Tibetan Word Segmentation Based on Word-position Tagging

Caijun Kang
Humanities and Communications College
Shanghai Normal University
Shanghai, China
e-mail: kang.caijun@gmail.com

Di Jiang
Lab. of Phonetics & Computational Linguistics,
Institute of Ethnology & Anthropology
Chinese Academy of Social Sciences
Beijing, China
e-mail: jiangdi@cass.org.cn

Congjun Long
Lab. of Phonetics & Computational Linguistics,
Institute of Ethnology & Anthropology
Chinese Academy of Social Sciences
Beijing, China
e-mail: longcj@cass.org.cn

***Abstract*—The best advantage of Tibetan word segmentation based on word-position is to reduce segmentation errors for unknown words. In this article authors upgrade usual 4-tag set to 6-tag set to fit in with the features of Tibetan characters, using CRF as tagging model to train and test corpus data, then building post processing modules to revise the result data. The experimental result shows that this method achieves a good performance and deserves further study, including expanding the corpus and optimizing the tag set and feature templates.**

Keywords-Tibetan; word-position; CRF; tagging model

I. Introduction

The mainstream methods of Tibetan word segmentation are based on dictionary matching, which mainly depend on dictionary, including maximum matching, reverse maximum matching, bi-directional maximum matching, optimal matching, etc. Recently, the segmentation theory based on statistics is accepted by researchers. Transplanted from Chinese word segmentation system Segtag, Tibetan word segmentation system Yangjin developed by Shi xiaodong and Lu Yajun achieved good accuracy via Hidden Markov Model [1]. Jiang Tao firstly used the 4-tag set in Tibetan word segmentation based on word-position [2]. And Liu Huidan used CRF with 8-tag set and claimed very good performance [3].

We use 6-tag set in Tibetan word-position tagging via conditional random field modeling. Due to lacking public Tibetan corpus, we extracted about 1,000,000-syllable corpus from the parallel corpus constructed by the Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences.

II. Word segmentation based on word-position tagging

Differed from dictionary-matching for word segmentations, the procedure of word-position based segmentation was regarded as a tagging task, such that each syllable is labeled as a position of a word. Once we get the probability assigned to a tag sequence for a particular word, we can infer the segmentation result.

The most advantage of word-position based segmentation is that it handles unknown words well. In word-position based tagging, unknown words are equally treated in the same process as known words. This feature simplifies the design of segmentation system. In the processing, A probability model is built by the word position features learned from all syllables according to corresponding feature function and predefined feature templates. The probability model is used to judge the correlation of adjacent syllables in the sequence of unsegmented syllables and output a sequence of tagged syllables , which demonstrates the word segmentation results.

The usual tag sets used by Chinese word segmentation are 4-tag set (B, M, E, S) [4] and 6-tag set (B1, B2, B3, M, E, S) [5,6]. In term of the special Tibetan bounded-variant forms, we expand the 4-tag set to 6-tag set: B (for word beginning), M (for word-medial), E (for word ending), E' (for word ending with bounded-variant form), S (for single word) and S' (for single word with bounded-variant form). Different from linear 6-tag set used by Chinese tagging method, our 6-tag set emphasizes the features of bounded-variant forms and more suits Tibetan. For example:

978-0-7695-5063-3/13 $26.00 © 2013 IEEE
DOI 10.1109/IALP.2013.74

མཚན་མོར་དོམ་དེས་ཕ་རིའི་ཐང་དཀྱིལ་ན་འོད་ཆེ་བ་ཞིག་ཡོད་པ་མཐོང་བས།

The result of word-position tagging:

མཚན་/Bམོར་/E’དོམ་/Sདེས་/Sཕ་/Bརིའི་/E’ཐང་/Bདཀྱིལ་/Eན་/Sའོད་/Sཆེ་/Bབ་/Eཞིག/Sཡོད་/Sཔ་/Sམཐོང་/Sབས་/S’།/S

The final word segmentation result:

མཚན་མོར་/དོམ་/དེས་/ཕ་རིའི་/ཐང་དཀྱིལ་/ན་/འོད་/ཆེ་བ་/ཞིག་/ཡོད་/པ་/མཐོང་/བས་/།

There are numerals and all kinds of punctuation marks in actual Tibetan texts, so the objects of word-position tagging are not only Tibetan syllables, but also numerals, Latin letters, punctuation marks and other non-syllable symbols. We label all non-syllable symbols as S.

III. Conditional Random Field

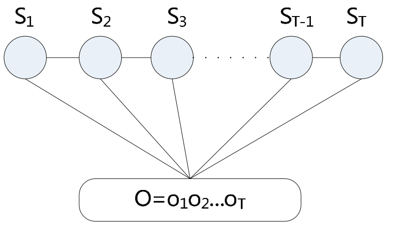

Figure 1. first order chain structure

In word-position based tagging methods, Conditional Random Field (CRF) model is one of the most effective models. Though originating from maximum entropy model, CRF minimizes the data sparseness problem. On the other hand, independence assumption is not necessary for CRF so that it is superior to HMM. CRF is an undirected graph model with the first order chain structure (shown as Figure.1) which is used to tag sequence data.

If $O=\{o_1,o_2,...,o_T\}$ is a sequence of unsegmented syllables , and $S=\{s_1,s_2,...,s_T\}$ is the tag sequence for the sentence whose tags are associated with a certain word position such as word beginning B or word ending with bounded-variant form E’. The probability assigned to a tag sequence for a particular sequence of syllables by a CRF with parameter $\Lambda=\{\lambda_1,\lambda_2,...,\lambda_K\}$ is given by the equation below:

$$p(S\,|\,O)=\frac{1}{Z(o)}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k\left(s_{t-1},s_t,o,t\right)\right)$$

where $Z(o)$ is a normalization term, $f_k\left(s_{t-1},s_t,o,t\right)$ is a feature function, and t indexes into syllables in the tag sequence.

.

IV. Building feature templates

In the word-position based segmentation, feature templates are used to define context dependencies. Usual feature templates are shown in Table 1. U01，U02 in Table 1 stand for the indexes of features, %x[0 , 0] stands for the Unigram feature of the current syllable, and %x[-1, 0]/%x[1, 0] stand for the Bigram feature of contextual syllables prior and after the current syllable, and so on.

For example:

ང་/Bཚོས་/E'མེས་/Bརྒྱལ་/Eལ་/Sདགའ་/Bཞེན་/Mབྱེད་/Eཀྱི་/Bཡོད/E།/S

Suppose that we use feature templates (U02:%x[0,0]，U07:%x[0,0]/%x[1,0]) to process the sentence above, we observe the feature (U02: (“ང་”, B), U07: (“ང་ཚོས་”, B)) from the first Tibetan syllable. Various features contains different information, for example: U06: (“ང་ཚོས་”, E) contain much more context information than U06: (“ཡོད།”, S) because “།” is the ending of a sentence and always tagged as S and has weak correlation with the word ahead.

To analyze the performance of different feature templates and select optimal ones for Tibetan word segmentation, we randomly pick out a Tibetan corpus with 23818 syllables to test the performance of each feature template respectively, and the result is shown in Table 2.

From the result, we find that the performance of Unigram feature templates is poor. The F1 of the most effective feature template U02:%x[0,0] is only 0.7626. Recall rate of Other Unigram ones are below 70%, some even less than 60%. By contrast, the performance of Bigram feature templates is much better. The most effective ones are U06:%x[-1,0]/%x[0,0], U07:%x[0,0]/%x[1,0], U09:%x[-1,0]/%x[1,0], all with F1 value near 0.94.

Through the statistics for Tibetan corpus, we find that the average weighted word length of Tibetan words is about 1.7, which is close to the value of the upper bound of Chinese average weighted word length distribution. The tag set we designed is more emphasis on the single contextual syllable and need more complicated feature information, so Bigram feature templates perform better than Unigram ones.

How to overall select tag set and feature templates to balance the efficiency and performance is our next work.

Table 1. Usual feature template set.

Type	Feature templates
Unigram	U00:%x[-2,0] U01:%x[-1,0] U02:%x[0,0] U03:%x[1,0] U04:%x[2,0]
Bigram	U05:%x[-2,0]/%x[-1,0] U06:%x[-1,0]/%x[0,0] U07:%x[0,0]/%x[1,0] U08:%x[1,0]/%x[2,0] U09:%x[-1,0]/%x[1,0]

Table 2. Experimental results of feature templates.

Feature template	Precision (%)	Recall (%)	F1
U00:%x[-2,0]	59.65	61.20	0.6042
U01:%x[-1,0]	66.65	67.72	0.6718
U02:%x[0,0]	75.97	76.56	0.7626
U03:%x[1,0]	66.12	67.61	0.6686
U04:%x[2,0]	57.88	59.30	0.5858
U05:%x[-2,0]/%x[-1,0]	84.40	84.87	0.8463
U06:%x[-1,0]/%x[0,0]	93.80	93.73	0.9376
U07:%x[0,0]/%x[1,0]	93.58	93.53	0.9355
U08:%x[1,0]/%x[2,0]	83.04	83.89	0.8346
U09:%x[-1,0]/%x[1,0]	94.64	94.58	0.9461
All feature templates	99.63	99.62	0.9963

V. EXPERIMENTAL RESULT AND POST-PROCESSING

Although the Tibetan language processing has been carried out for years, there is no public Tibetan corpus. Our training corpus is extracted from the parallel corpus constructed by the Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences. The training corpus contains about 1,000,000 Tibetan syllables. In order to ensure the accuracy of our training corpus, all the corpus data are artificially proofread. The experiment is divided into two parts: closed test and open test. The random closed test corpus includes a text of over 40,000 syllables, and open test corpus is formed from the Tibetan version of “2011 Government Work Report”, including a 20,000-syllable. The training and testing procedure adopt the open source software kit CRF++. And the experimental result is shown in Table 3.

There is neither measure standard and public Tibetan corpus, nor public Tibetan segmentation system. So we cannot compare our experimental result with that of other systems. Some scholars claimed that they achieved 99.83% accuracy rate in their Tibetan word segmentation system. That performance is close to the result of our closed. But our open test result cannot reach such performance.

From our experimental result, we find that CRF classifier can identify rare words and unknown words well with sufficient contextual information. And errors can be divided into two kinds: first kind of errors focus on bounded-variant form S’, which have unfixed position and can appear in different word position with similar probability; second kind of errors appear in the combinations of numerals and other punctuations, which are in great quantity but repeated rarely, lead to the lack of contextual information.

We can see the limitations of CRF classifier from the analysis above. To achieve better performance, we need post process to resolve the two kinds of errors. To handle with errors about bounded-variant form, we summarize the results of previous studies [7-10], and build a rule table. When the segmented result is in the low confidence interval, we match the result without rule table and take priority of the rule matching result. To handle with errors related with numerals, we build a digital component identification module with a digital component dictionary. The segmented result will be scanned by the module, and scanned consecutive digital components will be combined into the correct forms according to certain rules.

Table 3. Comparative test results.

	Syllable amount	precision (%)	Recall (%)	F1
Closed test	43861	98.20	98.34	0.9827
Open test	23818	91.27	90.85	0.9106

Open test is performed again after we add the post processing module, and the comparative result is shown in Table 4. From the result, we can see that the recall rate is improved obviously, while the precision rate is improved slightly. This phenomenon means the post processing module depending on rules cannot handle unknown words very well.

Table 4. Comparative open test results of original segmentation system & segmentation system with post-processing module.

	precision (%)	Recall (%)	F1
Original system	91.27	90.85	0.9106
System with post-processing	92.08	95.02	0.9352

VI. CONCLUSION

With the support of sufficient contextual features，we achieve high accuracy rate in the Tibetan word segmentation with the word-position tagging method based on statistics. This is the first time that word-position tagging method is tested in a large-scale actual corpus. It shows the feasibility and high performance of the method, which has the value of further study.

In the future study, we will expand the training corpus and improve the accuracy rate of low confidence interval result via post processing based on not only rules, but also statistics. In the meantime, we will overall optimize the tag set and feature templates to use features more effectively and try to reach the balance of performance and efficiency.

ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China.

REFERENCES

[1] Shi Xiaodong, Lu Yajun, A Tibetan Segmentation System-Yangjin, *Journal Of Chinese Information Processing*, 2011, 4.

[2] Jing Tao, Tibetan Word Segmentation System Based on Conditional Random Fields, Software Engineering and Service Science (ICSESS).2011:446-448.

[3] Liu Huidan, Nuo Minghua, Ma Longlong, Wu Jian And He Yeping, 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):168-177.

[4] N. Xue, Chinese Word Segmentation as Character Tagging[C]// International Journal of Computational Linguistics and Chinese Language Processing. 2003.

[5] Huang Changning, Zhao Hai, Chinese Word Segmentation: A Decade Review, Journal of Chinese Information Processing, 2007, 5: 8-19.

[6] Huang Changning, Zhao Hai, Character-Based Tagging:A New Method for Chinese Word Segmentation, Frontiers of Chinese Information Processing- 25th Anniversary Conference Proceedings Of Chinese Information Society Of China, Beijing: Tsinghua University Press: 2006: 53-63.

[7] Jiangdi, Kang Caijun, The Methods of Lemmatization of Bound Case Forms In Modern Tibetan[C]// 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering. IEEE Press, ISBN: 0-7803-7902-0.

[8] Cai Zhijie, The Recognition of Abbreviated Words in Tibetan Automatic Word Segmentation Systems. Journal of Chinese Information Processing, 2009.23(1).

[9] Basangjiebu, Yangmaozhuoma, Ouzhu, The Research on Tibetan Abbreviated Cases and Reduction of Tibetan Words in the Automatic Word Segmentation System. Tibet's Science and Technology. 2012,2:73-75,79.

[10] Kang Caijun, Long Congjun. Jiang Di, the segmentation of Tibetan abbreviated forms based on word position tagging. Computer Engineering and Applications, 2013.