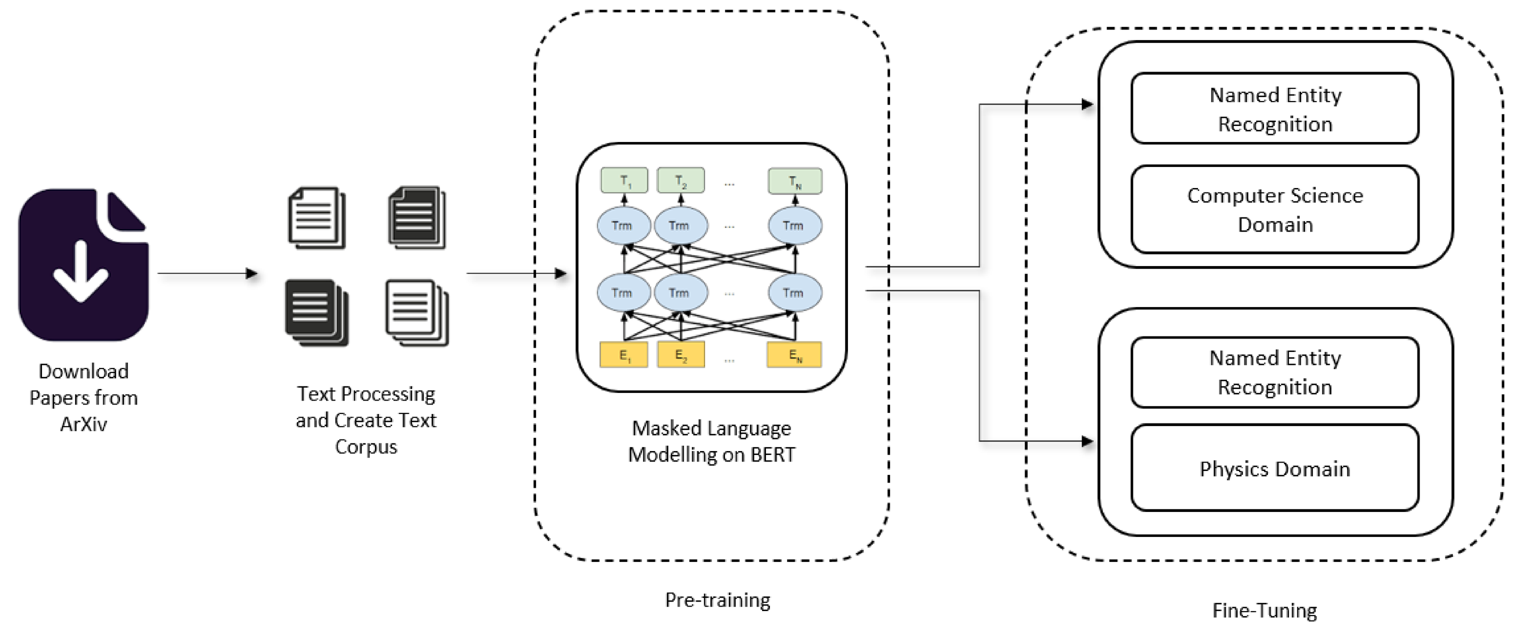

Figure 1. Workflow for producing PCBERT and PCSciBERT models.

Results

We find that the pseudo-perplexities of the models pre-trained in this work on physics and computer science texts have been reduced when compared to their original models. Table 1 shows the pseudo-perplexities achieved before and after pre-training for both cased and uncased variants of $\mathbf{BERT_{BASE}}$ and **SciBERT**. This shows that the language models can model the domain specific language used in computer science and physics texts better after pre-training.

Models	Pseudo-perplexity	
	Before Pre-training	After Pre-training
$\mathbf{BERT_{BASE}}$(cased)	8.26	3.15
$\mathbf{BERT_{BASE}}$(uncased)	10.04	3.40
SciBERT(cased)	7.95	3.46
SciBERT(uncased)	5.28	3.68

Table 1. Pseudo-perplexities of BERT models before and after pre-training on physics and computer science training corpus.

Tables 2 and 3 show the micro F1 scores of the models without the CRF layer in percentages for CS-NER and WIESP datasets, respectively. After fine-tuning on the datasets used for NER, we found that PCSciBERT(cased) outperformed the rest of the other models in terms of micro F1 scores for both datasets. This shows that the SCIVOCAB used by PCSciBERT does model the domain-specific language in the computer science and physics domain better than the general vocabulary used by $\mathbf{BERT_{BASE}}$. Moreover, this also implies that the cased variants of the models perform better than the uncased variants when it comes to NER tasks. Although PCSciBERT(cased) scored the highest F1 scores during fine-tuning, both cased and uncased variants of PCBERT experienced higher improvements in F1 scores than PCSciBERT, showing that pre-training **SciBERT** on texts of related domains can't improve the model as much as when pre-training a model from the general domain.

Models	Micro F1
$\mathbf{BERT_{BASE}}$(cased)	74.88
$\mathbf{BERT_{BASE}}$(uncased)	74.76
SciBERT(cased)	75.53
SciBERT(uncased)	75.41
PCBERT(cased)	75.32
PCBERT(uncased)	75.45
PCSciBERT(cased)	76.22
PCSciBERT(uncased)	75.67

Table 2. Micro F1 scores in percentage of models fine-tuned on CS-NER dataset.

Models	Micro F1
$\mathbf{BERT_{BASE}}$(cased)	77.46
$\mathbf{BERT_{BASE}}$(uncased)	79.12
SciBERT(cased)	80.7
SciBERT(uncased)	80.74
PCBERT(cased)	81.31
PCBERT(uncased)	81.04
PCSciBERT(cased)	82.19
PCSciBERT(uncased)	81.54

Table 3. Micro F1 scores in percentage of models fine-tuned on WIESP dataset.

Furthermore, we also found that the addition of a CRF layer to the models returned inferior performances as compared to the models without the CRF layer. We surmise that the hyperparameters chosen for fine-tuning the models with the CRF layer were sub-optimal and that the intrinsic sequence of the prediction tags was not completely modeled by the CRF layer. Tables 4 and 5 show the micro F1 scores of the models with a CRF layer in percentages for CS-NER and WIESP datasets, respectively.

Models	Micro F1
PCBERT(cased)+CRF	69.57
PCBERT(uncased)+CRF	71.29
PCSciBERT(cased)+CRF	71.46
PCSciBERT(uncased)+CRF	70.74

Table 4. Micro F1 scores of models with CRF layer fine-tuned on CS-NER dataset.

Models	Micro F1
PCBERT(cased)+CRF	79.74
PCBERT(uncased)+CRF	79.82
PCSciBERT(cased)+CRF	81.76
PCSciBERT(uncased)+CRF	80.43

Table 5. Micro F1 scores of models with CRF layer fine-tuned on WIESP dataset.

Conclusion

In this work, a new BERT language model for the computer science and physics domains has been built. The pseudo-perplexities of all the models pre-trained in this work were successfully reduced, showing that the language modeling capability of the models has been improved. Moreover, evaluation of the pre-trained models in this work on NER tasks from computer science and physics domains shows that the downstream performances of these models are better than their original models and that the cased variants perform better than the uncased variants. The addition of a CRF layer did not improve the model's performance and we assume a higher number of training epochs are needed.

References

[1] Open repository arxiv.
`https://arxiv.org/`.
Accessed: 2023-01-19.

[2] Workshop on information extraction from scientific publications ner dataset.
`https://huggingface.co/datasets/adsabs/WIESP2022-NER`.
Accessed: 2023-14-03.

[3] Muhammad Saleh Al-Qurishi and Riad Souissi.
Arabic named entity recognition using transformer-based-CRF model.
In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 262–271, Trento, Italy, 12–13 November 2021. Association for Computational Linguistics.

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

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[6] Jennifer D'Souza and Sören Auer.
Computer science named entity recognition in the open research knowledge graph, 2022.

[7] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo.
Portuguese named entity recognition using bert-crf, 2020.

Gitlab: https://code.fbi.h-da.de/stjemahz/mah-text-processing-ws2223

jeremymah69@gmail.com