Education, Sustainability & Society (ESS)

DOI: http://doi.org/10.26480/ess.01.2025.01.06

ISSN: 2617-9415 (Online)
CODEN: ESSDAX

REVIEW ARTICLE

LEXICON ENHANCED CHINESE NAMED ENTITY RECOGNITION WITH MULTI-TASK LEARNING

Nengfei Luo, Shanli Ye

School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China

**Corresponding author E-mail：slye@zust.edu.cn*

ARTICLE DETAILS

Article History:

Received 20 November 2024
Revised 23 December 2024
Accepted 22 December 2024
Available online 09 January 2025

ABSTRACT

Chinese named entity recognition, since it does not have separators such as spaces to separate entities from each other as in English, the ambiguity of entity boundaries makes it more challenging in prediction. According to the problem that most models integrated lexicon information in such a way that the noise would be introduced from lexicon and thus affects the model performance, in this study, we propose an effective multi-task learning method of lexical enhancement through word selection. By learning the scoring model in task 1, the more useful K words from the matched words in the lexicon will be selected; in task 2 we integrate the character-level and word-level features and feeds them into the BiLSTM to perform the sequence labeling. The two tasks are jointly learned with the same encoder. Comparative experiments were conducted on the public datasets Weibo, Resume, and MSRA, achieving F1-scores of 73.26%, 96.51%, and 95.77%, respectively. The performance showed improvement to some degree across several advanced baseline models. The model proposed in this study effectively integrates lexicon information while also addressing the interference caused by noisy words to a certain extent, thereby further enhancing the performance of named entity recognition.

KEYWORDS

Named entity recognition,lexical enhancement,scoring model, multi-task learning

1. INTRODUCTION

Named entity recognition (NER), as a fundamental task in natural language processing aims at identifying entities with specific meanings in text, such as names of people, places, organizations, etc., and plays an important role in applications such as information retrieva the research sentiment analysis and machine translation (Otter.,2020,Gua,2009, Barnes ,2021).

Due to the differences between English and Chinese texts, Chinese words do not use spaces as segmentation boundaries between words, which makes it difficult to accurately extract word boundary information for Chinese named entity recognition, and there are also challenges in terms of nested entities and ambiguous texts. Therefore, the fine granularity of input information is extremely critical in the Chinese named entity recognition task. If the characters of the text are used as the minimum granularity, it is possible to lose most of the semantic information (Li., 2021). If the vocabulary is used as the minimum granularity, it is easy to propagate the error information in the segmentation to the subsequent work, which affects the effectiveness of the model. Most of the current solutions are character-based models, which require joint learning by adding the task of Chinese word segmentation(CWS) to incorporate features such as glyphs, pinyin, and location after the character encoding layer, but they usually fail to capture the semantic features of the context, so these methods often have limited effect on recognizing entity boundaries. The other approach is a lexical enhancement model that does not adopt a word segmentation task and provides lexicon information by introducing an external lexicon, which provides an effective strategy for integrating lexicon information into character embeddings and has a wide applicability of the model. Although this method simply obtains lexicon information by means of lexicon matching, its performance mainly depends on the quality of the lexicon, and the applicability of the corpus.

As shown in Figure. 1, there may be some noise words when using lexicon information. “Wang Jinbu” in the example is a person's name, and also a golden entity, but when using lexicon matching, words such as “Wang Jin” and “Progress” interfere the recognition of the real entity. Therefore, if we can filter the matched words and select the more semantically related words, we can greatly reduce the interference of noise words on the model and effectively improve the performance of Chinese named entity recognition.

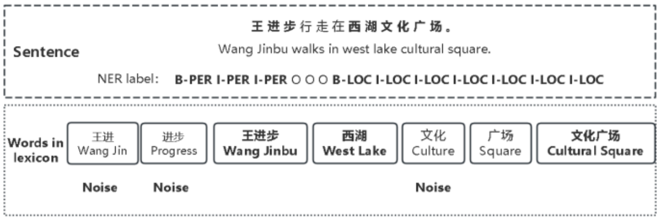
王进步行走在西湖文化广场。
Sentence
Wang Jinbu walks in west lake cultural square.
NER label: B-PER I-PER I-PER O O O B-LOC I-LOC I-LOC I-LOC I-LOC I-LOC I-LOC
Words in lexicon
王进 Wang Jin
进步 Progress
王进步 Wang Jinbu
西湖 West Lake
文化 Culture
广场 Square
文化广场 Cultural Square
Noise
Noise
Noise

Figure 1: The example shows that there are some noise words exist in the lexicon for named entity recognition.

Literally, the words like “progress” and “culture” is matched the lexicon, but their meaning do not match the sentence. In summary, according to the literature on the application of multi-task learning methods in Chinese named entity recognition, on the basis of BiLSTM-CRF, we propose the UWS-BiLSTM-CRF (Useful words selection-bidirectional long-short-term-

Quick Response Code	Access this article online	
	Website: www.educationsustability.com	**DOI:** 10.26480/ess.01.2025.01.06

Cite The Article: Nengfei Luo, Shanli Ye (2025). Lexicon Enhanced Chinese Named Entity Recognition With Multi-Task Learning. *Education, Sustainability & Society, 8(1): 01-06.*

memory- conditional random field): firstly, a series of heuristic rules are designed in task 1 to categorize all lexicon-matched words and select K useful words from them utilizing the word selection model; in task 2, the K useful words are fed into BiLSTM for sequence labeling together with character embeddings and positional embeddings, and finally decoding is accomplished through the CRF. These two tasks are jointly trained using the same BERT encoder in a multi-task learning approach, which significantly enhances the model's semantic extraction capability and the efficiency of operation. In this study, the model performance is validated on three commonly used Chinese named entity recognition datasets, and the results show that the proposed model outperforms other baseline models and can effectively improve entity recognition.

2. Related Work

A variety of efficient methods have been proposed to integrate lexicon information into the character-level feature representation layer, which have been shown to effectively improve the performance of Chinese named entity recognition. In addition, some studies have further improved the overall efficiency of the model through joint multi-task training. This study is inspired by these researches, and thus this section will introduce the research background of these two aspects in detail.

2.1 Lexicon-enhanced Chinese NER

In early Chinese named entity recognition tasks, some models only rely on character information to complete the labeling of entity type labels, such as BiLSTM+CRF and LSTM_CNN (Zhang et al., 2018.Huang et al.,2015). These methods have achieved good results, but the models do not take advantage of the semantic features of the context, and the information they learn is limited. Later, some scholars think that introducing external lexicon information is a feasible improvement direction(Huang et al.,2015).First tried the method in 2018 and proposed the Lattice LSTM, which accomplishes the fusion of character information and word information by designing a lattice structure for each character's potential word information(Zhang et al., 2018). A study designed the FLAT, which improved the performance of Chinese named entity recognition by setting head and tail position indexes for each character and word in the sentence, and converting the complex lattice structure into a planar one with the help of position indexes, and the improved method improved the performance of Chinese named entity recognition (Li., 2020).

In view of the problem that the LSTM-structured models have the inability to recognize long-distance dependencies, (Gui., 2019). Proposed the LR-CNN, which utilizes the characteristics of the dynamic CNN structure and not only well solves the problem that models such as Lattice LSTM have a weak capability of extracting features from the long sequence of input text, but also the Rethinking mechanism of the model can obtain the high-level semantic information. In addition, models such as CGN and multidimensional graph fusion Gazetteer proposed by respectively, both converted the CNER task into a node classification task for graphs by fusing lexicon information with characters through graph structures, and achieved relatively excellent performance (Sui 2019, Ding .,2019).

In order to avoid complex sequence modeling, the Soft-lexicon proposed by (Ma., 2020). Classifies lexicon words into four categories (B, M, E, S) for each character based on the position of the word matched in the lexicon, followed by compression of the four-dimensional word set and then fusion with the character information, which is a simple and efficient method that greatly reduces the complexity of the model. However, the Soft-lexicon method only uses the lexicon information extracted by the soft lexicon method to directly splice with the character information and does not address the interference of the lexicon-enhanced model due to the introduction of external lexicon resulting in noisy words, and the performance of the model largely depends on the quality of the lexicons. Inspired by this, in this paper we propose a simple and effective lexicon-enhanced Chinese named entity recognition method, which firstly divides lexicon matches into layers based on heuristic rules, and then fuses the features at the word level into the character feature representation layer by selecting useful words.

2.2 Multi-task learning Chinses NER

Multi-Task Learning (MTL) was first proposed by some study and this method is a learning mechanism in deep learning that aims to achieve joint training of multiple related tasks by designing models (Caruana.,1997). The method has been shown to solve the overfitting phenomenon during single-task training, because in multitask learning usually the parameters of the hidden layer are shared between tasks for joint optimization, speeding up the overall training efficiency. The advantages of multi-task learning make it widely adaptable to sub-tasks in various domains of natural language processing. The MTL-BERT proposed by while the multi-task learning is that in addition to performing the original CNER task, an additional sub-task is designed in the output of the feature extraction layer, namely the boundary prediction task to avoid the error propagation of word segmentation in Chinese named entity recognition.

In addition, there are more complex models, such as the BIFT proposed by (Meng Liao.,2023, Fang.,2023). Which improves the Transformer structure in the feature extraction layer and fuses the interaction information between characters and labels, and the decoding layer outputs the CNER loss and the Biaffine (biaffine) mechanism outputs the loss of multi-label classification task through CRF, respectively. Meanwhile, in the MTL-HWS designed by a scoring model for selecting the top K helpful words from the lexicon matching words is designed, and then the selected useful words are input into the NER model along with the character sequences to realize the character-level sequence annotation (Tian., 2023). We note that although this method alleviates the noise problem caused by introducing external lexicons to some extent, the model is still not flexible enough based on the fact that lexicons are usually static, when facing the emergence of new words and tasks that need to be dynamically adjusted.

3. Model

The structure of the named entity recognition model in this paper is shown in Figure 2. In general it can be divided into character representation layer, word selection model, feature fusion layer and decoding layer. The model is a simple multi-task learning method using a lexicon, whose task one is to input each word matched in the lexicon to a scoring model, then output its score representing the importance of these words in the sentence respectively, and then select K useful words based on the importance level. Task two is to perform feature fusion between the K useful words selected in task one and the BERT-encoded character embeddings, which are fed into the BiLSTM structure to complete the sequence labeling task.

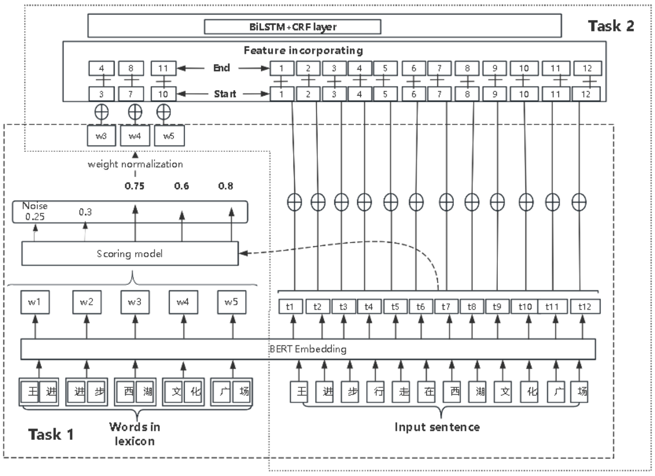

Figure 2: The model graph of the useful words selection-bidirectional long-short-term- memory- conditional random field (UWS-BiLSTM-CRF) based on multi-task learning.

3.1 Embedding layer

The proposal of pre-trained language models has attracted a lot of attention from researchers in the field of natural language processing. The development from Word2vec to BERT has established iconic milestones for natural language processing. The model BERT is able to fully extract semantic information from text through pre-training, and its core structure is transformers. The degree of correlation between individual characters in a sequence is computed through a self-attention mechanism, and then its weights are updated according to the computed degree of correlation to extract global text features.

In addition, the dependencies between different characters can be learned during the pre-training process through the single word mask strategy. In the whole BERT Embedding layer, Task 1 and Task 2 are trained by the same model BERT. Task 1 is to input both lexicon matching words and original sentences into the model BERT to complete the embedding representation of the semantics of each word and sentence; Task 2 is to input the original sentences into the model BERT, and the character sequences are mapped into vector sequences to get the embedding representation of the characters.

3.2 Word selection model

One of the major challenges in the current lexicon-enhanced models is the presence of noise in the candidate words when using external lexicons, which can affect the performance of model recognition. Therefore, we propose a new approach that removes the interference of noisy words as much as possible by designing a model that learns scoring based on heuristic rules, and by optimizing the scoring model can make a difference to selecting matching words that are highly related to the original semantics. For a given text sequence X=[$x_1, x_2, \dots, x_{|X|}$], its substring is denoted by w. w can be any n-gram, including golden entities and lexicon matching words, etc. The key of Task 1 is the optimization of the scoring model, firstly, the original sentence is identified using BERT, and the corresponding connectors [CLS] and separator tags [SEP] are typed to complete the character embedding.; all its substrings are encoded to get the word embedding. Finally, we get $T^x \epsilon R^{(|X|+2)\times d_b}$ for character embedding, and $T^w \epsilon R^{1\times d_b}$ for word embedding; where d_b is the hidden layer dimension of BERT.

Then, a linear activation function acts on the connection of T^x and T^w, the formula for scoring a specific substring in that sentence is obtained as:

$$\text{Score}\big(\text{X}, \ \text{w}\big) = \sigma([T^x \oplus T^w]\omega + \text{b}) \quad (1)$$

Where σ is the sigmoid function, ω and b are the weight and bias with dimension sizes ($2d_b$, 1) and (1, 1), respectively, which are trainable parameters; and ⊕ denotes the vector splicing operation.

In order to evaluate the semantic similarity of each substring in each sentence to the current sentence, we use the above scoring formula to complete the construction of the scoring model according to the heuristic rules defined below:

1. Heuristic rule 1: a golden entity should have score higher than any other word, which means score(X, a golden entity)>score(X, a random word matched in lexicon);

2. Heuristic rule 2: a noun in a lexicon matched word should have score higher than any other lexicon word, which means score(X, a noun matched in the lexicon) > score(X, a non-noun matched in the lexicon);

3. Heuristic rule 3: a word in the lexicon should have score higher than a n-gram of non-lexicon matches, which means score(X, a word matched in the lexicon) > score(X, a random non-lexicon n-grams).

According to the three heuristic rules, the n-grams of all sentences can be sorted into four levels: level 1 is all golden entities, level 2 is lexicon-matched nouns, level 3 is lexicon-matched all words except nouns, and level 4 is all n-grams except the first three. Their scores should also be reduced in order to implement the scoring model according to the learned sorting strategy. The detailed procedure is shown in Figure 3, for each high-level word in each sentence can at most sample 4 low-level. By this, we constructed pairs of ((w_{high}, w_{low}) word sets, as the training set, and input to the BERT sharing parameters. By learning the whole scoring model, we output the score$\big(\text{X}, \ w_i\big)$ of each substring w_i. In order to achieve the effect of high level word scores high versus low level words, the model is designed to be optimized based on marginal ranking loss (as in Formula (2)), where γ is predefined hyperparameter. The training set is trained by this loss function, and all matching word scores can be computed in parallel during the process.

$$LOSS_{score} = \sum \frac{e^{score(X,w_{\text{high}})}}{e^{score(X,w_{\text{low}})+\gamma}} \quad (2)$$

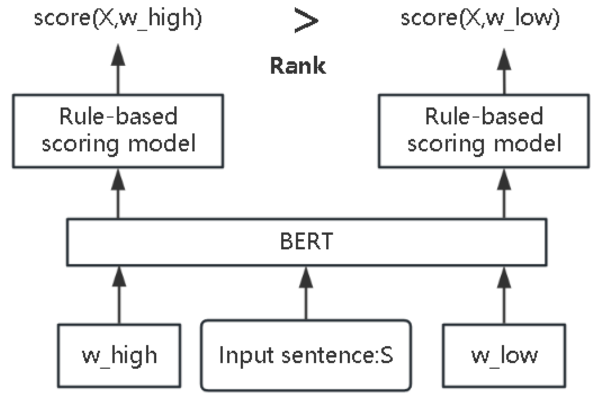

Figure 3: Model for rule-based learning of scoring

3.3 Feature incorporating layer

After the scoring of the lexicon matches is completed, the respectively weights are normalized and the matches that satisfy the conditions are selected as useful words according to the method of preset weight thresholds. Subsequently, these useful words along with the original sentences are input to Task 2 for character-level sequence tagging. The original sentence is embedded through the BERT Embedding layer, while the relationship between characters and words can be represented through the Positional Embedding layer.

Specifically, for the sequence X=[$x_1, x_2, \dots, x_{|X|}$], the character embedding can be represented as$T_i^x \epsilon R^{1\times d_b}$, i=1,2,...,|X|; assuming that the number of substrings in the sequence is represented by m, then each word embedding is represented by$T_j^w \epsilon R^{1\times d_b}$, j=1,2,...,m. The semantic embedding of the whole sequence is accomplished by concatenating the sequence with T_i^x and T_j^w to form a new sequence X$^c \epsilon R^{(|X|+m)\times d_b}$. The positional information needs to be supplemented by the positional embedding layer, and the positional indexes can be converted into vectors by constructing a lookup table. The start and end position embeddings can be represented as $P_{start} \epsilon R^{(|X|+m)\times d_p}$and $P_{end} \epsilon R^{(|X|+m)\times d_p}$ respectively, where d_p need to be predefined according to the sentence length. Finally, the three types of information are incorporated by simple concatenation:

$$x^c = X^C \oplus P_{start} \oplus P_{end} \ \epsilon R^{(|X|+m)\times(d_b+2d_p)} \quad (3)$$

In this way, the updated sequence gathers semantic information of characters and words, as well as positional information.

3.4 Bilstm Layer

For the traditional recurrent neural network (RNN) with the problem of gradient vanishing or gradient explosion, scholars proposed LSTM network (Figure 4), which successfully solved the problem. Through its core unit structure: memory gate i_t, forgetting gate f_t and output gate o_t. The three gating mechanisms can accomplish selective updating and forgetting of semantic information at the current moment to more stably extract the dependencies between contextual words to the text. However, the unidirectional LSTM can only learn historical information and cannot learn future information. Therefore, in this paper, a bidirectional LSTM network is used, where the output H of the feature fusion layer is fed into two LSTMs with opposite directions to obtain feature vectors with strong characterization ability. The forward LSTM is defined as follows:

$$\begin{bmatrix} \boldsymbol{i_t} \\ \boldsymbol{f_t} \\ \boldsymbol{o_t} \\ \widetilde{\boldsymbol{c}}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(\boldsymbol{W} \begin{bmatrix} \boldsymbol{x}_t^c \\ \boldsymbol{h}_{t-1} \end{bmatrix} + \boldsymbol{b} \right) \quad (4)$$

$$c_t = \widetilde{c}_t \odot i_t + c_{t-1} \odot f_t \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

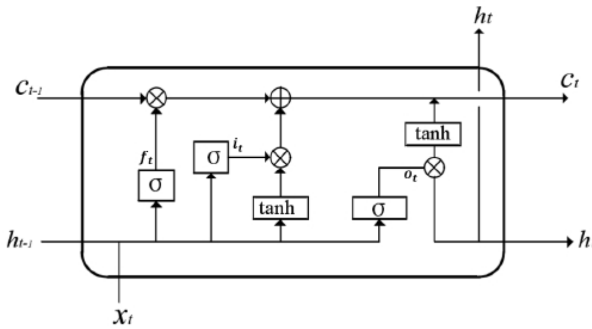

Figure. 4 Unit structure of LSTM, i_t represents the memory gate, f_t represents the forgetting gate, o_t represents the output gate.

Where σ denotes the activation function sigmoid, ⊙ is the vector dot product, W is the weight matrix, and b is the bias term. The backward LSTM structure is consistent with the forward LSTM, and for discrimination, the forward representation is $\overrightarrow{h_t}$ and the backward representation is $\overleftarrow{h_t}$. The two are spliced to form the final BiLSTM output H = [$\overrightarrow{h_t}$, $\overleftarrow{h_t}$], which forms the contextual representation of each character vector $x_t{}^c$.

3.5 Decode layer

In the Chinese named entity recognition task, rich textual semantic features are extracted by BiLSTM to complete the sequence labeling problem, but it cannot address the dependency between labels. Taking BIOS tagging as an example, the real tag sequence has the following three stipulations: (1) This study does not involve the labeling of nested entities as (Zhou G D refences.,2006). So the label of a certain entity type does not contain other types of entities; (2) the label "B"can only be the beginning of an entity, so "S" cannot appear after "B"; (3) the previous label of label "I" can be either "B" or "I". Since Conditional Random Field (CRF) can represent the dependencies between labels using a transfer matrix and obtain the globally optimal prediction results, this study adopts CRF to decode and predict the whole character sequence, and the specific process is as follows:

The probability of taking the input x = H and the original sequence y as the output sequence is denoted as $P(y|x)$ and is calculated as follows:

$$P(y|x) = \frac{e^{score(x,y)}}{\sum_{y' \in Y} e^{score(x,y')}} \qquad (7)$$

$$score(x, y) = \sum_{i=1}^{n} T_{i,y_i} + \sum_{i=0}^{n} A_{y_i y_{i+1}} \qquad (8)$$

$$T = \sigma(xw_t + b_t) \qquad (9)$$

Where y' is the true label sequence; Y is the set of all possible output sequences, T is the transfer matrix; T_{i,y_i} denotes the score of the ith character in the sequence corresponding to the label y_i, and $A_{y_i y_{i+1}}$ denotes the score of labels transferred from y_i to y_{i+1}. w_t and b_t are parameters defined in the fully connected layer.

The loss function for Task 2 is given by:

$$LOSS_{ner} = -log\, P(y^*|x) \qquad (10)$$

Where y^* denotes the sequence of true labels.

4. Experiments

4.1 Datasets

The experiments use three public datasets commonly used for Chinese named entity recognition tasks, including Weibo, Resume and MSRA (Levow, 2006,Zhang and Yang , 2018,Peng et al., 2006). The dataset adopts the BIO tagging system, where B denotes entity beginning, I denotes entity non-beginning, and O is non-entity. The Weibo dataset is obtained by filtering and tagging Sina Weibo data, and contains four entity types, namely, political entity (GPE), locality (LOC), organization (ORG), and person (PER), and its political entities can be further subdivided into general and specific references. The Resume dataset consists of resumes of executives of listed companies provided by Sina Finance as a corpus. It contains eight entity types: race (RACE), country (CONT), organization (ORG), name of person (NAME), education (EDU), title (TITLE), place of origin (LOC), and profession (PRO).The MSRA dataset is a dataset annotated by Microsoft Research Asia and specially used for the CNER task, which contains three types of entity types: name of a place, name of a person, and name of an organization, which are marked as LOC, PER, and ORG, respectively. The data sizes of the three datasets at the sentence level were counted and summarized as shown in Table 1.

Table 1: Statistics of dataset scale

Dataset	Training set	Development set	Test set
Weibo	1350	270	270
Resume	3821	463	477
MSRA	41728	4636	4365

4.2 Evaluation metrics

The three most common metrics used for model evaluation are precision (P), recall (R) and F1-score. These three metrics use the three parameters TP, FP, and FN, where TP denotes the samples that were actually predicted as positive for the positive category; FP denotes the samples that were actually misclassified as positive for the negative category; and FN denotes the samples that were actually misclassified as negative for the positive category.

The precision is the proportion of samples predicted by the model to be in the positive category that are actually labeled as positive:

$$P = \frac{TP}{TP+FP} \qquad (11)$$

Recall is the proportion of positive class samples that are predicted to be positive:

$$R = \frac{TP}{TP+FN} \qquad (12)$$

The F1-score is the reconciled mean of precision and recall:

$$F1 = \frac{2PR}{P+R} \qquad (13)$$

4.3 Experiment setup

For the pre-training language model BERT (Devlin, 2018). The most commonly used version of BERT-Base Chinese is used, with a hidden layer size of 768. If the baseline model is not pre-trained with BERT, the word embeddings are initialized with Giga-Word. The external lexicon used is "YJ", and the weight threshold for selecting high-level words as useful words is set to 0.6. In this study, BiLSTM is used to accomplish sequence tagging, so the hyperparameter settings of the BiLSTM layer mostly follow the Lattice-LSTM, and the size of the embedding layer is set to 50, with a dropout rate of the embeddings the word and character is set to 0.5, respectively. To adapt the model to different sizes of datasets, the LSTM hidden layer is appropriately adjusted, and the hidden layer size is set to 200 for the small datasets (Weibo and Resume), and 300 for the MSRA dataset. The training process is used for optimization using the Adam optimizer in order to prevent overfitting, the training is performed using an exponential decay learning rate, the initial learning rate is set to 4×10^{-4}, for the Weibo dataset, and 1.5×10^{-4} for the Resume and MSRA datasets, with a learning decay rate of 0.05. In addition, the hyperparameter γ in the $LOSS_{score}$ is set to 2×10^{-5}. The detailed parameter settings are shown in the Table 2(Diederik.,2014).

Table 2: Parameters of experiment

Parameter	Value
BERT Embedding dim	768
Char/Bigram emb size	50
Dim of the word vectors fused Lexicon	50
Single-layer LSTM hidden Layer dimension	200（Weibo、Resume）/300（MSRA）
Dropout	0.5
Learning rate	1.5×10^{-4}（Resume、MSRA）/4×10^{-4}（Weibo）
Batch size	18
Maximum length of sequence	128

4.4 Comparison study

4.4.1 Baseline models

For the purpose of validating the effectiveness of the present model, the more popular baseline models in recent years were selected for comparison, especially the lexicon enhanced CNER model.

- BiLSTM-CRF (Huang, 2015). Traditional network structure that does not use a lexicon and only inputs character-level features to obtain label sequence predictions.
- CAN-NER (Zhu et al., 2019). The model uses an improved CNN to get a Convolutional Attention Network (CAN) that captures dependencies between characters and information about the context of the sentence.
- Lattice-LSTM (Zhang and Yang, 2018). Introducing lexicon information at the character representation layer and fusing the lexicon information through an improved LSTM network.
- FLAT (Li et al., 2004). Lexicon-enhanced models using the Transformer structure convert the lattice structure of a Lattice-LSTM into a planar structure with excellent parallelization capabilities.
- LR-CNN (Gui et al.,2019). With the CNN-based lexicon approach, the model adds a new Rethinking mechanism, which is able to utilize CNNs to encode potential words with different size windows.
- Soft-lexicon (Ma et al., 2019). Words are introduced at the embedding layer, and word vectors are represented by BMES tagging classification to fuse lexicon information.

4.4.2 Experiment result

The experimental performance of the model proposed in this paper on Weibo, Resume and MSRA datasets are shown in Tables 3, 4 and 5, and it obtains 73.26%, 96.51%and 95.77% F1- score, respectively, which are better than other models on all three datasets. Especially for the Weibo dataset with irregular text format and small data size, the UWS-BiLSTM-

CRF achieves even better results, boosting the F1-score by 2.76% on the Soft-lexicon, which indicates that the model can better extract the dependency relationship between characters and words. Even on Resume, a text dataset with standardized words, the F1-score of the model is still improved by 0.39% compared with the best effect of Soft-lexicon, which shows that the word selection model plays a great role in fusing the lexicon information and removes the noise even further. Though in the large-scale dataset MSRA, the model improves 1.65% and 0.35% in F1-score than FLAT and Soft-lexicon, which are also pre-trained with BERT, proving that the model is very effective in the CNER task of lexicon enhanced.

Table 3: Experiment result on Weibo

Model	**Named entity**	**Notional entity**	**F1-score**
BiLSTM-CRF	60.80	52.90	56.58
CAN-NER	-	-	59.92
Lattice-LSTM	53.04	62.25	58.79
FLAT（+BERT）	-	-	60.32
LR-CNN	57.14	66.67	59.92
Soft-lexicon（+BERT）	70.94	67.02	70.50
UWS-BiLSTM-CRF（+BERT）	**73.10**	**73.42**	**73.26**

Table 4: Experiment result on Resume

Model	Precision	Recall	F1-score
BiLSTM-CRF	93.70	93.30	93.50
CAN-NER	95.05	94.82	92.56
Lattice-LSTM	94.81	94.11	94.46
FLAT（+BERT）	-	-	95.45
LR-CNN	95.37	94.48	95.11
Soft-lexicon（+BERT）	96.08	96.13	96.10
UWS-BiLSTM-CRF（+BERT）	**96.15**	**96.87**	**96.51**

Table 5: Experiment result on MSRA

Model	Precision	Recall	F1-score
BiLSTM-CRF	90.49	85.76	88.06
CAN-NER	93.53	92.79	93.16
Lattice-LSTM	93.57	92.79	93.18
FLAT（+BERT）	-	-	94.12
LR-CNN	94.50	92.93	93.71
Soft-lexicon（+BERT）	95.75	95.10	95.42
UWS-BiLSTM-CRF（+BERT）	**96.21**	**95.34**	**95.77**

4.5 Ablation study

In order to verify the effectiveness of each part of the UWS-BiLSTM-CRF, ablation experiments are conducted, and the results are shown in Table 6. "-YJ" refers to the method that does not utilize lexicon to integrate lexical features.; "All lexicon" refers to that the word selection model is not used, and all matching words are introduced; "Noun" denotes changing the word selection strategy to select only up to K nouns, according to their weights as selection criteria; "-Rule1" denotes not using heuristic rule 1, which sorts all n-grams according to levels 2, 3, and 4, and "-Rule2" denotes that heuristic rule 2 is not used and all n-grams are sorted only according to levels 1, 4, and "-Rule3" denotes that heuristic rule 3 is not used and all n-grams are sorted according to levels 1, 2, 3. In particular, in order to demonstrate the effectiveness of our proposed dynamic adjustment method of hyperparameter K, three more representative control experiments are set up with K=2, 4, 6, respectively.

Table 6: Ablation experiments

Performance（%）	Weibo			Resume			MSRA		
Model	P	R	F1	P	R	F1	P	R	F1
-YJ	69.13	69.40	69.27	94.02	95.46	94.73	92.79	92.30	92.54
All lexicon	71.82	72.12	71.97	95.37	95.89	95.63	93.73	93.41	93.57
Noun	71.48	71.97	71.72	95.39	96.38	95.88	94.40	93.73	94.06
-Rule1	70.69	71.22	70.95	93.80	94.77	94.28	93.81	92.33	93.57
-Rule2	69.92	70.83	70.37	94.65	95.88	95.26	92.47	93.15	92.85
-Rule3	71.67	72.36	72.01	94.70	96.27	95.48	93.59	93.11	93.35
K=2	71.37	69.70	70.53	95.02	95.43	95.22	93.48	93.67	93.58
K=4	70.28	73.42	71.82	95.79	96.55	96.17	94.12	93.89	94.05
K=6	72.99	73.16	73.07	95.87	**96.92**	96.39	95.20	94.55	94.87
Ours	**73.10**	**73.42**	**73.26**	**96.15**	96.87	**96.51**	**96.21**	**95.34**	**95.77**

From the results in the table, it can be seen that the model "-YJ" which removes the lexicon has the biggest performance degradation, and the performance of "All lexicon" is better than that of "-YJ", but worse than that of "Noun", indicating the necessity of setting a rule to select useful words for all matches, also proving the effectiveness of Heuristic Rule 2. By removing heuristic rules 1-3 from the underlying logic of the scoring model, we find that whichever rule is removed brings about a performance degradation, and almost all of them are lower than the "All lexicon" model. Specifically, removing Rule 2 has more severe consequences in the datasets Weibo and MSRA, which are more sensitive to nouns; removing Rule 1 has the largest impact on Resume, which is consistent with the characteristics of the datasets, as the golden entity in the Chinese resume dataset is the most important information in the scoring model. The removal of Rule 3 affects the performance less, but it is also helpful in the whole model. Finally, the effect of manual setting of K is compared, and it is obvious that the model performance is in an unstable state of fluctuation as the value of K changes, which greatly affects the robustness and generalization ability of the CNER model. Therefore, through a series of ablation experiments it can be concluded that all the structures of the UWS-BiLSTM-CRF model contribute to the performance of entity recognition, achieving the best results in all datasets.

5. Conclusion

In this study, for lexicon enhanced model in CNER task, how to integrate lexicon information into character embedding in a simpler way and solve the problem of noise brought by this process, UWS-BiLSTM-CRF model is proposed. The word vectors and semantic information of sentences are extracted for feature fusion by sharing BERT encoder through Embedding layer and input to BiLSTM-CRF structure, while the scoring model of Task 1 and the Chinese named entity recognition task of Task 2 are jointly trained in a multi-task learning approach. The experimental results show that the proposed model achieves excellent performance in all three datasets, with lexical enhancement and other multiple models on the accuracy rate and F1 value to obtain some improvement. This study is applicable to generalized datasets with few label types, and is only made in flat Chinese named entity recognition, which may not be able to ensure the accuracy of recognition in the face of recognition tasks containing complex nested entities. In the future, we will consider changing the fusion method of character features and useful words to capture the interaction information between words more effectively and enhance the generalization performance of the model in other entity recognition tasks.

References

Otter .,D ., W, Medina J R., Kalita J K,.2020 A survey of the usages of deep learning for natural language processing[J]. IEEE transactions on neural networks and learning systems, 32(2)Pp., 604-624.

Guo J., Xu G., Cheng X.,et al. 2009.,Named entity recognition in query[C]//Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.,267-274.

Barnes J, Velldal E, Øvrelid L. Improving sentiment analysis with multi-task learning of negation[J]. Natural Language Engineering, 2021, 27(2)Pp.,249-269.

Li.,qing, et al.,2021.,Enhancing Label Representations with Relational Inductive Bias Constraint for Fine-Grained Entity Ty." IJCAI. .

Huang Z,. Xu .,W.,Yu K. 2015 Bidirectional LSTM-CRF Models for Sequence Tagging [J]. arxiv preprint arxiv:1508.01991.

Zhang., Jiarui., et al.,"LSTM-CNN hybrid model for text classification." 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2018.

Zhang .,Y.,Yang J. Chinese NER using lattice LSTM[J]. arxiv preprint arxiv:1805.02023, 2018.

Li., X.,Yan H.,Qiu X, et al. FLAT., 2020 .,Chinese NER using flat-lattice transformer[J]. arxiv preprint arxiv:2004.11795,

Gui .,T.,Ma .,R.,Zhang Q.,et al.,2019 CNN-Based Chinese NER with Lexicon Rethinking[C]//ijcai. 2019.

Sui .,D.,Chen .,Y.,Liu K, et al., 2019 Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network[C]//Proceedings of the conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019: 3830-3840.

Ding.,R.,e .,P., Zhang.,X, et al. A neural multi-digraph model for Chinese NER with gazetteers[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 1462-1467.

Ma., R., Peng .,M, Zhang Q, et al. Simplify the usage of lexicon in Chinese NER[J]. arxiv preprint arxiv:1908.05969, 2019.

Caruana .,R.,Multitask learning[J]. Machine learning, 1997, 28: 41-75.

Fang.,Q., Li, Y., Feng, H., and Ruan, Y. Chinese Named Entity Recognition Model Based on Multi-Task Learning. Applied Sciences 13.8 (2023): 4770.

Xuetao .,T ,Xiaoxuan B ,Lu H . Multi-task learning with helpful word selection for lexicon-enhanced Chinese NER [J]. Applied Intelligence, 2023, 53 (16)Pp.,19028-19043.

Zhou .,G., D. 2006., Recognizing names in biomedical texts using mutual information independence model and SVM plus sigmoid[J]. International journal of medical informatics, , 75(6)Pp., 456-467.

Peng .,N., Dredze M.,2015.,Named entity recognition for chinese social media with jointly trained embeddings[C]//Proceedings of the conference on empirical methods in natural language processing. 2015: 548-554.

Levow .,G .,A., 2006 The third international Chinese language processing bakeoff: Word segmentation and named entity recognition[C]//Proceedings of the Fifth SIGHAN workshop on Chinese language processing.,108-117.

Devlin., J., 2018., BertPre-training of deep bidirectional transformers for language understanding[J]. arxiv preprint arxiv:1810.04805,

Diederik .,P .,K. Adam ., 2014A method for stochastic optimization[J]. (No Title)

Zhu., Y., Wang .,G, 2019., Karlsson B F. CAN-NER: Convolutional attention network for Chinese named entity recognition[J]. arxiv preprint arxiv:1904.02141,.