Tibetan Number Identification Based on Classification of Number Components in Tibetan Word Segmentation

Huidan Liu
Institute of Software, Chinese Academy of Sciences, Graduate University of the Chinese Academy of Sciences
huidan@iscas.ac.cn

Weina Zhao
Beijing Language and Culture University, Qinghai Normal University
weina@iscas.ac.cn

Minghua Nuo
Institute of Software, Chinese Academy of Sciences, Graduate University of the Chinese Academy of Sciences
minghua@iscas.ac.cn

Li Jiang
Institute of Software, Chinese Academy of Sciences
jiangli@iscas.ac.cn

Jian Wu
Institute of Software, Chinese Academy of Sciences
wujian@iscas.ac.cn

Yeping He
Institute of Software, Chinese Academy of Sciences
yeping@iscas.ac.cn

Abstract

Tibetan word segmentation is essential for Tibetan information processing. People mainly use the basic machine matching method which is based on dictionary to segment Tibetan words at present, because there is no segmented Tibetan corpus which can be used for training in Tibetan word segmentation. But the method based on dictionary is not fit to Tibetan number identification. This paper studies the characteristics of Tibetan numbers, and then, proposes a method to identify Tibetan numbers based on classification of number components. The method first tags every number component according to the class it belongs to while segmenting, and then updates the tag series according to some predefined rules. At last adjacent number components are combined to form a Tibetan number if they meet a certain requirement. In the testing result from 7938K Tibetan corpus, the identification accuracy is 99.21%.

1 Introduction

As a phonetic writing script, Tibetan syllables are separated with syllable dots. But like Chinese, there is no separator between Tibetan words. Tibetan word segmentation is essential for Tibetan information processing. In recent years, many experts did much work on Tibetan word segmentation. CHEN Yuzhong (2003) proposed a method based on case auxiliary words and continuous features to segment Tibetan text. Based on this method, using reinstallation rules to identify Abbreviated Words, CAI Zhijie (2009) designed and implemented the Banzhida Tibetan word segmentation system. QI (2006) proposed a three level method to segment Tibetan text. Dolha (2007), Zhaxijia (2007), CAI Rangjia (2009) and TASHI (2009) researched the word categories and annotation scheme for Tibetan corpus and the parts-of-speech and tagging set standards. At present, there is no corpus for Tibetan word segmentation. However, models which are used in Chinese word segmentation, such as HMM, ME, CRF, have to be trained with segmented corpus. As a result, we can't use them in Tibetan word segmentation. So people mainly use machine matching method based on dictionary in Tibetan word segmentation. But machine matching can not be used to identify Tibetan numbers because we can not include all numbers in the dictionary.

In Tibetan text, numbers have 3 different representations. The first is Arabic numbers, such as "2010". The second is Tibetan alphabet numbers composed with Tibetan digital characters: ཀ(0), ཁ(1), ག(2), ང(3), ཅ(4), ཆ(5), ཇ(6), ཉ(7),

Coling 2008: Poster Volume, pages 719–724,
Beijing, August 2010

ཉ(8), ཊ(9), such as “༢༠༡༠”(2010). The third is Tibetan syllable numbers(“Tibetan numbers” in short) which are composed with Tibetan syllables, such as བཅོ་ལྔ(fifteen).The former two classes of numbers can be identified by combining adjacent number characters. However, this method is not fit to the third class, because some Tibetan syllables are used not only in numbers but also in other common words.

According to papers written by Dolha (2007), Zhaxijia (2007), CAI Rangjia (2009) and TASHI (2009), Tibetan numbers should be taken as single words in Tibetan word segmentation, however, we haven’t found any paper on the issue of the identification of Tibetan numbers in Tibetan word segmentation.

In this paper, we propose a method which is based on classification of number components to identify the third class of numbers.

2 Composition of Tibetan numbers

In Tibetan, we use the following syllables (words) to express the meanings of number one to nine: གཅིག གཉིས། གསུམ། བཞི། ལྔ། དྲུག བདུན། བརྒྱད། དགུ།, and the following different syllables for ten, hundred, thousand, ten thousand, million, ten million and so on: བཅུ། བརྒྱ། སྟོང་། ཁྲི། འབུམ། ས་ཡ། ཨ་བ། དུང་ཕྱུར།…..Generally, Tibetan syllable numbers are composed by these syllables, but some syllables have variants, and sometimes we have to use different conjunctions according to the context. The composition of Tibetan syllable numbers has the following rules.

1. Number 1-10 are expressed with the syllables mentioned above, but sometimes variants are used: ཆིག(1),ཉིས།(2),སུམ།(3).
2. Number “tens” (20, 30, 40 …) have the form of “(2-9)+བཅུ”. but in “20”, “30”, variants of “2” and “3” are used, while in “60”, “70”, “80”, variant of “ten”(ཅུ།) is used.
3. Number 11-19 have the form of “བཅུ(10)+(1-9)”, but in “13” and “15” variant of “10”(བཅོ།) is used.
4. Number 21-99, except “tens”, have the form of “(tens)+conjunction+(1-9)”. Different conjunctions are used according to different “tens”: རྩ། སོ། ཞེ། ང་། རེ། དོན། ཙྭ། ཙོ. Sometimes, this form is abbreviated to “conjunction+(1-9)”.
5. In number which is larger then 100, conjunction (དང་) may be used, just like “and” in the reading of English number “115”. Sometimes, (མེད་) is used to express the meaning of vacancy. For example, number “507” is “ལྔ་བརྒྱ་བཅུ་མེད་བདུན”: ལྔ་(five) བརྒྱ་(hundred)བཅུ་ (ten)མེད་(has no)བདུན (seven).
6. Composition of numbers larger than 1000 can be deduced.
7. Ordinal numeral has the form of “(cardinal numeral)+(པor་པོ)”.
8. Multiples have the form of “ལྷག+ (cardinal numeral)”.
9. Fractions have the form of “(cardinal numeral) +ཆ+ (cardinal numeral)”.
10. Decimals have the form of “(cardinal numeral) +དོ+(གངས་རྫོགorཚེག)+ (cardinal numeral)”. “གངས་རྫོད” or “ཚེག” means the decimal point.
11. Approximate numbers have the form of “(cardinal numeral)+(suffix)”. Suffix can be one of (ཙོ། ཙམ། ཡས་མས། ལྷག་ཙམ། གངས་ཀ་ཤས། ཐུག་ཀ་ཤས།…) according to the meaning to be expressed.
12. Some Tibetan numbers don’t obey the above rules. They have no form of number, but have meanings of number, such as “དང་པོ” (first).

3 Tibetan number identification

In this paper, we call all syllables mentioned in the previous section “number components” in general. For some of these number components, we can take it as a part of number when we meet one of them. For others, we can’t, because they can be used to express non-number meanings. So we have to check whether it is a part of a number according to the context when we meet a number component.

Tibetan number identification is a part of Tibetan word segmentation. In Tibetan word seg-

mentation system, Tibetan text is segmented into words by maximum matching method. In this procedure, every Tibetan number is segmented into number components. Then, identification module combines adjacent number components when they meet a certain predefined rules.

In this section, we first briefly introduce the whole procedure of Tibetan word segmentation, then the classification of number components and the tagging method to identify Tibetan number.

3.1 Flow of Tibetan word segmentation

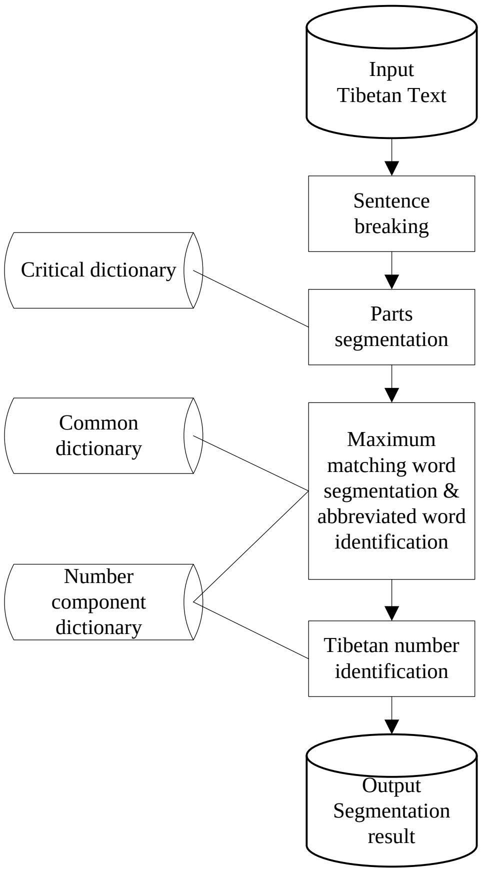

Figure 1. The flow chart of Tibetan word segmentation

As shown in Figure 1, for the input Tibetan text, we first segment it into sentences. Then we segment each sentence into parts with case-auxiliary words. In this procedure, a critical dictionary is used because case-auxiliary words can be a part of some Tibetan words (critical words). When we meet a critical word in Tibetan text, we should not segment it into shorter words. Next, we segment each part into words and identify the abbreviated word (CAI Zhijie, 2009) by maximum matching method with a common dictionary and the number components dictionary. At last, we identify Tibetan numbers and output word segmentation result.

In the procedure of segmenting a part into words, a Tibetan number is segmented into words; we must ensure every one of them is a number component. To do this, both common dictionary and number components dictionary are used. As we use maximum matching method, all Tibetan number words in the common dictionary should be obsolete.

Identification module tags each number component with a tag according to the class which it belongs to, then updates the tags and combines adjacent number components when they meet a certain predefined rules.

3.2 Classification of number components

We classify number components into the following five classes according to their functions and ambiguity.

Basic number: these number components are the basis of Tibetan numbers. Every one of them can be an independent number. If we meet it in context, we should take it as a part of a Tibetan number. Including: Tibetan number 1-9(གཅིག གཉིས། གསུམ། བཞི། ལྔ། དྲུག བདུན། བརྒྱད། དགུ།); ten, hundred, thousand, ten thousand, million, ten million and so on(བཅུ། བརྒྱ། སྟོང་། ཁྲི། འབུམ། ས་ཡ། བྱེ་བ། དུང་ཕྱུར།…); and their variants.

Number prefix: when it is used as a part of Tibetan number, the next word must be a basic number, while the previous word may be or may not be a number component. Including: abbreviations of “(tens)+conjunction” (སོ། ཞེ། ང་། རེ། དོན། གྱ། གོ།); variants of 1, 2, 3(ཅིག ཉིས། སུམ); decimal point(གྲངས་ཚུང་andཚེག).

Number linker: when it is used as a part of Tibetan number, both the previous word and the next word must be number components. These include (དང་། ཆ མེད་). Conjunctions (སོ། ཞེ། ང་། རེ། དོན། གྱ། གོ།) belong to number prefix class, so we don’t include them in this class. But Conjunction (རྩ) doesn’t belong to number prefix class, we include it in this class.

Number suffix: these number components are used to express the meaning of "total number", "approximate number", and "ordinal number" and so on. They follow basic number and should be taken as a part of Tibetan number word. Including: རྒྱ། ཙམ། ཡས་མས། ལྷག་ཙམ། གྲངས་ཀ་ཤས། ཐུག་ཀ་ཤས།…

Independent number: these number components have no form of number, but have meanings of number, such as "དང་པོ" (first).

The difference between "basic number" and "Independent number" is: a basic number can be a Tibetan number itself or a part of a Tibetan number, while an independent number is a Tibetan number itself, but it can't be a part of a Tibetan number.

3.3 Number identification

As shown in Figure 2, identification module tags each number component with a tag according to the class which it belongs to, then updates the tags and combines adjacent number components when they meet a certain predefined rules.

Figure 2. The flow of number identification

Class	Tag
Basic number	N **(**Number**)**
Number prefix	P **(**Prefix**)**
Number linker	L **(**Linker**)**
Number suffix	S**(**Suffix**)**
Independent number	I**(**Independent**)**
Other(non-number)	O **(**Other**)**

Table 1. Classes and their tags

We assign every class with a tag, as shown in Table 1. The tagging procedure screens every segmented part of Tibetan sentences, and tags every word with a tag according to the class which the word belongs to. If the word is not a number component, we tag it with "O" (Other).

As some number components can be used to express non-number meanings, (the cases exist in both number prefix class and number linker class), we have to check whether it is a part of a number according to the context. For number prefix, we take it as a part of number only if it is followed by a basic number, while for number linker, only if it follows a basic number and it is followed by another basic number. We define two rules to do this work.

Rule 1: update tag series "PN" to "NN".
Rule 2: update tag series "NLN" to "NNN".

The tags updating algorithm applies the rules to the current word series until no tag is updated. After tags updating, the tag of a number prefix ("P") is updated to "N" when it is a part of Tibetan number in the context, but the tag will still be "P" when it is not a part of Tibetan number. It is the same for number linkers.

Combination algorithm combines adjacent number components to form a Tibetan number word. It mainly combines continuous number components with tags "NN…N", and the following word is combined too if it has a tag "S". The tag of the number is updated to "N". All words with tag "N" or "I" are taken as Tibetan numbers after combination.

Then the segmentation result is output.

For example, for the following Tibetan sentence:

ལས་འཛོལ་མང་པོ་ཞིག་ནི་བརྒྱ་ཚ་གཅིག་གམ་ཐ་ན་བརྒྱ་ཚ་གྲངས་རྒྱུད་ལྷའི་ནང་ཁོངས་ཀྱི་ལྷ་ལག་གཅིག་ལ་སྟོན་མོར་ནས་བྱུང་འདུག (A considerable parts of accidents were due to the faults of 1% or even 0.5% of components.)

After parts segmentation and maximum matching word segmentation, it is segmented to:

ལས་འཛོལ/ མང་པོ/ ཞིག/ ནི/ བརྒྱ/ ཚ/ གཅིག/ གམ/ ཐ་ན/ བརྒྱ/ ཚ/ གྲངས་རྒྱུད/ ལྷ/ འི/ ནང་ཁོངས/ ཀྱི/ ལྷ་ལག་གཅིག/ ལ/ སྟོན་མོར/ ནས/ བྱུང/ འདུག/

After tagging:

ལས་འཛོལ/(O) མང་པོ/(O) ཞིག/(O) ནི/(O) བརྒྱ/(N) ཚ/(L) གཅིག/(N) གམ/(O) ཐ་ན/(O) བརྒྱ/(N) ཚ/(L) གྲངས་རྒྱུད/(P) ལྷ/(N) འི/(O) ནང་ཁོངས/(O) ཀྱི/(O) ལྷ་ལག་གཅིག/(O) ལ/(O) སྟོན་མོར/(O) ནས/(O) བྱུང/(O) འདུག/(O)

The corresponding tag series is:

OOOONLNOONLPNOOOOOOOOO

After the first run of tags updating, the tag series is changed to:

OOOONNNOONLNNOOOOOOOOO

After the second run of tags updating, the tag series is changed to:

OOOONNNOONNNNOOOOOOOOO

In the third run of tags updating, no tag is updated. Then, combination algorithm combines adjacent number components corresponding to the continuous "N" tags. The result is:

ལས་འཛོལ་/ མང་པོ་/ ཞིག་/ ནི་/ བརྒྱ་ཆ་གཅིག་/ གམ་/ ཐ་ན་/ བརྒྱ་ཆ་གྲངས་ཆུང་ལྔ་/ འི་/ ནང་ཁོངས་/ ཀྱི་/ ལྔ་ལག་གཅིག་/ ལ་/ སྐྱོན་ཤོར་/ ནས་/ བྱུང་/ འདུག/

The corresponding tag series is:

OOOONOONOOOOOOOOOO

It has two "N" tags, which means two Tibetan numbers are identified.

4 Experiment

Corpus	Byte	Sentence	BNS	TNS
Corpus 1	1624K	13957	2590	1667
Corpus 2	1334K	11441	1748	1076
Corpus 3	1408K	11923	1751	969
Corpus 4	1015K	8453	1212	672
Corpus 5	1311K	10445	1613	897
Corpus 6	1246K	10009	1474	880
Total	7938K	66228	10388	6161

Table 2. Information about the 6 corpuses

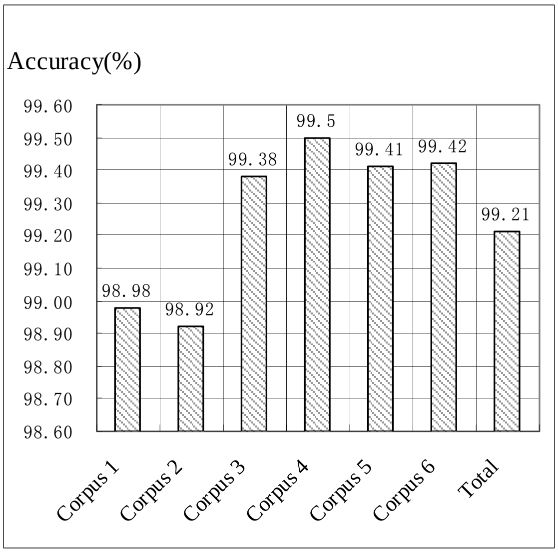

Figure 3. Accuracy of Tibetan number identification on 6 corpuses

As there is no corpus for Tibetan word segmentation, we have to make experiment on original Tibetan texts. We make use of several books which are written in Tibetan, and collect many web pages from several Tibetan web sites. After preprocessing, we get six corpuses. The basic information about the corpuses is shown in Table 2. Note that, in Table 2, the column "BNS" includes all sentences which have in it at least one number component belonging to basic number class, while the column "TNS" includes all sentences which have at least one Tibetan number in it. The count of the former is significantly larger than the count of the later because some basic numbers are used in idioms and proverbs which should be segmented as single words, thus we don't take them as number components under this circumstance. Figure 3 shows the results of our experiment. As we can see, the total identification accuracy is 99.21%. As we have included all basic numbers in our method, theoretically the recall is 100%.

After analyzing the results, we find that wrongly identified words can be divided into two classes. One is that there is a conjunction (དང) between two Tibetan numbers, but is taken as one Tibetan number, such as "བཅུ་དང་ཉི་ཤུ།" (ten and twenty), "ཁྲི་གཅིག་དང་ཁྲི་གཉིས" (ten thousand and twenty thousand). The other is that some Tibetan numbers has other non-number meanings in the context, but our algorithm takes them as numbers. For instance, "ཞེ་གཅིག" means 41 when it is used as a number, but it has another meaning of "similarly"; "དོན་ལྔ" means 75 when it is used as a number, but it has the meaning of "the five internal organs".

5 Conclusion

Tibetan syllables are separated with syllable dots. But like Chinese, there is no separator between Tibetan words. Tibetan word segmentation is essential for Tibetan information processing. People mainly use machine matching in Tibetan word segmentation base on dictionary. But machine matching can not be used to identify Tibetan numbers because we can not include all numbers in our dictionary. This paper proposes a method to tag number components according to the classes they belong to, and then apply predefined rules to update tag series, and next combine adjacent number components to

form a Tibetan number. In the testing result from 7938K Tibetan corpus, the identification accuracy is 99.21%, which means that this method is feasible to be applied to Tibetan word segmentation.

Acknowledgement

We thank the anonymous reviewers for their insightful comments that helped us improve the quality of the paper.

References

CHEN Yuzhong, LI Baoli, YU Shiwen, LAN Cuoji. 2003. An Automatic Tibetan Segmentation Scheme Based on Case Auxiliary Words and Continuous Features, *Applied Linguistics*, 2003(01): 75-82.

CHEN Yuzhong, LI Baoli, YU Shiwen. 2003. The Design and Implementation of a Tibetan Word Segmentation System, *Journal of Chinese Information Processing*, 17(3): 15-20.

CAI Rangjia. 2009. Research on the Word Categories and Its Annotation Scheme for Tibetan Corpus, *Journal of Chinese Information Processing*, 23(04):107-112

CAI Zhijie. 2009. Identification of Abbreviated Word in Tibetan Word Segmentation, *Journal of Chinese Information Processing*, 23(01):35-37.

CAI Zhijie. 2009. The Design of Banzhida Tibetan word segmentation system, *the 12th Symposium on Chinese Minority Information Processing*.

Dolha, Zhaxijia, Losanglangjie, Ouzhu. 2007. The parts-of-speech and tagging set standards of Tibetan information process, *the 11th Symposium on Chinese Minority Information Processing*.

QI Kunyu. 2006. On Tibetan Automatic Participate Research with the Aid of Information Treatment *Journal of Northwest University for Nationalities (Philosophy and Social Science)*, 2006(04):92-97.

SUN Yuan, LUO Sangqiangba, YANG Rui and ZHAO Xiaobing. 2009. Design of a Tibetan Automatic Segmentation Scheme, *the 12th Symposium on Chinese Minority Information Processing*.

TASHI Gyal, ZHU Jie. 2009. Research on Tibetan Segmentation Scheme for Information Processing, *Journal of Chinese Information Processing*, 23(04):113-117.

Zhaxijia, Dolha, Losanglangjie, Ouzhu. 2007. The theoretical explanation on "the parts-of-speech and tagging set standards of Tibetan information process", *the 11th Symposium on Chinese Minority Information Processing*.