A Bootstrapped Chinese Biomedical Named Entity Recognition Model Incorporating Lexicons

Liangping Ding
dingliangping@mail.las.ac.cn
National Science Library, Chinese Academy of Sciences
Beijing, China
Department of Library Information and Archives Management, University of Chinese Academy of Science
Beijing, China

Zhixiong Zhang*
zhangzhx@mail.las.ac.cn
National Science Library, Chinese Academy of Sciences
Beijing, China
Department of Library Information and Archives Management, University of Chinese Academy of Science
Beijing, China

Huan Liu
liuhuan@mail.las.ac.cn
National Science Library, Chinese Academy of Sciences
Beijing, China
Department of Library Information and Archives Management, University of Chinese Academy of Science
Beijing, China

Abstract

Biomedical named entity recognition (BioNER) is a subtask of named entity recognition, aiming at recognizing named entities in medical text to boost the knowledge discovery. In this paper, we propose a bootstrapped model incorporating lexicons, which takes advantage of pretrained language model, semi-supervised learning and external lexicon features to apply BioNER to Chinese medical abstracts. Extensive evaluation shows that our system is competitive on limited annotated training data, which surpasses the baselines including HMM, CRF, BiLSTM, BiLSTM-CRF and BERT for 54.60%, 37.92%, 55.46%, 48.67%, 7.99% respectively. The experimental results demonstrate that unsupervised pretraining makes pretrained language model acquire the ability that only a few annotated data can achieve great performance for downstream tasks. In addition, semi-supervised learning and external lexicon features can further compensate for the problem of insufficient annotated data.

CCS Concepts: • **Information systems** → **Entity relationship models**.

Keywords: Biomedical Named Entity Recognition, Bootstrapping, Feature Incorporation, Semi-Supervised Learning, Pretrained Language Model

ACM Reference Format:
Liangping Ding, Zhixiong Zhang, and Huan Liu. 2021. A Bootstrapped Chinese Biomedical Named Entity Recognition Model Incorporating Lexicons. In *EEKE 2022, June 20-24, 2022, Germany and online.* ACM, New York, NY, USA, 9 pages.

1 Introduction

Named entity recognition (NER) plays an important role in natural language processing (NLP) that entails spotting mentions of conceptual entities in text and classifying them according to a given set of categories. NER not only acts as a standalone tool for information extraction, but also lays foundations for a quantities of NLP tasks including information retrieval [1] [2], knowledge graph construction [3], text summarization [4], question answering [5] etc. Developing a well-performing, robust NER system can facilitate more sophisticated queries that involve entity types in information retrieval and more complete extraction of information for knowledge graph population.

There are over 32 million publications in PubMed[1] and over 27 million references in Medline[2]. The large number of unstructured scientific medical abstracts limit the large-scale knowledge discovery and application of medical literature. It is urgent to explore the automatic biomedical named entity recognition (BioNER) methods to transform unstructured literature into structured data to provide valuable information for researchers. However, because of the academic and innovative feature of medical literature, there are a number of formal and emerging medical terms in scientific medical abstracts, increasing the difficulty of BioNER. Another reason why BioNER is challenging is the non-standard usage of abbreviations, synonymous and the frequent use of phrases describing entities [6]. All of these reasons make BioNER a tricky problem, nevertheless need to be settled.

In recent years, deep neural networks have achieved significant success in named entity recognition and many other natural language processing tasks. Most of these algorithms are trained end to end, and can automatically learn features from large-scale annotated datasets. However, these data-driven methods typically lack the capability of processing rare or unseen entities. And BioNER in Chinese texts is more difficult compared to those in Romance languages due to the lack of word boundaries and the complexity of Chinese composition forms [7]. Previous statistical methods and feature engineering practice have demonstrated that human knowledge can provide valuable information for handling rare and unseen cases [8]. Lexicon or gazetteer as additional features introduce some linguistic

*Corresponding Author

[1] https://pubmed.ncbi.nlm.nih.gov/about/
[2] https://www.nlm.nih.gov/medline/medline_overview.html

and domain resources to the model and they are beneficial to identify the entity boundaries and further improve the performance of the model [9][10].

A lack of annotated training data for named entity recognition of Chinese medical abstracts is of particular concern when using neural architectures, which generally require large amounts of training data to perform well. Pretrained language model BERT [11] is a more recent approach to biomedical text mining tasks and has achieved successful model performances. It is trained on millions of unsupervised text and allows the downstream task to achieve excellent performance even though a small amount of annotated training data is available.

In this paper, we decided to take advantage of both the pretrained language model and semi-supervised learning to cope with BioNER with limited data. In addition, we considered auxiliary resources are often important to better understand the text and extract entities. Our contributions can be listed as follows:

1. To remedy the situation that it's difficult to detect accurate entity boundary and process unseen entities for Chinese NER, we collected more than 700,000 medical terms as lexicons and embedded them into BERT to improve the model performance.
2. To address the problem that less publicly available annotated dataset for BioNER of Chinese medical abstracts, we proposed a bootstrapped BioNER framework, combining the benefits of semi-supervised learning, pretrained language model and lexicon features.
3. We designed Application programming interface (API) to deploy our proposed model for convenient usage, which can be visited at: http://sciengine.las.ac.cn/NER_MED_CN

2 Related Work

Named entity recognition task is a fundamental task in information extraction and has received constant research attention over the recent years. Traditional named entity recognition focuses on identifying people, location and organization, which cannot meets the demands of valuable structured information extraction of specific domain. There has been a surge of interest in extracting named entities such as genes/proteins, drugs, diseases, organs in biomedical domain, which is defined as BioNER. BioNER plays a significant role in medical information mining and establishment of high-quality knowledge graph. Modeling methods in BioNER are broadly divided into four categories: Rule-based, Unsupervised Learning, Feature-based Machine Learning and Deep Learning.

Rule-based BioNER systems rely on heuristics and handcrafted rules including domain-specific gazetteers [12] and syntactic-lexical patterns to extract entities [13][14][15]. This approach was the dominant approach in the early BioNER system. However, it requires labor-intensive and skill-dependent design and always leads to a relatively low accuracy because it's not feasible to list all rules and the dictionary cannot cover all entities.

As for unsupervised learning approach, it is usually not the first choice to develop a BioNER system even though this kind of approach has the advantage of modeling without annotated data. However, Zhang and Elhadad introduced a system, which used an unsupervised approach to BioNER with the concepts of seed knowledge and signature similarities between entities [16].

For feature-based machine learning, BioNER is usually cast as a sequence labeling task where the goal is to find the best label sequence given an input token sequence [17][18]. Many algorithms have been applied in supervised BioNER, such as hidden Markov Models (HMMs) , Conditional Random Fields (CRFs) and Support Vector Machines(SVMs). This approach has been widely used and proven to achieve good performance in many studies [19][20]. Feature engineering is critical in machine learning based BioNER systems, which is mostly concerned with an abstraction over the given text where each token is represented by one or many Boolean, numeric or nominal values [21][22]. The most commonly used rich text features in BioNER are linguistic features such as POS tagging, orthographic features such as capitalization, morphological features such as suffix and prefix, contextual features such as n-grams, and lexicon features such as gazetteers[23].

In order to reduce the dependence on complicated feature engineering, the focus of BioNER has shifted to deep learning approaches in recent years [24][25][26]. Deep learning approaches are beneficial for modeling the highly non-linear features, while they are potential to overfit on the condition of insufficient annotated data compared to traditional machine learning based approaches. Semi-supervised learning is often used to make up for this problem [27][28]. Munkhdalai et al. used semi-supervised learning by extending the existing BioNER system BANNER [29]. The emergence of pretrained language model also addressed this issue to some extent. In 2018, Google released BERT [11], short for Bidirectional Encoder Representations from Transformers, which provided a new NLP paradigm that pre-train a language model using large amounts of unannotated data and then fine-tune the model based on a few annotated data of downstream task.

Lots of NER systems utilize gazetteers as a form of external knowledge, which can provide additional domain specificity to named entities [8]. In the context of natural language understanding, a gazetteer is simply a collection of common entity names typically organized by their entity type. Chiu and Nichols believed that gazetteers are crucial to NER performance and proposed a new gazetteer encoding scheme to concatenate gazetteer feature to the word embedding [10]. Even though gazetteers play a significant role in

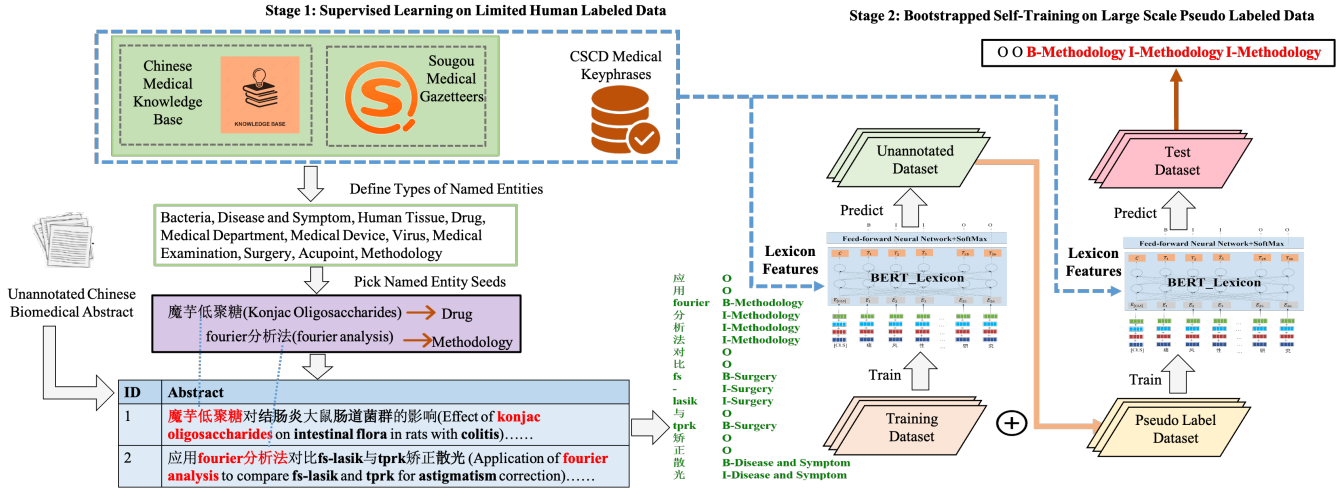

Figure 1. The Framework of Our Method

BioNER, it's hard to maintain large Chinese gazetteers with corresponding entity type in scientific domain. While for BioNER of Chinese medical abstracts, the collection of keyphrases can act as lexicons naturally and can be incorporated to the model to add in more information of entity boundary.

3 Methodology

In this paper, we regarded BioNER of scientific Chinese medical abstracts as a sequence labeling task and chose pretrained language model BERT as the backbone model. To address the problem that it's hard for the model to process unseen entities, we embedded medical lexicons into the feature vector space of BERT as domain features[30], which is a kind of human knowledge to instruct the training of model. Figure 1 shows the framework of our proposed method, which is a two-stage framework including supervised learning on a small amount of human labeled data and Bootstrapped self-training on large scale pseudo labeled data. The detailed description of our method is presented in the following sections.

3.1 Data Preprocessing

There are some publicly available datasets for Chinese BioNER, while they are mainly from Electronic Medical Records (EMRs), which have different linguistic features compared to scientific Chinese medical abstracts. For this reason, we decided to manually label a small amount of data. Before doing that, we defined the types of named entities according to Chinese medical knowledge base[3] and Sougou gazetteers[4], resulting in an eleven category scheme including Bacteria, Disease and Symptom, Human Tissue, Drug, Medical Department, Medical Device, Virus, Medical Examination, Surgery, Acupoint, Methodology.

To ensure more adequate samples of each entity type in the manually labeled dataset, we picked at least 10 named entity seeds for each category and then located these seeds in the abstract field of Chinese Science Citation Database (CSCD) to find the corresponding records to annotate. For example, for the category of drug, we set '魔芋低聚糖(Konjac oligosaccharides)', '益心舒胶囊(Yixinshu capsules)', '双歧杆菌四联活菌片(Bifidobacterium tetrakis tablets)', '护肠清毒微丸(Intestinal Protection and Cleansing Micro Pills)', '乐舒洗液(Lysol lotion)', '灵芝制剂(Ganoderma lucidum preparation)', '清肺解毒汤(Lung Clearing and Detoxifying Soup)', '阿米替林(Amitriptyline)', '碳青霉烯类(Carbapenems)', '重组人生长激素(Recombinant human growth hormone)' as seeds and located literature that contained corresponding seeds to annotate all the entities in the text. We concatenated the title and abstract metadata by period and transformed the original text to sequence labeling format.

Due to the lack of word boundaries in Chinese and the complexity of Chinese composition forms, the character-level formulation was used to avoid the segmentation errors of the Chinese tokenizer. And Chinese medical abstracts are more complicated with mixed Chinese characters, English words, numbers and punctuations, adding the difficulties of named entity recognition. To deal with this, we created a customized tokenizer which treats each Chinese character or English word as the basic element. After that, we manually annotated the dataset, tagging entity type for each token. In our annotating process, BIO (Beginning-Inside-Outside) tagging scheme was used as the reference, which *B-type* tags the first token of an entity, *I-type* the subsequent ones, and *O* tagging non-entity tokens.

[3] http://openkg.cn/
[4] https://pinyin.sogou.com

Table 1. An Example of the Annotated Dataset

Token	Lexicon	Entity Type
应	O	O
用	O	O
fourier	B	B-Methodology
分	I	I-Methodology
析	I	I-Methodology
法	I	I-Methodology
对	O	O
比	O	O
fs	O	B-Surgery
-	O	I-Surgery
lasik	O	I-Surgery
与	O	O
tprk	O	B-Surgery
矫	O	O
正	O	
散	O	B-Disease and Symptom
光	O	I-Disease and Symptom
的	O	O
准	O	O
确	O	O
性	O	O
。	O	O

Due to the complexity of the natural language and the specialty of medical scientific abstracts, some linguistic and domain resource features such as part-of-speech tagging feature, terminologies, dictionaries can be employed to improve the performance of the model. As a result, we explored the effect of lexicon feature, which introduced human knowledge to some extent. We built out lexicons based on keyphrases in CSCD, Chinese medical knowledge base and Sougou medical gazetteer. After removing duplicates and irrelevant words (e.g numbers, word length less than three), the number of medical terms reached to 704,507. Bi-direction maximum matching algorithm was used to match the text with the lexicon, capturing the longest possible match. Using BIO tagging scheme to generate lexicon feature was proven to be more effective than the traditional method proposed by Collobert et al.[31], which marked tokens with YES/NO. So in this paper, each token in the match was encoded by BIO tagging scheme and then a lookup table was used to generate the lexicon embedding. An example of the annotated dataset with lexicon feature is shown in Table 1.

3.2 Model Architecture

In this paper, the architecture of the BioNER model was a token-wise NER classifier on top of the pretrained BERT and lexicon features were embedded into the feature space to add in domain features[30]. We used feed-forward neural network as the classifier and the classifier took in the token-wise output embeddings from the pre-trained BERT layers and gave the prediction on the type for each token through SoftMax function. The model is denoted as *BERT_Lexicon*.

Formally, given an abstract with N tokens $X = [x_1, x_2, ..., x_N]$, an entity is a span of tokens $s = [x_i, ..., x_j](0 \le i \le j \le N)$ associated with an entity type. Based on sequence labeling formulation, the goal of the BioNER model is to assign a sequence of labels $Y = [y_1, y_2, ..., y_N]$ to the input X. For lexicon feature generation, we looked up the lexicon symbols from lexicon embedding matrix for each token and the lexicon features of the input abstract were represented as a sequence of vectors $Z = (z_1, z_2, ..., z_N)$. The whole framework of the model can be split into two stages.

Stage 1: Supervised Learning on Limited Human Labeled Data.

In the first stage, M abstracts that were already annotated by human at token level and lexicon features were generated by BIO tagging scheme, denoted as $\{(X_m, Z_m, Y_m)\}_{m=1}^{M}$. Let $f_1(X, Z; \theta)$ denote the *BERT_Lexicon* model in the first stage, which computes the probabilities for predicting the entity label sequence given an abstract with N tokens, where θ is the parameter of the model. We train *BERT_Lexicon* model by minimizing the cross-entropy loss over $\{(X_m, Z_m, Y_m)\}_{m=1}^{M}$:

$$\tilde{\theta} = argmin_\theta \frac{1}{M} \sum_{m=1}^{M} l(Y_m, f(X_m, Z_m; \theta)) \tag{1}$$

$$l(Y_m, f_1(X_m, Z_m; \theta)) = \frac{1}{N} \sum_{n=1}^{N} -log f_{n,y_n}(X_m, Z_m; \theta) \tag{2}$$

where $f_{n,y_n}(.;.)$ denotes the probability of the n-th token belonging to the y_n-th class.

Stage 2: Bootstrapped Self-Training on Large Scale Pseudo Labeled Data.

After we have learned the model of stage 1, we used it to inference on unannotated data to automatically create silver standard data, which we called pseudo labeled dataset. We proposed a strategy for bootstrapped semi-supervised algorithm (Algorithm 1), which iteratively train the model using chunks of the pseudo labeled dataset.

We denote human labeled training data as H, unannotated data as U, test data as T and the number of chunks that we split U into as k. The objective of the algorithm is to get the model after iterative training, named *FinalModel*. The whole process is further detailed in the following four steps. Noted that the model architecture in stage 2 is the same with that in stage 1.

Step 1: Train the M_{human} model using H and split the unannotated dataset U into k chunks; create an empty dataset *ProcessSet* to indicate the dataset to be inferenced;

Algorithm 1 Bootstrapped Semi-Supervised Algorithm

Input: H for human labeled training data, U for unannotated data, T for test data, k for the number of chunks
Output: $FinalModel$

Train NER model M_{human} using H
Split U into k chunks, each chunk named U_i
$TrainingSet \leftarrow H$
$ProcessSet$ is set empty
for $i = 1 \rightarrow k + 1$ **do**
  Get each chunk U_i in U
  if $i = 1$ **then**
    $TrainedModel \leftarrow M_{human}$
  else
    Train NER model M_{i-1} using $TrainingSet$
    $TrainedModel \leftarrow M_{i-1}$
  end if
  Evaluate the performance of $TrainedModel$ on T
  if $i = k + 1$ **then**
    Set $ProcessSet$ empty
    $FinalModel \leftarrow TrainedModel$
  else
    Add U_i to $ProcessSet$
    Use $TrainedModel$ to inference on $ProcessSet$ and obtain the pseudo labeled dataset P_i
    Add P_i to $TrainingSet$
  end if
end for

create the training dataset named $TrainingSet$, which is initiated by H.

Step 2: For the first iteration, add the first chunk of dataset U named U_1 to the $ProcessSet$ and use M_{human} to inference on it to get the pseudo labeled dataset P_1; update $TrainingSet$, which is the training dataset that will be used in the next generation, by adding P_1 to it.

Step 3: For the following iteration except the last round, retrain the NER model with $TrainingSet$ to get the new trained model M_{i-1}; add next chunk of U to $ProcessSet$ and inference on $ProcessSet$ to get new pseudo labeled dataset, which will be further added to $TrainingSet$.

Step 4: For the last iteration, all the data in U has been added to $ProcessSet$ and has been inferenced to get the final pseudo labeled dataset P_k; get the final model $FinalModel$ by training on $TrainingSet$, which combines H and P_k.

4 Experiments & Results

4.1 Dataset

In this study, we constructed training set and test set for BioNER of Chinese scientific literature manually, which contains sequence labeling format of 115 medical abstracts and 59 abstracts separately from CSCD. Eleven categories of entities were pre-defined and there is no overlapping between training set and test set. The distribution of entities in these two data sets is shown in Table 2. Eleven categories of entities were pre-defined and the corresponding examples are shown in Table 3.

Table 2. Entity Distribution on Dataset

Entity Type	Training Set	Test Set
Bacteria	39	39
Disease and Symptom	280	133
Human Tissue	102	76
Drug	243	114
Medical Department	42	6
Virus	50	20
Medical Examination	747	382
Surgery	117	66
Acupoint	48	15
Methodology	266	109
Medical Device	82	28

Except for the human labeled dataset, we constructed a large scale unannotated dataset for semi-supervised learning. We collected 99,885 abstracts from CSCD and transferred them into sequence labeling format, preparing as the pseudo labeled dataset. This unannotated dataset had no overlapping with the records in the training set and test set.

4.2 Experimental Design

In this paper, we cast BioNER as a sequence labeling task and implemented HMM, CRF, BiLSTM, BiLSTM-CRF, BERT models as baselines to prove the effectiveness of our framework. We designed two groups of experiments on human labeled dataset and pseudo labeled dataset.

The first group of experiments compared the performance of our BERT_Lexicon model with baseline models to explore the advantage of pretrained language model and the incorporation of lexicon under a small amount of training data. And in the second group of experiments, we used BERT model and BERT_Lexicon model to explore the effectiveness of our bootstrapped semi-supervised algorithms. We set the variable k, which is the number of chunks after splitting unannotated dataset, to 1 (meaning the whole unannotated dataset) and 5 to test the strength of bootstrapping. Noted that no matter in stage 1 or stage 2, we tested on the same test set to guarantee the comparability of results.

4.3 Experimental Settings

We implemented the neural network using transformers[5] library. Training and inference were performed on per-abstract level. Training was done by mini-batch stochastic gradient descent (SGD) with exponential learning rate decay and the initial learning rate was set to 1e-4. Each mini-batch consisted of 24 abstracts with the same 512 tokens. The

[5] https://github.com/huggingface/transformers

Table 3. Examples of Medical Named Entities

Category	Examples
Bacteria	流感嗜血杆菌(Haemophilus influenzae),革兰阳性菌(Gram-positive bacteria)
Disease and Symptom	环状胰腺(Circumferential pancreas),慢性肝损伤(Chronic Liver Injury)
Human Tissue	心肌细胞(Cardiomyocytes),晶状体(Crystalline lens)
Drug	益心舒胶囊(Yixinshu capsules),阿米替林(Amitriptyline)
Medical Department	妇产科(Obstetrics and Gynecology),神经外科(Neurosurgery)
Medical Device	载水淬灭荧光探针(Water-quenched fluorescent probes),色谱仪(Chromatographs)
Virus	鼻病毒(Rhinovirus),生殖道沙眼衣原体(Chlamydia trachomatis in the reproductive tract)
Medical Examination	腺苷酸活化蛋白激酶(Adenylate activated protein kinase),匹兹堡睡眠指数(Pittsburgh Sleep Index)
Surgery	胆囊切除术(Cholecystectomy),宫腔镜检查(Hysteroscopy)
Acupoint	足三里(Zusanli),足通谷(Zutonggu)
Methodology	多指标加权法(Multi-indicator weighting method),紫外-可见分光光度法(UV-visible spectrophotometry)

Bacteria	Disease and Symptom	Human Tissue	Drug	Medical Department	Medical Device	Virus	Medical Examination	Surgery	Acupoint	Methodology
革兰阴性菌 (Gram-negative bacteria)	肿瘤(Tumors)	肺组织(Lung tissue)	生理盐水 (Physiological saline)	呼吸科 (Respiratory)	流式细胞仪(Flow Cytometry)	慢病毒 (Lentiviral)	血清(Serum)	腹腔注射 (Intraperitoneal injection)	足三里(Zusanli)	rt-pcr法(Real-time PCR)
大肠埃希菌 (Escherichia coli)	糖尿病 (Diabetes)	脑组织(brain tissue)	抗菌药物 (Antibacterial drugs)	泌尿外科 (Urology)	色谱柱 (Chromatographic columns)	hbv(Hepatitis B virus)	阳性率 (Positive)	灌胃(Gavage)	三阴交 (Sanyinjiao)	mtt(MTT assay)
革兰阳性菌 (Gram-positive bacteria)	医院感染(Hospital Infections Diabetes)	肝组织(Liver tissue)	乙腈 (Acetonitrile)	普外科 (General Surgery)	透射电镜 (Transmission electron microscopy)	乙型肝炎病毒 (Hepatitis B virus)	tnf-α(Tumor necrosis factor-α)	灌胃给药 (Gavage drug delivery)	百会(Baihui)	cck-8法(Cell Counting Kit-8)
金黄色葡萄球菌 (Staphylococcus aureus)	高血压(High blood pressure)	肝脏(Liver)	亚胺培南 (Imipenem)	产科 (Obstetrics)	电镜(Electron microscopy)	lps(Lipopolysaccharides)	手术时间 (Surgery time)	皮下注射 (Subcutaneous injection)	内关(Neiguan)	rct(Randomized controlled trial)
铜绿假单胞菌 (Pseudomonas aeruginosa)	2型糖尿病 (Type 2 diabetes)	肿瘤细胞 (Tumor cells)	万古霉素 (Vancomycin)	重症监护室 (Intensive Care Unit)	扫描电镜(Scanning Electron Microscope)	慢病毒载体 (Lentiviral vectors)	耐药率(Drug resistance rate)	静脉注射 (Intravenous injection)	关元 (Guanyuan)	hplc(High-performance liquid chromatograph)

Figure 2. Top 5 Most Frequently Occurring Entities for Each Entity Type

lexicon lookup table was randomly initialized with values obedient to the standard normal distribution. We used Adam as the optimization algorithm to update the parameters of neural network. We used BERT-Base-Chinese model to initialize the model parameters in stage 1 and stage 2, which is composed of 12 Transformer blocks, the hidden size is 768 and the number of self-attention heads is 12. The model was trained for 100 epochs in stage 1 and 3 epochs in stage 2. It's worth noting that we adopted early stopping to prevent the model from overfitting.

4.4 Evaluation Metrics

In the experiments of BioNER, we were concerned with how many correct named entities we can identify from the given text rather than the label of each token. Therefore, we used CoNLL-2000 Evaluation Scripts[6] to calculate entity-level performance. As we can see from Table 2, there is a sever category imbalance in the dataset, so we decided to use weighted average precision, recall and F1-score to evaluate model performance. More specifically, we assigned different weights to each category according to its sample size.

[6]https://www.clips.uantwerpen.be/conll2000/chunking/conlleval.txt

Firstly, we calculated the evaluation metric for each category, in which the number of False Positives (FP), False Negatives (FN) and True Positives (TP) are used to compute precision (P), recall (R) and F1-score (F1) for category i (the total number of categories is n). The formulas for each metric are as follows.

$$P_i = \frac{\#TP_i}{\#(TP_i + FP_i)} \tag{3}$$

$$R_i = \frac{\#TP_i}{\#(TP_i + FN_i)} \tag{4}$$

$$F1_i = \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{5}$$

where *TP* denotes the entity that is predicted by the model and also appears in the ground truth; *FP* denotes the entity that is predicted by the model but does not appear in the ground truth; and *FN* denotes the entity that is not returned by the model but appears in the ground-truth.

And then we assigned weights to corresponding evaluation metrics according to sample size s_i of category i. The final formulas for weighted average are as follows:

Table 4. Experimental Results on Human Labeled Dataset

Method	P	R	F1
HMM	25.57%	23.55%	24.26%
CRF	59.34%	32.00%	40.95%
BiLSTM	20.82%	27.11%	23.40%
BiLSTM-CRF	28.65%	32.29%	30.19%
BERT	67.75%	74.55%	70.88%
BERT_Lexicon	**70.51%**	**75.67%**	**72.79%**

$$P = \frac{\sum_{i=1}^{n} s_i * P_i}{\sum_{i=1}^{n} s_i} \tag{6}$$

$$R = \frac{\sum_{i=1}^{n} s_i * R_i}{\sum_{i=1}^{n} s_i} \tag{7}$$

$$F1 = \frac{\sum_{i=1}^{n} s_i * F1_i}{\sum_{i=1}^{n} s_i} \tag{8}$$

4.5 Results

Experiment Results on Human Labeled Dataset.

We counted the precision, recall and F1-score metrics to evaluate the model performance. Table 4 shows the experimental results on human labeled dataset. We used F1-score to compare model performance, which takes both of precision and recall into consideration. As we can see, the neural networks including BiLSTM and BiLSTM-CRF achieved bad model performance with 23.40% and 30.19% F1-score respectively, 17.55% and 10.76% worse than CRF model. While pretrained language model outperformed other baselines models for more than 29% and the incorporation of external lexicon features brought a 1.91% improvement compared to BERT model.

This results demonstrated that although deep neural networks are capable of learning highly nonlinear features, they are prone to over-fitting on small amounts of data compared to traditional machine learning methods. While pretrained language model has already learned lots of semantic and syntactic knowledge in the unsupervised pretraining process. The captured knowledge from pretrained language model enabled downstream supervised learning tasks to achieve great model performance even with small amounts of annotated data. In addition, the introduction of lexicon features acted as the complement to BERT model, remedying the shortcoming of limited training data to some extent.

To see the extracted entities more clearly, we counted the top 5 most frequently occurring entities in each entity type on the pseudo labeled dataset, as is shown in Figure 2. As we can see, the BERT_Lexicon model of the first stage has already acquired relevant entity knowledge even with a small amount of annotated training data. In addition, the model can handle the English abbreviations well. For example, in the entity type of 'Methodology', all the top 5 frequently occurring entities are English abbreviations, indicating the value of the customized tokenizer and the strong capability of the model for capturing contextual information.

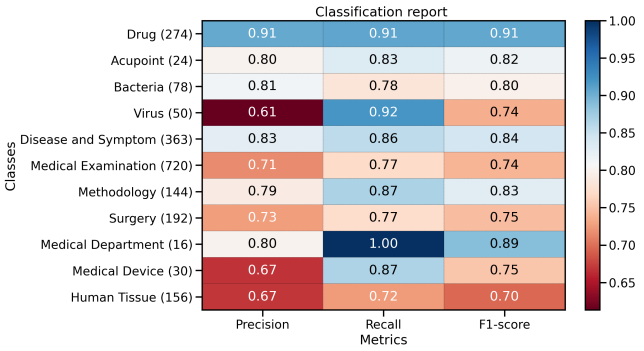

Figure 3. Classification Report Heatmap

Experiment Results on Pseudo Labeled Dataset.

Table 5 shows the experiment results on pseudo labeled dataset and the up-arrow column means the performance improvement compared to the last iteration[7]. Compared to using the pseudo labeled dataset of the whole unannotated data, using bootstrapping algorithm to iteratively generate pseudo labeled dataset can further boost the model performance. For $k = 5$, the model performance improved by 6.92% and 6.08% for BERT and BERT_Lexicon model respectively after 5 iteration rounds, outperforming the corresponding model performance for 3.4% and 3.1% respectively under $k = 1$.

As iteration round increased, both of BERT model and BERT_Lexicon model yielded better results and BERT_Lexicon model still outperformed BERT model in each iteration round, showing the effectiveness of lexicon features to assist in model training. Our final bootstrapped model with lexicons significantly improved the performance of baseline models of stage 1 including HMM, CRF, BiLSTM, BiLSTM-CRF, BERT model by 54.60%, 37.92%, 55.46%, 48.67%, 7.99% respectively.

For further analysis, we used the final model to evaluate the BioNER performance in each category and used heatmap to show the related information, as is shown in Figure 3. The x axis represents the evaluation metrics, and the y axis represents the category and the number of corresponding entities. We observed that the entity distribution was imbalanced and there was a relative big difference of performance among categories. In addition, the balance between precision and recall for some categories like Virus might also be big. We assumed that one of the reasons for these phenomena might be that the entity distribution is imbalanced in reality,

[7]The performance improvement for the first iteration means improvement over the corresponding model performance on human labeled dataset.

Table 5. Experiment Results on Pseudo Labeled Dataset

k	Round	#Data	Method	P	R	F1	↑	Method	P	R	F1	↑
5	1	20092	BERT	70.37%	76.40%	73.08%	+2.20%	BERT_ Lexicon	71.93%	78.41%	74.87%	+2.09%
	2	40069		72.63%	77.97%	75.02%	+1.94%		73.91%	79.04%	76.26%	+1.39%
	3	60046		73.76%	78.75%	76.01%	+1.00%		74.77%	79.63%	76.93%	+0.67%
	4	80023		74.89%	79.19%	76.80%	+0.79%		76.21%	80.70%	78.24%	+1.31%
	5	100000		75.94%	80.12%	77.80%	+1.00%		**76.53%**	**81.53%**	**78.86%**	+0.62%
1	1	100000		71.65%	77.53%	74.40%	+3.52%		73.70%	78.36%	75.76%	+2.98%

making it hard for the model to learn these categories simultaneously.

5 Practical Usage

In order to make our proposed model publicly available and widely used and tested, we built an online annotation tool for Chinese Biomedical Named Entity Recognition (Available at http://sciengine.las.ac.cn/NER_MED_CN). We used Flask to start the service in the background so that the model was preloaded and the model prediction process can be completed in a very short time. As is shown in Figure 4, the online annotation tool allowed users to type in any Chinese medical texts in the text box. The annotated results of the texts would be returned in real time after the 'NER' button was clicked. We used different colors to distinguish the named entities identified in the text and displayed their category labels.

In addition, for those who want to achieve batch annotation of a large number of documents, we provide an API service based on the http protocol, which can be accessed by GET or POST methods. The details of our API service are available at http://sciengine.las.ac.cn/API.

6 Conclusions

In this work, we took advantage of pretrained language model and external lexicon features to Chinese BioNER task and constructed a two-stage framework to provide a feasible path to compensate for the shortcomings of limited annotated data. A bootstrapped semi-supervised algorithm was proposed to generate pseudo labeled dataset iteratively, which can further improve the model performance. Our approach embodies a simple architecture that does not require a dataset-specific architecture or complicated feature engineering. In the future, we will explore more advanced pseudo-labeling methods to increase the model's ability to process noise, increasing the model's generalization capability.

7 ACKNOWLEDGMENTS

The work is supported by the Project "Artificial Intelligence Engine Construction Based on Scientific Literature" (Grant No.E0290906) and the project "Design and Research on a Next Generation of Open Knowledge Services System and Key Technologies" (Grant No.2019XM55).

Figure 4. Interface of the Online Demo for Chinese BioNER

References

[1] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274, 2009.

[2] Desislava Petkova and W Bruce Croft. Proximity-based document representation for named entity retrieval. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 731–740, 2007.

[3] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[4] Chinatsu Aone, Mary Ellen Okurowski, and James Gorlinsky. Trainable, scalable summarization using robust NLP and machine learning. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 62–66, Montreal, Quebec, Canada, August 1998. Association for Computational Linguistics.

[5] Diego Mollá, Menno Van Zaanen, Daniel Smith, et al. Named entity recognition for question answering. 2006.

[6] Ulf Leser and Jörg Hakenberg. What makes a gene name? named entity recognition in the biomedical literature. *Briefings in bioinformatics*, 6(4):357–369, 2005.

[7] Huanzhong Duan and Yan Zheng. A study on features of the crfs-based chinese named entity recognition. *International Journal of Advanced Intelligence*, 3(2):287–294, 2011.

[8] Qi Wang, Yangming Zhou, Tong Ruan, Daqi Gao, Yuhang Xia, and Ping He. Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition. *Journal of biomedical informatics*, 92:103133, 2019.

[9] Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. Towards improving neural named entity recognition with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5301–5307, 2019.

[10] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Dean Cheng, Craig Knox, Nelson Young, Paul Stothard, Sambasivarao Damaraju, and David S Wishart. Polysearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic acids research*, 36(suppl_2):W399–W405, 2008.

[13] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. Sr4gn: a species recognition software tool for gene normalization. *PloS one*, 7(6):e38460, 2012.

[14] Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, 12(6):e0179488, 2017.

[15] Alexandra Pomares Quimbaya, Alejandro Sierra Múnera, Rafael Andrés González Rivera, Julián Camilo Daza Rodríguez, Oscar Mauricio Muñoz Velandia, Angel Alberto Garcia Peña, and Cyril Labbé. Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science*, 100:55–61, 2016.

[16] Shaodian Zhang and Noémie Elhadad. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–1098, 2013.

[17] Jianbo Lei, Buzhou Tang, Xueqin Lu, Kaihua Gao, Min Jiang, and Hua Xu. A comprehensive study of named entity recognition in chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5):808–814, 2014.

[18] Jianbo Lei. Named entity recognition in chinese clinical text. 2014.

[19] Kaixin Liu, Qingcheng Hu, Jianwei Liu, and Chunxiao Xing. Named entity recognition in chinese electronic medical records based on crf. In *2017 14th Web Information Systems and Applications Conference (WISA)*, pages 105–110. IEEE, 2017.

[20] Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. Clinical entity recognition using structural support vector machines with rich features. In *Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics*, pages 13–20, 2012.

[21] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[22] Satoshi Sekine and Elisabete Ranchhod. *Named entities: recognition, classification and use*, volume 19. John Benjamins Publishing, 2009.

[23] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[24] Donghuo Zeng, Chengjie Sun, Lei Lin, and Bingquan Liu. Lstm-crf for drug-named entity recognition. *Entropy*, 19(6):283, 2017.

[25] Yonghui Wu, Min Jiang, Jianbo Lei, and Hua Xu. Named entity recognition in chinese clinical text using deep neural network. *Studies in health technology and informatics*, 216:624, 2015.

[26] Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. Bidirectional LSTM-CRF for clinical concept extraction. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 7–12, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[27] Zhucong Li, Zhen Gan, Baoli Zhang, Yubo Chen, Jing Wan, Kang Liu, Jun Zhao, and Shengping Liu. Semi-supervised noisy label learning for chinese medical named entity recognition. *Data Intelligence*, pages 1–10, 2021.

[28] H Sintayehu and GS Lehal. Named entity recognition: a semi-supervised learning approach. *International Journal of Information Technology*, 13(4):1659–1665, 2021.

[29] Tsendsuren Munkhdalai, Meijing Li, Khuyagbaatar Batsuren, Hyeon Ah Park, Nak Hyeon Choi, and Keun Ho Ryu. Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. *Journal of cheminformatics*, 7(1):1–8, 2015.

[30] Liangping Ding, Zhixiong Zhang, and Yang Zhao. Bert-based chinese medical keyphrase extraction model enhanced with external features. *International Conference on Asia-Pacific Digital Libraries*, 2021.

[31] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.