Transfer Learning in Biomedical Named Entity Recognition: An Evaluation of BERT in the PharmaCoNER task

Cong Sun
Dalian University of Technology
suncong132@mail.dlut.edu.cn

Zhihao Yang *
Dalian University of Technology
yangzh@dlut.edu.cn

Abstract

To date, a large amount of biomedical content has been published in non-English texts, especially for clinical documents. Therefore, it is of considerable significance to conduct Natural Language Processing (NLP) research in non-English literature. PharmaCoNER is the first Named Entity Recognition (NER) task to recognize chemical and protein entities from Spanish biomedical texts. Since there have been abundant resources in the NLP field, how to exploit these existing resources to a new task to obtain competitive performance is a meaningful study. Inspired by the success of transfer learning with language models, we introduce the BERT benchmark to facilitate the research of PharmaCoNER task. In this paper, we evaluate two baselines based on Multilingual BERT and BioBERT on the PharmaCoNER corpus. Experimental results show that transferring the knowledge learned from source large-scale datasets to the target domain offers an effective solution for the PharmaCoNER task.

1 Introduction

Currently, most biomedical Natural Language Processing (NLP) tasks focus on English documents, while only few research has been carried out on non-English texts. However, it is essential to note that there is also a considerable amount of biomedical literature published in other languages than English, especially for clinical documents. Therefore, it is of considerable significance to conduct NLP research in non-English literature. PharmaCoNER(Gonzalez-Agirre et al., 2019) is the first Named Entity Recognition (NER) task to recognize chemical and protein entities from Spanish biomedical texts. Biomedical NER task is the foundation of biomedical NLP research, which is often utilized as the first step in relation extraction, information retrieval, question answering, etc.

The existing biomedical NER methods can be roughly classified into two categories: traditional machine learning-based methods and deep learning-based methods. Traditional machine learning-based methods (Settles, 2005; Campos et al., 2013; Wei et al., 2015; Leaman et al., 2015, 2016) mainly depend on feature engineering, i.e., the design of useful features using various NLP tools. Overall, this is a labor-intensive and skill-dependent process. In contrast, deep learning-based methods are more promising in biomedical NER tasks. Since deep learning-based methods can automatically learn features, these methods no longer need to construct feature engineering and exhibit more encouraging performance. For examples, (Luo et al., 2017) proposed an attention-based BiLSTM-CRF approach to document-level chemical NER. (Dang et al., 2018) proposed a D3NER model, using CRF and BiLSTM improved with fine-tuned embeddings of various linguistic information to recognize disease and protein/gene entities. Recently, the language model pre-training (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019) has proven to be effective for improving many NLP tasks. The fine-tuning language model (Radford et al., 2018; Devlin et al., 2019) can transfer the knowledge learned from large-scale datasets to domain-specific tasks by simply fine-tuning the pre-trained parameters.

Inspired by the success of transfer learning with language models, we would like to make full use of the existing language model resources to implement the PharmaCoNER task. In this paper, we introduce the BERT (Devlin et al., 2019) benchmark to facilitate the research of PharmaCoNER task. We regard the large-scale dataset used to train the BERT model as the source do-

*Corresponding author

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 100–104

main, and the PharmaCoNER dataset as the target domain, thus considering the PharmaCoNER task as a transfer learning problem. We evaluate two baselines based on Multilingual BERT and BioBERT. Experimental results show that transferring the knowledge learned from source large-scale datasets to the target domain offers an effective solution for the PharmaCoNER task.

2 Related Work

2.1 Language Model

Learning widely used representations of words has been an active area of research for decades. To date, pre-trained word embeddings are considered to be an integral part of modern NLP systems, offering significant improvements over embeddings learned from scratch (Turian et al., 2010). Recently, ELMo (Peters et al., 2018) has been proposed to generalize traditional word embedding research (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017) to extract context-sensitive features. When integrating contextual word embeddings with existing task-specific architectures, ELMo achieves competitive performance for many major NLP benchmarks. More recent studies (Radford et al., 2018; Devlin et al., 2019) tend to exploit language models to pre-train some model architecture on a language model objective before fine-tuning that the same model for downstream tasks. BERT (Devlin et al., 2019), which stands for Bidirectional Encoder Representations from Transformers (Vaswani et al., 2017), is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. The pre-trained BERT can be fine-tuned to create competitive models for a wide range of tasks.

2.2 Transfer Learning

Many machine learning methods work well only under a common assumption: the training and test data are drawn from the same feature space and distribution (Pan and Yang, 2009). When the distribution changes, most models need to be rebuilt from scratch using newly annotated training data. However, it is an expensive and challenging process. Therefore, it would be meaningful to reduce the need and effort to recollect the annotated training data. In such scenarios, transfer learning between task domains would be useful. For example, (Cui et al., 2018) demonstrate the effects of transfer learning in the computer vision domain. They explore transfer learning via fine-tuning the knowledge learned from large-scale datasets to small-scale domain-specific fine-grained visual categorization datasets. For NLP tasks, (Conneau et al., 2017) and (McCann et al., 2017) also demonstrate the effects of transfer learning on the natural language inference and machine translation tasks, respectively. These methods demonstrate the significance of transfer learning in machine learning methods.

3 Methods

3.1 Problem Definition

The PharmaCoNER task is structured into two sub-tracks: 'NER offset and entity classification' and 'concept indexing'. Since we only participate in the first track, we will explain the fist track in detail. There are three entity types for evaluation in the PharmaCoNER corpus, namely 'normalizables', 'notnormalizables' and 'proteins'. Specifically, 'normalizables' is the mentions of chemicals that can be manually normalized to a unique concept identifier. 'notnormalizables' is the mentions of chemicals that could not be normalized manually to a unique concept identifier. 'proteins' is the mentions of proteins and genes. We used the extended BIO (Begin, Inside, Other) tagging scheme in our experiments. Formally, we formulate the PharmaCoNER task as a multi-class classification problem. Given an input sequence $S = \{w_1, \cdots, w_i, \cdots, w_n\}$ which has processed by WordPiece, the goal of PharmaCoNER is to classify the tag t of token w_i. Essentially, the model estimates the probability $P(t|w_i)$, where $T = \{$B-normalizables, I-normalizables, B-notnormalizables, I-notnormalizables, B-proteins, I-proteins, O, X, CLS, SEP$\}$, $t \in T$, $1 \leq i \leq n$.

3.2 Model Architecture

BERT (Devlin et al., 2019), which stands for bidirectional encoder representations from Transformers, is designed to learn deep bidirectional representations by jointly conditioning on both left and right context in all layers. The architecture of BERT is illustrated in Figure 1. The pre-trained BERT can be fine-tuned to create competitive models for a wide range of downstream tasks, such as named entity recognition, relation extraction, and question answering.

Here, we explain the architecture of BERT for

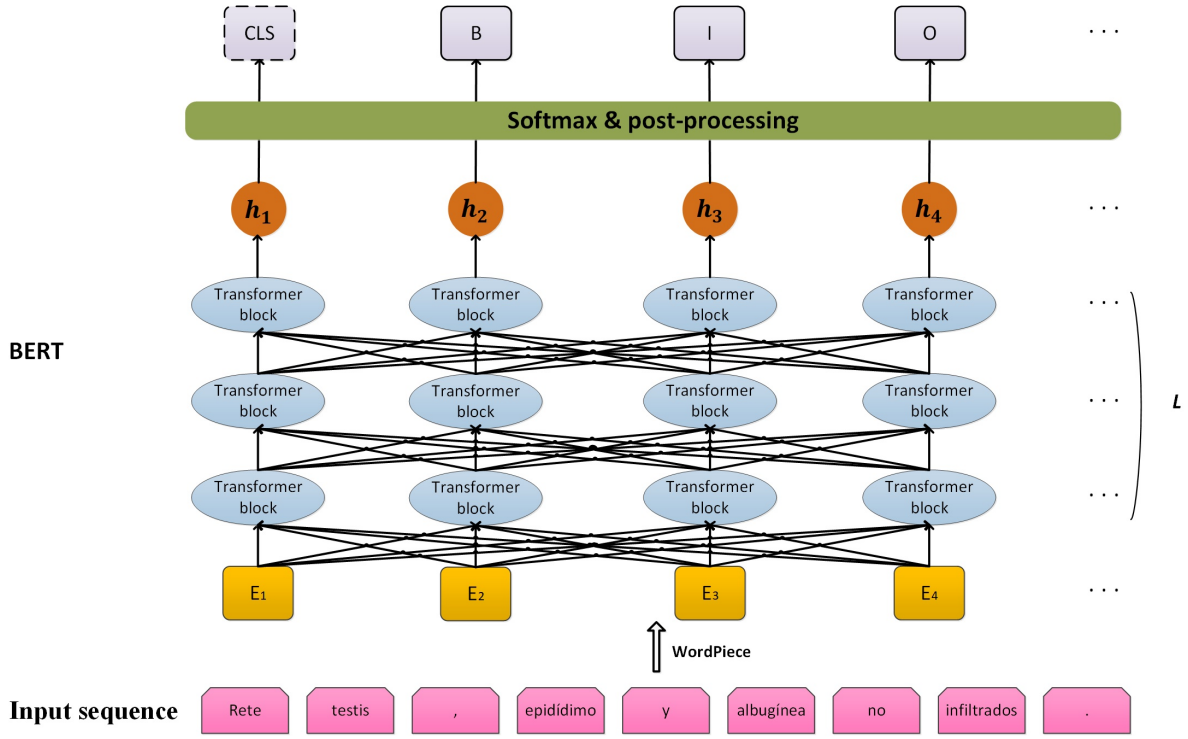

Figure 1: The Architecture of BERT.

NER tasks. The input of BERT can represent both a single text sentence or a pair of text sentences in one sequence. BERT differentiates the text sentences as follows: first, they separate them with a special token ([SEP]); second, they add the sentence A embedding to every token of the first text sentence and the sentence B embedding to every token of the second text sentence. Furthermore, every sequence starts with a special token ([CLS]). For a given token, the input representation is constructed by integrating the corresponding token, segment, and position embeddings. BERT provides two model sizes: BERT_{BASE} and BERT_{LARGE}. For the BERT model, the number of layers L, the hidden size H and the number of self-attention heads A are listed as follows:

- BERT_{BASE}: L=12, H=768, A=12, Total Parameters=110M.
- BERT_{LARGE}: L=24, H=1024, A=16, Total Parameters=340M.

During the shared task, we exploit Multilingual BERT (Devlin et al., 2019) and BioBERT (Lee et al., 2019) to implement the PharmaCoNER task. Both the multilingual BERT and BioBERT models are pre-trained based on the BERT_{BASE} size. The multilingual BERT model is pre-trained on Wikipedia in multiple languages. The BioBERT model is pre-trained on Wikipedia, BooksCorpus, PubMed (PubMed abstracts) and PMC (PubMed Central full-text articles). The pre-training process of Multilingual BERT and BioBERT is similar to the pre-training process of BERT_{base}. More details about Multilingual BERT and BioBERT can be found in the studies (Devlin et al., 2019; Lee et al., 2019).

For the output layer, we feed the final hidden representation h_i of each token i into the softmax function. The probability P is calculated as follows:

$$P(t|h_i) = softmax(W_o h_i + b_o) \quad (1)$$

where T = {B-normalizables, I-normalizables, B-notnormalizables, I-notnormalizables, B-proteins, I-proteins, O, X, CLS, SEP}, $t \in T$, W_o and b_o are weight parameters. Furthermore, during the training, we use the categorical cross-entropy as the loss function. Finally, as shown in Figure 1, we removed the special tokens (labeled by 'X', 'CLS' and 'SEP') and obtained the final BIO labels at the post-processing step.

4 Results and Discussion

4.1 Experimental Settings

In this section, we introduce the dataset, evaluation metrics and details of the training process of our model.

Dataset. The PharmaCoNER corpus has been randomly sampled into three subsets: the training set, the development set and the test set. The training set contains 500 clinical cases, and the development set and the test set include 250 clinical cases, respectively.

Evaluation Metrics. We apply the standard measures precision, recall and micro-averaged F1-score to evaluate the effectiveness of our model. These metrics are also adopted as the evaluation metrics during the PharmaCoNER task.

Training Details. During the PharmaCoNER task, we utilized the training set for training the model and exploited the development set to choose the hyper-parameters of our model. In the prediction stage, we combined the training and development sets for training our model, and the organizers used the gold-standard test set to evaluate the final results. The detailed hyper-parameter settings are illustrated in Table 1. 'Opt.' denotes optimal.

Parameters	Tuned range	Opt.
Sequence length	128	128
Train batch size	[8, 16, 32]	32
Dev batch size	8	8
Test batch size	8	8
Learning rate	[1e-5, 2e-5, 3e-5]	2e-5
Epoch number	[10, 50, 100, 200]	100
Warmup	0.1	0.1
Dropout	0.1	0.1

Table 1: Detailed Hyper-parameter Settings in the PharmaCoNER task.

4.2 Experimental Results

We applied Multilingual BERT and BioBERT on the PharmaCoNER corpus, respectively. The experimental results are shown in Table 2. 'P', 'R', 'F' denote precision, recall, and micro-averaged F1-score, respectively. It is encouraging to see that the performance of both models is quite competitive. For the multilingual BERT model, since the model learned the Spanish language information during the pre-training process, its F1-score is higher, reaching 89.24%. For the BioBERT model, it also achieves an F1-score of 89.02%. While BioBERT was only pre-trained on the English biomedical texts, applying it to the Spanish PharmaCoNER task still yields competitive performance. The primary reason may be that there are a large number of chemical and protein mentions sharing the same name in English and Spanish in biomedical literature. Therefore, it is feasible to use the existing model pre-trained on English biomedical corpora to fine-tune the PharmaCoNER task. These results indicate that transferring the knowledge learned from source large-scale datasets via fine-tuning to the target-specific domain is an effective solution to the PharmaCoNER task.

Models	P(%)	R(%)	F(%)
Multilingual BERT	90.46	88.06	89.24
BioBERT	90.70	87.41	89.02

Table 2: The Experimental Results of Multilingual-BERT and BioBERT.

Furthermore, we manually analyzed the errors generated by our models on the corpus test set after the PharmaCoNER task. The main errors can be classified into three categories: (1) incorrect boundaries, (2) missing the chemical/protein mention, (3) and incorrectly distinguishing the chemical and protein mentions. By analyzing these error examples, we infer that document-level information or biomedical knowledge may be helpful for the PharmaCoNER task.

5 Conclusion

In this paper, we introduce the BERT benchmark to facilitate the research of PharmaCoNER task. We evaluate two baselines based on Multilingual BERT and BioBERT on the PharmaCoNER corpus. It is encouraging to see that the performance of both models is quite competitive, reaching F1-scores of 89.24% and 89.02%, respectively. Experimental results demonstrate that transferring the knowledge learned from source large-scale datasets to the target domain offers an effective solution for the PharmaCoNER task.

In future work, we would like to explore an appropriate way to integrate document-level information or biomedical knowledge to improve the performance of the model.

Acknowledgments

This work was supported by the grants from the National Key Research and Development Program of China (No.2016YFC0901902), Natural Science Foundation of China (No.61272373, 61572102 and 61572098), and Trans-Century Training Program Foundation for the Talents by the Ministry of Education of China (NCET-13-0084).

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. Gimli: open source and high-performance biomedical name recognition. *BMC bioinformatics*, 14(1):54.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. 2018. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118.

Thanh Hai Dang, Hoang-Quynh Le, Trang M Nguyen, and Sinh T Vu. 2018. D3ner: biomedical named entity recognition using crf-bilstm improved with fine-tuned embeddings of various linguistic information. *Bioinformatics*, 34(20):3539–3546.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of the BioNLP Open Shared Tasks (BioNLP-OST)*.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmchem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, 7(1):S3.

Robert Leaman, Chih-Hsuan Wei, Cherry Zou, and Zhiyong Lu. 2016. Mining chemical patents with an ensemble of open systems. *Database*, 2016.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2017. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2227–2237.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

Burr Settles. 2005. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015.