DX-HITSZ at BioNLP-OST 2019: Trigger Word Detection and Thematic Role Identification via BERT and Multitask Learning

Dongfang Li[1], **Ying Xiong**[1], **Baotian Hu**[1], **Hanyang Du**[1]
Buzhou Tang[1,2], **Qingcai Chen**[1,2]
[1]Harbin Institute of Technology (Shenzhen), Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
{crazyofapple, xiongying0929, dhy1996525}@gmail.com
{hubaotian, tangbuzhou, qingcai.chen}@hit.edu.cn

Abstract

The prediction of the relationship between the disease with genes and its mutations is a very important knowledge extraction task that can potentially help drug discovery. In this paper, we present our approaches for trigger word detection (task 1) and the identification of its thematic role (task 2) in AGAC track of BioNLP Open Shared Task 2019. Task 1 can be regarded as the traditional name entity recognition (NER), which cultivates molecular phenomena related to gene mutation. Task 2 can be regarded as relation extraction which captures the thematic roles between entities. For two tasks, we exploit the pre-trained biomedical language representation model (i.e., BERT) in the pipe of information extraction for the collection of mutation-disease knowledge from PubMed. And also, we design a fine-tuning technique and extra features by using multi-task learning. The experiment results show that our proposed approaches achieve 0.60 (ranks 1) and 0.25 (ranks 2) on task 1 and task 2 respectively in terms of F_1 metric.

1 Introduction

Using the natural language processing methods to discover and mine drug-related knowledge from text has been a hot topic in recent years. For the goal of drug repurposing, an active gene annotation corpus (AGAC) was developed as a benchmark dataset (Wang et al., 2018b). The AGAC track is part of the BioNLP Open Shared Task 2019, aims to gather text mining approaches among the BioNLP community to propel drug-oriented knowledge discovery. It consists of three tasks for the extraction of mutation-disease knowledge from PubMed abstracts: trigger words NER, thematic roles identification, and mutation-disease knowledge discovery. We participated in the trigger words NER and thematic roles identification tasks.

Recently, pre-trained models have been the dominant paradigm in natural language processing. They achieved remarkable state-of-the-art performance across a wide range of related tasks, such as textual entailment, natural language inference, question answering, etc. BERT, proposed by Devlin et al. (2019), has achieved a better-marked result in GLUE leaderboard with a deep transformer architecture (Wang et al., 2018a). BERT first trains a language model on an unsupervised large-scale corpus, and then the pre-trained model is fine-tuned to adapt to downstream tasks. This fine-tuning process can be seen as a form of transfer learning, where BERT learns knowledge from the large-scale corpus and transfer it to downstream tasks. While BERT was built for general-purpose language understanding, there are also some pre-trained models following BERT architecture that effectively leverage domain-specific knowledge from a large set of unannotated biomedical texts (e.g. PubMed abstracts, clinical notes), such as SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2019), NCBI BERT (Peng et al., 2019), etc. These models can effectively transfer knowledge from a large amount of unlabeled texts to biomedical text mining models with minimal task-specific architecture modifications.

In this paper, we investigate different methods to combine and transfer the knowledge from the three different sources and illustrate our results on the AGAC corpus. Our method is based on fine-tuning BERT_{base}, NCBI BERT and BioBERT using multi-task learning, which has demonstrated the efficiency of knowledge transformation (Liu et al., 2019) and integrating models for both tasks with ensembles. The proposed methods are proved effective for natural language understanding in the biomedical domain, and we rank first place on task 1 (Trigger words NER) and second place on task 2 (Thematic roles identification).

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 72–76
Hong Kong, China, November 4, 2019.

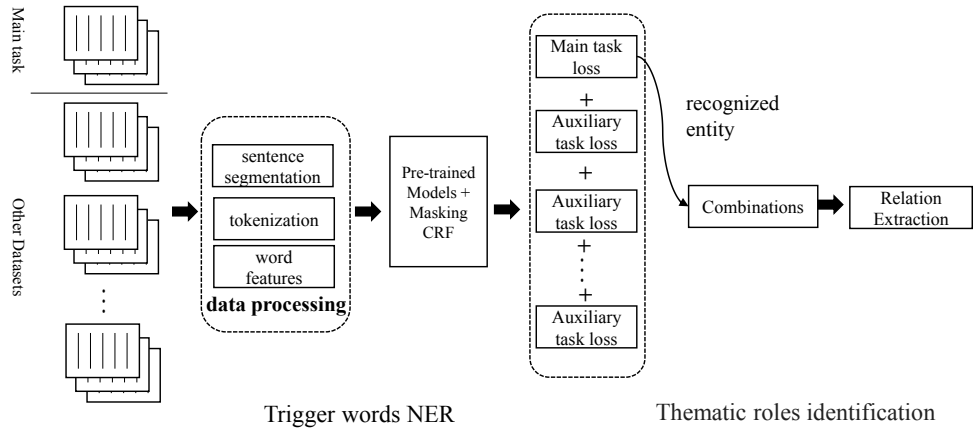

Figure 1: The pipeline of our approach. We first split PubMed abstracts into sentences, tokenize them into words and extract some features like POS tags, then a BERT-based method for NER offset and entity recognition, and finally predict relations for each potential entity pair.

2 Background

The model architecture of BERT (Devlin et al., 2019) is a multi-layer bidirectional Transformer encoder based on the original Transformer model (Vaswani et al., 2017). The input representation is a concatenation of WordPiece embeddings (Wu et al., 2016), positional embeddings, and the segment embedding. A special classification embedding ([CLS]) is inserted as the first token and a special token ([SEP]) is added as the final token. It is firstly pre-trained with two strategies on large-scale unlabeled text, i.e., masked language model and next sentence prediction. The pre-trained BERT model provides a powerful context-dependent sentence representation and can be used for various target tasks, i.e., text classification and machine comprehension, through the fine-tuning procedure.

Hence, the BERT model can be easily extended to the medical domain information extraction pipeline, first extracting the trigger words and then determining the relationship between them, as illustrated in Figure 1.

3 Our Approach

3.1 Task 1: Trigger Words NER

Task 1 aims to identify trigger words in the PubMed digest and annotating them as correct trigger markers or entities (Var, MPA, Interaction, Pathway, CPA, Reg, PosReg, NegReg, Disease, Gene, Protein, Enzyme). It can be seen as an NER task involving the identification of many domain-specific proper nouns in the biomedical corpus.

We first split each PubMed abstracts into sentences using '\n' or '.', and convert each sentence into words by NLTK[1] tokenizer. After that, words are further tokenized into its word pieces $\boldsymbol{x} = (x_1, \ldots, x_T)$. Then we use a representation based on the BERT from the last layer $\mathbf{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T)$. In order to make better use of the word-level information, POS tagging labels and word shape embedding representation (Liu et al., 2015) of each word [2] are also concatenated into the output of BERT, passing through a single projection layer, followed by a conditional random fields (CRF) layer with a masking constraint [3] to calculate the token-level label probability $\boldsymbol{p} = (p_1, \ldots, p_T)$. When fine-tuning the BERT, we found that the performance of the model performed better in the case of BIO for the selection of the tagging schemes compared to BIOES. We further extend our model to multi-task learning joint trained by sharing the architecture and parameters. Although the differences in different datasets, multi-task means joint learning with other biomedical corpora. The assumption is to make more efficient use of the data and to encourage the models to learn more generalized representations. More specially, the same token-level information and BERT encoder are shared and each data set has a specific output layer, e.g., CRF layer. Our final loss function is obtained as follows:

$$-\sum \lambda_{c_i} log P(y_{c_i}|x_{c_i}) + \lambda_r \|W\|_2 \quad (1)$$

[1] https://www.nltk.org/

[2] If a word is tokenized into several tokens, each token will be given the same tagging labels.

[3] Transition mask with invalid moves as 0 and valid as 1.

where y_{c_i} denote true tag sequence and x_{c_i} denote the input tokens for corpora c_i, λ_{c_i} and λ_r are weighted parameters.

3.2 Task 2: Thematic Roles Identification

Task 2 is to identify the thematic roles (ThemeOf, CauseOf) between trigger words.

We treat it as a multi-label classification problem by introducing "no relation (NA)" label. When constructing the training data of task 2, we use the relationship of two entities with a distance of no more than one sentence. For NA label, random sampling is performed. In the testing process, relation label will be assigned to the corresponding thematic role when its probability is maximum and larger than the threshold. Otherwise, it will be predicted as no relation. We also anonymously use a predefined tag (such as $\%Disease$) to represent a target named entity. And we additionally append two concrete predicted entity words separated by the [SEP] tag after each sentence. Following Shi and Lin (2019), we also add the token-level relative distance to the subject entity information for each token, i.e. 0 for the position t between two entities, $t - s$ for tokens before first entity and $t - e$ for tokens after second entity, where s, e are the starting and ending positions of first and second entity after tokenization, respectively. The relation logits of two entities are performed using a single output layer from the BERT, as

$$y = \text{softmax}(\mathbf{W}\boldsymbol{h}_{cls} + \boldsymbol{b}) \quad (2)$$

where $\boldsymbol{h}_{cls}$ denotes the hidden state of the first special token ([CLS]).

4 Experiments

In this section, we provide the leaderboard performance and conduct an analysis of the effect of models from different settings.

4.1 Experimental Setup

The AGAC track organizers develop an active gene annotation corpus (AGAC) (Wang et al., 2018b; Gachloo et al., 2019), for the sake of knowledge discovery in drug repurposing. The track corpus consists of 1250 PubMed abstracts: 250 for public, 1000 for final evaluation. We randomly split the public texts into train and development data sets with the radio of 8:2. The training set is used to learn model parameters, the development set to select optimal hyper-parameters. For evaluation results, we measure the trigger words recognition and thematic roles extraction performance with F_1 score. Table 1 shows the external data sets used under the joint learning method. The BIO form of these data sets is different from that of task 1, hence we use different projection and CRF layers. But not the more data sets, the better. We found that the NCBI disease (Doğan et al., 2014) and BC5CDR (Li et al., 2016) datasets are helpful for the final results, and the performance is reduced when using BC2GM (Smith et al., 2008) and 2010 i2b2VA dataset (Uzuner et al., 2011).

Dataset	#Train	#Dev	#Test
BC5CDR	4,559	4,580	4,796
NCBI disease	5,423	922	939
BC2GM	12,573	2,518	5,037
2010 i2b2/VA	16,315	-	27,626

Table 1: Datasets for joint learning in recognizing the trigger words.

4.2 Implementation and Hyperparameters

We tried the original BERT[4], BioBERT[5] and NCBI BERT[6] pre-trained models. Each training example is pruned to at most 384 and 512 tokens for named entity recognition (NER) and relation extraction (RE). We use a batch size of 5 for NER, and 32 for RE. We also use the hierarchical learning rate in the training process so that the pre-trained parameters and the newly added parameters converge at different optimization processes. For fine-tuning, we train the models for 20 epochs using a learning rate of 2×10^{-5} for pre-trained weights and 3×10^{-5} for others. The learning parameters were selected based on the best performance on the dev set. For NER, we ensemble 5 models from 5-fold cross-validation and 2 models using the normal training-validation approach. For RE, we ensemble 3 models that used all the construction data in training.

4.3 Main Results

Table 2 compares the results of the two tasks of the pre-trained model in trigger words NER and thematic roles identification. We report the impact of using different pre-training models on the

[4] https://github.com/google-research/bert

[5] https://github.com/dmis-lab/biobert

[6] https://github.com/ncbi-nlp/NCBI_BERT

Task	Model	P	R	F1
Trigger Words Recognition	BiLSTM+CRF	0.478	0.408	0.440
	BERT_{base}	0.497	0.448	0.471
	NCBI BERT	0.553	0.453	0.498
	BioBERT	0.511	0.529	**0.519**
Thematic Roles Identification	BERT_{base}	0.758	0.890	0.818
	NCBI BERT	0.778	0.879	0.826
	BioBERT	0.807	0.891	**0.847**

Table 2: Model comparision in development set with different pre-trained models

development set results. We found that even pre-trained models in the general field are superior to the classic BiLSTM+CRF tagging method (Lample et al., 2016). From the last three lines of each task, we can see that different pre-trained models have different results under the same experimental settings. It proves the effectiveness of pre-training tasks in specific domain. During the fine-tuning process of task 1, we found that the joint extraction of entities with other datasets improved our final results.

Label	P	R	F1
CPA	0.39	0.27	0.32
Disease	0.57	0.57	0.57
Enzyme	0.75	0.16	0.26
Gene	0.71	0.64	0.68
Interaction	0.50	0.29	0.36
MPA	0.46	0.47	0.47
NegReg	0.71	0.62	0.66
Pathway	0.83	0.36	0.50
PosReg	0.64	0.61	0.63
Protein	0.32	0.17	0.22
Reg	0.75	0.50	0.60
Var	0.64	0.63	0.64
ALL	0.63	0.56	0.60

Table 3: Precision (P), Recall (R) and F1 scores in test set of Task 1.

The results for task 1 is summarized in Table 3. The difference in the performance in the different labels is partly sourced by the imbalance distribution of trigger labels in the corpus. Our method ends up first place on the leaderboard and substantially improving upon previous state-of-the-art methods. The results for task 2 is summarized in Table 4. Our method ends up second place on the leaderboard. Our method has a large discrepancy between the development set performance and test set performance. It may be the test set is quite different from our constructed data set. This is also related to how we use recognized entities, sentence- or document-level combinations.

Label	P	R	F1
CauseOf	0.60	0.26	0.36
ThemeOf	0.63	0.11	0.19
ALL	0.61	0.16	0.25

Table 4: Precision (P), Recall (R) and F1 scores in test set of Task 2.

4.4 Ablation Study

As shown in Table 5, we found that adding a layer of BiLSTM behind the BERT encoder did not improve the performance of the model, resulting in a 0.04 loss of F_1. For NER tasks, external features are effective for the model's performance. So we verified the efficacy of word shape and POS tags on task 1, and we found that adding this information can increase the F_1 value of our model by more than 0.01.

Model	P	R	F1
BioBERT	0.511	0.529	0.519
+ BiLSTM	0.502	0.448	0.473
- Word shape	0.539	0.453	0.492
- POS tags	0.518	0.482	0.499

Table 5: Ablation study of Task 1 in development set.

5 Conclusion

In this paper, we have explored the value of integrating pre-trained biomedical language representation models into a pipe of information extraction methods for collection of mutation-disease knowledge from PubMed. In particular, we investigate the use of three pre-trained models, BERT_{base}, NCBI BERT and BioBERT, for fine-tuning on the new task and reducing the risk of overfitting. By considering the relationship between different data sets, we achieve better results. Experimental results on a benchmark annotation of genes with active mutation-centric function changes corpus show that pre-trained representations help improve baseline to attain state-of-the-art performance. In future work, we would like to train the entity recognition and relation extraction tasks simultaneously, reducing the cascading error caused by the pipeline model in biomedical information extraction.

Acknowledgment: This work was supported by Natural Science Foundation of China (Grant No.61872113), and the joint project with Baidu Inc.

References

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Mina Gachloo, Yuxing Wang, and Jingbo Xia. 2019. A review of drug knowledge discovery using bionlp and tensor or matrix decomposition. *Genomics & Informatics*, 17(2).

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Zengjian Liu, Yangxin Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Haodi Li, Jingfeng Wang, Qiwen Deng, and Suisong Zhu. 2015. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of biomedical informatics*, 58:S47–S52.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019.*, pages 58–65.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2):S2.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018a. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yuxing Wang, Xinzhi Yao, Kaiyin Zhou, Xuan Qin, Jin-Dong Kim, Kevin Bretonnel Cohen, and Jingbo Xia. 2018b. Guideline design of an active gene annotation corpus for the purpose of drug repurposing. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.