Tibetan Person Attributes Extraction Based on BP Neural Network

Lili Guo[1,2] ,Yuan Sun[1,2*]

[1]School of Information Engineering, Minzu University of China

[2]Minority Languages Branch, National Language Resource and Monitoring Research Center

100081, Beijing, China

guolili0305@163.com

*corresponding author: tracy.yuan.sun@gmail.com

Abstract: At present, Tibetan information is quickly connected with modernization and information, which results the expansive development of Tibetan information on the network. In the face of the massive network information, extracting the information that people want is an urgent problem to be solved. Currently, Chinese person attributes extraction studies have some good results, but there is still much space to Tibetan person attributes extraction. The paper uses person attribute keywords, case-auxiliary word, verbs and other related meaningful words as features to vector, constructs the error BP neural network model and utilizes this model to identification and classification for Tibetan person attributes, and achieved good results. This research has a very important role in the search engine, information security, machine translation and many other applications.

Key words: Tibetan; Person attributes extraction; BP neural network

1 Introduction

With the rapid development of Internet, a large number of electronic text information resources appear in front of people, and more and more online information in the form of multiple languages is released. At the same time, how to get the needed information quickly and accurately becomes a major problem. There is an urgent need for some automated tools to help people quickly find the information that you really need in the mass of information resources. According to incomplete statistics, the number of people uses the Tibetan language as much as 4.22 million. It is mainly distributed in China's Tibet autonomous region, Gansu, Qinghai, Sichuan and Yunnan and other Tibetan areas [1]. Research on Tibetan person attributes extraction can promote communication between nationalities, enhance mutual understanding between nationalities and drive the development of Tibetan economy, science and technology, culture and other fields to better serve the Tibetan people. Therefore, electronic and information processing of the Tibetan text becomes the focus of contemporary social issues.

In the early 1980s, National Institute of the Chinese Academy Social Sciences Zhang Lianshengtried to sort Tibetan vocabulary by computer, opened the precedent of Tibetan text processing. However, due to the form of Tibetan and English and Chinese is very different, so in the computer operating system platform will be difficult to develop. Some Tibetan studies so far has been a great progress. In Tibetan text resources and literature classification, text statistics [2] and entropy value calculation [3], Tibetan speech recognition and word segmentation method [4]other fields have a considerable part of the progress. These are good foreshadowing roles and the accumulation of relevant knowledge to study of Tibetan attributes extraction in this paper. The ultimate goal of information extraction is to turn the useful information in unstructured text into a structured text. The origin of information extraction technology is text understanding. But it is not fully to understand an entire text,

but on the part relevant information of the document for analysis. In Tibetan person attributes extraction is the same understanding.The first Tibetan corpus are obtained from the Tibetan websites, and in this paper, the sentence which contains the Tibetan person attributes is selected from the Tibetan corpus as the preprocessing data. In this paper, we select the person's father, birth place, gender, occupation and so on as the information point that needs to be extracted.

Person attributes extraction is an important part in information processing technology, and it is becoming a more and more hot topic. Tibetan person attributes extraction is still in its infancy, there is still a lot of work to be done. Typical methods in English are based on the feature vector method [5,6] and based on kernel function method [7,8]. There are two methods for the specific application of research in the Chinese language [9,10]. At present, there are many methods used in information processing, among which the machine learning algorithm is used to train and test is the most popular method. Vahideh Sadat Sadeghi,Khashayar Yaghmaie [11] uses a neural network that a combination as the input parameters to extract the edge characteristics of the vowel letters. Joachim Schenk and Gerhard Rigoll [12] used the neural network to feature extraction, and then applied to the standard HMM method, to improve the feature recognition of online handwritten research. Deng Bo et al. [13] introduced the lexical semantic matching technology to extract the Chinese entity relation on the basis of using pattern matching technology in Chinese information processing.Zelenko [14] early used the method to study the relationship between kernel extraction fields. Culotta [15] defined the kernel function based on the dependency tree and used the SVM classifier to extract the relationship by some conversion rules. Zhang et al. [16] designed a kind of compound convolution tree kernel function to carry on the relation extraction.

In Tibetan attributes extraction, there is very little research work. Natural language processing method used in Chinese can be used in Tibetan information processing [17]. However, the actual process of using must consider problems that Tibetan and Chinese compared to have a larger gap in the study of natural language processing. If this key issue has been resolved, Tibetan natural language processing technology will be further developed. At present the biggest difficulty is the lack of Tibetan corpus. According to the characteristics of the Tibetan language, this paper uses person feature keywords, case-marking, verbs and other related meaningful words to carry on the vector as the input to BP neural network [19] and extract results as output, training via BP learning arithmetic [20]. In the experiment, the vector of Tibetan corpus, neural network structure design, learning rate, training algorithm and other related parameters were adjusted to get the best results.

2 Person attributes extraction based on BP neural network

At present, BP neural network is the most representative and widely used model in artificial neural network model, which has the ability of self-learning, self-organization, self-adaptation and strong nonlinear mapping [18]. BP neural network has become a important tool for classification problems such as face recognition, character recognition and signal processing, etc. The entity relation extraction method based on BP neural network is usually transformed into classification problem, and the design of neural network structure is one of the key problems. The Tibetan corpus is obtained by configured crawler system from a Tibetan website, then selected articles about introducing the person introduction and processed these sentences such as artificial segmentation, marking word class. The tagged corpus select the relevant features to vector as input data,and reused to build neural network model identification and classification. The extraction process is illustrated in Figure 1.

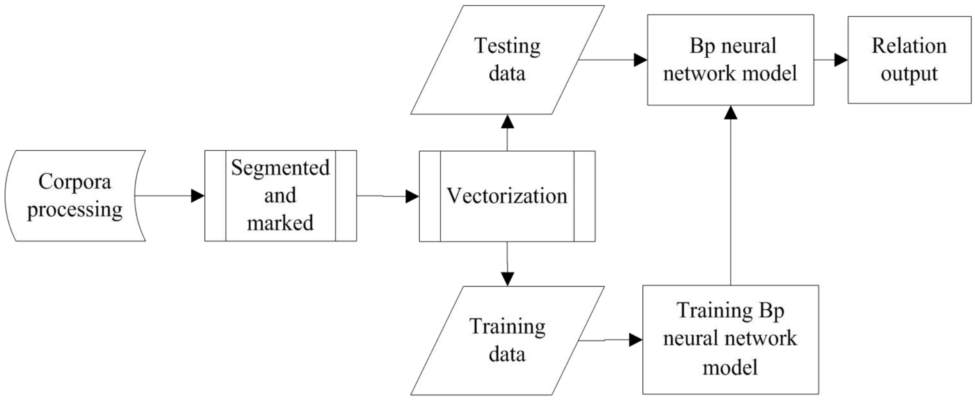

Figure 1 Extraction flow chart

2.1 Tibetan corpus processing

The Tibetan corpus is obtained by configured crawler system from a Tibetan website, such as Wikipedia (Tibetan Edition), Kangba media network, China Tibetan middle school network, etc, then selected articles about introducing the person introduction and processed these sentences such as artificial segmentation, marking word class, as shown in example 1-4. Among them, we choose 9,216 sentences with the person attributes, using 6,000 sentences as the training data, and the 3,216 sentence as the test data. Finally, the labeled corpus to vector is used as the input data to train neural network model.

Example 1:<e1>ཇོ་ནང་ཏཱ་ར་ནཱ་ཐ།/nh</e1>·ནི·/v<e2>ཐོ/n</e2>རབས་ཞིག་ཡིན་/v།/w

Example 2: <e1>གཙང་སྨྱོན་ཧེ་རུ་ཀ་རུས་པའི་རྒྱན་ཅན།/nh</e1>ནི་/v<e2>གྲུང་པོ།/n</e2>·ཡིན་/v།/w

Example 3: <e1>འགྲོ་མགོན་འདི་ཕྱི་/nh</e1>ནི་/v<e2>ལོ་༡༢༣༥པ་ཤིང་མོ་ལུག་/nt</e2>ལ་/kའཁྲུངས/v།/w

Example 4: <e1>སྲོང་བཙན་སྒམ་པོ་/nh</e1>ནི/v·<e2>རྒྱལ་པོ་/n</e2>ཡིན/v·ལ།/w

Vectoring of characteristics mainly uses keyword characteristics, annotation combination characteristics and marked characteristics of words nearby. High frequency keywords appear in a certain attribute category, with a high degree of recognition. There are many different Tibetan and Chinese and English. Tibetan is a post-predicate sentence. Verbs are the core of a sentence in Tibetan. The case-marking of verb nearby is rich semantic role information. To a certain extent, the case-marking reflects the relationship [1] between the predicate and the subject in the sentence, and the appearance of these markers has certain rules. The case-marking is an important effect for Tibetan attributes extraction. So the case-marking and named entity marking in combination may be helpful to the classification effect. The marking characteristics of keywords nearby have a certain effect on improving the classification performance like the role of entity words.

The pos tagging adopts "Modern Tibetan lexical category tag set specification for information processing" of "The national language resource monitoring and research center of minority language center" , where the "/nh" said the name, "/t" said the time, "/ns" said the name, "/n" means the common noun, "/k" said case markers, "/v" for verb.

2.2 BP Neural Network

Each layer of the BP neural network is made up of a lot of nerve cells. Each layer of neurons receives the input from the previous layer, and outputs the calculated results to the next layer by the

transfer function. If the error of the output value of the output layer and the actual output value is less than a predetermined value, the error will back propagation. By adjusting the connection weights and thresholds of each layer, the error between the calculated value and the actual value is gradually reduced until the error reaches the predetermined requirements. This algorithm has become an error back propagation algorithm. The neural network model based on BP algorithm is the BP neural network model.

The artificial neural network is similar to the neural system of the human brain, which is composed of a large number of artificial neural cells. BP neural network model, as shown in Figure 2, includes the input layer, hidden layer and output layer. The paper uses person attribute keywords, case-auxiliary word, verbs and other related meaningful words as features to vector. Through different experiments, the data of input layer selects a 12-dimensional vector : $x=(x_1,x_2,x_3\cdots x_{12})$. $w=(w_1,w_2,w_3\cdots w_{12})$ is the connection strength with the hidden layer connected to it and that means weight. Each of the inputs of the artificial nerve cell is associated with a weight, which will determine the overall positivity of the neural network. The sum of the input values being multiplied by the corresponding weights is the sum of the neurons input. If the sum exceeds a certain threshold θ_i , the cell is activated, the cell output signal $y_i=f(\sum wx-\theta_i)$. The function $f(\sum wx-\theta_i)$ is called activation function or transfer function and connected to the input and output of the nerve cell. It has a variety of types, often using S type logarithmic or tangent function and linear function. In this paper, we use S function as transfer function.

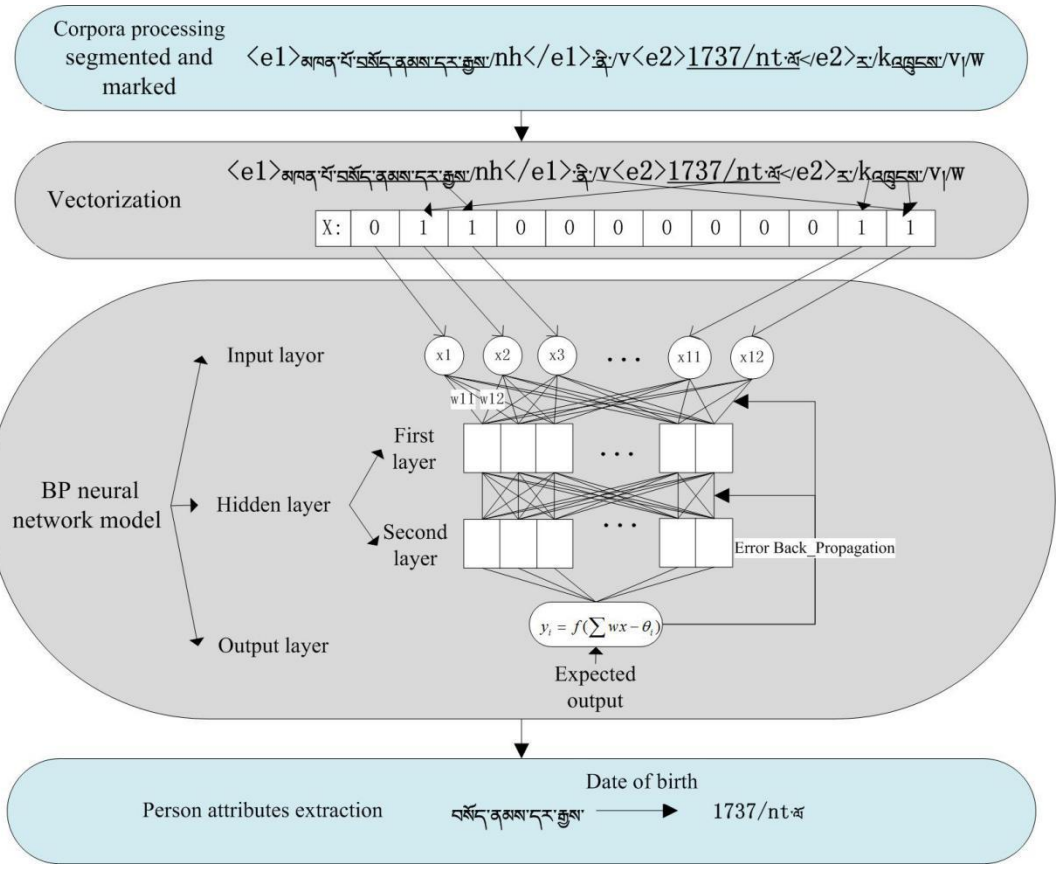

Figure 2 BP neural network model

The hidden layer is composed of any number of neurons. In this paper, we use two layer of hidden layer to distinguish the person attributes. The number of hidden layer nodes is determined according to the experience of the previous design and the experiment. The number of hidden layer nodes is directly related to Requirements for solving problems and the number of the input and output. In addition, too many hidden layer nodes can lead to long learning time, while too few nodes may identify the low ability of the sample without learning. The initial value of hidden layer nodes (L) is determined by one of the following two formulas [21].

$$L = \sqrt{m+n} + a \quad (1)$$

$$L = \sqrt{0.43mn + 0.12n^2 + 2.54m + 0.77n + 0.35 + 0.51} \quad (2)$$

Among them, n and m are the number of input nodes and the number of output nodes, and a is a constant of 0~10. In this paper, the number of input node m is 12, the number of output node n is 1. The number of two hidden layer nodes selected 12 neurons.

In addition to the input layer, hidden layer and output layer neurons need to activate the function. In this paper, the hidden layer uses Sigmoid function, and the output layer utilizes a linear function. Training network has two kinds of modes: one by one mode and the batch mode. In one by one mode, each input is applied to the network, and the weights and thresholds are updated. Batch mode variable does not need to set the training function for each layer's weights and thresholds, but only need to specify a training function for the entire network and is relatively easy to use. Many improved fast training algorithms can only use the batch mode, so this paper uses the batch mode to train the network's function. Training functions have trainlm algorithm, trainrp algorithm, trainbfg algorithm, traingdx algorithm, etc. In this paper, we use the traingdx algorithm which is suitable for simulation classification. Learning speed parameter can't be selected too large, otherwise the algorithm does not converge. Learning speed parameter can't too small to make the training process too long. Generally people choose the value to between 0.01 to 0.1. This paper uses 0.01. The training target error is 0.01.

3 Simulation results and analysis

The corpus of each attribute is divided into 3 parts, with 6,000 sentences as the training data, the 3,126 sentence as the test data. Performance evaluation of extraction can use the evaluation methods of information retrieval. The recall (R) can be roughly seen as a measure for the proportion of correctly extracted information, and the precision (P) is a measure of the correct amount of information extracted. There is an inverse relationship between recall and precision, that is to say, the increase of the precision will lead to the decrease of the recall, and vice versa. Evaluation of a performance should also consider the recall rate and precision, but at the same time comparing the two values will not achieve a clear effect. In this paper, the F value is used to evaluate the performance of the final system. In this way, we can see that the algorithm is good or bad with a numerical value. A value closer to 1 the better the result.

The results of the experiment are shown in Table I. The F value of the father, mother, gender and other attributes is relatively high, and the nationality and occupation and other attributes of the F value is relatively low. The key words such as father, mother and gender are more obvious and have high recognition, and the characteristics of nationality and occupation have some difficulty in the stage of feature vector. So the relative correct rate is lower.

Table I Tibetan attributes extraction results

Attribute class	Precision (%)	Recall (%)	F (%)
Father	91.05	89.32	90.18
Mother	89.26	86.24	87.69
Date of birth	86.28	84.25	85.25
Place of birth	87.94	84.18	86.00
Gender	88.56	86.04	87.28
Death day	82.65	78.91	80.73
Nationality	81.21	78.47	79.82
Occupation	79.89	75.49	77.58

During the experiment, the best experimental results are obtained by adjusting Feature vector, activation function, training algorithm, learning speed and so on to achieve the highest accuracy rate. In this paper, two hidden layers are used to construct the neural network model. Each hidden layer contains 10 neurons. Activation function selects sigmoid function. The algorithm of training neural network uses traingdx algorithm. Learning speed is set to 0.01. The training target error is set to 0.01.

In the experiment, with MATLAB simulation results are shown in Figure 4 Training error reduction diagram. With the increase of the iteration number, the value of the validation performance is reduced. The best validation performance is 0.0697 at epoch 100.

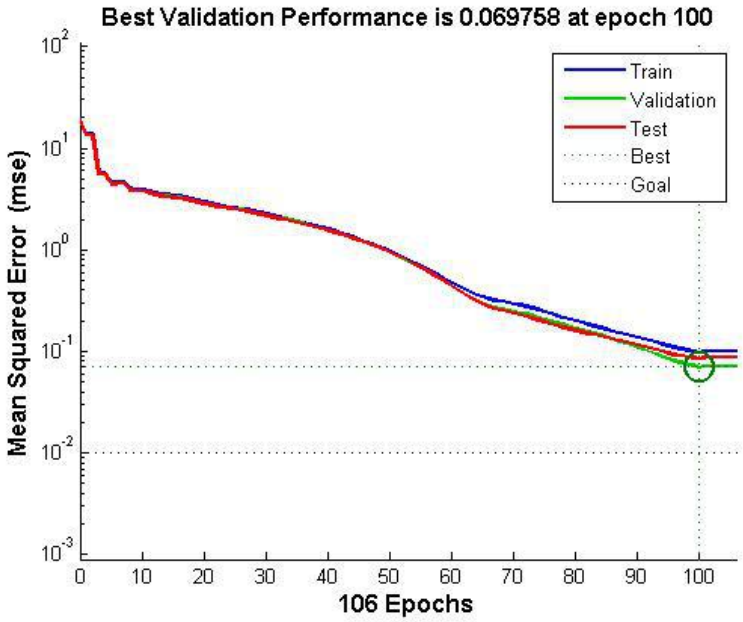

Figure 4 Training error reduction diagram

Figure 5 shows the changes in the gradient, validation checks and learning rate. The values of the gradient, verification checks and learning rate are 1.2863, 6, 0.0415, when the iteration is constant.

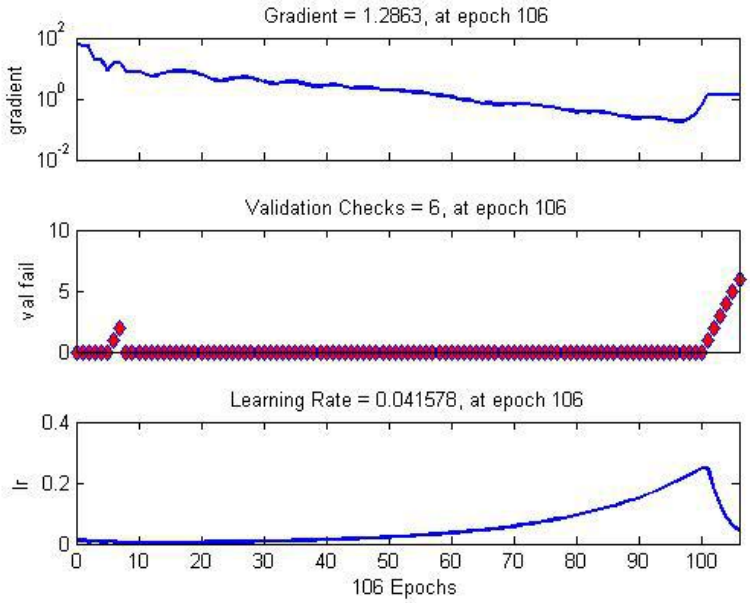

Figure 5 Training status chart

Figure 6 shows the linear regression of training, validation, testing and the three together. Linear regression function is $\text{output} = 0.98 * \text{target} + 0.13$.

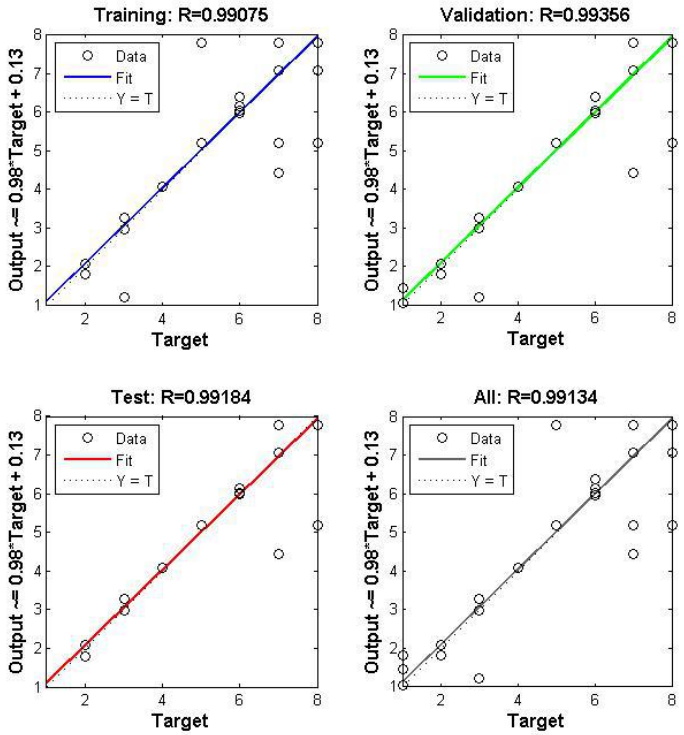

Figure 6 Training regression diagram

Figure 7 is a comparison of forecast and actual classification by BP neural network model. The green represents the label of the prediction category. The red indicates an annotated label before the testing. The one to eight represents the eight attributes person.

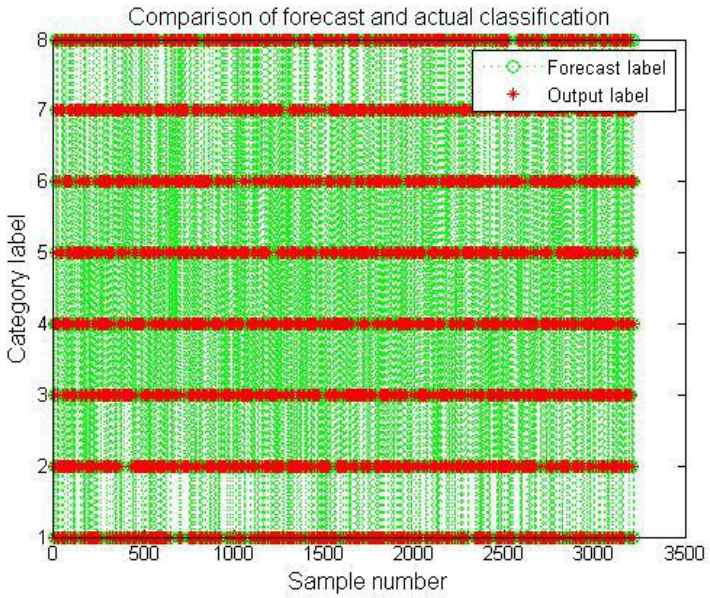

Figure 7 Comparison of forecast and actual classification

Figure 8 is the percentage of the error and the absolute error. A very small part of the prediction error is relatively large.

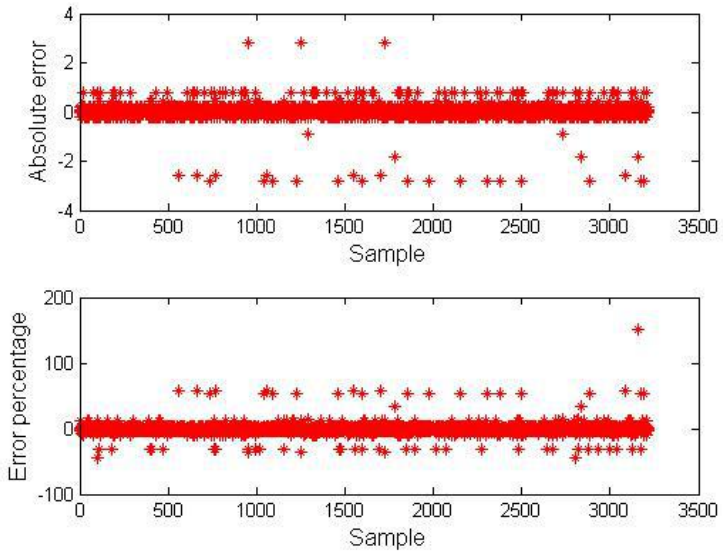

Figure 8 Error diagram

BP neural network to extract Tibetan person attributes as shown in Table II. It provides support for search engine, information security, machine translation and other researches.

Table II Results of Song zanganbu's attributes extraction

Attribute class	Extract attribute value	Sentences that contain attributes
Father	རྗེ་གནམ་རི་སྲོང་བཙན་	སྲོང་བཙན་སྐམ་ངའི་ཡབ་ནི་རྗེ་གནམ་རི་སྲོང་བཙན་ཡིན།
Mother	འཕྲི་བཟའ་ཐོད་དཀར་ཆ་སྦོང་བཟའ་	སྲོང་བཙན་སྐམ་ངའི་ཡུམ་ནི་འཕྲི་བཟའ་ཐོད་དཀར་ཆ་སྦོང་བཟའ་ཡིན།
Date of birth	བོད་ཅེ་སྤྲང་ （ ༥༢༩ ） ལོར་	སྲོང་བཙན་སྐམ་ངོ་ནི་བོད་ཅེ་སྤྲང་ （ ༥༢༩ ） ལོར་འཁྲུངས།
Place of birth	ཡོང་ཁྱེར་ཟླ་སའི་ཤར་ཐོག་སའི་ས་ཆུང་ཐག་ཚན་ཡོད་དབའ་རྒྱ་མཚ་ཀུན་ལ་ཨ་ཡོད་ཅེས་བྱ་བ་ཞིག་ཡོད་པ་ཞེས་	སྲོང་བཙན་སྐམ་ངོ་ནི་ཡོང་ཁྱེར་ཟླ་སའི་ཤར་ཐོག་སའི་ས་ཆུང་ཐག་ཚན་ཡོད་དབའ་བཀུ་ཐག་ཚན་ཡོད་པའི་དབྱ་རྒྱ་མཚ་ཀྱི་ཆུ་མ་ཡོ་བྱང་ཁྱམས་པ་ཞེ་འཁྱུར་ཐྲིང་ནས་འཁྲུངས
Gender	ཡོ	སྲོང་བཙན་སྐམ་ངོ་ནི་བོད་ཀྱི་ཡོ་རབས་ཤེཀ་ཡིན།
Death day	༦༤༠	སྲོང་བཙན་སྐམ་ཡོ་རབ་ཟླུང་གསུམ་པའི་ཚ་སྤྲང་༦༤༠ནམ་ལོ་ཆུ་གཅེཀས།
Nationality	ཀྱུང་གོ	སྲོང་བཙན་སྐམ་ངོ་ནི་ཀྱུང་གོའི་ཡིན།
Occupation	རྒྱལ་པོ།	སྲོང་བཙན་སྐམ་ངོ་ནི་རྒྱལ་པོ་ཡིན་ལ།

4 Summary

This paper introduces a kind of using BP neural network method for Tibetan person attributes extraction. The design and training of BP neural network model is the main part. In the training process, each implementation will produce different neural network models. It chooses the best model for prediction data after training the neural network model. And the experiment has achieved good results.At present, the corpus of person attributes extraction is not rich. The Tibetan data is relatively simple when compared with the test data of the Chinese in the experiment. And the number of labels, the expansion of the corpus content and inspection work is still further improved. The experimental results in a certain extent are obtained. In the study of Tibetan person attributes extraction is still great room for improvement. It provides support for search engine, information security, machine translation and other researches.

Acknowledgements

This work is supported by National Nature Science Foundation (No. 61501529, No. 61331013), National Language Committee Project (No. YB125-139, ZDI125-36), and Minzu University of China Scientific Research Project (No. 2015MDQN11, No. 2015MDYY069).

Reference

[1] Yuzhong Chen, Baoli Li, Shiwen Yu, et al. A Tibetan segmentation scheme based on case auxiliary words and continuous features [J]. Language application. 2003, (01): 75-82.

[2] Jinbao Liang. The vocabulary statistics of Tibetan history literature [D]. Beijing: China Academy of Social Sciences Institute of Ethnology and Anthropology, 2013.

[3] Hongzhi Yu, Yachao Li, Kun Wang, et al. Research of the maximum entropy Tibetan pos tagging of combining syllable features [J]. Journal of Chinese Information Processing. 2013, 27(5): 160-165.

[4] Kunyu Qi. Research on Tibetan word segmentation for information processing [J]. Journal of Northwest University for Nationalities (Philosophy and Social Science Edition), 2006, 26(04): 92-97.

[5] Guodong Zhou, Min Zhang. Extracting relation information from text documents by exploring various types of knowledge [J]. Information Processing and Management, 2007(43): 969-982.

[6] Nanda Kambhatla. Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations [C]. Proceedings of ACL, 2004: 178-181.

[7] Longhua Qian, Gougong Zhou, Fang Kong, et al. Exploiting constituent dependencies for tree kernel-based semantic relation extraction [C]. Proceedings of COLING, 2008: 697-704.

[8] Guodong Zhou, Min Zhang, Donghong Ji, et al. Tree kernel-based relation extraction with context-sensitive structured parse tree information [C]. Proceedings of EMNLP/CONLL, 2007: 728-736.

[9] Wanxiang Che, Jianmin Jiang, Zhong Su, et al. Improved-Edit-Distance kernel for Chinese relation extraction [C]. Proceedings of IJCNLP, 2005:132-137.

[10] Chenglong Zhuang, Longhua Qian, Guodong Zhou. Research on entity semantic relation extraction method based on tree kernel function [J]. Journal of Chinese Information Processing, 2009, 23(1): 3-9.

[11] Andreas Hotho KDE Group University of Kassel etc A Brief Survey of Text Mining 2005

[12] R.K.Aggarwal Mayank Dave Implementing a Speech Recognition System Interface for Indian Languages 2008

[13] Qing Deng, Xiaozhong Fan, Ligong Yang. Method for extracting entity relation with semantic pattern [J]. Computer Engineering, 2007, 33(10): 212-214.

[14] Weiru Zhang, Yue Sun, Xianpei Han. Solid relation extraction method based on Wikipedia and pattern clustering, Journal of Chinese Information Processing. 2012. 26(2): 75-81.

[15] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction [J]. Journal of Machine Learning Research, 2003(2): 1083-1106.

[16] Culotta A, Sorensen J. Dependency tree kernels for relation extraction [C]. Proceedings of ACL, 2004: 423-429.

[17] Zhang M, Zhang J, Su J, et al. A compo site kernel to extract relations between entities with both flat and structured features [C]. Proceedings of ACL, 2006: 825- 832.

[18] Sun Yuan, Zhao Xiaobing. Research on automatic recognition of Tibetan personal names based on multi-features [A]. Proceedings of International Conference on Natural Language Processing and Knowledge Engineering[C], 2010.Chinese Journal of information，2009，23(1): 3-9.

[19] FECIT Technological Product Research Center. Neural network theory and implementation of Matlab 7 [M]. Beijing: Publishing House of Electronics Industry. 2006:100- 105.

[20] Nickolai, S. R.: The layer-wise method and the back-propagation hybrid approach to learning a feed-forward

neural network, IEEE Transactions on Neural network 11(2) (2000), 295-305.

[21] Kaili Zhou, Yaohong Kang. Neural network model and Matlab simulation program design [M]. Beijing: Tsinghua university press, 2005.