Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060422

ARTICLE

A Chinese Named Entity Recognition Method for News Domain Based on Transfer Learning and Word Embeddings

Rui Fang and Liangzhong Cui*

Naval University of Engineering, Wuhan, 430033, China

*Corresponding Author: Liangzhong Cui. Email: 0909071003@nue.edu.cn

Received: 31 October 2024; Accepted: 04 March 2025; Published: 16 April 2025

ABSTRACT: Named Entity Recognition (NER) is vital in natural language processing for the analysis of news texts, as it accurately identifies entities such as locations, persons, and organizations, which is crucial for applications like news summarization and event tracking. However, NER in the news domain faces challenges due to insufficient annotated data, complex entity structures, and strong context dependencies. To address these issues, we propose a new Chinese-named entity recognition method that integrates transfer learning with word embeddings. Our approach leverages the ERNIE pre-trained model for transfer learning and obtaining general language representations and incorporates the Soft-lexicon word embedding technique to handle varied entity structures. This dual-strategy enhances the model's understanding of context and boosts its ability to process complex texts. Experimental results show that our method achieves an F1 score of 94.72% on a news dataset, surpassing baseline methods by 3%–4%, thereby confirming its effectiveness for Chinese-named entity recognition in the news domain.

KEYWORDS: News domain; named entity recognition (NER); transfer learning; word embeddings; ERNIE; soft-lexicon

1 Introduction

With the advent of the information age, social media and news websites have become primary sources for individuals to access news and information. However, the massive influx of online news data, characterized by lengthy content and varying quality, has significantly exacerbated the problem of information overload. In response, various personalized recommendation systems based on news data streams have emerged, utilizing intelligent algorithms for filtering and pushing content, thereby helping users quickly and accurately capture useful information. This trend has garnered widespread research interest from academic circles. In this context, efficiently extracting the core content of news reports is particularly crucial. In the field of NLP, information extraction techniques have increasingly become essential for handling news text data. Information extraction encompasses multiple domains, with NER serving as a fundamental and critical component. NER is vital for identifying and extracting key entities (such as time, places, and organizations) within the text, providing a data foundation for subsequent applications like information analysis and event tracking. For example, in news recommendation systems, identifying key entities within news articles enables a more accurate understanding of user interests, thereby facilitating more personalized content recommendations. In event monitoring systems, NER aids in the swift identification and tracking of key individuals and locations in news events, enhancing the speed of event response and processing efficiency. In

intelligent question-answering systems, NER technology can be employed to comprehend key entities within user inquiries, enabling the retrieval of related information from a news database to provide answers.

In 2018, Yan et al. [1] proposed a politically named entity recognition method based on bi-directional LSTM combined with CRF for the problem of easy misrecognition of irrelevant entities in the field of political news, but the method is less effective in recognizing ambiguous words. Hu et al. [2] proposed a long and short-term memory neural network based on the attention mechanism combined with the conditional random field model (AttBi-LSTM-CRF) for entity recognition in the Sohu news corpus. The Bi-LSTM network can obtain long-term contextual information, and the Attention module is added to obtain the important information from the input related to the labeling of the output. Zheng et al. [3] designed a neural network model based on Bi-Directional GRU-CRF to collect Chinese news data from Chinese embassies in ten ASEAN countries and perform named entity recognition for constructing the ASEAN 10 knowledge graph. Wei et al. [4] proposed an opinion entity recognition model based on the pre-trained language model RoBERTa and multilayer residual BiLSTM-CRF architecture for the problem of scarce data as well as complex entities in specific domains.

There are two main problems in named entity recognition in the news domain. One is the relative scarcity of labeled data, which usually requires the construction of datasets. Commonly used named entity recognition datasets are Resume dataset, Weibo dataset, etc. However, the datasets for the news domain are relatively scarce, which limits the effect of model learning and training. Second, the context of news text is more complex, and the complexity of words and contextual information must be taken into account. For example, the problem of multiple meanings of words: "navigation" can be used as a noun navigation equipment, but also can be understood as the action of guiding the direction; there is also the problem of multiple synonyms, to improve the model's ability to understand the context is the key to improve the recognition performance. Deep learning models such as LSTM and GRU, although capable of capturing rich linguistic features, still have limitations when dealing with complex contexts. These models usually require large-scale labeled data for training to learn patterns and regularities in language. However, high-quality labeled data in the news domain is often scarce and difficult to obtain, limiting the further optimization and generalization capabilities of deep learning models.

Therefore, the research gaps are mainly in the following two aspects: first, how to improve the model's ability to understand the context, and second, how to reduce the dependence on large-scale labeled data and improve the model's generalization performance.

To solve the above problems, this paper proposes a Chinese-named entity recognition method based on migration learning and word embedding, with the following main innovations:

(1) Reduce the dependence on large-scale labeled data and improve the generalization ability of the model by using the language knowledge already learned by the pre-trained model through the transfer learning technique.
(2) A novel model architecture is constructed by fusing the ERNIE pre-trained model and the Soft-lexicon word embedding method. The model can be compatible with and integrate semantic information from different granularities, realize the fusion of word vectors and character vectors, and provide richer semantic information, which helps the model better capture the complex semantic relations of the text and then improve the accuracy and robustness of recognition.

2 DL-Based Recognition of Chinese Named Entities

NER in Chinese text is challenging due to its complex linguistic properties. Traditional rule-based or statistical methods struggle with various linguistic phenomena and text variants. Recently, deep learning

(DL) methods have surfaced as a promising technique in Chinese NER. DL models such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Transformers can automatically learn semantic information and patterns within texts, thus improving the accuracy and reliability of NER.

The DL-based recognition framework comprises three main layers: embedding, encoding, and decoding, as shown in Fig. 1. The embedding layer converts raw word and vocabulary information into a unified vector representation. Each word or phrase is mapped to a vector in a high-dimensional space, enabling the capture of semantic information. The encoding layer delves deeper into contextual information by using complex neural network structures such as RNNs or CNNs, which helps in understanding the context of sentences or paragraphs and enhances the comprehension of specific words within the text. The decoding layer identifies and extracts regular features between sequences and classifies the feature vectors outputted by the encoding layer. Techniques such as CRFs are utilized by the decoding layer to consider the entire sequence's information, leading to more accurate prediction and recognition of named entities.

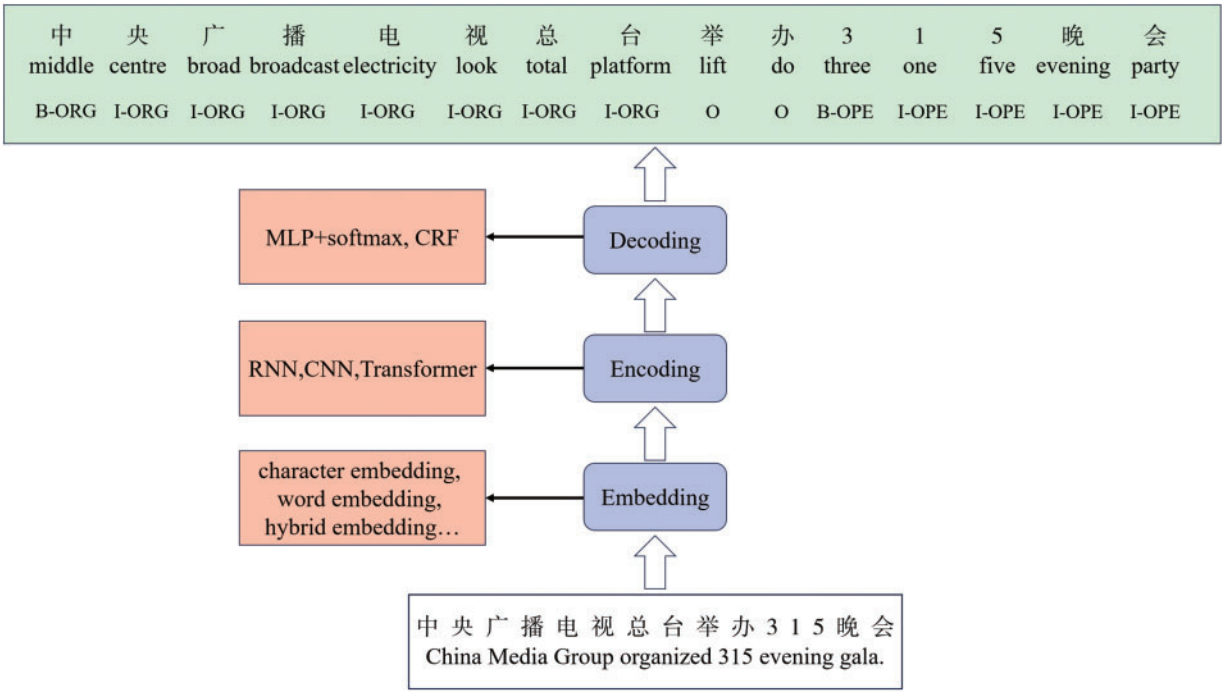

Figure 1: DL-based CNER model framework

2.1 Embedding Layer

In the embedding layer, original textual and lexical information is transformed into a unified vector representation to obtain semantic information. The embedding layer models can be classified into three types: character-based NER models, word-based NER models, and hybrid NER models.

2.1.1 Character-Based NER Models

Character-based NER models involve processing text by converting it into sequences of characters. This approach avoids errors caused by inaccurate Chinese text segmentation and can handle emerging vocabulary, pinyin, or entities with similar forms of Chinese characters.

Zhang et al. [5] designed an innovative, dynamic embedding strategy to address the issue of close connections between adjacent characters. This strategy integrates the attention method to optimize the

combination of character vector features and word vector features within the embedding layer. Sequence labeling strategies based on single-character features are widely used in Chinese NER tasks, making the optimization of character representation a key method for improving recognition effectiveness. Building on this, Luo et al. [6] proposed a Chinese character representation optimization method specifically for entity recognition tasks, which was implemented within the BiLSTM-CRF entity recognition framework. Experimental results show that this method significantly enhances entity recognition effectiveness, validating the efficacy of its character representation strategy. While character-based models demonstrate certain advantages in processing textual data, they still face challenges in directly conveying semantic information and effectively handling ambiguity issues [7].

2.1.2 Word-Based NER Models

Word-based models involve splitting Chinese text into words and preprocessing the text using a word-splitting tool. These models can recognize semantic links between words and perform well with long entity names. However, they suffer from word segmentation errors and struggle with irregular words and neologisms. To tackle the issue of ineffectively utilizing distant context information, Chen et al. [8] proposed a new neural network model designed explicitly for word segmentation. The model utilizes LSTM networks, which use memory units to retain previous information, thus overcoming the limitations of traditional fixed context window sizes. Ma et al. [9] combined Bi-LSTM, CNN, and CRF to create a hybrid neural network structure. This structure can automatically learn from word and character-level representations, achieving end-to-end processing for Named Entity Recognition (NER) without the need for feature engineering or data preprocessing, making it applicable to various sequence labeling tasks.

For NER tasks in Chinese electronic medical records, Zhang et al. [10] addressed the limitations of traditional methods that rely heavily on manual feature extraction by introducing word embedding techniques. They converted electronic medical record text sequences into vectorized representations containing semantic information. They also designed a composite model combining Bi-LSTM with CRF to enhance recognition accuracy further.

2.1.3 Hybrid NER Models

In the field of Chinese NER, in addition to methods that separately represent words and individual Chinese characters, researchers have explored hybrid representation techniques that combine character and word information, as well as representation strategies that introduce additional linguistic features. These strategies aim to enhance model performance but may reduce the model's generalization ability.

Word embedding representation techniques are categorized into structure fusion and data fusion. Structure fusion involves integrating word information into word-level representations by designing specific network structures, whereas data fusion utilizes multitask learning to process both word and phrase information simultaneously. Zhang et al. [11] integrated word information using an improved Lattice-LSTM network structure, enhancing the model's ability to handle long-distance dependencies and expanding the structure of the model from chain to graph. Liu et al. [12] proposed a new model WC-LSTM built upon Lattice-LSTM, solving the problem of Lattice-LSTM's inability to batch-process by providing single remote word information for each word node through four strategies. For data fusion, Cao et al. [7] used an adversarial transfer learning model to model words, words, and word-word fusion information by using three LSTM structures to achieve multitask learning and alleviate the problem of new word recognition. Furthermore, Wu et al. [13] used a CNN model to simultaneously perform word in-formation extraction and word segmentation with shared parameters.

Although incorporating additional linguistic features such as strokes and pinyin can improve the accuracy of Chinese NER, it may sacrifice the model's generalization ability. Then, pre-trained language models like BERT, XLNet [14], and ERNIE [15] have made significant progress by training on large-scale corpora and employing techniques such as positional coding.

In summary, this paper collectively refers to the above character fusion representation and the representation methods of adding additional features as the hybrid model, which can offer a more comprehensive and effective representation of Chinese NER.

2.2 Coding Layer

In Chinese NER, the coding layer converts text data from the embedding layer into high-dimensional feature vectors. These vectors aid subsequent classifiers in accurately classifying the text, helping to train the model to effectively represent text features and improve the accuracy of NER in texts. To extract text features and establish contextual relationships, the coding layer employs various network architectures such as CNNs, RNNs, and Transformers.

2.2.1 CNNs

CNNs are a widely used DL architecture originally designed for image recognition tasks. However, they also excel in the field of NLP, especially in capturing *n*-gram features in textual data. CNNs can extract abstract representations from localized text features through their convolutional layers, enabling them to understand dependencies in text.

To improve computing efficiency and fully utilize the parallel processing capability of GPUs, Gui et al. [16] proposed a CNN-based method. This method adopts a mechanism of rethinking and combines a dictionary to process sentence-matching problems parallelly. Experimental results demonstrate a significant improvement in recognition efficiency with this approach. Shi et al. [17] addressed the issues of polysemy and the time cost of dictionary matching in named entity recognition tasks by proposing a CNN-Head Transformer Encoder (CHTE) model. This model uses CNNs of various window sizes to generate value vectors of multiple attention heads within the Trans-former architecture. By doing so, it retains global semantic information, enhances local features, and represents potential word-level information, thereby improving Trans-former performance in named entity recognition.

2.2.2 RNNs

RNNs and their variants, such as LSTM and GRU, excel in processing serialized time-series data and can effectively capture temporal dynamic features in sequence data. Bidirectional RNNs utilize both forward and backward information in the text to gain a deeper understanding of contextual semantics.

Wu et al. [13] proposed the CNN-LSTM-CRF model in 2019, aiming to improve the effectiveness of Chinese NER by jointly training character-level named entity recognition (CNER) and word segmentation models. Sharing character embeddings and CNN network structures can effectively utilize segmentation information to enhance the ability to predict entity boundaries in the CNER task. In order to mitigate the gradient vanishing (exploding) problem of the LSTM model, Yang et al. [18] used a bidirectional GRU model for generic entity named entity recognition, and experiments showed that the bidirectional GRU model can capture text information over longer distances. Zhu et al. [19] designed the CAN network architecture, which integrates convolutional neural networks (CNNs) with attention mechanisms for the first time, aiming to accurately capture deep semantic features of locally adjacent characters in the text. At the same time, the CAN

network cleverly combines gated recurrent units (GRUs) with a global attention mechanism to effectively extract global semantic information from the text.

2.2.3 Transformer

The Transformer model differs from traditional RNNs and CNNs in that it incorporates an encoder with a self-attention mechanism and fully connected layers. The Transformer model has indeed demonstrated exceptional success in NLP and has significantly reduced the training time. Yan et al. [20] developed the TENER model, a novel NER architecture that employs an adaptive Transformer to extract both character and word-level features. Furthermore, Li et al. [21] introduced the FLAT model that transforms the lattice structure into a planar arrangement of spans. By leveraging the capabilities of the Transformer model and well-crafted positional encodings, it fully exploits lattice information and exhibits excellent parallel processing capability.

The advent of pre-trained models has further revolutionized NER. These models, pre-trained on massive text corpora, provide a robust foundation for fine-tuning specific NER tasks. Compared to other transfer learning approaches, pre-trained models offer several advantages, including efficient feature extraction and enhanced generalization, often resulting in faster convergence and superior performance. They are particularly effective in handling complex contextual dependencies and can be fine-tuned with relatively small amounts of labeled data, making them highly beneficial in scenarios where labeled data is scarce.

However, it is essential to acknowledge the limitations of pre-trained models. The substantial computational resources required for training and deployment can restrict accessibility, especially for resource-constrained environments. Moreover, domain-specific fine-tuning is still necessary to handle unique terminology and entities, as pre-trained models may not always generalize well to specialized domains. Despite these challenges, pre-trained models have markedly advanced the state-of-the-art in NER, serving as powerful tools for extracting meaningful entities from text data.

2.3 Decoding Layer

The decoding layer takes the contextual representation as input and generates a label sequence that accurately corresponds to the input text sequence. Currently, two main approaches are used for the decoding layer: multilayer perceptron (MLP) + Softmax and CRF.

2.3.1 MLP + Softmax

MLP + Softmax transforms the NER problem into a multiclassification problem by converting the context-rich feature vectors from the coding layer into a label sequence through the multilayer perceptron and Softmax layers. The advantage lies in its powerful nonlinear representation and learning ability for latent information. However, its disadvantage lies in the assumption that the labels are independent, whereas dependencies and rules often exist between labels, leading to potential information loss and affecting classification accuracy.

2.3.2 CRF

CRF is a discriminative probabilistic model that can directly model label dependencies in sequence annotation tasks and effectively address conflicts and ambiguities among labels. CRF models typically use predicted local label sequences to compute optimal global label sequences by modeling the dependencies among labels across the entire sequence, thereby enhancing the accuracy and robustness of sequence annotation.

Lample et al. [22] first introduced CRF as an output layer in a NER model and achieved outstanding performance across various datasets in diverse languages. Subsequently, numerous NER methods based on the CRF output structure have been proposed. Zhang et al. [23] proposed a NER method for Chinese electronic medical records, which jointly predicts medical record entities through the CRF layer, thereby avoiding the information loss problem that may arise from the label independence assumption.

3 Methodology

3.1 Model Structure

NER in scenarios with limited training data often involves the use of methods such as remote supervision and clustering. In this paper, transfer learning was employed to handle the issue of limited training data in scenarios with small sample sizes. The proposed model is based on ERNIE and utilizes an NER approach using transfer learning and word embedding. The model uses ERNIE to process the character vector input. Lexical information is then derived by constructing a lexicon by using a torch.nn.embedding. Next, the character vector and lexical information are combined, and a BiGRU is employed to extract both forward and backward contextual information related to the entity. As BiGRU does not consider the order relationship between tags (e.g., B-tags must precede I-tags), a CRF layer is incorporated to ensure that the output tags adhere to the correct tagging order, thereby improving prediction accuracy and recall. The proposed model comprises three main layers, and its overall structure is shown in Fig. 2.

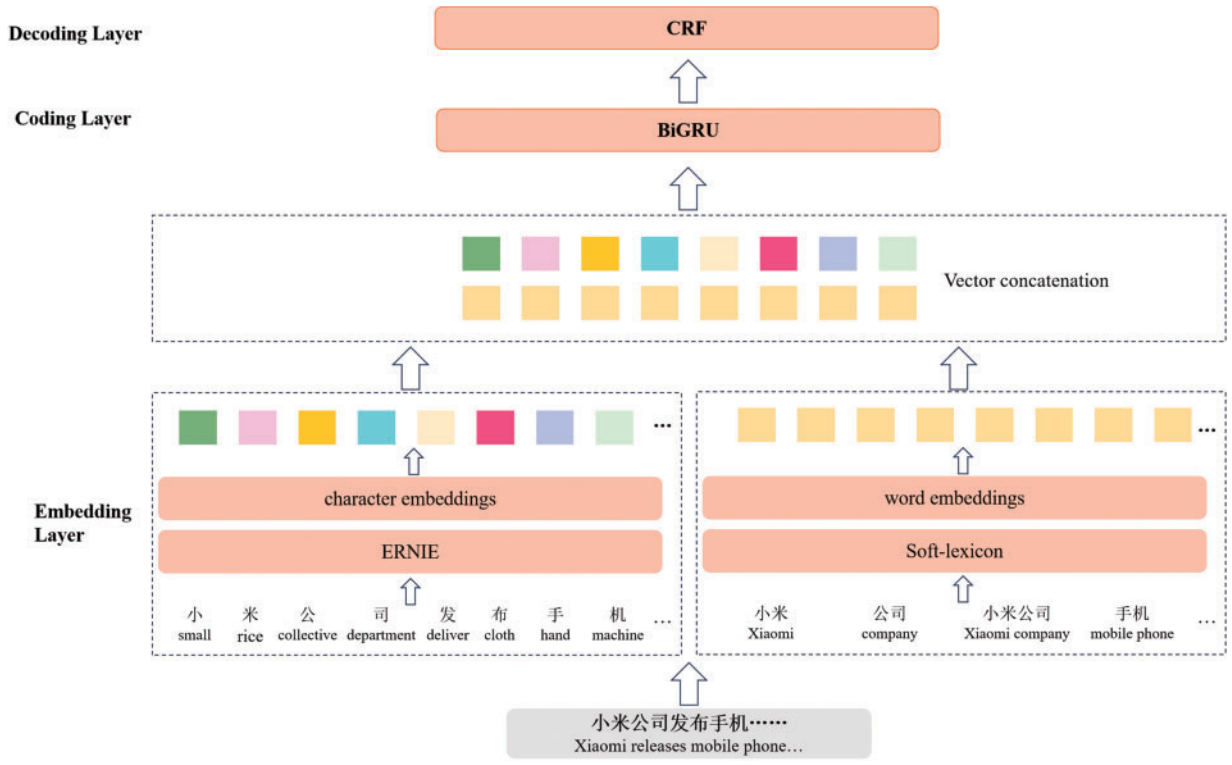

Figure 2: Overall structure of the model

3.2 Embedding Layer

The embedding layer consists of character embedding and word embedding. The ERNIE pretraining model is utilized to generate character vectors from the input. Lexical information of words is then

incorporated into the character representation by constructing a lexicon to enrich the character vector information and enhance the accuracy of NER.

3.2.1 ERNIE

ERNIE is a pre-trained model released by Baidu. Compared with BERT, it has revamped the masking strategy by introducing a hierarchical approach. In BERT, masking operates at the character level, while ERNIE divides masking into three granularities: character level, entity level, and phrase level. This hierarchical strategy enables ERNIE to simultaneously learn fine-grained character features and higher-level semantic structures, such as entity dependencies and syntactic phrase patterns. For example, when masking the entity "Beijing City" in a sentence, ERNIE masks the entire entity rather than individual characters, forcing the model to infer the entity's role based on its contextual relationships. This approach significantly enhances ERNIE's ability to model long-range semantic dependencies, making it particularly suited for news texts where entities often span multiple tokens and rely on distant context for disambiguation.

Additionally, ERNIE's training corpus integrates multi-domain knowledge, including Baidu Encyclopedia, news articles, and forum discussions. This diversity allows ERNIE to capture domain-specific entity patterns and adapt to varying linguistic styles. For instance, the term "Apple" might refer to the fruit in general texts but denote the company in business news. ERNIE's exposure to heterogeneous data helps it disambiguate such cases by leveraging entity-level masking and contextual cues from diverse corpora. This capability is critical for news domain NER, where entities frequently exhibit polysemy and domain-specific meanings.

ERNIE employs a multilayer Transformer architecture as its core encoder. For a sentence $s = \{c_1, c_2, \cdots, c_n\}$ consisting of n tokens (characters), unique tokens [CLS] and [SEP] are appended to signify the start and end of a sentence. As shown in Fig. 3, each token is represented as embeddings by the summation of Token Embeddings, Segment Embeddings, and Position Embeddings, which can be expressed as $E_{c_i} = E_{token} + E_{seg} + E_{pos}$. The sentence is then converted into a vector sequence and fed into the bidirectional Transformer to obtain features. Unlike BERT, which focuses on character-level masking, ERNIE's self-attention mechanism prioritizes entity and phrase-level relationships. For example, in the sentence "The CEO of Xiaomi Corporation announced a new product", ERNIE's attention heads would strengthen connections between "Xiaomi Corporation" and "CEO", thereby improving entity coherence recognition. By preserving contextual meaning across extended text spans, ERNIE reduces misclassification risks caused by fragmented or ambiguous contextual clues.

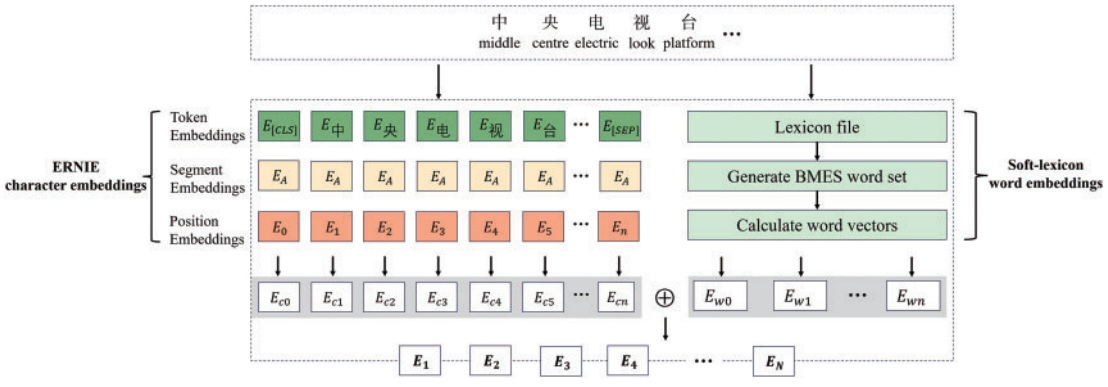

Figure 3: Embedding layer

3.2.2 Soft-Lexicon

Soft-lexicon is an effective method for integrating word information into character representations. It eliminates the need to design complex sequence modeling structures by making subtle adjustments to the character representation layer to incorporate dictionary information. In the embedding layer, each character is transformed into a high-dimensional vector, and soft dictionary features are constructed, which are then fused with the representations of each character. This approach effectively addresses the limitations of models that rely solely on character information in leveraging lexical data, while also achieving superior performance without compromising inference speed, making it more compatible with pre-trained models.

As shown in Fig. 4, to use the dictionary information, for the input text $S = (c_1, c_2, \cdots, c_n)$, we processed each character c_i by matching it with the words in the dictionary and categorized the characters that matched the words into the corresponding sets.

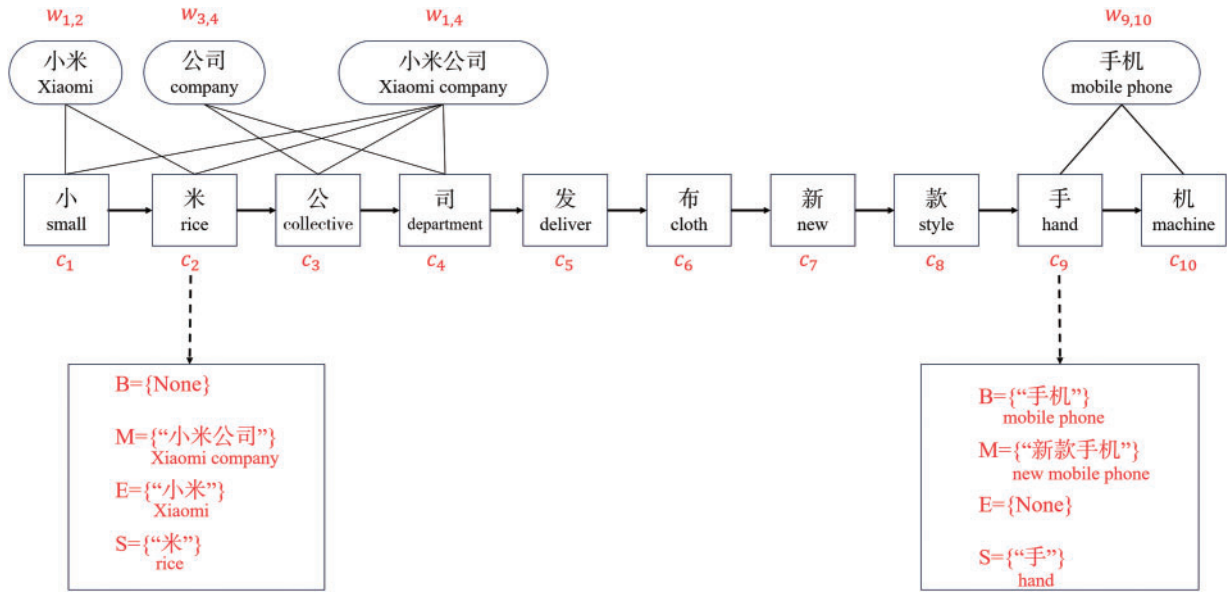

Figure 4: Soft-lexicon method

For the character c_i, the four sets formed are shown in Eq. (1):

$$\begin{aligned} B(c_i) &= \{w_{i,n}, \forall w_{i,n} \in W, i < n \leq k\} \\ M(c_i) &= \{w_{m,n}, \forall w_{m,n} \in W, 1 \leq m < i < n \leq k\} \\ E(c_i) &= \{w_{m,i}, \forall w_{m,i} \in W, 1 \leq m < i\} \\ S(c_i) &= \{c_i, \exists c_i \in W\} \end{aligned} \tag{1}$$

where W represents the lexicon, $w_{m,n}$ denotes the words in the lexicon (starting with c_m and ending with c_n), $B(c_i)$ denotes the words set starting with c_i, $M(c_i)$ denotes the words set with c_i as the middle character, $E(c_i)$ denotes the words set ending with c_i, and $S(c_i)$ denotes the set containing only a single character c_i. If no word is found in the lexicon that matches the $\{B, M, E, S\}$ structure, the set is empty.

After obtaining the character word set, the feature vectors of each class of word set for characters are calculated by counting the frequency of each word in the word set using the weighted average as shown

in Eq. (2):

$$v^s(E) = \frac{4}{Z} \sum_{w \in E} z(w)\, e^w(w) \tag{2}$$

where w is the word element in the word set E (the same for word sets B, M, and S), $v^s(S)$ represents the weight of the word set E, $z(w)$ represents the frequency of word w in the statistical word set, $e^w(w)$ represents the vector corresponding to word w in the word vector lookup table, and Z is the sum of the frequency of word w in the corresponding four types of word sets, and the formula of Z is as follows:

$$Z = \sum_{w \subseteq B \cup M \cup E \cup S} z(w) \tag{3}$$

Then the corresponding word set vectors of the character c_i are spliced and added to each character as shown in Eq. (4):

$$\begin{aligned} e^s(B, M, E, S) &= \{v^s(B), v^s(M), v^s(E), v^s(S)\} \\ x^c &\leftarrow \{x^c, e^s(B, M, E, S)\} \end{aligned} \tag{4}$$

This fusion enables the model to capture character-level details and word-level semantics simultaneously. For instance, in the ambiguous phrase "苹果手机" (Apple phone), the Soft-lexicon embedding for "苹" would incorporate weights from both the fruit-related ("苹果"—apple) and brand-related ("苹果公司"—Apple Inc.) contexts, aiding the model in disambiguating the entity type based on surrounding words.

By dynamically adjusting lexical contributions, Soft-lexicon enhances robustness to out-of-vocabulary terms and nested entity structures, which are prevalent in news texts. This adaptability complements ERNIE's contextual representations, creating a synergistic effect that improves both entity boundary detection and type classification.

3.3 Coding Layer

In this paper, the BiGRU serves as the coding layer. The GRU is a type of RNN that can capture the relationship between the current element and the preceding elements in a sequence, enabling the network to memorize contextual content, thus making it suitable for NER. Both LSTM and GRU are model variants of RNN, with GRU having a simpler structure than LSTM, thus reducing the amount of training required by simplifying the gate structure and delivering superior performance. Unlike conventional RNNs, which struggle with long-range dependencies due to vanishing gradients, BiGRU processes text bi-directionally, allowing the model to retain both past and future contextual information. This bidirectional structure is particularly beneficial for news texts, where entity meanings often depend on words appearing both before and after the target entity. By leveraging BiGRU, our model ensures that entity recognition is not solely reliant on immediate neighboring words but instead considers a more comprehensive semantic scope. The internal structure of GRU can be expressed as follows:

$$\begin{aligned} z_m &= \sigma\left(W_z \cdot [h_{m-1}, x_m]\right) \\ r_m &= \sigma\left(W_r \cdot [h_{m-1}, x_m]\right) \\ \tilde{h}_m &= \tanh\left(W_{\tilde{h}} \cdot [r_m * h_{m-1}, x_m]\right) \\ h_m &= (1 - z_m) * h_{m-1} + z_m * \tilde{h}_m \end{aligned} \tag{5}$$

where x_m represents the input of current time; h_{m-1} denotes the hidden state from the previous time step, serving as the neural network memory containing data from the prior node; h_m signifies the hidden state

passed to the next time step; $\tilde{h}_m$ denotes the candidate hidden state; r_m and z_m represent the reset and update gates; σ represents the sigmoid function; and the tanh function converts its input into the value in the range of $[-1, 1]$.

3.4 Decoding Layer

We used the CRF layer as the decoding layer. The CRF layer performs global label inference on the feature sequences outputted by the BiGRU coding layer to model consistency across the entire sequence. It considers dependencies between labels to better capture both local and global information in the sequence labeling task. Unlike sequence classification methods that assign labels independently to each token, CRF enforces label coherence by modeling interdependencies between adjacent tokens. This mechanism is particularly effective in handling entity boundaries and reducing errors where contextual cues play a crucial role in classification. The CRF decoding process is shown in Fig. 5.

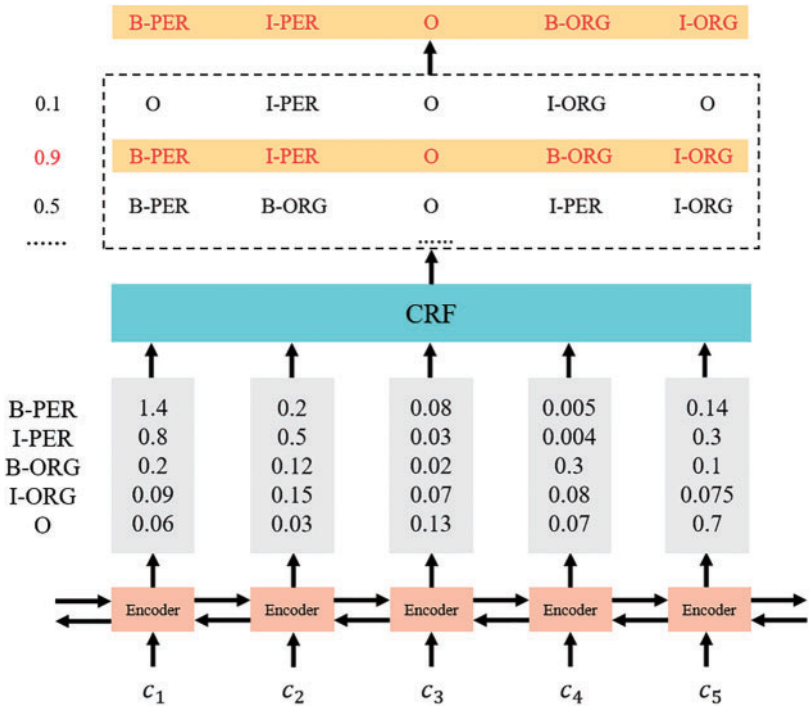

Figure 5: The CRF decoding process

The output sequence yielded by BiGRU is $X = \{h_1, h_2, \cdots, h_i\}$, and the score function for each input value at each position is calculated as follows:

$$S(h, y) = \sum_{i=1}^{N} P_{i,y_i} + \sum_{i=1}^{N} R_{y_i, y_{i+1}} \tag{6}$$

where N is the total length of the sequence, P_{i,y_i} denotes the probability value of the i-th sequence output label in the sequence to be y_i, and $R_{y_i,y_{i+1}}$ denotes the transfer probability from character label y_i to character label y_{i+1}.

It is then normalized using the Softmax to compute the probability value of the labeled sequence y:

$$P(y|h) = \frac{\exp[S(h, y)]}{\sum_{y' \in Y(h)} \exp[S(h, y')]} \tag{7}$$

where $Y(h)$ denotes the set of all possible labeling sequences.

4 Experiments

4.1 Datasets

The text uses the Chinese-named entity recognition dataset from SIGHAN Bakeoff 2006, sourced from People's Daily News data. It is labeled based on MSRA's rules, categorizing entities into organizations, locations, and person names. The labeling format follows the BIO method, with "B" for the first character, "I" for subsequent characters, and "O" for irrelevant characters. And seven types of labels have been annotated: "B-PER", "I-PER", "B-LOC", "I-LOC", "B-ORG", "I-ORG" and "O".

The labeling process was done manually, the dataset originally contained labeled content, which was manually reviewed and corrected to ensure the accuracy and consistency of the labeling. The annotation work relies heavily on the expertise and experience of the annotator, and thus there will be a certain degree of subjectivity and variability. In order to minimize this effect, multi-person annotation and cross-validation are used to improve the reliability and stability of the annotation results.

Meanwhile, the public datasets Resume and Weibo are selected as open test sets to verify the model's generalization ability. The details of the datasets are shown in Table 1.

Table 1: Datasets details

Datasets	Type	Train	Dev	Test
News	Sentences	19.7 k	2.2 k	4.3 k
	Characters	100.0 k	112.2 k	223.8 k
Resume	Sentences	3.8 k	0.46 k	0.48 k
	Characters	124.1 k	13.9 k	15.1 k
Weibo	Sentences	1.4 k	0.27 k	0.27 k
	Characters	73.8 k	14.5 k	14.8 k

4.2 Performance Measures

Named entity recognition is a multiclassification problem in which each element (e.g., word or character) needs to be correctly categorized as either an entity or a non-entity, as well as a specific entity type. Precision (P), Recall (R) and F1 Score (F1) are three evaluation metrics always used in classification problems, which measure the performance of the model from different perspectives, respectively; P reflects the reliability of the model in predicting as a positive class (i.e., an entity) and ensures the quality of the prediction results; whereas R reflects the model's ability to recognize all the actually existing entities, ensuring that no important information is missed; the $F1$ metric combines P and R to reflect the comprehensive performance of the model. In previous studies, the three metrics have also been widely used as evaluation metrics for named entity recognition, so this paper chooses these three metrics as the evaluation metrics for the model.

P refers to the proportion of truly positive samples among all samples predicted as positive by the model, measuring the reliability of the model in predicting positive samples. A model with high precision means that most of its predicted positive samples are indeed positive.

Recall, also referred to as the True Positive Rate (TPR) or Sensitivity, measures the proportion of actual positive samples that the model correctly identifies. A model with high recall means it can identify the most positive samples.

*F*1 is the harmonic mean of P and R, balancing between precision and recall to reflect the model's whole performance. A model with a high *F*1 Score performs well in both precision and recall.

The formulas for calculating P, R, and *F*1 are shown in Eqs. (8)–(10).

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{8}$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{10}$$

4.3 Experimental Environment

Details of the experimental environment are presented in Table 2.

Table 2: Experimental environment

Experimental environment	Configure
Operating system	Windows 10
CPU	Intel(R) Core(TM) i7-8700K CPU
Random access memory (RAM)	16 GB
Display card (computer)	NVIDIA GeForce RTX 3090
Display memory	24 GB
Development environment	Python 3.9
Development framework	PyTorch 1.13.0+cu116

4.4 Experimental Parameters

In the experiments of this paper, the hyperparameter settings of the training process are the same as those of the Soft-lexicon algorithm, such as character embedding size, vocabulary embedding size, dropout rate, and other parameters, while the hidden dimension is determined by gradually adjusting the number of hidden layer nodes through multiple experiments and observing the performance changes of the model to determine the optimal hidden dimension. The learning rate is adopted as an initial learning rate and dynamically adjusted according to the convergence and performance of the model during the training process.

To ensure experimental reproducibility, we used a fixed random seed for all training runs. This ensures that the reported results are consistent and replicable across multiple executions.

The specific values are shown in Table 3.

Table 3: Parameter values

Parameter	Value
Char embedding size	30
Gas embedding size	50
Char hidden dim	50
Hidden dim	300
Gaz dropout	0.5
Learning_rate	0.0015
Learning_rate decay	0.05
Dropout	0.5
Epoch	30

As depicted in Fig. 6, there is a discernible upward trend of the evaluated metrics as the hidden dimension increases. When the hidden dimension exceeds 300, the performance of the model starts to decrease. It indicates that when the hidden dimension is too large, it may lead to an overfitting problem, so the value of the hidden dimension is selected as 300 in the experiment.

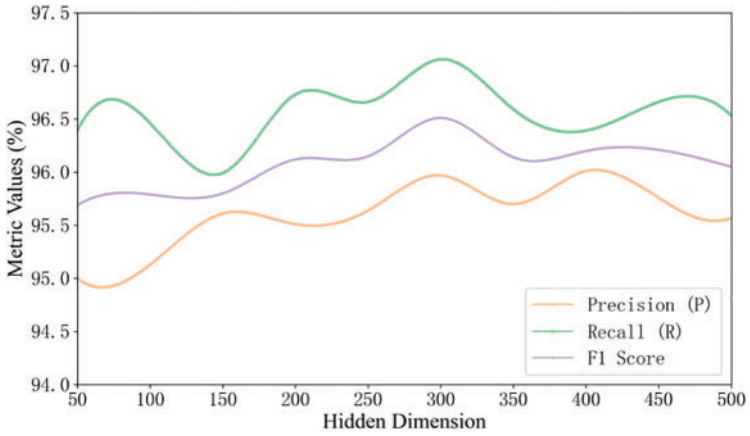

Figure 6: Effect of hidden dimension on model metrics

4.5 Experimental Design

Experiments are conducted in three main aspects to validate the effectiveness of our method for NER in the news domain.

The first part designs comparison experiments to train and test this paper's method and each baseline method on datasets News, Resume, and Weibo, respectively; records and compares the metrics of different methods on each dataset; and analyzes the performance advantages of our method over the baseline method and the reasons for them.

In the second part, ablation experiments are designed to set up ablation experiments for the two sub-modules of the pre-trained model and word embedding, and the complete model and ablation model are trained and tested on the news dataset, respectively; the metrics of the different models are recorded and

compared; and the effects of the pre-trained model and word embedding sub-module on the performance of NER are analyzed.

Section 3 performs computational efficiency analysis, which focuses on analyzing the computational efficiency of the model by recording the time required for one round of training the model on the news dataset.

In terms of baseline method selection, the main ideas are as follows:

This paper aims at named entity recognition by means of sequence annotation. The core idea of sequence annotation, a technique widely used in NLP, is to assign one or more labels to each element (e.g., character, word) in a text sequence to indicate its role or attribute in a specific task (e.g., named entity recognition, lexical annotation, event extraction, etc.). Therefore, the following classical sequence annotation methods are selected as a baseline to evaluate the effectiveness of our method. The specific methods are shown as follows:

Transformer: The Transformer model leverages self-attention mechanisms to capture long-distance dependencies within the text, demonstrating strong sequence modeling capabilities, making it suitable for tasks such as NER that require contextual understanding.

Lattice LSTM: Lattice LSTM addresses the unique challenges of processing Chinese text by incorporating lexical-level information, which compensates for the shortcomings of character-level models concerning segmentation errors. Its lattice structure permits simultaneous consideration of various combinations of characters and lexicons during the decoding process, thereby enhancing recognition accuracy. LR-CNN: The LR-CNN method employs CNN to extract local textual features and optimizes the weighting of lexical features through a lexical reconsideration mechanism. This improves the model's ability to extract key information and enables more accurate identification of relevant entity information.

WC-LSTM: WC-LSTM integrates lexical information using four encoding strategies: Shortest Word First, Longest Word First, Average, and Self-Attention. In comparison to Lattice LSTM, it partially addresses the issue of information loss and offers improved operational efficiency.

FLAT: The FLAT model combines the Transformer's encoding capabilities with the Lattice LSTM's lexical integration concepts. Through its flat lattice structure, it effectively merges lexical information and facilitates efficient computation. In NER tasks, the FLAT model fully utilizes the Transformer's global modeling strengths and the lattice structure's lexical fusion advantages, thereby improving recognition accuracy.

BERT, ALBERT, LERT: The pre-trained model is able to extract the contextual representation of the text and transform it into high-quality word vectors. These word vectors contain rich semantic information, which helps in the subsequent named entity recognition task. Combined with the sequence modeling capability of BiLSTM and the label inference capability of CRF, the model is able to better handle the sequence annotation task in natural language processing.

The above baseline methods are all publicly reproducible and have been successfully applied to NER tasks. They share commonalities in their ability to process textual data and have all explored the effectiveness of information fusion. A comparison with the information fusion of the methods in this paper will better demonstrate the strengths and improvements of the new methods over these baselines.

5 Results

5.1 Comparative Experimental Results

In this experiment, eight methods, Transformer, Lattice LSTM, LR-CNN, WC-LSTM, FLAT, BERT, ALBERT, and LERT, are selected as the baseline to compare with our method. The P, R, and F1 obtained

by several methods are detailed in Tables 4–6, and Fig. 7 shows the change of metrics with the number of training epochs on the news dataset.

Table 4: Results on news dataset

Model	P	R	F1
Transformer	84.23	79.84	81.98
Lattice LSTM	91.63	89.66	90.64
LR-CNN	91.78	91.62	91.70
WC-LSTM	92.47	91.36	91.91
FLAT	90.24	91.09	90.66
BERT-BiLSTM-CRF	94.02	93.49	93.76
ALBERT-BiLSTM-CRF	92.55	91.60	92.07
LERT	94.46	93.63	94.04
Ours	**95.07**	**94.37**	**94.72**

Table 5: Results of resume dataset

Model	P	R	F1
Transformer	91.48	91.60	91.54
Lattice LSTM	94.49	93.62	94.05
LR-CNN	94.30	94.36	94.33
WC-LSTM	95.08	94.91	95.00
FLAT	91.20	93.44	92.30
BERT-BiLSTM-CRF	94.60	95.64	95.12
ALBERT-BiLSTM-CRF	94.73	94.85	94.79
LERT	95.03	95.09	95.06
Ours	**95.74**	**96.63**	**96.18**

Table 6: Results of the weibo dataset

Model	P	R	F1
Transformer	55.44	52.90	54.14
Lattice LSTM	65.93	50.48	57.18
LR-CNN	60.98	50.97	55.53
WC-LSTM	68.42	34.54	45.91
FLAT	60.93	56.70	58.74
BERT-BiLSTM-CRF	67.44	63.53	65.42
ALBERT-BiLSTM-CRF	70.23	58.70	63.95
LERT	70.62	66.18	68.33
Ours	**72.97**	**67.15**	**69.94**

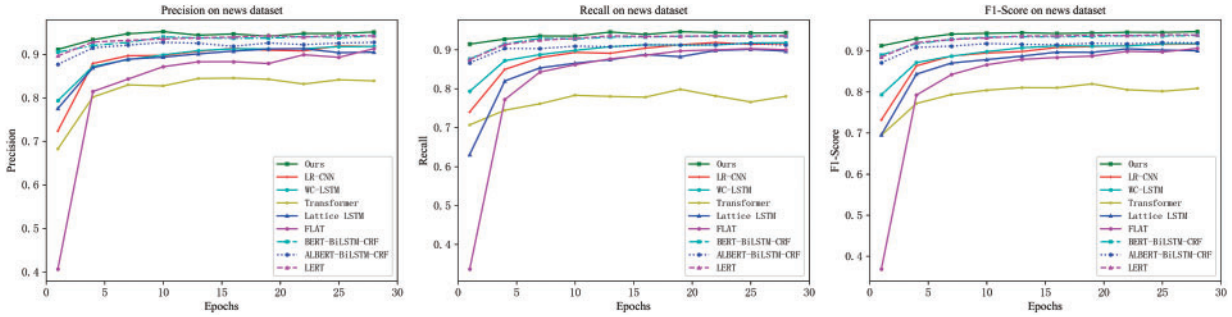

Figure 7: Changes in model metrics on news dataset

It can be concluded that this paper's model achieves the highest F1 value on three datasets, which are 94.72% (news dataset), 96.08% (Resume dataset), and 69.94% (Weibo dataset), which indicates that this paper's model has high accuracy and recall in recognizing named entities. The improvements can be attributed to the following factors:

(1) Soft-lexicon enhances entity recognition beyond self-attention mechanisms

Transformer-based models (BERT, ALBERT, LERT) leverage powerful contextual embeddings, but our method surpasses them. This indicates that explicit lexical information from Soft-lexicon provides an additional boost to entity recognition, especially for ambiguous terms.

(2) Pretrained knowledge combined with word-level features improves generalization

Compared to recurrent models (Lattice LSTM, WC-LSTM), our model improves the F1-score by 3%–4% on the news dataset, showing that integrating ERNIE-based transfer learning with Soft-lexicon word embeddings effectively captures diverse entity structures.

(3) The hybrid approach captures both character-and word-level semantics

The improvement over FLAT and LR-CNN suggests that our model's ability to combine character-based and word-based representations enhances robustness, especially for handling nested entities and contextual ambiguities.

These results confirm that our method effectively balances semantic representation learning (via ERNIE), contextual feature extraction (via BiGRU), and structured label inference (via CRF), leading to superior performance in Chinese NER.

5.2 Ablation Experimental Results

To investigate how transfer learning and word embeddings affect the NER, we conducted ablation experiments, and the results are presented in Table 7. BiGRU-CRF was used as the baseline model, with no transfer learning and word embeddings. Other configurations included "No ERNIE", indicating the use of only word embeddings and no transfer learning; "No Soft-lexicon", indicating the use of only transfer learning and no word embeddings; and the original model used in this paper.

Table 8 shows the results of the ablation experiments on ORG entities in the news dataset, which contains only one entity type, ORG.

In Table 7, we find that the BiGRU-CRF model without two sub-modules has precision, recall, and F1 scores of 86.47%, 80.74%, and 83.50%, respectively, which are significantly lower than the original model's scores of 95.07%, 94.37%, and 94.72%. The No ERNIE model, which is the model without the ERNIE module, shows a decrease of 3.23% in precision, 3.38% in recall, and 3.31% in F1 scores compared to the original model.

The No Soft-lexicon model, which is the model without the word embedding module, has precision, recall, and F1 scores of 92.87%, 92.78%, and 92.83%, respectively. Although there is a small decrease for the original model, the magnitude is not significant.

Table 7: Results of ablation experiment

Model	P	R	F1
Original model	95.07	94.37	94.72
No ERNIE	91.84	90.99	91.41
No Soft-lexicon	92.87	92.78	92.83
BiGRU-CRF	86.47	80.74	83.50

Table 8: Results of ablation experiment on news dataset (ORG)

Model	P	R	F1
Original model	91.52	89.13	90.31
No ERNIE	78.45	76.65	77.54
No Soft-lexicon	90.36	86.38	88.32
BiGRU-CRF	72.98	68.39	70.61

As can be seen from Table 8, ERNIE has the most significant impact on ORG entities, and the F1 value decreases by 12.77% after removing ERNIE. In comparison, the F1 value decreases by 1.99% after removing Soft-lexicon. ORG entities are usually the names of organizations or institutions, which have a more complex structure, involve multiple domains, and are more semantically diverse. ERNIE learns a large amount of linguistic knowledge and contextual information and is able to better capture the complex structure and semantic diversity of ORG entities. In contrast, Soft-lexicon provides additional semantic support for characters through lexical information, which is slightly less effective in targeting semantic diversity.

The results indicate that the performance of the model decreases when transfer learning and word embeddings are removed. It confirms the effectiveness of both components in enhancing NER performance. The removal of ERNIE resulted in a significant drop in performance across precision, recall, and F1 scores. This decrease underscores ERNIE's role in capturing long-range semantic dependencies and its ability to leverage multi-domain knowledge effectively, offering richer contextual representations that are crucial for distinguishing complex entities. Conversely, removing Soft-lexicon also led to a noticeable decline in performance, though less severe than the absence of ERNIE. This component's contribution lies primarily in its dynamic integration of lexicon information, which enhances the model's ability to disambiguate polysemous entities by embedding rich lexical features into character-level representations.

When either component is removed, the model struggles more with maintaining high accuracy in recognizing semantically diverse entities. This interaction between ERNIE's advanced contextual understanding and Soft-lexicon's lexical integration suggests a synergistic effect. Together, they complement each other; ERNIE strengthens the semantic backbone required to interpret broader context, while Soft-lexicon fills in the specific lexical gaps, ensuring comprehensive entity comprehension. This synergy is crucial for handling the variability and complexity inherent in news texts, as evidenced by the more significant performance drop when either is omitted. Therefore, while ERNIE alone significantly enhances model generalizability through transfer learning, the inclusion of Soft-lexicon ensures that specific contextual nuances are not overlooked, providing a more robust named entity recognition framework.

5.3 Comparative Results of Training Time

In order to compare the computational efficiency of different models, the time required to train one round of each model on the news dataset is recorded, and the training time of the FLAT model with the highest computational efficiency is taken as the benchmark, which is noted as 1, and the rest of the models' training time is used as a ratio with it, and the results are shown in Fig. 8. The larger the ratio, the longer the time required for a training round and the lower the efficiency.

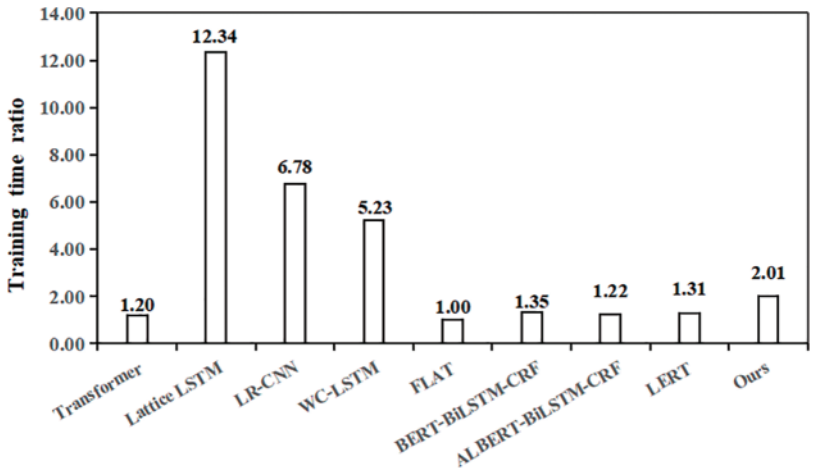

Figure 8: Comparative results of training time

The results reveal three key observations:

1. Traditional Architectures: Models like Lattice LSTM, LR-CNN, and WC-LSTM demand significantly longer training times, primarily due to their reliance on sequential processing and manual feature engineering.
2. Pre-trained Models: BERT, ALBERT, and LERT achieve comparable efficiency, as their Transformer-based architectures benefit from parallelized self-attention computations.
3. Our Method: The integration of ERNIE and Soft-lexicon embeddings introduces additional complexity, resulting in a higher training time ratio (2.3) compared to standalone pre-trained models. This overhead stems from two factors:

ERNIE Fine-tuning: The Transformer architecture requires backpropagation through deep neural pathways.

Soft-lexicon Encoding: Dynamic word set construction for each character.

While our method sacrifices some computational efficiency, this trade-off is justified by its superior recognition accuracy (Section 5.1). For latency-sensitive applications, future work could adopt lightweight pre-trained models (e.g., ALBERT) or optimize Soft-lexicon encoding through cached dictionaries.

6 Discussion

6.1 Discussion of Innovations

We will explore the innovations of this article from the following aspects:

(1) The innovation in model architecture

We have integrated the ERNIE pre-trained model with the Soft-lexicon method for word embedding to address the NER task in the news domain. This combination leverages the strengths of both approaches, resulting in a synergistic effect and achieving superior performance. The introduction of the ERNIE pre-trained model significantly boosts the model's contextual representation capacity and generalization ability.

In comparison to BERT-like models, such as those employed by Zhao et al. [24] (BERT-BiLSTM-CRF framework) and Guo et al. [25] (BERT-based dictionary-enhanced model called BERT-Pointer), which achieved F1 scores of 95.59% and 95.69% respectively on a Resume dataset, our method utilizing ERNIE achieves an even higher F1 score. This demonstrates the effectiveness of ERNIE in capturing nuanced contextual information relevant to NER tasks.

While many models, including ERNIE, utilize fine-tuning over pre-trained representations to adapt to specific tasks, pre-trained models bring several advantages over earlier transfer learning methods. Earlier methods relied heavily on handcrafted features or domain-specific data engineering. In contrast, pre-trained models like ERNIE benefit from being trained on large, diverse datasets, encapsulating a wide range of linguistic and commonsense knowledge. This comprehensive pre-training allows ERNIE and similar models to generalize well across different tasks, providing a strong foundation that requires minimal additional fine-tuning to perform effectively in new domains.

BERT, as a forerunner in self-attention-based models, captures contextual information primarily through self-attention mechanisms [26]. ERNIE builds upon this by integrating knowledge about entities and relations, allowing it to understand context and semantic relationships more deeply [27]. Furthermore, while BERT tackles character-level issues in languages like Chinese, potentially missing precise boundary identification, ERNIE is designed with optimizations such as multi-level masking, capturing character, phrase, and entity details to better align with the language's inherent characteristics [28].

The Soft-lexicon method enhances our architecture by sequentially obtaining all vocabulary sets corresponding to BMES (Beginning, Middle, End, Single) tagging for the current character and encoding these representations. This enables a more comprehensive utilization of lexical information, capturing the boundary and semantic information of entities more accurately and improving NER task performance. Additionally, the Soft-lexicon mechanism reduces domain-specific dependency by dynamically incorporating lexicon features without rigid architectural constraints, compensating for potential gaps in pre-trained knowledge.

Our proposed architecture is not limited to ERNIE; its modular design allows for the integration of any pre-trained encoder combined with Soft-lexicon embeddings. This adaptability is evidenced by studies [24] showing similar frameworks successfully employing models like RoBERTa or ALBERT. This level of flexibility distinguishes it from traditional methods, offering significant advantages in handling diverse datasets and linguistic features.

Moreover, our hybrid embedding approach, even without ERNIE, achieves impressive results. As shown in Table 7, the model attains an F1 score of 91.41% on news data, surpassing traditional LSTM-CRF baselines by 7.91%. This validates the standalone efficacy of our Soft-lexicon method in capturing lexical information and improving NER performance. Overall, our innovative architecture leverages the strengths of pre-trained models and Soft-lexicon embeddings to achieve superior results in NER tasks, demonstrating its robustness and adaptability across different domains and datasets.

Table 9 presents a comparative analysis of the Soft-lexicon method against several other approaches, highlighting its superior performance in NER tasks. This table underscores the profound impact that word embeddings have had on advancing NER technologies. By providing dense vector representations, word embeddings, particularly as used in Soft-lexicon, enable the model to grasp complex semantic relationships

within the text. Incorporating these rich representations allows Soft-lexicon to surpass traditional methods, achieving significantly improved precision and recall. This advancement reflects a broader trend in NER, where embedding techniques enhance model generalization and robustness across various linguistic contexts. The results exemplified in Table 9 not only affirm the effectiveness of Soft-lexicon but also illustrate the transformative role that word embeddings play in the ongoing evolution of NER systems. As models continue to evolve, leveraging such embeddings will be crucial for addressing the increasingly sophisticated demands of textual data analysis.

Table 9: Comparison of methods

Type	Method	Advantage	Shortcoming
Dynamic architecture: Design corresponding structures to incorporate lexical information	Lattice LSTM	Integrating lexical information through Lattice structure to improve recognition accuracy	Low inference efficiency and inability to capture long-range dependencies; a problem of information loss, and the complex structure limit the calculation speed
	LR-CNN	Using CNN to extract local features of text; optimize the weight of vocabulary features through the rewiring mechanism, and enhance the ability to capture key information	Feedback layers need to be added to each CNN layer to adjust the weights of the lexicon attention module and increase model complexity
	FLAT	Effective fusion and efficient computation of vocabulary information through Flat Lattice structure	Need to design position vectors to introduce vocabulary information, increasing the complexity of model design
Adaptive embedding: Model independent, with transferability	WC-LSTM	Integrating lexical information through four encoding strategies can to some extent solve the problem of information loss	The problem of information loss still exists; relying on LSTM for encoding limits the flexibility of the model.
	Soft-lexicon	Build soft lexicon features and add them to the representation of each character; without causing information loss, it can also introduce word embeddings	Each character's word set needs to be encoded.

(2) Comparison of experimental results

Fig. 9 shows the comparison of the metrics of the different models on the three datasets.

Figure 9: Comparison of results on three datasets

The comparison of results reveals that our model outperforms baseline methods in terms of P, R, and F1 scores across different datasets (news, Resume, Weibo). Relying solely on dictionary-based methods, such as Lattice LSTM and WC-LSTM, can assist in named entity recognition tasks; however, their performance is constrained by the completeness of the dictionary, resulting in limited semantic information. The ablation experimental results indicate that removing the word embedding module led to declines of 2.2%, 1.59%, and 1.89% in P, R, and F1, respectively, compared to the original model. Furthermore, eliminating the pre-trained model resulted in more significant decreases of 3.23%, 3.38%, and 3.31% in P, R, and F1, respectively. It suggests that singular feature representation can restrict the model's ability to handle unknown vocabulary or complex semantic relationships. Therefore, integrating multiple information sources and technical approaches can facilitate a complementary advantage.

Consequently, some researchers have combined unlabeled data with dictionary information to introduce additional prior knowledge, achieving better results with smaller dataset sizes; this approach also demonstrates improved performance when integrated with pre-trained models [29]. Additionally, other studies have incorporated multi-granularity information from characters, dictionaries, and entities into the BERT model, significantly enhancing its ability to handle complex contexts and ambiguities [30]. By integrating external knowledge from the ERNIE pre-trained model, which encompasses entity and phrase-level knowledge, knowledge from knowledge graphs, contextual information, and insights from multi-source training data [27], the proposed model can learn more accurate and enriched semantic representations, thus enhancing its overall effectiveness.

(3) Qualitative insights into contextual understanding

In the field of natural language processing, contextual understanding is the key for models to parse and infer the meaning of text accurately. In this paper, the model enhances the accuracy of entity recognition by combining the Soft-lexicon method and the ERNIE pre-trained model to enhance the model's ability to capture contextual information.

The Soft-lexicon method dynamically captures the multiple meanings of words by incorporating lexicon information into the model, thus enhancing contextual understanding. For example, the word "Xiaomi" in "Xiaomi has recently released a smartphone" can refer to a kind of food and represent the Xiaomi Technology Company. When the model encounters such ambiguity, for each character in the input text, the Soft-lexicon method performs a comprehensive dictionary lookup to identify all possible words that can be formed with that character, constructing a nuanced soft-lexicon feature that reflects these possibilities. For instance, for the character "Mi", the soft-lexicon feature will contain information about both "Xiaomi" and "Xiaomi Corporation". Therefore, the representation of each character is enriched not only by its intrinsic information but also by the lexicon it is associated with.

Additionally, the Soft-lexicon method assigns weights to each word based on its statistical frequency, ensuring that more frequently encountered words exert a dominant influence. As demonstrated in Fig. 10, for "Xiaomi Corporation", which appears four times among the related word sets for each character, compared to twice for "Xiaomi", greater weight is afforded. This ensures that the sequence modeling layer is profoundly attuned to the nuanced contextual information, allowing it to accurately deduce the entity type for each character or word during the augmented character representation modeling process.

Moreover, the ERNIE pre-trained model integrates extensive entity-specific information, including details pertinent to "Xiaomi Company". By merging this comprehensive entity knowledge with the contextual cues from Soft-lexicon, the model is adept at discerning "Xiaomi" as the entity "Xiaomi Corporation" within specific contexts, thereby reducing misrecognition and enhancing recognition accuracy. This integrated approach demonstrates how the model effectively manages intricate entity dependencies, showcasing its robustness in leveraging both statistical and semantic features to resolve complex scenarios, which supports the quantitative outcomes reported.

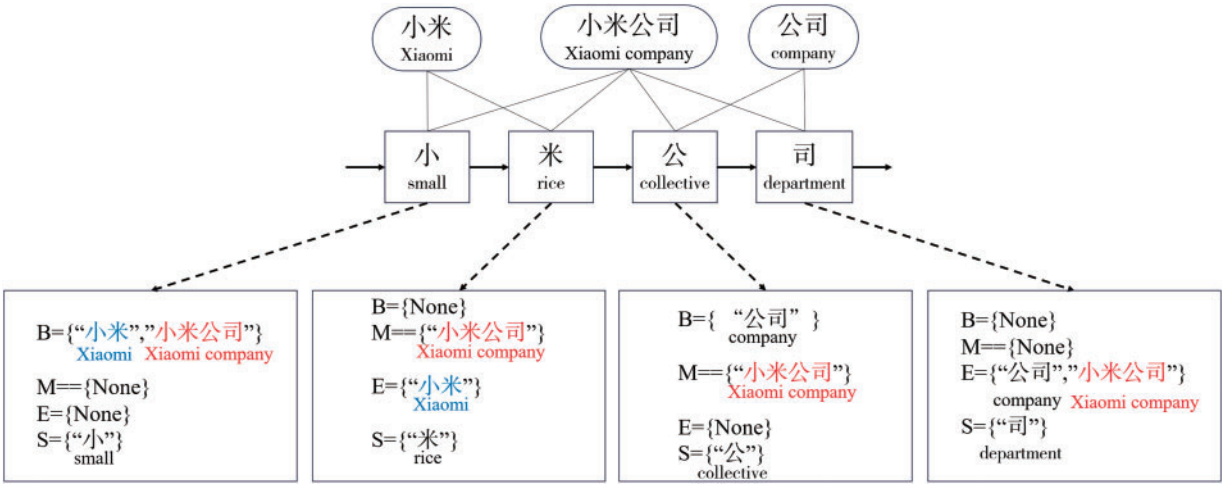

Figure 10: Weighting process of Soft-lexicon

6.2 Discussion of Applications

Our proposed method has broad applicability and can be applied to various news analysis fields, especially for complex texts appearing in the news. The method improves the accuracy and efficiency of processing complex texts by enhancing the model's contextual comprehension ability and the utilization of dynamic lexical information.

For example, in event extraction, it efficiently extracts the core elements of events from news texts, including explicit time and place, as well as event themes with fuzzy boundaries (e.g., "goal", "cause", "impact", etc.), thus supporting the understanding of event information and tracking trends [31]. Sentiment analysis analyzes the sentiment tendency of news texts, particularly in linguistically diverse texts like social media comments. The model leverages the abundant knowledge base of the pre-trained model and its deep contextual understanding to handle unique expressions such as slang, acronyms, and other online language features, enabling it to distinguish between positive, negative, or neutral sentiments. This capability is applicable to public opinion monitoring, brand reputation management, and other related applications [2].

In the task of news summary generation, this method extracts key semantic information from unstructured text and generates structured news summaries to improve the efficiency of information processing [32].

However, in practice, the method faces some potential problems, such as the need to consider its scalability in handling large-scale news streams and in resource-constrained environments. As the amount of news data continues to grow, the model needs to have the ability to efficiently process large amounts of data to meet the demands of real-time news analysis. From the previous time cost analysis, there is still some room for improvement in the computational efficiency of the model, thus affecting the real-time processing capability of the model when facing large-scale news streams. In the future, we can consider using lightweight pre-trained models to reduce the number of parameters and computational complexity of the model, or reduce the computational and memory occupation of the model through model compression techniques (e.g., knowledge distillation, quantization); we can also use incremental learning techniques to enable the model to be quickly updated and adapted when new data arrives, avoiding the need to re-train the whole model, so as to improve the scalability and resource efficiency of the model.

In addition to this, privacy issues and potential risks also require equal attention. Named entity recognition techniques can be used to identify and track key people and locations in the news, and if misused to monitor individual behavior, especially when dealing with large-scale news streams, it will constitute an invasion of personal privacy. At the same time, bias in the entity recognition process can exacerbate unfairness or discrimination, especially when it comes to data from different cultures, languages, and domains. To cope with these problems, the following strategies can be adopted: adopt data anonymization and privacy-preserving technologies to reduce the risk of privacy leakage; introduce fairness and bias detection mechanisms to ensure consistent model performance across different cultures; and improve the transparency and interpretability of the model to help users understand the model decisions, and by doing so, ensure the fairness and reliability of the model and reduce the risk of privacy.

6.3 Discussion of Limitations

Although our method performs well across multiple datasets, there remains room for improvement.

(1) High demand for computational resources

Despite the competitive performance of our methodology across diverse datasets, a critical evaluation of its computational requirements is warranted.. The integration of transfer learning and word embedding techniques enriches contextual representation capabilities but concurrently increases model complexity and resource dependency. Specifically, the Soft-lexicon method dynamically constructs word sets for each character, requiring additional encoding steps for lexical information integration. While our experiments demonstrate efficient inference speeds (Section 5.3), this quadratic complexity imposes significant overhead for lengthy texts, potentially limiting deployment in resource-constrained environments. Moreover, the reliance on pre-trained models and their extensive fine-tuning, especially when combined with sophisticated techniques like word embeddings, can exacerbate resource constraints, posing a challenge for widespread adoption in environments with limited computational capacity.

To address these challenges, we propose the following optimization strategies:

Lightweight Pre-trained Model Adoption: Replacing ERNIE with parameter-efficient architectures (e.g., ALBERT or DistilBERT) could reduce computational costs while preserving accuracy through knowledge distillation [33].

Model Compression: Techniques such as connection pruning (removing redundant neural pathways), parameter quantization (converting 32-bit floats to 8-bit integers), and teacher-student knowledge transfer [34] may substantially decrease memory footprints.

Distributed Computing: Frameworks like Apache Spark could parallelize Soft-lexicon encoding and model inference across GPU clusters, enhancing throughput for real-time applications.

Lexicon Optimization: Caching high-frequency word sets (e.g., top 10% frequent terms) and pruning rare entries (e.g., terms with corpus frequency <3) could streamline lexical processing without compromising recognition fidelity.

(2) Challenges in adapting to diverse data

While our methodology exhibits strong performance on standardized textual data, its efficacy diminishes when applied to non-standardized text, such as the Weibo dataset. This highlights limitations in handling real-world scenarios characterized by diverse linguistic variations and data scarcity. Specifically, the ERNIE pre-trained model, despite its robust contextual understanding, may inherit biases from its pretraining corpus, leading to recognition errors in domains with entity distributions that significantly diverge from those encountered during pretraining.

To evaluate the generalization capabilities of our approach, we assessed the model's performance not only on the SIGHAN Bakeoff 2006 dataset but also on the Resume and Weibo datasets. The results indicate that our method generalizes effectively across different types of Chinese text, achieving an F1 score of 96.18% on Resume and 69.94% on Weibo. This suggests that while the model generalizes effectively to structured, professional texts (e.g., Resume), it struggles with informal content, where entity mentions often appear in fragmented, ambiguous, or novel forms. Several factors contribute to this performance disparity:

Annotation scarcity: User-generated content, such as Weibo posts, contains evolving slang, ambiguous entity mentions, and informal expressions, which are underrepresented in existing annotated corpora.

Domain shift: Pre-trained models like ERNIE, primarily trained on formal texts (e.g., news and encyclopedias), lack exposure to domain-specific and conversational language, making generalization to informal settings difficult.

Entity complexity: As shown in Fig. 11, ORG-type entities achieve lower recall due to their structural variability and context-dependent semantics, which are more challenging to learn with limited labeled examples.

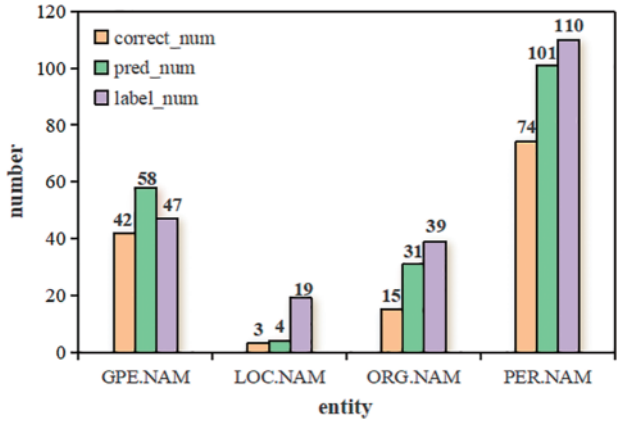

Figure 11: Comparison of the prediction results of various types of entity labels on the Weibo dataset

Fig. 11 compares entity prediction results on Weibo, showing that while GPE and PER entities achieve higher recall due to their clearer syntactic structures and lower context dependency, ORG entities exhibit greater variability and require deeper contextual reasoning, making them more challenging to recognize.

To improve generalization across diverse text domains, future research will focus on the following aspects:

Bias Mitigation through Domain-Adaptive Pretraining: Fine-tuning on domain-specific corpora could reduce biases introduced by pretraining on general-domain texts.

Multi-Granularity Fusion: Using dependency syntactic analysis and semantic role annotation to combine lexical and syntactic features to provide richer contextual information [35].

Knowledge Graph Augmentation: Combining knowledge graphs in order to utilize the relationship and attribute information between entities to enhance the accuracy of the model for entity recognition and linking [36].

Domain-Specific Rule Injection: Embedding industry-specific lexicon or syntactic templates could bolster specialized term recognition [37].

Adaptive Fine-Tuning and Semi-Supervised Learning: Exploring parameter-efficient tuning and weak supervision strategies to reduce reliance on extensive labeled data further.

(3) Limitations in recognizing nested and complex entities

Despite achieving substantial improvements over baseline approaches, our method encounters challenges in handling highly complex or nested entity structures. Soft-lexicon, while effective in incorporating lexical information, may struggle with deeply nested entities or ambiguous boundaries. For example, in news texts, entities such as "Shanghai Stock Exchange" and "Shanghai Securities" may appear in overlapping contexts, making accurate distinction difficult for Soft-lexicon. Furthermore, Soft-lexicon primarily relies on dictionary-based matching, which does not explicitly model hierarchical dependencies between nested entities. Although our method mitigates some of these issues through BiGRU's sequential modeling and CRF's global constraint mechanism, it does not fully resolve all cases of entity overlap. Future improvements could explore attention-based hierarchical modeling or dependency parsing techniques to enhance the recognition of complex entity structures further.

(4) Handling strong context dependencies in news texts

While our model effectively captures contextual information through ERNIE's pre-trained representations, BiGRU's bidirectional modeling, and CRF's global constraints, we acknowledge that strong context dependencies in news texts can still present challenges. Certain long-distance relationships and subtle contextual nuances may not always be fully captured, particularly in cases where entity references span multiple sentences or appear in highly ambiguous contexts. However, the integration of ERNIE allows our model to leverage pre-trained knowledge from large-scale corpora, significantly enhancing its ability to understand real-world contextual associations. Moreover, BiGRU mitigates the limitations of simple sequential models by preserving both past and future contextual cues. While improvements can still be made, our experimental results demonstrate that the combination of these techniques already achieves a high level of accuracy in handling context-dependent entity recognition, surpassing existing baseline methods.

Future improvements could explore multi-hop attention mechanisms to enhance long-distance dependency modeling, integrate discourse-aware pretraining strategies to improve entity linkage across sentences, and leverage graph-based methods to better capture global contextual relationships.

7 Conclusion

We propose a Chinese NER method that combines transfer learning and word embedding. By fusing different model architectures, we utilize the ERNIE pre-trained model for transfer learning and the Soft-lexicon mechanism for word embedding, achieving complementary advantages of different techniques. The

experimental results show that the external knowledge introduced by the pre-trained model, combined with the inherent feature representation of the data, enhances the model's contextual understanding and improves the ability to handle complex texts.

In the future, we will research the following directions: first, the application of lightweight pre-trained models. We aim to study the performance of lightweight models such as TinyBERT, MobileBERT, and DistilBERT in named entity recognition tasks by comprehensively evaluating their performance on NER tasks. This will help us determine the most suitable model for specific application scenarios and analyze how to maintain recognition accuracy while reducing resource consumption, thereby facilitating applications in resource-constrained environments. Second, the fusion of knowledge graphs with pre-trained models. We will consider designing effective knowledge fusion strategies, integrating entities, relationships, and other information from knowledge graphs into pre-trained models, and introducing more domain knowledge and common sense information to improve the representation capability of the models, thereby improving their robustness and interpretability.

Acknowledgement: The authors express heartfelt appreciation to Liangzhong Cui for his invaluable mentorship and un-flinching support throughout the entire research endeavor.

Funding Statement: This research was funded by Advanced Research Project (30209040702).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization: Rui Fang, Liangzhong Cui; methodology: Rui Fang; software: Rui Fang; validation: Rui Fang, Liangzhong Cui; formal analysis: Rui Fang; data curation: Rui Fang; writing—original draft preparation: Rui Fang; writing—review and editing: Liangzhong Cui. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data presented in this study are available on request from the corresponding author due to consideration for future research.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Yan HL, Shi SC. A named entity recognition method for political news. J Beijing Inf Sci Technol Univ. 2018;33(6):23–6, 43. (In Chinese). doi:10.16508/j.cnki.11-5866/n.2018.06.005.
2. Hu TT, Dan YB, Hu J, Li X, Li SB. News named entity recognition and sentiment classification based on attention-based bi-directional long short-term memory neural network and conditional random field. J Comput Appl. 2020;40(7):1879–83. (In Chinese). doi:10.11772/j.issn.1001-9081.2019111965.
3. Zheng YB, Xia ZC, Guo Z, Huang YZ, Liu WF. Named entity recognition of news texts in ten ASEAN countries. Sci Technol Eng. 2018;18(35):162–8. (In Chinese). doi:10.3969/j.issn.1671-1815.2018.35.027.
4. Wei H, Diao H, Kong L, Deng Y. Research on fine-grained named-entity-recognition method for public-opinion texts in northeast Asia. Comput Eng. 2024;50(5):354–62. (In Chinese). doi:10.19678/j.issn.1000-3428.0068955.
5. Zhang N, Li F, Xu G, Zhang W, Yu H. Chinese NER using dynamic meta-embeddings. IEEE Access. 2019;7:64450–9. doi:10.1109/ACCESS.2019.2916816.
6. Luo H, Lu L. Character embedding method for Chinese named entity recognition. J Chin Comput Syst. 2023;7:1434–40. (In Chinese). doi:10.20009/j.cnki.21-1106/TP.2021-0862.
7. Cao P, Chen Y, Liu K, Zhao J, Liu S. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31–Nov 4; Brussels, Belgium. p. 182–92. doi:10.18653/v1/d18-1017.

8. Chen X, Qiu X, Zhu C, Liu P, Huang X. Long short-term memory neural networks for Chinese word segmentation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015 Sep 17–21; Lisbon, Portugal. p. 1197–206. doi:10.18653/v1/d15-1141.
9. Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2016 Aug 7–12; Berlin, Germany. p. 1064–74. doi:10.18653/v1/p16-1101.
10. Zhang H, Kang X, Li B, Wang Y, Liu H, Bai F. Medical name entity recognition based on Bi-LSTM-CRF and attention mechanism. J Comput Appl. 2020;z1:98–102. (In Chinese). doi:10.11772/j.issn.1001-9081.2019081371.
11. Zhang Y, Yang J. Chinese NER using lattice LSTM. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018 Jul 15–20; Melbourne, Australia. p. 1554–64. doi:10.18653/v1/p18-1144.
12. Liu W, Xu T, Xu Q, Song J, Zu Y. An encoding strategy based word-character LSTM for Chinese NER. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics; 2019 Jun 2–7; Minneapolis, MN, USA. p. 2379–89. doi:10.18653/v1/n19-1247.
13. Wu F, Liu J, Wu C, Huang Y, Xie X. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation. In: The World Wide Web Conference; 2019 May 13–17; San Francisco, CA, USA. p. 3342–48. doi:10.1145/3308558.3313743.
14. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: generalized autoregressive pretraining for language understanding. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada. p. 5753–63.
15. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: enhanced language representation with informative entities. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. p. 1441–51. doi:10.18653/v1/p19-1139.
16. Gui T, Ma R, Zhang Q, Zhao L, Jiang YG, Huang X. CNN-based Chinese NER with lexicon rethinking. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence; 2019 Aug 10–16; Macao, China. p. 4982–8. doi:10.24963/ijcai.2019/692.
17. Shi Z, Ma Y, Zhao F, Ma B. Chinese named entity recognition based on CNN-head transformer encoder. Comput Eng. 2022;48:73–80. (In Chinese). doi:10.19678/j.issn.1000-3428.0062525.
18. Yang P, Dong W. Chinese named entity recognition method based on BERT embedding. Comput Eng. 2020;46(4):40–5. (In Chinese). doi:10.19678/j.issn.1000-3428.0054272.
19. Zhu Y, Wang G. CAN-NER: convolutional attention network for chinese named entity recognition. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers); 2019 Jun 2–7; Minneapolis, MN, USA. p. 3384–93.
20. Yan H, Deng B, Li X, Qiu X. TENER: adapting transformer encoder for named entity recognition. arXiv:1911.04474. 2019.
21. Li X, Yan H, Qiu X, Huang X. FLAT: chinese NER using flat-lattice transformer. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 6836–42. doi:10.18653/v1/2020.acl-main.611.
22. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12–17; San Diego, CA, USA. p. 260–70. doi:10.18653/v1/n16-1030.
23. Zhang FC, Qin QL, Jiang Y, Zhuang RT. Named entity recognition for Chinese EMR with RoBERTa-WWM-BiLSTM-CRF. Data Anal Knowl Discov. 2022;6(2):251–62. (In Chinese). doi:10.11925/infotech.2096-3467.2021.0910.
24. Zhao J, Cui M, Gao X, Yan S, Ni Q. Chinese named entity recognition based on BERT and lexicon enhancement. In: Proceedings of the 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence; 2022 Dec 16–18; Dongguan, China. p. 597–604. doi:10.1145/3584376.3584482.

25. Guo Q, Guo Y. Lexicon enhanced Chinese named entity recognition with pointer network. Neural Comput Appl. 2022;34(17):14535–55. doi:10.1007/s00521-022-07287-1.
26. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minneapolis, MN, USA. p. 4171–86.
27. Sun Y, Wang S, Feng S, Ding S, Pang C, Shang J, et al. Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. arXiv:2104.02137. 2021.
28. Wang Y, Sun Y, Ma Z, Gao L, Xu Y. An ERNIE-based joint model for Chinese named entity recognition. Appl Sci. 2020;10(16):5711. doi:10.3390/app10165711.
29. Huang S, Sha Y, Li R. A Chinese named entity recognition method for small-scale dataset based on lexicon and unlabeled data. Multimed Tools Appl. 2023;82(2):2185–206. doi:10.1007/s11042-022-13377-y.
30. Zhang L, Xia P, Ma X, Yang C, Ding X. Enhanced Chinese named entity recognition with multi-granularity BERT adapter and efficient global pointer. Complex Intell Syst. 2024;10(3):4473–91. doi:10.1007/s40747-024-01383-6.
31. Niu F, Zhong S, Liu N, Zhong W, Yang D, Ye X, et al. Research on an improved extraction method for three elements of disaster news. J Saf Sci Technol. 2023;19(2):13–9. doi:10.11731/j.issn.1673-193x.2023.02.002.
32. Li K, Chen YR, Zheng WJ, Hua BL. News timeline mining and presentation based on machine reading comprehension. Inf Stud Theory Appl. 2022;45(4):184–9. (In Chinese). doi:10.16353/j.cnki.1000-7490.2022.04.025.
33. Yu D, Huang J, Dang T, Zhang K. Recognition of named entity in Chinese resume based on ALBERT. Comput Eng Des. 2024;45:261–7. (In Chinese). doi:10.16208/j.issn1000-7024.2024.01.033.
34. Zhao H, Tang H, Zhang Y, Sun X, Lu M. Named entity recognition model based on k-best viterbi decoupling knowledge distillation. J Front Comput Sci Technol. 2024;18(3):780–94. (In Chinese). doi:10.3778/j.issn.1673-9418.2211052.
35. Xu XB, Wang T, Kang R, Zhou G, Li TN. Multi-feature Chinese named entity recognition. J Sichuan Univ Nat Sci Ed. 2022;59(2):51–7. (In Chinese). doi:10.19907/j.0490-6756.2022.022003.
36. Jin ZG, He XY, Yue SM, Xiong YL, Luo J. Named entity recognition in medical domain combined with knowledge graph. J Harbin Inst Technol. 2023;55(5):50–8. (In Chinese). doi:10.11918/202201126.
37. Wang J, Wang Z, Cao S. A named entity recognition model based on lis domain knowledge. Libr Trib. 2023;43:15–25. (In Chinese). doi:10.3969/j.issn.1002-1167.2023.07.004.