Asia-Pacific Journal of Information Technology and Multimedia 13(1) 2024: 47 – 60
https://doi.org/10.17576/apjitm-2024-1301-04

Leveraging Transfer Learning and Label Optimization for Enhanced Traditional Chinese Medicine Ner Performance

Memanfaatkan Pembelajaran Pemindahan Dan Pengoptimuman Label Untuk Meningkat Prestasi Pengekstrakan Entiti Nama Perubatan Tradisional Cina

Saidah Saad, Huang Zikun*

Faculty of Information Science and Technology, University Kebangsaan Malaysia, Malaysia

**Corresponding author: saidah@ukm.edu.my*

Received 11 September 2023
Accepted 4 December 2023, Available online 1 June 2024

ABSTRACT

Named Entity Recognition (NER) is a crucial component in various domains, including medical and financial fields, as it helps identify text fragments belonging to predefined categories from unstructured text. Over time, NER algorithms have evolved from dictionary-based approaches to machine learning and deep learning techniques. Transfer learning, a novel deep learning method, has shown impressive results in NER tasks. However, transfer learning models still face challenges, such as limited entity labels and the impact of noisy datasets. To address these challenges, this research aims to optimise the application of deep learning models for NER and achieve enhanced results. The research initially applied the BERT+CRF model to the WanChuang dataset, resulting in an F1-measure of 89.1%. This established the feasibility of using transfer learning models for NER on Chinese medical data and served as a baseline for comparison in the project. To address label-related issues in the baseline model, a scheme was proposed to improve the learning rate of the CRF layer, resulting in an increased F1 measure of 91.0%. Additionally, to mitigate the impact of noisy training data, a 10-fold retraining scheme was introduced to optimise the training set. By retraining the model using the optimised training set, an optimal F1 measure of 92.7% was achieved. The experiments demonstrated that the transfer learning model enhances NER entity extraction capabilities while the optimised CRF layer effectively captures the internal relationships of entity tags, thus improving overall performance. This research contributes to advancing NER techniques and their application in various domains.

Keywords**:** Named Entity Recognition, Traditional Chinese medicine, transfer learning, BERT, CRF

ABSTRAK

Abstrak: Pengekstrakan Entiti Nama (NER) ialah komponen penting dalam pelbagai domain, termasuk bidang perubatan dan kewangan, kerana ia membantu mengenal pasti serpihan teks yang tergolong dalam kategori yang dipratentukan daripada teks tidak berstruktur. Dari masa ke masa, algoritma PEN telah berkembang daripada pendekatan berasaskan kamus kepada pembelajaran mesin dan teknik pembelajaran mendalam. Pembelajaran pemindahan, kaedah pembelajaran mendalam yang baru, telah menunjukkan hasil yang mengagumkan dalam tugasan PEN. Walau bagaimanapun, model pembelajaran pemindahan masih menghadapi cabaran, seperti label entiti terhad dan kesan set data yang hingar. Untuk menangani cabaran ini, penyelidikan ini bertujuan untuk mengoptimumkan aplikasi model pembelajaran mendalam untuk PEN dan mencapai hasil yang dipertingkatkan. Penyelidikan ini pada mulanya menggunakan model BERT+CRF pada dataset WanChuang, menghasilkan ukuran F1 sebanyak 89.1%. Ini membolehkan ianya diguna sebagai model pembelajaran pemindahan untuk NER untuk data perubatan Cina dan berfungsi sebagai garis dasar untuk perbandingan. Untuk menangani isu berkaitan label dalam model garis dasar, satu skema telah dicadangkan untuk meningkatkan kadar pembelajaran lapisan CRF, menghasilkan peningkatan ukuran F1 sebanyak 91.0%. Selain itu, untuk mengurangkan kesan data latihan yang hingar, skim latihan semula 10 kali ganda telah diperkenalkan untuk mengoptimumkan set latihan. Dengan melatih semula model menggunakan set latihan yang dioptimumkan, ukuran F1 optimum sebanyak 92.7% telah dicapai. Eksperimen menunjukkan bahawa model pembelajaran pemindahan meningkatkan keupayaan pengekstrakan entiti NER manakala lapisan CRF yang dioptimumkan secara berkesan dapat menjejak perhubungan dalaman tag entiti, sekali gus meningkatkan prestasi keseluruhan. Penyelidikan ini menyumbang kepada memajukan teknik NER dan aplikasinya dalam pelbagai domain.

Kata kunci: Pengekstrak Entiti Nama, Perubatan Tradisional Cina, pembelajaran pemindahan, BERT, CRF

INTRODUCTION

Named Entity Recognition (NER) holds significant importance within Natural Language Processing. It handles structured and unstructured data by categorising named entities into predefined classes. Identifying named entities in Traditional Chinese Medicine (TCM) literature is crucial for extracting knowledge from extensive unstructured texts in this field. This process entails extracting instances of concepts and determining their respective types. In Traditional Chinese Medicine, numerous ancient books and medical records contain numerous clinical terms related to TCM. These terms encompass valuable information such as medicines, symptoms, and diseases. This information proves instrumental in developing a TCM expert system, knowledge map, and question-answering system (Yu et al., 2017).

Before the advent of deep learning methods, traditional NER methods were deficient for general domain and TCM texts. NER methods based on traditional dictionary methods are challenging to scale and optimise. Because the domain characteristics of TCM texts are not considered, the generalisation ability of the NER method for TCM texts in general domain texts is weak, and the recognition results are not satisfactory. The NER algorithm based on machine learning requires manual definition of templates and adjustment of parameters. Compared with other methods, the NER model based on deep learning has achieved better results in TCM texts.

Learning models based on deep migration have successfully succeeded in Natural language processing (Luo et al., 2020). These models can use text data to embed words in context and achieve better accuracy in multilingual understanding tasks. Among them, BERT uses masking language modelling to pre-train the bidirectional encoder in a large-scale universal domain corpus for sentence prediction, and the expression effect is the most significant. This model provides a pre-trained model for the Chinese, which can be better transferred to the field of

TCM. Therefore, this study believes that applying BERT to NER tasks in the field of TCM can achieve better results, and its performance is evaluated by training on the dataset.

This study proposes utilising transfer learning algorithms as a crucial and more efficient approach for Named Entity Recognition (NER) tasks. Transfer learning involves leveraging existing knowledge to acquire new knowledge, aiming to identify similarities between established and novel information for effective knowledge transfer. In the context of Traditional Chinese Medicine (TCM), annotated data is dispersed, and the scarcity of domain-specific terms in the medical field poses challenges for data annotation, often leading to noise. Transfer learning is employed to address the substantial amount of training data required for in-depth learning. This involves pre-learning knowledge from publicly available text data and transferring it to traditional Chinese medicine. Researchers have successfully applied this method to NER tasks in TCM, employing transfer learning models like BERT, resulting in commendable outcomes. (Zhang et al., 2022).

This paper consists of five sections. The first section discusses the background of this study, including the problems in TCM NER. The second section summarises historical research and NER methods. The third section elaborates on the methods used in the study. The fourth section introduces the results of the work and discussion. Finally, the fifth section summarises the research results and suggests future work.

LITERATURE REVIEW

A. TRADITIONAL CHINESE MEDICINE (TCM)

Traditional Chinese Medicine (TCM) generally refers to a kind of traditional medicine created by the Han nationality in China. The study of TCM encompasses human physiology, pathology, diagnosis, treatment, prevention, and management of diseases. TCM originated in primitive societies and developed its theoretical framework during the Spring, Autumn and Warring States periods. In addition, TCM has significantly influenced other countries with Chinese character civilisations, such as Japanese Chinese medicine (Cheng, 2014), Korean medicine (Ju-Ah et al., 2017), Vietnamese medicine (Bui, 2019) and so on. TCM has a long history, has a complete theoretical system and unique treatment, and has accumulated much clinical experience. There is a vibrant reserve of TCM prescriptions and literature on treating diseases.

The advancement of information technology infrastructure has propelled post-structured exploration into information related to Traditional Chinese Medicine (TCM), emerging as a predominant trend. This involves documenting traditional Chinese medicine prescriptions' ingredients, dosage, and associated treatment symptoms. By leveraging this data, one can distil and present the diagnostic and treatment experiences and compatibility plans within TCM, serving as a valuable reference for future learning and application.

Extracting medical entity information from texts, utilising electronic medical records and textual information on Chinese herbal medicine as corpora, represents a pivotal task in the contemporary medical landscape. This process, facilitated by natural language processing technologies, is underscored as essential in the medical domain (Wang et al., 2020). NER work is an essential and crucial foundational step that can serve as a tool for traditional Chinese medicine artificial intelligence to assist clinical decision-making, with enormous theoretical and applied research value. In the field of modern TCM, we mainly face some problems.

Although there are certain norms in medical terminology, as a natural language, it is still a free text expression. Doctors often use different expressions of traditional Chinese medicine terminology when expressing the same meaning (Lei et al., 2014). Due to the long history of TCM development, the descriptions and terminology of the same drug may differ in different historical documents, and drugs with similar names may also be quite different (Sarkar et al., 2023). This makes the dissemination and extraction of information difficult. It is necessary to propose a scheme that can uniformly extract the diseases in the literature.

In summary, proposing a work that does not rely on manual recognition and can complete NER tasks in TCM is meaningful.

B. NER FOR TRADITIONAL CHINESE MEDICINE (TCM)

In Traditional Chinese Medicine (TCM), Named Entity Recognition (NER) involves applying research focused on identifying specific entities within TCM text. These entities hold distinct significance, such as classifying "cold" as an illness, "cough" as a symptom, and "ginseng" as a drug. TCM entities serve as fundamental elements in TCM text, and delving into their internal relationships constitutes essential work in TCM research.

Traditionally, TCM research relied on manual extraction methods from unstructured texts like TCM literature and clinical diagnosis records. However, with the rapid progress of Natural Language Processing (NLP) technology in recent years, there has been a growing integration of NER and TCM. The adoption of NER technology, replacing traditional manual approaches for extracting ingredients in traditional Chinese medicine, has significantly reduced time and manpower requirements. This shift enhances the efficiency of research on Chinese medicine ingredients.

The NER task in traditional Chinese medicine can be seen as a subdomain NER task. The traditional Chinese medicine NER task is mainly based on a Chinese text corpus—public traditional Chinese medicine societies, such as CMeEE. However, because most medical NER tasks need to focus on more subdivided fields in their own medicine, such as diabetes, more medical NER literature uses their own data sets.

C. NER METHOD FOR CHINESE LANGUAGE

The existing NER methods can be divided into rule-based, statistical machine-learning, and deep-learning methods.

1. Rules Based Method

Previous studies mainly used rules and dictionaries to identify named entities, which were constructed manually. Select keywords as features and use pattern matching to extract text corresponding entities. This approach is limited by the need for manual rules to differentiate entity types and its lack of portability. This method's effectiveness depends on the dictionary's capacity and pattern rules (Krstev et al., 2014). Furthermore, an extensive dictionary can be challenging, and updating it requires time and effort. Although this method appears straightforward, it has limitations. Rule-based systems mainly rely on manually crafted semantic and grammatical rules to identify entities, which is limited by dictionary capacity and results in high accuracy but low recall (Chen et al., 2020). Analysing the lexical features and collocation habits of named entities, construct artificial recognition rules for named entities. In this process, there is a need for continuous improvement and completion of rules. However, each rule should be written before actual use to solve the problem of multiple named entities using context.

Utilising a dictionary for entity recognition offers the advantage of minimal errors, as the entire vocabulary is pre-included in the dictionary. However, this approach cannot make predictions beyond the dictionary's scope, as the algorithm cannot recognise content outside of it. In contrast, grammatical rule-based algorithms can extend recognition beyond the dictionary; however, in Traditional Chinese Medicine (TCM), prescription compositions do not consistently adhere to the grammatical rules of human English or Chinese.

2. Machine Learning-Based Method

NER tasks utilising machine learning can be categorised into supervised and unsupervised learning. In supervised learning, NER can take the form of multiclass and sequential tagging tasks. Carefully constructed features, represent each sample in tagged data and machine learning methods are employed to model the labelled data. The resulting trained model is then utilised to recognise unknown data rapidly (Bin et al., 2016). Commonly employed machine learning methods for supervised NER tasks include Hidden Markov Models (HMM), Support Vector Machines (SVM), and Conditional Random Fields (CRF).

On the other hand, unsupervised NER methods rely on clustering techniques or similarity judgments between entities and seed terms. Entity recognition is accomplished through statistical analysis using lexical features on large-scale unlabeled corpora. This process yields different text clusters based on similarity, representing distinct entity groups. Frequently utilised features or auxiliary information in unsupervised NER include lexical resources, Term Frequency-Inverse Document Frequency (TF-IDF), and shallow semantic information (blocked NP-chunking). The widespread adoption of machine learning in addressing NER tasks is attributed to its robust mathematical foundation and strong interpretability.

3. Deep Learning Based Method

The advancement of computer hardware and word embedding technology has empowered neural networks to address numerous challenges in natural language processing effectively. This approach applies to Named Entity Recognition (NER): discrete word representations are mapped to low-dimensional space, yielding dense word embeddings. Subsequently, these embeddings and word order information are input into a recursive neural network, allowing the network to extract features and predict the maximum value.

Compared to linear Hidden Markov Models (HMM) and Conditional Random Fields (CRF), deep learning algorithms demonstrate superior capabilities in feature extraction from original data. The deep learning solution for NER tasks is structured into three key steps: distributed text representation, context encoding architecture, and tag decoder.

D. OPTIMIZATION

The optimisation scheme is mainly divided into two aspects: data optimisation and model optimisation.

Data optimisation in the context of NER refers to improving the quality and effectiveness of training data used to train NER models. It involves various techniques aimed at enhancing the data to enhance the model's performance and generalisation capabilities. Data optimisation techniques involve data cleaning, label consistency, data augmentation, accurate-consistent annotation, etc. Whether it is supervised learning or unsupervised learning, data is always the most essential driving force. More data types can bring better stability and predictability to good models of unknown data. For the model, the data encountered is more likely to be recognised than the previously unseen data. However, adding data is not blind.

Model optimisation is mainly aimed at the modification of the model itself or the adjustment of hyperparameters. Training a neural network model involves utilising optimisation algorithms to solve the parameter optimisation problem by minimising the cost function. Training a neural network model may require several hundred or thousands of machines to train simultaneously for several months. Using an optimisation algorithm can save training time to accelerate the model convergence. Standard optimisation algorithms include the gradient descent algorithm, RMS prop algorithm (Bekoulis et al., 2018) and Adam algorithm (Roth & Hongxia, 2016). Random gradient descent can help us converge to the global optimal value and eliminate local minimum and saddle points. However, this is only a theoretical case. No algorithm can guarantee such a large number of parameters to find the optimal solution.

RESEARCH MODEL AND RESEARCH QUESTIONS

The NER task is still an open research field in Chinese NLP because it brings many challenges, including the migration and scalability of NER algorithms and the problem of noisy data annotation in domain-specific terms in the medical field.

This project employs the BERT+CRF transfer learning model as a baseline and applies it to the NER task. The learning rate of the CRF layers is independently adjusted to achieve improved results within a limited number of training epochs. Additionally, to address the data noise problem, a 10-fold training method is proposed to optimise the training set and enhance the model's recognition effectiveness.

The rationale behind selecting the BERT+CRF model as the baseline for this study is grounded in its demonstrated state-of-the-art performance in natural language processing tasks, including Named Entity Recognition (NER), due to BERT's ability to capture contextual information bidirectionally. Integrating CRF with BERT is particularly beneficial for sequential labelling tasks like NER, as the CRF layer models’ dependencies between adjacent labels, which is crucial for accurate entity recognition in the complex and context-rich medical domain. BERT's pre-trained contextual embeddings are well-suited for medical texts, which often contain specialized terminology and complex syntactic structures. Previous successes of BERT+CRF in related domains further support its efficacy. Additionally, fine-tuning BERT on domain-specific data enables adaptation to the nuances of traditional Chinese medicine datasets, enhancing model performance and generalizability. This rationale underpins the choice of BERT+CRF as the baseline model and underscores its suitability for addressing the challenges inherent in NER tasks within the medical domain.

The research scope of this project is the application and improvement of a transfer learning model based on Chinese traditional Chinese medicine dataset. This study uses the deep learning model BERT+CRF for NER tasks based on the WanChuang dataset. As far as the improvement content of this project is concerned, these include increasing the Learning rate of the CRF layer and 10-fold retraining modes. The following are the questions that will be investigated in this project:

1. How to improve the learning ability of the CRF layer in the BERT+CRF model structure to improve the restriction ability of the entity relationship?
2. How to reduce the impact of dataset noise without expert knowledge?

The main purpose of this study is to practice the feasibility of the BERT+CRF model on the NER task and improve it. The objectives of the research are as follows:

1. Implementing the BERT+CRF model to perform NER tasks on the Chinese traditional Chinese medicine dataset to verify its feasibility and serve as a baseline model.
2. To propose optimisations for the learning rate of the CRF layer in the baseline model and compare the experimental results after retraining.
3. To propose the 10-fold optimisation scheme to optimise the noise of the training set and compare it with the experimental results after retraining.

METHODOLOGY

The flowchart shown in Figure 1 is the method used in this study. The NER task of this study is based on the BERT+CRF model and is divided into 7 steps for model training. This includes dataset preprocessing, improved CRF learning rate, 10-fold label optimisation, content embedding, training of the BERT and CRF models, and final prediction output and evaluation. This study modifies the parameters based on the prediction obtained from the baseline model BERT+CRF model structure and retrains them to obtain better results.

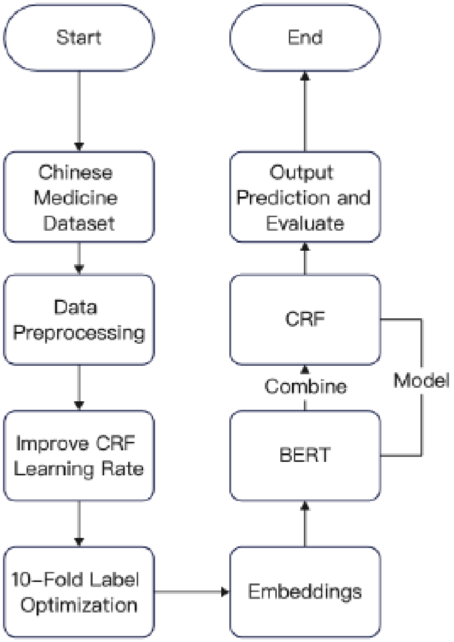

Start
Chinese Medicine Dataset
Data Preprocessing
Improve CRF Learning Rate
10–Fold Label Optimization
Embeddings
BERT
Combine
CRF
Model
Output Prediction and Evaluate
End

FIGURE 1. Overview of Methodology

The first step includes data collection and preprocessing. Based on literature references and the collection of publicly available datasets, and based on the focus of this project, we have chosen the publicly available dataset Wanchuang (Tianchi, 2020) as our dataset for our work. A publicly available dataset is more conducive to comparing and evaluating our work with other studies. This step will preprocess the dataset based on its content, including data cleaning, segmentation, and evaluation.

The second stage evaluates based on the data processed in the previous stage. Based on the model super parameters and label errors and omissions in the dataset, the project first improved the learning rate of the CRF layer and then used a 10-fold method to supplement and delete the labels in the training set to achieve better NER task performance.

The third step is to embed the dataset through a converter, including word embedding, statement embedding, and positional embedding. This step aims to transform the natural language dataset into the input part of the model through embedding for BERT model training.

The fourth step involves training the model, which includes fine-tuning the pre-trained BERT model and training the CRF model. Model training was conducted using the Colab Notebook, an online work platform provided by Google that offers GPU computing power. The "Bert-base-Chinese" branch was chosen for this Chinese dataset, leveraging the pre-trained BERT model to achieve improved results. The BERT model captures label relationships in the dataset, while the CRF layer ensures predicted values adhere to logical constraints.

Following model training, a validation set is utilised to assess the model's effectiveness and compare it with the baseline BERT+CRF model. This project primarily examines the influence of different annotated datasets on experimental outcomes. Finally, all experimental records and results are compared, culminating in a comprehensive research summary.

RESULTS AND DISCUSSION

All model training was conducted using Google Collab laptops, leveraging the advantages of powerful GPU acceleration. The training process for 30 epochs is only completed within 1 hour. Initially, the baseline model was trained for 30 epochs with a Learning rate of 1e-6. F1-Measure was evaluated in each round of training, including on the current and test sets, as performance indicators. It can be seen from Figure 2 that the F1-Measure on the test set has reached stability over 20 epochs. The loss curve in Figure 3 indicates that the loss value has not continued to decrease. The experimental results of the baseline model indicate that training for more than 20 epochs is unnecessary. The baseline model achieves the highest F1-Measure of 89.1% at training epoch 15 out of 30 training epochs and remains stable after that.

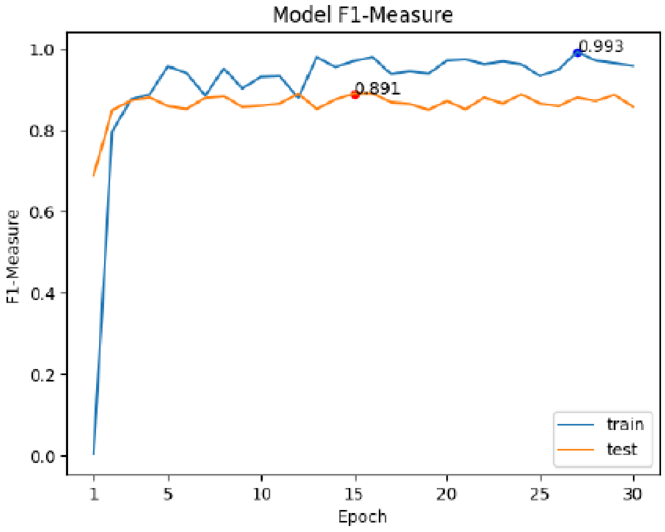


FIGURE 2. Baseline Model F1-Measure

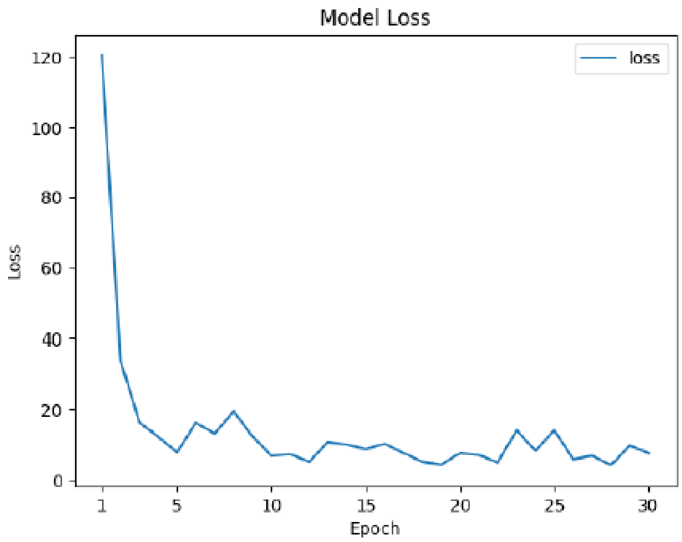


FIGURE 3. Baseline Model Loss

In the methodology analysis, the CRF layer of the baseline model exhibits insufficient learning of a robust transition matrix for evaluating the constraint relationship of BIO. Consequently, the decision is made to augment the learning rate of CRF. In the experiment, the CRF layer's learning rate is expanded to 100 times and 1000 times that of the BERT layer. Figure 4 illustrates a stable 1.9% increase in the F1-Measure of the model on the test set. Notably, it is observed that a learning rate of 1000 times is not significantly different from a learning rate of 100 times. Manual observation of the prediction results on the test set reveals a reduction in annotations that fail to meet lexical continuity, such as instances where Begin-Drug should not be linked to Symptoms.

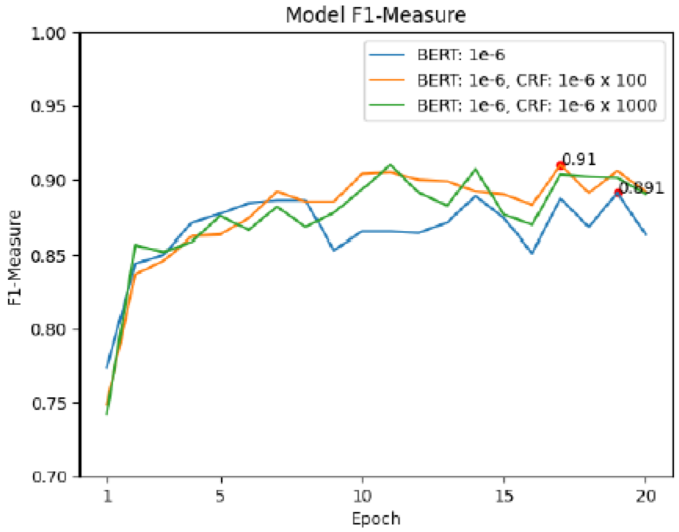


FIGURE 4. F1-Measure Comparison of Different Learning Rates

The project employs a 10-fold label supplement solution for addressing missing and incorrect labels. The 6,000-piece training set is partitioned into ten segments, with each iteration involving nine pieces for training, totalling 5,400 documents. Subsequently, the model is trained on all ten segments after optimising the learning rate. Utilizing these ten models to predict the original training set results in ten predicted training sets.

In cases where a label in the original training set is absent in any prediction results, we eliminate that label. Conversely, if all predictions are labelled with a valid label but do not appear in the training set, we incorporate this label into the training set. This approach optimises the training set by addressing labelling issues stemming from a lack of medical knowledge.

Two updated datasets are maintained—one with solely new annotations added, and another with annotations both added and removed. Retraining the model using these datasets yields stable improvements in F1-Measure compared to the baseline model. Notably, the modified labelled dataset achieves the highest F1-Measure at 92.7%, a 3.6% increase over the baseline model. The dataset incorporating deletion and addition operations outperforms the dataset solely adding annotations. This suggests that blindly adding labels without addressing labelling errors is not conducive to better results. Removing erroneous data, including inconsistencies in the original annotations, can enhance recognition efficiency. Figure 5 illustrates a line graph depicting different training sets and models in the F1-Measure evaluation.

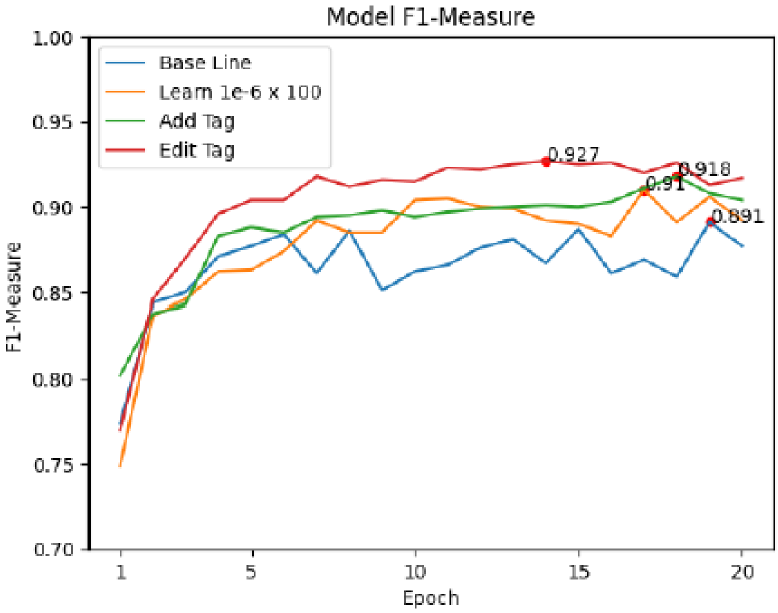


FIGURE 5. F1-Measure Comparison of Different Datasets

At the same time, by observing the loss curve of the model through Figure 6, it can be found that the modified data set can obtain faster and more stable loss indicators. At the same time, it can be reduced to a lower level in the later stage of training, Achieving a loss value of 0.116. It means that the modified data set can obtain better fitting ability, which aligns with our expectation of reducing noise by processing the training set. In addition, by observing the accuracy rate and recall rate curves, it can be found that improving the learning rate of the CRF layer can slightly improve the accuracy rate, which is consistent with the role of the CRF layer in restricting label relationships in the model. The model trained using a 10-fold optimised training set can better improve the recall rate.

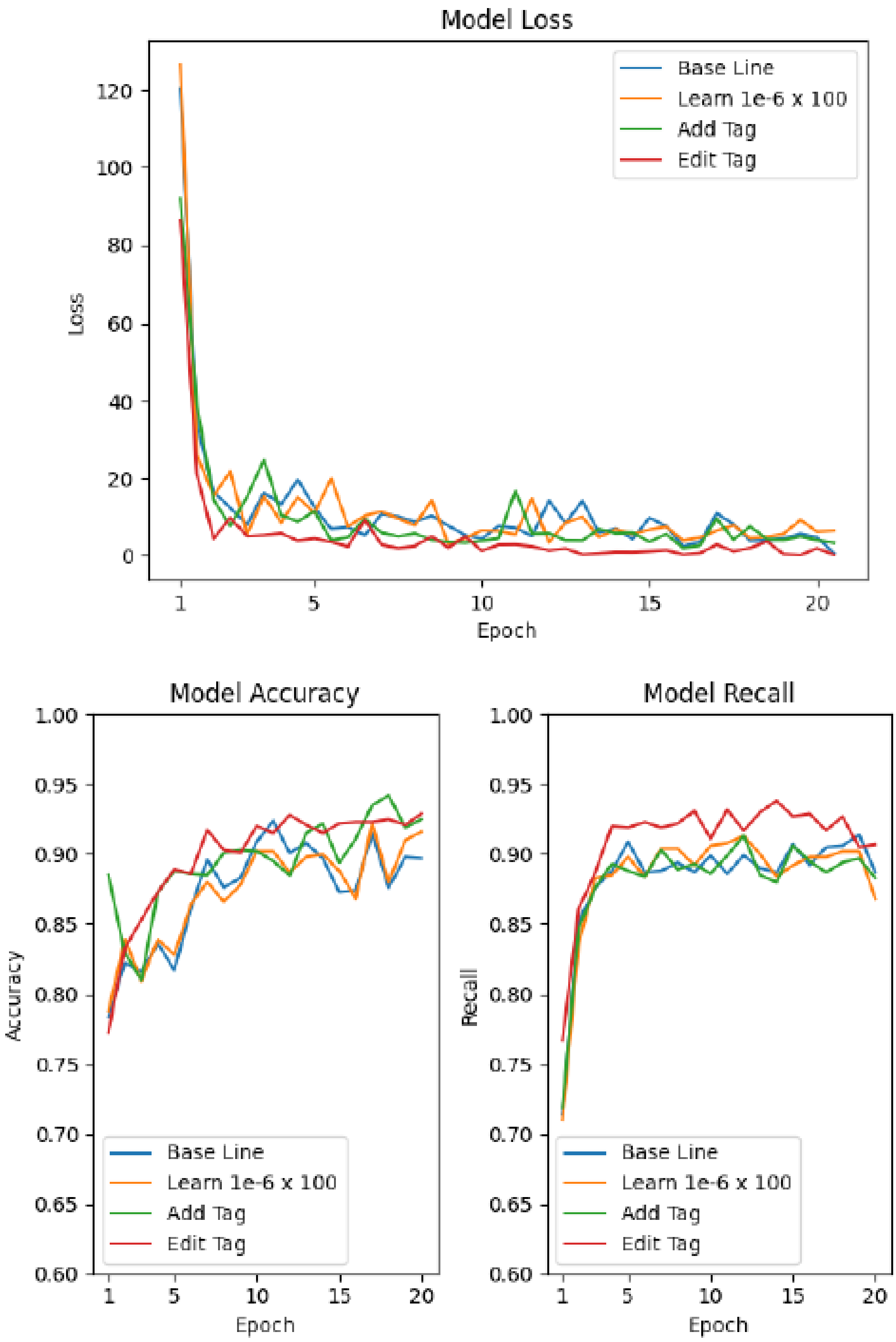
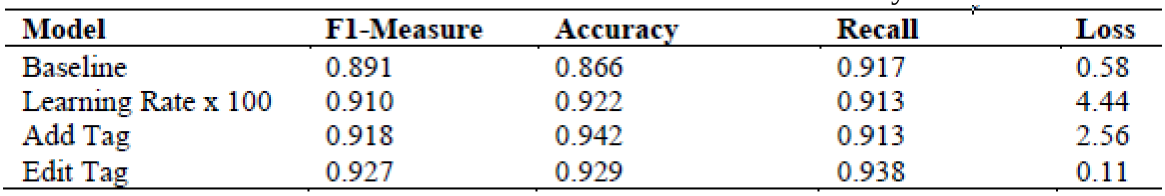


FIGURE 6. Loss Comparison of Different Datasets

By observing the accuracy rate and recall rate curves, it can be found that improving the learning rate of the CRF layer can slightly improve the accuracy rate, which is consistent with the role of the CRF layer in restricting label relationships in the model. The model trained using a 10-fold optimised training set can better improve the recall rate.

To summarise our experiments, we first used 30 epochs of training in the baseline model to confirm that the model fitting ability did not improve after 20 epochs, so we used 20 epochs as a hyperparameter in subsequent experiments. The highest F1 measure in the Add Tag training set appeared at the 18th epoch, and the highest F1 measure in the Edit Tag training set appeared at the 20th Epoch. This is due to the randomness included in the deep learning algorithm. It can be confirmed that the subsequent F1-Measure has achieved stability.

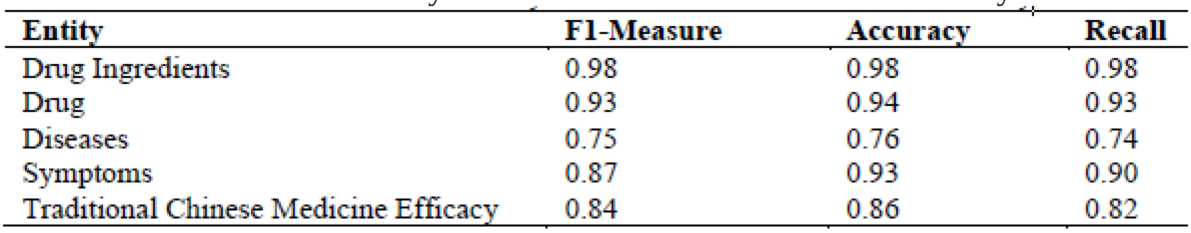
TABLE 1. Best Performance Evaluation Summary



Model	F1-Measure	Accuracy	Recall	Loss
Baseline	0.891	0.866	0.917	0.58
Learning Rate x 100	0.910	0.922	0.913	4.44
Add Tag	0.918	0.942	0.913	2.56
Edit Tag	0.927	0.929	0.938	0.11

Accurate to the entity, the F1-Measure of the drug ingredient is the highest, reaching 98%. The efficacy of traditional Chinese medicine, diseases, symptoms and medicines reached 84%, 75%, 90% and 93% respectively. Among them, the recognition rate of diseases is slightly lower than that of other entities, and one of the reasons is that there are fewer disease labels in the original training set. Evaluation metrics for each entity are presented in Table 2.

TABLE 2. Summary of Performance Evaluations for Each Entity



Entity	F1-Measure	Accuracy	Recall
Drug Ingredients	0.98	0.98	0.98
Drug	0.93	0.94	0.93
Diseases	0.75	0.76	0.74
Symptoms	0.87	0.93	0.90
Traditional Chinese Medicine Efficacy	0.84	0.86	0.82

CONCLUSION

NER is an important basic tool in Knowledge graphs and other application fields. It has played a crucial role in promoting the development of Natural language processing technology towards practicality. With the popularisation of deep learning today, the effect of NER has been improved more universally than the previous dictionary-based method. However, the improvement of NER still faces challenges in the Chinese natural language, such as the problems caused by word segmentation and the improvement of the effect.

In order to improve the effectiveness of the Chinese dataset, this project uses the WanChuang dataset for model training and research. This is a NER dataset in the field of traditional Chinese medicine. In order to study more general methods, we limit the entity categories to 5: drugs, drugs, traditional Chinese medicine efficacy, symptoms, and diseases. This project proposes to use a pre-trained model BERT+CRF as the baseline model, and two methods were proposed to achieve the research goal of improving NER extraction efficiency.

The first approach proposed in this study is to increase the learning rate of the CRF layer. We find that the learning rate of the pre-trained model BERT cannot learn enough constrained content for the CRF layer in a limited epoch. Since the CRF layer plays a role in limiting the rationality of the prediction results output by the BERT layer in the model structure we use, the entity category should start with the BEGIN label. The continuous entity categories should be the same; by increasing the learning rate of the CRF layer to 100 times that of the BERT layer, we obtained about 1.9% improvement in experiments. The highest F1 measure reached 91%. It can be seen that the unreasonable label relationship has been significantly reduced by predicting the output of the result.

The second approach proposed in this study is to use a 10-fold training to optimise the training set. Through the analysis of the training set, we found that the training set contains more noise, including wrong and missing labels. Correcting these annotations manually requires substantial expertise and unpredictable time and labour costs. In order to optimise the noise problem of the training set, we propose a 10-fold model optimisation scheme. By dividing the training set into 10 parts and taking 9 parts each time to train a baseline model, we will obtain 10 completed models. Use 10 models to predict the training set and obtain 10 predicted training sets. When an entity annotation appears in the original training set but does not appear in any predicted training set, we delete the entity annotation. When an entity annotation does not appear in the original training set but appears in all predicted training sets, we will supplement the annotation of that entity. Using the modified training set to retrain the model, this experiment achieved the highest F1-measure of 92.7%, about 3.6% higher than the baseline model, and an improvement of about 1.6% compared to the model that improved the abovementioned learning rate. At the same time, by observing the loss curve of the model, it can be found that the modified data set can obtain faster and more stable loss indicators. At the same time, it can be reduced to a lower level in the later stage of training, Achieving a loss value of 0.116. It means that the modified data set can obtain better fitting ability, which aligns with our expectation of reducing noise by processing the training set.

In terms of future work related to this research, three parts can be applied to expand the work of this project. The first part can effectively increase the exposure of deep learning models by integrating datasets from other Chinese medical fields. The second part optimises the model's ability in Chinese word segmentation by adding vocabulary in the medical field. The third part is that some drugs in the medical field use their main components as their drug names. How to solve the problem of nested labelling is also the main focus of future work.

REFERENCE

Bekoulis, Giannis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. "Joint entity recognition and relation extraction as a multi-head selection problem”.

Bin, Liu, Wu Zhiqiang, Wang Jianhong, Deng Lizong, Peng Yousong, and Jiang Taijiao. 2016. "Extracting Clinical entities and their assertions from Chinese Electronic Medical Records Based on Machine Learning." Atlantis Press.

Bui, Anita. 2019. "Traditional vietnamese medicine between chinese heritage and national tradition." Chinese Medicine & Culture 2 (1):21-5.

Chen, Xianglong, Chunping Ouyang, Yongbin Liu, and Yi Bu. 2020. "Improving the Named Entity Recognition of Chinese Electronic Medical Records by Combining Domain Dictionary and Rules." International journal of environmental research and public health 17 (8).

Cheng, Hu. 2014. "The Modernization of Japanese and Chinese Medicine (1914-1931)." Chinese Studies in History 47 (4):78-94.

Ju-Ah, Lee, Choi Tae-Young, Lee Myeong Soo, Ko Mimi, Kang Byoung-Kab, Liu Huan, Jiang Jun-Jie, and Li Yuan-Yuan. 2017. "Protocol for Systematic Review of Controlled Trials of Korean and Chinese Herbal Treatments for Stroke." Journal of Acupuncture Research 34 (4):169-71.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. "A system for named entity recognition based on local grammars." Journal of Logic & Computation 24 (2):473-89.

Lei, Jianbo, Buzhou Tang, Xueqin Lu, Kaihua Gao, Min Jiang, and Hua Xu. 2014. "A comprehensive study of named entity recognition in Chinese clinical text." Journal of the American Medical Informatics Association : JAMIA 21 (5):808-14.

Luo, Ling, Zhihao Yang, Mingyu Cao, Lei Wang, Yin Zhang, and Hongfei Lin. 2020. "A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature." Journal of Biomedical Informatics 103.

Roth, Adam, and Zhang Hongxia. 2016. "A Dialogue on Traditional Medicine: East Meets West." China Media Research 12 (4):85-92.

Sarkar, Chayna, Biswadeep Das, Vikram Singh Rawat, Julie Birdie Wahlang, Arvind Nongpiur, Iadarilang Tiewsoh, Nari M. Lyngdoh, Debasmita Das, Manjunath Bidarolli, and Hannah Theresa Sony. 2023. "Artificial Intelligence and Machine Learning Technology Driven Modern Drug Discovery and Development." International Journal of Molecular Sciences 24 (3):2026.

Tianchi. 2020. " Entity Recognition of Traditional Chinese Medicine Instructions." Wang, Yu, Yining Sun, Zuchang Ma, Lisheng Gao, and Yang Xu. 2020. "Named Entity Recognition in Chinese Medical Literature Using Pretraining Models." Scientific Programming 2020:8812754.

Yu, T., J. Li, Q. Yu, Y. Tian, X. Shun, L. Xu, L. Zhu, and H. Gao. 2017. "Knowledge graph for TCM health preservation: Design, construction, and applications." Artificial Intelligence in Medicine 77:48-52-.

Zhang, Xuan, Lin Zhang, Jacky C. P. Chan, Xihong Wang, Chenchen Zhao, Ying Xu, Weifeng Xiong, Wai Chak Chung, Feng Liang, Xu Wang, Jiangxia Miao, and Zhaoxiang Bian. 2022. "Chinese herbal medicines in the treatment of ulcerative colitis: a review." Chinese Medicine 17 (1).