Medical Named Entity Recognition using CRF-MT-Adapt and NER-MRC

Hengyi Zheng[1], Rui Wen[2] (✉), Xi Chen[2] (✉),
Ziheng Zhang[2], Yifang Yang[2], Yunyan Zhang[2], and Ming Xu[1] (✉)

[1] Shenzhen University, Shenzhen, China
[2] Tencent Jarvis Lab, Shenzhen, China
{ruiwen,jasonxchen}@tencent.com
xuming@szu.edu.cn

Abstract. Medical Named Entity Recognition is a fundamental component of understanding the medical free-text notes in Electronic Health Records, and it has become a popular research topic in both academia and industry. The China Conference on Knowledge Graph and Semantic Computing (CCKS) organizes a challenge for Medical Named Entity Recognition, aiming at extracting medical entity mentions and categorizing them into pre-defined classes. We propose a Multi-Task sequence labeling model with Adaptive Loss Weighting (*CRF-MT-Adapt*) to address the issue of low recall and a Named Entity Recognition model based on Machine Reading Comprehension (*NER-MRC*) to address the issue of long-span entity mentions. We experimentally demonstrate the state-of-the-art performance of the two proposed models and the ensemble even surpasses the strong baselines by at least 2% F-score. On the official test set, our best submission achieves an F-score of 90.51% and 95.96% under strict and relaxed criteria respectively.

Keywords: Electronic Health Records · Named Entity Recognition · Machine Reading Comprehension

1 Introduction

The Electronic Health Records (EHRs) contain a large number of medical free-text notes, particularly covering important health-related information such as demographics, medical history, medication and allergies, laboratory test results, radiology images, and vital signs. Medical Named Entity Recognition (NER) is of fundamental importance to understanding medical free-text notes, mining useful information, and discovering knowledge. Therefore, the 2020 China Conference on Knowledge Graph and Semantic Computing (CCKS 2020) organizes the Medical Named Entity Recognition task aiming at semantic computing of Chinese EHRs, or specifically, extracting medical entity mentions and categorizing them into six pre-defined classes (i.e., disease, anatomy, imaging examination, drug, operation, and laboratory examination) from any given Chinese EHR.

Named Entity Recognition (NER) is usually considered as a sequence labeling task [5]. According to [5], most frequently used NER models are either statistical models or neural networks. Statistical models, such as Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM) [9], and Conditional Random Field (CRF) [10], were proposed to infer the entire annotation sequence with the optimal joint probability as the objective. Collobert *et al.* [1] were the first to apply neural networks to address the sequence labeling task, in which pre-trained word vectors were obtained as the input features, greatly eliminating labor-intensive feature engineering. With the rapid development of neural networks, the combination of neural networks and statistical models has become the baseline models for general-purpose NER, such as BiLSTM-CRF [8] or CNN-CRF [1].

However, previous CRF-based models typically combines entity extraction and categorization, which leads to higher model coupling and weak model interpretability. Moreover, CRF itself is based on a linear Markov chain and is not ideal for extracting long-span entity mentions. To address the aforementioned issues, we propose two novel models for Medical NER, including 1) a Multi-Task sequence labeling model with Adaptive Loss Weighting (*CRF-MT-Adapt*) and 2) a Named Entity Recognition model based on Machine Reading Comprehension (*NER-MRC*). Specifically, CRF-MT-Adapt divides the NER task into two sub-tasks: entity mention extraction and entity categorization, to simplify the NER task and to make possible the manual intervention for performance improvement. Since the entity mention extraction solely focus on the recognition of medical entities and decoupled from entity categorization, thus significantly increasing the recall of the model. NER-MRC, on the other hand, applies machine reading comprehension to recognize entities, which just mark start and end span of entities from origin passage rather than sequence labeling, thus increasing model performance for long-span entities. We further propose an ensemble of CRF-MT-Adapt and NER-MRC and it achieves the state-of-the-art performance on the CCKS 2020 Medical NER dataset. On the official test set, our best submission achieves an F-score of 90.51% and 95.96% under the strict and relaxed criteria respectively.

2 Model Description

2.1 CRF-MT-Adapt Model

In sophisticated medical corpora, we find that some baselines, such as BERT +BiLSTM+CRF, often suffer from poor recall performance. One of the accountable reasons is that these baselines cannot work well on long-tail entities. The other could be that these baselines, typically sequence labeling models, combines entity extraction and categorization, thus increasing the model complexity. Our proposed Multi-Task sequence labeling model with Adaptive Loss Weighting (*CRF-MT-Adapt* for short) decouples the NER task into two stages, namely entity mention extraction and entity categorization, as shown in Fig. 1. CRF-MT-Adapt has two benefits that it firstly reduces the model complexity in each stage and it also makes possible the manual or rule-based intervention in each stage, hopefully improving model performance.

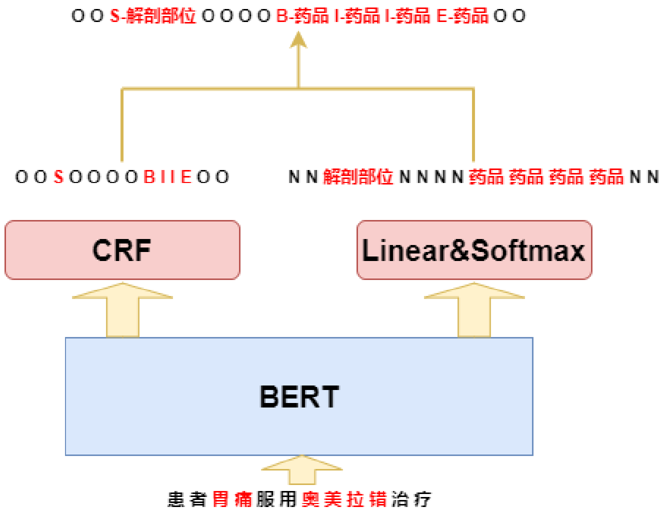
O O S-解剖部位 O O O O B-药品 I-药品 I-药品 E-药品 O O
O O S O O O O B I I E O O
N N 解剖部位 N N N N 药品 药品 药品 药品 N N
CRF
Linear&Softmax
BERT
患者胃痛服用奥美拉错治疗

Fig. 1. CRF-MT-Adapt Model.

The first stage of CRF-MT-Adapt, entity mention extraction, is shown on the left part of Fig. 1. The word vectors of the input sequence are inferred by a pre-trained language model, such as BERT [2] or RoBERTa [7], which represent the semantics of each character in the input sequence. These word vectors are then fed into a CRF layer after linear transformation to predict the position and the span of entity mentions. The second stage of CRF-MT-Adapt, entity categorization, is shown on the right part of Fig. 1, and it shares the same word vectors, inferred by the pre-trained language model, with the first stage. The model then maps the word vectors, with semantic information contained, into the output layer via linear transformation as a multi-class classification task. For each token, we obtain the category by calculating the maximum category probability after Softmax. Through combining the labelled results from two stages, the NER results, including mention spans and entity categories, of the input sequence are obtained.

Generally speaking, in multi-task models, the weights of different task losses greatly influence the model convergence and the final performance. To avoid the cumbersome manual tuning of weights, we adopt an adaptive weight based on the uncertainty of the same variance [3]. The loss function is therefore composed of two parts: the loss of entity mention extraction denoted as $\mathcal{L}_{CRF}$ and the loss of entity categorization denoted as $\mathcal{L}_{CE}$. The final loss function is the weighted sum of the above two parts, as shown in Eq. 1.

$$\mathcal{L}_{Total} = \frac{1}{\sigma_1^2}\mathcal{L}_{CRF} + \frac{1}{\sigma_2^2}\mathcal{L}_{CE} + \log\sigma_1 + \log\sigma_2 \tag{1}$$

where σ is a learnable parameter that reflects the uncertainty (or noise) of the same data in different tasks, and σ_1 controls the weight of $\mathcal{L}_{CRF}$ and σ_2 controls the weight of $\mathcal{L}_{CE}$. The greater noise would lead to the lower confidence of the sub-task result and then lower proportion of the sub-task loss. The logarithmic term is designed for regularization, meaning that during the gradient descent, the weight part $\frac{1}{\sigma^2}$ of the loss expects σ to increase, while the logarithmic part $\log\sigma$ expects σ to decrease, which regularizes the learning of σ.

Compared with traditional sequence labeling methods, our CRF-MT-Adapt model has the following two advantages:

1. CRF-MT-Adapt separates the NER task into entity mention extraction and entity categorization to reduce the model complexity and avoid the high coupling, and CRF-MT-Adapt improves the F-scores of both sub-tasks (cf. Section 3.5 for experimental results).
2. In practice, CRF-MT-Adapt is more interpretable because of low coupling. Two sub-tasks, namely entity mention extraction and entity categorization, are separately controllable, which makes possible the corresponding manual intervention or strategies and is thus more suitable to industrial applications.

2.2 NER-MRC Model

We propose another novel model for the Medical NER task, namely Named Entity Recognition model based on Machine Reading Comprehension (*NER-MRC* for short) as shown in Fig. 2, which transforms the NER task into BERT-based MRC task [6].

The core idea of the proposed model is to construct *Question* for each entity category (six pre-defined categories), to utilize original texts as *Passage*, and to predict the start and end positions of given entity in the *Passage*. In this work, we construct a total of six *Questions* for each entity category based on their medical definition and labeling guidelines; for instance, the *Question* for the category of "disease" is shown as the input in Fig. 2. The loss function for NER-MRC model consists of the Binary Cross-Entropy (BCE) loss at the start position and the end position, as shown in Eq. 2.

$$\begin{aligned}\mathcal{L}_{start} &= BCE(P_{start}, Y_{start})\\ \mathcal{L}_{end} &= BCE(P_{end}, Y_{end})\\ \mathcal{L}_{Total} &= \mathcal{L}_{start} + \mathcal{L}_{end}\end{aligned} \tag{2}$$

We summarize two advantages of the proposed NER-MRC method over the traditional NER models:

1. NER-MRC model introduces the prior information via the designed *Question*. For example, for the category of "anatomy", *Question* is constructed

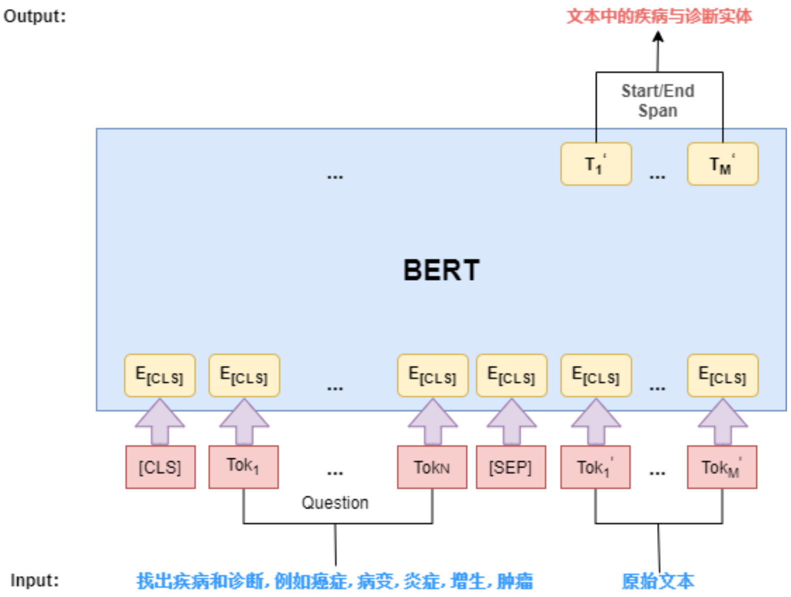

Fig. 2. NER-MRC Model.

as follows “to find out the anatomy or anatomical parts of the human body where all diseases, symptoms and signs occur”. Since the entity category definition already contains implicit prior knowledge of the category, it becomes much easier for NER-MRC model to categorize the entities.

2. Because CRF-based models are based on a linear Markov chain, the labeling is related to the state transfer, thus not friendly to long-span entities. NER-MRC model is more effective in extracting long-span entities as it is based on the prediction of the start and end positions.

3 Experiments

3.1 Dataset

In the CCKS 2020 Medical NER challenge, the organizing committee provides 1,050 labeled EHRs as the training set with six pre-defined categories, including disease (DI), anatomy (AN), image (IM), drug (DR), operation (OP), and laboratory (LA). In addition, the organizing committee provides another 300 unlabeled EHRs as the test set for the model evaluation. In our experiments, we split the 1,050 labeled EHRs into two partitions, 950 EHRs as the training set and 100 EHRs as the validation set, and the 300 unlabeled EHRs are used as the test set.

3.2 Dataset Analysis and Data Pre-processing

Before designing the models, we firstly conduct quantitative analysis of the original dataset provided by the CCKS 2020 organizing committee. We present the statistics of the Medical NER dataset (only the training set of 1,050 EHRs) in Table 1, including character-level text length distribution and category-based entity distribution.

Table 1. Statistics of the Medical NER dataset provided by CCKS 2020.

Character-level Text Length Distribution							
Metric	Mean	Std.	Min.	25%	50%	75%	Max.
Length	409	154	172	307	380	488	1436
Category-based Entity Distribution							
Abbr.⋆	*DI*	*AN*	*IM*	*DR*	*OP*	*LA*	Overall
Count	4345	8811	1002	1935	923	1297	18313

⋆ The abbreviation is as follows: disease (*DI*), anatomy (*AN*), image (*IM*), drug (*DR*), operation (*OP*), and laboratory (*LA*).

Based on the aforementioned statistical analysis, we pre-process the dataset as follows:

1. We find that the dataset contains some meaningless characters, such as "\n", "\r", etc; we therefore replace these characters with "[UNK]" before applying word segmentation to the dataset.
2. The average text length of the dataset is 409 while the maximum text length is 1436. It is necessary to divide the texts before feeding them to the model. In this paper we set the maximum text length as 256, and for texts longer than 256, we employ a dynamically planned text segmentation method[3] to reduce longer texts while retaining as much information as possible. An example of the text segmentation is displayed in Fig. 3.

3.3 Ensemble of CRF-MT-Adapt and NER-MRC

In addition to CRF-MT-Adapt and NER-MRC as single models, we also propose an ensemble of both CRF-MT-Adapt and NER-MRC with voting strategies to further improve the performance. To address some obvious issues in preliminary fusion results and improve ensemble strategy, we formulate some simple rules into the ensemble:

[3] https://github.com/caishiqing/joint-mrc

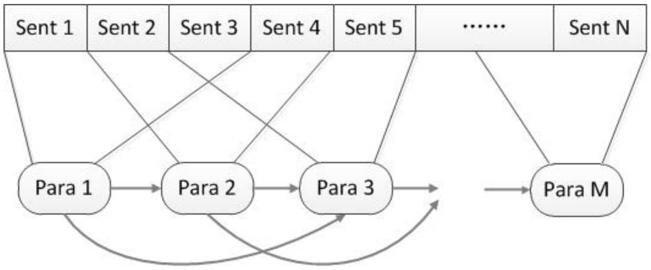

Fig. 3. Example of Text Segmentation.

1. Overlapped entities removing after fusion: The final results after voting and fusion could have overlapped entities, which is divided into two phenomena: 1) same mention boundary but different entity categories, and 2) overlapped boundary. The first phenomenon does not exist in the dataset provided by CCKS 2020 and the predicted results of both proposed models. The second phenomenon of overlapped boundary is the main problem. If multiple entities have overlapped boundaries, we retain the most frequent entity and discard the rest, thus addressing the second phenomenon.
2. Vocabulary matching of drug entities: We find that drug entities are mostly glossaries and do not suffer from "inclusion" issue (i.e. one entity in another entity). Therefore, we filter out drug entities from the entity vocabulary provided by CCKS 2020 organizing committee, use string matching (by regular expression) to search for drug entities in the input sequence, and then append the string matching results into the model results to obtain the final ensemble results.

3.4 Experimental Setting

For our model, we apply a grid search for hyper-parameters and find the best configuration: the optimizer is Adam [4] with weight decay rate of 0.01, the learning rate is 1e-4, the batch size is 24, and the maximum length of input sequence is 256. In our experiment, the BERT hyperparameters are set as BERT$_{large}$ with 24 layers, 1024 hidden dimensions, and 16 multi-heads. The evaluation metrics are F-scores under two criteria: 1) strict criteria in which a correct match requires same mention, same boundaries and same entity type 2) relaxed criteria in which a correct match requires same entity type and overlapped boundaries.

3.5 Experimental Results

Table 2 lists both strict and relaxed F-scores of various models on the official test set from overall aspect and from entity category aspect[4]. We implement

[4] The experimental results are obtained via the CCKS 2020 official evaluation platform https://www.biendata.xyz/competition/ccks_2020_2_1/.

three strong baselines, namely {BERT/RoBERTa/NEZHA}+BiLSTM+CRF[5], all of which are built on the state-of-the-art pre-trained language models [2, 7, 11]. We compare our proposed CRF-MT-Adapt, NER-MRC and the ensemble against these strong baselines

Table 2. Experimental results (F-scores) of various models.

Models	Entity category						Overall
	DI	*AN*	*IM*	*DR*	*OP*	*LA*	
BERT+BiLSTM+CRF †	.8580	.8683	.8601	.9166	.9144	.8014	.8723
RoBERTa+BiLSTM+CRF †	.8492	.8763	.8758	.9177	.9327	.7930	.8755
NEZHA+BiLSTM+CRF †	.8811	.8775	.8537	.9152	.9312	.7839	.8808
CRF-MT-Adapt (Ours) †	.8763	.8966	.8590	.9235	.9202	.8207	.8916
NER-MRC (Ours) †	.8578	.8808	.8576	.9108	.8903	.7700	.8736
Ensemble †	**.8992**	**.9043**	**.8769**	**.9310**	**.9375**	**.8503**	**.9051**
Ensemble ‡	.9717	.9592	.9046	.9731	.9687	.9015	.9596

† indicates strict F-score and ‡ indicates relaxed F-score.
The best performance is in **bold** and the second best performance is underlined.

We find that CRF-MT-Adapt outperforms the baselines by at least 1% F-score in the overall evaluation and performs more balanced across different entity categories. Furthermore, CRF-MT-Adapt has the best performance in half entity categories (*AN*, *DR*, and *LA*) when compared with other single models. While NER-MRC does not have the highest F-score in any entity categories or the overall evaluation, it still shows competitive performance to the strong baselines. In particular, it provides more differentiated results which has been proved to bring better results in model ensemble. The ensemble of CRF-MT-Adapt and NER-MRC shows significant performance improvement (up to 3% in F-score) over single models, which demonstrates the effectiveness of the ensemble strategy, namely incorporating the long-span entity mentions extracted from NER-MRC into the final results.

4 Conclusions

In this paper, we propose two novel models, CRF-MT-Adapt and NER-MRC, to the Medical NER task. Different from traditional CRF-based models, CRF-MT-Adapt model shows performance improvement and better interpretability, and NER-MRC model shows better capability in extracting the long-span entities. In future work, we plan to investigate how to extract entity boundaries

[5] The hyperparameters for RoBERTa and NEZHA are set as RoBERTa$_{large}$ and NEZHA$_{large}$ respectively.

more accurately and how to effectively merge multiple entities. We also plan to attempt model distillation and pruning approaches to increase the suitability of our proposed NER models in real-world industrial applications.

References

1. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**(null), 2493–2537 (Nov 2011)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
3. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. CoRR **abs/1705.07115** (2017), http://arxiv.org/abs/1705.07115
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. CoRR **abs/1812.09449** (2018), http://arxiv.org/abs/1812.09449
6. Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J.: A unified MRC framework for named entity recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5849–5859. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.519, https://www.aclweb.org/anthology/2020.acl-main.519
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 **abs/1907.11692** (2019)
8. Ma, X., Hovy, E.H.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. CoRR **abs/1603.01354** (2016), http://arxiv.org/abs/1603.01354
9. McCallum, A., Freitag, D., Pereira, F.C.N.: Maximum entropy markov models for information extraction and segmentation. In: Proceedings of the Seventeenth International Conference on Machine Learning. p. 591–598. ICML ’00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2000)
10. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. pp. 188–191 (2003), https://www.aclweb.org/anthology/W03-0430
11. Wei, J., Ren, X., Li, X., Huang, W., Liao, Y., Wang, Y., Lin, J., Jiang, X., Chen, X., Liu, Q.: Nezha: Neural contextualized representation for chinese language understanding (2019)