A New Algorithm for Component Decomposition and Type Recognition of Tibetan Syllable*

Jie Zhu[1,2], Shugen Wang[3,4] Yanru Wu[1,2], Meijing Guan[1,2], and Yang Feng[3,4]

[1] Department of Computer Science, Tibetan University, Lhasa, China
[2] National-Local Joint Engineering Research Center for Tibetan Information Technology, Lhasa, China
{rocky_tibet,1106539970,zimogmj}@qq.com
[3] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS), Beijing, China
[4] University of Chinese Academy of Sciences, Beijing, China
{wangshugen,fengyang}@ict.ac.cn

Abstract. In this paper, aiming at the problems including but not limited to Tibetan sorting, Tibetan syllable component attribute statistics, Tibetan speech recognition in the application field of component recognition of Tibetan syllable, we propose a new algorithm for component decomposition and type recognition of Tibetan syllable based on TSRM (Tibetan Syllable Rule Model). Experimental results on mixed-arranged complex Tibetan texts show that our newly proposed algorithm can achieve a score around 90% both accuracy rate and recall rate for Component Decomposition and Type Recognition of Tibetan Syllable.

Keywords: Tibetan Syllable · Component Decomposition · Type Recognition.

1 Introduction

Tibetan Information Processing (TIP) has yielded heartening fruits in both character and word level with endeavor of experts in the past three or four decades. International and National Standard for Tibetan Coding for Information Exchange, National Standard for keyboard layout and font style of Tibetan encoding character set have been promulgated successively one after another. In 2018, National Standards for Information Processing of Tibetan Word Segmentation and Character Sorting have been promulgated. The publication of these standards indicates that certain research areas of TIP have achieved recognized results and are extending to the application field gradually. Although

* Supported by National Science Foundation of China(No. 61751216), Key Projects of National Key Research and Develop Plan(No. 2017YFB140220), National Team and Key Laboratory Construction Project for Computer and Tibetan Information(TEF[2018[81).

great achievement has been made in TIP in character and word processing aspect, appropriate algorithms and models for different application scenarios and fields still need to be designed and developed to carry out meticulous research, e.g., the granularity of Tibetan word segmentation in the search engine field can be different from speech synthesis, Tibetan syllables should be separated while attributes and categories of each character should be recognized in fields of sorting and spelling, and so on.

The reason why it is important to research in component decomposition and type recognition of Tibetan Syllable is listed as below:

Firstly, as sorting is one of the basic problems that computer must deal with, which widely exists in spreadsheets and database table of computers, the analogous problem, i.e. sorting of Tibetan character, without doubt, plays an important role in TIP. The sorting of Tibetan syllable is the core in sorting of Tibetan characters, in which the identification of the category of each component is of vital importance, after which, Tibetan strings are sorted according to the sorting rules.

Secondly, Tibetan syllables are spelled automatically in the order of Prefix Character (PC), Super Character (SC), Base Character (BC), Under Character (UC), Vowel (V), First Postfix (FP) and Second Postfix (SP) after the category of each component identified.

Finally, statistics in various attributes need to be gathered on large scale amount of Tibetan texts, e.g., word frequency, syllable frequency, attribute of components in syllables, and so on. In particular, the statistics of attribute and the entropy calculation in Tibetan syllables are fundamental and have reference value for Tibetan coding and design of keyboard layout.

This paper mainly studies the algorithm for component decomposition and type recognition of Tibetan syllable. In this paper, we propose a new algorithm improving the efficiency and execution speed of the algorithm for component decomposition and type recognition of Tibetan syllable with Tibetan Syllable Rule Model (TSRM). The rest of the paper is organized as follows.

In the second part, we introduce the research basis of this topic. In the third part, we study and design the algorithm for component decomposition and type recognition of Tibetan syllable. In the fourth part, we carry out relevant experiments on the new algorithm proposed. The fifth part is the conclusion and prospect of this paper.

2 Related Work

The research on algorithm for component decomposition and type recognition of Tibetan syllable arises with the sorting of Tibetan in computer. As early as the 1990s, Zhaxi Ciren [1] studied the algorithm for sorting of Tibetan syllable and proposed decomposition and sorting of Tibetan syllable according to Unicode Tibetan characters. Jiang Di et al. [2] proposed the concepts of structural order and character order, and discussed the structure of Tibetan syllable in detail. Huang Heming et al. [3] discussed the method of Tibetan sorting based

on DUCET, the basic idea is that other six position of Tibetan syllables except the BC are replaced with spaces if there are no Tibetan characters in them.

Based on a definition of Tibetan component priority, Bianba Wangdui et al. [4] proposed a sorting algorithm for cotemporary Tibetan syllable by Cartesian product. Zhu Jie et al. [5] studied the sorting algorithm of Tibetan and proposed the algorithm of location for BC based on GB for Tibetan Coded. In these articles, Tibetan syllable are all sorted by splitting the syllable and getting BC from the string coding sequence by Tibetan grammar rules, and then determining each component of them gradually.

People found that the key of the sorting algorithm for Tibetan is component recognition of Tibetan syllable during research on sorting algorithm. Bianba Wangdui et al. [6] studied and proposed an algorithm for components recognition of Tibetan syllable to recognize BC with Tibetan grammar rules, the number of syllable and position of component according to Tibetan structure, writing rules and grammatical rules. In general, this algorithm of recognition is a step-by-step process from left to right. Cai Hua [7] studied component recognition automatically of Tibetan syllable and proposed 7 types of structures, while each structure was divided into several different sub-categories and then components of Tibetan syllables were determined according to structures of sub-categories. Renqing Zhuome et al. [8] studied the structure of Tibetan syllables and put forward 7 types of structures, while each structure was divided into different sub-categories and 5-tuple structure were divided into 11 sub-categories.

Another method for component recognition of Tibetan syllables was proposed by Huang Heming et al. [9]. According to the feature of Tibetan encoding, the concepts of placeholder and non-placeholder were proposed, then Tibetan syllables were divided into syllables of placeholder and non-placeholder. For the case of syllables of placeholder, every component of the syllable was judged by the restrictions of the PC, but leave over misjudgment of syllable components like གངས, and others alike. For the case of syllables of non-placeholder, every component of the syllable was judged by the combination rule between SC and BC, appending with UC and BC, and rule between PC and Superposition Character (SPC), and positions among non-placeholder codes, and so on.

In the literature mentioned above, the general method of component recognition of Tibetan syllable is by judgment of the syllable structure from left to right as writing order with the number of character elements. With category structure obtained of component of Tibetan syllable, the sub-category structure of it was judged according to rules of Tibetan grammar, especially rules of syllable composition. The category of component of each character is determined with order of positions of elements in the sub-category structure in the end. The feature of this kind of method is that not only the judgment of number of elements is contained, but also the judgment of structures of both category and sub-category are contained. As every coin has two sides, with various judgments of grammatical rules, complex structures of both branch and cycle structure lead to a high time complexity.

In this paper, we propose a novel algorithm for component decomposition

and type recognition of Tibetan syllable with TSRM [10] from another aspect. For each Tibetan syllable, without Sanskrit translated Tibetan, it can be divided into three parts of prefix, vowel and suffix. Then, PC, SC, BC and UC can be recognized from the part of prefix, while FP and SP can be recognized from the part of suffix. The characteristic of the algorithm is, with complex grammatical judgment process placed in TSRM, only category recognition of the part of prefix need to be considered in component recognition, so the time complexity can be reduced greatly.

3 Algorithm for Component Recognition of Tibetan Syllable

3.1 Problem Analysis

A complete modern Tibetan syllable consists of 7 parts, as shown in Fig. 1. Among them, 1 represents PC, 2 represents SC, 3 represents BC, 4 represents UC, 5 represents V, 6 represents FP, 7 represents SP. In practice, except BC in position 3, other 6 positions can be empty. Therefore, a Tibetan syllable is composed of at least 1 character and at most 7 characters. Tibetan syllable is a

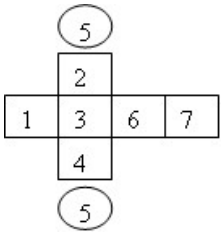


Fig. 1. Tibetan syllable.

sequence of Tibetan character codes stored in computer, as shown in Tab. 1. A Tibetan syllable is composed of at least 1 code and at most 7 codes while the coding position of BC is uncertain in such sequence and can appear in the first, second or third position. The complexity of component recognition increases greatly due to the uncertainty of BC's position, which directly affects the other codes' position in a coding sequence. In the previous literature, BC is determined firstly in component recognition of Tibetan syllable, according to the constitute rule of Tibetan syllable and the grammar theory of Tibetan. Therefore, the category of each component can be recognized with the aid of structural types and grammatical rules, but the time complexity of this cluster of algorithms is very high due to the algorithm complexity. In this paper, a Tibetan syllable is divided into three parts of prefix, vowel and suffix with TSRM, and then PC, SC, BC and UC are recognized from the part of prefix while FP and SP are

Table 1. Relationships among Tibetan syllables, the position of base character and its coding.

Syllable	Coding	BC/NUM	Syllable	Coding	BC/NUM
ག	0F42	1/1	བྱུན	0F56 0F62 0F92 0FB1	3/4
གད	0F42 0F51	1/2	ཀྱམ	0F42 0FB1 0F74 0F62	2/4
ཚམ	0F66 0F92	2/2	ཀྱལ་བ	0F42 0FB1 0F74 0F62 0F51	2/5
གངས	0F42 0F44 0F66	1/3	ཁྱེད	0F62 0F92 0FB1 0F74 0F42	2/5
འཚོ	0F60 0F42 0F7C	2/3	བྱུངས	0F56 0F62 0F92 0FB1 0F63	3/5
བཅུ	0F56 0F62 0F92	3/3	ཁྱེད་ཀྱི	0F62 0F92 0FB1 0F74 0F42 0F66	2/6
རྒྱུ	0F62 0F92 0F53	2/3	བཀའ་གྲས	0F56 0F66 0F92 0FB2 0F42 0F66	3/6
འབངས	0F60 0F42 0F44 0F66	2/4	བསྟན་གྲས	0F56 0F66 0F92 0FB2 0F72 0F42 0F66	3/7

recognized from the part of suffix. Finally, the category of each component is determined after the component decomposition of Tibetan syllable.

3.2 Algorithm for Syllable Decomposition of Tibetan Syllable

According to TSRM, the set of pre-combination charter of Tibetan syllable is established with Tibetan grammar. The set will consist of three parts of prefix set Trule, vowel set Tvowel, and suffix set Tpostfix. The content of Trule, Tvowel and Tpostfix are shown as below:

Trule={ག , ང , ད , ན , བ , མ , འ , ར , ལ , ས , ཨ , ཀ , ཁ , ག , ང , ཅ , ཆ , ཇ , ཉ , ཏ , ཐ , ད , ན , པ , ཕ , བ , མ , ཙ , ཚ , ཛ , ཝ , ཞ , ཟ , འ , ཡ , ར , ལ , ཤ , ས , ཧ , ཨ , ཀྱ , ཁྱ , གྱ , ཅྱ , ཆྱ , ཇྱ , ཏྲ , ཐྲ , དྲ , ནྲ , པྲ , ཕྲ , བྲ , མྲ , ཙྲ , ཚྲ , ཟྲ , ཤྲ , སྲ , ཧྲ , ཀྲ , ཁྲ , གྲ , ཤྭ , སྭ , ཧྭ , ཀྭ , ཁྭ , གྭ , ཀྱྭ , ཁྱྭ , གྱྭ , ཧྲྭ , ཤྲྭ , སྲྭ , ཀྵ , ཧྨ , ཧྨྱ , ཧྭྱ , ཧྲྭྱ , ཤྲྭྱ , སྲྭྱ , ཀྵྱ , ཧྨྲ , ཧྨྲྱ , ཧྭྲ , ཧྭྲྱ , ཤྲྭྲ , ཤྲྭྲྱ , སྲྭྲ , སྲྭྲྱ , ཀྵྲ , ཀྵྲྱ , ཧྨྲྭ , ཧྨྲྭྱ , ཧྭྲྭ , ཧྭྲྭྱ , ཤྲྭྲྭ , ཤྲྭྲྭྱ , སྲྭྲྭ , སྲྭྲྭྱ , ཀྵྲྭ , ཀྵྲྭྱ , ཧྨྲྭྲ , ཧྨྲྭྲྱ , ཧྭྲྭྲ , ཧྭྲྭྲྱ , ཤྲྭྲྭྲ , ཤྲྭྲྭྲྱ , སྲྭྲྭྲ , སྲྭྲྭྲྱ , ཀྵྲྭྲ , ཀྵྲྭྲྱ , ཧྨྲྭྲྭ , ཧྨྲྭྲྭྱ , ཧྭྲྭྲྭ , ཧྭྲྭྲྭྱ , ཤྲྭྲྭྲྭ , ཤྲྭྲྭྲྭྱ , སྲྭྲྭྲྭ , སྲྭྲྭྲྭྱ , ཀྵྲྭྲྭ , ཀྵྲྭྲྭྱ }

Tvowel={ϕ, ི , ུ , ེ , ོ }, where ϕ denotes an implied vowel character.

Tpostfix={ϕ,ག ,ང , ད ,ན , བ ,མ , འ ,ར , ལ ,ས , གས ,ངས , དས , ནས ,མས , ཏས ,རད , ལད }, where ϕ denotes space without FP or SP.

As per TRSM, our algorithm splits a Tibetan syllable into three parts: prefix, vowel and suffix. The algorithm for syllable decomposition of Tibetan syllable is shown in Algorithm. 1.

Algorithm 1 Algorithm for syllable decomposition of Tibetan syllable.

Input:
Tibetan syllable;
Output:
prefix, vowel and suffix;
Creating object fr to read Tibetan syllables;
Creating variables of output string: Tv, Tr, Tp;
Creating string: syllable;
syllable=fr.read();
while syllable ≠ null **do**
  if Tv ∈ syllable in Tvowel **then**
    Tv ⇐ the vowel;
    Tr ⇐ the string before the vowel;
    Tp ⇐ the string after the vowel;
  else
    Tr ⇐ syllable;
    if Tr is not in Trule **then**
      Tp ⇐ the last character of syllable;
      Tr ⇐ the surplus string of syllable;
    end if
    if Tr is not in Trule **then**
      Tp ⇐ the last two characters of syllable;
      Tr ⇐ the surplus string of syllable;
    end if
  end if
end while

Table 2. The relationship between the number of syllable prefix and the position of BC.

BC	BC in prefix/NUM	BC	BC in prefix/NUM
ག	1/1	བསྒ	3/3
གྷ	2/2	བྒྱ	1/3
སྐ	2/2	དཀྱ	2/3
ཀྱ	1/2	སྐྲ	2/3
དྲ	1/2	བསྒྲ	3/4

3.3 Algorithm for Component Recognition of Tibetan syllable

The algorithm for component recognition of Tibetan syllable can be carried out into three steps of component recognition of prefix, vowel and suffix. The second step and the third step are simple while the first step as component recognition of prefix is difficult. The part of prefix in TARM consists of at least 1 character and at most 4 characters. The relationship between the number of syllable prefix and the position of BC is shown in Tab. 2. The prefix is recognized and divided into four different categories by number of characters in the first step of component recognition of prefix while the component recognition of prefix with character number of 1 and 4 is simple and the algorithm focuses on prefix with character number of 2 and 3.

As can be seen from Tab. 2, the syllable prefix can be judged as BC directly when there is only 1 character in it, while the syllable prefix can be judged as the PC, SC, BC and UC when there are 4 characters.

It is difficult to judge the position of BC for it can be in the positions of 1 and 2 of the encoding sequence when the character number of prefix is 2, while it can be in the positions of 1, 2 and 3 of the encoding sequence when the character number of prefix is 3.

For the case of prefix with character number of 2, the first character should be judged whether in set {ར་མགོ, ལ་མགོ, ས་མགོ}, if the first character is in this set, then the first character is determined as SC and the second character is determined as BC. Next, the second character should be judged whether in set UC {◌ྱ , ◌ྲ , ◌ླ , ◌ྭ }, if the second character is in set UC, then the first character is determined as BC and the second character is determined as UC. However, the second character is not in set UC, then the first character is determined as PC and the second character is determined as BC.

For the case of prefix with character number of 3, the third character should be judged whether in set UC {◌ྱ , ◌ྲ , ◌ླ , ◌ྭ }, if the third character is in set UC, then the second character should be judged whether in set UC {◌ྱ , ◌ྲ , ◌ླ , ◌ྭ } as well, if the second character is also in set UC, then the first character is determined as BC and the second and third characters are determined as UC. However, if the second character is not in the set UC, then the first character should be judged whether in set SC {ར་མགོ, ལ་མགོ, ས་མགོ}, if the first character is in set SC, then the first character is determined as SC, the second character is determined as BC and third characters is determined as UC. But, if the third character is not in the set UC, then the first character is determined as PC, the second character is determined as SC and third characters is determined as BC.

The algorithm for component recognition of Tibetan syllable is shown in Algorithm. 2.

4 Experiment

The experiment is divided into two groups: one is the experiment on item, and the other one is the experiment on original corpus. In the experiment on

Algorithm 2 Algorithm for component recognition of Tibetan syllable.

Input:
Tv, Tr, Tp, syllable;
Output:
Type of each component for syllable;
Creating output string variables TrP, TrU, TrB, TrD, Ap, they are PC, SC, BC, UC and second UC respectively;
Tv is V, TpF and TpS are FP and SP respectively;
Creating string: syllable;
while Tr $\neq$ null **do**
  if Tr.lenth=1 **then**
    TrP $\Leftarrow \phi$, TrU $\Leftarrow \phi$, TrB $\Leftarrow$ Tr, TrU $\Leftarrow \phi$;
  else
    if Tr.lenth=2 **then**
      if the first character $\in$ SC **then**
        TrP $\Leftarrow \phi$, TrU $\Leftarrow$ the first character, TrB $\Leftarrow$ the second character, TrD $\Leftarrow \phi$;
      else
        if the second character $\in$ UC **then**
          TrP $\Leftarrow \phi$, TrU $\Leftarrow \phi$, TrB $\Leftarrow$ the first character, TrD1 $\Leftarrow$ the second character;
        else
          TrP $\Leftarrow$ the first character, TrU $\Leftarrow \phi$, TrB $\Leftarrow$ the second character, TrD $\Leftarrow \phi$;
        end if
      end if
    else
      if Tr.lenth=3 **then**
        if the third character $\in$ D AND the second character $\in$ D **then**
          TrP $\Leftarrow \phi$, TrU $\Leftarrow \phi$, TrB $\Leftarrow$ the first character, TrD $\Leftarrow$ the second character, Ap $\Leftarrow$ the third character;
        else
          if the third character $\in$ D AND the first character $\in$ U **then**
            TrP $\Leftarrow \phi$, TrU $\Leftarrow$ the first character, TrB $\Leftarrow$ the second character, TrD $\Leftarrow$ the third character; Ap $\Leftarrow \phi$;
          else
            TrP $\Leftarrow$ the first character, TrU $\Leftarrow$ the second character, TrB $\Leftarrow$ the third character, TrD $\Leftarrow \phi$; Ap $\Leftarrow \phi$;
          end if
        end if
      else
        if Tr.lenth=4 **then**
          TrP $\Leftarrow$ the first character, TrU $\Leftarrow$ the second character, TrB $\Leftarrow$ the third character, TrD $\Leftarrow$ the fourth character;
        else
          print err
        end if
        Tv $\Leftarrow$ Vowel;
        if Tp!=ϕ **then**
          if Tp.lenth=1 **then**
            TpF $\Leftarrow$ Tp, TpS $\Leftarrow \phi$;
          end if
          if Tp.lenth=2 **then**
            TpF $\Leftarrow$ the first character, TpS $\Leftarrow$ the second character;
          end if
        end if
      end if
    end if
  end if
end while

item, the original corpus was classified into three categories of mixing with other languages, non-vowel syllable and vowel syllable. The first experimental file of category of mixing with other languages, named TEST1, is established from the Tibetan original corpus with other linguistic symbols. The second experimental file of category of non-vowel syllable, named TEST2, is established from the Tibetan original corpus with consisting of vowel syllables. The third experimental file of category of vowel syllable, named TEST3, is established from the Tibetan original corpus with consisting of vowel syllables. What's more, the fourth experimental file, named TEST4, is established by download Tibetan corpus from Internet, which contains all the features of three experimental files TEST1, TEST2 and TEST3 above.

4.1 Experimental Corpus

The experimental corpus consists of 100 articles download from China Tibetan Net (www.tibet3.com), which contains news, culture, writing, education, economy, law, pilgrimage, general knowledge and folkways, and so on. The size of TEST1, TEST2, TEST3 and TEST4 are 120 KB, 790 KB, 1.65 MB and 1.71 MB, respectively. TEST1 consists of four articles with mixing of many linguistic symbols. The content of TEST4 is complex for it consists of 100 original articles with not only all kinds of Tibetan coding, e.g., Tibetan symbols, Tibetan characters, Sanskrit translated Tibetan and mistaken Tibetan syllables, but also characters and symbols in Chinese coding and symbols and characters in English.

4.2 Evaluation Criteria and Evaluation

Documents can be divided into two categories of related and unrelated, while the retrieval results can be divided into retrieved and non-retrieved according to the definition of Accuracy Rate (A), Recall Rate (R), and Precision Rate (P) in Information Retrieval (IR). The confusion matrix can be shown in Tab. 3 by evaluation criteria in IR. The Tibetan syllables that consist of correct Tibetan syllable, mistaken Tibetan syllable and Sanskrit translated Tibetan can be divided into two categories of modern Tibetan syllable and non-modern Tibetan syllable. The modern Tibetan syllable are syllables conformed to Tibetan grammar, while the non-modern Tibetan syllable are syllables not conformed to Tibetan grammar including Sanskrit translated Tibetan, special Tibetan syllables in ancient articles corpus, and mistaken Tibetan syllable as well, for convenience. The modern Tibetan syllables correspond to relevant documents, while the non-modern Tibetan syllables correspond to non-related documents. The Tibetan syllable with component recognized correctly corresponds to retrieved documents, while the Tibetan syllable with component recognized incorrectly corresponds to non-retrieved documents.

Let TP represent the number of modern Tibetan syllable components that recognized correctly. Let FP represent the number of non-modern Tibetan syllable components that recognized correctly. FP can be considered as recognized

Table 3. The incidence table of Tibetan syllable decomposition.

	Modern Tibetan syllable	Non-modern Tibetan syllable
Identified correctly	TP	FP
Identified incorrectly	FN	TN

correctly accidentally. Let FN represent the number of modern Tibetan syllable components that recognized incorrectly. Let FN represent the number of non-modern Tibetan syllable components that recognized incorrectly. A, R, and P are used to evaluate the experimental result of component recognition.

A is defined as:

$$A = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

R is defined as:

$$R = \frac{TP}{TP + FN} \tag{2}$$

P is defined as:

$$P = \frac{TP}{TP + FP} \tag{3}$$

We check the experimental result manually, numbers of syllable decomposition and component recognition TP, FP, FN, TN, and evaluation criteria of A, R, P are shown in Tab. 4.

Several screenshots of experimental result selected randomly are shown in Fig. 2. Screenshots of four experimental files TEST1, TEST2, TEST3 and TEST4 are shown from left to right and from top to bottom respectively.

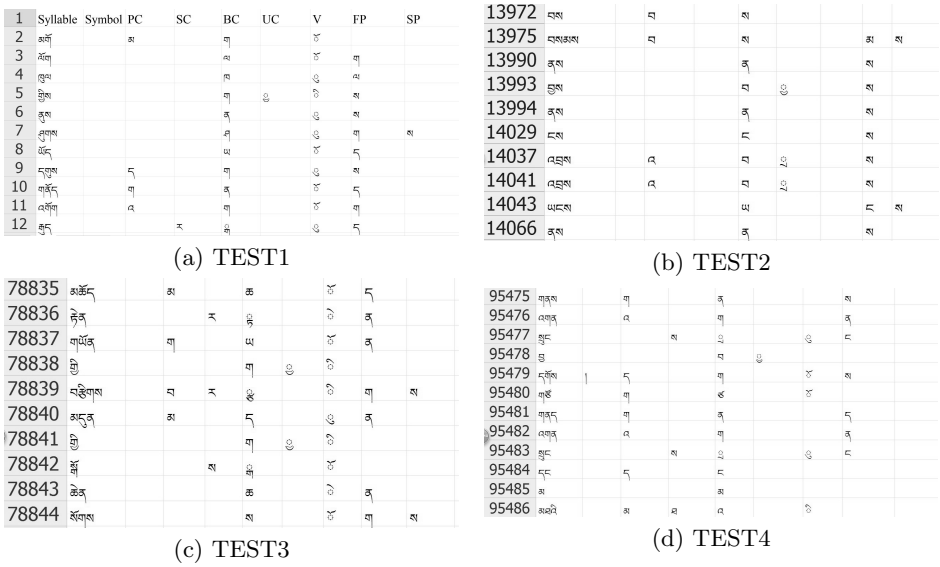
Table 4. Experimental results of syllable decomposition and component recognition.

File	TP	FP	FN	TN	A/%	R/%	P/%
TEST1	8823	0	769	5	91.98	91.98	I00
TEST2	75126	0	8673	165	89.67	89.65	100
TEST3	121612	0	10879	458	91.81	91.78	100
TEST4	196738	0	19552	623	90.98	90.96	100

Through the analysis of the experimental result, the reason for syllable decomposed incorrectly can be summarized as following:

(1) The Tibetan syllable is split incorrectly. Because there are many compact cases and stick to other syllables in Tibetan words, e.g., ས་ར་འ་འི་ཨུ་འོ་འང་འམ་, that cannot restore to two syllables and result in syllable decomposed incorrectly. The occupancy of compact cases as འི་ཨུ་འོ་འང་འམ་ among them is more than 90% in experimental result of FN.

(2) The ambiguity rule characters are judged incorrectly. Rule characters



(a) TEST1

(b) TEST2

(c) TEST3

(d) TEST4

Fig. 2. Screenshots of experimental result.

from {གང་,གན་ , གས་,དག་ , དང་,དབ་ , དཔ་,བག་ , བད་,བས་ , མག་,མང་ , མང་,མན་ } as ambiguity rule characters can either be prefix part of a syllable or be a syllable separately and lead to syllable decomposed incorrectly. The occupancy of experimental file TEST2 is around 70% in experimental result of FN.

(3) The non-modern Tibetan syllable is decomposed incorrectly. For non-modern Tibetan syllable of Sanskrit translated Tibetan, special Tibetan syllables in ancient article, and mistaken Tibetan syllable are not conformed to TSRM lead to syllable decomposed incorrectly, e.g., ཁཀྵི is decomposed to ཁ as PC, ཀ as SC, ྵ as BC, and ི as V. FP is 0 for it is a non-Tibetan syllable and not conformed to modern Tibetan grammar, either in the same grammatical system generally.

5 Conclusion and Prospect

In this paper, we propose a new algorithm for component decomposition and type recognition of Tibetan syllable which can decompose each Tibetan syllable very well and recognize the type of each component correctly by application of TSRM respectively. The algorithm for spelling of Tibetan syllables automatically which can spell most Tibetan syllables that conformed to Tibetan grammar is proposed ahead based on TSRM in another article ever before, while the algorithm for spelling checking of Tibetan syllable which can detect mistaken Tibetan syllables very well is proposed before is also on basis of TSRM in another former article. Compared with other algorithms alike, the algorithm proposed in this work can improve efficiency of component decomposition and type recognition

greatly with implemented simply. Both TSRM previously and the algorithm for component decomposition and type recognition of Tibetan syllable proposed in this paper are in accordance with grammar rule. In the future, machine learning algorithm maybe used in TIP for further research.

References

1. Zhaxi,F. Tibetan Sorting Rules and Come True Automatic Sorting in Computer. China Tibetology **4**,128–135 (1999)
2. Jiang, D., Zhou, J.W. On the Sequence of Tibetan Words and the Method of Making Sequence. Journal of Chinese Information Processing **14**(1), 56–62 (2000)
3. Huang, H.M., Zhao, C.X. A DUCET-based Tibetan Sorting Algorithm. Journal of Chinese Information Processing **22**(4), 109–113 (2008)
4. Banba, W., Drolkar, Dong, Z.C., et al. Study on the Sorting Algorithm of Tibetan Dictionary. Journal of Chinese Information Processing **19**(1), 191–196 (2015)
5. Druggye, Ngogdrup. A Method for Ordering Tibetan Text Based on Tibetan Coded GB. Journal of Tibetan University(Natural Science Edition) **23**(1), 33–35 (2008)
6. Banba, W., Zhuo, G., Chen, Y.L., et al. Study on Recognition Algorithms for Tibetan Construction Elements. Journal of Chinese Information Processing **28**(3), 104–111 (2014)
7. Tshedpa. Research on the Automatic Recognition and Sorting of Tibetan Word Components on the Unicode. Journal of Tibetan University(Natural Science Edition) **29**(2), 80–86 2014
8. Renqing, Z., Qi, K.Y., Gongbao, Z. Research the Types of Seven-tuple Syllables in Tibetan. Journal of Northwest University for Nationalities(Natural Science)**36**(97) 32–36 (2015)
9. Huang, H.M., Da, F.P. Collation-based judgment of modern Tibetan syllable. Journal of Computer Applications **9**(7), 2003–2005 (2009)
10. Zhu, J., Li, T.R., Ge, S., et al. Tibetan Syllable Rule Model and Applications. Acta Scientiarum Naturalium Universitatis Pekinensis **49**(1), 68–74 (2013)