Named Entity Recognition and Event Extraction in Chinese Electronic Medical Records

Cheng Ma[1] and Wenkang Huang[2](✉)

[1] Fudan University, Shanghai, China
19210240239@fudan.edu.cn
[2] Ant Group, Hangzhou, China
wenkang.hwk@alibaba-inc.com

Abstract. The China Conference on Knowledge Graph and Semantic Computing (CCKS) 2021 Evaluation Task 4 presented clinical named entity recognition and event extraction for the Chinese electronic medical records. Two annotated data sets for the two subtasks were provided for participators. Our model on the test dataset achieves the strict F1-Measure of 0.7684 which ranked the first place.

Keywords: CCKS · Named entity recognition · Event extraction

1 Introduction

With the advent of the information age, electronic medical records have become more and more popular. Electronic medical records contain a large amount of medical semantic knowledge. It is particularly important to use effective natural language processing technology to extract the knowledge contained in electronic medical records. CCKS 2021 sets up five evaluation themes and a total of fourteen evaluation tasks, task 4 focuses on named entity recognition (NER) and event extraction (EE) in the Chinese electronic medical records (EMR).

NER and EE are commonly used techniques to extract structured knowledge from unstructured text. The most popular NER method is sequence labeling, which can be based on long short-term memory (LSTM) [1,2] or bidirectional encoder representation from transformers (BERT) [3]. Sometimes, medical event extraction can be transformed into a medical entity recognition task.

In this paper, we use the medical named entity recognition data set and medical event extraction data set provided by CCKS 2021 task 4. By using the pre-trained BERT model and fusing multiple models based on the expansion of the BERT model, we obtained a strict F1 score of 0.7684 based on the data set provided by CCKS 2021 Task 4.

B. Qin et al. (Eds.): CCKS 2021, CCIS 1553, pp. 133–138, 2022.
https://doi.org/10.1007/978-981-19-0713-5_15

2 Task Formalism

2.1 Clinical Named Entity Recognition (CNER)

This task is a Chinese medical record medical entity recognition task, that is, for a given set of plain text documents of electronic medical records, identify and extract entity mentions related to medical clinics, and classify them into pre-defined categories, such as disease, drug, operation, etc.

Formalized Definition. We define this task formally.
INPUT:

1) A document collection from EMR: $D = \{d_1, d_2, ..., d_N\}$, where $d_i = (w_{i1}, ..., w_{in})$
2) A set of pre-defined categories: $C = \{c_1, ..., c_m\}$.

OUTPUT:
Collections of entity mention-category pairs: $\{(m_1, c_{m1}), ..., (m_i, c_{mi}), (m_p, c_{mp})\}$.

The $m_i = (d_i, b_i, e_i)$ represent the entity mention in document d_i, where b_i and e_i is the start and end position of m_i, respectively. $c_{mi} \in C$ represents the category of m_i. The overlap between mentions is not allowed, which is $e_i < b_{i+1}$.

Pre-defined Categories. There are 6 categories that are defined as follows.

1) Disease and diagnose (Dis)
2) Imaging examination (ImgExam)
3) Laboratory examination (LabExam)
4) Operation
5) Drug
6) Anatomy

2.2 Clinical Event Extraction (CEE)

This task is the task of extracting medical events from Chinese medical records, that is, given the main entity of the electronic medical record text data of the tumor, define several attributes of the tumor event, such as tumor size, tumor primary site, etc., identify and extract events and attributes, and perform text structure change.

Formalized Definition. This task is formally defined as follows.
INPUT:

1) Event entity.
2) A document collection from EMR: $D = d_1, \ldots, d_N$, where $d_i = (w_{i1}, \ldots, w_{in})$
3) A set of pre-defined attributes: $P = p_1, p_2, \ldots, p_m$

OUTPUT:
Collections of attribute entities: $\{[d_i, (p_j, (s_1, s_2, \ldots, s_k))]\}$, and $1 \leq i \leq N, 1 \leq j \leq m$. The s_k is the entity of attribute p_j from document d_i. There could be 0 or more than one entity for each attribute.

Pre-defined Categories. The 3 pre-defined attributes are:

1) Tumor Primary Site
2) Tumor Size
3) Tumor Metastatic Site

3 Methods

3.1 BERT Encoder

Bidirectional encoder representation from transformers (BERT) is a pre-trained language mode based on a large-scale universal corpus, which has two self-supervision task, next sentence prediction and masked language model. BERT learns a large amount of general knowledge through the task of self-supervision, and it only needs simple fine-tuning to transfer the knowledge to downstream tasks, so as to get better results.

3.2 Conditional Random Fields (CRF)

A conditional random field (CRF) is a type of discriminative, undirected probabilis tic graphical model, which has been widely used for sequence labeling problems. For a given character sequence $z = z_1, ..., z_n$, where z_n is the input vector composed of the char and features of ith character, and a given label sequence $y = y_1, ..., y_n$ for z. $\gamma(z)$ represent the all of possible labels for z. The CRF model define the formula of the probability of character sequence y with given label sequence y is:

$$p(y|z;\theta) = \frac{\sum_{t=1}^{n} exp(S(y^{(t)}, z^{(t)}, \theta))}{\sum_{t=1}^{n} \sum_{j \in \gamma(z)} exp(S(y_j, z^{(t)}, \theta))} \tag{1}$$

Where $S(y^{(t)}, z^{(t)}, \theta)$ are potential function, and θ is the parameters of CRF. In our work, we use the character as a unit for sequence labeling model rather than use the word. Log likelihood function was used to get the loss of the CRF layer. Finally, the viterbi algorithm was used to decode.

3.3 Transform of Event Extraction

In the event extraction task, we need to identify the three positions of the tumor primary site, tumor metastasis site, and tumor size in a piece of text. We convert these three positions into entity type tags, and then turn this task into entity Identify the task.

Because the original data only gives the characters of the tumor primary site, tumor metastasis site, and tumor size, but not the position in the text. When transforming the event extraction task into an entity recognition task, we need to mark the location of the tumor primary site, tumor metastasis site, and tumor size in the text.

We noticed that the word “转移” (transfer) appears in many texts, which is very important for the identification of the metastasis site. Therefore, for the confirmation of the physical location of the metastasis site, the method we choose is to select the entity location closest to the word “转移”.

4 Evalution Metrics

4.1 Clinical Named Entity Recognition

This task uses Precision, Recall and F1-Measure as evaluation metrics. The extracted entities set is denoted as $S = s_1, s_2, ..., s_m$ and the gold entities set is denoted as $G = g_1, g_2, ..., g_n$. The set element is an entity mention, expressed as a four-tuple $<d, pos_b, pos_e, c>$, where d represents a document, pos_b and pos_e respectively correspond to the start and end of the entity mention in document d, c indicates that the entity mentions the predefined category to which it belongs. There are two evaluation metrics, the strict metric and relaxed metric.

Strict Metric. For the strict metric, $s_i \in S$ is equal to $g_j \in G$, which means they are exactly the same:

1) $s_i[d] = g_j[d]$
2) $s_i[pos_b] = g_j[pos_b]$
3) $s_i[pos_e] = g_j[pos_e]$
4) $s_i[c] = g_j[c]$.

The strict Precision, Recall and F1 can be calculated as follows:

$$P_s = \frac{|S \cap_s G|}{|S|} \tag{2}$$

$$R_s = \frac{|S \cap_s G|}{|G|} \tag{3}$$

$$F1_s = \frac{2P_sR_s}{P_s + R_s} \tag{4}$$

Relaxed Metric. The relaxed metric does not require that $s_i \in S$and $g_j \in G$are exactly the same, and they only need to meet the following requirements:

1) $s_i[d] = g_j[d]$
2) $max(s_i[pos_b], g_j[pos_b]) \leq min(s_i[pos_e], g_j[pos_e])$
3) $s_i[c] = g_j[c]$.

The relaxed Precision, Recall and F1 can be calculated as follows:

$$P_r = \frac{|S \cap_r G|}{|S|} \tag{5}$$

$$R_r = \frac{|S \cap_r G|}{|G|} \tag{6}$$

$$F1_r = \frac{2P_r R_r}{P_r + R_r} \tag{7}$$

4.2 Clinical Event Extraction

There could be more than one attribute entity for an event attribute. The Precision, Recall and F1 are calculated based on the attribute entity rather then attribute.

5 Experiments

5.1 Datasets

The CCKS 2021 Task 4 provided annotated data set for Clinical Named Entity Recognition and Clinical Event Extraction. The statistics of CNER and CEE data set are shown in Table 1 and Table 2 respectively.

Table 1. The statistics of clinical named entity recognition data set.

	Docs	Dis	ImgExam	LabExam	Operation	Drug	Anatonmy	Total
Train	1050	4345	1002	1297	923	1935	8811	18313
Valid	450	1834	481	575	406	894	3861	8051
Unlabeled	1000	–	–	–	–	–	–	–

Table 2. The statistics of clinical event extraction data set.

	Docs	TumorPrimarySite	TumorSize	TumorMetastaticSite	Total
Train	1000	1075	1025	1878	3978
Valid	400	269	260	638	1167
Unlabeled	1000	–	–	–	–

5.2 Settings

By adjusting the hyper-parameters of the training model through the validation datasets, the best hyper-parameters in CRF model was obtained and described below. The model are trained by Adam optimization algorithm [4].

1) Learning-rate of CRF layer is 5e-5.
2) Learning-rete of BERT layer is 2e-5.

We also used adversarial training (such as FGM [5]) and other tricks to improve the model results.

5.3 Results

The results of our model on the validation set are shown in the Table 3 and Table 4 respectively. Compared strict and relaxed results in Table 3, we find that the operations don't have a high strict F-measure but have a high relaxed F-measure. It means that the right position of entities has been found without the right boundary. It can be seen from Table 4 that the model does not recognize the primary site of the tumor very well, possibly because the word "转移" does not provide enough information for the recognition of the primary site.

Table 3. The results of clinical named entity recognition on Valid data sets.

	Dis	ImgExam	LabExam	Operation	Drug	Anatonmy	All
Strict	0.870	0.893	0.884	0.876	0.942	0.868	0.880
Relaxed	0.943	0.935	0.935	0.950	0.970	0.940	0.944

Table 4. The statistics of clinical event extraction on Valid data set.

	TumorPrimarySite	TumorSize	TumorMetastaticSite	All
Strict	0.736	0.912	0.814	0.800

6 Conclusion

This paper presents a detailed introduction of CCKS 2021 Task4 for clinical named entity recognition and clinical event extraction for Chinese EMRs. Our team won the first place in the Task 4 evaluation with a combined score of 0.7684. We will focus on the more effective extraction of entities' boundary in the future.

References

1. Lample, G., et al.: 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, pp. 260–270 (2016)
2. Ma, X.Z., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1064–1074 (2016)
3. Devlin, J., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Miyato, T., Dai, A.M., Goodfellow, I.: Adversarial Training Methods for Semi-Supervised Text Classification. arXiv preprint arXiv:1605.07725 (2017)