Research on Tibetan Language Resource Construction Based on Tibetan Natural Language Processing

Hao Chen¹, Yonghao Liu², Hong Li¹, Suonan Jiancuo¹,*

¹School of Information Science and Technology, Tibet University, China, 850000
²Tibet University School of Science, China, 850000

*Corresponding author e-mail: xcz199612@163.com

Abstract. With the rapid growth of science and technology, the research and construction of Tibetan has also accessed a period of rapid growth. Both the expansion of the surface or the depth of the point have reached a considerable level, the theory and means of research have also advanced from the previous, the number of research team, the number and quality of published works greatly improved, the areas involved, the problems solved are unprecedented. At the same time, due to the continuous iteration and updating of the research technology of language and text information processing in the field of Chinese and English, the Tibetan information processing technology has gradually expanded from the processing of text information to the processing of language voice information. However, this is still far from enough, the processing of natural Tibetan language is still not large-scale development. Therefore, this paper uses computer technology to analyze the urgency and importance of the construction of Tibetan language resources from the perspective of natural language processing of Tibetan, and there are few typical examples in the entire research process.

Keywords: Information Technology, Tibetan Language Construction, Language Resources, Natural Language

1. The current situation of Tibetan language resources

The current Tibetan language has a history of more than 1300 years. In this historical process, due to the continuous improvement and standardization of Tibetan language in different historical periods, it can keep pace with social development and better meet the needs of social progress and development in Tibet. In particular, it played a great role in absorbing and learning from the cultural achievements of other nationalities and promoting the development of social civilization in Tibet. In order to strengthen the organization and leadership of the standardization of Tibetan neologisms, in 2006, with the approval of the autonomous region government, the Tibetan Translation Standardization Committee of Tibetan neologisms was formally established. The working rules for the Examination and Approval of Tibetan Neologisms and the Rules for the Translation of Neologisms and the Use of Loanwords in Tibet Autonomous Region were formulated and promulgated. Nowadays, it has become an urgent task to formulate Tibetan standard language. With economic development and living
standards, the pursuit of knowledge of modern cultural aspirations growing, the requirements for means of communication of information have become more sophisticated. The standardization of language and characters is the premise of realizing informationization[1]. Without the standardization standard of language and characters, it is impossible to realize informationization in the true sense. There are still differences among the three dialects in Tibetan, which seriously restrict the development of standardization and informationization of Tibetan, and thus affect the exchange and development of politics, economy and culture. Therefore, it is imperative to formulate Tibetan standard language.

2. The relevant content of natural language processing technology

Natural language processing is an important direction in the field of computer science and artificial intelligence. It studies various theories and methods that can realize effective communication between people and computers in natural language. Natural language processing is a subject that integrates linguistics, computer science and mathematics. The research on natural language management is divided into two parts: basic research and applicability research, both of which are carried out from pronunciation and text. Basic research mainly focuses on linguistics, mathematics, computer science and other fields, such as disambiguation, grammar formalization, theoretical basis of computational linguistics, and language resource library. Applied research mainly focuses on some fields that need to apply natural language processing technology, such as information retrieval, text classification, automatic summarization, machine translation and so on. All technologies and applications of natural language processing are inseparable from language resources, which is of decisive significance to the development and competition of natural language information processing. In a word, with the popularization of Internet and the emergence of massive information, natural language processing is playing an increasingly important role in people's daily life[2]. (see Figure 1)

Figure 1. Examples of Tibetan patterns

2.1. Recognition and analysis of voice text

People have always dreamed of being able to communicate with machines by voice so that they can understand the meaning of people's words. Chinese school-enterprise networking alliance was the speech recognition than as a "machine of the auditory system." Voice recognition technology is an advanced technology that permits machines to recognize and understand people's voice signals and convert them into consistent texts. Voice recognition technology chiefly includes three aspects: characteristic abstraction technology, mode matching criteria and pattern training technology. The voice recognition technology of the Internet of Vehicles has also been fully cited. For example, in the Yi Truck Internet, you can set the destination and direct navigation by just pressing the one-click to talk to the customer service staff, which is secure and suitable. In speech recognition research and development process, relevant researchers designed and produced speech databases in every language according to the pronunciation characteristics of different languages, which can provide adequate and scientific training for Chinese continual speech distinguish procedure research, method design and industrialization work for relevant scientific research institutes and universities at home and abroad[3].

At the same time, after speech recognition, the computer will use grammar to analyse the sentence, one is to identify whether a sentence conforms to the grammar, that is, to complete the sentence recognizer, on the other hand is the internal structure of the sentence, to determine the grammatical component of the sentence, that is, to complete the syntax analysis. Syntax analysis is based on a given grammatical system, automatically derives the grammatical structure of sentences, analyses the relationship between the grammatical units contained in sentences and these discourse units, and
transforms sentences into a structured grammar tree. Because of the characteristics of Tibetan itself, it has many theoretical and technical problems, so far has not yet developed a system as widely accepted as other language systems, and has not formed the information technology Tibetan word-splitting norms, information technology Tibetan word labelling and other national basic standards. Because the researchers of various units do not have a unified premise and foundation, lack of evaluation criteria, it is difficult to accurately compare and evaluate the results of various word-splitting algorithms and words, and have a negative impact on resource sharing[4]. (see Figure 2)

Figure 2. Example of machine translated Tibetan keyboard

2.2. Machine translation
In the early stage of machine translation research, direct translation or translation method based on intermediate language is generally adopted. Since the end of 1980s, the extensive application of corpus technology and statistical machine learning method in machine translation research has broken the long-standing deadlock in the unification of analytical methods. Machine translation research has entered a new era, and a number of machine translation methods based on corpus have come out one after another and developed rapidly. A series of important progress has been made in statistical translation methods based on large-scale corpus, which fundamentally changes people's initial understanding and viewpoint of statistical translation methods. Practice has proved that compared with traditional translation systems which have been optimized for decades, statistical translation methods can produce competitive translation results. Therefore, it is inconceivable that the machine translation technology will advance without the support of a large number of bilingual corpora[5].

3. The problem of the construction of Tibetan language resources.
(1) It is urgent to formulate Tibetan standard language. With economic development and living standards, the pursuit of knowledge of modern cultural aspirations growing, the requirements for means of communication of information have become more sophisticated. The standardization of language and characters is the premise of realizing informationization. Without the standardization standard of language and characters, it is impossible to realize informationization in the true sense. There are still differences among the three dialects in Tibetan, which seriously restrict the development of standardization and informationization of Tibetan, and thus affect the exchange and development of politics, economy and culture. Therefore, it is imperative to formulate Tibetan standard language.

(2) It is necessary to unify the terminology of Tibetan neologisms examined and approved by different provinces. At present, all provinces and autonomous regions have their own standardization institutions for Tibetan neologisms, which examine and approve a large number of neologisms every year. However, the lack of communication platforms and channels between provinces and autonomous regions leads to the disunity of some neologisms and needs to be standardized.

(3) It is necessary to strengthen the examination and approval of translation of professional dictionaries and promote the standardization of Tibetan terms in various professional fields. In recent years, we have examined and approved Tibetan words in the fields of law, medicine, physics,
chemistry, mathematics and computer, which have played a very good role. As there are many disciplines and professional terms in many fields, we have not examined and approved them.

4. The methods of building Tibetan language resources
In view of the need sing-up of Tibetan language resources, I put forward the following suggestions:

(1) According to the definition, principles and assumptions drawn up in the Tibetan standard language program, we should heed the development of Tibetan standard language. It is suggested that the formulation of Tibetan standard language should be included in the national language construction planning project, and the National Language Commission should take the lead in setting up a research and development working group to carry out the research and development work in a planned and deployed way. Establish the phonetic system of Tibetan standard language, standardize the vocabulary of Tibetan standard language and determine the grammar of Tibetan standard language. Or carry out research in the form of a single subject.

(2) Strengthen the standardization of Tibetan neologisms. At present, there are endless neologisms. However, many of these neologisms are used inconsistently in Tibetan and Tibetan. How to standardize and unify these neologisms in Tibetan and Tibetan is an important issue to improve the scientific and cultural level of the Tibetan people and share information resources. First, it is necessary to strengthen the functions of the Tibetan Language Committee of the National Terminology Standards Committee. The National Terminology Standards Committee should strengthen the guidance and coordination of minority sub-committees in standardizing and unifying the terminology of new words in Chinese, especially in the standardization and unification of Tibetan new words. Secondly, it is suggested that the State Language Committee should take the lead in establishing the coordination mechanism of Tibetan language norms and standards authority in Tibetan areas of China. Help to formulate and organize the implementation of the long-term plan for the standardization of Tibetan language in China.

(3) Strengthen the basic and applied research of Tibetan language. By integrating resources, cooperating in tackling key problems, and striving for support through multiple channels, we should do a good job in publishing the Dictionary of Chinese-Tibetan Contrastive Neologisms, formulating the Principles and Methods of Tibetan Abbreviations, Rules of Tibetan Abbreviations, and Modern Tibetan Punctuations.

(4) Strengthen the training of Tibetan language standardization professionals. At present, most of the staff engaged in the standardization of Tibetan language in our district are translators and other professionals, who have not received systematic study and training in standardization, and have encountered many problems and difficulties in practical work. We suggest that the relevant state departments should organize and carry out the specialized training of standardized talents, improve the professional quality of staff, and further promote the standardization of Tibetan language[6].

5. Specific examples of construction methods.
The following is to take the Department of Information on language resources and information of Gangjie Tibetan language developed by the University of Tibet as an example, and introduce the method of the construction of Tibetan language resources. Gangjie Tibetan language resource information system using B/S (browser/server) software architecture mode users through the network access to the web server (i.e., Gangjie Tibetan language resource information system), system access database server, database server, resource information and related documents stored separately, documents encrypted in the form of text files, the two are linked by resource classification numbers.(see Figure 3)
Figure 3. B/S architecture mode of Gangjie Tibetan language resource information system

6. Conclusion
Throughout the Tibetan language information technology twenty years of development, many researchers have made unremitting efforts and beneficial explorations in the research of Tibetan information processing and the formulation of related standards, and have made many achievements, which is beyond doubt. However, we should be soberly aware that if we take the overall development level and research status of language information processing technology at home and abroad as a benchmark to measure the development level and research status of Tibetan information processing technology, it seems that the gap is far from the same, and it is self-evident to seize the time to catch up. However, we can't narrow the gap by comparing in general terms. We should calm down and make a survey to find out where the gap lies. The next question is how to grasp the essence or the overall situation as soon as possible, and how to take practical measures to solve it in a limited time, so as to shorten the gap and even catch up with the world development trend of language information processing. This article is trying to put forward my thoughts and opinions by answering these questions.

Acknowledgments
This work was financially supported by Tibet Autonomous Region Department of Education, Tibetan Finance Index [2018] No. 81, National Computer and Tibetan Information Technology National Team and Key Laboratory Construction Project.

References
[1] Lubei Zhang, Linda T. H. Tsung. Tibetan bilingual education in Qinghai: government policy vs family language practice[J]. International Journal of Bilingual Education and Bilingualism, 2019, 22(3), 290-302.
[2] George van Driem. Evidential Systems of Tibetan Languages[J]. Bulletin of the School of Oriental and African Studies, 2019, 82(1), 190-192. Bouvrie J. Notes on Convolutional Neural Networks[J]. NeuralNets, 2006.
[3] Julia Escher, Yves Trachsel. Shimunek, Andrew: Languages of Ancient Southern Mongolia and North China – A Historical-Comparative Study of the Serbi or Xianbei Branch of the Serbi-Mongolic Language Family, with an Analysis of Northeastern Frontier Chinese and Old Tibetan Phonology[J]. Asiatische Studien - Études Asiatiques, 2018, 72(4), 1231-1234.
[4] Miaoyan Yang. Discourses on ‘authenticity’: language ideology, ethnic boundaries, and Tibetan identity on a multi-ethnic campus[J]. Journal of Multilingual and Multicultural Development, 2018, 39(10), 925-940.
[5] Hiroyuki Suzuki, Sonam Wangmo. Language evolution and vitality of Lhagang Tibetan: a Tibetic language as a minority in Minyag Rabgang[J]. International Journal of the Sociology of Language, 2017, 2017(245), 63-90.
[6] Manuel Widmer. Review of Lauren Gawne & Nathan W. Hill. (eds). 2016. Evidential systems of Tibetan languages[J]. Linguistics of the Tibeto-Burman Area, 2017, 40(2), 285-303.