Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 2, 2023

Chinese Named Entity Recognition based on ERNIE

Xing Qi

School of Electrical Engineering, Southwest Minzu University, Chengdu, China

Abstract: The traditional named entity recognition model based on neural network uses static word vector, which can't represent the ambiguity of the word in the context. The ERNIE-BiLSTM-CRF model is proposed. The ERNIE pre-training model can output different word vectors for different contexts by using multiple layers of Transformer, obtaining dynamic word vectors that contain overall sequence information. Secondly, the word vectors are input into the BiLSTM layer, which can obtain sentence context information through forward and backward LSTM and obtain more sentence features, thereby improving the model's effectiveness. Finally, the sequence is labeled through the CRF layer to obtain the globally optimal labeling information and complete the named entity recognition task. The experimental results show that compared with the traditional model, the F1 score of this model has significantly improved.

Keywords: Named Entity Recognition; Enhanced Representation through Knowledge Integration; Gated Recurrent Unit; Conditional Random Field.

1. Introduction

Named Entity Recognition (NER) [1]is a fundamental task in natural language processing, aimed at identifying entities with specific meanings, such as names of people, places, and organizations, from unstructured text data. This provides a foundation for subsequent applications such as information extraction, question-answering systems, and machine translation. Chinese entity structures are complex, diverse in form, and have fuzzy boundaries [2]. Moreover, Chinese characters can have multiple meanings or be polysemous in different contexts. These factors increase the difficulty of Chinese NER, making it more valuable for research and practical applications. The ERNIE-BiLSTM-CRF model is proposed. The ERNIE pre-training model can output different word vectors for different contexts by using multiple layers of Transformer, obtaining dynamic word vectors that contain overall sequence information. Secondly, the word vectors are input into the BiLSTM layer, which can obtain sentence context information through forward and backward LSTM and obtain more sentence features, thereby improving the model's effectiveness. Finally, the sequence is labeled through the CRF layer to obtain the globally optimal labeling information and complete the named entity recognition task. The experimental results show that compared with the traditional model, the F1 score of this model has significantly improved.

2. Related Work

Named Entity Recognition (NER) is the extraction of required entities from structured or unstructured text. In 1996, the MUC-6 conference first proposed the task of named entity recognition and specified the main task as identifying entities in the text to be processed [3]. The entities that needed to be identified were divided into three major categories (entity, time, and numerical) and seven subcategories (person, organization, location, time, date, currency, and percentage). The named entity recognition task proposed by the MUC-6 conference has been of great interest to people, and subsequently, conferences such as ACE and CoNLL-2003 have also proposed research on named entity recognition [4]. In these conferences, not only were the entities specified by MUC-6 identified, but new entity types were also added. For example, the ACE conference added multiple entity types such as geopolitical, facility transportation, and weapons. Additionally, ACE and CoNLL-2003 introduced multiple languages in the named entity recognition task, such as Spanish and Chinese in ACE, and English and German in CoNLL-2003[5]. Moreover, there were differences in the research tasks and mainstream methods. ACE not only carried out the task of named entity recognition but also included relationship recognition, reference, and anaphora resolution as part of the research [6]. In contrast to MUC-6, the methods that performed well in CoNLL-2003 directly or indirectly used the maximum entropy model or some statistical machine learning methods, while in MUC-6, rule-based and dictionary-based methods were the mainstream methods at the time [7]. Bootstrapping [8] is a classic method that can automatically generate rules. The rule-based and dictionary-based methods require manual customization by relevant personnel. In the current field and scope of corpora, these methods need to gradually increase in size to adapt to new situations, resulting in significantly increased workload. Therefore, rule-based and dictionary-based methods are rarely used alone nowadays but are mainly used in combination with other methods to improve model accuracy.

After the use of rule-based and dictionary-based methods, people began to use statistical learning methods for named entity recognition research. The Conditional Random Fields (CRF) model was used later than the Support Vector Machine (SVM) model, but it was more effective, and it is now one of the most commonly used models. Researchers such as Li[9] and Jiang[10] have conducted a series of model comparison studies, and from the experimental results, it can be concluded that CRF is a better model for named entity recognition. Stanford University has also developed the StandorfNER tool for named entity recognition based on CRF. Multi-model mixing is a choice for improving model performance. Li[11] and others used multiple SVM models to improve model performance. The use of external knowledge bases to solve new entities is a method chosen by many researchers, such as Cukierman [12] using the Wikipedia database for semantic disambiguation.

Because traditional machine learning models heavily rely on feature engineering, the selection of feature engineering is very complex. In recent years, with the rise of deep learning, representation learning has become a research hotspot. The method of word vector representation in deep learning solves the problem of data sparsity, and word vectors contain semantic information. Compared with the cumbersome manual selection of features, using word vectors is more suitable for the development of named entity recognition. Under the trend of deep learning, various neural network models have been proposed and improved, and have been used in named entity recognition tasks, such as convolutional neural network (CNN), long short-term memory network (LSTM), and bi-directional long short-term memory network (BiLSTM). In recent years, the BiLSTM-CRF model based on deep learning has been more effective in medical named entity recognition. Collobert et al. proposed the most representative deep learning model and designed the SENNA [13] system, which requires small memory and can efficiently solve various problems such as part-of-speech tagging and named entity recognition. Sahu[14] et al. generated word embedding features by cascading CNN and RNN, which can effectively reduce workload without too much feature engineering. Vaswani [15] et al. proposed the Transformer model, which abandons traditional CNN and RNN, and the entire network structure is composed entirely of the attention mechanism, which reduces complexity and allows for parallel computation. Yan [16] et al. optimized the Transformer-based model and proposed the TENER model. Incorporating language features such as phonetic and character features into the model is also a way to improve the model. Bharadwaj [17] et al. used LSTM to incorporate phonetic features to improve the model's effectiveness.

With the development of the machine learning field, neural network design has become increasingly complex, leading to more and more workload. People use a large amount of data to train the network structure to obtain pre-trained models, which can be fine-tuned on their own dataset, reducing workload. There are already many pre-trained models, such as the BERT [18] model proposed by Devlin et al. based on Transformer, the Roberta [19] model proposed by the Facebook team the following year, and the Albert [20] model proposed by Lan et al. Improvements to the BERT model are mainly achieved by modifying the Next Sentence Prediction and Masked LM to improve model performance [21]. Dong [22] et al. proposed the UNILM model, and Song [23] et al. proposed the MASS model. Tsai [24] et al. proposed a BERT-based model and used knowledge distillation to improve model training time. Nested named entities refer to the phenomenon of entities nested in entities. There are more studies on this aspect abroad, and the methods for nested named entities can be divided into several types: based on hypergraphs, stack-based model region models, and based on reading comprehension. LU [25] et al. first proposed a method for nested named entity recognition based on hypergraphs in 2015, which merges tokens with the same value and makes each token contain an O label, and then the decoder outputs all possible labels. Muis[26] et al. proposed a multi-graph representation method, which uses separators to detect nested entities. JU [27] et al. extracted nested entities by stacking BiLSTM+CRF layers. Jue[28] et al. stacked each token from bottom to top, generating longer entities that exceed the length of the lower layers, and detected all entities in the text by traversing all entities. Named entity recognition technology has gradually matured, especially with the application of neural networks, which has led to a gradual improvement in the F1 score of named entity recognition. Pre-trained models can solve certain problems, but new issues have also emerged, such as excessive resource consumption [29].

3. ERNIE-Bigru-CRF Model

The overall architecture of the ERNIE-BiGRU-CRF model is shown in Figure 1. The model first obtains the semantic representation of the input through the ERNIE pre-trained language model, which enhances knowledge-based semantic representations. The obtained word vectors are inputted into a bidirectional GRU layer to extract sentence-level features, and finally, the CRF layer performs sequence labeling to obtain the globally optimal label sequence.

Compared with previous mainstream named entity recognition models, the most significant difference of the ERNIE-BiGRU-CRF model is the incorporation of knowledge-enhanced semantic representations from the ERNIE pre-trained language model. The ERNIE model learns a comprehensive semantic representation of concepts by masking semantic units such as words and entities. This representation captures the ambiguity of words and enhances the semantic representation capacity of the model.

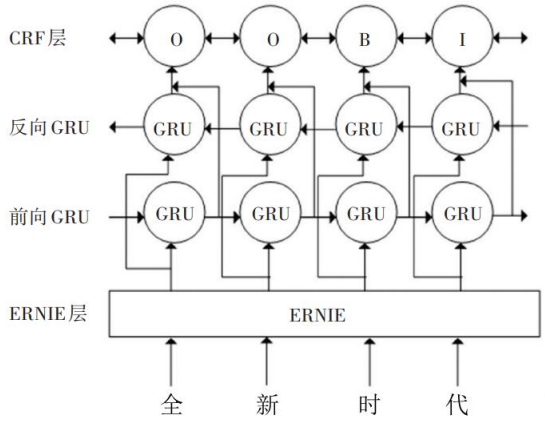

Fig. 1 Named entity recognition framework based on ERNIE - BiGRU-CRF

3.1. ERNIE Pre-trained Language Model

ERNIE is a knowledge-enhanced semantic representation model that models prior semantic knowledge of words, entities, and entity relationships in massive data to learn a comprehensive semantic representation of concepts [30]. ERNIE and BERT are both pre-trained language models constructed with a multi-layer bidirectional Transformer encoder as the basic unit, as shown in Figure 2 [31].

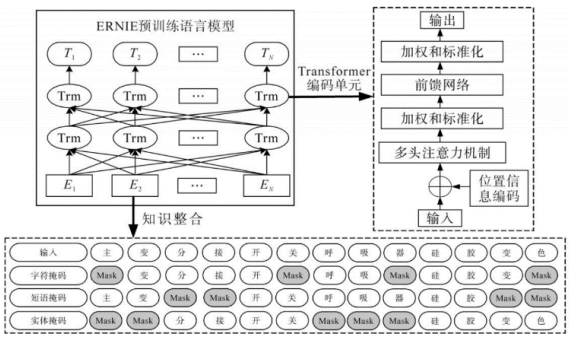

Fig. 2 Structure of ERNIE

ERNIE consists of two parts: encoding and knowledge integration. In the encoding part, the Transformer encoder is used to generate text word vectors that fuse contextual semantic information. As shown in the diagram, the Transformer encoder is modeled based on an attention mechanism. The attention mechanism can reflect the importance of different vocabulary in the text and improve the weight of information relevant to defect text classification, thereby further improving the accuracy of feature extraction by the model. In the knowledge integration part, ERNIE proposes a multi-stage knowledge masking strategy that integrates semantic knowledge of defect text at the character, phrase, and entity levels. Knowledge masking refers to randomly masking some characters and training the model to predict the masked part, thereby effectively learning the contextual information of the masked part. Compared to the single-character masking strategy of the BERT model, ERNIE introduces three levels of masking, including characters, phrases, and entities, and through a multi-stage knowledge masking strategy, ERNIE can generate word vectors that contain rich semantic information of the defect text and effectively preserve the correlation between various components of the defect text, thus ensuring that important semantic information is not lost.

3.2. Recurrent Neural Network

Traditional machine learning methods rely on manual feature extraction before text classification, which can easily lead to loss of contextual information. To solve this problem, Long Short-Term Memory (LSTM) uses gate units to control the process of long-term information transfer and enhance the correlation of context. LSTM has three gate structures, including input gate (i), forget gate (f), output gate (o), and memory cell (c), which can effectively overcome the problem of gradient vanishing that exists in general neural networks. The structure of LSTM is shown in Figure 3.

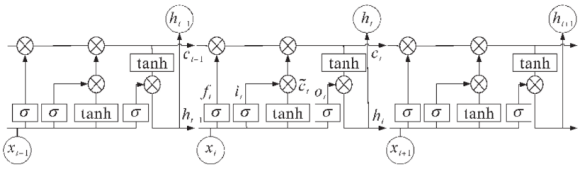

Fig. 3 LSTM structure

In Figure 3, xt is the input at time t; ft is the output for t+1 at time t; ot is the output at time t; ht is the hidden layer representing the output at time t; σ is the sigmoid function; ct is the cell state at time t. The gate unit calculation formulas in LSTM are shown below:

$$i_t = \sigma(w_i \cdot [h_t, x_t] + b_i) \quad (1)$$

$$f_t = \sigma\left(w_f \cdot [h_t, x_t] + b_f\right) \quad (2)$$

$$o_t = \sigma(w_o \cdot [h_t, x_t] + b_o) \quad (3)$$

$$h_t = O_t \tanh C_t \quad (4)$$

$$c_t = f_t C_{t-1} + i_t \tanh(w_c \cdot [\mathrm{h_{t-1}}, \mathrm{x_t}] + \mathrm{b_c}) \quad (5)$$

In the formula, Wi, Wf, and Wo are weight matrices that connect to the gate unit; bi, bf, and bo are the bias values.

4. Experiments and Results

4.1. Experimental environment and experimental data set

The specific experimental environment is as follows: Python was chosen as the development language, the CPU model is AMD Ryzen 7 4800H, and Visual Studio Code was chosen as the development tool.

The experiment used the People's Daily corpus from January 1998 as the dataset. This corpus is a tagged corpus produced jointly by the Peking University Computational Linguistics Laboratory and the Fujitsu Research Center, and has been used as raw data in a large number of research studies and papers. The experimental process involved randomly dividing the dataset into training set, validation set, and test set, with 80%, 10%, and 10% respectively. The dataset was annotated using the BIO tagging scheme (Beginning, Inside, Outside), which includes three entity types: person names (PER), location names (LOC), and organization names (ORG), with a total of seven tags ('B-PER', 'I-PER', 'B-LOC', 'I-LOC', 'B-ORG', 'I-ORG', 'O'). Precision (P), recall (R), and F1-score were used as evaluation metrics for the model in the experiment, and were calculated using the following formulas:

$$P = \frac{\text{模型正确标注的实体个数}}{\text{模型标注的所有实体个数}} \times 100\% \quad (6)$$

$$R = \frac{\text{模型正确标注的实体个数}}{\text{样本中所有实体个数}} \times 100\% \quad (7)$$

$$F = \frac{2 \times P \times R}{P+R} \times 100\% \quad (8)$$

4.2. Experimental result

The experiment compares the model in the article with other models to validate its effectiveness. Several models with good performance and mainstream popularity were selected for comparison with the model in the article, including CRF, LSTM-CRF, and GRU-CRF, and the results are shown in Table 1. The experimental results show that the model in the article has improved in all metrics compared to other models. By comparing the experimental results on the same dataset, it is demonstrated that applying the model in the article can improve the recognition performance in entity recognition tasks.

Table.1 Comparison of experimental results of different models %

模型	*P*	*R*	F1
CRF	72.41	63.53	67.68
BiLSTM-CRF	76.83	69.06	72.74
RNN-CRF	76.32	70.17	73.11
GRU-CRF	78.94	71.26	74.90
ERNIE-BiLSTM-CRF	83.74	75.52	79.41

5. Conclusion

To address the problem that traditional word vectors cannot represent the polysemy of characters and models have difficulty in obtaining complete semantic representations, this paper proposes the ERNIE-BiGRU-CRF model. The model uses multi-layer bidirectional Transformers as encoders to extract features, and adopts three levels of masking strategies including character masking, phrase masking, and entity masking to dynamically generate context semantic representations of characters. Compared to traditional word vectors, ERNIE-BiGRU-CRF can enhance the model's semantic representation ability and improve named entity recognition performance. However, in specific domains that lack large-scale annotated data, the model may extract incorrectly due to insufficient context information and the presence of abbreviations and ambiguous entities. Therefore, the next research direction can consider combining deep

learning with transfer learning methods to address these issues.

References

[1] MARRERO M, URBANO J, S Á NCHEZ-UADRADO S, et al. Named Entity Recognition: Fallacies, Challenges, and Opportunities [J]. Computer Standards and Interfaces, 2013, 35 (5): 482.

[2] Zhang Libang Chinese electronic medical record segmentation and name entity mining based on semi supervised learning [D] The following is: [Master's Thesis] Harbin: Harbin Institute of Technology, 2014

[3] Chinchor N. MUC-6 Named Entity Task Definition (Version 2.1) [C]. Proceedings of the 6th Conference on Message Understanding, Columbia, Maryland, 1995: 142-194.

[4] LDC Corporation. Entity Detection and Tracking Phase 1 Experimental Study Task Definition [EB/OL].

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/edt-phase1-v2.2.pdf , 2017-03-10.

[5] Sang, Eric, Meldfen. CoNLL-2003 Shared Task Introduction: Language Independent Named Entity Recognition [C]. Calculate natural language learning progress. Association for Computational Linguistics, 2003:142-147.

[6] Doddington G. Tasks, Data, and Evaluation of Automated Content Extraction (ACE) Programs [J]. Proc Lrec, 2004, 34(3): 123-153.

[7] Liu Liu, Wang Dongbo Overview of Named Entity Recognition Research [J]. Journal of Information Technology, 2018, 37 (03): 329-340

[8] Silviu Cucerzan, David Yarowsky. Language independent named entity recognition combining morphological and contextual evidence [C]. Proceedings of the 1999 SIGDAT EMNLP and VLC Joint Meeting, 1999:90-99.

[9] Li D, Savova G, Kipper-schuler K. Conditional Random Fields and Support Vector Machines for Recognition of Named Obstacle Entities in Clinical Text [C]. Proceedings of the 2008 Symposium on Current Trends in Biomedical Natural Language Processing. 2008:94-95.

[10] Jiang M, Chen Y, Liu M. Research on machine learning based methods to extract clinical entities and their assertions from a large number of abstracts [J]. Journal of the American Medical Informatics Association, 2011, 18 (5): 601-606.

[11] Li L, Mao T, Huang D, et al. A Hybrid Model for Chinese Named Entity Recognition [C]. Proceedings of the 5th SIGHAN Symposium on Chinese Language Processing. Strauss burg: Computational Linguistics Association, 2006:72-78.

[12] Cucerzan S. Large scale named entity disambiguation based on Wikipedia data [C]. Proceedings on Empirical Methods in Natural Language Processing. Czech Republic: Prague, 2007:708-716.

[13] Ronan Collabert, Jason Weston, L é on Bottou, et al. Natural language processing starts (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12 (76): 2493-2537.

[14] Sahu S K, Anand A. Recursive Neural Network Model for Disease Name Recognition Using Domain Invariant Features [C]. Proceedings of the 54th Annual Conference of the Computational Linguistics Association. Strauss burg: Computational Linguistics Association, 2016: 2216-2225.

[15] Vaswani A, Shazeer N, Parmar N, etc. Attention is what you need [J]. ar Xiv, 2017, (1706): 37-62.

[16] Yan H, Deng B, Li X, et al. TENER: Adaptive converter encoder for name entity recognition [J]. ar Xiv, 2019, (1911): 44-74.

[17] Bharadwaj A, Mortensen D, Dyer C, etc. Speech Perceptual Neural Model for Named Entity Recognition in Low Resource Transfer Environments [C]. Proceedings of the 2016 Conference on Empirical Methods of Natural Language Processing. Strauss burg: Computational Linguistics Association, 2016:1462-1472.

[18] Devlin J, Chang M W, Lee K, et al. Bert: Pre training of deep bidirectional converters for language understanding [J]. ar Xiv, 2018, (1810): 48-50.

[19] Liu Y H, Ott M, Goyal N, et al. Roberta: A robust optimization Bert pre training method [EB/OL].

https://arxiv.org/abs/1907.11692 , 2019-07-26.

[20] Lan Z, Chen M, Goodman S, et al. ALBERT: Simplified BERT [EB/OL] for language representation self supervised learning.

https://openreview.net/pdf?id=H1eA7AEtvS , 2019-07-26.

[21] Wang Naiyu, Ye Yuxin, Liu Lu, Feng Lizhou, Bao Tie, Peng Tao. Research on Language Models Based on Deep Learning Progress [J]. Journal of Software, 2021, 32 (04): 1082-1115

[22] DongL, YangN, WangW, et al. A unified language model for pre training and generation of natural languages [EB/OL].

https://arxiv.org/abs/1905.03197 ,2019-0 5-08.

[23] Song K, Tan X, Qin T, et al. MASS: Language generated masking sequence to sequence pre training [EB/OL].

https://arxiv.org/pdf/1905.02450.pdf , 2019-06-21.

[24] Tsai H, Riesa J, Johnson M, et al. A small practical BERT model for sequence tagging [C]. Procedure. 2019 Conference on Natural Language Processing Experiences and Methods and the Ninth International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 3623 – 3627.

[25] Lu W, Dan R. Joint reference extraction and classification using reference hypergraphs [C]. Proceedings of the 2015 Conference on Empirical Methods of Natural Language Processing, 2015:857-867.

[26] Muis A O, Lu W. Mark gaps between words: Use reference separators to identify overlapping references [C]. Proceedings of the 2017 Conference on Empirical Methods of Natural Language Processing, 2017:2608-2618.

[27] Ju M, Miwa M, Ananiadou S. Neural hierarchical model for nested named entity recognition [C]. Proceedings of the 2018 North American Branch of the Computational Linguistics Association: Human Language Technology, Volume 1 (Long Paper), 2018: 1446-1459.

[28] Jue W, Shou L, Chen K, et al. Pyramid: A Layered Model for Nested Named Entity Recognition[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020:5918-5928.

[29] Li Zhoujun, Fan Yu, Wu Xianjie. Overview of Research on Pretraining Technology for Natural Language Processing [J]. Computer Science, 2020,47 (03): 162-173

[30] Sun Y, Wang S, Li Y, et al. ERNIE 2.0: A continual pre-train-ing framework for language understanding[J]. Proceedings of theAAAI Conference on Artifical Intelligence, 2020, 34(5): 8968-8975.

[31] Luo Xiao. Overview of Natural Language Processing Research Based on Deep Learning [J]. Intelligent Computers and Applications, 2020, 10 (4): 133-137. Luo Xiao A survey of natural language processing based on deep learning[J]. Intelligent Computer and Application, 2020, 10(4): 133-137.