# Research on Automatic Terminological Dictionary Extraction Based on Bilingual Parallel Corpus

**Fangting Liu***

*College of Foreign Languages, Bohai University, Jinzhou, 121013, China*
*liufangting1984@126.com*
**Corresponding author*

***Abstract:*** *Bilingual dictionaries have always been a basic resource in the fields of machine translation and cross-language information retrieval. The text in parallel corpus is a pair of sentences translated into each other. It is a set of text composed of the source language text and the corresponding translated text, which has strong text alignment characteristics. By making full use of rich corpus information in parallel corpora, a bilingual automatic term extraction system is constructed by designing reasonable algorithms to realize automatic or semi-automatic term extraction, which is used to solve the difficult problems in machine translation and cross-language natural language processing. Based on the method of deep learning, this paper constructs the LSTM model based on RNN. Based on LSTM, a new BLSTM model is formed, which has the advantages of comprehensive information, strong robustness, and the ability to take into account both front and back data in natural language processing, so that the trained machine can learn more abstract samples.*

***Keywords:*** *Bilingual Parallel Corpus; Terminological Dictionary; Automatic Extraction; BLSTM; Recurrent Neural Network*

## 1. Introduction

The advancement of the economic society and global communication has led to an increase in the use of multiple languages in written documents, consequently creating a growing demand for translation and retrieval services across different languages. Among them, bilingual dictionaries have always been a basic resource in the fields of machine translation and cross-language information retrieval. Therefore, the work of term extraction in bilingual dictionaries has become the focus of research. The text in parallel corpus is a pair of sentences translated into each other. It is a set of text composed of the source language text and the corresponding translated text, which has strong text alignment characteristics. The intertransliterability of parallel corpus provides corpus resources for automatic extraction of equivalent pairs in phrase translation, and plays an important role in machine translation, assisted machine translation and cross-language information retrieval [1,2]. As the core carrier of knowledge, the mutual translation of terms has become one of the biggest obstacles in the exchange of knowledge and technology between countries. This paper studies the automatic extraction of term dictionaries based on bilingual parallel corpora, which is of great significance for the construction of bilingual term dictionaries and cross-language retrieval.

## 2. The BLSTM Model

Recurrent Neural Network (RNN) is a kind of neural network that models and predicts sequence data, and is a very important model in the field of deep learning. RNN overcomes many limitations of traditional machine learning methods on modeling data and is widely used in a variety of tasks, such as speech recognition, machine translation, text classification, time series data prediction and other sequences dependent scenarios [3,4]. The BLSTM model used in the term dictionary automatic extraction algorithm studied in this paper is based on RNN and evolved from LSTM.

### *2.1 LSTM Model*

RNNS often suffer from gradient disappearance, or gradient explosion, which makes it impossible to achieve long sequence memory. After years of research, two scientists, Hochreiter and Schmidhuber, invented Long Short-Term Memory (LSTM). LSTM is a deformation on the basis of RNN, which changes the internal computing structure network and adds a memory unit to store the useful contents of previous sequences and apply them to subsequent sequences, solving the problem that RNN cannot realize the memory of long sequences [5,6].

LSTM adds "gate" to the original RNN model to control information transmission, to avoid gradient disappearance and explosion problems to a certain extent, so as to obtain the long-distance dependent information of text semantics better.

(1) Forgetting gate. The forgetting gate acts on the state of the memory cell under the previous unit, and the purpose is selective forgetting, forgetting the information in the memory cell is to select useful information and discard useless information.

$$i_t = \sigma\left(W^{(i)}x_t + U^{(i)}h_{t-1}\right) \tag{1}$$

(2) Input gate. The input gate is also a memory cell state, the purpose of which is to selectively record new information into the memory cell and transmit it to the next level.

$$f_t = \sigma\left(W^{(f)}x_t + U^{(i)}h_{t-1}\right) \tag{2}$$

(3) Output gate. The output gate acts on the input and the hidden output. After passing through the output gate, the final output, including neither the cell state nor the input, transmits the result to the next layer.

$$o_t = \sigma\left(W^{(o)}x_t + U^{(o)}h_{t-1}\right) \tag{3}$$

(4) New memory unit. The new memory unit uses the current word and the previous moment to hide the layer state, generating content that includes new information about the current word.

$$a_t = \tanh\left(W^{(c)}x_t + U^{(c)}h_{t-1}\right) \tag{4}$$

(5) Final memory unit. The final memory unit refers to the suggestions given by the forgetting gate and the input gate, appropriately forgets part of the past memory, controls the content of the new memory unit, and then adds the two results.

$$c_t = f_t \times c_{t-i} + i_t \times a_t \tag{5}$$

(6) Hidden layer. The nodes between the hidden layers of the recurrent neural network are connected, and the input of the hidden layer includes not only the output of the input layer, but also the output of the hidden layer at the previous time.

$$h_t = o_t \times \tanh\left(c_t\right) \tag{6}$$

The input gate uses the current word and the state of the previous hidden layer to determine how important the current word is, and thus how important the current word is for generating that new memory. The function of the forgetting gate is similar to the function of the input gate, to determine the importance of the past memory, and thus the extent to which the past memory is involved in the formation of new memories. The function of the output gate is to separate the contents of the final memory unit from the hidden layer state, and many contents of the final memory unit do not need to be passed to the hidden layer.

### *2.2 The BLSTM Model*

Although LSTM can solve the problem of gradient disappearance and long-term information dependence to a certain extent, because the information is one-way transmission, the degree of information loss will become obvious when the transmission time increases. Bidirectional Long Short-Term Memory (BLSTM) is a relatively new model based on LSTM, as shown in Figure 1.

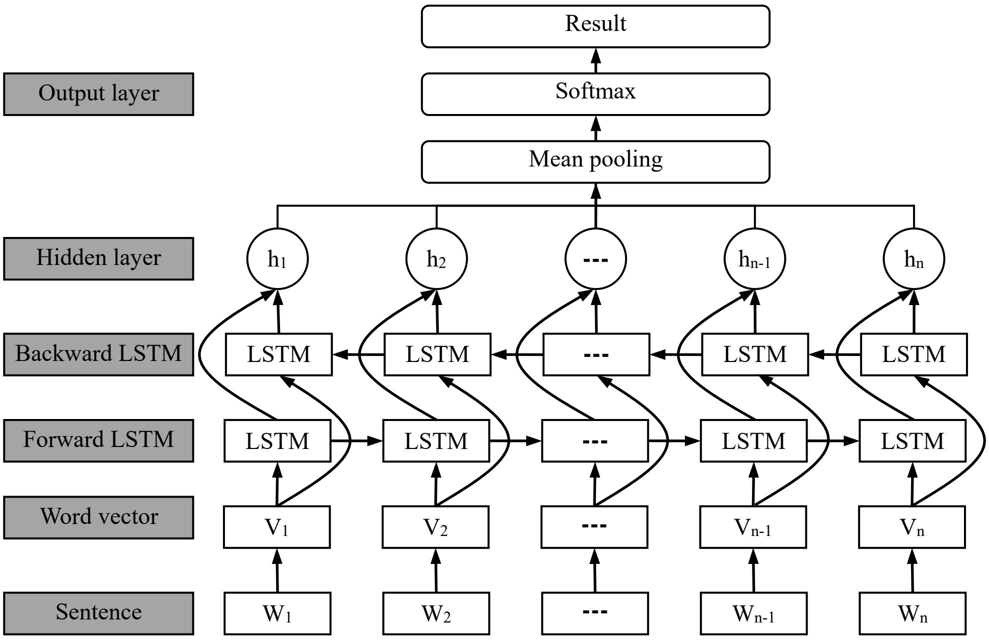

*Figure 1. BLSTM model structure*

BLSTM is composed of a forward propagating LSTM neural network and a negative propagating LSTM neural network. At the end of the network, there is an information fusion layer, which is reflected in the mathematical level as vectors and merges, so that the vectors fused by the information fusion layer contain the past and generalized information. Compared with multi-layer LSTM neural network, it has the advantages of comprehensive information, strong robustness, and the ability to take into account both front and back data in natural language processing, which enables the trained machine to learn more abstract samples [7].

### *2.3 Softmax Activation Function*

The activation function is the function that runs on the neurons of the artificial neural network and is responsible for mapping the input of the neuron to the output. Activation function is very important for artificial neural network models to learn and understand very complex and nonlinear functions. The activation function has the properties of nonlinearity, differentiability, monotonicity and range of output values. This article uses Sigmoid activation function, the general expression is:

$$f(x)=\frac{1}{1-e^{-x}} \tag{7}$$

Softmax regression is used to predict the vector of the target word and backpropagation is based on the loss function of the predicted vector and the actual vector. Hierarchical Softmax or Negative Sampling is usually introduced in the output layer to optimize the calculation. Hierarchical Softmax constructs a Huffman Tree using dictionary words and reduces the output dimension from dictionary dimension *V* to log (*V*) by mapping the output vector to the Huffman Tree.

For the BLSTM model, Softmax activation function is used as the classification result, and the activation formula of the *i*-th neuron is as follows:

$$S_i=\frac{e^{V_i}}{\sum_j e^{V_j}} \tag{8}$$

Softmax normalizes all category labels and takes the category label corresponding to the maximum probability as the output result. The training of the neural network uses backpropagation iteration to modify the weights, and the cross-entropy loss function is used for multi-classification problems. The cross entropy function is defined as:

$$C = -\frac{1}{n}\sum\nolimits_{x}\left[y\log a + (1-y)\ln(1-a)\right] \tag{9}$$

In the above formula, $n$ is the total number of training data, $a$ is the output of each neuron, and $y$ is the label corresponding to the data. The neural network uses gradient descent to update the internal weights. For the learning rate setting, this paper selects the Adadelta optimizer to update the weights.

### *2.4 Adadelta Optimizer*

In the field of machine learning and artificial intelligence, an optimizer is an important tool for tuning and optimizing the parameters of a model to improve its performance and accuracy. Adadelta optimizer is a common optimization algorithm that adaptively adjusts the learning rate to improve the effectiveness of model training. The Adadelta optimizer uses first-order derivative information, has good dynamic adaptability, and has less computational overhead than the original stochastic gradient descent algorithm. The Adadelta optimizer does not need to manually adjust the learning rate, and it shows strong robustness to the noise gradient information, different model structures, different data modes, and the selection of hyperparameters.

Instead of accumulating all past squared gradients, Adadelta limits the window in which it accumulates past gradients to some fixed size. Instead of storing previous squared gradients inefficiently, the sum of gradients is defined recursively as the decay average of all past squared gradients. The running average of the time step then depends only on the previous average and the current gradient:

$$E\left[g^2\right]_t = \gamma E\left[g^2\right]_{t-1} + (1-\gamma)g_t^2 \tag{10}$$

Set it to a value similar to the momentum item, and for clarity, rewrite the normal SGD update according to the parameter update vector:

$$\begin{cases}\Delta\theta_t = -\eta \times g_{t,i} \\ \theta_{t+1} = \theta_t + \Delta\theta_t\end{cases} \tag{11}$$

Adagrad's parameter update vector has the following form:

$$\Delta\theta_t = -\frac{\eta}{\sqrt{G_t + \varepsilon}} \times g_i \tag{12}$$

Simply replace the diagonal matrix with the decay average of the past squared gradient:

$$\Delta\theta_t = -\frac{\eta}{\sqrt{E\left[g^2\right]_t + \varepsilon}} \times g_i \tag{13}$$

## 3. Extraction Model Based on BLSTM

Based on a large-scale bilingual parallel corpus, a deep neural network is applied to the automatic extraction of a term dictionary. The model framework is shown in Figure 2.

In Figure 2, the word vector that has been trained in advance is divided into two parts: training set and test set. Firstly, the training set is introduced into the neural network for training, and the values of each parameter in the network are obtained. Then, the test set is imported into the trained extraction model for testing, and the bilingual dictionary is obtained and the extraction effect of the model is tested.

In the research process of this topic, the representation of text and words is no longer represented by the classical vector space model, but by the form of word vector, that is, the word vector is represented. Before training the initial parameters, the neural network generally needs to represent the words in the input layer vectorially. Since distributed word vectors contain rich semantic information, this project will use the pre-trained distributed word vectors as input and output. In the initial stage of training, the word vector contains rich information, that is, the input layer of the neural network contains rich information in the initial state. On this basis, the model training can significantly improve the training effect to a certain extent.

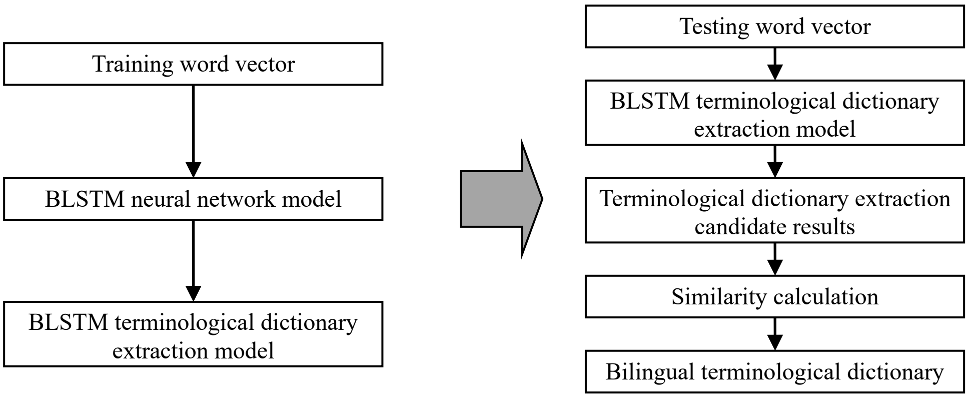
Training word vector
BLSTM neural network model
BLSTM terminological dictionary extraction model
Testing word vector
BLSTM terminological dictionary extraction model
Terminological dictionary extraction candidate results
Similarity calculation
Bilingual terminological dictionary

*Figure 2. Terminological dictionary extraction model based on BLSTM*

## 4. Similarity Calculation

There are many methods to calculate the similarity, this topic chooses Euclidean distance similarity, cosine similarity or modified cosine similarity according to the actual situation.

### *4.1 Euclidean Similarity*

Euclidean distance, also known as Euclidean metric, refers to the true distance between two points in M-dimensional space, or the natural length of a vector

The Euclidean distance in two-dimensional space is calculated as follows:

$$d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{14}$$

The calculation formula of Euclidean distance in three-dimensional space is as follows:

$$d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \tag{15}$$

The Euclidean distance between two points and is calculated as follows:

$$d_E = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (i = 1,2,\cdots,n) \tag{16}$$

### *4.2 Cosine Similarity*

Calculate the Angle between two vectors to reflect whether two vectors are similar. The dimensions of vectors are not limited, and any dimension can be compared. In the user-based collaborative filtering recommendation algorithm, cosine similarity treats the scoring matrix as multiple scoring vectors [8]. Suppose vectors *u* and *v* are rating vectors of two users, and the similarity calculation formula is as follows:

$$sim(u,v) = \cos(\bar{u},\bar{v}) = \frac{\bar{u} \cdot \bar{v}}{|\bar{u}||\bar{v}|} = \frac{\sum_{i \in I_{uv}} R_{ui} R_{vi}}{\sqrt{\sum_{i \in I_{uv}} R_{ui}^2} \sqrt{\sum_{i \in I_{uv}} R_{vi}^2}} \tag{17}$$

### *4.3 Modified Cosine Similarity*

The cosine similarity does not consider the difference between the scoring habits of two users, so if the similarity is calculated directly according to the scoring, the recommendation result may not be accurate. Subtract the mean of all of the user's ratings from the user's ratings, and this is the modified cosine similarity. The similarity calculation formula is as follows:

$$sim(u,v)=\frac{\sum_{i\in I_{uv}}\left(R_{ui}-\overline{R}_u\right)\left(R_{vi}-\overline{R}_v\right)}{\sqrt{\sum_{i\in I_u}\left(R_{ui}-\overline{R}_u\right)^2}\sqrt{\sum_{i\in I_v}\left(R_{vi}-\overline{R}_v\right)^2}} \tag{18}$$

## 5. Conclusions

As one of the representative carriers of domain information, extracting domain terms from a large number of domain text data has become a hot topic in natural language processing. Bilingual term extraction is widely used in the fields of dictionary compilation and data mining. In view of its important role in the field of natural language processing, it is an important problem that needs to be solved in this field to build a bilingual automatic term extraction system by using existing corpus resources and computer resources and designing reasonable algorithms to realize automatic or semi-automatic term extraction. Compared with traditional models, deep learning models can not only reduce the manual feature workload, but also automatically obtain effective semantic and grammatical features from words or sentences. At the same time, the word vector training tool is used in the input layer of the network to represent the words vectorically. At the beginning of the network training, the words contain rich semantics, and the single word vector space is re-integrated to make the word vector space of the source language and the target language have stronger correlation, so as to further improve the extraction performance of the bilingual dictionary.

## Acknowledgements

This work is supported by 2023 annual social science planning fund project of Liaoning province (L23BYY014): Compilation and application of terminological dictionary for Japanese software outsourcing document translation based on corpus.

## References

*[1] Li Y, Li Z.Research on College English Translation Teaching Based on Parallel Corpus[J].International Journal of New Developments in Education, 2023, 5(18): 55-59.*

*[2] L. Sun, Y. B. Jin, L. Du, et al. Automatic Extraction of Bilingual Term Lexicon from Parallel Corpora[J]. Journal of Chinese Information Processing, 2000, 15(06): 34-39.*

*[3] Prakash C, Sumit C. An intelligent chatbot design and implementation model using long short-term memory with recurrent neural networks and attention mechanism[J]. Decision Analytics Journal, 2023, 9(1): 100359-100359.*

*[4] Aschale A A, Demilie M M, Olalekan A S. Towards audio-based identification of Ethio-Semitic languages using recurrent neural network[J]. Scientific reports, 2023, 13(1): 19346-19346.*

*[5] Nilkanth N P, Suresh B K, S. P P. Adaptive membership enhanced fuzzy classifier with modified LSTM for automated rainfall prediction model[J]. Intelligent Decision Technologies, 2023, 17(4): 1031-1060.*

*[6] Yongping Z, Achyut S. Enhancing Supply Chain Transparency and Risk Management Using CNN-LSTM With Transfer Learning[J]. Journal of Organizational and End User Computing (JOEUC), 2023, 35(1): 1-22.*

*[7] KafiKang M, Hendawi A. Drug-Drug Interaction Extraction from Biomedical Text Using Relation BioBERT with BLSTM[J]. Machine Learning and Knowledge Extraction, 2023, 5(2): 669-683.*

*[8] J. Liu, J. Yang, S. S. Song. Collaborative Filtering Recommendation Algorithm Based on Purchasing Intention of Users[J]. Journal of Jilin University(Science Edition), 2021, 59(6): 1432-1438.*