Tibetan Word Segmentation as Sub-syllable Tagging with Syllable's Part-Of-Speech Property

Huidan Liu[1], Congjun Long[1,2], Minghua Nuo[1], and Jian Wu[1]

[1] Institute of Software, Chinese Academy of Sciences, Beijing, China, 100190
[2] Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing, China, 100081

Abstract. When Tibetan word segmentation task is taken as a sequence labelling problem, machine learning models such as ME and CRFs can be used to train the segmenter. The performance of the segmenter is related to many factors. In the paper, three factors, namely strategy on abbreviated syllables, tag set, and the syllable's Part-Of-Speech property, are compared. Experiment data show that: first, if each abbreviate syllable is separated into two units for labelling rather than one, the F-measure improves 0.06% and 0.10% on 4-tag set and 6-tag set respectively. Second, if 6-tag set is used rather than 4-tag set, the F-measure improves 0.10% and 0.14% on the two strategies on abbreviated syllables respectively. Third, when the syllable's Part-Of-Speech property is take into account, F-measure improves 0.47% and 0.41% respectively than the other two methods without using it on 4-tag set, while it improves 0.45% and 0.35% on 6-tag set, which is much more higher than the former improvements. So it's a better choice to take advantage of the syllable's Part-Of-Speech property information while using the sub-syllable as the tag unit.

Keywords: Tibetan word segmentation, Tibetan, sub-syllable tagging, CRFs, syllable's POS property.

1 Introduction

Tibetan text is written without natural word delimiters, so word segmentation is an essential, fundamental and foremost step in Tibetan language processing. In recent years, people take Tibetan word segmentation as a sequence labelling problem and use machine learning models to train a word position tagger for it. However, as there are many abbreviated syllables in Tibetan text, and abbreviation exists in nearly all Tibetan sentences. It's still a not thoroughly solved task to recognize whether a syllable is abbreviated in a certain sentence. Thus, it's a problem to find which tagging unit is the best for Tibetan when taking the word segmentation task as a sequence labelling problem.

In this paper, we compare the influence of different tagging methods to the performance of word segmenter. The paper is organized as follows: In Section 2 we recall related work on Tibetan word segmentation. In Section 3, we simply introduce the concept of Tibetan syllable and abbreviated syllable. We explain different approaches which take syllable and sub-syllable as the tagging unit respectively or with syllable's POS property in Section 4. Then, in Section 5 we make experiments to compare the performances of those approaches. Section 6 concludes the paper.

2 Related Work

In this section, we recall the research history and current situation on Tibetan word segmentation. As Tibetan script is also used to write Dzongkha language, Dzongkha word segmentation related work is also included.

Jiang analysed the problems existing in Tibetan word segmentation in last century, including which word should be included in the dictionary [16]. Zhaxiciren designed and implemented a machine assisted Tibetan word segmentation and new word registration system [42]. However, Jiang thinks it's not an applicable system [16] . Chen *et al.* proposed a method based on case auxiliary words and continuous features to segment Tibetan text [6, 8, 10]. As grammar rules are used, the method is very general and can be used in different domains [6, 8, 10]. Caizhijie designed and implemented the Banzhida Tibetan word segmentation system based on Chen's method, using reinstallation rules to identify the abbreviated words (syllables) [2, 4, 5]. Qi proposed a three level method to segment Tibetan text [28]. Sun *et al.* researched Tibetan Automatic Segmentation Scheme and disambiguation method of overlapping ambiguity in Tibetan word segmentation [32–35]. Dolha *et al.* , Zhaxijia *et al.* , Cairangjia , Gyal and Zhujie made researches on the word categories and annotation scheme for Tibetan corpus and the part-of-speech tagging set standards [1, 12, 14, 43]. Liu proposed a rule based method to process identify Tibetan numbers [23], which is a post procedure of Tibetan word segmentation. Norbu *et al.* described the initial effort in segmenting the Dzongkha scripts [26]. They proposed an approach of Maximal Matching followed by bigram techniques. Experiment shows that it achieves an overall accuracy of 91.5% on all 8 corpora in different domains [26]. Chungku *et al.* described the application of probabilistic part-of-speech taggers to the Dzongkha language, and proposed a tag set containing 66 tags which is applied to annotate their Dzongkha corpus [11].

Before 2010, people mainly use maximum matching method based on dictionary in Tibetan word segmentation [3, 4, 7, 9, 34, 35] accompanying with some grammar rules sometimes. Meanwhile, machine learning models which are used in Chinese word segmentation, such as HMM, ME, CRFs, are widely used in Chinese word segmentation task.

Xue reformulated Chinese word segmentation as a tagging problem [37–39], which is a reform of Chinese word segmentation. The approach uses the maximum entropy tagger to label each Chinese character with a word-internal position tag, and then combines characters into word according to their tags. Ng and Low used the same method in their segmenter [24, 25]. Peng *et al.* first used the CRFs for Chinese word segmentation by treating it as a binary decision task, such that each character is labeled either as the beginning of a word or the continuation of one [27]. Tseng *et al.* presented a Chinese word segmentation system submitted to the closed track of Sighan bakeoff 2005 [36]. This segmenter uses a conditional random field sequence model which provides a framework to use a large number of linguistic features such as character identity, morphological and character reduplication features [36]. In the two International Chinese Word Segmentation Bake-offs held in 2003 and 2005, character based tagging methods using machine learning models quickly rose in two Bakeoffs as a remarkable one with state-of-the-art performance [13, 30]. Especially, two participants, Ng and Tseng gave the best results in almost all test corpora [24, 25, 36].

Tibetan researchers draw inspiration from the methods of Chinese word segmentation by character tagging after 2010, and begin to use machine learning models in Tibetan word segmentation.

Shi ported the Chinese word segmentation system named Segtag to Tibetan word segmentation task and got the Yangjin system in which the Hidden Markov Model (HMM) is used. They get 91% precision on a test set including 25KB text [29]. Jiang used the Conditional Random Fields (CRFs) model with 4 tags (BMES), 10 basic features and 2 additional features to train the word position tagger on a training set including 2500 sentences. He got a 93.5% precision on a test set including 225 sentences. It also showed that the method is better than the maximum matching method [17]. Liu trained a word position tagger for Tibetan with CRFs on a training set including 131903 sentences which are generated by another rule based word segmenter [21, 22] and got a 95.12% F-measure precision. He also compared the influence of different tag sets to the performance the the segmenter. In Liu's method, two additional tag are used to label abbreviated syllables. Li implemented a Tibetan word segmentation system with CRFs and compared the influence of tag sets (tagging units actually) and different processing strategies of the abbreviated syllable recognition [20]. Sun proposed a discriminative model based approach for Tibetan word segmentation. He compared the influence of different word-formation units to the performance, and found syllable is the best unit [31]. He compared different machine learning models, namely CRFs, Maximum Entropy (ME) and Max-Margin Markov Networks (M^3N), and found CRFs is the best one, which get a F1-measure of 94.33% [15].

Generally speaking, when machine learning models such as ME and CRFs are used to train a Tibetan word segmenter. The performance of the segmenter is related to many factors. In the paper, we will compare the influence of three factors, namely strategy on abbreviated syllables, tag set, and the syllable's Part-Of-Speech property.

3 Tibetan Syllable and Abbreviated Syllable

3.1 Tibetan syllable

The Tibetan alphabet is syllabic, like many of the alphabets of India and South East Asia. A syllable contains one or up to seven character(s). Syllables are separated by a marker known as "tsheg", which is simply a superscripted dot. Linguistic words are made up of one or more syllables and are also separated by the same symbol, "tsheg", thus there is a lack of word boundaries in the language. Consonant clusters are written with special conjunct letters. Figure 1 shows the structure of a Tibetan word which is made up of two syllables and means "show" or "exhibition".

Tibetan sentence contains one or more phrase(s), which contain one or more words. Another marker known as "shed" indicates the sentence boundary, which looks like a vertical pipe. (a) shows a Tibetan sentence and (b) is its translation.

(a) དཔོལ་སྤྱི་ཚོགས་རིང་ལུགས་ཀྱི་སྤྱི་ལ་དབང་བའི་ལམ་ལུགས་དང་རྟོག་ བསྟུན་ཐོག་སྤྲོད་ཀྱི་ཐ་དད་མཐའ་འཁོར་ལས་ བྱས་ཡོད།

(b) We have always followed the principles of socialist public ownership and distribution according to work.

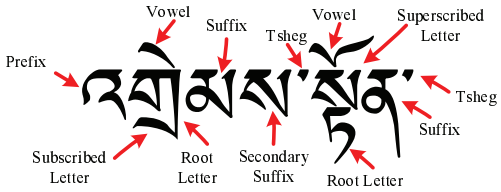

Fig. 1. Structure of a Tibetan word.

3.2 Abbreviated syllables

In Tibetan text, some words, including " འི ", " ས ", " ར ", " འང ", " འམ ", " འོ "(We call them abbreviation marker (AM) in this paper), can glue to the previous word without a syllable delimiter "tsheg", which produce many abbreviated syllables.

For example, when the genitive case word " འི " follows the word " རྒྱལ་པོ " (king), we don't put a "tsheg" between them and get the fused form " རྒྱལ་པོའི " (king[+genitive], king's), in which " པོའི " is an abbreiated syllable. When the ergative case word " ས " follows the word " དཀྱོ " (we), it forms " དཀྱོས " (we[+ergative]), in which " ཀྱོས " is an abbreviated syllable. In the above two examples, either abbreviated syllable should be broken into two parts while segmenting, and the left part has to be combined with the previous syllable(s) to form a word, while the right part is a 1-syllable word. In addition, the word before the AM can be 1-syllable word. For instance, if " འི་ " follows " ང " (I), it forms " ངའི་ " (I [+genitive], my), and the abbreviated syllable should be broken into two 1-syllable words. Table 1 shows more examples.

Table 1. Examples of Tibetan abbreviated syllables. When Tibetan words are followed by one of the abbreviation markers, the suffix letter /a/ (if any) and the tsheg may be omitted.

word	AM	result	explanation
ང་	ས་	ངས་	Tsheg is omitted.
གཡལ་ཆེ་	འི་	གཡལ་ཆེའི་	Tsheg is omitted.
གོ་	འང་	གོའང་	Tsheg is omitted.
རྡོ་བ་	འམ་	རྡོ་བའམ་	Tsheg is omitted.
སྐྱ་	འམ་	སྐྱའམ་	Tsheg is omitted.
དབང་པོ་ལྟ་	འོ་	དབང་པོ་ལྟའོ་	Tsheg is omitted.
སྐུ་མཐའ་	འི་	སྐུ་མཐའི་	འ (/a/) and tsheg are omitted.
ནམ་མཁའ་	འི་	ནམ་མཁའི་	འ (/a/) and tsheg are omitted.
བཤད་པ་	ར་	བཤད་པར་	Tsheg is omitted.
རྒྱལ་པོ་	འི་	རྒྱལ་པོའི་	Tsheg is omitted.

4 Different Methods

4.1 Method 1: Syllable as the tagging unit

The idea of word segmentation as tagging assigns a tag to each unit, namely B (Begin), M (Middle), E (End) and S (Single), according to its position in the word in a certain context. The rules are as follow:

- It's tagged B if it's the left boundary of a word.
- It's tagged M if it's at middle of a word.
- It's tagged E if it's the right boundary of a word.
- It's tagged S if it's a word by itself.

Tags are used to combine the tag units into words in the subsequent procedures. As presented in section 3.2, Abbreviated syllable should be broken into two parts, thus Liu [21] used another two tags ES (End and Single) and SS (Single and Single). Kang [18] used the same strategy. The rules to use the two additional tags are as follow:

- It's tagged ES if it comes from a multiple-syllable word and an AM.
- It's tagged SS if it comes from a single-syllable word and an AM.

Using the above tags, the Tibetan sentence in (a) can be tagged as (c), and the segmentation result is (d):

(c) ང/B ཚོས/ES སྤྱི/B ཚོགས/M རིང/M ལུགས/E ཀྱི/S སྤྱི/B ལ/M དབང/M བའི/M ལམ/M ལུགས/E དང/S རྩོལ/S བསྟུན/S ཐོག/S སྤྲོད/S ཀྱི/S ཙ/B དོན/E མཐའ/B འཁྲོངས/E བྱས/S ཡོད/S །/S

(d) ང་ཚོ/ ས/ སྤྱི་ཚོགས་རིང་ལུགས/ ཀྱི/ སྤྱི་ལ་དབང་བའི་ལམ་ ལུགས/ དང/ རྩོལ/ བསྟུན/ ཐོག/ སྤྲོད/ ཀྱི/ ཙ་དོན/ མཐའ་འཁྲོངས/ བྱས/ ཡོད/ །/

Model training stage: When Tibetan word segmentation corpus is being converted into training format, abbreviation marks are combined with the former syllable to form an abbreviated syllable, and get the tag ES or SS. All other syllables, which have an abbreviated mark at the tail but aren't abbreviated syllables actually, will get the tag B, M, E or S. Thus, the model is able to make the disambiguation.

Model applying stage: In the method, as each unit will have a tsheg with it, when applying the model, Tibetan text are segmented into syllables by tshegs. Then, the model assigns each syllable a tag. Syllables are combined into words according to their tags and meanwhile abbreviated syllables are segmented into tow syllables.

4.2 Method 2: Sub-syllable as the tagging unit

As abbreviation marks plays an important role in Tibetan text, they provide much clearer informations to the tagger than those abbreviated syllables. Thus if abbreviation marks are taken as the tag units, more context information will be provided to the model. We call this method "sub-syllable as the tagging unit", because syllables which have an abbreviation mark as suffix will be segmented into two units. So the unit is smaller than a syllable actually. Using the method, the above mentioned Tibetan sentence will be tagged as follow:

(e) ར/B ཚོ/E ས/S སྤྱི/B ཚོགས/M རིང/M ལུགས/E ཀྱི/S སྤྱི/B ལ/M དབང/M བ/M འི/M ལམ/M ལུགས/E དང/S རྩོལ/S བསྟུན/S ཐོག/S སློབ/S ཀྱི/S ཙ/B དོན/E མཐའ/B འཛོམས/E ཟ/B ས/E ཡོད/S །/S

The differences between (e) and (c) are as follow:

- ཚོས་ is tagged as ཚོས་/ES in (c) but ཚོ/E ས/S in (e) because it's an abbreviated syllable.
- བའི་ is tagged as བའི་/M in (c) but བ/M འི/M in (e) because it's an abbreviated syllable but occurs at the middle of the word.
- ཟས་ is tagged as ཟས་/S in (c) but ཟ/B ས/E in (e) because it's not an abbreviated syllable but has an AM (ས) as the suffix.
- ཚོགས་ ,ལུགས་ ,ལུགས་ and འཛོམས་ are tagged as the same in (c) and (e) because the ས is at the secondary suffix position rather than suffix position in those syllables, so it can't be an AM by checking spelling rules and those syllables can't be abbreviated syllables. (See Figure 1.)
- The two tags ES and SS used in (c) aren't used in (e) because abbreviated syllables are segmented into sub-syllables in (e).

Model training stage: When Tibetan word segmentation corpus is being converted into training format, abbreviation marks in abbreviated syllable get the tag S. Normal syllables which have a abbreviation mark in its suffix position will be segmented into two units, and the abbreviation mark in them gets a tag M or E. The model is able to make the disambiguation too.

Model applying stage: When applying the model, Tibetan text are segmented into syllables by tshegs and syllables with an AM as the suffix are segmented into sub-syllables further. Then, the model assigns each syllable a tag. Syllables are combined into words according to their tags.

4.3 Method 3: Sub-syllable tagging with syllable's POS property

In many cases, a syllable is a word itself and has semantic meanings. All multi-syllable word can be taken as formed by several monosyllable words. As a syllable is a word, it also has the part-of-speech property. So we can assign a POS tag to each syllable denoting it's a noun, verb or others. The POS property may contribute to word segmentation.

So, based on Method 2, we assign each sub-syllable with a combined tag of the syllable's POS property tag and the in-word position tag. This is Method 3.

(f) 【ར/rh ཚོ/pl 】【ས/ka 】【སྤྱི/a ཚོགས/n རིང/n ལུགས/n 】【ཀྱི/kg 】【སྤྱི/a ལ/kp དབང/v བ/h འི/kg ལམ/n ལུགས/n 】【དང/c 】【རྩོལ/n】【བསྟུན/v 】【ཐོག/v 】【སློབ/v 】【ཀྱི/kg 】【ཙ/n དོན/n 】【མཐའ/n འཛོམས/v 】【ཟས/v 】【ཡོད/ve 】【།/xp 】

(g) ར/rh-B ཚོ/pl-E ས/ka-S སྤྱི/a-B ཚོགས/n-M རིང/n-M ལུགས/n-E ཀྱི/kg-S སྤྱི/a-B ལ/kp-M དབང/v-M བ/h-M འི/kg-M ལས/n-M ལུགས/n-E དང/c-S རྩོལ/n-S བསྟུན/v-S ཐོག/v-S སྒྲུབ/v-S ཀྱི/kg-S ཪ/n-B དོན/n-E སབཞང/n-B འཚེངས/v-E ཐ/v-B ས/v-E ཡོད/ve-S ।/xp-S

Model training stage: The corpus are prepared as (f), and converted to (g), which is used to train a tagger. As the POS property is on syllable level, but the tag unit is a sub-syllable, we have to split a certain syllable into two sub-syllables. It occurs when a normal syllable has an abbreviation mark as suffix. In the above sentence, ཐས/v is broken into two sub-syllables and tagged as ཐ/v-B ས/v-E . The first part of the combined tag denotes the POS property while the last part denotes the word boundary property.

Model applying stage: When applying the model, Tibetan text are segmented into syllables by tshegs and syllables with an AM as the suffix are segmented into sub-syllables further. Then, the model assigns each syllable a combined tag. Syllables are combined into words according to the last part of the combined tag.

5 Experiments and Results

5.1 Corpus

A corpus from some textbooks used in primary school and middle school is used in this work. Sentences are segmented into words and syllables and tagged with the POS property tag. Word boundaries are marked by special characters. About $1/5$ of the corpus are randomly selected as the test set, 3,983 sentences (47,332 words) in total. The remaining 15,931 sentences (191,852 words) forms the training set. The OOV rate of the test set is 5.34%.

5.2 Tag set

Generally, tag set $\{B, M, E, S\}$ (4-tag set) and $\{B, B2, B3, M, E, S\}$ (6-tag set) [21, 40, 41] are used in the experiments to compare the influence of different tag sets on the performance. The difference is that two additional tags are used for units at the middle of a word. The first middle unit is tagged as B2, while the second middle unit is tagged as B3. Table 2 compares the tag results by the two tag sets. As needed another two additional tags ES and SS are used on Method 1 to tag the abbreviated syllables.

5.3 Machine learning Model

Maximum Entropy (ME) tagger was used in early character-based tagging for Chinese word segmentation [24,25,37–39], In recent years, more and more people choose linear-chain CRFs as the machine learning model in their studies [27, 36, 40, 41].

CRFs model is firstly introduced into language processing by Lafferty [19]. Peng *et al.* [27] first used this framework for Chinese word segmentation by treating it as a

Table 2. Tag results by 4-tag set(4nt) and 6-tag set(6nt) on words with different lengths. The first, second and other middle units in a word are tagged as B2, B3 and M respectively by 6-tag set, while all of them are tagged M by 4-tag set.

word	tag set	tag sequence	sub-syllable/tag sequence
དེ	4nt	S	དེ /S
	6nt	S	དེ /S
རྟ་ཅང་	4nt	B-E	རྟ་ /B ཅང་ /E
	6nt	B-E	རྟ་ /B ཅང་/E
སྒྲིབ་ལུགས་ཅན་	4nt	B-M-E	སྒྲིབ་ /B ལུགས་ /M ཅན་ /E
	6nt	B-B2-E	སྒྲིབ་ /B ལུགས་ /B2 ཅན་/E
ཐུབ་ཅེ་གཅིག་སྟེམས་	4nt	B-M-M-E	ཐུབ་ /B ཅེ་ /M གཅིག་ /M སྟེམས་ /E
	6nt	B-B2-B3-E	ཐུབ་ /B ཅེ་ /B2 གཅིག་ /B3 སྟེམས་/E
གནམ་རིག་མཁས་པ	4nt	B-M-M-M-E	གནམ་ /B རིག་ /M མཁ /M ས་ /M པ /E
	6nt	B-B2-B3-M-E	གནམ་ /B རིག་ /B2 མཁ /B3 ས་ /M པ/E
རྩེ་དགའི་རྣམ་འགྱུར་	4nt	B-M-M-M-M-E	རྩེ་ /B དགཱ /M འི་ /M རྣམ་ /M འགྱུ /M ར་ /E
	6nt	B-B2-B3-M-M-E	ཙེ་ /B དགཱ /B2 འི་ /B3 རྣམ་ /M འགྱུ /M ར་/E
ཚར་འགྲོ་རྒྱུར་འདོད་བྱེད་	4nt	B-M-M-M-M-M-E	ཚ /B ར /M འགྲོ་ /M རྒྱུ /M ར /M འདོད་ /M བྱེད་ /E
	6nt	B-B2-B3-M-M-M-E	ཚ /B ར /B2 འགྲོ་ /B3 རྒྱུ /M ར /M འདོད་ /M བྱེད་/E

binary decision task, such that each Chinese character is labelled either as the beginning of a word or not.

The probability assigned to a label sequence for a tagging unit sequence by a CRFs is:

$$p_\lambda(Y|W) = \frac{1}{Z(W)} exp\left(\sum_{t\in T}\sum_{k}\lambda_k f_k(y_{t-1}, y_t, W, t)\right). \quad (1)$$

where $Y = y_i$ is the label sequence for the sentence, W is the sequence of unsegmented units, $Z(W)$ is a normalization term, f_k is a feature function, and t indexes into units in the label sequence.

As theory and research practices on many sequence labelling tasks show that CRFs is better than ME, we use CRFs to train the taggers in this work. The CRF++ toolkit 0.58 [3] by Taku Kudo is used.

5.4 Feature Template

A 5-unit context window and TMPT-10 defined in Table 3 are used in this work.

5.5 Comparison and analysis

In the work, several experiments are made to compare factors that impact the performance.

[3] http://taku910.github.io/crfpp

Table 3. Feature templates TMPT-10 used in this paper. A 5-unit context window is used. The unigrams and bigrams of the units are used to express the context of the current unit.

Feature	Explanation
$C_n,\ n = -1, 0, 1$	The previous, current and next unit
C_{-2}	The unit before the previous unit
C_2	The unit after the next unit
$C_nC_{n+1},\ n = -1, 0$	The previous (next) unit and current unit
$C_{-1}C_1$	The previous unit and next unit
C_1C_2	The next two units
$C_{-2}C_{-1}$	The previous two units

Table 4. Performance comparison of the 3 methods. Method 2 (m2) improves the F-measure by 0.06% and 0.10% on 4-tag set (4nt) and 6-tag set (6nt) respectively over Method 1 (m1). Method 3 (m3) improves the F-measure by 0.47% and 0.41% over the other methods without using it on 4-tag set, while it improves 0.45% and 0.35% on 6-tag set.

	R(%)	P(%)	F1(%)	R(OOV)	R(IV)		R(%)	P(%)	F1(%)	R(OOV)	R(IV)
m1-4nt	93.99	93.78	93.88	70.84	95.29	m1-6nt	94.14	93.82	93.98	70.45	95.48
m2-4nt	94.12	93.77	93.94	69.70	95.51	m2-6nt	94.32	93.83	94.08	69.51	95.73
m3-4nt	94.63	94.07	**94.35**	68.08	96.13	m3-6nt	94.88	93.99	**94.43**	68.04	96.39

Influence of different strategies on abbreviated syllables. Table 4 also shows that Method 2 gets an F-measure improvement of 0.06% and 0.10% on 4-tag set(4nt) and 6-tag set (6nt) respectively. The recall improvements are 0.13% on 4-tag set and 0.18% on 6-tag set, while the precisions almost keep the same. It seems that Method 2 leads to a more significant improvement on recall than on precision. Comparing with Method 1, the OOV recall rate declines 1.14% on 4-tag set and 0.94% on 6-tag set, while the IV recall rate improves 0.22% and 0.25%.

Influence of different tag sets. Table 4 shows that 6-tag set (6nt) gets an F-measure improvement of 0.10% on Method 1 and 0.14% on Method 2 respectively compared with 4-tag set (4-nt). Meanwhile, the recall and precision are both improved when 6-tag set is used rather than 4-tag set, which shows that 6-tag set is slightly better than 4-tag set.

Influence of syllable's POS property. Table 4 lists the performance data of the three methods on 4-tag set. It shows that Method 3 outperforms the other two methods. Comparing with Method 1 and Method 2, the recall and precision improve both significantly, thus Method 3 gets an F measure improvement of 0.47% and 0.41% respectively over the other two methods without using it on 4-tag set, while it improves 0.45% and 0.35% on 6-tag set, which is much more higher than those former improvements. Consequently, the syllable's POS property is a more important factor to improve the

overall performance than the strategy on abbreviated syllables and the tag set. However, comparing with the other two methods, Method 3 has a worse performance on the Out-Of-Vocabulary words.

6 Conclusion

The performance of Tibetan word segmenter is related to many factors. Three factors are compared in the paper, namely strategy on abbreviated syllables, tag set, and the syllable's Part-Of-Speech property. Experiment data show that: first, if each abbreviate syllable is separated into two units for labelling rather than one, the F-measure improves 0.06% and 0.10% on 4-tag set and 6-tag set respectively. Second, if 6-tag set is used rather than 4-tag set, the F-measure improves 0.10% and 0.14% on the two strategies on abbreviated syllables respectively. Third, when the syllable's Part-Of-Speech property is taken into account, F-measure improves 0.47% and 0.41% respectively over the other two methods without using it on 4-tag set, while it improves 0.45% and 0.35% on 6-tag set, which is much more higher than the former improvements. So it's a better choice to take advantage of the syllable's Part-Of-Speech property information while using the sub-syllable as the tag unit.

Acknowledgements We thank the reviewers for their critical and constructive comments and suggestions that helped us improve the quality of the paper. The research is partially supported by National Science Foundation (No.61202219, No.61202220, No.61303165) and Informationization Project of the Chinese Academy of Sciences (No.XXH12504-1-10).

References

1. Cairangjia: Research on the word categories and its annotation scheme for tibetan corpus. Journal of Chinese Information Processing 23(04):107-112 (2009)
2. Caizhijie: The design of banzhida tibetan word segmentation system. In: Researches and Advancements of Information Processing for Chinese Minority Languages and Characters (2009)
3. Caizhijie: The design of banzhida tibetan word segmentation system. In: the 12th Symposium on Chinese Minority Information Processing (2009)
4. Caizhijie: Identification of abbreviated word in tibetan word segmentation. Journal of Chinese Information Processing 23(01), 35–37 (2009)
5. Caizhijie: The design of banzhida tibetan word segmentation system. Journal of Ethic Normal College of Qinhai Normal University 2, 75–77 (2010)
6. Chen, Y., Li, B., Yu, S.: The design and implementation of a tibetan word segmentation system. Journal of Chinese Information Processing 17(3), 15–20 (2003)
7. Chen, Y., Li, B., Yu, S.: The design and implementation of a tibetan word segmentation system. Journal of Chinese Information Processing 17(3): 15-20 (2003)
8. Chen, Y., Li, B., Yu, S., Lan, C.: An automatic tibetan segmentation scheme based on case auxiliary words and continuous features. Applied Linguistics 1, 75–82 (2003)
9. Chen, Y., Li, B., Yu, S., Lancuoji: An automatic tibetan segmentation scheme based on case auxiliary words and continuous features. Applied Linguistics (01): 75-82 (2003)

10. Chen, Y., Yu, S.: The present situation and prospect of the study of technological methods concerning handling the information in tibetan script. China Tibetology (04):97-107 (2003)
11. Chungku, C., Rabgay, J., Faa?, G.: Building nlp resources for dzongkha: A tagset and a tagged corpus. In: Proceedings of the 8th Workshop on Asian Language Resources. pp.103-110. Beijing, China (2010)
12. Dolha, Zhaxijia, Losanglangjie, Ouzhu: The parts-of-speech and tagging set standards of tibetan information process. In: the 11th Symposium on Chinese Minority Information Processing (2007)
13. Emerson, T.: The second international chinese word segmentation bakeoff. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pp.123-133. Jeju Island, Korea (2005)
14. Gyal, T., Zhujie: Research on tibetan segmentation scheme for information processing. Journal of Chinese Information Processing 23(04):113-117 (2009)
15. He, X., Li, Y., Ma, N., Yu, H.: Study on tibetan automatic word segmentation as syllable tagging. Application Research of Computers 32(1):61-65 (2015)
16. Jiang, D.: History and progress of tibetan text information processing. In: Frontiers of Chinese information processing –proceedings of the 25th anniversary conference of Chinese information processing society. pp. 83–97. Beijing:Press of Tsinghua university (2006)
17. Jiang, T.: Tibetan word segmentation system based on conditional random fields. Software Engineering and Service Science (ICSESS) pp. 446–448 (2011)
18. Kang, C., Jiang, D., Long, C.: Tibetan word segmentation based on word-position tagging. In: 2013 International Conference on Asian Language Processing (IALP). pp. 239–242. IEEE (2013)
19. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp.282-289 (2001)
20. Li, Y., Jam, Y., Zong, C., Yu, H.: Research and implementation of tibetan automatic word segmentation based on conditional random field. Journal of Chinese Information Processing 27(4):52-58 (2013)
21. Liu, H., Nuo, M., Ma, L., et al: Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Fields. In: Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 2011). pp. 168–177 (2011)
22. Liu, H., Nuo, M., Zhao, W., et al: SegT: A practical tibetan word segmentation system. Journal of Chinese Information Processing 26(1):97-103 (2012)
23. Liu, H., Zhao, W., Nuo, M., Jiang, L., Wu, J., He, Y.: Tibetan number identification based on classification of number components in tibetan word segmentation. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 719–724. Association for Computational Linguistics (2010)
24. Low, J.K., Ng, H.T., Guo, W.: A maximum entropy approach to chinese word segmentation. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pp.161-164. Jeju Island, Korea (2005)
25. Ng, H.T., Low, J.K.: Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based. In: Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing, pp.277-284 (2004)
26. Norbu, S., Choejey, P., Dendup, T., Hussain, S., Mauz, A.: Dzongkha word segmentation. In: Proceedings of the 8th Workshop on Asian Language Resources. pp.95-102. Beijing, China (2010)
27. Peng, F., Feng, F., McCallum, A.: Chinese segmentation and new word detection using conditional random fields. In: Proceedings of the 20th International Conference on Computational Linguistics, pp.562-568. Geneva, Switzerland (2004)

28. Qi, K.: Research on tibetan automatic word segmentation for information processing. Journal of Northwest University for Nationalities(Philosophy And Social Science) (04):92-97 (2006)
29. Shi, X., Lu, Y.: A tibetan segmentation system – yangjin. Journal of Chinese Information Processing 25(4), 54–56 (2011)
30. Sproat, R., Emerson, T.: The first international chinese word segmentation bakeoff. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp.133-143. Sapporo, Japan (2003)
31. Sun, M., Huaquecairang, Caizhijie, Jiang, W., et al: Tibetan word segmentation based on discriminative classification and reranking. Journal of Chinese Information Processing 28(2):61-66 (2014)
32. Sun, Y., Luosangqiangba, Yang, R., Zhao, X.: Design of a tibetan automatic segmentation scheme. In: Researches and Advancements of Information Processing for Chinese Minority Languages and Characters - Proceedings of the 12th Symposium on Chinese Minority Information Processing. pp. 228–237 (2009)
33. Sun, Y., Luosangqiangba, Yang, R., Zhao, X.: Study of segmentation strategy on tibetan crossing ambiguous words. In: Researches and Advancements of Information Processing for Chinese Minority Languages and Characters. pp. 238–243 (2009)
34. Sun, Y., Wang, Z., Zhao, X., et al: Design of a tibetan automatic word segmentation scheme. In: Proceedings of 2009 1st IEEE International Conference on Information Engineering and Computer Science. pp. 1–6 (2009)
35. Sun, Y., Yan, X., Zhao, X., et al: A resolution of overlapping ambiguity in tibetan word segmentation. In: Proceedings of 2010 3rd International Conference on Computer Science and Information Technology. pp. 222–225 (2010)
36. Tseng, H., Chang, P., Andrew, G., Jurafsky, D., Manning, C.: A conditional random field word segmenter for sighan bakeoff 2005. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pp.168-171. Jeju Island, Korea (2005)
37. Xue, N.: Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing 8(1): 29-48 (2003)
38. Xue, N., Converse, S.P.: Combining classifiers for chinese word segmentation. In: Proceedings of the First SIGHAN Workshop on Chinese Language Processing, pp.63-70. Taipei, Taiwan (2002)
39. Xue, N., Shen, L.: Chinese word segmentation as lmr tagging. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL03, pp.176-179. Sapporo, Japan (2003)
40. Zhao, H., Huang, C.N., Li, M.: An improved chinese word segmentation system with conditional random field. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp.108-117. Sidney,Australia (2006)
41. Zhao, H., Huang, C., Li, M., Lu, B.: Effective tag set selection in chinese word segmentation via conditional random field modeling. In: Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, pp.87-94 Wuhan, China (2006)
42. Zhaxiciren: The design of a machine assisted tibetan word segmentation and new word registeration system. In: Proceedings of modernization of Chinese minority nationality languages (1999)
43. Zhaxijia, Dolha, Losanglangjie, et al: Theoretical explanation on the parts-of-speech and tagging set standards of tibetan information processing. In: Procedings of the 11th China national conference on minority language information processing. pp. 441–452 (2007)