Article

Not peer-reviewed version

A New Chinese Named Entity Recognition Method for Pig Disease Domain Based on Lexicon Enhanced BERT and Contrastive Learning

Cheng Peng , Xiajun Wang , Qifeng Li * , Qinyang Yu , Ruixiang Jiang , Weihong Ma , Wenbiao Wu , Rui Meng , Haiyan Li , Heju Huai , Shuyan Wang , Longjuan He

Posted Date: 23 July 2024

doi: 10.20944/preprints202407.1804.v1

Keywords: pig disease; Chinese named entity recognition; lexicon enhanced BERT; contrastive learning; small sample

Article

A New Chinese Named Entity Recognition Method for Pig Disease Domain Based on Lexicon Enhanced BERT and Contrastive Learning

Cheng Peng [1,2,3], Xiajun Wang [1,4], Qifeng Li [1,2,3,*], Qinyang Yu [1,2,3], Ruixiang Jiang [1,2,3], Weihong Ma [1,2,3],Wenbiao Wu [1,2,3], Rui Meng [1,2,3], Haiyan Li [1,2,3], Heju Huai [1,2,3], Shuyan Wang [1,2,3] and Longjuan He [5]

[1] Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China
[2] National Innovation Center of DigitalTechnology in Animal Husbandry, Beijing 100097, China
[3] National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
[4] Faculty of Resources and Environmental Science Hubei University, Wuhan 430061, China
[5] Institute of Agricultural Economics and Development, CAAS,Beijing 100081, China
* Correspondence: liqf@nercita.org.cn

Featured Application: Our work provides reliable technical support for the information extraction of pig diseases in Chinese. It can be applied to other domain-specific fields, thereby facilitating seamless adaptation for named entity identification across diverse contexts.

Abstract: Named Entity Recognition (NER) is a fundamental and pivotal stage in the development of various knowledge-based support systems, including knowledge retrieval and question-answering systems. In the domain of pig diseases, Chinese NER models encounter several challenges such as the scarcity of annotated data, domain-specific vocabulary, diverse entity categories, and ambiguous entity boundaries. To address these challenges, we propose PDCNER, a Pig Disease Chinese Named Entity Recognition method leveraging lexicon-enhanced BERT and contrastive learning. Firstly, we construct a domain-specific lexicon and pre-train word embeddings in the pig disease domain. Secondly, we integrate lexicon information of pig diseases into the lower layers of BERT using a Lexicon Adapter layer, which employs char-word pair sequences. Thirdly, to enhance feature representation, we propose a lexicon-enhanced contrastive loss layer on top of BERT. Finally, a Conditional Random Field (CRF) layer is employed as the model's decoder. Experimental results show that our proposed model demonstrates superior performance over several mainstream models, achieving a precision of 87.76%, a recall of 86.97%, and an F1-score of 87.36%. The proposed model outperforms BERT-BiLSTM-CRF and LEBERT by 14.05% and 6.8%, respectively, with only 10% of the samples available, showcasing its robustness in data scarcity scenarios. Furthermore, the model exhibits generalizability across publicly available datasets. Our work provides reliable technical support for the information extraction of pig diseases in Chinese and can be easily extended to other domains, thereby facilitating seamless adaptation for named entity identification across diverse contexts. The codes have been open-sourced at https://github.com/tufeifei923/pdcner.

Keywords: pig disease; Chinese named entity recognition; lexicon enhanced BERT; contrastive learning; small sample

1. Introduction

In the context of large-scale and intensive pig breeding practices, it is of great significance to establish intelligent diagnostic and preventive measures for pig diseases. Early prevention and timely

 doi:10.20944/preprints202407.1804.v1

diagnosis are pivotal for maintaining swine health and mitigating potential losses. Named Entity Recognition (NER) assumes a critical role in this endeavor by identifying specific entities within textual corpora, serving as the cornerstone for numerous downstream tasks in natural language processing. These tasks include but are not limited to information retrieval, intelligent question answering, and knowledge graph construction. However, the existing entity recognition methods mostly focus on recognition of person, location and organization, etc. Given the pressing need to bolster disease surveillance and management in swine, there arises an urgent imperative to develop specialized NER methodologies tailored to the specific lexicon of pig disease terminology in Chinese.

The early NER methods include rule-based recognition methods and statistics-based machine learning recognition methods. In recent years, with the rapid development of neural networks, methods of deep learning are more suitable for the task of NER and become the mainstream method [1–5].

The rule-based NER method requires the rules which are formulated manually by experts. This method has high accuracy when dealing with small datasets, but it is difficult to expand it on a large scale and apply it in different domains because the rules are based on manual construction, which is a time-consuming task [6].

The statistics-based NER method select the appropriate training model according to the specific research background. Commonly used statistical models include hidden Markov models(HMM), conditional random field model(CRF), branch support vector machine (SVM) and maximum entropy model (ME), etc. Compared to the rule-based model, this method omits many tedious rule designs and are fast, portable and convenient to use [7,8]. However, the statistics-based method requires a large number of manually labeled datasets to train model parameters, which is gradually replaced by deep learning method.

The deep learning based NER method can learn more complex features and achieve good results. In contrast to the preceding two approaches, deep learning-based NER methods do not necessitate an abundance of artificial features. Therefore, the deep learning-based methods has been widely concerned by researchers. Common deep learning models include convolutional neural network (CNN), recurrent neural network (RNN), graph neural network (GNN), deep neural network (DNN), generative adversarial network (GAN), long short-term memory network (LSTM), Transformer and BERT(bi-directional encode representation from transformers) and so on [1,9]. Compared to the rule-based and statistics-based models, deep learning models are dominant and achieve state-of-the-art results in NER. However, the scalability of deep learning models applied in specific domain remains a significant challenge.

The lexicon-based NER method can effectively avoid segmentation errors and improve the accuracy of entity boundary recognition by integrating potential word information into feature vectors. Currently, a large number of lexicon enhanced Chinese entity extraction methods have been proposed, with better performance than methods based on character embedding or word embedding. Lattice-LSTM [10] has achieved new benchmark results on several public Chinese NER datasets. However, the Lattice-LSTM model architecture is complex, which limits its application in many industrial areas requiring real-time NER responses. A convolutional neural network based method that incorporates lexicons using a rethinking mechanism was proposed, which can model all the characters and potential words that match the sentence in parallel [11]. A lexicon-based graph neural network with global semantics was proposed to tackle word ambiguities. In this model, the lexicon knowledge is used to connect characters to capture the local composition, while a global relay node can capture global sentence semantics and long-range dependency [12]. A Lexicon Enhanced BERT (LEBERT) for Chinese sequence labeling was put forward [13]. The model integrates external lexicon knowledge into BERT layers directly by a Lexicon Adapter layer and achieves better performance than both lexicon enhanced models and BERT baseline in Chinese datasets. More character-word association models have been proposed, such as SoftLexicon [14], FLAT [15], PLTE [16].

The pre-trained model-based NER method effectively leverages deep bidirectional contextual information. It demonstrates superior performance with shorter training times, reduced labeling data requirements, and improved results compared to traditional models. Currently, BERT [17] is widely

 doi:10.20944/preprints202407.1804.v1

used, followed by ELMo [18], RoBERTA [19], ERNIE [20], ALBERT [21], and others. At present, the pre-trained models and lexicon are integrated by utilizing their respective strengths. Li proposed Flat-LAttice Transformer for Chinese NER, which converts the lattice structure into a flat structure consisting of spans [15]. Li proposed the LEBERT-BiLSTM-CRF model for elementary mathematics text NER, which integrates external lexicon knowledge into BERT layers directly by a lexicon adapter layer and performs better than other NER models [22].

Contrastive learning acquires feature representations of samples by comparing positive and negative samples in feature space. This approach has garnered significant attention in the fields of computer vision (CV) and natural language processing (NLP). ConSERT (Contrastive Framework for Self-supervised Sentiment Representation Transfer) and SimCSE(Simple Contrastive Learning of Sentiment Embedding) model, which use different data enhancement methods and comparative learning loss function to learn the representation of sentences, obtain SOTA results on the task of text semantic similarity [23,24]. COntrastive learning with Prompt guiding for few-shot NER (COPNER) was proposed and outperforms state-of-the-art models with a significant margin in most cases. This method introduces category specific words COPNER composed of prompts as supervised signals for contrastive learning to optimize entity token representation [25]. Moreover, Named Entity Recognition in low-resource scenarios based on contrastive learning has also received considerable attention [26–28]. He proposed a novel prompt-based contrastive learning method for few-shot NER without template construction and label word mappings [26]. Li proposed a multi-task learning framework CLINER for Few-Shot NER [27].

In the field of livestock husbandry, text mining, Named Entity Recognition (NER), intelligent question-and-answer systems, and artificial intelligence (AI) technologies have been gradually applied. However, this field faces numerous challenges, including the prevalence of technical terms, complex knowledge structures, fine knowledge granularity, and a lack of labeled datasets [29]. Seok created a BERT-DIS-NER model that adds a CRF layer to BERT for the disease named entity recognition and used syllable unit-based named entity recognition that can reflect the characteristics of disease names. The F1-score is 0.81 trained with human data and fine-tuned with animal data [30]. Kung designed and implemented an intelligent knowledge question-and-answer system for pig farming based on bi-GRU and SNN methods, combined with the LTSM deep-learning method [31].

NER methods have been found extensive applications in the agricultural domain and other vertical fields [32–37]. Nonetheless, there remains a apparent gap in current research concerning the accurate recognition of named entities within the domain of pig diseases in Chinese. Pig disease data is characterized by complex entities, fuzzy boundaries and domain-specific vocabulary, which encompasses specialized terminologies drawn from the domains of animal husbandry and veterinary science.

Furthermore, the resources in the field of pig diseases are confined and dispersed, exacerbating the scarcity of publicly available benchmark corpora and labeled datasets specific to this domain in Chinese. While considerable research has been devoted to NER systems in human medicine [38,39], it remains impractical to directly transfer such models to the domain of pig diseases due to the domain-specific rules and vocabulary governing this domain. Hence, named entity recognition in the field of pig diseases needs to be further explored. A model of Pig Disease Chinese Named Entity Recognition(PDCNER) is proposed in this paper. The main contributions of the paper are as follows:

(1) We propose a simple yet effective NER model that integrates enhanced lexicon and contrastive learning for the complex pig disease domain, making the model more sensitive to texts in this domain and improving predictions for entities. The lexicon-enhanced BERT facilitates the direct integration of external lexicon knowledge of pig diseases into BERT layers via a Lexicon Adapter layer.

(2) To enrich the semantic feature representation and improve performance under data scarcity conditions, we propose a lexicon-enhanced contrastive loss layer on top of the BERT encoder. Experimental results on small sample scenarios and common public datasets demonstrate that our model outperforms other models.

 doi:10.20944/preprints202407.1804.v1

(3) Given the lack of an annotated corpus for the pig disease domain, we collected and annotated a new Chinese corpus and annotated datasets consisting of 7,518 entities. To address the insensitivity of word segmentation caused by the specialization of the pig disease domain, we constructed a lexicon for identifying specific terms in pig diseases using frequency statistics methods under the guidance of veterinarians.

The remainder of the paper is organized as follows: Section 2 introduces the data set and method proposed in this paper. Section 3 provides a detailed description of the our experiments and analyzes the results. Finally, the conclusion are presented in Section 4.

2. Materials and Methods

2.1. Materials

2.1.1. Corpus Collection and Pre-Processing

Due to the lack of NER public benchmark datasets in pig disease domain, a new Chinese pig disease corpus was constructed and annotated under the guidance of animal disease experts and veterinarians. To ensure the quality of data, we collected information on pig diseases from professional books, published standards, Baidu Encyclopedia and official websites. The data source details are mentioned in Appendix A.

After the data acquisition, we performed basic data per-processing steps. Firstly, the optical character recognition(OCR) technology was used to convert the books and standards into text format. Secondly, useless data such as garbled characters or special symbols, were manually deleted and wrong words were modified in raw text. Thirdly, the duplicate data and invalid data were removed. Ultimately, an comprehensive and effective text corpus of containing 1.45 million characters was obtained(Corpus I of pig disease).

2.1.2. Corpus Annotation

152,596 characters were selected from Corpus I to form Corpus II for entity labeling. Label Studio tool and BIEO labeling method were used to label entities. B represents the start position of the entity, I represents the inside of the entity, E represents the end position of the entity, and O represents the other. Six types of pig disease such as pig type, disease name, body part, symptom, medicine, prevention and control measures were labeled in the corpus text. Finally, the annotated pig disease corpus containing 7518 entities was obtained under the guidance of pig disease experts. The statistical information of labeled entities is presented in Table 1.

Table 1. Statistics of annotated entities.

Category	Category definition	Examples	Numbers	Proportion of the total
Type	Name of different types of pig	妊娠母猪，仔猪 (Pregnant sows,piglets)	735	9.78%
Disease	Name of pig disease	猪丹毒，胸膜炎 (Porcine erysipelas, pleurisy)	958	12.74%
Body parts	Body position, organs and system of pigs	心脏，巨噬细胞 (Heart, Macrophages)	2063	27.44%
Symptom	External performance caused by diseases	气喘，咳嗽，水肿 (Asthma, cough,swollen)	2973	39.55%
Medicine	Medications for treating diseases	替米考星，克林霉素 (Timicosin, clindamycin)	789	10.49%
Total			7518	100%

2.1.3. Construction of Lexicon and Pre-Training Word Embedding

 doi:10.20944/preprints202407.1804.v1

We constructed lexicon in pig disease domain based on Corpus II and professional books. Firstly, the most commonly used professional terms were extracted from the Corpus II by word segmentation and frequency statistics. Then, some professional words in the glossary of professional books were manually added to the dictionary under the guidance of veterinarians, such as “猪副嗜血杆菌病 (Haemophilus parasuis)”,“噻苯达唑 (Thiabendazole)”,“内阿米巴原虫 (Entamoeba spp)”. Finally, the pig disease lexicon comprising 2391 professional terms was obtained for understanding the specific words and technical term. Subsequently, this lexicon was incorporated into the built-in dictionary of Jieba to avoid incorrect segmentation of words.

To obtain a high-quality embedded representation of pig diseases, we trained Corpus I and the lexicon. The Gensim tool was used to train Word2Vec model with a word vector dimension of 200. The construction process of the lexicon and the pre-training of word embeddings are illustrated in Figure 1.

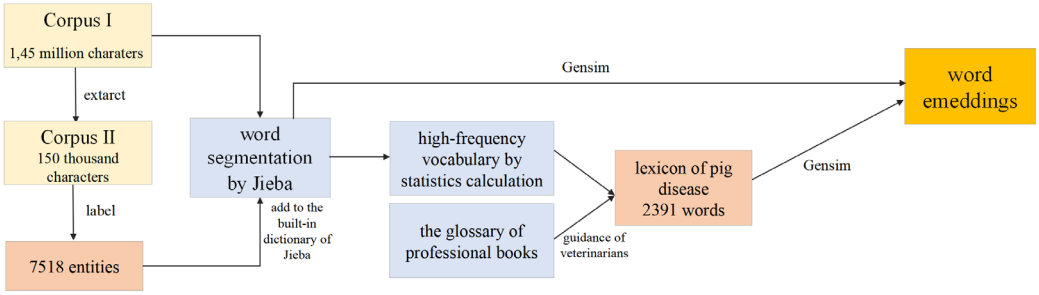


Figure 1. Construction process of lexicon and pre-training word embedding.

2.2. Methods

The structure of PDCNER model is shown in Figure 2. Firstly, the Chinese sentences in the pig disease corpus are converted into a character-words pair sequence, and both Chinese character features and lexicon features are used as inputs. Secondly, a lexicon adapter is added between Transformer layers, which is used to dynamically extract the most relevant matching items. The word of each character uses the bi-linear attention mechanism from character to word, and the lexicon adapter is applied between adjacent Transformer in BERT. The lexicon features and BERT representations are fully interacted through multi-layer encoders in BERT, so that lexicon knowledge can be effectively integrated into BERT. The contrastive loss layer is above the Lexicon Enhanced BERT encoder, ensuring that similar samples are as close as possible, while dissimilar samples are as far apart as possible. Embeddings of the same type of entity are treated as positive samples, whereas embeddings of different types of entities are treated as negative samples. Considering the correlation between consecutive labels, a Conditional Random Field (CRF) layer is employed to label the sequence.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 July 2024 doi:10.20944/preprints202407.1804.v1

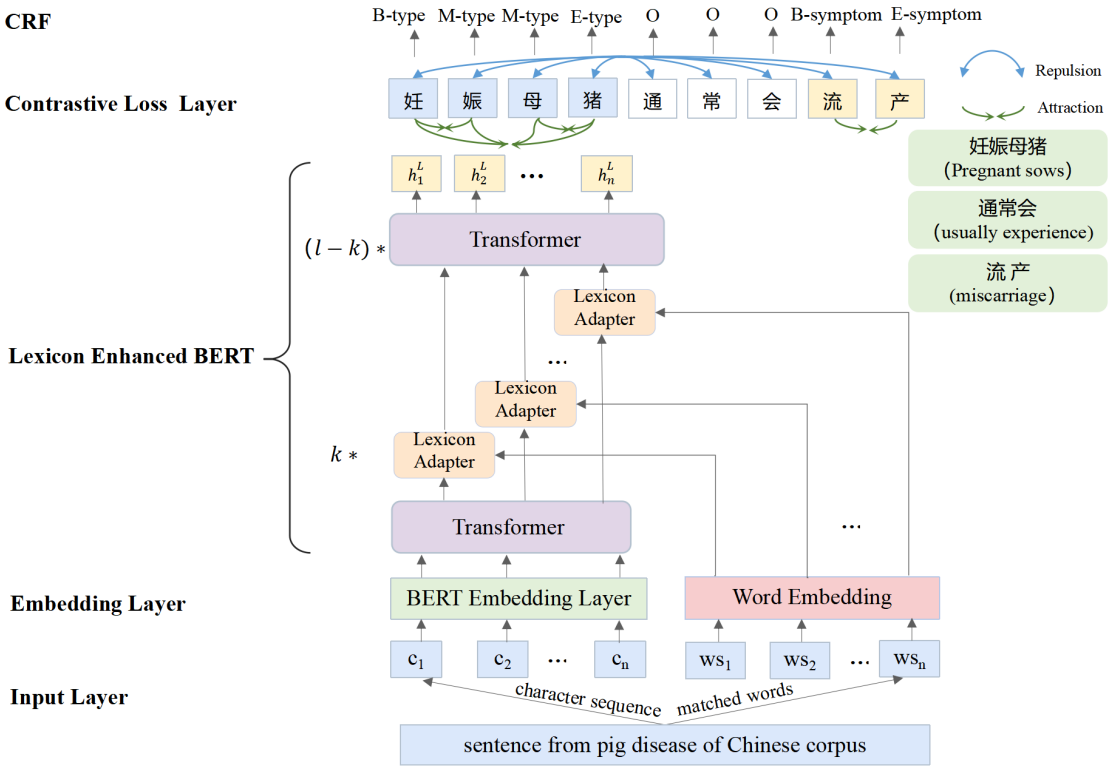


Figure 2. Structure of PDCNER.

2.2.2. Char-Words Pair Sequence

According to the Lexicon Enhanced BERT in [13], we firstly expand the character sequence into a sequence of character-word pairs for applying lexical information of pig disease.

A Chinese sentence with n characters, $S_C = \{c_1, c_2, \cdots, c_n\}$. We identify all potential words within the sentence by comparing the character sequence with the lexicon of pig disease. To achieve this, we first create a Trie data structure based on the lexicon. Then we examine all character subsequences within the sentence, matching them with the Trie to identify all potential words. For instance, the truncated sentence "非洲猪瘟 (African swine fever)" as an example. We can identify five distinct words: "非洲 (Africa)", "非洲猪 (African pig)", "猪瘟 (swine fever)", "瘟 (epidemic disease)" and "非洲猪瘟 (African swine fever)". In the field of pig disease, African swine fever is a complete disease name and should not be separated. Following this, for each matched word, we associate it with the characters that compose it. In conclusion, we pair each character with its associated words and transform the Chinese sentence into a sequence of character-word pairs, represented as:

$$S_{Cw} = \{(c_1, ws_1), (c_2, ws_2), \cdots, (c_n, ws_n)\} \quad (1)$$

where c_i denotes the i-th character in the sentence, and ws_i signifies the words matched and assigned to c_i.

2.2.3. Lexicon Adapter

Using the lexicon adapter proposed in LEBERT [13], the pig disease lexicon information is directly injected into BERT for integrating lexical features.

For the i-th character in a character-word sequence, the input is (h_i^c, x_i^{ws}), h_i^c represents the character vector, the output of a transformation layer in BERT. $x_i^{ws} = \{x_{i1}^w, x_{i2}^w, \cdots, x_{im}^w\}$ represents a group of words embeddings. The j-th word in x_i^{ws} is represented as following:

$$x_{ij}^w = e^w(w_{ij}) \quad (2)$$

where e^w is the pre-trained word embedding list and w_{ij} represents the j-th word in ws_i.

 doi:10.20944/preprints202407.1804.v1

To align these two different representations, a nonlinear transformation is used for each word vector:

$$v_{ij}^{w} = W_2\left(\tanh\left(W_1 x_{ij}^{w} + b_1\right)\right) + b_2 \quad (3)$$

where $W_1 \in R^{d_c \times d_w}$, $W_2 \in R^{d_c \times d_c}$, d_w and d_c represent the dimension of word embedding and BERT's hidden size respectively. b_1 and b_2 are scaler bias.

Each character is associated with a variety of words, but the degree of contribution from each word differs. For instance, in the field of pig diseases, the words "非洲 (Africa)" and "猪瘟 (swine fever)" are more important than "非洲猪 (African pigs)" and "瘟 (epidemic disease)".In order to find the most relevant words, the character-word attention mechanism is used. The correlation of each word is calculated as follows:

$$a_i = \mathrm{softmax}\left(h_i^{c} W_{attn} V_i^{T}\right) \quad (4)$$

where $W_{attn} \in R^{d_c \times d_c}$ is the bi-linear attention mechanism. The all v_{ij}^{w} assigned to i-th character $V_i = (v_{i1}^{w}, \cdots, v_{im}^{w})$, where m is the total number of assigned words.

Lastly, the lexicon information is integrated into the vector representation of the character.

$$\tilde{h}_i = h_i^{c} + \sum_{j=1}^{m} a_{ij} v_{ij}^{w} \quad (5)$$

2.2.4. Lexicon Enhanced BERT

The lexicon adapter is attached between transformer layers in BERT so that the knowledge of pig disease lexicon can be injected into BERT. A sequence of characters $\{c_1, c_2, \cdots, c_n\}$ is input into the input embedding of BERT and then $E = \{e_1, e_2, \cdots, e_n\}$ is obtained by adding token, segmentation and position embedding. After that, E is input into the Transformer encoders. Each layer is as follows.

$$\begin{aligned} G &= \mathrm{LayerNormalization}\left(H^{l-1} + Multiheadattention\left(H^{l-1}\right)\right) \\ H^{l} &= \mathrm{LayerNormalization}(G + \mathrm{FFN}(G)) \end{aligned} \quad (6)$$

where $H^{l} = \{h_1^{l}, h_2^{l}, \cdots, h_n^{l}\}$ represents the output of l-th layer and H^0=E. FFN represents the two-layer feed-forward network with RELU as the hidden activation function.

The lexicon information were input between the k-th and (k+1)-th layer Transformer. $H^{k} = \{h_1^{k}, h_2^{k}, \cdots, h_n^{k}\}$ are got first after k consecutive Transformers layers. Subsequently, each character-word pair (h_i^{k}, x_i^{ws}) was processed through the lexicon adapter to obtain a new hidden layer representation and the i_{th} pair was converted into $\tilde{h}_i^{k}$ accordingly.

$$\tilde{h}_i^{k} = \text{Lexicon Adapter}\left(h_i^{k}, x_i^{ws}\right) \quad (7)$$

$H^{k} = \{h_1^{k}, h_2^{k}, \cdots, h_n^{k}\}$ are input into the remaining (L-K) Transformer as there are 12 layers of Transformers in BERT. Finally, the output of L-th Transformer H^{L} used for the name entity recognition task is obtained.

2.2.5. Lexicon Enhanced Contrastive Learning

The normalized temperature-scaled cross-entropy loss, denoted as NT-Xent was used as our contrastive loss function [40]. For each training iteration, we randomly select N texts from the datasets to form a mini-batch, which yields 2N feature representations. The model is then trained to identify each data point's corresponding pair among the 2(N-1) negative samples present within the batch.

$$\mathcal{L}_{i,j} = -\log \frac{\exp(s(r_i, r_j)/T)}{\sum_{k}^{2N} I_{[k \neq i]} \exp(s(r_i, r_k)/T)} \quad (8)$$

where r_i, r_j represent the embedding for entities of the same type, whereas r_i, r_k denote the embedding for entities of different types. s() refers to the cosine similarity function, where T acts as the

 doi:10.20944/preprints202407.1804.v1

temperature parameter, and I serves as an indicator function. Ultimately, we calculate the final contrastive loss by averaging the classification losses of all 2N instances within the batch.

3. Experiment and Results Analysis

3.1. Evaluation

To identify a named entity for pig diseases, it is necessary to correctly identify both the boundaries of the entity and its corresponding categories. The proposed PDCNER model is evaluated using standard measures, Precision(P), Recall(R), and F1, which are computed using the Eqs. (9), (10) and (11).

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{9}$$

$$R = \frac{T_P}{T_P + F_n} \times 100\% \tag{10}$$

$$F1 = \frac{2PR}{P+R} \times 100\% \tag{11}$$

where Tp represents the number of positive samples that are accurately predicted, while Fp denotes the count of positive samples that are inaccurately predicted. Additionally, Fn stands for the number of negative samples that are incorrectly predicted.

3.2. Experimental Settings

Experiments. The hardware environment that the experimental research relied on was Intel(R) Xeon(R) Silver4116 CPU@2.10GHz, GPU@NVIDIA Tesla P100. The software environment was Python3.8 and tensorflow 2.0.0. The model parameters were set as follows: based on $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019)[17] version, with 12 transformer layers, 768 hidden layers, and 12 multi-head attention mechanisms. The lexicon corresponds to the vocabulary of pre-trained word embeddings in field of pig disease. During the training process, we incorporate the Lexicon Adapter between the first and second Transformer layers of BERT [13]. Meanwhile, the parameters of BERT and the pre-trained word embedding were fine-tuned.

Hyperparameters.The learning rate was 1e-5, the training batch_size was 16, the dropout was 0.5, the optimizer chose Adam, and the number of iterations was 100.

Dataset.This study randomly divided the datasets into training, validation, and test sets according to a ratio of 7:2:1.

3.3. Experiment Results and Analysis

3.3.1. Comparison with Baseline Models

We evaluate the effectiveness of PDCNER against widely-used NER models on the pig disease corpus. The overall findings of our experiments are presented in Table 2.

According to Table 2, several inferences can be drawn: Firstly, the F1 scores of pre-trained models are higher than those of models without pre-trained by more than 10%, such as BILSTM_CRF model. This indicates that pre-trained models, with their deeper network structures, have learned more language features and enhanced their ability to recognize entities. Secondly, the pre-trained model incorporating lexicon information demonstrates better performance than models without lexicon integration. This improvement is primarily due to the integration of dictionary information specific to the pig disease domain, which effectively captures entity boundaries and word information. Thirdly, the recognition performance of our model significantly outperforms other models, achieving a micro-average F1-score of 87.36% in recognizing five major entities. This experimental result indicates that our model can effectively improve entity recognition performance in the domain of pig disease.

(1) Effectiveness of the lexicon enhanced BERT

 doi:10.20944/preprints202407.1804.v1

Comparative analysis with the results of BERT-BiLSTM-CRF reveals notable improvement in the precision, recall, and F1-score of PDCNER, with improvements of 7.47 percentage points, 2.24 percentage points, and 4.91 percentage points, respectively. PDCNER leverages the lexicon adapter to make full use of pig disease feature information, seamlessly integrating it into the BERT architecture. Specifically, the Lexicon Adapter is attached between the 1-st and 2-nd transformers within BERT, facilitating the infusion of pig disease lexicon knowledge into the model's representation.

(2) Effectiveness of the contrastive learning

Through comparative evaluation utilizing the same datasets and downstream model, PDCNER demonstrates superior accuracy in identifying pig disease entities compared to LEBERT, exhibiting improvements in precision, recall, and F1-score by 0.58 percentage points, 0.52 percentage points, and 0.55 percentage points, respectively. This underscores the efficacy of lexicon enhanced contrastive learning, which improve the model's capacity for semantic representation of text on the basis of the normalized temperature-scaled cross-entropy loss function. This loss function enhances the model's ability to identify similar entities by minimizing the distance between positive samples (i.e., similar or related samples) and simultaneously maximizing the distance from negative samples (i.e., unrelated or different types of samples). Consequently, this reduces the risk of the model incorrectly classifying different entities. Furthermore, it allows the model to adjust the weighting of distance measurements, enabling the model to focus more on distinguishing subtle differences between different entities during the training process.

Table 2. Comparison of experimental results for different NER models.

Model category	Model	P(%)	R(%)	F1(%)
baseline model without pre-trained	BILSTM_CRF	71.58	67.51	69.49
pre-trained model	BERT-BiLSTM-CRF	80.29	84.73	82.45
	BERT-CRF	81.62	84.39	82.98
	BERT-CNN-CRF	82.44	80.55	81.48
	BERT-WWM-ext-BiLSTM-CRF	81.73	85.47	83.56
	RoBERTa-BiLSTM-CRF	81.64	85.31	83.43
pre-trained model with lexicon	BERT-BILSTM-CRF-SoftLexicon	82.99	84.73	83.85
	LEBERT	87.18	86.45	86.81
	PDCNER(ours)	**87.76**	**86.97**	**87.36**

3.3.2. The Recognition Effect on Different Entities

For better understanding of the proposed approach, we evaluate the PDCNER model separately on the five entities type, disease, body parts, symptom and medicine, which are presented in Figure.3.

We found that the F1-scores for type, disease, and medicine all exceeded 90%, with the F1-score for type being the highest at 95.41%. Conversely, the lowest F1-score was for the entity of symptom, at 80.92%. The primary reason for this disparity is that the boundaries of pig type and disease entities are very clear, whereas the boundaries of symptom entities are more ambiguous. For instance, type entities typically end with terms like 'pigs (猪)' (e.g., sick pigs (患病猪), conservation pigs (保育猪), fattening pigs (育肥猪)), while disease entities usually end with terms such as 'disease (病),' 'inflammation (炎),' and 'plague (瘟)' (e.g., Porcine blue ear disease (猪蓝耳病), Necrotic enteritis (坏死性肠炎), African swine fever (非洲猪瘟)). In contrast, partial symptom entities are generally composed of complex verbs, conjunctions and modifiers, such as 'feed intake continue to decrease (采食量持续下降)' and 'continuous spasmodic cough (连续痉挛性咳嗽)'. And the average length of complex symptom entities is 8 Chinese characters, which contributes to a low overall recognition rate.

On the other hand, the F1-scores of disease entities and medicine entities were 92.96% and 90.05%, respectively. Both disease and medicine entities include a large number of technical terms, yet the method proposed in this paper achieves a good recognition effect on these two entities. The

 doi:10.20944/preprints202407.1804.v1

results demonstrate that PDCNER fully utilizes both Chinese character features and lexicon knowledge in the pig disease domain at the input level, and the lexicon adapter can effectively leverage pig disease knowledge.

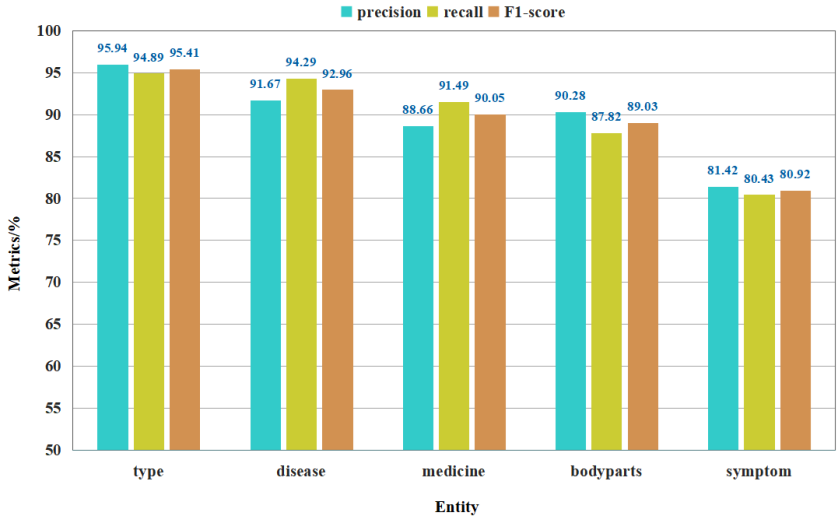
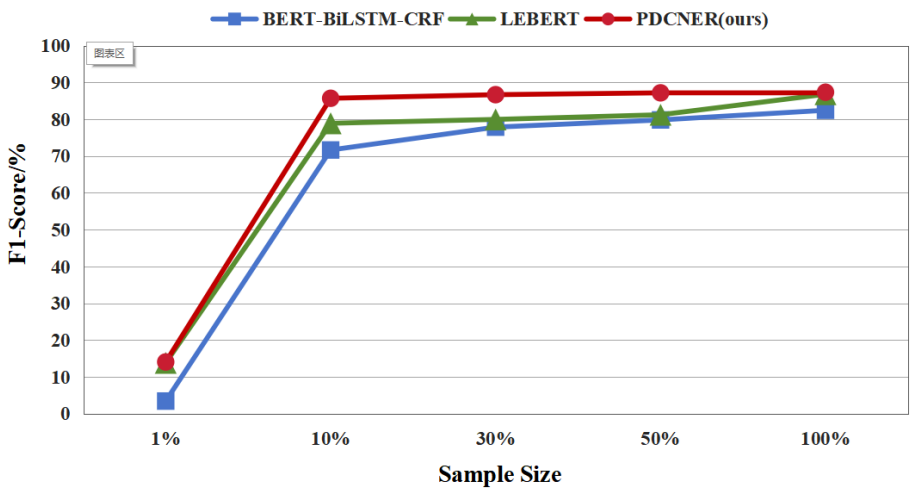


Figure 3. Precision, recall, and F1-score of PDCNER in recognizing five major entities.

3.3.3. The Recognition Effect on Small Sample

In order to verify the reliability and robustness of PDCNER in the condition of scarce data for entity recognition, we used 1%, 10%, 30% and 50% of labeled samples for experimentation. The results can be found in Figure 4 and Table 4. The result shows that the PDCNER model has obvious improvement compared to BERT-BILSTM-CRF and LEBERT.

The F1-score of PDCNER reaches 85.71% when the sample size is 10%, which is only 1.65% lower than that of the full sample. As the sample size increases to 30%, the F1-score of the PDCNER model further improves to 86.69%, showing a marginal decrease of only 0.67% compared to the full sample. Moreover, it outperforms the BERT-BiLSTM-CRF and LEBERT models by 14.05% and 6.8% respectively with only 10% samples available. These results demonstrate the PDCNER model's capability to achieve higher recognition accuracy even under data scarcity scenarios. The incorporation of lexical information in the bottom layer of BERT enables efficient utilization of BERT's representational capabilities. Additionally, the adoption of contrastive learning enhances the semantic representation space, facilitating effective feature capture without extensive training.



 doi:10.20944/preprints202407.1804.v1

Figure 4. F1-score of 3 models on different sample size.

Table 4. Results of the small sample experiments.

Model	1%			10%			30%		
	P	R	F1	P	R	F1	P	R	F1
BERT-BiLSTM-CRF	31.79	1.80	3.41	67.92	75.84	71.66	74.59	81.44	77.87
LEBERT	17.39	11.43	13.79	74.81	83.47	78.91	74.05	80.14	76.97
PDCNER(ours)	18.18	11.43	**14.04**	85.00	86.44	**85.71**	87.22	86.17	**86.69**
Model	**50%**			**100%**					
	P	R	F1	P	R	F1			
BERT-BiLSTM-CRF	79.18	83.45	81.26	80.29	84.73	82.45			
LEBERT	81.09	83.4	82.23	87.18	86.45	86.81			
PDCNER(ours)	87.68	86.71	**87.19**	87.76	86.97	**87.36**			

3.3.4. The Recognition Effect on Public Datasets

To assess the generalization capability of PDCNER, we conducted evaluations across three public datasets: Weibo, Ontonotes, and Resume. As illustrated in Table 5, the PDCNER model achieved the highest F1 scores across all three datasets. Specifically, the F1 values of PDCNER were 9.78%, 4.33%, and 0.82% higher than those of the BERT-BiLSTM-CRF model for the Weibo, Ontonotes, and Resume datasets, respectively. Additionally, compared to the LEBERT model, the PDCNER method showed improvements of 2.53%, 0.37%, and 0.06% for the same datasets. These results indicate that PDCNER exhibits superior performance not only on the pig disease corpus but also demonstrates remarkable generalization capability across different domains.

Table 5. F1-score for each model on public datasets.

Model	**Weibo**			**Ontonotes**			**Resume**		
	P	R	F1	P	R	F1	P	R	F1
BERT-BiLSTM-CRF	71.29	67.1	69.13	84.11	80.2	82.11	96.56	97.23	95.89
LEBERT	78.2	74.64	76.38	89.76	82.68	86.07	95.89	97.42	96.65
PDCNER(ours)	81.42	76.56	**78.91**	89.68	83.43	**86.44**	96.07	97.36	**96.71**

4. Conclusions

High-quality extraction of entity related to pig diseases is critical for intelligent consultation, question answering, technical recommendations, and other application scenarios.

In this study, we constructed a corpus, labeled datasets and lexicon for Chinese named entity recognition specific to pig diseases, encompassing 152,596 characters, 7,518 entities and 2,391 professional terms. To tackle the challenges of entity identification in the pig disease domain, such as the scarcity of annotated data, numerous technical terms, and fuzzy boundaries, we propose the PDCNER model. This model integrates lexicon information from the pig disease domain into the BERT's Transformer layers at the lower level and employs contrastive learning to enhance representation quality and generalization capability. The results indicate that the PDCNER model surpasses the performance of BERT-BiLSTM-CRF and other mainstream models, achieving precision, recall, and F1-score of 87.76%, 86.97%, and 87.36%, respectively. This demonstrates high-quality entity recognition in the field of pig diseases. Moreover, small-sample experiments confirm that our model is more suitable than other models for completing the named entity recognition task in data-scarce scenarios. Experiments on public datasets also verify its generalization ability. Our approach provides a reference for improving NER performance in domain-specific applications.

 doi:10.20944/preprints202407.1804.v1

In future work, we plan to focus on the identification of more fine-grained entity types in animal disease domain, such as appearance symptoms and anatomical symptoms, as well as special types of entities, such as nested entities and discontinuous entities.

Author Contributions: Conceptualization, Cheng Peng and Qifeng Li; Data curation, Rui Meng, Haiyan Li, Shuyan Wang and Longjuan He; Formal analysis, Wenbiao Wu and Heju Huai; Investigation, Shuyan Wang; Methodology, Cheng Peng and Xiajun Wang; Software, Xiajun Wang; Validation, Qinyang Yu, Ruixiang Jiang and Weihong Ma; Writing – original draft, Cheng Peng and Xiajun Wang; Writing – review & editing, Cheng Peng and Qifeng Li.

Funding: This research was funded by National Science and Technology Major Project(2021ZD0113802).

Data Availability Statement: As the datasets used in this manuscript will be used for other technical research, they are available on request from the corresponding author upon reasonable request.

Acknowledgments: We thank the editors and the anonymous reviewers for their valuable suggestions..

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A Data Source Details

No	Type	Example
1	Professional books	Jeffrey J. Zimmerman, Locke A.Kerri, Alejandro Raminez.etc,editor in chief. Hanchun Yang,main translation. Disease of Swine. North United publishing media Media Co., Ltd., Liaoning science and technology publishing house:Beijing, China, 2022. Yousheng, Xu. Primary color atlas of scientific pig raising and pig disease prevention and control. China Agricultural Publishing House: Beijing, China,2017. Changyou,Li, Xiaocheng,Li. Prevention and control technology of swine epidemic disease. China Agricultural Publishing House:Beijing, China,2015. Jianxin Zhang. Diagnosis and control of herd pig epidemic disease. He ‘nan Science and Technology Press:Zhengzhou, China, 2014. Chaoying, Luo, Guibo, Wang. Prevention and treatment of pig diseases and safe medication. Chemical industry press:Beijing, China,2016, etc. 《猪病学》，《科学养猪与猪病防制原色图谱》，《猪群疫病防治技术》，《群养猪疫病诊断与控制》，《猪病防治及安全用药》等
2	Standard specification	Technical Specification for Quarantine of Porcine Reproductive and Respiratory Syndrome (SN/T 1247-2022), Diagnostic Techniques for Mycoplasma Pneumonia in Swine (NY/T 1186-2017), Diagnostic Techniques for Infectious Pleuropneumonia in Swine (NY/T 537-2023), Diagnostic Techniques for Swine Dysentery (NY/T 545-2023), Technical Specification for Quarantine of Porcine Rotavirus Infection (SN/T 5196-2020), etc. 《猪繁殖与呼吸综合征检疫技术规范》 (SN/T 1247-2022), 《猪支原体肺炎诊断技术》 (NY/T 1186-2017), 《猪传染性胸膜肺炎诊断技术》 (NY/T 537-2023), 《猪痢疾诊断技术》(NY/T 545-2023), 《猪轮状病毒感染检疫技术规范》(SN/T 5196-2020) 等
3	Technological specification	Technical specification for prevention and control of highly pathogenic blue ear disease in pigs, technical specification for prevention and control of foot-and-mouth disease, technical specification for prevention and control of classical swine fever, etc. 《高致病性猪蓝耳病防治技术规范》，《口蹄疫防治技术规范》，《猪瘟防治技术规范》等

 doi:10.20944/preprints202407.1804.v1

4	Policy paper	Ministry of Agriculture and Rural Affairs "List of Class I, II and III Animal Diseases", The Ministry of Agriculture issued the "Guiding Opinions on Prevention and Control of Highly Pathogenic Porcine Blue Ear Disease (2017-2020)", Notice of National Guiding Opinions on Prevention and Control of Classical Swine Fever (2017-2020), etc. 农业农村部《一，二，三类动物疫病病种名录》，农业部关于印发《国家高致病性猪蓝耳病防治指导意见（2017—2020年）》，《国家猪瘟防治指导意见（2017—2020年）》的通知
5	Relevant industry website.	China Veterinary Website(https://www.cadc.net.cn/sites/MainSite/), Big Animal Husbandry Website(https://www.dxumu.com/), Huinong Website(https://www.cnhnb.com/), etc. 中国兽医网，大畜牧网，惠农网等

References

1. Li, J; Sun, AX; Han, JL;Li, CL. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*. **2022**,34(1):50-70.
2. Cheng, JR; Liu, JX; Xu, XB; Xia, DW; Liu L; Sheng V. A review of Chinese named entity recognition. *KSII Transatctions on Internet and Information Systems*. **2021**,15(6):2012-2030.
3. Mi, BG; Fan, Y. A review: development of named entity recognition (NER) technology for aeronautical information intelligence. *Artificial Intelligence Review*.**2022**,56(2):1515-1542.
4. Liu P; Guo Y; Wang F, et al. Chinese named entity recognition: The state of the art. *Neuro computing*. **2022**, 473: 37-53.
5. Qiu X; Sun T; Xu Y, et al. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*. **2020**, 63(10): 1872-1897.
6. Zhang, SD; Elhadad, N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*. **2013**,46(6):1088–1098.
7. Zhang, YJ; Zhang, T. Research on Me-based Chinese NER model. In Proceeding of 7th International Conference on Machine Learning and Cybernetics(ICMLC), IEEE, Kunming,China, 12-15 July, 2008, 5,2597-2602.
8. Hu, HP; Zhang, H. Chinese Named Entity Recognition with CRFs: Two Levels. In Proceeding of International Conference on Computational Intelligence & Security, IEEE, Suzhou, China, 2008, 6, 1-6.
9. Kang, Yilin; Sun, Lubing; Zhu; Rongbo; Li Mengyao. Survey on Chinese named entity recognition with deep learning. *Journal of Huazhong University of Science and Technology (Natural Science Edition)*. **2022**, 50(11):44-53.
10. Zhang Yue; Yang Jie. Chinese NER using lattice LSTM. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,Melbourne, Australia, 2018,1: 1554-1564.
11. Gui Tao; Ma Ruotian; Zhang Qi, et al. CNN-Based Chinese NER with lexicon rethinking. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China: AAAI Press,2019,8: 4982-4988.
12. Gui, Tao; Zou, Yicheng; Zhang, Qi; Peng, Minlong; Fu, Jinlan; Wei, Zhongyu; Huang, Xuanjing. A Lexicon-Based Graph Neural Network for Chinese NER. In Proceeding. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),Hong Kong,China,3-7 Nov,2019: 1040-1050.
13. Liu, W ; Fu, X; Zhang, Y; Xiao, W. Lexicon enhanced Chinese sequence labeling using BERT adapter. In Proceeding of 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Bangkok, Thailand, 1-6 August 2021. arXiv2021,arXiv:2105.07148v3.
14. Ma Ruotian; Peng Minling; Zhang Qi, et al. Simplify the usage of lexicon in Chinese NER. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics.**2020**: 5951-5960.
15. Li Xiaonan; Yan Huang; Qiu Xipeng, et al. FLAT: Chinese NER using flat-lattice transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: ACL Press. **2020**: 6836–6842.
16. Xue Mengge;Yu Bowen; Liu Tingwen, et al. Porous lattice transformer encoder for Chinese NER. In Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics. Barcelona, Spain, 13-18 Sep 2020; pp: 3831-3841.
17. J Devlin; MW Chang; K Lee, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv **2018**, arXiv:1810.04805.

18. Peters, Matthew E; Neumann, Mark; Iyyer, Mohit, et al. Deep contextualized word representations. arXiv **2018**, arXiv:1802.05365.
19. Yinhan Liu; Myle Ott; Naman Goyal; Jingfei Du; Mandar Joshi; Danqi Chen; Omer Levy; Mike Lewis; Luke Zettlemoyer; Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXiv **2019**, arXiv:1907.11692.
20. Sun, Y; Wang, S; Li, Y; Feng, S; Wu, H. ERNIE: enhanced representation through knowledge integration. arXiv **2019**, arXiv:1904.09223v1.
21. Lan Z Z; Chen M D; Goodman S, et al. ALBERT: a lite BERT for self-supervised learning of language representations. 8th International Conference on Learning Representations(ICLR),Addis Ababa, Ethiopia,26-30 April 2020.
22. Li S; Bai Z Q; Zhao S; Jiang GS; Shan LL; Zhang L. A LEBERT-based model for named entity recognition. In Proceedings of the 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture(AIAM), ACM International Conference Proceeding Series, Manchester, UK, 23-25 Oct **2021**;pp: 980-983.
23. Yan YM; Li RM; Wang SR; Zhang Fz; Wu W; Xu Wr. ConSERT: A contrastive framework for self-supervised sentence representation transfer. arXiv **2021**, arXiv:2105.11741.
24. Gao T; Yao X; Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. EMNLP, Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 7-11 Nov 2021:6894-6910.
25. Huang Y; He K; Wang Y, et al. COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition[C]. In Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea, 12-17 Oct 2022; pp: 2515-2527.
26. Kai He; Rui Mao; Yucheng Huang; Tieliang Gong; Chen Li; Erik Cambria.,Template-Free Prompting for Few-Shot Named Entity Recognition via Semantic-Enhanced Contrastive Learning. *IEEE transactions on neural networks and learning systems*, **2023**(9). DOI10.1109/TNNLS.2023.3314807.
27. Li, XW; Li, XL; Zhao, MK; Yang, M; Yu, RG; Yu, M ; Yu, J. CLINER: exploring task-relevant features and label semantic for few-shot named entity recognition. *Neural Computing & Applications*. **2023**,36(9):4679-4691.DOI 10.1007/s00521-023-09285-3.
28. Chen, P; Wang, J; Lin, HF; Zhao, D; Yang, ZH; Wren, J. Few-shot biomedical named entity recognition via knowledge-guided instance generation and prompt contrastive learning. *Bioinformatics*.**2023**,39(8). DOI 10.1093/bioinformatics/btad496.
29. Sahadevan, S; Hofmann-Apitius; M, Schellander; K, Tesfaye; D, Fluck; J, Friedrich, CM. Text mining in livestock animal science: Introducing the potential of text mining to animal sciences. *Journal of Animal Science*. **2012**, 90(10): 3666-3676.
30. Oh, Han-Seok; Lee. Hyunah. Named Entity Recognition for Pet Disease Q&A System. *Journal of Digital Contents Society*. **2022**,23(4):765-771.
31. Hsu-Yang Kung; Ren-Wu Yu; Chi-Hua Chen,et, al. Intelligent pig-raising knowledge question-answering system based on neural network schemes. *Agronomy Journal*. **2021**,113(2):906-922.
32. Zhang, D.; Zheng, G.; Liu, H.; Ma, X.; Xi, L. AWdpCNER: Automated Wdp Chinese Named Entity Recognition from Wheat Diseases and Pests Text. *Agriculture,* ***2023,*** 13, 1220. https://doi.org/10.3390/agriculture13061220.
33. Veena G.; Vani Kanjirangat; Deepa Gupta. AGRONER: An unsupervised agriculture named entity recognition using weighted distributional semantic model. *Expert Systems With Applications.* **2023**, 229.https://doi.org/10.1016/j.eswa.2023.120440.
34. Zhang L; Nie X; Zhang M; Gu M; Geissen V; Ritsema CJ; Niu D; Zhang H. Lexicon and attention-based named entity recognition for kiwifruit diseases and pests: A Deep learning approach. *Front. Plant Sci.* **2022**, 13:1053449. doi: 10.3389/fpls.2022.1053449.
35. Guo, XC; Lu, SH; Tang, Z; Bai, Z; Diao, L; Zhou, H ; Li, L. CG-ANER: Enhanced contextual embeddings and glyph features-based agricultural named entity recognition. *Computers and Electronics in Agriculture,* **2022**:106776.
36. Huang, B.; Lin, Y.; Pang, S.; Fu, L. Named Entity Recognition in Government Audit Texts Based on ChineseBERT and Character-Word Fusion. Appl. Sci. 2024, 14, 1425. https://doi.org/10.3390/app14041425.
37. Guo, Y.; Feng, S.; Liu, F.; Lin, W.; Liu, H.; Wang, X.; Su, J.; Gao, Q. Enhanced Chinese Domain Named Entity Recognition: An Approach with Lexicon Boundary and Frequency Weight Features. Appl. Sci. 2024, 14, 354. https://doi.org/10.3390/app14010354.
38. Jia YC; Zhu DJ. Medical Named Entity Recognition Based on Deep Learning. *Computer Systems and Applications*, **2022**, 31(9): 70-81 (in Chinese).
39. DU Jin-hua; YIN Hao; FENG Song. Research and Development of Named Entity Recognition in Chinese Electronic Medical Record. *Acta Electronica Sinica,* **2022**,50(12):3030-3053.
40. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations.**2020**, arXiv preprint arXiv:2002.05709.

 doi:10.20944/preprints202407.1804.v1