Sequence labeling model based on hierarchical features and attention mechanism

Lijuan Yao¹*, Yanfen Cheng¹ and Chao Li¹

¹ College of Computer Science and Technology, Wuhan University Of Technology, Wuhan, Hubei, 430063, China

*Corresponding author’s e-mail: yaolj@whut.edu.cn

Abstract. Sequence labeling is a basic task in natural language processing, which is of great help to processing text information. Conventional sequence labeling approaches heavily rely on hand-crafted or language-specific features, which requires a lot of time. Therefore, most of the existing methods are based on the BiLSTM-CRF model, but how to use a neural network to extract useful information for each unit or segment in the input sequence becomes the main factor limiting the efficiency. Several BiLSTM-CRF based models for sequence labeling have been presented, but the major limitation is how to use neural networks for extracting useful representations for each unit or segment in the input sequence. In response to this problem, this paper proposes a sequence labeling algorithm based on hierarchical features and attention mechanism, which uses a hierarchical structure to integrate character-level and word-level information, and applies different attention mechanisms to these two layers of information. According to the structural characteristics of different levels, excavate more potential information. Finally, the previously captured and guided features are used for sequence tag prediction using CRF. Finally, the proposed model is subjected to comparative experiments and the results obtained are analyzed.

1. Introduction
The sequence labeling task is a basic task in the field of natural language processing, which is mainly to label each unit in the sequence with the corresponding label. This kind of labeling task can be regarded as a combination of several independent classification tasks. Common sequence labeling tasks include name entity recognition, part-of-speech, semantic role labeling, and chunking analysis. Most traditional high-performance sequence labeling methods are based on basic statistical machine learning models, such as Hidden Markov Model (HMM), Maximum Entropy Model (MEM), etc. Although traditional methods based on supervised learning have achieved great success, these methods require a lot of engineering skills and domain expertise to design hand-crafted features. With the rise of deep learning, task feature representation methods based on neural networks have been extensively studied.

Traditional high-performance sequence annotation models are mostly linear statistical models, including HMM, MEM, and CRF. These models rely heavily on artificial features and task-specific resources. However, the cost of knowledge development for this specific task is very high [1], making it difficult for the sequence labeling model to adapt to new tasks or new fields. With the development of neural networks, deep learning methods have gradually been applied to sequence labeling tasks. With the development of deep learning, many studies are devoted to improving the efficiency of sequence labeling algorithms by automatically extracting features from different types of neural...
networks [2], in which various features of word information are encoded as distributed representations of the input, and sentence level The context means that it is learned during end-to-end training. Many recent neural network-based sequence labeling models use variants of recurrent neural networks (RNN), such as LSTM and gated recurrent units (GRU), to solve the vanishing gradient problem. At present, the method of learning sequence labeling with BiLSTM has become the mainstream. Huang et al. [3] proposed a sequence labeling model based on BiLSTM-CRF, using BiLSTM and CRF in the output layer, and considering adjacent labels when predicting labels. The design of BiLSTM-CRF considers the long correlation between the input word sequence and BiLSTM, as well as the internal tag correlation. Many recent models have proposed variants of BiLSTM-CRF. Ma [4] proposed the BiLSTM-CNNs-CRF model. The BiLSTM layer is used to capture the context information of the sentence, the CNN layer is used to capture the morphological information of the word, and the CRF layer is used to capture the dependency between tags.

Most of the existing methods are based on the BLSTM-CRF framework to achieve simple integration with the attention mechanism, but still have defects such as local dependence and inaccurate character information acquisition. Therefore, this paper proposes a sequence labeling model based on hierarchical features and attention mechanism, which can better capture the semantic dependence information, make up for the local dependence of LSTM modeling, and obtain better results in sequence labeling.

2. Materials and Methods
Sequence labeling refers to the process of labeling the corresponding label at each position of the sequence given an input sequence. This section will introduce the sequence labeling model based on hierarchical features and attention mechanism. The proposed model contains three levels, namely character-level coding layer, word-level coding layer and label decoding layer. The bottom character-level encoder is used to extract the character-level information of each word, and use the attention mechanism to learn the attention weight of each character. The extracted character-level information is connected with the word embedding vector to form the input of the word-level encoder. Then, the word-level encoder is used to extract word-level information from each input vector. Aiming at the long-distance semantic loss problem of the cyclic neural network, the multi-head attention mechanism is introduced to model the semantic relationship between any two words in the sentence. Finally, the previously extracted features are sent to the CRF model for label prediction.

2.1. Character-level coding layer
In natural language processing tasks, a vocabulary is usually loaded or customized in advance. If after loading a data set, there are some words in the data set that are not in the existing vocabulary, then these words are called Out-of-vocabulary, or OOV for short. For sequence labeling tasks, such as named entity recognition (NER), OOV words are very common. If a word is not found in the pre-trained word vector, the model will give the word a general OOV representation for trade-off. For English texts, certain morphological features of words, such as prefixes and suffixes, contain a lot of hidden information, and they are more likely to represent the part-of-speech tags of these words. For example, "ing" and "ed" are important features of adjectives, but word-level The encoder easily ignores these characteristic information. Therefore, it is necessary to use a character-level encoder to capture the morphological information of words. The character-level encoder is shown in Figure 1.

Given a sentence, suppose the first word in the sentence. The single-layer BiLSTM is used to extract surface features. In order to extract deeper information features, this section uses a two-layer neural network to extract character-level features. The character-level encoder first uses a BiLSTM to obtain the context vector from the input sequence, calculates the representation of each character, and each character is represented as the forward and backward connection of the bottom BiLSTM on the original character input sequence. Then use another BiLSTM to calculate the representation of each word to get the character-level word vector representation. Then directly connect the final state in the two directions.
The hidden variables obtained after the forward and backward LSTM are spliced to obtain the hidden variable $h_t$ of the $i$-th sequence at time $t$. In order to perform feature extraction on important character-level information, an attention mechanism is used here. The hidden variable $h_t$ undergoes a nonlinear transformation to obtain its implicit representation $u_{it}$, as shown in formula (1).

$$u_{it} = \tanh (W_w h_{it} + b_w) \tag{1}$$

The importance weight of the current character $\alpha_{i}$ is the similarity between $u_{it}$ and the character-level attention mechanism matrix $w_u$. The point multiplication operation of $w_u$ and $u_{it}$ is performed and the softmax function is used to normalize it to obtain the weight coefficient $\alpha_{it}$, as shown in formula (2) Shown.

$$\alpha_{it} = \frac{\exp(u_{it}^T + u_w)}{\sum \exp(u_{it}^T + u_w)} \tag{2}$$

After that, based on the corresponding weight $\alpha_{it}$, the word representation $s_i$ is calculated as the weighted sum of hidden variables $h_w$ at each moment, as shown in formula (3).

$$s_i = \sum \alpha_{it} h_w \tag{3}$$

![Figure 1. Character-level encoder](image-url)

### 2.2. Word-level coding layer

LSTM is a chain structure, although local context information can be obtained through BiLSTM, this sequential input processing method may limit the ability to capture the non-continuous relationships marked in sentences and weaken the semantic relevance between long-distance words. The potential relationship between words is not well extracted.

Therefore, the self-attention mechanism is introduced in this layer to enhance the semantic interaction between words. Given the current word, the attention mechanism can capture relevant information from a global scope, without being restricted by the sequential method. In order to learn diversified semantic relationships, a multi-head attention mechanism is used to model the semantic relationship between any two words in a sentence.
First, the matrices $Q$, $K$, and $V$ are linearly mapped to the K-dimensional subspace, and the calculation formula for each subspace is:

$$\text{head}_i = \text{Attention}(Q W^i, K W^i, V M_i)$$  \hspace{1cm} (4)

$W^i_j \in R^{2d_i \times d_i}$, $W^m_i \in R^{2d_i \times d_i}$, $W^r_i \in R^{2d_i \times d_i}$, $Q=K=V=H$. However, the corresponding weight matrix is still set to be different, in order to obtain a richer feature expression. Using the dot product computer system, the formula is:

$$\text{Attention}(Q, K, V) = \text{soft max} \left( \frac{Q K^T}{\sqrt{d_i}} \right) W$$  \hspace{1cm} (5)

Then all the attention heads are spliced together to get:

$$Z = \text{Concat}(\text{head}_1, ..., \text{head}_l) W^r$$  \hspace{1cm} (6)

### 2.3. Label decoding layer

For sequence labeling tasks, there is usually a certain dependency between adjacent labels. For example, in part-of-speech tagging, an adjective is more likely to be followed by a noun, rather than a verb. Therefore, this section uses the CRF model to model the tag sequence instead of decoding each tag independently.

CRF defines a transition matrix $A$ to model the interaction relationship between adjacent tags, and defines a state matrix $P$ to model the interaction relationship between words and tags. Suppose $Y = (y_1, y_2, ..., y_n)$ it is a label sequence, $P$ is a matrix of $n \times k$, $n$ is the length of the sequence, $k$ is the number of different labels, and the transposition of the $t$-th row is the vector $m_t$ after the above feature fusion. $P_{ij}$ represents the possibility that the $i$-th word is the $j$-th label. The predicted value of the defined label sequence is:

$$s(X, Y) = \sum_{t=0}^{n} A_{y_t, y_{t+1}} + \sum_{t=1}^{n} P_{y_t}$$  \hspace{1cm} (7)

Perform a calculation on all possible label sequences to obtain a conditional probability of the current label sequence:

$$P(Y | X) = \frac{e^{s(X,Y)}}{\sum_{Y \in Y} e^{s(X,Y)}}$$  \hspace{1cm} (8)

Maximize the log probability of the correct sequence by function:

$$\log(p(Y | X)) = s(X, Y) - \log(\sum_{Y \in Y} e^{s(X,Y)})$$  \hspace{1cm} (9)

Among them, $Y$ is all possible sequence labels of sentence $X$ in the exponential space. By maximizing the above formula, the model will learn the correct label sequence. Since optimizing the above formula in the exponential space is an NP problem, for reference, McCallum et al. [10] used forward and backward algorithms in linear conditional random fields to effectively solve this problem. During the test, the model uses the maximum posterior probability to predict the output:

$$Y' = \arg\max_{Y \in Y} p(Y | X)$$  \hspace{1cm} (10)

### 3. Results & Discussion

In the above, the sequence labeling algorithm based on hierarchical features and attention mechanism is studied, and the detailed calculation process is given. In order to evaluate the effect of the model,
this section will conduct experiments on the subtasks of English sequence labeling, naming Entity recognition task.

In order to verify the validity of the model, this article will conduct comparative experiments with other models. The following is a brief introduction to each model.

1) charLSTM-BiLSTM-CRF: a sequence labeling model based on character-level information, which is used as the experimental baseline model in this article.

2) MEM: Maximum entropy model, a maximum likelihood method for automatically constructing a maximum entropy model, using similar sparse features in the MEM model.

3) Neural sequence labeling model (NSLM): A hierarchical LSTM-LSTM-CRF neural network with an attention mechanism. The attention mechanism is used to weight and sum the original character vector and the pre-trained word vector instead of directly performing the connection.

4) BiLSTM-CNN-CRF: This model uses the CNN layer to capture the morphological information of the word, the BiLSTM layer captures the semantic information of the word, and finally uses a CRF layer to predict the label.

5) Bert: It is a pre-trained language model that can generate deep two-way language representations.

The model proposed in this chapter is represented by HASL. Table 2-4 shows the performance of the test model on the CoNLL2003 dataset in the named entity recognition task.

| Model                        | F - Score |
|------------------------------|-----------|
| charLSTM-BiLSTM-CRF(baseline) | 91.28     |
| MEM                          | 82.74     |
| NSLM                         | 84.69     |
| BiLSTM-CNN-CRF               | 91.21     |
| Bert                         | 91.40     |
| HASL                         | **91.73** |

It can be seen from the above table that the MEM method does not perform well on this data set, because its performance depends to a large extent on feature engineering. The HASL model proposed in this chapter has the best overall performance. Compared with the word-level neural models NSLM and Bert, the HASL model can capture more character-level features, which indicates that character-level information is essential for sequence labeling tasks. Information such as prefixes and suffixes and capitalization of words can improve the accuracy of sequence labeling tasks. Compared with the baseline model, the model proposed in this chapter has increased by 0.45, which shows that the introduction of the attention mechanism is of great help in improving performance. In named entity recognition, capital letters are often a significant sign of a person's name. By introducing an attention mechanism, capital letters can gain more attention weight and improve the accuracy of recognition. At the same time, using the multi-head attention mechanism in the word layer can model the relationship between discontinuous words, which is of great help to the recognition of named entity types.

4. Conclusions

This paper proposes a sequence labeling method based on hierarchical features and attention mechanism. First, character-level related information is obtained through the character-level encoding layer, and the attention mechanism is used to weight important characters. After splicing, it is sent to the word-level coding layer as input. The multi-head attention mechanism is introduced in the word-level coding layer to solve the long-distance dependence problem of the LSTM network, and to better model the dependence between words in the sequence. Finally, the previously extracted features are sent to the label decoding layer for sequence label prediction. Finally, in order to verify the effectiveness of the model proposed in this chapter, experiments are carried out on the task of named entity recognition, and a certain degree of improvement has been achieved.
Acknowledgments
National Social Science Foundation Project (15BGL048)

References
[1] Ma X,Xia F.(2014) Unsupervised Dependency Parsing with Transferring Distribution via Parallel Guidance and Entropy Regularization.In:52nd Annual Meeting of the Association for Computational Linguistics .New York. pp. 165-178.
[2] X. Ma, E. Hovy, End-To-End Sequence Labeling Via Bi-Directional LSTM-CNNs-CRF, The Association for Computational Linguistics, 2016, pp.1064–1074.
[3] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging.in: Conference on Empirical Methods in Natural Language Processing , 2015. pp.1064–1074.
[4] Y. Liu, W. Che, J. Guo, Q. Bin, T. Liu, Exploring segment representations for neural segmentation models, in: International Joint Conference on Artificial Intelligence, 2016, pp. 2880–2886.