Named entity recognition in steel field based on BiLSTM-CRF model

Zhai Chenhao¹, Wang Chengyao¹*
¹ School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing, 100083, China
*Corresponding author’s e-mail: wangchengyao@ustb.edu.cn

Abstract. NER is a basic research in the field of NLP. Most of the Chinese NER tasks are for the recognition of names of people, places, and institutions. The recognized content is single and cannot be applied in many fields. In this paper, we label and build the NER dataset for the steel field. Based on the BiLSTM-CRF model, we proposed character and word combination embedding method. Experiments show that our method reduces the OOV problem, which has an increase in 7.20% and 5.03% in the F1 score compared with the traditional character embedding and word embedding.

1. Introduction
Named entity recognition [1] is a technique for recognizing specific nouns in unstructured free text. The general task will recognize the person’s name, place name, and organization name from the text. The content recognized and extracted to vary from the task. As an upstream task in the NLP field, NER technology directly affects the accuracy of downstream tasks such as relation extraction [2] and entity linking. At the same time, NER technology is widely used in Knowledge Graph, intelligent questions and answers [3] and machine translation.

At present, the NER task has shown good performance in the field of general research, and there is no relevant research results in the steel field. This paper proposes the technology of named entity recognition of the steel field, especially for the recognition of entities in the steel product category. Then, we can extract key information about product texts in the steel industry to build knowledge graph.

The difficulty in entity recognition of the steel fields lies in: (1) The text in the steel field is Chinese, and there is no explicit segmentation between words. (2) There are more entities and types. (3) There are many mixed cases of Chinese and English in the entity names of the steel field.

This paper builds the data set required for the NER task in the steel field. Based on the BiLSTM-CRF model, we propose a combination of character and word embedding methods. And introduce the CNN module in the embedding layer. Compared with word embedding and character embedding, our approach reduces the OOV problem with word granularity and improves the performance of entity recognition.

2. Related Work
This section gives a brief introduction to the method of NER task. The methods of NER are mainly divided into three categories.

The first category, based on rules and feature templates, requires a large number of rules to be manually defined, and the generalization is extremely poor. Zhou Kun et al. [4] used pre-defined rules
for Chinese named entity recognition, and constantly expanded the rule base to achieve better results. Riaz K [5] et al. used rule matching to recognize Urdu entities. Fang X [6] et al. propose a hybrid machine learning method and a rule-based method, which improved the extraction effect of names by 14%.

The second category, based on the traditional machine learning method, although the generalization is improved compared to the first category, it still needs a lot of feature engineering. Usually researchers use HMM [7], CRF [8] and SVM [9] algorithms for named entity recognition. Fu G [10] et al. used the lexicalized HMM method to improve the effect of entity recognition of unknown vocabulary. Li L [11] et al. used the CRF-based hybrid model to achieve 93.61 and 91.75 F1 scores for person names and place names on the MSRA data set, respectively.

The third category, based on the deep learning method, has the advantages of less feature engineering and higher generalization than the first two categories. Sun [12] et al. used the LSTM-CRF model to achieve entity recognition of the fisheries filed. Huang Z [13] et al. used BiLSTM-CRF to model the context of the input sequence and achieved good results in the NER task.

3. Methodology
This section mainly introduces the labeling content of the dataset and the labeling method, the architecture of the model, and the implementation details of each module.

3.1. Data set label
Since there is no public data set in the steel field, we label the existing text data before the entity recognition task. The strategy used for labeling is the "BIO" method. "B" represents the beginning of the entity, "I" represents the non-first part of the entity, and "O" represents the non-entity part. The specific labeling contents include product name, institution, performance, strength, use, specification, production process, origin, brand and content.

We manually label and build a seed collection. Then, the labeling result is extracted to form a domain dictionary, and the dictionary is used for re-labeling. Finally, the labeling result is manually optimized. Loop iterations ultimately result in better quality data sets. The process is shown in Figure 1.

![Figure 1. Data set labeling process diagram.](image)

3.2. Architecture
The model proposed in this paper is based on BiLSTM-CRF. At the input, we add character and word embedding module to enrich the input information. The architecture has three main modules from input to output: Embedding layer, BiLSTM layer and CRF layer. The input of the model is the sequence of words \( X = \{x_1, x_2, \ldots, x_n\} \). The output is the predicted tag sequence \( Y = \{y_1, y_2, \ldots, y_j, \ldots, y_n\} \). The architecture is shown in Figure 2.
3.3. Embedding layer

In this paper, we proposed character and word combination embedding, compared with character embedding and word embedding. In the Chinese NER task, the character embedding method can avoid the performance degradation caused by the inaccuracy of the word segmentation, and the word embedding method can contain more information. The input to the model is a sequence of word granularity. CNN module is introduced into the embedding layer. Local features are extracted from the character sequence of each word. After max pooling layer, the features of the character sequence are obtained, and joint with the word embedding to get the final embedding. The specific module details are shown in Figure 3.

![Figure 2. Architecture diagram.](image)

![Figure 3. Embedding layer.](image)
3.4. BiLSTM layer
LSTM (Long Short-Term Memory) [14] is a variant of RNN. It has three gate units: input gate, output gate and forget gate. LSTM improves the problem of RNN gradient disappearance and gradient explosion, and the sequence modeling effect is better. Chinese NER task can be regarded as sequence labeling task. The bidirectional LSTM can effectively use the information of the forward and backward, and can extract the timing information inside the sequence. The role of BiLSTM in this paper is to further extract the feature of embedding.

\[ i_t = \sigma(x_t \cdot o_{lh}^i + h_{t-1} \cdot o_{hh}^i + b_i^f) \]  
\[ f_t = \sigma(x_t \cdot o_{lh}^f + h_{t-1} \cdot o_{hh}^f + b_f^f) \]  
\[ o_t = \sigma(x_t \cdot o_{lh}^o + h_{t-1} \cdot o_{hh}^o + b_o^o) \]  
\[ c_t = \tanh(x_t \cdot o_{lh}^c + h_{t-1} \cdot o_{hh}^c + b_c^c) \]  
\[ h_t = o_t \otimes \tanh(c_t) \]

The above is the operation process of the LSTM cell. Equation (1) is the input gate, equation (2) is the forget gate, and equation (3) is the output gate. \( \otimes \) is a point multiplication operation. \( \sigma \) is the Sigmod function. \( w \) and \( b \) represent weights and bias. \( c_t \) is the intermediate state of the cell. \( h_t \) is the final state of the cell. \( h_{t-1} \) is the state of the previous moment. \( h_t \) is the final output of the current moment.

The output of Bi-LSTM is the splicing of forward output and reverse output \( h_t = [\tilde{h}_t, \tilde{h}_t] \).

3.5. CRF layer
CRF (Conditional Random Field) [8] mentioned in the previous section is a traditional method to solve the sequence labeling task, but many feature engineering needs to be done. In the model mentioned in this paper, the role of CRF is to solve the problem of label deviation. BiLSTM and the previous embedding layer can be regarded as feature engineering in traditional machine learning. However, the output of BiLSTM is independent, which will lead to the appearance of illegal labeling results, that is, results similar to \{O, O, B, O, I, I, O\}, obviously I tag can only appear Behind the B tag or the I tag. The specific form of CRF is shown in equations (7) and (8):

\[ P(y | x) = \frac{1}{Z(x)} \exp \left( \sum_{i,k} \lambda_k t_k (y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l (y_i, y_{i+1}, x, i) \right) \]  
\[ Z(x) = \sum \exp \left( \sum_{i,k} \lambda_k t_k (y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l (y_i, y_{i+1}, x, i) \right) \]

Where \( t_k \) and \( s_l \) are the feature functions, \( \lambda_k \) and \( \mu_l \) are the weights, which \( Z(x) \) is the normalization factor. The CRF considers the relationship of adjacent tags during the learning phase to obtain a state transition matrix \( A \). In the prediction phase, a first-order Viterbi algorithm is used to decode, and a global optimal solution is obtained.

4. Experimental results and comparison
In this section, we will compare our proposed model with the baseline model and detail the experimental dataset, evaluation indicators, experimental results and analysis.

4.1. Dataset and evaluation indicators
The data set used in the experiment of this paper is a self-built data set in the steel field. The experimental data set was formed by manual labeling of 100 documents. The total number of sentences in the data set is 2302, the total number of entities is 4564, and the number of characters is 64955. The allocation ratios of the training and test sets are 85% and 15%. The specific corpus statistics are shown in Table 1.
4.2. Experimental result

In this paper, three sets of experiments were set up, hyperparameter experiment, model comparison experiment and main control experiment. The embedding matrix used in the experiment is the result of random initialization, the pre-trained Word2Vec is not used, and some training tricks are not used. The purpose is to ensure that the single change principle of the experimental variables and the interpretability of the results.

4.2.1. hyperparameter experiment. This experiment is based on the model proposed in this paper. The experiment is mainly for optimizer and learning rate. The specific experimental results are shown in Table 2.

| Optimizer | Learning rate | Accuracy | Precision | Recall | F1  |
|-----------|---------------|----------|-----------|--------|-----|
| SGD       | 0.1           | 53.97    | 3.41      | 11.31  | 5.24|
|           | 0.03          | 90.39    | 66.38     | 62.26  | 64.26|
|           | 0.01          | 90.32    | **69.04** | 58.60  | 63.39|
|           | 0.003         | 89.95    | 67.18     | 56.05  | 61.11|
|           | 0.001         | 89.17    | 62.02     | 50.96  | 55.94|
| Adam      | 0.1           | 48.45    | 3.21      | 13.22  | 5.17|
|           | 0.03          | 75.33    | 7.74      | 4.14   | 5.39|
|           | 0.01          | 86.53    | 57.78     | 53.82  | 55.73|
|           | 0.003         | **90.88**| 66.72     | **64.17**| **65.42**|
|           | 0.001         | 90.39    | 64.97     | 60.83  | 62.83|

It can be seen from Table 2 that no matter which optimizer is used, if the learning rate is too large or too small, the optimal solution will not be obtained. SGD performs better in terms of precision, and Adam is better at the overall performance of the model. Finally, the various hyperparameters used in our experiments are shown in Table 3.

| Parameter | Value |
|-----------|-------|
| Batch_size| 128   |
| Cell_num  | 1     |
4.2.2. **Model comparison experiment.** This experiment compares the results of four models of LSTM, LSTM-CRF, BiLSTM and BiLSTM-CRF, all using character and word embedding. This experiment uses the same dataset and hyperparameters. The comparison results are shown in Table 4. The word &char in the table stands for character and word embedding.

| Model Name               | Accuracy | Precision | Recall | F1     |
|--------------------------|----------|-----------|--------|--------|
| LSTM+word&char           | 88.83    | 56.11     | 55.57  | 55.84  |
| LSTM-CRF+word&char       | 89.70    | 62.20     | 57.64  | 59.83  |
| BiLSTM+word&char         | 90.23    | 62.79     | 60.99  | 61.87  |
| BiLSTM-CRF+word&char     | 90.88    | 66.72     | 64.17  | 65.42  |

As can be seen from Table 4, BiLSTM is better than LSTM regardless of whether or not the CRF layer is used. The reason for the above results is that BiLSTM adds backward information compared to LSTM. The CRF layer constrains the output sequence, so adding a CRF layer to LSTM or BiLSTM can significantly improve the model. In summary, the BiLSTM-CRF model has the highest F1 score, so the model we used for the main control experiment was BiLSTM-CRF.

4.2.3. **Main control experiment.** Through the hyperparameter experiment and the model comparison experiment, we can determine the various parameters used in the experiment and the basic model of the experiment. The main control experiment in this paper is based on the BiLSTM-CRF model, which compares the effects of three different embedding methods: character embedding, word embedding and character and word embedding. The specific experimental results are shown in Table 5.

| Embedding Mode           | Accuracy | Precision | Recall | F1     |
|--------------------------|----------|-----------|--------|--------|
| Char_embedding           | 87.96    | 57.23     | 59.24  | 58.22  |
| Word_embedding           | 89.54    | 63.99     | 57.17  | 60.39  |
| Word&Char_embedding      | 90.88    | 66.72     | 64.17  | 65.42  |

From the experimental results, the F1 score based on char embedding is 58.22. The F1 score based on character and word embedding is 65.42, and the performance is the best. The word embedding method has a 6.76% improvement in precision compared to the char embedding method, because the basic meaning unit of Chinese is vocabulary, and the vocabulary carries more information than the word carries, so the model can accurately learn the true meaning of vocabulary. The char embedding method has a 2.07% improvement in the recall rate than the word embedding method. Since the number of vocabulary in Chinese is much larger than the number of words, the OOV problem is easy to appear in the word-based model, and the character embedding method alleviates the problem of OOV.

The model proposed in this paper combine the advantages of word embedding and character embedding. The precision rate is 2.73% higher than that of word embedding, and the recall rate is 4.93% higher than char embedding. F1 score is 7.20% higher than char embedding, and 5.03% higher than word embedding.
5. Conclusion
In this paper, for the NER task in the steel field, the data set is marked according to the text data, and
the data set of the NER task in the steel field is constructed. Based on the BiLSTM-CRF model, we
propose the char and word embedding method. Using CNN to extract features from char embedding.
Then, it joint with word embedding to get the final embedding. Without any prior data, experiments
have shown that our method yields the best results.

Of course, our method is not without drawbacks, and the increase in the amount of parameters will
take more training time. In the future, our work will expand the dataset in the steel field and automate
labeling. We intend to introduce more features, such as part of speech, pinyin, radicals, etc.

Acknowledgments
This work was financially supported by the Innovation Method Fund of China (No. 2016IM010300).

References
[1] Li C, Weng J, He Q, et al. (2012) TwiNER: Named entity recognition in targeted twitter stream. 
In: International ACM SIGIR Conference on Research & Development in Information Retrieval. Portland.ACM, 2012. 721-730.
[2] Lee J Y, Dernoncourt F, Szolovits P. (2017) MIT at SemEval-2017 Task 10: Relation Extraction
with Convolutional Neural Networks. Association for Computational Linguistics, SemEval-2017: 978-984.
[3] Lukovnikov D, Fischer A, Lehmann J. (2017) Neural Network-based Question Answering over
Knowledge Graphs on Word and Character Level.In: International Conference on World Wide
Web. Perth. 1211-1220.
[4] Zhou Kun. (2010) Research on rule-based named entity recognition. Hefei University of
Technology (in Chinese).
[5] Riaz K. (2010). Rule-based named entity recognition in Urdu. In: ACL 2010.Uppsala. 126-135.
[6] Fang X , Sheng H. (2002) A Hybrid Approach for Chinese Named Entity Recognition. Discovery
Science. 2534: 297-301.
[7] Rabiner L R . (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech
Recognition. Proceedings of the IEEE. 77: 257-286.
[8] Lafferty J, McCallum A, Pereira F. (2001)Conditional Random Fields: Probabilistic Models for
Segmenting and Labeling Sequence Data. International Conference on Machine Learning. 3: 282-289.
[9] Joshi A , Lal R , Finin T , et al. (2013) Extracting Cybersecurity Related Linked Data from Text.
In: IEEE International Conference on Semantic Computing. Irvine. 252-259
[10] Fu G , Luke K K . (2005) Chinese named entity recognition using lexicalized HMMs. ACM SIGKDD Explorations Newsletter. 7(1):19-25.
[11] Li L, Ding Z, Huang D, et al. (2008) A Hybrid Model Based on CRFs for Chinese Named Entity
Recognition. In: International Conference on Advanced Language Processing & Web
Information Technology. Dalian. 127-132.
[12] Sun J, Yu H, Feng Y, et al. (2018) Named Entity Recognition in Fisheries Based on Deep
Learning. Journal of Dalian Ocean University, 33(02): 132-136.
[13] Huang Z , Xu W , Yu K . (2015) Bidirectional LSTM-CRF Models for Sequence Tagging.
Computer Science. arXiv:1508.01991.
[14] Hochreiter S, Schmidhuber J. (1997) Long short-term memory. Neural Computation, 9(8):1735-
1780.