Current Status of Tibetan Sentiment Analysis and Cross-Language Analysis

Li Liang[1], Fang Tian[2, *], Benwang Sun[1]

[1]Department of Computer Technology and Applications, Qinghai University, Xining, China

[2]Information Technology Center of Qinghai University, Qinghai University, Xining, China

Keywords: Sentiment analysis; Tibetan text; cross-language sentiment analysis; Tibetan sentiment dictionary

Abstract: Summarizes the status of Tibetan text sentiment analysis, and analyzes that its current problems are mainly the deficiency of basic tools for text processing and the lack of annotated sentimental corpus. This paper proposes to apply cross-language techniques to the construction of Tibetan sentiment dictionary and sentiment analysis. Based on the Chinese emotional dictionary and the Tibetan-Chinese dictionary, cross-language techniques are used to establish the mapping between Chinese and Tibetan. The Chinese sentimental resources were migrated to Tibetan and a Tibetan sentiment dictionary was constructed. Comparison of different sentiment classification methods proves the effectiveness of cross-language technology in Tibetan sentiment analysis.

1. Introduction

Text sentiment analysis is the process of analysis and induction of sentimental comment texts. It is a branch of natural language processing. Tibetan, as a human language, is extremely important in Tibetan daily cultural exchanges and information transmission. The Tibetan-speaking areas include: Tibetans in China, and some in Nepal, Bhutan, India, and Pakistan. Tibetan, refers to the Tibetan language used by Tibetans. The Tibetan language has been widely promoted and used in Tibetan areas, and Tibetan-language teaching schools have spread throughout Tibetan areas. There are more than 20 Tibetan newspapers and periodicals in the country. There are 8 publishing houses publishing Tibetan books, and there are more than 20 modern printing presses for printing Tibetan books and newspapers. The Tibetan language text sentiment analysis and the mining of Tibetan textual hidden information are conducive to understanding Tibetan cultural characteristics and Tibetan sentiment analysis level [1]. At present, research on the sentiment analysis of Chinese and English texts has been relatively mature at home and abroad [2]. The research on the sentimental tendency analysis of Tibetan texts has lagged behind.

After long-term accumulation of corpus resources and research, the research on sentiment analysis in both English and Chinese is relatively mature, and the current research in Tibetan has also achieved good result. This paper summarizes the status of Tibetan sentiment analysis, and proposes a Tibetan-language text sentiment analysis based on cross-language techniques in response to the lack or non-disclosure of the basic tools for Tibetan text processing and sentimental corpus.

2. Tibetan Sentiment Analysis Status

2.1 Tibetan Text Processing Basic Tools

Tibetan textual emotion processing usually requires text processing basic tools such as word segmentation and word vector. After the author has summarized the different phases of Tibetan texts using segmentation and part-of-speech tagging, these segmentation tools or methods are not public.

After more than ten years of development, the research on Tibetan word segmentation has achieved many important results. The earliest research on the Tibetan word segmentation system can be traced back to 1997, and Jiang Di studied regular word segmentation techniques [3]. In 2003, Chen Yuzhong et al. [4] proposed a written Tibetan automatic word segmentation method based on case-auxiliary and continuous features. This method has profound implications in solving unknown words and improving the effect of Tibetan word segmentation. In 2010, Qi Kunyu [5] proposed a lexical-segmented Tibetan word segmentation dictionary mechanism based on the structural characteristics of the Tibetan basic set encoding strings. The dictionary mechanism was relatively complete, but was affected by insufficient training corpus. In 2011, Shi Xiaodong et al. [6] transplanted Segtag, a Chinese word segmentation system based on Hidden Markov Models (HMM), into the Tibetan language and proposed the Yangjin word segmentation system based on Hidden Markov Model (HMM). In 2014, Sun Meng [7] proposed a Tibetan word segmentation method based on a discriminative model, introduced a non-local feature using a reordering algorithm based on word graphs, and used the shortest path algorithm to generate optimal word segmentation results. In 2015, Luosang Gadeng et al. [3] adopted the knowledge fusion conditional random field (CRFs+knowledge) method. The experimental results show that this method is superior to the method using only the CRF model. In 2015, He Xiangzhen et al. [8] adopted a Tibetan word segmentation method based on syllable annotation. It achieved a comparison between the four word segmentation systems under the same conditions, based on conditional random field (CPF), based on maximum entropy (ME), based on maximum interval Markov model (M3N), and based on maximum entropy post-processing (ME+D) , the results show that the CRF-based Tibetan word segmentation system works best. In 2015, Gesang Duoji et al. [9] designed and developed a Tibetan-based automatic word segmentation system based on a lexicon library. The dictionary matching algorithm used was a forward maximum matching algorithm (MM). In 2018, Li Bohan et al. [10] applied deep learning techniques to Tibetan text segmentation to compare the effect of multiple deep neural network model word segmentation, including recurrent neural network (RNN), bidirectional recurrent neural networks (BiRNN), and stacked recurrent neural networks. (StackedRNN), Long short term memory network (LSTM) and Encoder-Label LSTM.

The technical ideas of the above word segmentation system are mainly rule-based, statistics-based, and based on rules and statistics. Rule-based methods use dictionaries or grammar rules to find out the value output that matches a dictionary or rule. This method usually requires a large enough dictionary to use the maximum matching algorithm. The algorithm is simple to implement and has high efficiency, but it does not have strong recognition ability for unregistered words. Statistical-based machine learning methods use corpus to train statistical models, which are then based on model segmentation. The statistical models mainly include CRF, ME, and HMM. The accuracy of ME or CRF is slightly higher than that of HMM, but its training is relatively complex and is affected by the size of the corpus. These statistical models are mainly to solve the problem that the dictionary is not large enough to be updated in time. In addition, the latest research based on deep neural network method has achieved good results in Tibetan word segmentation.

2.2 Tibetan Text Emotional Corpus

Emotional corpus is data used for emotional classification. The size of emotional corpus determines whether the effect of emotion classifier training is excellent or not. Tibetan language lacks marked emotional corpus. Du Xuefeng [11] uses two sources of emotional corpus: one is to crawl through Tibetan news or commentary articles, and the other is to crawl simple sentences with sentimental tendencies in Chinese microblogs and translate them into Tibetan. The final corpus totaled 10,000 affective sentences, of which there were 600 test corpus. Jiang Tao et al. [12] built a Tibetan sentiment corpus by means of a microblogging interface and manual annotation. The corpus used by Pu-Ci Ren et al. [13] is a member of Tibet University. It consists of 44,000 Tibetan phrases selected by Sina Weibo and Tencent Weibo. The sentiment tends to be positive and negative. Li Miaomiao's [14] emotional sentiment-based Tibetan textual sentiment analysis determines the emotional tendency of the entire text by counting the number of positive, neutral, and negative sentences in the text, and then divide these chapters into positive, neutral, and negative sentiments. Yang Zhi[15] grabs data from microblogs, conducts manual classification and screening, and then emotionally annotates corpus.The emotional corpus used by Yan Xiaodong et al. [16] was collected from various Tibetan language forums and Tibetan microblogs. The corpora contains 988 sentences, including 423 positive sentences, 376 negative sentences, and 189 neutral sentences.

According to the above summary, it can be seen that the collection of Tibetan sentiment corpora generally involves crawling data from the Internet, and then building sentimental corpora by word segmentation and manual annotation methods, which consumes a lot of manpower and financial resources.

2.3 Emotion Analysis

Sentiment analysis is the judgment of sentimental tendency with subjective texts. This section collates and summarizes Tibetan sentiment analysis methods. Yuan Bin et al. [17] constructed a semantic feature space based on Tibetan syntactic structures and semantic feature vectors, and proposed a Tibetan sentiment analysis method based on semantic space. Experiments show that the semantic space+TF-IDF emotion classification method is superior to SVM+TF-IDF and naiveBayes+ME. Jiang Tao et al [12] proposed a multi-feature-based sentiment orientation analysis algorithm and tested the support vector machine (SVM) classification model to have better robustness and higher accuracy than the naive Bayesian (NB) classification model. . At the same time, compared with Zhang Jun's [2] classification method based on sentiment dictionary, SVM classification algorithm has greatly improved accuracy and recall rate. TSSRAE was proposed by Pu-Ci Ren et al. [13] and introduced the recursive self-encoding algorithm in the deep learning domain into Tibetan sentiment analysis. Utilizing an unsupervised recursive self-coding algorithm to vectorize the matrix of word vectors, the output layer classifier is trained to predict the emotional tendencies of Tibetan sentences. Experiments show that TSSRAE classification algorithm is superior to semantic space model, feature fusion model and SVM sentiment classification model.

At present, the commonly used Tibetan sentiment analysis methods are mainly based on Tibetan dictionaries and machine learning algorithms. From the above summary, it can be seen that the method based on machine learning is generally superior to the classification method based on the Tibetan dictionary. Based on the classification algorithm of Tibetan dictionaries, a sentiment dictionary marked with polarity is used to make judgments by means of emotional word weighting. Based on the classification algorithm of machine learning, text classification is through training classifiers. The rankings of commonly used classification algorithms from high to low are semi-supervised recursive self-encoding, semantic space, feature fusion, SVM, and NB.

In summary, the current emotional analysis of Tibetan texts is in its infancy, but its starting point is relatively high. Relying on the existing Chinese methods and combining the characteristics of Tibetan itself, it has achieved good results. There are two main problems that lead to the slow development of Tibetan sentiment analysis. The first is that the word segmentation system is not disclosed, and the second is that the annotation of emotional language materials requires a lot of manpower and financial resources. The solution to these two problems has to go through a long and long process, and it sometimes takes two to three years to establish a signed sentimental corpora. To avoid these two issues, this paper proposes the application of cross-language sentiment analysis techniques to Tibetan sentiment analysis where relative emotional resources are scarce.

3. Cross-language sentiment analysis

Cross-language sentiment analysis (CLSA) uses the source language with emotion-distinguishing tags to train data, overcomes differences in vocabulary cross-language distribution, and assists target language data in the classification of emotional tendencies [18]. The source language here is a language with large-scale emotion-marking corpus, and the target language is a language with a small-scale emotional corpus [19]. At present, the cross-language sentiment analysis research generally establishes the mapping between the source language and the target language. Then, the sentimental resources of the source language are migrated to the target language through the emotional resource migration method to learn the target language sentiment analyzer. The existing CLSA technology uses machine translation technology to translate different languages into the same language, and then applies a sentiment analysis method in monolingual. Lu et al. [19] translated the target language training set into the source language using machine translation technology, and then applied the AdaBoost algorithm on the joint training set to obtain a classifier suitable for the target language emotion recognition. Wan et al. [20] translated the marked English text and unlabeled Chinese text, and then used Co-Training algorithm to recognize Chinese emotions. The accuracy of the classification of this method is heavily influenced by the quality of machine translation; there is also a mapping between languages obtained from non-annotated parallel corpus. Meng et al. [21] use unsupervised learning methods to discover unobserved affective words from non-parallel corpus and enhance the recognition coverage of sentiment words to achieve the purpose of enhancing the mapping relationship between languages. Since there are few parallel corpus between the two languages, this method is not practical.

3.1 Construction of Tibetan emotion dictionary based on cross-language

The Tibetan sentiment dictionary used in this paper is based on Chinese sentiment dictionary and Tibetan-Chinese Dictionary[22] to transfer Chinese sentimental resources into Tibetan through machine translation.Chinese sentiment dictionary include Tsinghua University's derogatory meaning dictionary(Contains positive and negative emotions, 5567 derogatory words, 4468 demeaning words), Taiwan University's NTUSD dictionary(Contains positive and negative emotions, 8276 derogatory words, 2810 demeaning words), Hownet dictionary(Contains positive and negative emotions, 4766 derogatory words, 4370 derogatory words, degree words are divided into the most, very, more, slightly, not enough, super, and negative words), and HIT's stopword dictionary(Contains 767 stop words).

The specific construction process of the Tibetan emotional dictionary:First, the Hownet dictionary was merged with Tsinghua University's derogatory meaning dictionary, and then the merger result was merged with Taiwan University's NTUSD dictionary to eliminate duplicates. The final result of the merger uses the Tibetan-Chinese Dictionary to construct Tibetan sentiment dictionary through machine translation.The use of cross-language technology to construct Tibetan stopword dictionary through the HIT's stopword dictionary and the Tibetan-Chinese Dictionary.The final Tibetan emotional dictionary contains basic emotional words, degree words, negative words, transition words, double negative words, and Tibetan stop words.

With the continuous expansion of the dictionary, there are 27,361 total word counts in the Tibetan emotional dictionary, including 220 degree adverbs, basic emotional words(10670 positive emotional words, 10402 negative emotional words, and 5711 neutral emotional words), and 385 stop words.

3.2 An Sentiment Analysis based on Tibetan Emotion Dictionary

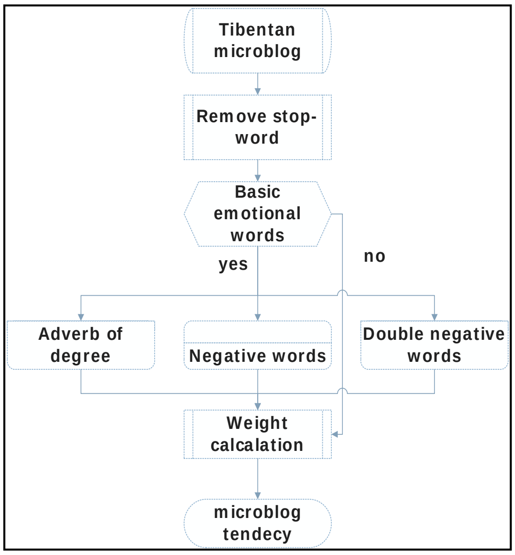

Fig.1 Emotional calculation

The analysis of Tibetan microblog sentiment based on sentiment dictionary is generally to judge the sentimental tendency of a microblog through the accumulation of the weight of emotional words or phrases in the microblog.The transitional words in the Tibetan microblogging can change the sentimental tendency of the whole microblog, so it is also necessary to treat the Tibetan microblogs with transitional words. If the microblog contains a transitional word, the microblog behind the translated word is taken for sentimental calculation. The sentimental calculation steps are shown in Figure 1.

Using 700 Tibetan microblogs to achieve a sentiment analysis based on Tibetan sentiment dictionary, and comparing the results with different deep learning models. The results are shown in Table 1.

TABLE I. Sentiment analysis results

Algorithm	Precision %	Recall %	F-measure %
Emotional Dictionary	68.84	66.11	67.45
CNN	71.30	69.67	70.48
LSTM	72.51	71.36	71.93

From Table I, it can be seen that the dictionary constructed by cross-language technology is also meaningful and basically reaches usability. The F-measure of the two deep learning models were 3.03% and 4.55% higher than those based on sentiment lexicon respectively. The dictionary-based emotion calculation loses certain semantic relationships, while the deep learning model can better preserve the emotional characteristics of the micro-blog sentences. The original semantic links achieve better classification results.

The convolutional neural network can retain the connections between the global words of the microblog, but it can not mine the deep semantic relations. The LSTM can excavate deeper semantic relations. How to combine these two advantages will be the next step.

4. Conclusion

This article summarizes the current situation of Tibetan text sentiment analysis, and analyzes that the existing problems in Tibetan sentiment analysis are mainly the lack of marked emotional corpus and the non-disclosure of the word segmentation system. In order to solve these problems, a cross-language sentiment analysis technique was applied to the Tibetan textual sentiment analysis to shorten the maturity of Tibetan sentiment analysis. Our laboratory has expanded Tibetan sentiment vocabulary through mapping between languages, and based on this dictionary we have performed Tibetan text sentiment analysis. At the same time, this method is compared with the method based on depth model, which verifies the effectiveness of cross-language technology in Tibetan sentiment analysis. The next step is to improve the quality of machine translation.

Acknowledgment

The work is supported by the National Science Foundation of China (Grant No. 61461045) and also supported by the foundational application research of Qinghai Province Science and Technology Fund (No. 2016-ZJ-743). Thank Tibetan students Bading and Xiangqianjia for emotional dictionary and corpus annotation.

References

[1] Cao Hui, Dong Xiaofang, Meng Xianghe. Statistical research on Tibetan newspaper words [J]. Journal of Northwest University for Nationalities: Natural Science, 2012, 33(3): 50-54.

[2] Zhang Jun, Li Yingxing. Sentiment analysis of Tibetan microblog based on sentiment dictionary [J].Silicon Valley,2014,7(20):220+222.

[3] Luosang Gadeng, Yang Yuanyuan, Zhao Xiaobing. CRFs Tibetan Word Segmentation System Based on Knowledge Fusion [J]. Chinese Journal of Information, 2015, 29(06):213-219.

[4] Chen Yuzhong, Li Baoli, Yu Shiwen. Design and Implementation of Tibetan Automatic Word Segmentation System [J]. Chinese Journal of Information, 2003(03):15-20+65.

[5] Qi Kunyu. Research on Tibetan Word Segmentation Mechanism Based on International Standard Coding System [J]. Journal of Northwest University Nationalities (Natural Science), 2010, 31(04):29-32.

[6] Shi Xiaodong, Lu Yajun. The Yangjin Tibetan Word Segmentation System[J].Journal of Chinese Information Processing,2011,25(04):54-56.

[7] Sun Meng, Hua Cancai, Cai Zhijie, Jiang Wenbin, Lu Yajuan, Liu Qun. Tibetan participle based on discriminant classification and reordering techniques [J]. Chinese Journal of Information, 2014, 28(02):61-65+ 90.

[8] He Xiangzhen, Li Yachao, Ma Ning, Yu Hongzhi. Research on Tibetan word segmentation based on syllables annotation [J]. Journal of Computer Applications, 2015, 32(07): 1989-1991.

[9] Gesang Duoji, Qiao Shaojie, He Zedong. Research on Tibetan Word Segmentation System Based on Dictionary [J]. Electronic Technology and Software Engineering, 2015(08):80-82.

[10] Li Bohan, Liu Huidan, Long Congjun, Wu Jian. A Method of Tibetan Word Segmentation Based on Deep Learning [J]. Computer Engineering and Design, 2018, 39(01): 194-198.

[11] Du Xuefeng. An Analysis of the Tendency of Tibetan Sentences [D]. Central University for Nationalities, 2015.

[12] Jiang Tao, Yuan Bin,Yu Hongzhi, Gao Yangji. Analysis on Affective Tendency of Tibetan Microblog Based on Multiple Features[J].Journal of Chinese Information Processing, 2017, 31(03): 163-169.

[13] Pu-Ci Ren, Hou Jialin, Liu Yue, Ji Donghai. Application of Deep Learning Algorithm in Tibetan Sentiment Analysis [J].Computer Science and Exploration, 2017, 11(07):1122-1130.

[14] Li Miaomiao. Study on Sentiment Analysis Method of Tibetan Text [D]. Tibet University, 2017.

[15] Yang Zhi.Sentiment Analysis of Tibetan Microblog Based on Dictionary and Machine Learning [J].Software, 2017, 38(11):46-48+94.

[16] Yan Xiaodong, Huang Tao. Emotional Classification of Tibetan Text Sentences Based on Emotional Dictionaries [J]. Chinese Journal of Information, 2018, 32(02):75-80.

[17] Yuan Bin, Jiang Tao, Yu Hongzhi. A Tibetan microblog sentiment analysis method based on semantic space. Research on Computer Applications, 2016, 33(3); 682-685

[18] Zhang Peng, Wang Suge, Li Deyu. A Method of Cross-Language Text Emotional Tendency Discrimination Based on Strategy Fusion [J]. Chinese Journal of Information, 2016,30(02):32-40.

[19] LU Ling, YANG Wu, CAO Qiong. A Cross-Language Emotional Resource Migration Method Based on AdaBoost[J]. Computer Applications and Software, 2015, 32(11): 77-79+87.

[20] Wang D. Approaches for Transportation Mode Detection on Mobile Devices[C]//Seminar on Topics in Signal Processing,2014: 77-82.

[21] Meng X, Wei F, Liu X,et al. Cross-lingual mixture model for sentiment classification[C]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, 2012: 572-581.

[22] Sun Yizheng. Tibetan-Chinese Dictionary [M]. National Publishing House, 1985.