# Tibetan Unknown Word Identification from News Corpora for Supporting Lexicon-based Tibetan Word Segmentation

**Minghua Nuo**[1]
minghua@iscas.ac.cn

**Huidan Liu**[1]
huidan@iscas.ac.cn

**Congjun Long**[1,2]
congjun@nfs.iscas.ac.cn

**Jian Wu**[1]
wujian@iscas.ac.cn

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

**Abstract**

In Tibetan, as words are written consecutively without delimiters, finding unknown word boundary is difficult. This paper presents a hybrid approach for Tibetan unknown word identification for offline corpus processing. Firstly, Tibetan named entity is preprocessed based on natural annotation. Secondly, other Tibetan unknown words are extracted from word segmentation fragments using MTC, the combination of a statistical metric and a set of context sensitive rules. In addition, the preliminary experimental results on Tibetan News Corpus are reported. Lexicon-based Tibetan word segmentation system SegT with proposed unknown word extension mechanism is indeed helpful to promote the performance of Tibetan word segmentation. It increases the F-score of Tibetan word segmentation by 4.15% on random-selected test set. Our unknown word identification scheme can find new words in any length and in any field.

## 1 Introduction

Tibetan is a phonetic writing script; it is syllabic, like many of the alphabets of India and South East Asia. Tibetan sentences are strings of syllables with no delimiters to mark word boundaries. Therefore the initial step for Tibetan processing is word segmentation. However, occurrences of unknown words, which are not listed in the dictionary, degraded significantly the performances of most word segmentation methods.

Currently, the lexicon-based Tibetan word segmentation scheme is widely adopted. In general, any lexicon is limited and unable to cover all the words in real texts. According to our statistics on a 326,062,576-bytes news corpus from the website *Tibet Daily*, there are about 2.89% unknown words. Therefore, unknown word identification (UWI) became a key technology for Tibetan segmentation.

The rest of this paper is organized as follows. In Section 2 we recall related work on UWI methods. Semi-automatic Tibetan UWI method is provided in Section 3. Section 4 gives the description of experimental results for evaluation, and Section 5 offers concluding remarks.

## 2 Related Work

For unknown words with more regular morphological structures, such as personal names, morphological rules are commonly used for improving the performance by restricting the structures of extracted words (Chen et al. 1994, Sun et al. 1995, Lin et al. 1993, Ma & Chen 2003). However, it is not possible to list morphological rules for all kinds of unknown words, especially those words with very irregular structures. Therefore, statistical approaches usually play major roles on irregular UWI in most previous work (Sproat & Shih 1990, Chiang et al. 1992, Tung & Lee 1995, Palmer 1997, Chang et al. 1997, Sun et al. 1998, Ge et al. 1999).

Many statistical metrics have been proposed, including point-wise mutual information (MI) (Church et al., 1991), mean and variance, hypothesis testing (t-test, chi-square test, etc.), log-likelihood ratio (LR) (Dunning, 1993), statistic language model (Tomokiyo et al., 2003), context-entropy (on each side) and frequency ratio against background corpus (Luo & Song 2004), DCF (Hong et al., 2009), and so on. Point-wise MI is often used to find interesting bigrams (collocations). However, MI is actually better to think of it as a measure of independence than of dependence (Manning et al., 1999). LR is one of the most stable methods for automatic term extraction so far, and more appropriate for sparse data than other metrics. However, LR is still biased to two frequent words that are rarely adjacent, such as the pair (the, the) (Pantel et al., 2001). On the other aspect, MI and LR metrics



*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 451–457, Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

are difficult to extend to extract multi-word terms.

There are also many hybrid methods combined statistical metrics with linguistic knowledge and machine Learning algorithms, such as Part-of-Speech filters (Smadja, 1994; Asanee, 1997), roles tagging based (Zhang et al., 2002), syntactic discriminators (Chen & Ma 2002), max-margin Markov networks (Qiao and Sun, 2010; Li and Chang, 2010), Unsupervised Learning Strategy (Sun et al., 2004), Latent Discriminative Model(Sun et al., 2011), boosting-based ensemble learning (TeCho et al., 2012). But POS filters, roles tagging, machine learning algorithms does not work for Tibetan UWI. So far, there is no Tibetan POS tagger and Tibetan parser. We have built large scale Tibetan text resources recently, and we are tagging Part-Of-Speech and labeling role right now, these corpora can form training set in the near future.

Previous research and work in Tibetan word segmentation have made great progresses. However, cases with unknown words are not satisfactory. In recent years, researchers mainly use maximum-matching method accompanying with some grammar rules (Chen et al. 2003a, Chen et al. 2003b, Cai 2009a, Cai 2009b, Qi 2006, Dolha 2007, Zha 2007, Tashi 2009) to segment Tibetan text. Liu et al. (2012) designed and implemented a Tibetan word segmentation system named "SegT" which is lexicon-based practical system with a constant lexicon. However, it has the difficulty of identifying unknown words in newspaper articles and web documents which are highly changeable texts with time.

The research on Tibetan UWI is, however, still at its initial stage. There is no public report of performance of Tibetan new word or unknown word identification. This paper introduces Tibetan UWI work which is in progress.

## 3 Tibetan Unknown Word Identification from News Corpus

Generally, Tibetan location name and organization names are formed from a shorter word or proper noun adding a morpheme, river(ཆུ་), lake(མཚོ), beach(ཐང་), gorge(འགག་), ministry(ལྷན), bureau(ཇོའུ), association(ཁང་), company(ཀུང་སི), province(ཞིང་ཆེན), city(གྲོང་ཁྱེར), county(རྫོང་) etc.; some are also followed by modifiers, such as postposition, size, color, shape. We also observe that often, these morphemes are segmented separately during the first-time segmentation process. "Natural annotation" in our news articles also indicates the occurrence of unknown words. This section simply introduces Tibetan script first and then aims to detail the two key procedures in Tibetan UWI from Tibetan web resources, that is, detect unknown words based on natural annotation and based on context sensitive rules.

### 3.1 Characteristics of Tibetan Script

The Tibetan alphabet is syllabic; a syllable contains one or up to seven character(s). Syllables are separated by a marker known as intersyllabic marks (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". Consonant clusters are written with special conjunct letters. Tibetan texts consists of a string of syllables without any blanks to mark word boundaries except for punctuation'།', called shad, at the end of each sentence, and '་', called tsheg, within syllables. 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

Tibetan sentence consists of one or more words, phrases or multi-word units. Another marker known as "shad" indicates the sentence boundary, which looks like a vertical pipe. Figure 2 shows a Tibetan sentence. It is segmented in line 2 and word by word translation is given in line 3.

| ཁ་སང་མི་སྤྱག་པོ་འདིའི་ཁང་པ་གོང་ཆེན་པོ་ཉོག་གཟིགས་སོང་། | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ཁ་སང་ | མི་ | སྤྱག་པོ་ | འདིའི་ | ཁང་པ | གོང་ཆེན་པོ་ | ཉོག་ | 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

### 3.2 Natural Annotation based Identification

Tibetan unknown word covers both named entity and emerging new words in Tibetan web corpus. Special attention is paid to those noticeable named entities in order to suggest strong word candidates. "naturally annotated" means different type of annotations on varieties of Web resources which are "unconsciously handcrafted" by Web users for their own purposes, but can be used by

computational linguists in a conscious and systematic way for various tasks of natural language processing, for examples, punctuation marks in Tibetan can benefit word boundaries identification, social tags in social media can benefit keyword extraction, "categories" given in News Corpus can benefit text categorization.

"Space", "punctuation" and "Tibetan auxiliary words" always appear next to a word. Hyperlink in web text is a useful explicit natural annotation too. In addition, <head> tag of html pages including meta data as keywords, author, source, description; these are also quite useful natural annotation for UWI. Meanwhile, in our Tibetan News Corpus, English and Chinese in brackets give the hints for their corresponding Tibetan translation words. Sentences including this kind of annotation are as follow.

- ཇ་༢༠༠༠འཇའ་ཚོ་མཚོ(janet gyatso)དྲན་པ།
- འཇར་མེ་ནི(germany)རྒྱལ་ཁབ་ཀྱི་སྲུན་ཆེན(münchen)གོང་ཁྲེར་ཀྱི(schloss hohenkammer)མཐོ་སློབ་ཀུང་ཚོགས།
- ཨྱི་ཡོ་ཀཽཽཡོར་ཨིར་ནི་མི་ཏི་མི་ཏེ་ན་ཀེལ་ནར་(Ernst Steinkellner)ཀྱི་ཨོ་འུ་མི་ཁྲི་ཡ(Austria)རྒྱལ་ཁབ་ཀྱི་རྒྱ་རྡོ(Graz)གོང་ཁྲེར་ཆེན་མོའི(Schloss Segau)མཐོ་སློབ་ཀུ་གོ་སྒྲིག་མཛད།
- ཨྱི་ཡོ་ཀཽཽཤོར་སྐྱང་ཚོགས་པ་ཨེ་ལི་ཡོ་ཏི་མི་པར་ལིང(Elliot Sperling)གིས་ཨ་རི་ཟོ་གྲོང་ཁྲེར་བྷ་ལུ་མིང་ཏོན(Bloomington)ཀྱི་ཨིན་དི་ཡ་ན/དི་མཐོ་སློབ(Indiana University)ཀུ་བཙུགས།
- ཕ་མེས་ཨེར（帕米尔）མཐོ་སྨད་ནས་ཤར་ཡོགས་ཀྱི་ཆིན་ལེན་རི་རྒྱུད（祁连山脉）བར་འཕེད་ཐིག་ཀུཌ་31ཡག་ཡོད་ལ།

These brackets in news texts point out the right boundary of lots of location name and organization names. We confirm left boundaries relying on pre-established transliteration table. Thus following named entities such as རྡེ་རྡོ(Graz), བློ་ལུ་མིང་ཏོན(Bloomington), ཨིན་དི་ཡ་ནའི་མཐོ་སློབ (Indiana University), ཕ་མེས་ཨེར(帕米尔) can be extracted from examples given above.

### 3.3 Contextual Rule based Identification

We will use a hybrid method MTC, that is, combination of statistical metric and context sensitive rules, to recognize the boundary of an unknown word. It is applied to segmented texts.

Beforehand, we analyse the lexicon-based pre-segmentation of a sentence. Unknown words in the text would be incorrectly segmented into pieces of single syllable or shorter words through pre- segmentation.

Figure 3 illustrates two possible pre-segmented results of syllable string, that is, explicit unknown words in above expression or hidden unknown word in below expression.

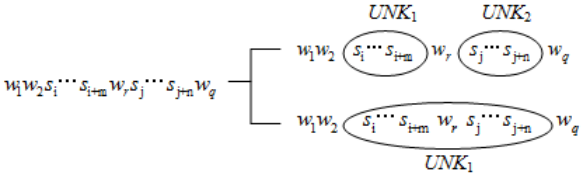


Figure 3. Categories of Tibetan unknown words

In Figure 3, *UNK*, *w* and *s* denotes unknown word, word and syllable respectively. Only explicit unknown words are discussed in this paper.

Assume *S* is a sentence; and the right side of following equation represents its pre-segmented result.

$$S = w_1 w_2 s_1 s_2 s_3 w_3 s_4 s_5 s_6 w_4$$

$$\text{where} \quad w_1, w_2, w_3, w_4 \in \text{Lexibase}$$

$$s_1 s_2 s_3, s_4 s_5 s_6 \notin \text{Lexibase}$$

We name consecutive monosyllables (i.e. $s_1 s_2 s_3$) after the first-time word segmentation as segmentation fragments. Table 1 gives examples of Tibetan word segmentation fragments.

| segmentation fragments | Correct segmentation | Translation of terms |
|---|---|---|
| ཐུ་ཀུ་/ རལ་/ ཝི་/ ལི་/ ཡི/ | ཐུ་ཀུ་རལ་ཝི་ལི་ཡི | Turrell wylie |
| ཏེ་/ ཀི་/ ཡོ/ | ཏེ་ཀི་ཡོ | Tokyo |
| ཀོ་/ ལུམ་/ བྷི་/ ཡ་/ མཐོ་/ སློབ/ | ཀོ་ལུམ་བྷི་ཡ་མཐོ་སློབ | Columbia university |

Table 1: Example of segmentation fragments.

Column II in Table 1 is the correct segmentation of these unknown words. After maximum-matching word segmentation, it is segmented to the content in column I. Almost all these unknown words in our corpus are segmented into monosyllables because these words are not included in our Tibetan word segmentation lexicon.

At detection stages, the contextual rules were applied to detect fragments of unknown words, i.e. monosyllabic morphemes. Since it is hard to derive a set of morphological rules, which exactly cover all types of unknown words, statistical rules are designed without differentiate their extracted word types.

A corpus-based learning method is proposed to derive a set of rules for monosyllabic words and monosyllabic morphemes. The idea is that if two consecutive morphemes are highly associated then combines them to form a new word.

For each bi-seed-gram, the mutual information MI and t-score are calculated. These scores reflect the co-occurrence affinity between the two tokens of the bi-gram. These two scores are calculated by the following formulas:

$$MI^2 = \log_2 \frac{a^2}{(a+b)(a+c)} \quad (1)$$

$$t = \frac{P_r(w_a, w_b) - P_r(w_a) \times P_r(w_b)}{\sqrt{\frac{1}{N} P_r(w_a, w_b)}} = \sqrt{a} - \frac{(a+b)(a+c)}{\sqrt{a}(a+b+c+d)} \quad (2)$$

where, *a*, *b*, *c* and *d* are elements of a contingency table. For example, given a bi-gram containing tokens *x* and *y*,

*a* = number of bi-grams in which both *x* and *y* occur;

*b* = number of bi-grams in which only *x* occurs;

*c* = number of bi-grams in which only *y* occurs;

*d* = number of bi-grams in which neither *x* nor *y* occurs.

Another measure for Tibetan UWI is seed extension confidence. Denote Tibetan word (or syllable) grouping of n-grams as $S_T(n)$, where *n* indicates the length of current word; Extend it to an adjacent Tibetan syllable and get $S_T(n+1)$, so the seed extension confidence $C_n$ defined as:

$$C_n = \lambda_1 |MI_{mean}(n) - MI_{mean}(n+1)| + \lambda_2 |T_{mean}(n) - T_{mean}(n+1)| \quad (3)$$

in which $MI_{mean}$ and $T_{mean}$ indicates the mean of *MI* and t-value in the scope of extended Tibetan word respectively.

To characterize Tibetan unknown words and their boundaries the extension step will be held. For each extension-ready Tibetan seed word, note the extension confidence $C_n$; if $C_n$ is greater than the threshold, current Tibetan word is accepted, and extension continues; when $C_n$ is less than the threshold extension stops. Boundary for Tibetan unknown word is obtained at the end of extension. Figure 4 shows the detail of extension process. High frequency bi-seed-gram can be extended to an unknown word (which is in brackets in Figure 4) using $C_n$.

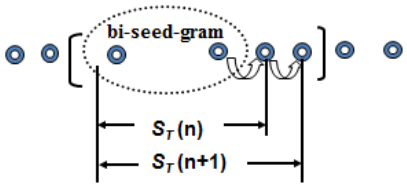


Figure 4. Concept of bi-seed-gram extension

## 4 Evaluation

In this section, we first evaluate performance of Tibetan unknown word identification; then present the performance of Tibetan word segmentation system SegT with unknown word discovery to show the positive effect of UWI.

### 4.1 Experimental Data

We have built the largest Tibetan text resources over the internet via an automatic crawler. They are from three web sites, that are, *Tibet Daily*, *People's Daily* and *Qinghai Daily*. This News Corpus includes different fields such as politics, science, technology, education, language and culture, religion, tourism, environment and Tibetan medicine. Presently, other types of text, especially informal discussion on social network like Twitter and Wikipedia in Tibetan is in small size. Thus, we will utilize above Tibetan News Corpus to extract likely new words in this paper. Our evaluation data contains 12,027 words from 737 randomly selected sentences which have word checking results (the proportion of unknown word is more than 1%).

### 4.2 Performance of Tibetan UWI

We will use the precision, recall, f-score of unknown word ($P_{unk}$, $R_{unk}$, $F_{unk}$) to evaluate the performance of Tibetan UWI. In our 3-fold cross validation, 70% of evaluation data is selected as training set, and the remainder is test set. Table 1 shows the Tibetan unknown word identification results on our evaluation dataset.

| Method | $P_{unk}$ | $R_{unk}$ | $F_{unk}$ |
|---|---|---|---|
| MT | 0.8205 | 0.7091 | 0.7607 |
| MTC | 0.8323 | 0.7606 | 0.7948 |

Table 1. 3-fold cross validation Results of Tibetan unknown word identification.

In Table 1, MT denotes statistical metric, and MTC denotes the combination of MT and context sensitive rules; the given result is the average of 3-fold cross validation. As shown in Table 1, combination of contextual rules with statistical measure can promote the performance of Tibetan UWI; the f-score reaches 79.48%.

After analyzing the results, we find that wrongly identified words can be divided into two classes, i.e., Tibetan person name and transliterated names. We will add deictic words into context sensitive rule and supplement transliteration table to promote identification accuracy of these kinds of unknown words.

### 4.3 Evaluation for Tibetan Word Segmentation with the Extended Lexicon

In order to validate the effect of our unknown word identification on Tibetan word segmentation, we conduct following experiments.

In a typical word segmentation system, once a text is segmented using the available lexicon or heuristic rules, the segmentation process is finished. We observe that unknown words make up 0.5% to 4% of all the words in our Tibetan news articles. Therefore, UWI is an important issue for a word segmentation algorithm. We add a semi-automatic unknown word identification component to the back-end of the whole segmentation process.

We will evaluate the precision ($P_{\text{seg}}$), recall ($R_{\text{seg}}$), f-score ($F_{\text{seg}}$) of Tibetan word segmentation in this subsection.

$$P_{seg} = N_{seg1} / N_{seg2}$$

$$R_{seg} = N_{seg1} / N_{seg3}$$

$$F_{seg} = 2P_{seg}R_{seg} / (P_{seg} + R_{seg})$$

where $N_{seg1}$ denotes the number of correctly segmented Tibetan words; $N_{seg2}$ denotes total number of segmented Tibetan words; $N_{seg3}$ denotes the total number of Tibetan words in the testing texts.

The segmentation of original web texts uses a basic Segmentor (SegT (Liu et al, 2012)) and a general lexicon (with 220,000 Tibetan entries). Unknown words (out of our lexicon) are segmented into pieces in this step. The following process is to detect possible unknown words from word segmentation fragment which are very likely to be words. We will compare lexicon-based Tibetan segmenter with and without unknown word identification component on our evaluation data. Presently, there is no Tibetan word segmentation specification and standard; in addition, there is no large and publicly available Tibetan training corpus. Thus make comparison with other research papers is difficult. We choose the best Tibetan word segmentation system Liu's SegT (Liu et al. 2012) as baseline.

Table 2 illustrates the results of Tibetan segmentation by SegT with general lexicon and SegT with lexicon extension on evaluation. SegT+MTC, denotes Tibetan word segmenter SegT with lexicon extension; the proposed method in section 3 has been applied to semi-automatically extend the lexicon of Tibetan word segmentation system SegT.

| | $P_{seg}$ | $R_{seg}$ | $F_{seg}$ |
|---|---|---|---|
| SegT | 0.7769 | 0.8638 | 0.8181 |
| SegT + MTC | 0.8197 | 0.8872 | 0.8521 |

Table 2: Effects of Tibetan word segmentation.

Experimental results show that the maximum word segmentation performance is got using general lexicon extended by MTC. As we see from Table 2, the precision, recall and f-score are increased by 5.49%, 2.71%, 4.15% respectively compared with SegT. The score of SegT+MTC is increased significantly because of the higher proportion of unknown words. The experimental results demonstrate that the Tibetan word segmentation system SegT with proposed unknown word extension mechanism is indeed helpful to promote the accuracy and recall rates of Tibetan word segmentation.

## 5 Conclusion

In this paper, we present a hybrid method for Tibetan unknown word identification. Its f-score reaches around 80%. Compared with English or Chinese unknown word recognition work, the proposed methods doesn't achieve satisfactory results, however, preliminary experimental results demonstrate that SegT with proposed unknown word extension mechanism is indeed helpful to promote Tibetan word segmentation performance. In the future, the evaluation of proposed method needs to be extended to large-scale test corpus and detailed context sensitive rules are used to identify Tibetan unknown words.

## 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.61303165, No.61202219, and No.61202220), Major Science and Technology Projects in Press and Publishing (No.0610-1041BJNF2328/23), and Informationization Project of the Chinese Academy of Sciences (No.XXH12504-1-10).

## Reference


Kawtrakul Asanee, Thumkanon Chalatip, Poovorawan Yuen, Varasrai Patcharee, Suktarachan Mukda. 1997. Automatic Thai Unknown Word Recognition.

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

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

Zhi-jie Cai. 2009b. The Design of Banzhida Tibetan word segmentation system. In: proceedings of *the 12th Symposium on Chinese Minority Information Processing*.

Jing-Shin Chang and Keh-Yih Su. 1997a. An Unsupervised Iterative Method for Chinese New Lexicon Extraction. *International Journal of Computational Linguistics & Chinese Language Processing.*

Hsin-Hsi Chen and Jen-Chang Lee. 1994. The Identification of Organization Names in Chinese Texts. *Communication of Chinese and Oriental Languages Information Processing Society*, 4(2), Singapore, 1994, pp131-142 (in Chinese).

Keh-Jiann Chen and Wei-Yun Ma, 2002. Unknown Word Extraction for Chinese Documents. In: Proceedings of *COLING 2002*, pp 169-175.

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

Yu-Zhong Chen, Bao-Li Li, Shi-Wen Yu and Lancuoji. 2003b. An Automatic Tibetan Segmentation Scheme Based on Case Auxiliary Words and Continuous Features, *Journal of Applied Linguistics*, (01): 75-82.

Tung-Hui Chiang, Jing-Shin Chang, Ming-Yu Lin and Keh-Yih Su. 1992. Statistical Models for Word Segmentation and Unknown Word Resolution. In: Proceedings of *ROCLING V*, pp 121-146.

Lee-Feng Chien. 1999. PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval. *Information processing and management* 35:501-521.

Kenneth Church, William Gale, Patrick Hanks and Donald Hindle. 1991. Using Statistics in Lexical Analysis. In: *Zernik ed. Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Hillsdale, N J: Erlbaum, pp 115-164.

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

Ted E. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19 (1): 61-74.

Xian-Ping Ge, Wan-Da Pratt, and Padhraic Smyth. 1999. Discovering Chinese Words from Unsegmented Text. In: proceedings of *SIGIR '99*, pp 271-272.

Chin-ming Hong, Chih-ming Chen, Chao-yang Chiu. 2009. Automatic extraction of new words based on Google News corpora for supporting lexicon-based Chinese word segmentation systems. *Expert Systems with Applications*, 36:3641-3651.

Yun-Lun Li, Bao-Bao Chang. 2010. Maximum Margin Markov Networks-Based Chinese Word Segmentation Method. *Journal of Chinese Information Processing*, 24(l):8-14.

Ming-yu Lin, Tung-hui Chiang and Keh-Yih Su. 1993. A preliminary study on unknown word problem in Chinese word segmentation. In: Proceedings of *1993 R.O.C. Computational Linguistics Conference*, Taiwan, pp 119-137.

Hui-dan Liu, Wei-na Zhao, Ming-hua Nuo, Li Jiang, Jian Wu, Ye-ping He. 2010. Tibetan Number Identification Based on Classification of Number Components in Tibetan Word Segmentation. In: Proceedings of *the 23rd International Conference on Computational Linguistics - poster volume* (*COLING 2010*), pp 719-724.

Hui-dan Liu, Ming-hua Nuo, Wei-na Zhao, Jian Wu, Ye-ping He. 2012. SegT: A Practical Tibetan Word Segmentation System. *Journal of Chinese Information Processing*, 26(1):97-103.

Zhi-Yong Luo, Rou Song. 2004. An Integrated Method for Chinese Unknown Word Extraction. In: Proceedings of *3rd ACL SIGHAN Workshop*. Barcelona, Spain. pp 148-154.

Wei-Yun Ma and Keh Jiann Chen. 2003. A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction. In: Proceedings of *the Second SIGHAN Workshop on Chinese Language Processing*, pp 31-38.

Christopher D. Manning, Hinrich Schutze. 1999. Foundations of Statistical Natural Language Processing, MIT Press.

Palmer D. David. 1997. A Trainable Rule-based Algorithm for Word Segmentation. In: Proceedings of *the 35th Annual Meeting of ACL and 8th Conference of the European Chapter of ACL*. Madrid.

Patrick Pantel and De-kang Lin. 2001. A statistical corpus based term extractor. In E. Stroulia and S.

Matwin, editors, *Lecture Notes in Artificial Intelligence*, pp 36-46. Springer-Verlag.

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

Wei Qiao, Mao-song Sun. 2010. Joint Chinese word segmentation and named entity recognition based on max-margin Markov networks. *Journal of Tsinghua University* (*Science & Technology*), 50(5): 758-762.

Richard Sproat and Chilin Shih. 1990. A Statistical Method for Finding Word Boundaries in Chinese Text. *Computer Processing of Chinese and Oriental Languages*, 4, 336-351.

Mao-song Sun, Chang-ning Huang, Benjamin K. Tsou, Fang Lu and Da-yang Shen.1997. Using Character Bigram for Ambiguity Resolution in Chinese Word Segmentation. *Computer Research & Development*. 34(5):332-339.

Mao-song Sun, Chang-ning Huang, Hai-yan Gao, Jie Fang. 1995. Identifying Chinese Names in Unrestricted Texts. *Journal of Chinese Information Processing*, 9(2):16-27.

Mao-song Sun, Da-yang Shen and Benjamin K. Tsou. 1998. Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data. In: Proceedings of *COLING-ACL '98*, pp1265-1271.

Xiao Sun, De-gen Huang, Hai-yu Song et al. 2011. Chinese new word identification: a latent discriminative model with global features. *Journal of computer science and technology*, 26(1): 14-24.

Yuan Sun, Luosangqiangba, Rui Yang and Xiao-Bing Zhao. 2009. Design of a Tibetan Automatic Segmentation Scheme. In: proceedings of *the 12th Symposium on Chinese Minority Information Processing.*

Yuan Sun, Xiao-Dong Yan, Xiao-Bing Zhao and Guo-Sheng Yang. 2010. A resolution of overlapping ambiguity in Tibetan word segmentation. In: Proceedings of *the 3rd International Conference on Computer Science and Information Technology*, pp 222-225.

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

Jakkrit TeCho, Cholwich Nattee, Thanaruk Theeramunkong. 2012. Boosting-based ensemble learning with penalty profiles for automatic Thai unknown word recognition. *Computers and Mathematics with Applications* 63, pp 1117-1134.

T. Tomokiyo and M. Hurst. 2003. A Language Model Approach to Keyphrase Extraction. In: Proceedings of *ACL-2003 workshop on multiword expressions*. Sapporo, Japan. pp 33-40.

C.H. Tung and H. J. Lee. 1995. Identification of unknown words from corpus. *International Journal of Computer Processing of Chinese and Oriental Languages*, Vol. 8, Supplement, pp 131-146.

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

Hua-ping Zhang, Qun Liu, Xue-qi Cheng. 2003. Chinese lexical analysis using hierarchical hidden Markov model. In: proceedings of *Second SIGHAN workshop affiliated with 41th ACL*. Sapporo Japan, pp 63-70.

Kevin Zhang (Hua-Ping Zhang), Qun Liu, Hao Zhang, Xue-qi Cheng. 2002. Automatic Recognition of Chinese Unknown Words Based on Role Tagging, In: Proceedings of *SigHan 2002 Workshop attached with the 19th International Conference on Computational Linguistics*, Taipei, September. pp 71-77.

Ying Zhang, Ralf D. Brown, Robert E. Frederking, Alon Lavie. 2001. Pre-processing of Bilingual Corpora for Mandarin-English EBMT. In: Proceedings of *MT Summit VIII*, Santiago de Compostela, Spain.