Joint Entity and Relation Extraction from Clinical Records Using Pre-trained Language Model

Xintao FANG†, Yuting SONG††, and Akira MAEDA††

† Graduate School of Information Science and Engineering,
Ritsumeikan University 1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577 JAPAN
†† College of Information Science and Engineering,
Ritsumeikan University 1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577 JAPAN
E-mail: †gr0475vx@ed.ritsumei.ac.jp

Abstract For entity and relation extraction tasks in the general domain, joint models have achieved state-of-the-art performance. Despite their success in the general domain, joint models have not been applied to extract clinical entities and relations. In this paper, we aim to employ a joint model to extract entities and their relations in clinical records. Inspired by the success of pre-trained language models (e.g., BERT) in many NLP tasks, we use the BERT models pre-trained on a biomedical dataset to integrate a joint model for clinical entity and relation extraction. We conducted experiments to evaluate our method on the clinical dataset of the 2010 i2b2 challenge. The results demonstrated the effectiveness of the BERT-based joint model.

Key words Entity recognition, relation extraction, joint model, clinical records, pre-trained language model

1 Introduction

With the development of hospital information systems, medical institutes have begun to use electronic medical records (EMR) to track patient's conditions (e.g., diagnosis and treatment information). The wide adoption of EMR has produced large amounts of digital text concerning patients. Correspondingly, the importance of natural language processing (NLP) techniques in medical fields has radically increased, since they can process clinical texts and obtain knowledge that may assist precise and timely treatments.

This paper focuses on extracting entities and their relations from the narrative texts of clinical records. The goal of the entity and relation extraction is to discover relational structures of entity mentions from unstructured texts. It is a crucial step in building a knowledge graph and also be helpful to question answering.

The methods of entity and relation extraction can be classified into two major categories: (1) pipeline method [1], [17] and (2) joint method [13], [16]. The pipeline method first extracts the entities, and then recognizes their relations. This separated framework makes the task easy to deal with. However, it ignores the relevance between the two sub-tasks since each sub-task has an independent model. The results of entity recognition may affect the performance of relation classification and lead to erroneous delivery. On the other hand, recent studies propose to use joint models for entity and relation extraction [13], [16]. It helps to relieve the aforementioned erroneous delivery issues and has achieved state-of-the-art performance on general domains (e.g., news, Wikipedia articles).

Despite the success of the joint model in the general domain, joint models have not been applied to extract clinical entities and relations. In this paper, we employ a joint model to extract entities and their relations in clinical records. Inspired by the success of pre-trained language models (e.g., BERT) in many NLP tasks [6], we use the BERT model pre-trained on a biomedical dataset to integrate a joint model for clinical entity and relation extraction. We evaluate the the effectiveness of the proposed method on the clinical dataset of the 2010 i2b2 challenge. The results show that our BERT-based joint model outperforms previous work and achieve higher F1 scores.

The remainder of this paper is organized as follows. Section 2 introduces the related work of named entity recognition and relation extraction. Section 3 explains the proposed method. Section 4 presents the experimental settings and results analysis. Section 5 concludes this paper and outlines future work.

2 Related work

In this section, we first present some previous work on entity and relation extraction in the NLP field. Since our work focuses on extracting entities and their relations from clinical texts, we also present related work on clinical entity recognition and relation extraction. Entity and relation extraction is a fundamental task in NLP. There are two main frameworks to deal with this task: the pipeline method [2], [7], [8] and the joint learning method [3], [15], [16]. The pipeline method treats this task as a pipeline of two sub-tasks, i.e., named entity recognition (NER) and relation classification.

2.1 Pipeline method for entity and relation extraction

The NER is the first sub-task in the pipeline method of entity and relation extraction. Early NER was mainly based on statistical machine learning methods, such as Hidden Markov Network [22], Maximum Entropy Model [24], Support Vector Machine [23], etc. In recent years, with the advance of deep learning, neural network models such as CNN+CRF [5] and RNN+CRF [14], [27] have also been proposed for NER. These neural network models no longer rely on manually constructed features. They automatically learn semantic features of texts and achieve state-of-the-art results on multiple benchmark datasets. In more recent years, the emergence of BERT [19] has greatly improved the NER accuracy.

The relation extraction is the second sub-task. Early work on relation extraction mainly relies on manual feature engineering. Similar to NER, with the development of deep learning techniques, some methods based on CNN [21], [26] and RNN [12], [14] have been proposed for relation extraction. These neural network models can better deal with sentence-level feature extraction. With the emergence of BERT model, the BERT-based method have achieved state-of-the-art results [19].

2.2 Joint of entity and relation extraction

The joint learning method is expected to simultaneously extract entities and their relations from texts. Early joint models require manually designed features [9] or the existing NLP tools (e.g., POS taggers) [9], [20]. For example, [15] proposed a method of fusing sequence information and dependency tree information and adding it to the neural network. In addition, [10] introduced a pointer decoding method based on Bi-LSTM. [28] proposed a new annotation framework to combine named entity recognition and relation extraction tasks. However, these methods can not solve the problem of one entity corresponding to multiple relations. [4] proposed a joint neural network model to overcome the aforementioned issues and to automatically perform end-to-end relation extraction without the need for any manual feature engineering or the use of additional NLP components. This method regarded relation extraction as a multi-head selection problem. In other words, any entity may have one or more relations with other entities.

2.3 Clinical entity recognition and relation extraction

In the field of clinical NLP, there is less related work on clinical entities and relations extraction using joint models. Most previous studies regard this task as two independent tasks. Some methods only focus on one of these two tasks, i.e., clinical entity recognition or relation classification given clinical entities. For instance, the i2b2 challenge provided the clinical dataset and proposed two tasks of clinical entity recognition and relation classification separately [25]. Thus, most research using this dataset only focused on a single task of NER or relation classification. Recently, [19] proposed a benchmark to evaluate various NLP tasks in the biomedical field. This work provided four BERT models that were trained using biomedical corpus, including PubMed and MIMIC-III. Inspired by the best performance of biomedical BERT models on various NLP tasks in the biomedical field, we integrate these biomedical BERT models into a joint model for clinical entity and relation extraction.

3 Proposed method

In this section, we introduce our BERT-based joint model for clinical entity and relation extraction. We follow the previous work [4] which proposed a joint neural model that formulates the relation extraction as a multi-head selection problem. The core of the multi-head selection method is that any particular entity may be involved in multiple relations with other entities. By using the multi-head selection method, we can deal with the problem of overlapping relations. Since a clinical entity may be involved in multiple relations with other entities (e.g., a treatment entity may cause multiple medical problems), we utilize the multi-head joint model to extract clinical entities and their relations. Besides, inspired by the good performance of BERT model on NLP tasks, we propose to leverage pre-trained biomedical BERT model as the encoder and fine-tune it with multi-head joint model for clinical entity and relation extraction.

Our model consists of a pre-trained BERT as an encoding layer, a CRF layer for clinical entity recognition, and a sigmoid layer for relation extraction, which is shown in Figure 1.

3.1 BERT-based encoding layer

BERT is a multi-layer bidirectional transformer based language representation model [6], which is pre-trained on a large unlabeled corpus and can be fine-turned on various NLP tasks, such as word segmentation, and sentiment analysis. Since our work focuses on entity and relation extraction in the clinical domain, we use a BERT model pre-trained on biomedical field databases PubMed and MIMIC-III [19] as the encoding layer of our model to extract features. As

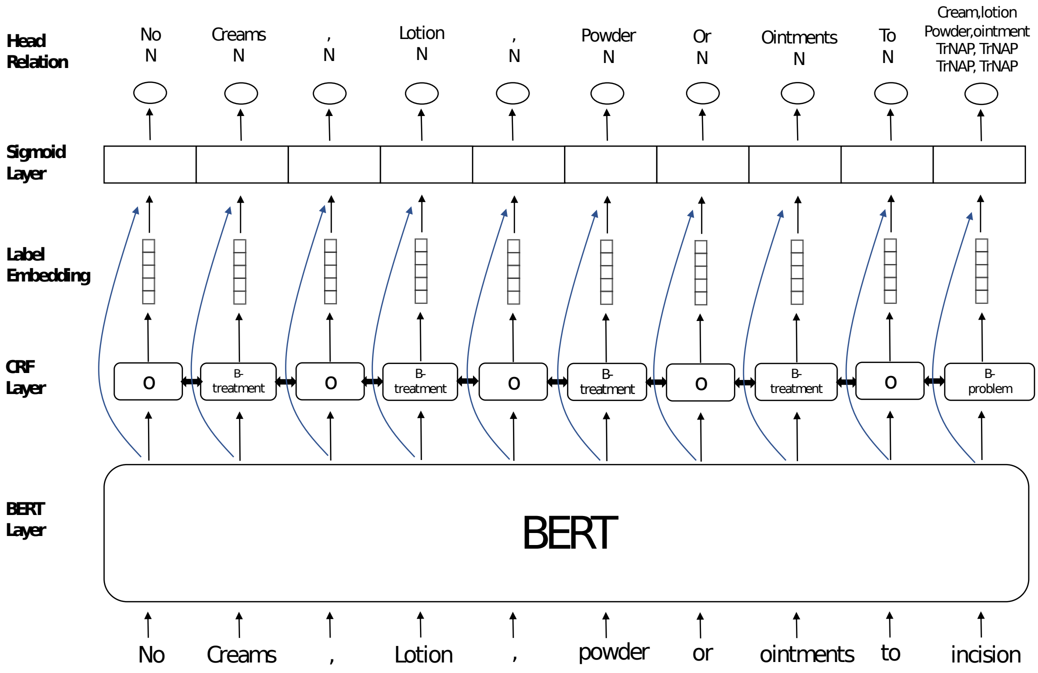

Figure 1 The BERT-based multi-head joint model for clinical entity and relation extraction

a comparison, we also use the BERT model proposed by Google as the encoding layer of our model, which is pre-trained on BookCorpus and English Wikipedia [6].

3. 2 CRF for clinical entity recognition

A CRF layer is added to the BERT model for NER. The word embeddings obtained from BERT model are used as the input of the CRF layer. As the output of CRF, we predict the probability of the entity category label to which the current word may belong. When in training mode, we use the label of the actual entity as the input for the next step. When in predicting mode, we use the predicted label as the input for the next step.

3. 3 Sigmoid layer for relation extraction

In the relation extraction part of the joint model, the relation extraction is regarded as a multi-head selection problem. Each word may have multiple relations with other words. The input to the sigmoid layer here is the combination of the word embeddings of the output of BERT model and the vectors of labels that are generated by the CRF layer. The output is the predicted tuple $< y, c >$, where y represents the head vector, and c is the vector of the corresponding relations for each token.

4 Experiments

4. 1 Dataset

We used the clinical dataset provided by i2b2/VA 2010 challenge [1]. The original dataset consists of 871 discharge summaries including 394 training data and 477 test data. However, part of the original dataset is not available to the public. We can only download a subset of the original dataset from the website of the i2b2/VA 2010 challenge. The avaiable dataset consists of 426 discharge summaries including 170 training data and 256 test data. This dataset contains three types of clinical entities (i.e., medical problem, test, and treatment) and eight types of relations, which are shown in Table 1 and Table 2.

Table 1 Three types of clinical entities

Catagory	Medical problem Test Treatment
Example 1	**Hypertension** (problem) was controlled on **hydrochlorothiazide** (treatment).
Example 2	**An echocardiogram** (test) revealed **a pericardial effusion** (problem).

Table 2 Eight types of relations between clinical entities

Catagory	Description
TrIP	Treatment improves medical problems
TrWP	Treatment worsens medical problems
TrCP	Treatment causes medical problems
TrAP	Treatment is applied to medical problems
TrNAP	Treatment is not applied to medical problems
TeRP	Tests reveal medical problems
TeCP	In order to prove medical problems, need to be checked
PIP	The relation between medical problems

1 : https://www.i2b2.org/NLP/DataSets/Main.php

In our experiments, we mixed the 170 training data and 256 test data downloaded from the website of the i2b2/VA 2010 challenge. Then, we randomly split it into training, validation, and test datasets with a ratio of 3:1:1. We have made statistic information of entities and relationships in the experimental dataset that are shown in Table 3 and Table 4.

4.2 Implementation details

We developed our model by using the PyTorch [18]. In the training process, we performed using the Adam optimizer [11] with a learning rate of 10^{-3}. We experimented with a tanh activation function. As a comparison with our method that uses BERT, we conducted experiments using Bi-LSTM as the encoding layer. When using the LSTM-based model for training, we used the 200-dimensional word embeddings for this dataset. We set the size of the LSTM as 300. And both of train batch and evaluation batch were set as 32. When using the BERT-based model for training, we keep the hyperparameter correspond as before.

We conducted experiments using three different BERT models: 1) BERT(Books+Wiki) is the 'BERT-base-uncased' model, which was released by Google [6]. It was pre-trained on BookCorpus and Wikipedia articles. 2) BERT(PubMed) was released by [19]. This BERT model was initialized with pre-trained BERT provided by Google. Then, it continued to be pre-trained using the PubMed abstracts, which is a biomedical domain corpus. 3) BERT (PubMed+MIMIC) was also released by [19]. Similar to the second BERT model, this model was initialized with pre-trained BERT provided by Google. Then, it continued to be pre-trained using the PubMed abstracts and MIMIC-III clinical notes. We obtain the best hyperparameters after 45 to 130 epochs. We select the best epoch according to the results in the validation set.

4.3 Evaluation metrics

We defined that an entity is considered correct if only the boundaries of the entity are correct (entity type is not considered), a relation is correct when the type of the relation and the argument entities are both correct. The experimental results are evaluated by using precision, recall, and F1 score, which are shown as follows:

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{1}$$

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{2}$$

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

4.4 Results

Table 5 shows the precision, recall and F1 scores for named entity recognition and relation extraction tasks using the Bi-LSTM multi-head joint model and the BERT-based multi-head joint model. Both in the task of named entity recognition and relation extraction, the BERT-based models are better than the Bi-LSTM models. In addition, by comparing the results of using three different BERT models, we can see that the BERT model that is pre-trained on PubMed and MIMIC-III performs better than the BERT model pre-trained only on PubMed. And both of these two biomedical models perform better than the BERT model which is pre-trained on BooksCorpus and English Wikipedia [6].

5 Conclusion

In this paper, we proposed a BERT-based joint model for clinical entity and relation extraction. We compared the results of using Bi-LSTM and BERT as the encoding layer of multi-head joint model. In addition, we conducted experiments by using the BERT models in the general domain and the biomedical domain. The results showed that the biomedical BERT-based joint model achieved the highest F1 score. In the future, we will evaluate our proposed method on other clinical datasets, and further improve the accuracy of our method based on the BERT model.

References

[1] N. Bach and S. Badaskar. A review of relation extraction. *Literature review for Language and Statistics II*, 2:1–15, 2007.

[2] G. Bekoulis, J. Deleu, T. Demeester, and C. Develder. Reconstructing the house from the ad: Structured prediction on real estate classifieds. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 274–279, 2017.

[3] G. Bekoulis, J. Deleu, T. Demeester, and C. Develder. An attentive neural architecture for joint segmentation and parsing and its application to real estate ads. *Expert Systems with Applications*, 102:100–112, 2018.

[4] G. Bekoulis, J. Deleu, T. Demeester, and C. Develder. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45, 2018.

[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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, 2019.

[7] K. Fundel, R. Küffner, and R. Zimmer. Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

[8] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo. Development of a benchmark corpus to support the automatic extraction of

Table 3 Number of reports, sentences, and clinical entities in training, validation, and test datasets

Dataset	Reports	Sentences	Clinical entities		
			Problem	Treatment	Test
Training	256	3024	11555	8097	8043
Validation	85	935	3858	3062	2962
Test	85	1105	4252	3029	2828

Table 4 Number of each relation type in training, validation, and test datasets

Dataset	Eight types of relations between clinical entities								Total
	TrIP	TrWP	TrCP	TrAP	TrNAP	TeRP	TeCP	PIP	
Training	151	108	338	1726	111	2038	335	1428	6235
Validation	27	13	83	448	35	562	108	396	1672
Test	24	11	101	437	27	431	58	359	1448

Table 5 Precision (P), recall (R), and F1 scores of NER and relation extraction

Model	NER			Relation extraction		
	P	R	F1	P	R	F1
Bi-LSTM (Wiki+Gigaword)	0.8398	0.8287	0.8342	0.4326	0.2128	0.2852
Bi-LSTM (PubMed+PMC)	0.8428	0.8236	0.8331	0.4322	0.2235	0.2946
BERT (Books+Wiki)	0.8991	0.9532	0.9253	0.3717	0.2973	0.3304
BERT (PubMed)	0.9290	0.9295	0.9292	0.3891	0.3731	0.3809
BERT (PubMed+MIMIC)	0.9359	0.9501	0.9430	0.4077	0.4211	0.4143

drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892, 2012.

[9] N. Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 178–181, 2004.

[10] A. Katiyar and C. Cardie. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 917–928, 2017.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, 2016.

[13] Q. Li and H. Ji. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, 2014.

[14] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, 2016.

[15] M. Miwa and M. Bansal. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, 2016.

[16] M. Miwa and Y. Sasaki. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1858–1869, 2014.

[17] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[19] Y. Peng, S. Yan, and Z. Lu. 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*, pages 58–65, 2019.

[20] B. Rink and S. Harabagiu. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 256–259, 2010.

[21] C. N. d. Santos, B. Xiang, and B. Zhou. Classifying relations by ranking with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 626–634, 2015.

[22] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. *Advances in neural information processing systems*, 16:25–32, 2003.

[23] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 104, 2004.

[24] T. J. Ulrych and T. N. Bishop. Maximum entropy spectral analysis and autoregressive decomposition. *Reviews of Geophysics*, 13(1):183–200, 1975.

[25] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[26] K. Xu, Y. Feng, S. Huang, and D. Zhao. Semantic relation classification via convolutional neural networks with sim-

ple negative sampling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 536–540, 2015.

[27] Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1785–1794, 2015.

[28] S. Zheng, Y. Hao, D. Lu, H. Bao, J. Xu, H. Hao, and B. Xu. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59–66, 2017.