NER-PS-MS: Medical Attribute Extraction based on Medical Named Entity Recognition

Yawen Song, Ling Luo, Nan Li, Zeyuan Ding, Zhihao Yang* and Hongfei Lin

College of Computer Science and Technology, Dalian University of Technology, Dalian, China 116024
* Corresponding Author: yangzh@dlut.edu.cn

Abstract. Focusing on the inter-hospital transfer issue caused by the differences of clinical expressions among different health care facilities or among different doctors in the same facility, the 2019 China conference on knowledge graph and semantic computing (CCKS) organized the medical attribute extraction (inter-hospital transfer) task for the first time. In this task, we propose an approach combining a BiLSTM-CRF model for Medical Named Entity Recognition(MNER) and a downstream PS-MS model for Medical Attribute Extraction(MAE), which we call NER-PS-MS. On the official test set, our best submission achieves an F-score of 70.69% considering all three attributes.

Keywords. Attribute Extraction, Chinese Medical Text, Named Entity Recognition, Downstream

1 Introduction

Observing the increasing clinical data from multiple health care facilities carefully, we can find that the habits of writing medical records vary from one facility to another, even among doctors in the same department in the same facility not only in words, but also in language patterns [1]. It means that the NLP model we trained in one medical institution may have very different performance when tested on the data of another medical institution. If the data from every medical institution were to be re-labeled, we would need expensive labor and time investment. Therefore, it becomes particularly important to transfer the model of a field trained with a lot of annotated data to a new field with a little annotated data.

Focusing on this issue, the 2019 China conference on knowledge graph and semantic computing (CCKS-2019) organized a MAE task to identify and extract the answer entity of the medically relevant target fields predefined from the given clinical plain text documents.

The corpus provided by organizers do not contain accurate location annotations of attribute answers, which means we are not able to resolve it as a sequence labeling task directly. In this paper, we propose an NER-PS-MS model based on named entity recognition. This model performs NER firstly to obtain the boundary information of anatomic site entities and then uses the downstream PS-MS modules to extract attrib-

ute results with NER output as input. That is, we transfer the model for NER task to be the upstream model of MAE task.

2 Method

The NER-PS-MS method is designed to handle the problem of medical attribute extraction with limited annotations.

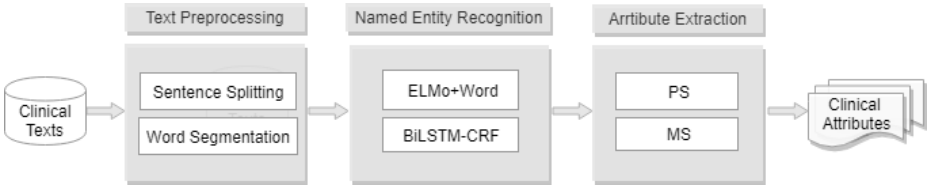

Fig. 1. The processing flow of our NER-PS-MS method

According to Fig.1, we first split the original texts into sentences and send them into upstream BiLSTM-CRF model for named entity recognition. Secondly, we feed the BIOES labeled sentences output by the upstream model into the Primary Tumor Site(PS) Module and the Metastasis Site(MS) Module and obtain the extraction results of three attributes. Note that the size of the primary tumor site is extracted from PS module.

2.1 Problem Formulation

We formalize attribute extraction as the following definition. Let *x* be a plain medical text and let $(x_1, x_2, \dots, x_n)$ be the character sequence of x. Given an attribute α, medical attribute extraction is the process of discovering a function E such that $\mathrm{E}(x) = E(x_1, x_2, \dots, x_n) = (x_i, x_{i+1}, \dots, x_k)$ for $1 \le \mathrm{i} \le \mathrm{k} \le \mathrm{n}$ where $\mathrm{a} = (x_i, x_{i+1}, \dots, x_k)$ is a particular value of α. In this task, attributes predefined are the size of the primary tumor site(SP), the primary tumor site(PS) and the metastasis site(MS). Fig.2 shows us an instance of MAE.

左肺癌化疗后，对比 2015-05-05 片： 右上肺门区肿物较前稍大，右上肺内转移瘤较前无明显变化。 两锁下、两下气管旁、血管前间隙、两上气管旁、主-肺动脉窗、主动脉旁、隆突上及两肺门多发转移淋巴结，较前无明显变化 右侧胸膜多发转移瘤，部分较前稍大；右侧胸腔积液较前稍增多。 肝内数个囊肿。 肝 S5 肝内胆管结石。左肺癌化疗后，对比 2015-05-05 片： 左上肺肺门区见一团块状肿物，大小约 52mm×43mm×42mm，密度不均匀，中心见坏死低密度区，增强扫描见不均匀强化，较前稍增大，病变内缘紧贴纵隔胸膜，包绕右上肺各基底段支气管开口段，管腔稍狭窄。

Fig. 2. An instance of MAE (The highlighted parts are MS, PS and SP successively)

Analyzing the problem and data provided, we realize that as a subtask of the task MNER, MAE has some connections with MNER: two of the three attributes are belong to the anatomic site type of entity in MNER, as shown in Fig.2 and Fig.3; The SP attribute is related to the PS attribute, i.e., the SP is corresponding to the PS attribute rather than the MS attribute and there will be no sizes if there is no extracted primary tumor site.

左肺癌化疗后，对比 2015-05-05 片： 右上肺门区肿物较前稍大，右上肺内转移瘤较前无明显变化。 两锁下、两下气管旁、血管前间隙、两上气管旁、主-肺动脉窗、主动脉旁、隆突上及两肺门多发转移淋巴结，较前无明显变化 右侧胸膜多发转移瘤，部分较前稍大；右侧胸腔积液较前稍增多。 肝内数个囊肿。 肝 S5 肝内胆管结石。左肺癌化疗后，对比 2015-05-05 片： 左上肺肺门区见一团块状肿物，大小约 52mm×43mm×42mm，密度不均匀，中心见坏死低密度区，增强扫描见不均匀强化，较前稍增大，病变内缘紧贴纵隔胸膜，包绕右上肺各基底段支气管开口段，管腔稍狭窄。

Fig. 3. An example of NER results (The highlighted mentions are anatomic site entities recognized).

2.2 NER model

The architecture of our bidirectional long short-term memory with a CRF layer (BiLSTM-CRF) model is depicted in Fig.4.

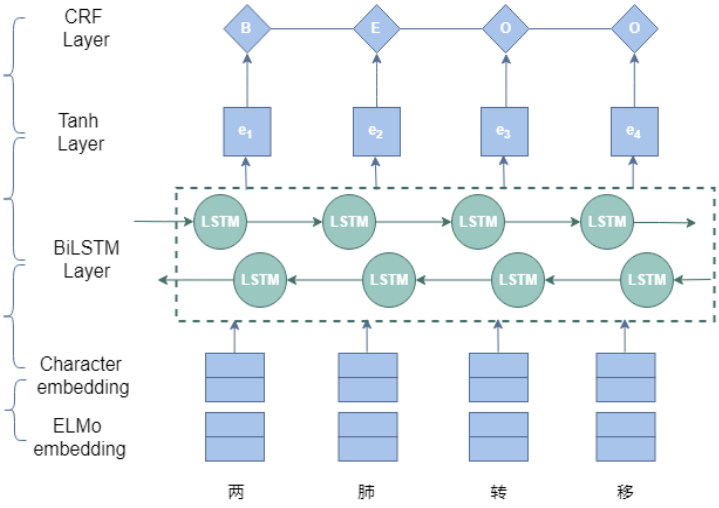

Fig. 4. NER model

Since character-based methods outperform word-based methods for Chinese NER [2,3], we use character embedding directly. Besides, after exploring the characteristics of ELMo [4], our team develop a novel stroke ELMo embedding as an additional

feature. The sentence representation consists of the above two embeddings is fed into a BiLSTM [5] encoder in which the forward LSTM computes a representation of the sequence from left to right and the backward one computes in reverse, generating the representation of every word in the sentence by concatenating the word's left and right context representations. Then we add a tanh layer on top of the BiLSTM layer to learn higher features. Finally, we obtain the best sequence path in all possible tag paths by adding a CRF layer.

2.3 PS Module and MS Module

PS Module. PS module is devised for extracting the primary tumor site and the size of the primary tumor site (as shown in Fig.5).

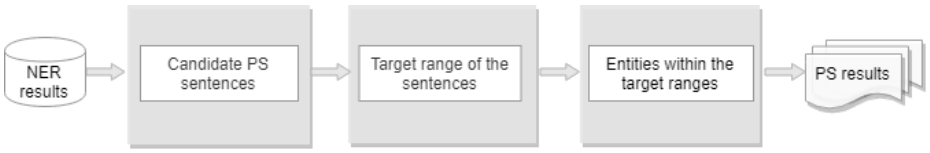

Fig. 5. The processing flow of PS module

This module requires the output of NER model as input and chooses candidate sentences according to the keywords we picked out for PS in advance. Then it selects the target-range in the candidate sentence in the light of the keywords we set before hand and extracts all anatomic site entities within the range. In this module, keywords for choosing candidate sentences(sentence-keywords) are '癌', '切除术', 'HCC' and 'MT', which are all trigger words of cancer, and main keyword for selecting target-range (range-keyword) is '转移', which means we only extract entities before sentence-keywords and after range-keywords if range-keyword appears in current sentence. There is one exception that we will elect the entity after sentence-keywords, not before, as PS when the sentence with sentence-keywords is describing the size of the primary tumor site.

After extracting all PSs in one clinical text, we scan all sentences in the medical text and extract the size nearest to the PS in one sentence if there is a PS in that sentence. Sizes have to match with regexes we constructed in advance. In the exception case of the last paragraph, the size mentioned will be selected directly as the SP corresponding to the PS mentioned.

MS Module. Similar to PS module, MS module is designed for extracting the metastasis site. We also input NER results into it and receive candidate sentences filtered by the sentence-keywords. MS module extracts all anatomical site entities within the target-range in the candidate sentence. Keyword for candidate sentences is '转移', and main range-keyword is '癌'. It means we only extract entities before sentence-keywords and after range-keywords if there is range-keyword in the candidate sen-

tence. Note that when it comes to '淋巴结转移', we have to merge all entities in the target-range of one sentence.

3 Experiments and Results

3.1 Dataset

We collected a total of 3,005 clinical texts from CCKS-2017 challenge and CCKS-2018 challenge and trained 100-dimensional character embedding by the cw2vec tool [6] as pre-trained character embedding. Then we merged the training and development dataset of CCKS-2018 challenge, and randomly split 20% of them as development set to train the BiLSTM-CRF model.

For MAE evaluation, we use the training set provided by the organizers in the CCKS-2019 challenge. The training set consists of 900 non-target-condition and 100 target-condition medical records annotated with three attributes of the primary tumor sites, the size of the primary site and the sites of the metastasis. As for the test set, all we know is that it covers 400 target-condition medical records.

3.2 Evaluation method

We follow the evaluation method presented by the organizer, which considers all of the entities of all three attributes extracted rather than three attribute values, since there will be more than one entity within one attribute value.

For each entity of an attribute, we only accept the correct match strictly that the ground truth and extraction result share same mention and same boundaries. In the next section, for better understanding, we give the micro-average precisions (Prec.), recalls (Rec.) and F-scores (F) of three attributes in the training sets, respectively.

3.3 Results

Since the accurate attribute answers of the 400 texts in the test set are not released yet, we are incapable of computing the Prec., Rrec. and F separately. For best submission of the test set, we got only the integral F of three attributes is 70.69%. The results of three attributes in the training set are listed respectively in Table 1, Table 2 and Table 3.

Table 1. Extraction results of the primary tumor site

	texts number	Prec.(%)	Rrec. (%)	F(%)
Target(training)	100	48.81	80.51	60.78
Non-target(training)	900	46.79	62.24	53.42

We can easily find that the same strategy performs differently when the condition changes. For example, the F of the tumor primary sites on target condition is about 7

percent higher than that on the non-target condition. Conversely, the performance of the metastasis site on non-target condition is much better, achieving an F of 70.41%. Experimental results reveal that the Rrec. of the tumor primary site extraction are all higher than Prec., while the metastasis site results are just the reverse. Higher Prec. reinforces that we extracted a lot of primary sites, but we could not assure the accuracy; even though we got not enough metastasis sites, two thirds of them were correct.

Table 2. Extraction results of the metastasis site

	texts number	Prec. (%)	Rrec. (%)	F(%)
Target(training)	100	60.14	58.04	59.07
Non-target(training)	900	73.57	67.51	70.41

Table 3. Extraction results of the size of primary tumor site

	texts number	Prec. (%)	Rrec. (%)	F(%)
Target(training)	100	63.63	45.16	52.83
Non-target(training)	900	42.85	33.73	38.26

4 Conclusion

In this paper, we propose a medical attribute extraction approach (NER-PS-MS) based on medical NER. In this approach, we explored the performance of MAE fed by MNER labeled sentences. The experimental results suggest that this approach is effective to identify and extract the answer of the predefined attributes. At last, our best submission achieves the 70.69% of F considering all three attributes.

The NER-PS-MS approach is dependent on the performance of NER model, influenced by the error accumulation problem. In the future, we will explore an independent machine learning approach for attribute extraction.

5 Reference

1. Jun Y: https://mp.weixin.qq.com/s/Z6UfCC5vSF74K6qE9E6gCA, 2019.
2. He J, Wang H: Chinese named entity recognition and word segmentation based on character. *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing:2008*.
3. Li H, Hagiwara M, Li Q, Ji H: Comparison of the Impact of Word Segmentation on Name Tagging for Chinese and Japanese. *LREC: 2014. 2532-2536*.
4. Peters M. E, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L: Deep Contextualized Word Representations. *NAACL-HLT 2018: 2227-2237.*
5. Huang Z, Xu W, Yu K: Bidirectional LSTM-CRF Models for Sequence Tagging[J]. *Computer Science, 2015.*
6. Uzuner Ö, Solti I, Cadag E: Extracting medication information from clinical text. *Journal of the American Medical Informatics Association 2010, 17(5):514-518.*