Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Field *

Huidan Liu[a,b], Minghua Nuo[a,b], Longlong Ma[a], Jian Wu[a], and Yeping He[a]

[a] Institute of Software, Chinese Academy of Sciences,
No.4 South Fourth Street, Zhong Guan Cun, Haidian District, Beijing 100190, China
[b] Graduate University of the Chinese Academy of Sciences,
No.80 Zhongguancun East Road, Haidian District, Beijing 100190, China
{huidan, minghua, longlong, wujian, yeping}@iscas.ac.cn

Abstract. In this paper, we proposed a novel approach for Tibetan word segmentation using the conditional random field. We reformulate the segmentation as a syllable tagging problem. The approach labels each syllable with a word-internal position tag, and combines syllable(s) into words according to their tags. As there is no public available Tibetan word segmentation corpus, the training corpus is generated by another segmenter which has an F-score of 96.94% on the test set. Two feature template sets namely TMPT-6 and TMPT-10 are used and compared, and the result shows that the former is better. Experiments also show that larger training set improves the performance significantly. Trained on a set of 131,903 sentences, the segmenter achieves an F-score of 95.12% on the test set of 1,000 sentences.

Keywords: Tibetan word segmentation, Tibetan, syllable tagging, CRF.

1 Introduction

During the 7th Century AD, Thume Sambota reputedly devised a script for Tibetan based on the Devanagari model. At present, Tibetan script is used to write both Tibetan language and Dzongkha (Bhutanese) language. The former is a language spoken by about 6 million people in China, India, Bhutan, Sikkim, Ladakh and Nepal, while the later is spoken by about 130,000 people in Bhutan as a national and official language.

Just like Chinese, Tibetan text is written without natural word delimiters, so word segmentation is an essential, fundamental and foremost step in Tibetan language processing. Researchers mainly use maximum matching method and some grammar rules to segment Tibetan text at present.

In this paper, we reformulate Tibetan word segmentation as a syllable tagging problem, and propose an approach using the conditional random field (CRF) for Tibetan word segmentation. The paper is organized as follows: In Section 2 we recall related work on Tibetan word segmentation and Chinese word segmentation methods by character tagging. In Section 3, we simply introduce Tibetan script. We propose the approach in Section 4. Then, in Section 5 we make experiments to evaluate the segmenter, compare it with other Tibetan segmenters. Section 6 concludes the paper.

2 Related Work

In this section, we first recall the research history and current situation on Tibetan/Dzongkha word segmentation. We also recall research on Chinese word segmentation by character tagging which we have drawn inspiration from.

* The research is partially supported by National Science & Technology Major Project (No.2010ZX01036-001-002, No.2010ZX01036-001-002) and CAS Action Plan for the Development of Western China (No.KGCX2-YW-512).

2.1 Tibetan/Dzongkha Word Segmentation

Chen *et al.* (2003a; 2003b) proposed a method based on case auxiliary words and continuous features to segment Tibetan text. Caizhijie (2009a; 2009b) designed and implemented the Banzhida Tibetan word segmentation system based on Chens method, using reinstallation rules to identify Abbreviated Words. Qi (2006) proposed a three level method to segment Tibetan text. Sun *et al.* (2009; 2010) researched Tibetan Automatic Segmentation Scheme and disambiguation method of overlapping ambiguity in Tibetan word segmentation. Dolha *et al.* (2007), Zhaxijia *et al.* (2007), Cairangjia (2009), Gyal and Zhujie (2009) made researches on the word categories and annotation scheme for Tibetan corpus and the part-of-speech tagging set standards.

Norbu *et al.* (2010) described the initial effort in segmenting the Dzongkha scripts. 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. Chungku *et al.* (2010) 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.

Models which are used in Chinese word segmentation, such as HMM, ME, CRF, need to be trained with corpus. However, there is no public available corpus for Tibetan word segmentation at present. So people mainly use maximum matching method based on dictionary in Tibetan word segmentation(Chen *et al.*, 2003a; Chen *et al.*, 2003b; Caizhijie, 2009a; Caizhijie, 2009b; Sun *et al.*, 2009; Sun *et al.*, 2010), accompanying with some grammar rules sometimes.

2.2 Chinese Word Segmentation by Character Tagging

Xue reformulated Chinese word segmentation as a tagging problem (Xue and Converse, 2002; Xue, 2003; Xue and Shen, 2003), 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 (Ng and Low, 2004; Low *et al.*, 2005) used the same method in their segmenter.

Peng *et al.* (2004) first used the CRF 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.

Tseng *et al.* (2005) presented a Chinese word segmentation system submitted to the closed track of Sighan bakeoff 2005. 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.

To give a comprehensive comparison of Chinese segmentation on common test corpora, two International Chinese Word Segmentation Bake-offs were held in 2003 and 2005, and there were 12 and 23 participants respectively (Sproat and Emerson, 2003; Emerson, 2005). In all of proposed methods, character based tagging method quickly rose in two Bakeoffs as a remarkable one with state-of-the-art performance. As reported by Emerson (2005), the results of Bakeoff-2005 shows a general trend to a decrease in error rates from 3.9% to 2.8% compared to the results of Bakeoff-2003. Especially, two participants, Ng (Ng and Low, 2004; Low *et al.*, 2005) and Tseng (Tseng *et al.*, 2005) gave the best results in almost all test corpora.

Zhao (2006a; 2006b) considered both feature template selection and tag set selection, instead of feature template focused only method. They made an empirical comparison study of performance among different tag sets, and found that there is a significant performance difference as different tag sets are selected. Based on the proposed method, their system gave the state-of-the-art performance several years ago. They also proposed a criterion called "average weighted word length distribution" to choose tag set.

In this paper, we will adopt the method used in Chinese word segmentation by tagging to segment Tibetan text.

3 Tibetan Script at a Glance

3.1 Delimiters in Tibetan

The Tibetan alphabet is syllabic, like many of the alphabets of India and South East Asia. Each letter has an inherent vowel /a/. Other vowels can be indicated using a variety of diacritics which appear above or below the main letter. 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".

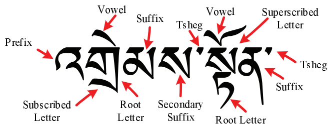

Figure 1: Structure of a Tibetan word.

ཁ་ས་མི་ཕྱུག་པོ་འདིའི་ཁང་པ་གོང་ཆེན་པོ་ཞིག་གཟིགས་སོང་།									
ཁ་ས་	མི་	ཕྱུག་པོ་	འདིའི་	ཁང་པ	གོང་ཆེན་པོ་	ཞིག་	གཟིགས་	སོང་	།
Yesterday	man	rich	this	house	expensive	an	bought	did	.
Yesterday this rich man bought an expensive house.									

Figure 2: A Tibetan sentence and its translation.

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. Figure 2 shows a Tibetan sentence.

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 dont put a "tsheg" between them and get the fused form "རྒྱལ་པོའི" (king[+genitive]), in which "པོའི" is an abbreiated syllable. When the ergative case word "ས" follows the word "ང་ཚོ" (we), it forms "ང་ཚོས" (we[+ergative]), in which "ཚོས" 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]), and the abbreviated syllable should be broken into two 1-syllable words.

4 Proposed Method

4.1 Reformulating Tibetan Word Segmentation as a Syllable Tagging Problem

Just like Chinese hanzi which is analysed by Xue (2003), many Tibetan syllables can also occur in different positions within different words. Table 1 shows how the Tibetan syllable "ཁ" (mouth) can occur in four different positions.

Table 1: The Tibetan syllable can occur in multiple word-internal positions.

Position	Example	Meaning	Tag
Word by itself	ཁ་	mouth	S
Begin	ཁ་གསབ་	supplement	B
Middle	མི་ཁ་ཤས་	someone	M
End	གྲུ་ཁ་	ferry	E

Before applying the machine-learning algorithm, we convert the segmented words in the corpus into a tagged sequence of Tibetan syllables. We tag each syllable with one of the four tags, B (Begin), M (Middle), E (End) and S (Single) depending on its position within a word.

- It is tagged B if it occurs on the left boundary of a word, and forms a word with the syllable(s) on its right side.
- It is tagged M if it occurs in the middle of a word.
- It is tagged E if it occurs on the right boundary of a word, and forms a word with the syllable(s) on its left side.
- It is tagged S if it forms a word by itself.

As presented in section 3.2, Abbreviated syllable should be broken into two parts, thus we need another two tags ES (End and Single) and SS (Single and Single):

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

Then tags for the syllable "ཁ" in the former example are shown in the last column in Table 1.

Then we can use the tags to label Tibetan word. The usage of the tags on words in different length is shown in Table 2. Using the above tags, the Tibetan sentence in (a) can be tagged as (c),

Table 2: Usage of the tag set.

Word Type	Example	Tag Sequence	Word Type	Example	Tag Sequence
1-syllable	ཁ་	S	1-syllable + AM	དམ་	SS
2-syllable	ཁ་གསང་	B-E	2-syllable + AM	གནས་པའི་	B-ES
3-syllable	མི་ཁ་ཤས་	B-M-E	3-syllable + AM	མ་བྱས་པའམ་	B-M-ES
4-syllable	རེས་མེད་ཚོལ་རྩྭ	B-M-M-E	N/A	N/A	N/A

and the segmentation result is in (d):

(a) འཚོས་སྤྱི་ཚོགས་རིང་ལུགས་ཀྱི་སྤྱི་ལ་དབང་བའི་ལམ་ལུགས་དང་མྱོས་བསྡུན་པོག་སྤྲོད་ཀྱི་རྩ་རོན་མཐའ་འཛོམས་བྱས་ཡོད།

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

(c) འ/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 མྱོས་/S བསྡུན/S པོག/S སྤྲོད/S ཀྱི/S ཙ/B རོན/E མཐའ/B འཛོམས་/E བྱས/S ཡོད/S །/S

(d) འཚོ/ ས་/ སྤྱི/ཚོགས་རིང་ལུགས་/ ཀྱི/ སྤྱི/ལ་དབང་བའི་ལམ་ ལུགས་/ དང/ མྱོས་/ བསྡུན/ པོག/ སྤྲོད/ ཀྱི/ ཙ་རོན/ མཐའ་འཛོམས་/ བྱས/ ཡོད/ །/

4.2 Tag Set Selection

In the former subsection, we presented our basic idea with a tag set B, M, E, S, ES, SS. In this subsection, we will decide how many tags to use. In Chinese word segmentation, people used different tag sets as shown in Table 3.

Table 3: Definitions of different tag sets.

Tag set	Tags	Words in tagging
2-tag	B, E	$B, BE, BEE, ...$
4-tag	B, M, E, S	$S, BE, BME, BMME, ...$
5-tag	B, B_2, M, E, S	$S, BE, BB_2E, BB_2ME, BB_2MME, ...$
6-tag	B, B_2, B_3, M, E, S	$S, BE, BB_2E, BB_2B_3E, BB_2B_3ME, BB_2B_3MME, ...$

Zhao *et al.* (2006b) proposed an effective method to choose the best tag set. They used a criterion called "average weighted word length distribution" to decide the tag set. They define L_k as follow:

$$L_k = \frac{1}{N}\sum_{i=k}^{K} i \times N_i \tag{1}$$

where L_k is the average weighted word length for all $i \geq k$; N_i is the word count with word length i, and K is the maximum word length in the corpus. N is the total word count. Then L_1 is the average weighted word length of the whole corpus.

Statistics on 8 Chinese word segmentation corpora shows that the average weighted word lengths of the 8 corpora are between 1.51 and 1.71, thus they chose a 6-tag set in their system. It can represent a 5 characters context window which can contain 3 words in average: $(1.51, 1.71) \times 3 = (4.53, 5.13)$. Experiments show that the 6-tag set is more effective when it acts in concert with an appropriate template(Zhao *et al.*, 2006a; Zhao *et al.*, 2006b).

Table 4: Statistical data on Tibetan corpus.

Word Length	L_k	Cumulative frequency	Word Length	L_k	Cumulative frequency
1	1.5294	0.5735	4	0.1432	0.9946
2	0.9559	0.9396	5	0.0307	0.9975
3	0.2238	0.9665	6	0.0161	0.9994

Base on this method, we made statistics on Tibetan corpus, and the data are shown in Table 4. In the corpus, the average weighted word length is 1.5294, which is also in the interval of [1.51, 1.71], so the 6-tag set is also suitable to Tibetan, and it can label 99.75% words in the corpus with different tags to every syllable in each word.

Table 5: Usage of the 8-tag set.

Word Type	Example	Tag Sequence
1-syllable	ཁ་	S
2-syllable	ཁ་གས་	$B-E$
3-syllable	མི་ཁ་ཤས་	$B-B_2-E$
4-syllable	རེས་མེད་རྒོལ་རྩྭ་	$B-B_2-B_3-E$
5-syllable	གཙོ་ཁྲིད་ཀྱི་ཚོ་གཤས་	$B-B_2-B_3-M-E$
6-syllable	སྤྱི་ལ་དབང་བའི་ལམ་ལུགས་	$B-B_2-B_3-M-M-E$
1-syllable + AM	དས་	SS
2-syllable + AM	གཤས་པའི་	$B-ES$
3-syllable + AM	མ་བྱས་པའམ་	$B-B_2-ES$
4-syllable + AM	འཛམ་གླིང་གསུམ་པའི་	$B-B_2-B_3-ES$

As analysed in former subsection, we need another 2 tags ES and SS for abbreviated syllables. Actually, it forms an 8-tag set: B, B_2, B_3, M, E, S, ES, SS. The usage of the 8-tag set on words in different length is shown in Table 5. With the 8-tag set, the Tibetan sentence mentioned in the former subsection is tagged as (e):

(e) ར་/B ཆོས་/ES སྤྱི་/B ཆོགས་/B_2 རིང་/B_3 ལུགས་/E ཀྱི་/S སྤྱི་/B ལ་/B_2 དགང་/B_3 བའི་/M ལམ་/M ལུགས་/E དང་/S རྩོལ་/S བསྡུན་/S ཐོག་/S སྒྲིལ་/S ཀྱི་/S ཪ་/B དོན་/E མཐའ་/B འཚེངས་/E བྱས་/S ཡོད་/S །/S

4.3 CRF Tagger for Tibetan

Maximum Entropy (ME) tagger was used in early character-based tagging for Chinese word segmentation (Xue and Converse, 2002; Xue, 2003; Xue and Shen, 2003; Ng and Low, 2004; Low *et al.*, 2005), In recent years, more and more people choose linear-chain CRF as the learning model in their studies (Peng *et al.*, 2004; Tseng *et al.*, 2005; Zhao *et al.*, 2006a; Zhao *et al.*, 2006b).

CRF is firstly introduced into language processing by Lafferty (2001). Peng *et al.* (2004) first used this framework for Chinese word segmentation by treating it as a binary decision task, such that each Chinese character is labeled either as the beginning of a word or not.

The probability assigned to a label sequence for a syllable sequence by a CRF is:

$$p_\lambda(Y|W) = \frac{1}{Z(W)} exp\left(\sum_{t \in T}\sum_{k} \lambda_k f_k(y_{t-1}, y_t, W, t)\right) \quad (2)$$

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

4.4 Feature Templates

The probability model and corresponding feature function is defined over the set $H \times T$, where H is the set of possible contexts (or any predefined condition) and T is the set of possible tags. Generally, a feature function can be defined as follow:

$$f(h,t) = \begin{cases} 1, & if \ \ h = h_i \ \ and \ \ t = t_j \\ 0, & otherwise \end{cases} \quad (3)$$

where $h_i \in H$ and $t_j \in T$.

Table 6: The two template sets TMPT-6 and TMPT-10.

Template Set	Type	Feature	Function
TMPT-6	Unigram	$C_n,\ n = -1, 0, 1$	The previous, current and next syllable
	Bigram	$C_nC_{n+1},\ n = -1, 0$	The previous (next) syllable and current syllable
		$C_{-1}C_1$	The previous syllable and next syllable
TMPT-10	Unigram	$C_n,\ n = -1, 0, 1$	The previous, current and next syllable
		C_{-2}	*The syllable before the previous syllable*
		C_2	*The syllable after the next syllable*
	Bigram	$C_nC_{n+1},\ n = -1, 0$	The previous (next) syllable and current syllable
		$C_{-1}C_1$	The previous syllable and next syllable
		C_1C_2	*The next two syllables*
		$C_{-2}C_{-1}$	*The previous two syllables*

In Chinese word segmentation by character tagging, their are two often used template sets (Xue and Converse, 2002; Xue, 2003; Xue and Shen, 2003; Ng and Low, 2004; Low *et al.*, 2005; Peng *et al.*, 2004; Tseng *et al.*, 2005; Zhao *et al.*, 2006a; Zhao *et al.*, 2006b), namely TMPT-6 and TMPT-10, which are shown in Table 6. In this work, we use and compare the two template sets to see their effect on the segmentation results. Note that there is a slight difference, the character "C" in the templates denotes a Tibetan *syllable* rather than a *character* like in Chinese word segmentation.

5 Experiments and Results

In this paper, we evaluate the approach by the Precision (P), Recall (R) and F-score (F1). Their definitions are the same as which are used in Chinese word segmentation. We used the CRF++ toolkit 0.54 from sourceforge `http://crfpp.sourceforge.net`.

5.1 Corpora

At present, there is no public available Tibetan segmentation corpus. In out experiments, the training set are machine-generated, and the test set is manually segmented.

Test Set: We randomly selected 1000 out of more than 230,000 Tibetan sentences from our Tibetan text corpus which consists of several political books and many news stories, legal texts and articles, then manually segmented them to form the test set.

Combining methods used by Chen (2003a, 2003b), Caizhijie (2009a, 2009b) with Sun (2009, 2010), we implemented a Tibetan word segmenter (SegT). Experiment on the test set shows that the precision, recall and the F-score are 96.99%, 96.91% and 96.94% respectively. SegT is used to generate the training set.

Corpus A: We get 1990 web pages in the domain of Tibetan arts, news, notice, religious, traditional culture, and history from the website `http://tb.tibet.cn`, and get 64,419 Tibetan sentences after preprocessing, which takes 9,739 K bytes in UTF-8 encoding plain text.

Corpus B: We get another 3292 web pages in the domain of news from another website `http://tb.chinatibetnews.com`. The preprocessing is similar to Training set A. and we get 67,484 Tibetan sentences which takes 18,372 K bytes in UTF-8 encoding plain text.

With SegT, the sentences are segmented into words to form the training set. We haven't made any manual corrections. As a machine-generated corpus, it includes about 3% words which are not correctly segmented. The two training set will be used to inspect the effect of training set scale change on the segmentation performance.

5.2 Comparison of TMPT-6 and TMPT-10

While training the model with the command crf_learn, we tried different values of the two parameters c and f. The former (c) trades the balance between overfitting and underfitting. The results will significantly be influenced by this parameter. The latter (f) sets the cut-off threshold for the features. CRF++ uses the features that occurs no less than f times in the given training data.

We set $f = 2$ and change the value of c, make the training and test on TMPT-6 and TMPT-10 respectively. We assign c with the values from 1.0 to 10.0 with a step length of 1.0. The data are shown in Table 7 and Figure 3. The F-score is the highest one at $c = 4.0$ on both template sets. Figure 3 shows that the F-scores are significantly higher on TMPT-6 than on TMPT-10, which is similar to the result in Chinese word segmentation.

Table 7: F-scores with different values of c on TMPT-6 and TMPT-10.

	TMPT-6			TMPT-10		
c	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
1.0	92.44	93.67	93.05	92.26	93.37	92.81
2.0	92.57	93.71	93.13	92.39	93.33	92.86
3.0	92.55	93.72	93.13	92.42	93.49	92.95
4.0	92.58	93.70	**93.14**	92.46	93.50	**92.98**
5.0	92.54	93.66	93.10	92.40	93.45	92.93
6.0	92.57	93.65	93.10	92.38	93.40	92.88
7.0	92.50	93.47	92.98	92.42	93.40	92.90
8.0	92.57	93.63	93.10	93.34	93.34	92.84
9.0	92.51	93.53	93.02	92.39	93.39	92.89
10.0	92.45	93.48	92.96	92.39	93.41	92.90

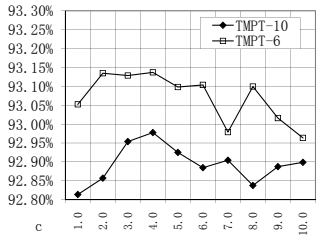

Figure 3: F-scores on TMPT-6 and TMPT-10.

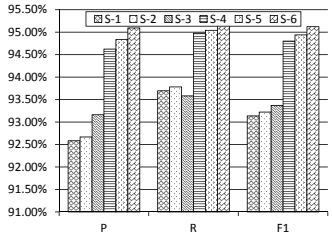

Figure 4: Performance with different settings.

5.3 Effect of Training Set Scale

As we get a better performance on TMPT-6 than TMPT-10, we make further experiments on it. We make the training and test with different f, c and training sets. The data are shown in Table 8 and

Figure 4. With $f = 2$, we get the highest F-score (93.22%) at $c = 4.9$ on Corpus A (S-2). With the same f and c, we train the segmenter with Corpus B, and the performance improves slightly (S-3). Then we use both Corpus A and Corpus B in the training, and get significant performance improvement (S-4). The best performance is achieved at $f = 1$ and $c = 4.0$.

Table 8: Performance of CRF-SegT with different settings.

Setting	Training Set	f	c	P(%)	R(%)	F1(%)
S-1	Corpus A	2	4.0	92.58	93.70	93.14
S-2	Corpus A	2	4.9	92.67	93.78	93.22
S-3	Corpus B	2	4.9	93.16	93.58	93.37
S-4	Corpus A+B	2	4.9	94.63	94.98	94.80
S-5	Corpus A+B	1	4.9	94.84	95.04	94.94
S-6	Corpus A+B	1	10.6	**95.09**	**95.15**	**95.12**

5.4 Comparison with other segmenters

We collected the test data of several Tibetan segmenters from related papers, and they are shown in Table 9. Chen and Sun evaluate their segmenters only by accuracy (Recall actually), so we can compare them only by recall. Compared with them, CRF-SegT outperforms only two of these sgementers. However, these segmenters are evaluated on different test sets, so it's not much meaningful to directly compare the Recalls here. All the segmenters are using maximum-matching method based on dictionary except our CRF-SegT. All of them have inherent weakness on segmentation disambiguation and Out-Of-Vocabulary word identification. Our CRF-SegT doesn't have those weakness. Considering these factors and the low quality of the training corpus, our approach is effective and acceptable though it doesn't have the highest F-score compared with other Tibetan segmenters.

Table 9: Comparison of different Tibetan/Dzongkha segmenters.

	Sentence Count	Word Count	P(%)	R(%)	F1(%)
Chen 2003a	N/A	N/A	N/A	≥ 0.96	N/A
Chen 2003b	500	5890	N/A	97.21	N/A
Sun 2009	435	4067	N/A	87.02	N/A
Norbu 2010	N/A	714	N/A	91.50	N/A
Our/SegT	1000	13977	96.99	96.91	96.94
Our/CRF-SegT	1000	13977	94.09	95.15	95.12

As collected by Zhao et al. (2006b), the F-scores of a state-of-art Chinese segmenter on most of 8 corpora are up to 95.3%-97.4%. Compared with them, CRF-SegT achieves a lower F-score. It mainly results from the low quality of the training corpus. As the F-score of SegT and CRF-SegT are 0.9694 and 0.9512 respectively, it means the latter $0.9512/0.9694 = 0.9812$ fits the former. Thus, the CRF-SegT fits the training corpus at a percentage of 98.12%. It's already greater than the upper bound of the interval 95.3%-97.4%, which shows that the result is acceptable.

6 Conclusion

In this paper, we reformulate Tibetan word segmentation as a syllable tagging problem and implemented a Tibetan word segmenter (CRF-SegT) which is trained by the CRF model on a corpus generated by another Tibetan word segmenter (SegT). Two feature template sets namely TMPT-6 and TMPT-10 are used and compared, and the result shows that the former is better. Experiments also show that larger training set improves the performance significantly. Trained on a set

of 131,903 sentences, the segmenter achieves an F-score of 95.12% on the test set of 1,000 sentences. At present, almost all Tibetan segmenters are using maximum matching method based on dictionary except the proposed one. All of them have inherent weakness on segmentation disambiguation and Out-Of-Vocabulary word identification. The proposed approach doesn't have those weakness. What is more, our segmenter is test on a larger test set. Considering these factors, our approach is effective and acceptable.

References

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

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

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

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

Yuzhong Chen, Baoli Li, Shiwen Yu and Lancuoji. 2003b. An Automatic Tibetan Segmentation Scheme Based on Case Auxiliary Words and Continuous Features, *Applied Linguistics*, 2003(01): 75-82.

Chungku Chungku, Jurmey Rabgay and Gertrud Faa. 2010. Building NLP Resources for Dzongkha: A Tagset and a Tagged Corpus. *Proceedings of the 8th Workshop on Asian Language Resources*. pp.103-110. Beijing, China.

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

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pp.123-133. Jeju Island, Korea.

Jianfeng Gao, Mu Li, Andi Wu and Chang-Ning Huang. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computational Linguistics*, 31(4): 531-574.

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

John Lafferty, Andrew McCallum and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pp.282-289.

Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pp.161-164. Jeju Island, Korea.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based. *Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing*, pp.277-284.

Sithiar Norbu, Pema Choejey, Tenzin Dendup, Sarmad Hussain and Ahmed Mauz. 2010. Dzongkha Word Segmentation. *Proceedings of the 8th Workshop on Asian Language Resources*. pp.95-102. Beijing, China.

Fuchun Peng, Fangfang Feng and Andrew McCallum. 2004. Chinese Segmentation and New Word Detection Using Conditional Random Fields. *Proceedings of the 20th International Conference on Computational Linguistics*, pp.562-568. Geneva, Switzerland.

Kunyu Qi. 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.

Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pp.133-143. Sapporo, Japan.

Yuan Sun, Luosangqiangba, Rui Yang and Xiaobing Zhao. 2009. Design of a Tibetan Automatic Segmentation Scheme, *the 12th Symposium on Chinese Minority Information Processing*.

Yuan Sun, Xiaodong Yan, Xiaobing Zhao and Guosheng Yang. 2010. A resolution of overlapping ambiguity in Tibetan word segmentation. *Proceedings of the 3rd International Conference on Computer Science and Information Technology*, pp.222-225.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pp.168-171. Jeju Island, Korea.

Hanna M. Wallach. 2004. Conditional Random Fields: An Introduction. *Technical Report MS-CIS-04-21*. Department of Computer and Information Science, University of Pennsylvania,

Kun Wang, Chengqing Zong and Keh-Yih Su. 2009. Which is more suitable for Chinese word segmentation, the generative model or the discriminative one? *Proceedings of the 23th Pacific Asia Conference on Language, Information and Computation*, pp.827-834, Hong Kong, China.

Nianwen Xue and Susan P. Converse. 2002. Combining classifiers for Chinese word segmentation. *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, pp.63-70. Taipei, Taiwan.

Nianwen Xue and Libin Shen. 2003. Chinese Word Segmentation as LMR Tagging. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL03*, pp.176-179. Sapporo, Japan.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*. 8(1): 29-48.

Ruiqiang Zhang, Genichiro Kikui and Eiichiro Sumita. 2006. Subword-based Tagging for Confidence-dependent Chinese Word Segmentation. *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, pp.961-968, Sydney, Australia.

Yue Zhang and Stephen Clark. 2007. Chinese Segmentation with a Word-Based Perceptron Algorithm. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp.840-847, Prague, Czech Republic.

Hai Zhao, Chang-Ning Huang and Mu Li. 2006a. An Improved Chinese Word Segmentation System with Conditional Random Field. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pp.108-117. Sidney,Australia.

Hai Zhao, Changning Huang, Mu Li and Baoliang Lu. 2006b. Effective tag set selection in Chinese word segmentation via conditional random field modeling. *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pp.87-94 Wuhan, China.

Zhaxijia, Dolha, Losanglangjie and 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*.