Medical Named Entity Recognition from Un-labelled Medical Records based on Pre-trained Language Models and Domain Dictionary Post-print

Authors: Wen Chaojie, Tao, Chen, Xudong, Jia, Jiang, Zhu, Tao, Chen

Date: 2022-11-27T00:00:00+00:00

Abstract

Medical named entity recognition (NER) is an area in which medical named entities are recognized from medical texts, such as diseases, drugs, surgery reports, anatomical parts, and examination documents. Conventional medical NER methods do not make full use of un-labelled medical texts embedded in medical documents. To address this issue, we proposed a medical NER approach based on pre-trained language models and a domain dictionary. First, we constructed a medical entity dictionary by extracting medical entities from labelled medical texts and collecting medical entities from other resources, such as the Yidu#2; N4K data set. Second, we employed this dictionary to train domain-specific pre-trained language models using un-labelled medical texts. Third, we employed a pseudo labelling mechanism in un-labelled medical texts to automatically annotate texts and create pseudo labels. Fourth, the BiLSTM-CRF sequence tagging model was used to fine-tune the pre-trained language models. Our experiments on the un-labelled medical texts, which were extracted from Chinese electronic medical records, show that the proposed NER approach enables the strict and relaxed F1 scores to be 88.7% and 95.3%, respectively.

Full Text

Preamble

Medical Named Entity Recognition from Un-labelled Medical Records based on Pre-trained Language Models and Domain Dictionary

Chaojie Wen, Tao Chen†, Xudong Jia & Jiang Zhu
Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen 529020, China

Keywords: Medical named entity recognition; Pre-trained language model; Domain dictionary; Pseudo labelling; Un-labelled medical data

Citation: Wen, C.J., et al.: Medical named entity recognition from un-labelled medical records based on pre-trained language models and domain dictionary. Data Intelligence 3(3), 402-417 (2021). doi: 10.1162/dint_a_{00105}

Received: February 15, 2021; **Revised:** March 29, 2021; **Accepted:** April 22, 2021

Corresponding author: Tao Chen (Email: chentao1999@gmail.com; ORCID: 0000-0002-3634-0854)

Abstract

Medical named entity recognition (NER) is the task of identifying medical entities from clinical texts, including diseases, drugs, surgical procedures, anatomical parts, and examination reports. Conventional medical NER methods do not fully utilize the vast amounts of un-labelled medical texts embedded in electronic health records. To address this limitation, we propose a medical NER approach based on pre-trained language models and a domain dictionary. First, we constructed a medical entity dictionary by extracting entities from labelled medical texts and incorporating additional medical entities from external resources such as the Yidu-N4K dataset. Second, we leveraged this dictionary to train domain-specific pre-trained language models using un-labelled medical texts. Third, we employed a pseudo-labelling mechanism to automatically annotate un-labelled texts and generate pseudo-labels. Finally, we fine-tuned the pre-trained language models using a BiLSTM-CRF sequence tagging model. Experiments conducted on un-labelled medical texts extracted from Chinese electronic medical records demonstrate that our proposed NER approach achieves strict and relaxed F1 scores of 88.7% and 95.3%, respectively.

---

1. Introduction

Medical records are critical resources that document patients' diagnostic and treatment activities in healthcare facilities. In recent years, numerous medical institutions have made substantial progress in digitizing electronic medical records, gradually replacing handwritten records with digital formats. Many researchers have been working to extract medical knowledge from this digital data, aiming to help healthcare professionals understand potential causes of various symptoms and build medical decision support systems.

Medical named entity recognition (NER) has recently gained significant attention in the medical community as an important technique for extracting named entities from clinical texts, such as diseases, drugs, surgical reports, anatomical

parts, and examination documents [1]. For instance, the Chinese Information Processing Society of China (CIPSC) has established the CCKS 2020 Medical Entity and Event Extraction from Chinese Electronic Medical Records (Chinese EMRs) program. This initiative aims to identify medical entities from archived Chinese EMRs in plain text format, grouping them into six pre-defined categories: disease & diagnosis, imaging examination, laboratory examination, surgery, drug, and anatomic part (CCKS 2020 Medical Entity and Event Extraction from Chinese EMRs Task 1: Medical Entity Recognition) . As a major contributor to this program, Yidu Cloud Co., Ltd. has recently constructed the first medical dataset from Chinese electronic medical records (referred to hereafter as the CCKS-2020-CEMR dataset). This dataset contains a dictionary of 6,292 words, 1,000 un-labelled texts, and 1,500 labelled texts, with 26,414 medical entities annotated in the labelled texts. This dataset is expected to evolve into a benchmark for medical named entity recognition, and we used it to test and evaluate our NER approach.

The CCKS-2020-CEMR dataset is summarized in Table 1 , and an example of labelled texts is shown in Figure 1 [Figure 1: see original paper]. While various medical NER methods and models have been developed for extracting medical knowledge from datasets, these approaches are predominantly supervised methods that learn features from labelled data. Consequently, un-labelled data cannot be effectively utilized and is often discarded, causing valuable medical information hidden in un-labelled texts to remain unrecognized. To address this issue, we propose a medical NER approach that combines pre-trained language models with a domain dictionary, employing pre-trained language models and in-domain training techniques to extract medical knowledge from un-labelled medical texts.

The main contributions of our work are summarized as follows:

- We constructed a medical domain entity dictionary and integrated it into pre-trained language models, significantly improving the recognition rate (measured by F1 score) of named entities.
- We fed un-labelled medical texts to pre-trained language models for in-domain training. Experimental results demonstrate that the proposed NER approach with in-domain training can improve the performance of named entity recognition in the medical domain.
- We employed a pseudo-labelling method to augment the CCKS-2020-CEMR dataset by automatically annotating un-labelled texts and marking them as pseudo-labels.

The remainder of this paper is organized as follows: Section 2 reviews previous research work related to medical NER. With a solid understanding of the state-of-the-art in medical knowledge extraction, we present our medical NER approach in Section 3. Section 4 describes the evaluation of our medical NER approach using the CCKS-2020-CEMR dataset and analyzes the results obtained with pre-trained language models and an entity dictionary. Section 5 concludes the paper by summarizing the research contributions and outlining

future research directions.

2. Previous Research Work on Medical NER

Medical NER [2] is a fundamental task for extracting knowledge representation from electronic medical records. Solutions to medical NER can be broadly categorized into two groups: traditional machine learning-based methods and neural network-based methods.

Traditional machine learning-based methods primarily treat NER as a sequence labeling task. Wang et al. [3] conducted a study using conditional random fields (CRFs), support vector machines (SVM), and maximum entropy (ME) to recognize symptoms and pathogenesis in Chinese EMRs. Wang et al. [4] investigated the use of CRF with different feature sets to identify symptom names from clinical notes in traditional Chinese medicine. Xu et al. [5] proposed a joint model that integrated segmentation and NER simultaneously to improve performance in Chinese discharge summaries. These methods rely heavily on hand-crafted features.

Unlike conventional machine learning-based methods, neural network-based methods employ an end-to-end training process that automatically learns features from data. Numerous studies have addressed medical NER using neural network-based approaches. Wu et al. [6] developed a deep neural network approach for NER in Chinese clinical texts. Yang et al. [7] combined characteristics of Chinese electronic language structure, developed labeling rules for named entity recognition in Chinese EMRs, and completed natural language processing research for knowledge extraction from Chinese EMRs. Yang et al. [8] used a combination of Bi-directional Long Short-Term Memory (BiLSTM) and CRF to extract medical entities from admission records and discharge summaries. Chowdhury et al. [9] proposed a multitask bi-directional Recurrent Neural Network (RNN) model to recognize diseases, symptoms, and other entities in Chinese electronic medical records, employing two different task layers (word embedding and character embedding) to improve extraction accuracy. Furthermore, Wan et al. [10] proposed a Chinese medical NER method based on joint training of Chinese characters and words, which considered Chinese language features and incorporated semantic information of words into the NER process.

However, all these medical NER methods and models do not fully utilize un-labelled medical texts embedded in medical documents. Therefore, developing a medical NER approach that can recognize named entities from un-labelled texts is crucial.

3. NER Methodology for Medical Texts

This section presents our approach for medical NER using pre-trained language models and a domain dictionary. An overview of the approach is shown in Figure 2 [Figure 2: see original paper]. First, we preprocessed labelled medical texts (Section 3.1) and constructed a medical entity dictionary (Section 3.2). Second, we introduced a pseudo-labelling method to automatically annotate un-labelled texts with pseudo-labels (Section 3.3). Third, we used the BiLSTM-CRF technique to fine-tune the pre-trained language models (Section 3.4). Fourth, we employed medical domain pre-trained language models (including the WoBERT model) and trained them using un-labelled medical texts (Section 3.5). Finally, we combined results from different pre-trained models for named entity recognition (Section 3.6).

3.1 Data Preprocess

Before implementing our medical NER approach to extract knowledge from the labelled texts in the CCKS-2020-CEMR dataset, we identified errors in the dataset that required preprocessing. The data preprocessing task consists of four steps:

Data cleaning. We removed obvious labeling errors in the dataset. For example, we eliminated extra spaces in terms like "直肠癌根治术" and "胃角高级别下皮内瘤变" (meaning "radical resection of rectal cancer" and "gastric angle with high grade endothelial neoplasia" in English). We also fixed entity boundary errors.

Text normalization. We converted full-width Chinese symbols to half-width symbols and changed uppercase English letters to lowercase.

Corpus format conversion. We converted labelled texts from CEMR format to BIO format. In BIO format, each element is labeled as "B-X", "I-X", or "O", where "B-X" indicates the beginning of an entity of type X (type X refers to one of the six medical categories: disease & diagnosis, imaging examination, laboratory examination, surgery, drug, and anatomic part), "I-X" indicates the middle of an entity of type X, and "O" indicates that the word is not a medical entity. The 1,500 BIO-formatted texts were used as training data for our proposed medical NER approach.

Corpus segmentation. Considering the limitation of LSTM methods in handling long sequences, we segmented medical records into shorter sequences while ensuring the relative integrity of sentences. Each sequence was limited to 200 characters, with periods and other symbols used to retain semantic information. [CLS] and [SEP] tokens were also added.

3.2 Domain Dictionary

The CCKS-2020-CEMR dataset contains a medical dictionary with 6,292 medical words. To expand this dictionary, we crawled related medical terms from

public medical websites . Duplicate or highly similar medical words were removed, resulting in a collection of 11,000 medical words related to drugs, diseases, anatomical parts, and other medical information types.

We used the word-piece [11] format to represent medical words in our dictionary. Word-piece is a text representation method that divides words into sub-word units. First, we divided words into sub-words in the training set, then added special word boundary symbols before testing and evaluating our NER approach. This allowed original word sequences to be recovered from text sequences while keeping them unchanged. For example, the word-piece representation of "the highest mountain"is "the high ##est moun ##tain" , where "highest"is divided into "high" and "##est" , and "mountain" into "moun" and "##tain" .

The word-piece format is also suitable for Chinese word representation. For instance, the Chinese medical entity "左侧颈动脉中段" (mid left carotid artery) is split into four pieces: "左侧" (left), "## 颈" (neck), "## 动脉" (artery), and "## 中段" (middle). Using the word-piece format, we segmented Chinese words from the 11,000 medical entities in our dictionary, converted them into 30,000 word pieces, and incorporated all these word pieces into the original dictionaries of pre-trained language models.

3.3 Pseudo Labelling

During the in-domain training process with the WoBERT_{FPT} model, we adopted a pseudo-labelling method to augment the labelled texts for this study. Pseudo-labelling [12] is an unsupervised learning method that improves NER model performance when extracting knowledge from un-labelled texts. Our pseudo-labelling method first uses labelled texts to train the BiLSTM-CRF model, then applies the trained model to predict labels for un-labelled texts. These predicted labels are called pseudo-labels. Labelled texts and pseudo-labelled texts are then combined as a new training set for named entity recognition.

The framework for the pseudo-labelling method is shown in Figure 3 [Figure 3: see original paper]. Labelled texts are used to train the RoBERTa-wwm-ext-large-BiLSTM-CRF model, which then predicts pseudo-labels for un-labelled texts. Pseudo-labelled texts with high confidence levels are added to the original training set.

3.4 Pre-trained Language Models

We developed three pre-trained models in our NER approach to handle un-labelled medical texts: "WoBERT_{FurtherPreTraining}*{BiLSTM}-CRF*" *(W*{FPT}*{BC}),* " *RoBERTa-wwm-ext-large*" *(RWL), and* " *RoBERTa-wwm-ext-large*{BiLSTM}-CRF" (RWL_{BC}). All three models use labelled and pseudo-labelled texts (described in previous sections) as the training set and incorporate our medical dictionary as the dictionary file.

All these pre-trained models are derived from the RoBERTa model. RoBERTa [13] is a pioneering pre-trained language model that trains itself from a large-scale text corpus in an unsupervised manner, using Chinese characters as the basic processing unit. The RoBERTa-wwm-ext-large model improves upon RoBERTa by implementing Whole Word Masking (wwm) and masking Chinese characters that form the same words [14], effectively using Chinese words as the basic processing unit.

Pre-trained language models expand their dictionaries by adding words from the domain dictionary (described in Section 3.2). When training set texts are mapped to the expanded dictionary, they are assigned identifiers (IDs). Different words receive different IDs. All out-of-vocabulary (OOV) words in the training set that cannot be mapped to any word in the expanded dictionary are assigned a unique ID. For example, the word "维生素 D" (vitamin D) is not in the original dictionary of a pre-trained language model but exists in the domain dictionary. Before merging the domain dictionary into the original dictionary, "维生素 D" (vitamin D) in our training dataset is considered an OOV word and assigned an ID of "100". After embedding the domain dictionary into the original dictionary, "维生素 D" (vitamin D) can be found in the expanded dictionary and is assigned an ID of "7818".

Building on these three pre-trained models, our medical NER approach uses conventional sequence tagging models such as Bi-LSTM and CRF to fine-tune the pre-trained language models.

3.5 In-domain Training

Directly applying pre-trained language models to Chinese medical NER may yield unsatisfactory results due to word distribution shifts from general domain texts to Chinese medical texts. One way to alleviate this issue is to obtain domain-specific knowledge through in-domain training. In-domain training, also known as continuous training [15], is a technique that uses pre-trained language models to extract domain-specific knowledge from texts, providing substantial gains over models trained on general or mixed-domain knowledge.

In this paper, we used the WoBERT model for in-domain training. WoBERT is a pre-trained language model derived from the RoBERTa-wwm-ext model. It leverages the RoBERTa-wwm-ext model and optimizes its pre-training process for un-labelled Chinese texts. The medical domain dictionary described in Section 3.2 serves as the training dictionary for the WoBERT model, while the un-labelled texts in the CCKS-2020-CEMR dataset are used as the training set. After training, a domain-specific language model is obtained, referred to as the WoBERT with Further Pre-Training model (WoBERT_{FPT}). The WoBERT_{FPT} model can be adapted for Chinese medical texts.

3.6 NER Models for Medical Texts

In this study, we developed nine NER models: 1) BiLSTM-CRF, 2) BERT_{BC}, 3) W_{BC}, 4) W_{FPT}*{BC}, 5) W*{FPT}*{BC}+pseudo labelling, 6) W*{FPT}*{BC}+domain dictionary+pseudo labelling, 7) RWL*{BC}+domain dictionary+pseudo labelling, 8) RWL+domain dictionary+pseudo labelling, and 9) fusion model.

The BiLSTM-CRF model is a traditional model without a pre-trained module before sequence tagging and cannot extract hidden knowledge from un-labelled medical texts. The BERT_{BC} model uses BERT as its pre-trained module for named entity recognition. The W_{BC} model employs the WoBERT model as its in-domain pre-trained module before the BiLSTM-CRF (BC) sequence tagging process. The W_{FPT}*{BC} model also adopts the WoBERT model and expands its pre-training capabilities before BC-based sequence tagging. The W*{FPT}*{BC}+pseudo labelling model adds pseudo-labelled texts to the training set before executing the W*{FPT}*{BC} model—this is our first NER model for handling un-labelled medical texts. The W*{FPT}_{BC}+domain dictionary+pseudo labelling model incorporates both the domain dictionary and pseudo-labels when handling un-labelled medical texts—this is our second NER model for extracting knowledge from un-labelled medical texts.

The RWL_{BC}+domain dictionary+pseudo labelling model uses the "RoBERTa-wwm-ext-large" (RWL) model as its pre-trained module, working with the domain dictionary and pseudo-labels to recognize named medical entities through the BC method—this is our third NER model for un-labelled medical texts. The RWL+domain dictionary+pseudo labelling model uses the domain dictionary and pseudo-labels in the RWL process—this is our fourth NER model. The fusion model combines results from our NER models (specifically, "WoBERT_{FurtherPreTraining}*{BiLSTM}-CRF" (W*{FPT}*{BC}), " RoBERTa-wwm-ext-large" (RWL), and " RoBERTa-wwm-ext-large*{BiLSTM}-CRF" (RWL_{BC})) that deal with un-labelled medical texts.

---

4. Evaluation

We computed precision (P), recall (R), and F1 score to evaluate the performance of the nine medical NER models, where S refers to the predicted result set of the model and G refers to the golden result set.

We adopted two evaluation indicators—strict and relaxed F1 scores—to assess the effectiveness of the nine medical NER models in extracting medical knowledge. According to the boundary definition of recognized results, the strict indicator requires the predicted entity to exactly match the golden result, while the relaxed indicator only requires the predicted entity' s boundary to be covered by the golden result' s boundary.

We used the Adam Optimizer as our optimization algorithm for stochastic gradient descent when training our nine NER models. The Adam Optimizer combines the best properties of AdaGrad and RMSProp algorithms, providing an optimization method that can handle sparse gradients on noisy problems. In this research, the learning rate was set to 5e-5, the dropout rate to 0.5, and the batch size to 5. The BiLSTM model was configured with 128 hidden layers. The strict and relaxed F1 scores for each model are shown in Tables 2 and 3, respectively.

The results in Tables 2 and 3 demonstrate that pre-trained models improve the effectiveness of traditional NER models. For example, when the BERT pre-trained model is used before the BiLSTM-CRF model, the overall F1 score increases by 6.9% (strict indicator) and 5.1% (relaxed indicator). Using optimized pre-trained models further improves recognition accuracy. For instance, the overall F1 score of the NER model combined with the WoBERT model is 2.2% (strict indicator) and 3.8% (relaxed indicator) higher than the NER model combined with the BERT model. After applying in-domain training, we increased the overall F1 score of the WoBERT_{BiLSTM}-CRF model by 0.8% (strict indicator) and 0.9% (relaxed indicator). Adding the domain dictionary improved the overall F1 score of the WoBERT_{BiLSTM}-CRF model by 0.2% (strict indicator) and 0.5% (relaxed indicator). These results indicate that optimized pre-trained models, in-domain training technology, and the domain dictionary significantly improve the performance of traditional medical NER models (such as BiLSTM-CRF and BERT_{BC}).

Table 2 shows that the pseudo-labelling technique helps the WoBERT_{BiLSTM}-CRF model increase its strict F1 score by 0.5%. However, it decreases the relaxed F1 score by 0.1% compared to the W_{FPT}_{BC} model in Table 3 , indicating that pseudo-labelling does not necessarily improve medical NER model performance. One possible reason is that generated pseudo-labels may introduce noise into the original training data. For example, "前列腺" (prostate) is an "anatomic part" entity that should be labeled as such in the clause "复发侵及前列腺" (recurrent invasion of prostate). However, in "前列腺增生" (prostatic hyperplasia), "前列腺" combines with other characters to form a "disease & diagnosis" entity. The pseudo-labelling algorithm used in our paper incorrectly labeled "前列腺" (prostate) in this clause as an "anatomic part" entity.

Table 3 reveals that the W_{FPT}*{BC} model' s F1 scores are 0.4%, 0.9%, and 0.6% higher than those of the RWL model for imaging examination, laboratory examination, and surgery categories, respectively. This indicates that the W*{FPT}*{BC} model can recognize more short entities than the RWL model. By observing predictions from the three models ("WoBERT_{FurtherPreTraining}*{BiLSTM}-CRF" (W*{FPT}*{BC}), " RoBERTa-wwm-ext-large" (RWL), and " RoBERTa-wwm-ext-large*{BiLSTM}-CRF" (RWL*{BC})), we found that the RWL model performs very well on long entities, while W_{FPT}*{BC} excels at recognizing short entities. For example, in the clause "慢性肾衰, 尿毒症期" (chronic renal failure, uremic phase),

which should be recognized as a single "disease & diagnosis" entity, the RWL model precisely recognizes the entire entity, while the W{FPT}_{BC} model only recognizes "慢性肾衰" (chronic renal failure) as a "disease & diagnosis" entity.

The RWL_{BC} model demonstrates better overall recognition performance than the other two models. Table 2 shows that the RWL_{BC} model' s total strict F1 score is 0.9% higher than both the W_{FPT}_{BC} and RWL models. Based on this observation, we created a set of rules for the fusion model to merge predictions from the three models. For example, one rule states: "When the RWL and W_{FPT}*{BC} models identify an entity at a certain location and the entity length exceeds 6 characters, use the RWL model' s prediction; otherwise, use the W*{FPT}*{BC} model' s prediction." Another rule specifies: " If the RWL*{BC} model predicts a result at a certain position while the other two models do not, the RWL_{BC} model' s prediction is considered final for the NER process."

After integrating predictions from the three models ("W_{FPT}*{BC} + domain dictionary + pseudo labelling", "RWL*{BC} + domain dictionary + pseudo labelling" , and "RWL + domain dictionary + pseudo labelling"), Tables 2 and 3 show that all strict F1 scores increased across all medical categories except "anatomic part," while relaxed F1 scores decreased in many categories. This indicates that integrating predictions from the three models is effective for strict indicators but not for relaxed indicators. The reason may be that many "anatomic part" entities have nested structures, causing boundary errors in the fusion model. For example, "右上肺支气管" (right upper lobe bronchus) is an "anatomic part" entity, but "右上肺" (right upper lung) and "支气管" (bronchus) might be recognized as two separate entities after integration. Since the relaxed F1 indicator focuses on the number of recognized entities rather than entity boundaries, recognized entities like "右上肺" (right upper lung) and "支气管" (bronchus) are removed from predictions, resulting in decreased relaxed F1 scores.

Tables 2 and 3 also show that for the laboratory examination category, both strict and relaxed F1 scores are significantly lower than other categories. One possible reason is that some "laboratory examination"entities, such as "中性粒细胞百分比" (percentage of neutrophils) and "随机血糖" (random blood glucose), appear only in the test set but not in the training set. To address this issue, we need a larger medical domain dictionary containing more "laboratory examination" entities.

5. Conclusion and Future Work

This paper presents a medical NER approach that incorporates a medical domain dictionary and pre-trained language models for recognizing medical named

entities. Focusing on extracting knowledge from un-labelled medical texts, we performed the following tasks:

Medical dictionary construction. We built a medical domain entity dictionary and integrated it into pre-trained language models, significantly improving the recognition rate (measured by F1 score) of named entities.

In-domain training. We fed un-labelled medical texts to pre-trained language models for in-domain training. Experimental results show that the proposed NER approach with in-domain training improves the performance of named entity recognition in the medical domain.

Pseudo-labelling. We used a pseudo-labelling method to augment the CCKS-2020-CEMR dataset by automatically annotating un-labelled texts and marking them as pseudo-labels. Experimental results demonstrate that pseudo-labelling can improve NER performance in terms of strict F1 scores.

Medical named entity recognition using un-labelled texts and medical records is a challenging task. This research establishes a method for using medical dictionaries, in-domain pre-trained models, and pseudo-labelling techniques to extract knowledge from un-labelled medical texts. Future work will include: 1) exploring other natural language processing (NLP) techniques that can produce more effective performance for named entity recognition on un-labelled medical data, and 2) developing methods to efficiently identify out-of-vocabulary entities.

Acknowledgements

This work is supported in part by the Guangdong Science and Technology grant (No. 2016A010101033) and the Hong Kong and Macao joint research and development grant with Wuyi University (No. 2019WGAH21).

Author Contributions

C.J. Wen (klaywen15@163.com) initiated the proposed medical NER approach and conducted the experiments. C.J. Wen, T. Chen (chentao1999@gmail.com), and X.D. Jia (xudong.jia@csun.edu) contributed to the final version of the manuscript. C.J. Wen and Jiang Zhu (pacer0112@gmail.com) contributed to data preparation. T. Chen and X.D. Jia provided project guidance.

Data Availability Statement

The datasets generated and/or analyzed during this study are not publicly available because they were produced by medical expert consultants of Yidu Cloud

based on their professional experience. Public release of the datasets requires consent from all expert consultants, but they are available from the corresponding author upon reasonable request.

References

[1] Lei, J., et al.: A comprehensive study of named entity recognition in Chinese clinical text. Journal of the American Medical Informatics Association 21(5), 808–814 (2014)

[2] Wu, G., et al.: An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition. IEEE Access 7, 113942–113949 (2019)

[3] Wang, S., Li, S., Chen, T.: Recognition of Chinese medicine named entity based on condition random field. Journal of Xiamen University (Natural Science) 48, 349–364 (2009)

[4] Wang, Y., Liu, Y., Yu, Z.: A preliminary work on symptom name recognition from free-text clinical records of traditional Chinese medicine using conditional random fields and reasonable features. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, pp. 223–230 (2012)

[5] Xu, Y., et al.: Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. Journal of the American Medical Informatics Association 21(e1), e84–e92 (2014)

[6] Wu, Y., et al.: Named entity recognition in Chinese clinical text using deep neural network. Studies in Health Technology and Informatics 216, 624–628 (2015)

[7] Yang, J., et al.: Chinese electronic medical record named entity and entity relationship corpus construction. Journal of Software 27(11), 2725–2746 (2016)

[8] Yang, H., et al.: Named entity recognition based on bidirectional long short-term memory combined with case report form. Chinese Journal of Tissue Engineering Research 22(20), 3237–3242 (2018)

[9] Chowdhury, S., et al.: A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. BMC Bioinformatics 19, 449 (2018)

[10] Wan, L., Luo, Y., Zhi, L.: The recognition of naming entity of Bi-LSTM Chinese electronic medical records based on the joint training of Chinese characters and words. China Digital Medicine 14(2), 54–56 (2019)

[11] Wu, Y., et al.: Google' s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)

[12] Lee, D.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Proceedings of ICML 2013 Workshop: Challenges in Representation Learning (WREPL), pp. 1–6 (2013)

[13] Liu, Y., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[14] Cui, Y., et al.: Pre-training with whole word masking for Chinese BERT. arXiv preprint arXiv:2004.13922 (2020)

[15] Gururangan, S., et al.: Don' t stop pretraining: Adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8342–8360 (2020)

Author Biography

Chaojie Wen received his B.S. degree in Computer Science and Technology from Wuyi University in 2019. He is currently pursuing his M.S. degree in Electronic and Communication Engineering at Wuyi University, Guangdong, China. His research interests include natural language processing, named entity recognition, and relation extraction. ORCID: 0000-0002-9325-9147

Tao Chen received his B.Eng. degree in Communication Engineering in 2003 from Nanjing Institute of Communication Engineering, China, his M.Eng. degree in Computer Application in 2013 from Wuyi University, Guangdong, China, and his Ph.D. degree in Computer Application from Harbin Institute of Technology, China in 2018. He joined Wuyi University, China, as a lecturer in 2018. He is the author of one book, 17 articles, and 3 patents. His research interests include natural language processing, deep learning, knowledge acquisition and reasoning, and sentiment analysis. ORCID: 0000-0002-3634-0854

Xudong Jia is a Visiting Scholar at Wuyi University, China, and a Professor and Associate Dean of the College of Engineering and Computer Science at California State University, Northridge. He received his B.S. in 1983 and M.S. in 1986 from Beijing Jiaotong University, his M.S. in 1992 from the University of Toronto, Canada, and his Ph.D. in 1996 from the Georgia Institute of Technology. His research interests include intelligent transportation systems (ITS) standards, geographic information system (GIS) applications in transportation, traffic safety, transportation information systems, travel demand management, and air quality. He is an associate editor of IEEE Intelligent Transportation Systems Society and IEEE Open Journal of Intelligent Transportation Systems.

Jiang Zhu received his B.S. degree in Electronic Information Science and Technology from Inner Mongolia University for Nationalities in 2018. He is currently pursuing his M.S. degree in Pattern Recognition and Intelligent System at Wuyi University, Guangdong, China. His research interests include natural language processing, named entity recognition, and data mining.

Footnotes:

https://www.biendata.xyz/competition/ccks_{2020}_{2_1}

Yidu-N7K dataset (dataset of CHIP 2019 clinical terminology standardization task): http://openkg.cn/dataset/yidu-n7k

WoBERT: https://github.com/ZhuiyiTechnology/WoBERT

Note: Figure translations are in progress. See original paper for figures.