CCKS2020 Medical Event Extraction Based on Named Entity Recognition

Xinnan Zhang, Xinyu Zhao, Shen Ge, and Xian Wu

Tencent Medical AI Lab, China
{xinnanzhang, xinyuzhao, shenge, kevinxwu}@tencent.com

Abstract Event extraction of electronic medical records attracts much attention in recent years. The 2020 China conference on knowledge graph and semantic computing (CCKS 2020) leads an open challenge that recognizes three kinds of tumor related attributes from the Chinese electronic medical records (CEMRs). In this challenge, we utilize a pretrained BERT model to catch the semantic infomation of the sentences which is then fed into a BiLSTM-CRF model to do Named Entity Recognition (NER). To achieve better performance, we also use data augmentation and post processing heuristic rules. In the official test set, our approach achieves an F1 score of 0.74579.

Keywords: Named Entity Recognition · BERT · Data Augmentation.

1 Introduction

In the big data era, how to apply Natural Language Processing (NLP) in medical filed has become a crucial issue. As a fundamental task in NLP, NER aims to find named entities mentioned in unstructured texts and classify them into predefined categories. Medical NER can handle massive electronic medical records which are continuously generated by the hospitals and help doctors grasp patient information rapidly, saving their time in reading medical records from different hospitals, different doctors, and in different writing styles.

The 2020 China Conference on Knowledge Graph and Semantic Computing (CCKS 2020) organizes a variety of open challenges and provides a series of high-quality annotated data to promote the development of Chinses NLP applications. The task of event extraction for Chinese electronic medical records (CEMRs) aims to extract tumor primary site (TPS), primary tumor size (PTS) and tumor metastatic site (TMS) from CEMRs. In this paper, we employ a pretrained RoBERTa [7] model which trains Bidirectional Encoder Representations from Transformers (BERT) [3] in a robust way, to convert CEMRs data to high-dimensional representation and then feed it into a bidirectional Long Short-Term Memory model with a conditional random field layer (BiLSTM-CRF) [5] to generate the output predictions.

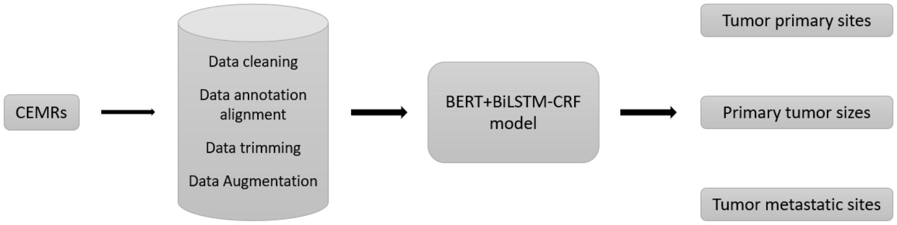


Fig 1. Training data flow

2 Method

The training data flow of our NER method is shown on Fig. 1. Firstly, we perform several pre-processing operations on the original dataset, including data cleaning, data annotation alignment, data trimming, and data augmentation as well. Secondly, all training data is then split equally into 5 parts, since 5-fold cross validation is used in our training. We employ three different parameter settings of BiLSTM models, leading to 15 models generated in total. Finally, we fed the ensemble of 15 models to a post-processing module to produce the final inferenced results. The architecture of our BERT-BiLSTM-CRF model is illustrated in Fig. 2.

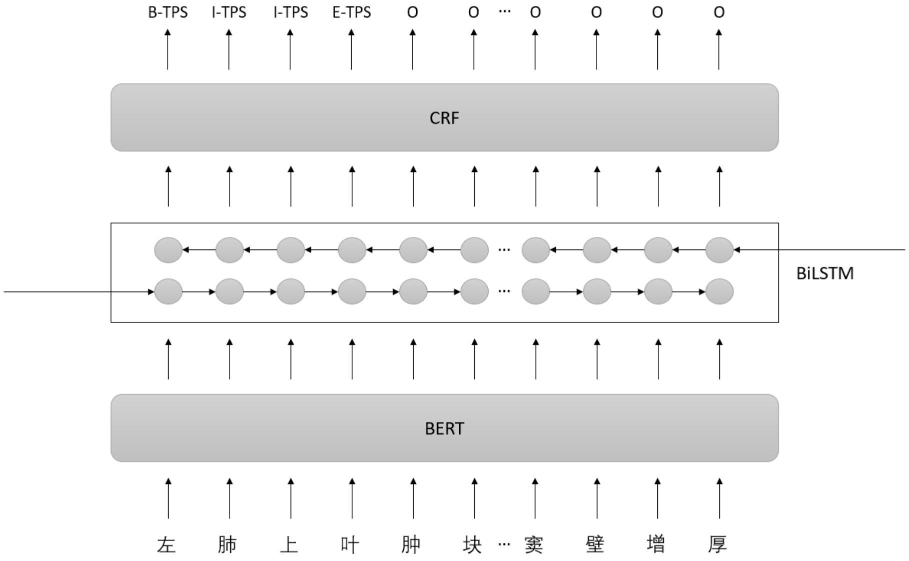
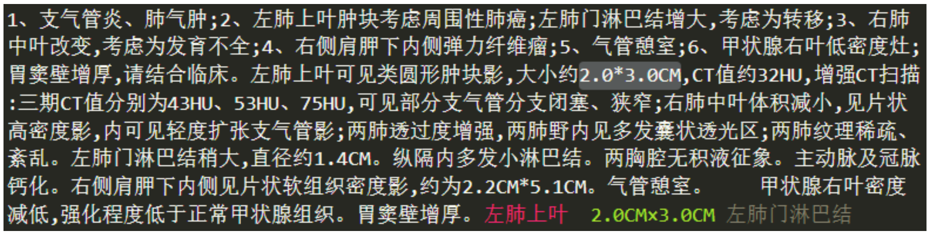

B-TPS I-TPS I-TPS E-TPS O O ··· O O O O
CRF
BiLSTM
BERT
左 肺 上 叶 肿 块 ··· 窦 壁 增 厚

Fig 2. Architecture of BERT-BiLSTM-CRF model

2.1 Pre-processing

Data Cleaning In the original data, we found that there are some meaningless characters such as "\xa0", "\t". In order to reduce the data noise and the length of CEMRs, we drop these characters in pre-processing.

Data Annotation Alignment As shown in Fig. 3, the data annotation only gives the target attribute entities without mentioning the location of them in the original medical records. In order to perform NER tasks, we need to align the ground truth labels back with CEMRs by string match. However, the format of primary tumor size label is different than that in the original text. For example, in Fig. 3, the medical record describes the primary tumor size to be "2.0*3.0cm", while the ground truth label is "2.0CM×3.0CM". As a result, we add special processing to normalize these sizes descriptions, making sure that the data annotations are align well with the original texts.



1、支气管炎、肺气肿;2、左肺上叶肿块考虑周围性肺癌;左肺门淋巴结增大,考虑为转移;3、右肺中叶改变,考虑为发育不全;4、右侧肩胛下内侧弹力纤维瘤;5、气管憩室;6、甲状腺右叶低密度灶;胃窦壁增厚,请结合临床。左肺上叶可见类圆形肿块影,大小约2.0*3.0CM,CT值约32HU,增强CT扫描:三期CT值分别为43HU、53HU、75HU,可见部分支气管分支闭塞、狭窄;右肺中叶体积减小,见片状高密度影,内可见轻度扩张支气管影;两肺透过度增强,两肺野内见多发囊状透光区;两肺纹理稀疏、紊乱。左肺门淋巴结稍大,直径约1.4CM。纵隔内多发小淋巴结。两胸腔无积液征象。主动脉及冠脉钙化。右侧肩胛下内侧见片状软组织密度影,约为2.2CM*5.1CM。气管憩室。 甲状腺右叶密度减低,强化程度低于正常甲状腺组织。胃窦壁增厚。左肺上叶 2.0CM×3.0CM 左肺门淋巴结

Fig 3. Example of Annotated CEMR

Data Trimming Limited by the input sequence length of BERT model, which only accepts upto 512 tokens, we manually drop some parts of texts that we think do not harm the completeness of CEMRs infomation in this NER task. Actually, around 1/3 of the data has more than 512 tokens and thus must be manually trimmed. It is possible that we cannot find a shortened version of CEMRs with complete information, and such CEMRs will be truncated to 512 tokens for the input of the model.

Data Augmentation The original training set only constains around 1000 instances, which is far from enough for the model to understand CEMRs data in different conditions. In order to enhance the robustness of our model, we randomly re-arrange the sentence order in each instance and generate a corresponding new training sample, doubling the whole training set. Meanwhile, we use the trained model to produce predictions on the validation set. With an F1-score of 0.8404, these 400 validation samples are mixed into the original 1000 training set. With the help of sentence re-arrangement policy, we get a total of 2800 training instances after data augmentation.

2.2 BERT

In 2018, Google proposed BERT [3], which is designed to pretrain deep bidirectional representations from unlabeled text, and is a language model based on multi-layer bidirectional Transformers trained by Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks. Opening a new era of NLP, the pretrained BERT model can be fine-tuned with downstream tasks and obtain state-of-the-art results on eleven NLP tasks, e.g., Named Entity Recognition [3]. However, many works [6, 10] pointed out that the NSP task may harm the model performance, and it is more effective to train the BERT model with the whole long sentences directly. Inspired by such observations, Facebook proposed a model called Robustly Optimized BERT Pretraining Approach (RoBERTa) [7] which achieves a better performance than the original BERT model. RoBERTa is trained with bigger batches over more data, with the NSP task removed due to its proven unbeneficial effect to some downstream tasks. Other tricks, such as model training with whole documentation and the introduction of dynamic mask mechanism are also used in RoBERTa training.

2.3 BiLSTM-CRF

BiLSTM-CRF is the most commonly used model to tackle sequence tagging tasks, especially NER. LSTM [4] is a special version of RNN structure, which alleviates gradient vanish problem. BiLSTM adds bidirectional mechanism to traditional LSTM, empowering it with the capability of using both past and future input features efficiently. Meanwhile, as a traditional discriminant model, CRF is typically used to handle sequence tagging tasks by introducing extra features to consider global loss and constraint classification of each sequence token [1, 2, 8]. Here we pre-defined 12 kinds of tag categories based on ”BIOES” strategy, as shown in Table 1.

2.4 Model Ensemble and Post-processing

Model Ensemble We use two strategies for the ensemble of all 15 predicted results of models. **(Strategy A):** For each tumor relevant attribute, if one entity is extracted by more than 8 models which accounts for more than 1/2 of the total number of models, it is then chosen to be included in the output. Otherwise, the entity is discarded in the prediction. **(Strategy B):** As mentioned above, we train the model in a 5-fold cross validation fashion, so any fold has 3 model variants with different hyper-parameters. Next, for each fold, we pick up the best model out of 3 with the highest F1-score. Using such strategy, we obtain 5 models and the voting threshold for entities is 3.

Post-processing After generating two results from strategies A and B, a series of heuristic rules are then applied on these two results for more meaningful results. We list some major ones below:

Table 1. Pre-defined categories of the model

Head tags	Tail tags	Combined tags
Begin	tumor primary site	B-TPS
	primary tumor size	B-PTS
	tumor metastatic site	B-TMS
Intermediate	tumor primary site	I-TPS
	primary tumor size	I-PTS
	tumor metastatic site	I-TMS
End	tumor primary site	E-TPS
	primary tumor size	E-PTS
	tumor metastatic site	E-TMS
Single	tumor primary site	S-TPS
	tumor metastatic site	S-TMS
Other		O

- For each predicted TPS “X”, we try to match it back to the original instance to find patterns like “X癌”, “XCA” or “X术后”. If none of above patterns are found, we consider to discard the TPS prediction of “X”.
- For PTSs, inspired by some Computed Tomography (CT) knowledge, word patterns like “肿块影”, “结节影” and “密度影” etc helps us to keep the generated PTSs.
- TMSs also need some constraints to produce meaningful predictions. For example, patterns like “转移”, “侵犯”, “多发.*(破坏|病灶)”, “不除外” or “待除外” help us to confirm the correctness of TMS predictions, while patterns like “除外” help us to discard the TMS predictions.
- For the two generated results from strategies A and B, we mainly use the TPS predictions from strategy B, and switch to strategy A when B predicts empty TPSs. We use PTSs exclusively from strategy A and TMSs from strategy B.

3 Experiments and Results

In our experiments, we use pretrained RoBERTa model which named `hfl/chinese-roberta-wwm-ext` and `hfl/chinese-roberta-wwm-ext-large` [9] in huggingface transformers. These models are pretrained by Joint Laboratory of HIT and iFLYTEK Research (HFL), where `hfl/chinese-roberta-wwm-ext` is a BERT-base model containing 12 layers, 768 hidden states and 12 heads with 110M parameters and `hfl/chinese-roberta-wwm-ext-large` is a BERT-large model containing 24 layers, 1024 hidden states and 16 heads with 330M parameters. For the BiLSTM layer, we use 3 different output dimension sizes as 512, 768 and 1024. All models are trained by Adam optimizer with the BERT learning rate 3e-5 and the BiLSTM-CRF learning rate 1e-3. During inference stage, if the total number of tokens is more than 512, then we truncate the CEMR two times, one from the head and the other from the tail, and pass both truncated texts into the model. The final prediction result is an aggregation of these two inference results. Our models are trained with Tesla P40 with a batch size of 8.

3.1 DataSet

CCKS 2020 CEMRs event extraction challenge provides 1000 annotated medical records as the training set, 400 unannotated medical records as the validation set and 300 unannotated medical records as the test set. Applying the aforementioned data augmentation method, we get 2800 annotated CEMRs as the training set.

3.2 Evaluation

Our evaluation method is the same one with the organizer's approach. For all three target attributes, an entity recognized by model and strictly match one of the golden entities is regarded as a True-Positive (TP), otherwise it is regarded as a False-Positive (FP) if it fails to satisfy the aforementioned conditions. And a False-Negative (FN) is a golden entity that has not been recognized. The precision (P), recall (R) and F1-score are computed by

$$P = \frac{TP}{TP + FP}, \tag{1}$$

$$R = \frac{TP}{TP + FN}, \tag{2}$$

$$F1 = \frac{2P \times R}{P + R} \tag{3}$$

3.3 Experimental Results

Table 2 and Table 3 show our experimental results on the offical validation and test dataset, respectively. Models with BERT-base use `hfl/chinese-roberta-wwm-ext` as pretrained BERT model, and BERT-large models use `hfl/chinese-roberta-wwm-ext-large`. It is well known that a large pretrained BERT model may improve the performance, which is also confirmed by our experiments. Meanwhile, our data augmentation stategy works well on this task.

4 Conclusion

In this paper, we present a neural network approach to extract tumor related entities from Chinese electronic medical records. In this approach, two special data augmentation methods is employed based on common BERT-BiLSTM-CRF architecture model. The experimental results show that our strategies could effectively improve the performance of this model. Our best submission achieves an F1-score of 0.7458 on the test data set.

Table 2. Model performance on validation data

Models	With Data augmentation	With post-processing	F1-score
BERT-base+BiLSTM+CRF	False	False	0.7757
	True	False	0.8162
	True	True	0.8278
BERT-large+BiLSTM+CRF	False	False	0.7873
	True	False	0.8236
	True	True	0.8404

Table 3. Model performance on test data

Models	With Data augmentation	With post-processing	F1-score
BERT-base+BiLSTM+CRF	False	False	0.6815
	True	False	0.7184
	True	True	0.7375
BERT-large+BiLSTM+CRF	False	False	0.6875
	True	False	0.7283
	True	True	0.7458

References

1. Chen, A., Peng, F., Shan, R., Sun, G.: Chinese named entity recognition with conditional probabilistic models. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. pp. 173–176 (2006)
2. Chen, Y., Zhou, C., Li, T., Wu, H., Zhao, X., Ye, K., Liao, J.: Named entity recognition from chinese adverse drug event reports with lexical feature based bilstm-crf and tri-training. Journal of biomedical informatics **96**, 103252 (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
5. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
6. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics **8**, 64–77 (2020)
7. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
8. Wallach, H.M.: Conditional random fields: An introduction. Technical Reports (CIS) p. 22 (2004)
9. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. ArXiv pp. arXiv–1910 (2019)

10. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems. pp. 5753–5763 (2019)