A New Method for Abbreviation Prediction via CNN-BLSTM-CRF

To cite this article: Jianyu Zheng et al 2019 J. Phys.: Conf. Ser. 1267 012001

View the article online for updates and enhancements.
A New Method for Abbreviation Prediction via CNN-BLSTM-CRF

Jianyu Zheng\textsuperscript{1,3}, Xinge Xiao\textsuperscript{2}, Bihua Wang\textsuperscript{1,3}, Yun Zhu\textsuperscript{3} and Lijiao Yang\textsuperscript{1,3,*}

\textsuperscript{1}Institute of Chinese Information Processing, Beijing Normal University, China
\textsuperscript{2}Archives of Central China Normal University, Central China Normal University, China
\textsuperscript{3}UltraPower-BNU Joint Laboratory for Artificial Intelligence, Beijing Normal University, China

*zheng\_jianyu@126.com

Abstract. It is a crucial problem to process abbreviation in the field of natural language processing. The most commonly used way to cope with this problem is to construct the reference database by predicting the abbreviation through its fully expanded form. Previous work on abbreviation prediction mostly rely on traditional machine learning algorithms, which inevitably requires a large number of manual annotations or expert knowledge to establish a feature system. In this paper, a neural network model based on CNN-BLSTM-CRF is proposed, which can predict Chinese abbreviations better without relying too much on the feature system: Firstly, convolutional neural network extracts phrase and Chinese character information from the fully expanded form, and then BLSTM-CRF deep network is constructed to annotate the fully expanded form, so as to extract its corresponding abbreviation form. The experimental results show that the method in this paper can perform better than the state-of-art method in traditional machine learning, and the results provide a reference for abbreviation research and the construction of resource repository.

1. Introduction
Abbreviations are usually obtained by compressing and omitting longer words or phrases, such as "AI" is the abbreviation of "Artificial Intelligence" and "Bei Da" (北大) is the abbreviation of "Bei Jing Da Xue" (北京大学). Abbreviations have more concise language expression, so that they are widely used in natural language.

On the other hand, the use of abbreviations also brings some challenges to Chinese information processing. Therefore, it is helpful to the study of Chinese Word Segmentation, POS Tagging, Named Entity Recognition, Machine Translation and Information Retrieval, etc. by processing abbreviations correctly [1]. A commonly used way to cope with the challenge is to construct a large-scale fully expanded form-abbreviation reference database. The key to construct the database is to predict abbreviations through its fully complete form.

Scholars have made many efforts on Chinese abbreviations prediction. The most representative one is Xu Sun, Hou-Feng Wang, Bo Wang (2008), which used SVR to predict abbreviations: different abbreviation candidates were scored and rearranged, and the candidate with the highest score was the final choice [2]. Jiao Yan, Wang Houfeng, Zhang Longkai [3] used CRF model and Web information to predict abbreviation; Xu Sun, Wenjie Li, Fanqi Meng, Houfeng Wang [4] starting with Negative Full
Forms (NFFs), used Simple Heuristic System and Unified System to improve the accuracy of abbreviation prediction. In addition, Xu Sun, Naoaki Okazaki, Jun'ichi Tsujii [5] tried to introduce hidden variables on the basis of sequence annotation, which finally achieved good performance. Yi Zhang, Xu Sun [6] proposed a novel abbreviation prediction dataset with NFFs and evaluated it using CRF and BLSTM models.

Inspired by previous work, this paper proposes a new method based on neural network to predict abbreviations, which combines CNN, BLSTM and CRF together. Firstly we randomly generate a character embedding table and transform the fully expanded form into a character embedding matrix by looking up the table at first. Then, CNN and BLSTM are used to extract more efficient and high-dimension features and recognize characters in sequences that may exist in abbreviations respectively. Finally, CRF is used to decode the character sequence and obtain the optimal one. Through experiments, we find that this new method can improve the accuracy of abbreviation prediction, for instance, all accuracy is 0.62% higher than the state-of-the-art method of previous work. The main contributions of our work are:

1. proposing a new neural network structure for abbreviation prediction;
2. detailed discussion on the role of each module in the network;
3. improving the accuracy of abbreviation prediction to a certain extent.

2. Problem definition

There are many ways to form Chinese abbreviations, which can be summarized as Reduction (such as "Bei Jing Da Xue-Bei Da" (北京大学-北大)), Elimination (such as "Qing Hua Da Xue-Qing Hua" (清华大学-清华)) and Generalization (such as "Hei longjiang Sheng, Jilin Sheng, Liaoning Sheng- Dong San Sheng" ("黑龙江省、吉林省、辽宁省-东三省"). However, even if there are a limited number summed up in abbreviation, the actual situation still remains complicated. As far as the "Reduction" method is concerned, unlike English and other alphabetic writings, which form abbreviations usually by combining initial letters, the way of Reduction in Chinese is much more complicated. It is unnecessary to take the first Chinese character of each word to form a new abbreviation, for example, the "Li Shi/Yu Yan/Yan Jiu Suo" is abbreviated to "Shi Yu Suo" (历史/语言/研究所-史语所), which is related to the fact that Chinese characters have a certain meaning but alphabets do not represent special concepts. Moreover, the relationships between the fully expanded forms and its abbreviations are not always one-to-one relationship. For example, "Ren Da" (人大) can not only refer to "Ren Min Dai Biao Da Hui" (人民代表大会) but also "Ren Min Da Xue" (人民大学) [7]. In short, because the formation of abbreviations is influenced by many different factors, it is difficult for us to seek out the rules that meet all situations.

Therefore, the problem of abbreviation prediction is often transformed into a sequence labeling problem. We can make different marks on fully expanded forms according to its different characteristics. After marking the fully expanded forms, we can choose appropriate model to predict the corresponding abbreviations eventually.

Figure 1 shows a mark method commonly used in the previous research of abbreviation prediction:

```
| 语言研究所 |
|----|----|----|----|
| P  | P  | S  | S  | P  |
```

**Figure 1.** Chinese abbreviation generation as a sequential labeling problem.

In figure 1, each Chinese character is assigned as P (produce the current character) or S (skip the current character). The Chinese character marked P will be output finally, while the Chinese character marked S will be skipped. So the abbreviations corresponding to the fully expanded forms will be successfully generated.

Obviously, it is difficult to train an ideal model simply by using that labeling method. Xu Sun, Naoaki Okazaki, Jun'ichi Tsujii [5] took the lead by introducing Global Information labelling (GI) in the study.
of abbreviation prediction. Specifically, as shown in figure 2, this labeling method can better consider the global information. Experiments show that it improves the accuracy of abbreviations prediction.

According to the description of the paper, the label \( y_i \) at position \( i \) attaches the information of the abbreviation length generated by its previous labels, \( y_1, y_2, \ldots, y_{i-1} \). In this encoding, a label not only contains the produce or skip information, but also the abbreviation-length information, i.e., the label includes the number of all P labels preceding the current position [5].

In this paper, we not only use these two labeling methods, but also consider the phonetic information, tone information, location information, repetition of adjacent or separated characters and other features of each Chinese character in the fully expanded form.

3. CNN-BLSTM-CRF model

In this section, we will introduce our network model proposed in this paper. We will explain each layer of this network in top-down order.

3.1. Neural network architecture

Figure 3 shows the framework of CNN-BLSTM-CRF Model, which consists of three modules: CNN module, BLSTM module and CRF module. First, a table of character embedding is generated randomly, and the fully expanded form is transformed into the corresponding character embedding matrix by looking up the table. This matrix is convolved twice and pooled once by CNN to extract the higher dimensional features of the sequence based on the character level. A layer of BLSTM is added to recognize the characters in the sequence that may exist in the abbreviations. Finally, the output of BLSTM is decoded by CRF module, and the optimal label sequence is obtained.

![Figure 3. Model architecture diagram of CNN-BLSTM-CRF.](image)

3.2. CNN module

The advantage of convolutional neural networks is to better extract local features of the data. Chiu and Nichols et al. [8] used CNN to extract the morphological information of characters or words and used them in the task of named entity recognition. In the CNN module, the specific network architecture is shown in figure 4:
In figure 4, each fully expanded form is transformed into a character embedding matrix by looking up table according to the character. In order to solve the problem of different size of character embedding matrix caused by different length of fully expanded form. This paper adopts the principle of “cutting off from the long and adding to the short”. That is, the full expanded form exceeding the set length will be truncated from the left, and the full expanded form that does not reach the set length will be supplemented with paddings at the left, so that all character embedding matrices are uniform in size. Finally, as shown in Char Embedding in the figure 4, the character embedding matrix is then input into the first convolution layer. Because two-character words appear more frequently in the problem of abbreviation prediction, the kernel size is set to 2. So that it can extract effective information better regardless of whether it is reduction or elimination. Next, the output of the first convolutional layer is treated as input to the next layer. The purpose of this convolution layer is to extract the corresponding character in the two-character word which is more likely to be retained in abbreviation, so the kernel size is set to 1. Finally, the important features in the sequence are better preserved and invalid information is filtered through the maxpooling layer.

3.3. BLSTM module

Recurrent neural network is a kind of depth model for simulating sequential data. Because the traditional recurrent neural network can not solve the “long-range dependence” problem of sequence data well, the phenomenon of gradient disappearance or gradient explosion will occur in practical [9].

In order to make up this disadvantage, Hochreiter et al. proposed Long short-term memory network [10]. LSTM allows the model to selectively retain contextual information through a specially designed gate structure. In general, the formalized structure of an LSTM unit at time t is as follows:

\[
\begin{align*}
    i_t &= \sigma(W_i h_{t-1} + U_i x_t + b_i) \\
    f_t &= \sigma(W_f h_{t-1} + U_f x_t + b_f) \\
    \tilde{c}_t &= \tanh(W_c h_{t-1} + U_c x_t + b_c) \\
    c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
    o_t &= \sigma(W_o h_{t-1} + U_o x_t + b_o) \\
    h_t &= o_t \odot \tanh(c_t)
\end{align*}
\]

Where, \(\sigma\) is the activation function sigmoid; \(\odot\) is point multiplication operation; \(\tanh\) is the hyperbolic tangent function; \(x_t\) is the input vector at time \(t\); \(h_t\) is the hidden state of the LSTM unit at time \(t\); \(i_t, f_t, o_t\) represent the input gate, the forget gate and the output gate at time \(t\); the matrix \(U\) corresponds to the weight matrix of the input vector \(x_t\) in each gate; \(W\) and \(b\) respectively correspond to
the weights and biases of the hidden states in each gate; \( \tilde{c} \) is an intermediate state obtained only according to the input at time \( t \), which is used to update the current state; \( c_t \) represents the state of time \( t \) [11].

In traditional LSTM networks, the implicit state can only deal with forward transmitted sequence data. In this paper, to make better use of contextual information, we have adopted bidirectional LSTM (BLSTM) [12]. The basic idea is to obtain two different hidden states for each fully expanded form in sequence and inverse order respectively to represent the past and future information of the sequence. Finally, the two hidden states are concatenated together as the output and passed to the next module.

### 3.4. CRF module

In the task of sequence labeling, CRF can obtain a globally optimal label sequence by considering the relationship between adjacent labels. For example, in POS task, adjectives are always followed by nouns rather than verbs. Specifically, the output of BLSTM is used as an input to the CRF layer for further solution. \( X = \{x_1, x_2, \cdots, x_n\} \) is assumed to be a full expanded form of the input sequence vector, where \( x_n \) is the vector representation of the \( n \)th Chinese character in the sequence after processing by the previous module. \( Y = \{y_1, y_2, \cdots, y_n\} \) represents the corresponding label sequence, \( y_n \) is the label corresponding to the \( n \)th Chinese character \( x_n \). Then, CRF probability model can be regarded as calculating the probability \( p(y | x; W, b) \) by the corresponding label \( y \) under the condition of a given variable \( x \). The formula is as follows:

\[
p(y | x; W, b) = \frac{\prod_{i=1}^{n} \psi(y_{i-1}, y_i, x)}{\sum_{y \in \mathcal{Y}} \prod_{i=1}^{n} \psi(y_{i-1}, y_i, x)} \tag{2}
\]

Where \( \psi(y', y, x) = \exp(W_{y',y}^T x_i + b_{y',y}) \) is a potential function, \( W_{y',y}^T \) and \( b_{y',y} \) represents the weight vector and bias of the tag pair \( (y', y) \), respectively. The formula for its log-likelihood function is as follows:

\[
L(W, b) = p(y | x; W, b) \tag{3}
\]

The optimal solution of the formula can be obtained by maximum likelihood estimation. Finally, the label sequence \( Y^* \) corresponding to the optimal solution can be obtained by decoding.

In summary, the specific network architecture of the BLSTM and CRF modules is shown in figure 5.

![Figure 5](image-url)

**Figure 5.** The specific network architecture of CNN module and CRF module.
4. Experiment

4.1. About the corpus

In this paper, 12338 abbreviations and their corresponding fully expanded forms are extracted from the Chinese Abbreviations Dictionary [13] and Practical Abbreviations Knowledge Dictionary [14]. Among them, 11112 abbreviations belong to reduction and elimination, which can be used in this experiment. In this paper, the experimental data are shuffled, the training set and test set are divided according to 4:1.

4.2. Experimental settings

About the optimization algorithm, this paper uses Adam as the optimizer, and the learning rate is set to 0.001. Table 1 shows the parameters of the entire network architecture. Besides, the parameters of the network can be fine-tuned according to the actual situation.

| Hyper-parameter      | Value       |
|----------------------|-------------|
| Embedding            | char_embedding+BMEU 200+200 |
|                      | max_sequence_length 12 |
| Conv1(1_D)           | window size 2 |
|                      | number of filters 250 |
|                      | stride 1 |
|                      | padding mode same |
| Conv1(1_D)           | window size 1 |
|                      | number of filters 400 |
|                      | stride 1 |
|                      | padding mode same |
| MaxPooling(1_D)      | window size 2 |
|                      | stride 1 |
|                      | padding mode same |
| BLSTM                | cell size 350 |
|                      | drop out 0.4 |

4.3. Results

To assess the effectiveness of this model, we will measure it from two indicators, all-match accuracy and character accuracy, as explained below:

- all-match accuracy: The number of correct outputs generated by the system divided by the total number of full forms in the test set [5].
- character accuracy: The number of correct labels generated by the system divided by the total number of characters in the test set [5].

We compare our model with the previous state-of-art method [5]. The operation of CRF is the same as the previous method in the experiment, considering the complete form of pinyin, tone, whether the corresponding Chinese character is a Chinese number, whether the adjacent Chinese character is the same, whether it is the same after a Chinese separated character, and the position information (BMEU feature) of each character in the word after the segment. The specific testing results are as follows in Table 2:

| Model                  | all acc(%) | char acc(%) |
|------------------------|------------|-------------|
| CRF                    | 37.98      | 79.48       |
| CNN-BLSTM-CRF          | 38.66      | 76.16       |
| CRF+GI                 | 46.4       | 79.64       |
| CNN-BLSTM-CRF+GI       | 47.02      | 78.45       |
In addition, to verify the validity of our model, Table 3 shows the effects of different modules on abbreviation prediction when used separately:

**Table 3. Prediction effect table of different modules of Neural Network.**

| Model                   | all acc(%) | char acc(%) |
|-------------------------|------------|-------------|
| LSTM                    | 31.99      | 75.09       |
| BLSTM                   | 33.71      | 75.89       |
| CNN-BLSTM               | 35.28      | 76.1        |
| BLSTM-CRF               | 35.61      | 76.15       |
| LSTM+GI                 | 40.96      | 76.66       |
| BLSTM+GI                | 42.03      | 77.18       |
| CNN-BLSTM+GI            | 43.69      | 77.97       |
| BLSTM-CRF+GI            | 44.5       | 77.56       |

4.4. **Analysis**

It can be seen from Table 2 that compared with the original method, the model proposed in this paper has a certain improvement on all acc. This also shows that the end-to-end neural network model can avoid the complicated feature engineering in the traditional machine learning, and can extract some implicit high-dimensional features from the original data through the nonlinear operation of the network. It is helpful to predict the final sequence.

The effect of BLSTM in Table 3 is significantly higher than that of LSTM, which indicates that BLSTM can make full use of the context information of the sequence. CNN-BLSTM has a slightly better effect on all acc than BLSTM, and it also shows the effectiveness of CNN in extracting character-level features. In addition, BLSTM-CRF is 2% better than BLSTM in all acc, which means that linear CRF can make full use of the relationship between adjacent labels and output an optimal label sequence.

It can be seen from Tables 3 and 4 that the corresponding models are greatly improved compared with the original ones after the labels are encoded with Global Information. This is because the label incorporates non-local information when encoded, and the non-local information can not only confirm whether the current character exists in the abbreviation, but also reflect the relative position information of the character in the entire sequence [5].

It can also be seen through experiments that this model performs worse than the traditional CRF algorithm in terms of char acc indicator. This is due to the “cutting off from the long and adding to the short” operation on the fully expanded form of the input sequence: If we take the longest sequence in the whole data set as the standard to complete all other sequence, most of the sequences in the corpus will be filled with a large number of paddings, thus the prediction effect of the model will decline. However, to some extent, the truncated part of the long sequence can not be labeled by the “cutting off from the long and adding to the short” method, which results in a bad char acc result.

Finally, the experimental results in Table 2 and 3 are low mainly due to uneven data. The abbreviations are collected from Chinese Abbreviations Dictionary and Practical Abbreviations Knowledge Dictionary which cover short name, abbreviations of organization names and omissions of fixed phrases, and the distribution of abbreviations is different. As a result, there are some differences in the distribution of abbreviations between the training set and the test set. In addition, the dataset also contains some long fully expanded form, among which the longest sequence contains 49 Chinese characters.

5. **Conclusion**

On the basis of previous researches, we propose a new abbreviation prediction model combined with CNN, BLSTM and CRF to obtain better prediction results. In experiments, we also use abbreviation corpus to evaluate some methods of abbreviation prediction proposed by predecessors and the CNN-BLSTM-CRF model with Global Information shows competitive performance. However, the amount of data in the abbreviation corpus is insufficient if we want to further get good results. To solve this
problem, we will try to increase the amount of data sets, improve our model and further enhance the effect of abbreviation prediction in our future work.

Acknowledgments
This work is supported by National Language Committee Research Program of China (No.ZDI135-42) and National High-Tech Research and Development Program of China (2012AA011104).

References
[1] Jiao Y, Wang H F, 2011, Abbreviation Prediction Using Machine Learning Method and Search Engine Verification, Proceedings of The Tenth China National Conference on Computational Linguistics, August 20, 64-69
[2] Sun X, Wang H F, Wang B, 2008, Predicting Chinese Abbreviations from Definitions: An Empirical Learning Approach Using Support Vector Regression, Journal of Computer Science & Technology, Vol 23, Issue 4, 602-611
[3] Jiao Y, Wang H F, Wang H F, Zhang L K, 2012, Abbreviation Prediction Using Conditional Random Field and Web Data, Journal of Chinese Information Processing, Vol 26, Issue 2, 62-69
[4] Sun X, Li W J, Meng F J, Wang H F, 2013, Generalized Abbreviation Prediction with Negative Full Forms and Its Application on Improving Chinese Web Search, International Joint Conference on Natural Language Processing, October 14-18,641-647
[5] Sun X, Okazaki N, Tsujii J,2009 Robust Approach to Abbreviating Terms: A Discriminative Latent Variable Model with Global Information, Proceedings of the Meeting of the Association for Computational Linguistics, August 2-7 905-913
[6] Zhang Y, Sun X, 2017, A Chinese Dataset with Negative Full Forms for General Abbreviation Prediction, Computer Science, 1-5
[7] Wang H F, 2011, Survey: Abbreviation Processing in Chinese Text, Journal of Chinese Information Processing, Vol 25, Issue 5, 60-68
[8] Chiu J P C, Nichols E, 2015, Named Entity Recognition with Bidirectional LSTM-CNNs, Transactions of the Association for Computational Linguistics, Vol 4, Issue 1, 1-14
[9] Pascanu R, Mikolov T, Bengio Y, 2013, On the difficulty of training recurrent neural networks, International Conference on International Conference on Machine Learning, June 16-21, 1310-1319
[10] Graves A, 1997, Long Short-Term Memory, Neural Computation, Vol 9, Issue 8,1735-1780
[11] Ma X, Hovy E, 2016, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, August 7-15,1064-1074
[12] Dyer C, Ballesteros M, Wang L, et al, 2015, Transition-Based Dependency Parsing with Stack Long Short-Term Memory, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, July 26-31 ,334-343
[13] Shi B Y, Xu Y W, 1990, Chinese Abbreviations Dictionary, Foreign Language Teaching and Research Press
[14] Ya K, Xiu Y, 1992, Practical Abbreviations Knowledge Dictionary, New World Press