Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base

Boran Hao[*]
Boston University
brhao@bu.edu

Henghui Zhu[‡*]
AWS AI
henghui@amazon.com

Ioannis Ch. Paschalidis [†]
Boston University
yannisp@bu.edu

Abstract

Domain knowledge is important for building Natural Language Processing (NLP) systems for low-resource settings, such as in the clinical domain. In this paper, a novel joint training method is introduced for adding knowledge base information from the Unified Medical Language System (UMLS) into language model pre-training for some clinical domain corpus. We show that in three different downstream clinical NLP tasks, our pre-trained language model outperforms the corresponding model with no knowledge base information and other state-of-the-art models. Specifically, in a natural language inference task applied to clinical texts, our knowledge base pre-training approach improves accuracy by up to 1.7%, whereas in clinical name entity recognition tasks, the F1-score improves by up to 1.0%. The pre-trained models are available at https://github.com/noc-lab/clinical-kb-bert.

1 Introduction

Natural language processing (NLP) is increasingly becoming an important tool in medical research. Different NLP applications have been developed for assisting physicians, helping to increase the efficiency of performing a variety of tasks. For example, the clinical concept extraction task (Uzuner et al., 2011) labels clinical findings, treatments, and tests contained in a clinical report, whereas the temporal relations extraction task (Sun et al., 2013) identifies temporal expressions of the clinical concept. The need for these applications calls for better clinical natural language models.

The recent advances in language model pre-training greatly boosted the performance of NLP models. Compared with the traditional NLP- or word2vec-based techniques, the pre-trained language models capture the meaning of the words based on the context of each sentence. Some pre-trained language models, including BERT (Devlin et al., 2019), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019), achieved state-of-the-art performance on most NLP tasks. These models can be further fine-tuned on a domain-specific corpus for specific tasks.

In addition to a language model, knowledge base information can help to build a better NLP model. During language model pre-training, the representation of the words is learned from some given corpus. Thus, the language model would not be able to produce a meaningful representation of a concept missing from the corpus. This problem is critical, especially in a low-resource setting like the clinical domain. On the other hand, a knowledge base contains a plethora of concepts and relations. Leveraging such information has the potential to yield models that can generalize better without resorting to extremely large training corpora. For example, the Unified Medical Language System (UMLS) provides 54 types of relations between biomedical concepts of 127 semantic types. It contains 27.7 million relations, which, as we show, can be leveraged to improve clinical NLP models for various downstream tasks.

[*] Equal contribution. ‡ The work was done when Henghui Zhu was at Boston University.

[†] Research partially supported by the NSF under grants IIS-1914792, DMS-1664644, and CNS-1645681, by the ONR under grant N00014-19-1-2571, and by the NIH under grants R01 GM135930 and UL54 TR004130.

Proceedings of the 28th International Conference on Computational Linguistics, pages 657–661
Barcelona, Spain (Online), December 8-13, 2020

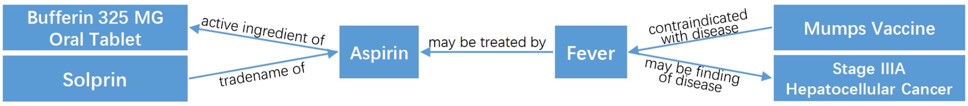
Bufferin 325 MG Oral Tablet
Solprin
active ingredient of
tradename of
Aspirin
may be treated by
Fever
contraindicated with disease
may be finding of disease
Mumps Vaccine
Stage IIIA Hepatocellular Cancer

Figure 1: A snippet of the UMLS knowledge base graph. The nodes (blue boxes) are clinical concepts and the edges indicate concept relations.

In this paper, we propose a novel method for fusing the knowledge base information into the language model pre-training stage. Specifically, we use a negative sampling method to convert the relations in the knowledge base into a classification problem and introduce a new loss function during the pre-training that shapes the representation of the concept according to the relation graph in the knowledge base. We obtained two different pre-trained language models from BioBERT (Lee et al., 2020) and ALBERT (Lan et al., 2019) and achieved state-of-the-art performance in two downstream NLP tasks. Finally, an ablation study is performed for showing the effectiveness of our proposed method.

2 Related Work

Pre-trained language model in the clinical domain. There are many previous works for pre-trained language models using biomedical domain corpora. For example, (Zhu et al., ; Alsentzer et al., 2019; Lee et al., 2020) pre-trained different language models on the biomedical and clinical corpora and the models showed significant improvement on various downstream tasks. Also, (Si et al., 2019) compared the performance of a model pre-trained with an open domain corpus to a model pre-trained with a clinical domain corpus, and concluded that the latter achieved superior performance.

Knowledge base integrated NLP models. There are also some works that integrated knowledge base information into NLP models and improved the performance in downstream tasks. (Zhang et al., 2019) built some NLP models that integrated information from the UMLS knowledge base and showed a better performance compared to baselines. Also, the knowledge base information can be utilized during language model pre-training and some works, including (Lauscher et al., 2019; Peters et al., 2019; Yao et al., 2019), have shown promising results in this direction.

3 Methods

3.1 Dataset

We use two datasets in this work. The MIMIC-III (Johnson et al., 2016) was used for mask language model pre-training. It consists of almost 2 million de-identified Intensive Care Unit (ICU) reports. In addition, the UMLS (ver. 2019AA) knowledge base (Bodenreider, 2004) is used for introducing the knowledge information into the pre-trained language model. UMLS provides a multi-relational graph; nodes correspond to clinical concepts and edges to relations. Each edge is represented as a triplet (concept1, relation, concept2), indicating the relation between two concepts. A snippet of the UMLS knowledge base is illustrated in Figure 1. In this example, a triplet (Fever, may be treated by, Aspirin) is presented.

3.2 Pre-trained Language Model with Knowledge Base Information

We start with a standard way of training language models on a clinical domain corpus. Two types of language models are considered, BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2019). The masked language model and the next sentence prediction loss are used during pre-training.

We utilize information from the UMLS knowledge base as follows. For two medical concepts and a relation in the UMLS, the transformer model aims to predict if the relation exists between the two concepts. Specifically, we build a binary classifier based on the transformer model for a triplet (concept 1, relation, concept 2). The input of the transformer is of the format `[CLS] concept1 [relation] concept2 [SEP]`, where `[relation]` is a special token for the relation in the vocabulary. Different relations are represented with different types of `[relation]` tokens. We use the output of `[CLS]` token to predict if the relation exists.

The pre-training technique we applied is as follows. The final loss consists of the masked language model loss, L_m, the next sentence prediction loss, L_n, for the language model pre-training, and the triplet classification loss $\tilde{L}_t$. Specifically, we use the approach of (Devlin et al., 2019) for sampling the mask tokens in pre-training. We shuffle the training examples in each batch and set the ratio of the language model pre-training examples over the knowledge base examples to roughly 4/3. The final loss used is $L = 4L_m + L_n + \tilde{L}_t$.

3.3 Negative Sampling for Knowledge Base Triplets

The knowledge graph in the UMLS is very sparse. In particular, if one randomly samples a relation and two concepts, it is extremely rare for a triplet to exist. Instead, we propose a negative sampling method to overcome the imbalance when using random triplet sampling. Specifically, given a positive triplet (concept 1, relation, concept 2), we sample two concepts, say concept 3 and concept 4 with no such relation as a negative triplet, i.e., (concept 3, relation, concept 4). However, if the concepts of the negative triplet are randomly sampled from the UMLS, then the classification problem becomes trivial since classifying concept types suffices. As a result, such a sampling procedure would not necessarily result in a model substantially different from the one without the knowledge base. To address this issue, we force concepts 1 and 3, and concepts 2 and 4, respectively, to be of the same UMLS semantic type. Therefore, the resulting triplet classification is less trivial, helping to build better pre-trained models.

4 Experimental Results

To evaluate our pre-trained clinical models, we selected two *Named Entity Recognition (NER)* tasks: i2b2 2010 (Uzuner et al., 2011) and i2b2 2012 (Sun et al., 2013), and one *Natural Language Inference (NLI)* task MedNLI (Romanov and Shivade, 2018). Table 1 shows the number of classes and how the dataset was split for each task.

Table 1: Number of classes and training/test samples in each dataset.

	MedNLI	i2b2 2010	i2b2 2012
Classes	3	7	13
Training samples	11232	16315	7446
Test samples	1422	27625	5665

4.1 Experimental Setup

For the language model and knowledge base pre-training, we consider the following two settings. We call *Clinical KB-BERT* the language model starting from the BioBERT (Lee et al., 2020) model and *Clinical KB-ALBERT* the model starting from the ALBERT xxlarge (Lan et al., 2019) model and, in both cases, using the proposed pre-training method on MIMIC-III and UMLS. We use the Adam optimizer (Kingma and Ba, 2014) with learning rates 5×10^{-5} and 2×10^{-5}, and a batch size of 32 and 16 for the Clinical KB-BERT and Clinical KB-ALBERT, respectively. In addition, when training the Clinical KB-ALBERT, we use a stochastic gradient descent optimizer for a warmup training over 10,000 steps to prevent the model from diverging. For the downstream tasks, the same batch size and learning rates are used, as described above. We follow the NER and NLI settings from (Devlin et al., 2019) for the downstream tasks. All the tasks are fine-tuned using the corresponding datasets for 3 epochs. For the Clinical KB-BERT model, we applied a max sequence length of 150, which is the same as in (Alsentzer et al., 2019). For the ALBERT-based model, we used a max sequence length of 128, due to the constraint of graphic memory and the fact that ALBERT models use a different sentence-piece tokenizer.

4.2 Experimental Results

For the two NER tasks, we use the exact F1-score of the span (Uzuner et al., 2011) as the performance metric. And for the MedNLI task, the classification accuracy is used. The results are shown in in Table 2. We achieved state-of-art performance for 2 out of 3 tasks.

Table 2: Accuracy (MedNLI) and Exact F1-score (i2b2) across various clinical NLP tasks.

Model	MedNLI	i2b2 2010	i2b2 2012
BioBERT (Lee et al., 2020)	80.8%	86.5%	78.9%
Bio+Discharge Summary BERT (Alsentzer et al., 2019)	82.7%	87.8%	78.9%
Clinical BERT + biLSTM (Si et al., 2019)	-	**90.3%**	80.9%
Our Clinical KB-ALBERT	**84.4%**	89.7%	**81.9%**

5 Discussion

The following ablation study is performed to verify if the knowledge base helps improve the performance of the pre-trained model. We pre-trained an ALBERT model using MIMIC-III without knowledge base information, and we reported the downstream model performances for different model configurations (see Table 3; LM stands for Language Model; KB stands for Knowledge Base). Also, in order to compare against the results in (Alsentzer et al., 2019), we added span correction for NER by default, which converts the prediction of an I's tag after a O's tag into the corresponding B's tag. As shown in Table 3, the span correction hurts the NER performance.

Table 3: Ablation study for pre-trained models.

Model	MedNLI	i2b2 2010	i2b2 2012
Clinical KB-BERT - span correction	**84.1%**	**88.8%**	**81.5%**
Clinical KB-BERT	84.1%	87.9%	80.0%
- KB pre-training	82.7%	87.2%	78.9%
- in-domain LM pre-training	77.6%	83.5%	75.9%
Clinical KB-ALBERT - span correction	**84.4%**	**89.7%**	**81.9%**
Clinical KB-ALBERT	84.4%	89.3%	80.6%
- KB pre-training	84.0%	88.9%	80.5%
- in-domain LM pre-training	82.3%	87.8%	79.1%

From Table 3, in-domain language modeling (LM) pre-training contributes the most in boosting performance. This has already been shown in (Alsentzer et al., 2019). Adding the UMLS knowledge helps the models pre-trained in the clinical domain. Under BERT's setting, our Clinical KB-BERT has an accuracy of 84.1% for the MedNLI task, outperforming the model without knowledge base pre-training by 1.4%. Also, the exact F1-scores for the i2b2 2010 and 2012 tasks were improved by 1.6% and 2.6%, respectively, by adding the knowledge base information. For ALBERT, the improvement is less significant since the original ALBERT model already achieves a good performance; specifically, the improvement we achieve over in-domain LM pre-training is 0.4% for the MedNLI task and 0.8% and 1.4%, respectively, for the i2b2 2010 and 2012 tasks.

For the NER tasks, we follow the NER setting in (Devlin et al., 2019). We have achieved a similar performance compared to models with additional bidirectional LSTM layers (Si et al., 2019). For the i2b2 2012, task, our ALBERT model outperforms previous baselines. This suggests the effectiveness of the pre-trained language model with knowledge base information.

6 Conclusion

We proposed a simple but effective way to include knowledge base information into clinical domain-specific language model pre-training. Two pre-trained models, Clinical KB-BERT and Clinical KB-ALBERT, are obtained and are shown to achieve state-of-art results for two types of downstream tasks. An ablation study was performed to show the effectiveness of including the knowledge base. Future work will evaluate the proposed models on additional clinical downstream tasks.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

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

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2019. Informing unsupervised pretraining with external linguistic knowledge. *arXiv preprint arXiv:1909.02339*.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. 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)*, pages 43–54.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Xiao Zhang, Dejing Dou, and Ji Wu. 2019. Learning conceptual-contexual embeddings for medical text. *arXiv preprint arXiv:1908.06203*.

Henghui Zhu, Ioannis C Paschalidis, and Amir M Tahmasebi. Clinical concept extraction with contextual word embedding. In *NIPS Machine Learning for Health Workshop*. Available as arXiv:1810.10566.