A method of constructing syllable level Tibetan text classification corpus

Jizhaxi Dao[1,3,4,*], *Zhijie* Cai[1,2,3,4], Rangzhuoma Cai[1,2,3,4], *Maocuo* San[1,3,4], and *Mabao* Ban[1,3,4]

[1]College of Computer Science and Technology, Qinghai Normal University, Qinghai Xining, 810016, China
[2]School of Computer Science and Technology, Southwest Minzu University, Sichuan Chengdu 610041, China
[3]Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province, Qinghai Xining 810008, China
[4]Key Laboratory of Tibetan Information Processing, Ministry of Education, Qinghai Xining 810008, China

Abstract. Corpus serves as an indispensable ingredient for statistical NLP research and real-world applications, therefore corpus construction method has a direct impact on various downstream tasks. This paper proposes a method to construct Tibetan text classification corpus based on a syllable-level processing technique which we refer as TC_TCCNL. Empirical evidence indicates that the algorithm is able to produce a promising performance, which may lay a starting point for research on Tibetan text classification in the future.

1 Introduction

Corpus is a general language material stored in a computer which can be effectively indexed, retrieved, inquired and analyzed. It is an ideal language knowledge resource. Its construction process includes collecting and preprocessing. The scale, coverage and preprocessing workflow of corpus determine the system performance which is modeled on the corpus. In recent years, with the development of advanced technology in machine learning and especially deep learning, the demand for large-scale and high-quality corpus is growing. Corpus preprocessing should not only provide computational insights for the language but also emphasize its application domains.

From the perspective of preprocessing, Tibetan corpus can be divided into syllable level, word level and phrase level. For a specific task, the selection of different levels of corpus has a greater impact on its modeling domain. In the classification of English and Chinese texts, the corpus of word level or phrase level is generally selected. Theoretically speaking, it is more appropriate to select word level or phrase level corpus in Tibetan text classification task, which can directly translate the relevant technologies of English and Chinese text classification to Tibetan text classification. However, the accuracy of Tibetan word segmentation and phrase recognition hardly meets the practical requirements,

* Corresponding author: 1336786645@qq.com

therefore, it is difficult to establish large-scale and high-quality word level or phrase level corpus. There are obvious boundary characters between syllables in Tibetan text, and syllable segmentation is more convenient, so it is relatively easy to establish a syllable level Tibetan corpus. We select 10MB corpus from the natural language processing group of Qinghai Normal University, and observe the influence of syllable level and word level corpus on Tibetan text classification performance. When the corpus size is the same, syllable level and word level text classification performance are basically the same, which indicates that syllable level corpus can be used to solve Tibetan text classification problems. Based on the analysis of the current situation of corpus construction, this paper designs a syllable level Tibetan text classification corpus construction model, and gives the core module text normalization algorithm TC_TCCNL, which lays the foundation for the construction of Tibetan text classification corpus.

2 Background

Corpus is the basic resource of statistical natural language processing. Since the 13th ACL in 1990 took the realization of large-scale real text processing as the strategic goal of computational linguistics, large-scale corpora have been established in many countries.

Tibetan corpus preprocessing generally includes removing non-Tibetan characters, filtering out stop words, syllable segmentation, word segmentation and Part-of-speech tagging. In 2003, Chen Yuzhong[2] and others designed a Tibetan word segmentation algorithm (BCCF) based on case auxiliary words and continuity features by using dictionary method, and Jiang Di [3] proposed Tibetan word segmentation method. From 2009 to 2019, Cai Zhijie et al. [4-8] studied the Tibetan word segmentation technology based on the dictionary, and made a more comprehensive research on the design of dictionary database, block technology, query algorithm and contraction word recognition. In 2009, sun yuan et al. [9] also studied Tibetan word segmentation technologies such as bidirectional maximum matching method to detect intersection ambiguity and word frequency information disambiguation based on case auxiliary word block method. In 2011, Shi Xiaodong and Lu Yajun [10] transplanted the Chinese word segmentation system SegTag based on HMM into Tibetan word segmentation, and designed and implemented the Yangjin Tibetan word segmentation system; in order to make the Tibetan language corpus standardized, unified and practical, Cai rangjia [11] and others put forward the Tibetan word category and part of speech marker set, and established a segmentation tagging dictionary. From 2015 to 2017, Li Yachao et al. [12-14] realized Tibetan word segmentation system based on syllable Tagging Based on conditional random field model. Only in 2020 did Cairang Zhuoma et al. [15] propose Tibetan word segmentation strategy and algorithm based on Part-of-Speech constraints, which can better solve the problems of ambiguity and unknown words.

Although scholars have studied the Tibetan word segmentation from various aspects, due to the complexity of the language, the highest accuracy rate P, recall rate R and F of the open evaluation of Tibetan word segmentation are 93.14, 92.17 and 92.66 (MLWS2017) [16], and there is still a certain gap in meeting the actual needs.

3 Methods

3.1 TC_TCC corpus collection

With the rapid development of the Internet, Tibetan text has changed from massive paper text content to web text content. With the increasing number of Tibetan web pages, the way

scholars collect corpus has shifted from traditional paper-based content manual input and machine scanning to the most popular web crawler technology. According to the form of information resources and the content of the text, it is a key work to choose an effective way to collect corpus.

In this paper, the strategy of crawler based and manual input supplemented is used to collect Tibetan text corpus of news, novel, scripture and medicine. The main sources of the corpus are China Tibet network, China Tibetan Netcom and Qiongmai literature network. Among them, the collection of news and fiction corpus is completed by crawler, while the acquisition of scripture and medical corpus adopts the method of manual input (due to the lack of text corpus of scripture and medical web page). A total of 153.1MB Tibetan text corpus is collected in this paper, which contains a total of 13760350 Tibetan syllables. The source, size and number of syllables are shown in Table 1.

Table 1. TC_ TCCsource distribution of corpus.

Serial number	source	data size(MB)	Number of syllables
1	Qiongmai Literature Network(http: //www.tibetcm.com/)	15.89	1350133
2	China Tibetan Netcom(http: //www.tibet3.com/)	71.75	5731600
3	China Tibet Online(http://tb.tibet.cn/)	1.12	90196
4	Manual entry	64.34	6588416
total		153.1	13760350

3.2 TC_TCC build model

With the rapid development of computer technology, deep learning has become the mainstream technology of natural language processing. However, deep learning requires the scale and quality of corpus, so it is very important to construct large-scale and high-quality corpus.

Text classification is one of the most classic scenarios in the field of natural language processing. It uses computers to automatically classify text according to a certain classification system or standard. Since Maron published his first paper on automatic text classification in Journal of ASM in 1960, scholars have studied text classification technology from different perspectives and put forward many classical mathematical models for text classification. Text classification has experienced knowledge-based method, traditional machine learning method and the most popular deep learning method. These approaches establish classifiers according to the knowledge formation rules provided by experts or train the formation rules on the pre-classified corpus, so as to automatically classify the samples of unknown categories. Knowledge based text classification method needs to design rules manually, so it is difficult to construct classifier. The text classification method based on machine learning can automatically obtain features from the pre classified text, and it is convenient to construct a classifier, which has become the mainstream method of text classification. When using machine learning based text classification method, the granularity setting and preprocessing of corpus features are extremely important. According to the analysis in the introduction section in this paper, it is more appropriate to construct a syllable level Tibetan text classification corpus under the current technical conditions. The syllable level Tibetan text classification corpus is called the syllable level Tibetan text classification corpus (Tibetan Characters Text Classification Corpus, TC_TCC) .

Build TC_ TCC time, First, we should collect and acquire Tibetan students' corpus, and then preprocess the raw materials. TC_ TCC construction model is shown in Figure 1.

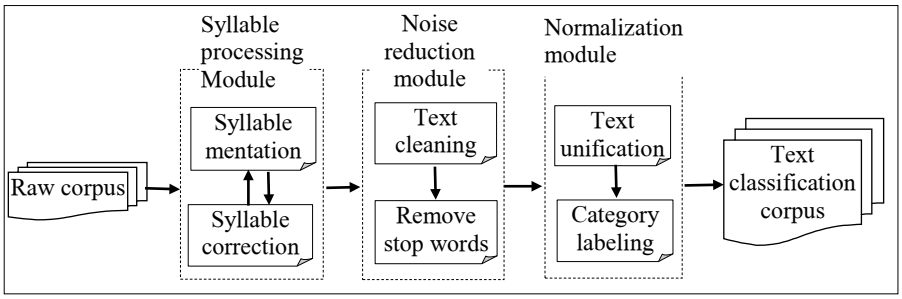

Fig. 1. TC_ TCC construction model.

TC_ TCC construction model consists of three modules, namely, syllable processing, noise reduction and normalization annotation. The syllable processing module includes two parts: syllable segmentation and syllable correction. The main task of syllable segmentation is to use the Tibetan syllable separator "་" as the cut-off point to segment the syllabic of corpus, such as sentence "ང་ནི་སློབ་ཕེན"(I am a student), according to the syllable separator"་"cut into"ང/ནི/སློབ/ཕེན/"; the main task of syllable correction It is to correct the errors caused by the missing syllable separator "་", such as "བཀྲཤིས" (Zhaxi) to "བཀྲ་ཤིས". The noise reduction module includes two parts: text cleaning and removing stop words. The main function of text cleaning is to replace single vertical character, double vertical character and four vertical character at the end of Tibetan sentence with single vertical character; the other is to replace non Tibetan characters such as numbers, English, Chinese characters and punctuation marks with "N" to ensure that the text structure will not be affected; the main function of text cleaning is to replace single vertical character, double vertical character and four vertical character at the end of Tibetan sentence according to the stop list, the Tibetan function words are eliminated. The normalization annotation module consists of two parts: text normalization and category annotation. The main function of text normalization is to normalize the text size to ensure that each text size is consistent; the main function of category annotation is to label the text category.

Using TC_TCC to classify Tibetan text has the following advantages:

1）The vector dimension can be reduced. The total number of Tibetan syllables is far less than the total number of Tibetan words. When training the language model and other downstream tasks, the vector dimension can be reduced and the calculation amount of the model can be reduced.

2）High quality training corpus can be obtained. There is an obvious syllable separator "་" between Tibetan syllables, which can be used as a segmentation point to obtain high-quality training corpus and improve the performance of text classification.

3.3 TC_TCC normalization method

Because there are obvious separators between syllables in Tibetan text, syllable segmentation is relatively simple. Syllable correction mainly corrects the non-true word errors in the text, which can be corrected according to the method in the literature [17]. In the process of text noise reduction, stop words and non-Tibetan character table can be directly replaced with "N", and single vertical character, double vertical character and four vertical character can also be replaced with single vertical character. Tibetan language corpus is mainly obtained by crawler and manual input, so the length of the corpus is inconsistent. In the text classification model training based on machine learning, text normalization is an indispensable basic work, which will directly affect the results of model learning. There are a series of problems such as too much calculation and loss of information when training the language model for long texts. The ideal model can be

obtained by training the corpus with the same size. Therefore, the core of text classification corpus is to normalize the text. Text category annotation is carried out at the same time of normalization.

The goal of TC_TCC text normalization is to keep the text size consistent. Text normalization can use text size (the number of bytes of text) or the number of syllables contained in the text as the standard for cutting. It is difficult to ensure that the cutting point is the end of the sentence (single vertical character) or adjust the position of the cutting point. In this case, the text content will be incomplete. Taking the number of syllables in the text as the standard cutting can not guarantee the cutting The cutting point is a sentence ending character, but it is easy to find the sentence ending character nearest to the tangent point. The sentence ending character found can be used as the cutting point, which can not only maintain the consistency of the text size, but also ensure the integrity of the text content. The size of text is usually described by the number of bytes, and it is more appropriate to use the size of text when the number of syllables is used as the standard for cutting text. In order to reveal the relationship between the size of the text and the number of Tibetan syllables contained in the text, we examine the relationship between the size of the text and the number of Tibetan syllables contained in the text. According to the statistics of 30512.8KB corpus, it is found that there is a relationship between the size of Tibetan text and the number of syllables it contains. There are about 100 Tibetan syllables in 1KB text. The relationship between the size of the text and the number of syllables is shown in Table 2.

Table 2. Relationship between Tibetan text size and syllable number.

Text name	Size(KB)	Number of syllables	Syllable/1KB
Sakya motto	150.6	15213	101.02
Kanjur	2750	280130	101.87
Script	326.6	33022	101.11
News	3700	366961	99.18
Literary review	5700	565983	99.30
Academic	5400	537383	99.52
Fiction	12200	1231509	100.94
Common sense	115.5	11682	101.14
History	170.1	16910	99.41
Average value	3390.31	339865.89	100.25

From the above analysis, TC_TCC normalization algorithm(TC_TCCNL, Tibetan characters Text classification corpus Normalization) can be obtained. The basic idea is calculating the position of the cutting point according to the given file size λ. If the cutting point is at the end of the sentence, the text will be segmented by the cutting point, otherwise, the closest sentence ending character to the cutting point will be found to segment the text.

We used the TC_TCCNL algorithm to perform a normalization experiment on a 35.86MB Tibetan text classification corpus containing 3585995 syllables for scriptures, medicine, news, and novels, and achieved the expected results. The text normalized data is shown in Table 3.

Table 3. TC_TCCNL experimental data table(λ=6).

Serial number	Text type	Text size(MB)	Number of syllables contained	Number of split text
1	Scripture	12.00	1,200,000	2000
2	Medicine	6.31	630,848	1052
3	News	9.14	914,121	1524
4	Fiction	8.41	841,026	1402
total		35.86	3,585,995	5978

As can be seen from the table above, the scripture text size is 12.00MB and contains 1,200,000 Tibetan syllables. When the normalization parameter λ=6, a total of 2,000 segmented texts are obtained, and the text size is between 5.7-6.2KB; The class text size is 6.31MB and contains 630848 Tibetan syllables. When the normalization parameter λ=6, a total of 1052 segmented texts are obtained, of which the size of the first 1051 texts is between 5.8-6.3KB, and the 1052th text The size of the news text is 4KB; the size of the news text is 9.14MB, containing 914,121 Tibetan syllables. When the normalization parameter λ=6, a total of 1524 segmented texts are obtained, of which the size of the first 1523 texts is between 5.8-6.2KB The size of the 1524th text is 2KB; the size of the novel text is 8.41MB, containing 841,026 Tibetan syllables, when the normalization parameter λ=6, a total of 1402 segmented texts are obtained, of which the size of the first 1401 texts Between 5.9-6.3KB, the size of the 1402th text is 4KB. Experimental data shows that the TC_TCCNL algorithm can normalize a given text to a specified size. When the size of the source text is not an integer multiple of the parameter λ, the size of the last text is inconsistent with the size of the previous text.

4. Conclusion

Corpus serves as the most important ingredient of statistics and machine learning, and the quality and distribution of corpus have a great impact on the further research. From the perspective of constructing corpus, Tibetan corpus can be divided into syllable level, word level and phrase level. Due to the restriction of word segmentation and phrase technology, Tibetan word level and phrase level corpus construction technology can not meet the specifications in real-world applications. Based on the analysis of the current situation of Tibetan corpus construction, this paper studies the construction method of Tibetan text classification corpus, and proposes a syllable level Tibetan text classification corpus construction method, including the syllable level Tibetan text classification corpus construction model and the core algorithm TC_TCCNL. Experimental data show that the algorithm achieves the expected effect, which lays the foundation for the construction of Tibetan text classification corpus. In the future, the Tibetan text classification technology based on syllable level could be studied on the basis of this achievement.

The National Natural Science Foundation of China (618666032,61966031), Projects funded by the Department of science and technology of Qinghai Province (2019-SF-129), innovation team funding project of "Yangtze River scholars and innovation team development plan" (IRT1068), Qinghai Provincial Key Laboratory Project (2013-Z-Y17, 2014-Z-Y32, 2015-Z-Y03), Key Laboratory of Tibetan information processing and machine translation (2013-Y-17).

References

1. Ouzhu, zhaxiga. Tibetan computational linguistics [M]. Southwest Jiaotong University Press, 2013.
2. Chen Yuzhong, Li Baoli, Yu Shiwen. Design and implementation of Tibetan automatic word segmentation system [J]. Chinese Journal of information technology, 2003,17 (3): 15-20.
3. Jiang Di. Modern Tibetan chunk segmentation method and process [J]. National language, 2003 (4): 31-39.
4. Cai Zhijie. Recognition of contraction words in Tibetan automatic segmentation system [J]. Chinese Journal of information technology, 2009,23 (1): 35-37.

5. Cai Zhijie. Design and implementation of Banzhida Tibetan automatic word segmentation system [J]. Journal of Qinghai Normal University for nationalities, 2010,21 (02): 75-77.
6. Cai Zhijie, Cai rang Zhuoma. Design of Tibetan automatic word segmentation system [J]. Computer engineering and science, 2011,33 (5) 151-154.
7. Cai Zhijie, Cai rang Zhuoma. Design of Tibetan tagging dictionary database [J]. Chinese Journal of information technology, 2010.24 (5): 46-49.
8. Lama Zhaxi, Cai Zhijie, Zhaxiji. Tibetan contraction lattice recognition method [J]. Computer application research, 2019, 36 (4): 1080-1083.
9. Sun yuan, Luosang Qiangba, Yang Rui, et al. Design of Tibetan automatic word segmentation scheme [C]. Research and progress of Chinese minority language information processing. Nationalities Press, 2009, 228-237.
10. Shi Xiaodong, Lu Yajun. Yangjin Tibetan word segmentation system [J]. Chinese Journal of information technology, 2011,25 (04): 54-56.
11. Cai rangjia. Research on processing methods of Tibetan corpus [J]. Computer engineering and application, 2011 (06): 142-143 + 150.
12. Li Yachao, Jia Yangji, Zong Chengqing, et al. Research and implementation of Tibetan automatic word segmentation method based on conditional random field [J]. Chinese Journal of information technology, 2013,27 (4): 51-58.
13. Li Yachao, Jiang Jing, Jia Yangji, Yu hongzhi.tip-las: an open source Tibetan part of speech tagging system [J]. Chinese Journal of information technology, 2015,29 (06): 204-207.
14. Li Yachao, Jia Yangji, Jiang Jing, et al. Tibetan word segmentation method based on unsupervised features [J]. Chinese Journal of information technology, 2017,31 (02): 72-75.
15. Cai rang Zhuoma, Cai Zhijie. Tibetan word segmentation strategy and algorithm based on part of speech constraints. Chinese Journal of information technology, 2020, 34 (2): 33-37.
16. Cai Zhijie. Research on key technology of Tibetan word vector representation [D]. Qinghai Normal University, 2018.
17. Cai Zhijie, sun Maosong, cairang Zhuoma. A vector model based spelling checking method for Tibetan characters [J]. Chinese Journal of information technology, 2018,32 (9): 47-55.