ISSN 2616-7433 Vol. 1, Issue 8: 117-124, DOI: 10.25236/FSST.2019.010819

The Research on Method and Difficulty Analysis of Automatic Word Segmentation in Yi Language for Information Processing

Wang ChengPing

Southwest Minzu University, Chengdu Sichuan 610041, China

ABSTRACT. *Yi word automatic segmentation is a basic Yi language information processin research. The article first analyzes the characteristics of Yi language. Second, the concept of word of Yi and application, and a variety of Yi segmentation algorithm has done a systematic introduction. Then combined with the characteristics of Yi automatic segmentation to achieve the difficulties faced where the word Yi pointed out the future direction of technology.*

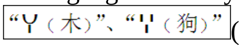
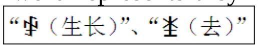
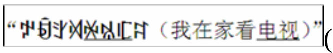
KEYWORDS: *Yi language; Automatic Segmentation; Segmentation algorithm; Difficulties analysis; Development direction*

1. Introduction

Yi nationality is a member with a large population in our national family. According to statistics in 2010, there were more than 9 million people in Yunnan, Sichuan, Guizhou and Guangxi provinces. Yi language belongs to the Yi language branch of the Tibeto-Burman language family of Sino-Tibetan language family. It is divided into six dialect areas. Although some dialects differ greatly, they all have obvious common historical origins. Yi language used to be a syllable character, called “Wei language” or “Luoluo language”, “Lu language”, “Bimo language”, “Xibo language”, commonly known as “ancient Yi language”. now there are about 80,000 glyphs in the ancient Yi language. In 1980, the State Council promulgated the “Yi language Standard Plan” and implemented it in Yi nationality areas in Sichuan. In 2010, the National Yi language Terminology Standardization Committee put forward a general Yi Language program and implemented it in Yunnan, Sichuan, Guizhou and Guangxi provinces.

Yi language belongs to a large character set, therefore, Yi language information processing adds two large tasks: large character set processing and string-to-string processing. Yi language information processing application system needs words as the basic unit as long as it involves retrieval, machine translation, summarization, proofreading, etc. However, due to the complexity of Yi language itself, Yi language word segmentation has become a difficult point in the language segmentation technology.

2. Characteristics of Yi language

On the glyph, Yi language is mostly a single character, which cannot be further divided, such as: (Han-Yi Contro).Syllable, Yi language is a monosyllabic word, a word represents a syllable, that is, a word with independent meaning, such as:  (Han-Yi Contro).Grammatically, Yi language uses word order and function words as the main means of expressing grammatical meaning, such as: (Han-Yi Contro).

Yi language is the same as Chinese, and each character is basically the same size. Chinese is called “square character” and Yi language is called “stone character”. In addition, there is an obvious boundary between words, and there is no obvious boundary between words or between words, which is not divided into words, such as:. In addition, the words in Yi language have no fixed or obvious word segmentation marks such as prefix, suffix and gender, number and case change. These same characteristics of Yi language determine that Yi language also faces the technical problem of automatic word segmentation in Yi language information processing field.

3. The Concept and Research Method of Yi Language Automatic Word Segmentation

3.1 The Concept of Yi Language Automatic Word Segmentation

Word is an important knowledge carrier and basic operation unit in natural language processing system. Yi language information processing technology only involves morphology, syntax analysis and semantic analysis. For example: automatic annotation, information retrieval, automatic proofreading, automatic summarization, machine translation, etc., it is necessary to analyze with words as the basic unit. Yi language automatic word segmentation refers to the use of computers to automatically segment words in Yi language text, that is, like English and Chinese, words in Yi language sentences are marked with spaces between them. Yi language automatic word segmentation is considered as one of the most basic links in Yi language natural language processing. It is the basis of Yi language POS tagging, information retrieval, text classification, machine translation, and Yi language speech recognition and synthesis. It is an essential element for computer Yi language to analyze grammar and understand semantics. It can be analyzed from two aspects. From the perspective of application requirements, the main purpose of Yi language automatic word segmentation is to determine the basic analysis unit of Yi language information processing, which is to do basic work for further development of Yi language automatic analysis, text recognition, automatic summarization, text understanding, machine translation and other technologies. From the point of view of the processing process, the automatic word segmentation of Yi language can be regarded as the process of automatically recognizing words in

Yi language text by computer and adding obvious segmentation marks between words.

3.2 Research Method of Automatic Word Segmentation in Information Processing of Yi Language

At present, the research methods of automatic word segmentation in Yi Language can be summarized into the following three types:

3.2.1 Mechanical Word Segmentation

Mechanical word segmentation mainly includes maximum matching method, reverse maximum matching method, word-by-word matching method, component dictionary method, word frequency statistics method, marking method, parallel word segmentation method, word bank division and association matching method. For example, the experimental center of national language and character information processing of southwest university for nationalities has adopted the forward maximum matching method to design and develop the “Yi language automatic word segmentation system based on the established word list”, and the accuracy of word segmentation has reached more than 85%.

3.2.2 Semantic Word Segmentation

Semantic analysis is introduced to process more linguistic information of natural language itself. For example, extended transfer network method, knowledge word segmentation semantic analysis method, adjacency constraint method, comprehensive matching method, suffix word segmentation method, feature word library method, constraint matrix method, grammar analysis method, etc. For example, the experimental center for ethnic language and word information processing of southwest university for nationalities has adopted Yi language grammar analysis and comprehensive matching method to design and develop “Yi language automatic word segmentation system based on corpus features”. the accuracy rate of word segmentation has reached more than 95%, which is also a representative of the current research on yi language automatic word segmentation technology.

3.2.3 Artificial Intelligence Method

Artificial intelligence is a mode of intelligent processing of information, also known as understanding word segmentation. There are mainly two processing methods: one is the symbol processing method based on psychology. Simulate the function of human brain, like expert system. That is to say, we hope to simulate the function of human brain, construct inference network, and carry out interpretation processing through symbol conversion. One is a physiological-based simulation method. The purpose of neural network is to simulate the operation mechanism of nervous system mechanism of human brain to realize certain functions. At present, the “Yi language Intelligent Word Segmentation System for Information Processing” being developed by the National Language Information Processing Experimental Center of Southwest University for Nationalities adopts this method.

These three methods can be divided into two categories: one is rule-based, and most of the Yi language automatic word segmentation methods currently use this method, such as the Yi language automatic word segmentation system based on the established vocabulary. One is corpus-based, such as Yi language automatic word segmentation system based on corpus features. The calculation models of rule-based word segmentation algorithms are Markov processes in probability theory, also known as meta-grammar, hidden Markov processes and channel noise models in communication. However, whether it is Markov process or channel noise model, they all come down to statistical information for calculating the word frequency of Yi language. string frequency and mutual information are another manifestation of word frequency.

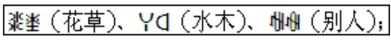
4. Analysis of Difficulties in the Implementation of Yi Language Automatic Word Segmentation

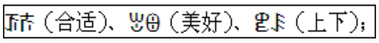
Because the text of Yi Language is composed of continuous characters, there is no space in the middle, and there is no obvious separator like the western language, so it is very difficult for Yi Language to segment words automatically. In view of the characteristics of Yi Language and the current research on automatic word segmentation in computational linguistics, there are two main difficulties in the study of automatic word segmentation in Yi Language.

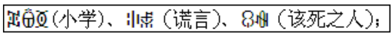
4.1 Linguistic Difficulties

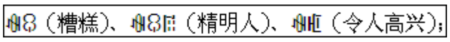
4.1.1 Inconsistent Definitions of Words

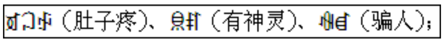
Words are the smallest language unit that can be used independently, which is the formal definition of words in linguistic circles. The specific definition of words has been erratic, so far there is no recognized and authoritative vocabulary. Yi Language also has this difficulty: not only does it not have a unified and strict informal definition, but it also has some problems with formal or ABSTRACT.definitions. This difficulty is caused by the demarcation of words and morphemes on the one hand, and the demarcation of words and phrases (phrases) on the other.

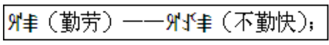
For example: noun+noun structure:  （花草）、（水木）、（别人）；(Han-Yi Contro)

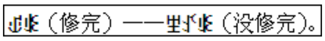
Adjective+adjective Structure:  （合适）、（美好）、（上下）； (Han-Yi Contro)

Adjective+noun Structure:  （小学）、（谎言）、（该死之人）；(Han-Yi Contro)

Noun+adjective Structure:  （糟糕）、（精明人）、（令人高兴）； (Han-Yi Contro)

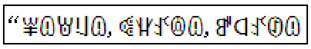
Noun+verb structure:  （肚子疼）、（有神灵）、（骗人）；(Han-Yi Contro)

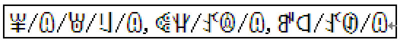
Adjective+negative Structure:  （勤劳）——（不勤快）；(Han-Yi Contro)

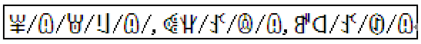
Verb+complement+negative structure:  （修完）——（没修完）。 (Han-Yi Contro)

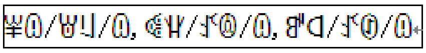
4.1.2 Yi language participle has not yet formed a recognized standard for word segmentation

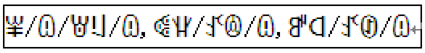
In this way, the same text may be divided into several different results by different people.

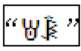
For example: (Should not contend, contend that buckwheat cake is ripe or not, sour soup is warm or not.)has at least four segmentation results:

Word segmentation result 1: 

Word segmentation result 2: 

Word segmentation result 3: 

Word segmentation result 4: 

In addition, there is a large number of (Erbi) in Yi language, namely proverbs and idioms. Its structure is compact and its semantics is complete, but many of its characters can be separated into words individually or combined with other characters or strings to form words, which is also a difficulty in Yi language automatic word segmentation.

4.2 Difficulties in Computer Technology

4.2.1 The grammar knowledge rule base and semantic knowledge rule required by Yi language automatic word segmentation are not perfect yet

Yi language information processing is not a linguistic study in a simple sense. The research on automatic word segmentation technology for information processing also involves many subjects such as computer science, information science, mathematics, automation technology, artificial intelligence, etc. At present, there is no matching and authoritative recognized word segmentation grammar rule in the Yi language information field.

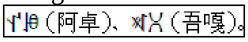
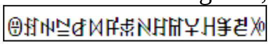
4.2.2 No Reasonable Formal Model of Natural Language

The research on automatic word segmentation technology of Yi Language is still in its infancy. There is no theory or practice about the language model of Yi Language for reference.

4.2.3 Identification and Segmentation Ambiguity Resolution of Unregistered Words

Among the existing Yi language automatic word segmentation methods, the mechanical word segmentation method based on word list and dictionary occupies the leading position. However, the main difficulty of Yi language word segmentation lies not in the matching of entries in dictionaries, but in the identification of unregistered words and the disambiguation of segmentation.

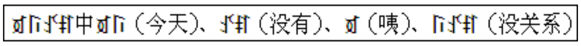
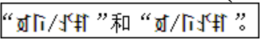
1) Identification of Unregistered Words

In large-scale real text processing, due to the limited capacity of the computer-readable word segmentation dictionary, there must be words not included in the actual culture, mainly including proper nouns such as names of people, places and organizations, as well as new words such as common words and professional terms that emerge with the continuous development of society. The most typical unregistered word in the Yi language participle is a person's name, such as:  （阿卓）、（吾嘎）。 (Han-Yi Contro).Unregistered words also include place names, product names, organization names, trademark names, abbreviations, internet neologisms, etc. For example, organization name:  (National Yi Language Terminology Standardization Committee); Place name: (Double stream); Product name: (Recording pen); Abbreviation: (China and America) and so on.

Since the Yi language text does not have natural word intervals like spaces in western languages, the problem of correct word recognition needs to be solved first when dealing with unregistered words. In a specific context, there may be different recognition and segmentation between unregistered words themselves, between unregistered words and between unregistered words and context. Therefore, the processing of unregistered words becomes a key problem in the use of word segmentation. Improving the recognition performance of unregistered words will definitely improve the overall performance of the analysis system. Nowadays, the accuracy rate of unknown words has become one of the important marks to evaluate the advantages and disadvantages of a word segmentation system. The identification of unregistered words is not only of direct use significance to various Yi language information processing systems, but also plays a fundamental role. At present, there is no special research on the identification of unregistered words in Yi language information processing, but it is simply solved when encountered in Yi language word segmentation. I believe there will be a new breakthrough after the further development of Yi language information processing technology.

2) Disambiguation of Tangent Ambiguity

Ambiguity refers to the same sentence, there may be two or more segmentation methods. One of the difficulties in Yi language word segmentation is the ambiguity processing, because ambiguity is divided into many types, and different solutions should be adopted for different types of ambiguity. At the same time, there is also the problem of ambiguous segmentation in unregistered words, which also increases the difficulty of ambiguous segmentation. The grammar and syntactic structure of Yi language are more complex than that of English. A sentence can be understood as different word strings, phrase strings, etc. and have different meanings in different scenes or different contexts. Therefore, there are many ambiguities and polysemy in Yi language sentences.

For example: ꀋꆏꀕꐥ中ꀋꆏ（今天）、ꀕꐥ（没有）、ꀋ（咦）、ꆏꀕꐥ（没关系） (Han-Yi Contro) are both words. This phrase can be divided into“ꀋꆏ/ꀕꐥ”和“ꀋ/ꆏꀕꐥ”.

Therefore, ambiguity processing is an important factor that affects the segmentation accuracy of the word segmentation system. If the disambiguation problem can be properly handled, the accuracy of word segmentation will be improved accordingly, which is also the direction that needs further research in the future in the design of Yi language automatic word segmentation system.

5. Conclusion

Since there is no obvious segmentation mark between words in Yi language, the research field of Yi language segmentation emerges as the times require in Yi language information processing and becomes one of the basic topics in yi language information processing. Yi language automatic word segmentation technology will have a wide application prospect in Yi language information retrieval, text recognition, machine translation, speech recognition and synthesis and other fields. This paper mainly makes a systematic introduction to the existing automatic word segmentation methods of Yi language. at the same time, combining with the characteristics of Yi language, it analyzes the difficulties in realizing the automatic word segmentation of Yi language from the perspectives of linguistics and computer information processing technology. It is hoped that the preliminary exploration in the cross-cutting field of Yi language modernization and computer Yi language and character information processing can play a role in attracting valuable contributions and jointly develop and perfect this technology.

Acknowledgement

This paper is one of the research results of the project funds:

The National Natural Science Foundation project (71774134), Ministry of Education Project (17YJA740051), the Sichuan social science planning project (SC15B096), the project funded by the innovation team construction plan of southwest minzu university (13TD0058), and the construction project of the minzu language and character information processing laboratory, the key laboratory of Sichuan University.

References

[1] Feng Zhiwei (2001). Computer Chinese Information Processing. Beijing Publishing House, pp. 20-145.
[2] Chamarat Yi (2000). Computer Yi language Information Processing.

Electronic Industry Press, pp. 21-67.

[3] Chen Xiaohe (2000). Automatic Analysis of Modern Chinese. Beijing Language and Culture University Press, pp.35-80.

[4] Deng Hongtao (2005). Design model of Chinese automatic word segmentation system. Computer and Digital Engineering, no.4, pp.138-140.

[5] Sun Tieli, Liu Yanji (2009). Research Status and Difficulties of Chinese Word Segmentation Technology. Information Technology, no.7, pp.187-189

[6] Zhou Wenshuai, Feng Su (2006). Research Status and Application Prospect of Chinese Word Segmentation Technology. Journal of Shanxi Normal University (Natural Science Edition), no.3, pp.32-35.

[7] Daijianying (2005). Research and Implementation of Chinese Automatic Word Segmentation System. Chongqing University, pp.30-50.

[8] Chamarat (2011) .Yi language Information Processing Technology: Thirty Years of Development and Prospect. Journal of chinese information, no.6, pp.170-174.

[9] Wang Chengping (2012). Research on Yi language Automatic Word Segmentation Technology Based on Established Word List. Science, Technology and Engineering, no.10, pp.2328-2332.