ABSTRACT
Tibetan person name recognition is one of the most difficult tasks in the area of Tibetan information processing, and the effect of recognition impacts directly on the precision of Tibetan word segmentation and the performance of related application systems, including Tibetan-Chinese machine translation, Tibetan information retrieval, text categorization, etc. Based on the analysis of wording rules and features of Tibetan person names, this paper proposes a method which combines maximum entropy and conditional random fields to identify Tibetan person names. The experiment shows that this approach works quite well, with the value of F1-measure reaching 93.29%.

KEYWORDS
Tibetan name recognition, maximum entropy, conditional random fields
A hybrid approach using maximum entropy models and conditional random fields to identify Tibetan person names

Yangji Jia  Yachao Li  
Northwest University for Nationalities

Chengqing Zong  Hongzi Yu  
Chinese Academy of Sciences  Northwest University for Nationalities

1 Introduction

Named entity recognition is a basic problem in natural language processing. Its main task is to identify named entities name such as person names, place names, organization names, quantity, and time expressions and so on in the text. Named entity recognition is very difficult. The recognition precision directly affects segmentation accuracy and the performances of related systems.

This paper introduces a new method of Tibetan name recognition. Unlike Chinese names, the syllables of Tibetan person names are likely not to be contextually delimited. Moreover, many common words are used as person names directly, which poses an obstacle for Tibetan personal name recognition.

Tibetan names discussed here include Tibetan name per se and names translated into Tibetan (mostly Chinese names and foreign names translated via Chinese).

The main difficulties in Tibetan named entity recognition can be summarized as followings:

1) Tibetan syllables are delimited by syllable points (tsheg) in the text and there are no intervals or other boundary identification between words. Thus, word segmentation and tagging are necessary before recognizing named entities.

2) Tibetan person names have no obvious morphological characteristics, and they are not distinguished by an orthographic convention such as the capitalization of the first letter, as in English.

3) A large number of common words are used as person names, particularly popular are concepts from the natural world, the day of someone's birth, place in birth order, etc., for example “ལྷ་མཚོ” (“ocean”), “ཁོ་རོ་(flower)”, “ལོངས་ཀར་(longevity)”, “བོད་(happiness)”, “ཁོ་རོ་(Friday)”, “བོད་(first)”. Such names increase the difficulty of name recognition.

4) Syllable length is not normative. Tibetan person names are mainly formed by two, three and four syllables, but there are also some single syllable names (like “བོད་”), and long multi-syllable names (like “བོད་ཀུན་བོད་ཀུན་བོད་ཀུན་བོད་ཀུན་”) (Mitra 1991). Tibetan names can extend up to 26 syllables (Wang 1991).
5) Lack of relevant resources such as name dictionaries. Currently we have rule-based approaches and statistical-based methods for name recognition. Many scholars have done work on name entity recognition in China, but mainly for Chinese and English. Zheng Jiaheng (2000) has researched on extraction and analysis of the Chinese first and last names with word frequency on large scale corpus, and thereby studies the evaluation function of Chinese Name Recognition. The method used in approaching named entity recognition in recent years is gradually shifting from earlier rule based methods to machine learning methods (Zong 2008). Li Zhongguo (2006) first identifies possible names by using boundary templates, then make an edge community correction by applying local context statistics and heuristic rules. Zhang Huaping (2004) pursues on role-word labels by taking V algorithms, puts forward a role labeling based automatic Chinese names identification method. Zhang Suxiang (2009) proposed a Chinese name identification method based on CRFs (conditioned random fields) by using physical features and internal particle characteristics. Mao tingting (2007) has also proposed a Chinese automatic name recognition method by combining SVM (support vector machines) and probabilistic model. In addition, there are studies on Chinese name recognition using combinations of maximum entropy model and rules, which obtained high recalls (Qinn 2006; Jia 2007).

Gazang Zhuoma (2008) provides a more comprehensive analysis, including a deep exploration of the cultural significance of Tibetan names. Luo Zhiliang et al. (2003) give a statistical analysis of naming words and rules based on examples of names and corpus. These authors studied Tibetan names with currently available methods, and filtered out some high-frequency words, such as “ཟ”, “ི”, “ཤ”. Finally, they put forward a credibility based model on the Tibetan name recognition, and it confirmed the contribution of Chinese Word Segmentation System in Tibetan name recognition module (Liu 2009). We believe the constituent units of Tibetan person names are including monosyllabic words and two-syllable words from a structural view. In addition, Tibetan Names have rich boundary information and there is a pattern to follow. Therefore, we research Tibetan name recognition by using name characteristics wording and boundary information.

The rest of this article is organized as follows: Section 2 describes the constitute characteristics of Tibetan Names; Section 3 presents a hybrid approach to Tibetan person name identification by maximum entropy model and conditional random fields; The experimental results are analyzed in section 4 and we offer some conclusions and outlooks in section 5.

2 Constitute Features of Tibetan names

2.1 Types of names

A name is a symbol which differentiates among people. Names are rich linguistic and cultural phenomena. Tibetan naming practices follow several patterns. For example:

- named in religious terms: རྡོ་ རྗུ་;
- named after natural phenomena: རྲ་ོ ཡི་རྒྱལ;
- named by expression of good wishes: རྩ་ རིང་དབང་ཆེན;
- named with commemorating the birth: བོད་ཀྱི་ རིང་གོས;
- named by animal name: རྒྱ་མོ་ར་གུ;
- names of historic events: རོ་བཅུ་ རིག་གནས་ཐབས་.
We analyzed all Tibetan names appearing in a corpus consisting of the January 2007 issues of the Tibet Daily Newspaper, among which, 91% of Tibetan names are of the first three types discussed above.

### 2.2 Characteristics wording

1) The constituent units of Tibetan names include monosyllabic words and two-syllable words. Names wording collection are more in dispersion. However, the length of Tibetan names varies in a small range. 95% of Tibetan names appearing in the corpus consist of 2 to 4 syllables. The recognition of names with two syllables is very important because most longer names are formed by adding together two-syllable names.

2) According to the statistics, about 92% of four-syllable names in the corpus tend to be composed of 2 two-syllable names. Just as W1: "དོན་ǿབ" and W2: "ཚǃ་རིང" are both two-syllable names, but they could constitute a four-syllable name.

3) Name positions are relatively fixed. In general, they can appear at the head or the middle of a sentence, but not at the end. Tibetan has SOV word order, and the verb is always at the end of a sentence.

   For example:
   
   [ཅན་མ་ཞིང་ཉི་མ་བདེ་Dzིད མཚǑ ངལ] (name appears at the head of sentence.)
   
   [བན་ཞི་གཞོན་པ་] (name appears at the middle of sentence.)

4) From the perspective of part of speech, Tibetan names not only contain full lexemes "ོལ་དཀར ཉི་མ་བདེ་Dzིད མཚǑ ངལ", but also function words such as དི་ and དོ་ in དི་མོ་བཏོན་ or དོ་ནས་འཇོག་.

5) Names of four syllables may be abbreviated. For example: the four-syllable name "ཚǃ་རིང་Ɉན་ཚǑགས" can be abbreviated into a two-syllable name "ཚǃ་Ɉན". Other examples are "བསོད་ཚǃ། ངོལ། བྱ་བȪན། ཉི་དོན།". Most such abbreviated names are two-syllable words.

### 2.3 Boundary information

While analyzing name boundary information, we take single vertical character "\n\n" as a punctuation identifier, and understand as a complete sentence a span of text that ends with "\n\n", using this span as a processing unit. Boundary information plays a very important role in name recognition. For example:

1) <znr>ལྷིན་ཐེལ་བ་</znr><nr>སྣང་བཟང་དཔལ་</nr><ynr>བཟང་པོ</ynr>

2) ེེ་ལམ་<znr>གསར་འགོད་པ</znr><nr>བཀྲ་ཤིས་དོན་ǿབ</nr><ynr>ཀྱིས་</ynr><ynr>བཀོད་པའི་གསར་འǽར་ནང་ȭ།</ynr>

In the above examples, tag "nr" Indicates the target names, while tag "znr" indicates the left boundary, and tag "ynr" indicates the right boundary.

In Tibetan, many words can indicate the boundary of names, such as "ོབོ་མ་ (comrade)", "ོཡོང་ (chairman)", "ོཆེ་ཐུན (teacher)", and "ོབོ་ (an honorific word)". These words are boundary word which has help for inspiration and instruction for person names. When these words appear in corpus, the credibility of name recognition will be improved.

We extracted 1403 person names from 2007 Januarys corpus (about 3.5MB) of the Tibet Daily Newspaper, and found 995 Tibetan names and 408 translated names. As in the following sentence:

<znr>བི་ཐོང་</znr><nr>སྣང་བཟང་དཔལ་</nr><ynr>བཟང་པོ</ynr><ynr>ལྷིན་ཐེལ་བ་</ynr>
The left boundary word “འིན་ེལ་བ་” and right boundary word “ཀྱིས” are extracted. From the corpus, 117 left boundary words and 84 right boundary words are extracted. Their occurrences are collected. The top ten of left and right boundaries words are listed in the following tables.

| S  | SNR | དང་ | ཉི་ | ཐོ་ | གཞི | སྒོ | མ། | ཞི | དཔ | ཊོ | ངོ | ནད | ཐོ | ལོ | རོ | བ་ | འིན་ | གེལ་ | བ་ | ཌོ | ངོ | ནད | ཐོ | ལོ | དཔ | ཊོ | ངོ | མ། | ཞི |
|----|-----|------|------|------|------|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 590 | 297 | 82   | 50   | 29   | 27   | 25   | 23  | 18  | 15  |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

Table 1. frequency of left boundary examples

| S  | དང་ | ཉི་ | ཐོ་ | གཞི | སྒོ | མ། | ཞི | དཔ | ཊོ | ངོ | མ། | ཞི | ཐོ | ལོ | རོ | བ་ | འིན་ | གེལ་ | བ་ | ཌོ | ངོ | ནད | ཐོ | ལོ | དཔ | ཊོ | ངོ | མ། | ཞི |
|----|------|------|------|------|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 568 | 383  | 194  | 53   | 51   | 45   | 14  | 12  | 7   | 4   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

Table 2. frequency of right boundary examples

The SNR in table 1 indicates the personal name appears on the head of sentence, i.e. when the left boundary word is null. Function word “ཀྱིས” in table 2 indicates that implementation function word has appeared 5 times in the corpus, the frequencies in descending order are “ཀྱིས (140)”, “ཀྱིས (88)”, “ཀྱི (76)”, “ཀྱི (61)” and “ཀྱི (18)”. The Tibetan function word “(ནོ)” in table 2 indicates the five genitive function words, and the frequencies in descending order are “ནོ (13)”, “ནོ (9)”, “ནོ (7)” and “ནོ (3)”.

The word sequence of Tibetan names containing sentences can be expressed as follows: \( W_1W_0W_1 \), \( W_0 \) represent the headword name, \( W_1 \) represent the left boundary word and \( W_1 \) represent the right boundary word. \( W_1 \) could be none under normal circumstances, when it is, \( W_1=SNR \), indicates the head of the sentence is a person name.

We have summarized and compiled statistics for the boundary words frequently occurring in the corpus, and find that the left boundary word of Tibetan name is generally an occupation name or title such as chairman, secretary, uncle, journalist, teacher, county head and herdsman, and the structure is: “<“teacher”><nr>, “<“uncle”><nr>, “<“villager”><nr>”. Clearly we can find that there is no modifier among it, and the tag “nr” represents a person name. The right boundary words are function words and only seldomly honorific suffixes, modal particle and other words. The right boundary information is more dispersed compared to the left boundary information, which has broad vocabulary using relatively, and the recognition performance is much lower. Therefore, the article is based on the left boundary information and then adds the right boundary information. The structure is as follows: “<“པོའི་ཐོས་ཀྱིས”><nr>, “<“ཤེས་་”><nr>, “<“ཅོག་”><nr>, “<“བཤད་ཀྱིས”><nr>, “<“ཐོབ་”><nr>”.

3 A hybrid approach to Tibetan person name identification using maximum entropy models and conditional random fields

3.1 The maximum entropy principle

Maximum Entropy was originally put forward by E. T. Jaynes in 1950. It is applied to natural language processing model for the first time by Della Pietra in 1992. The basic idea is as follows. Firstly, by using a given training sample, one chooses a consistent probability distribution as the training sample, and they must meet all known facts. It will be given a uniform probability
distribution for those unsure of the part in the absence of more constraints and assumptions of the case. Entropy is used to indicate the uncertainty of a random variable. The greater the uncertainty, the greater the entropy, and the more evenly distributed.

maximum entropy model:

\[ P^* = \arg \max_{P \in \mathcal{C}} H(P) \]  
(1)

where \( H(P) \) is the entropy of the model \( P \), and \( \mathcal{C} \) is the model collection when it satisfies the constraint conditions. Under the surface we need to seek \( P^* \). The form of \( P^* \) such as formula shows:

\[ P^*(y|x) = \frac{1}{Z(x)} \exp(\sum_i \lambda_i f_i(x,y)) \]  
(2)

where \( Z(x) \) is a normalization constant and its form is as shown in formula (3)

\[ Z(x) = \sum_y \exp(\sum_i \lambda_i f_i(x,y)) \]  
(3)

Among it, \( \lambda_i \) is the weight parameter of a feature.

3.2 CRFs principle

CRFs model is a new classification method which is made by Lafferty in 2001. It models the target sequence on the basis of sequence observing. Define \( O = \{O_1, O_2, ..., OT\} \) as the observed input data sequence; define \( S = \{S_1, S_2, ..., ST\} \) as the predicted state sequence. Then, in the case of a given input data sequence, the linear chains (CRF) of the parameter \( = \{\lambda_1, \lambda_2, ..., \lambda_T\} \), and the conditional probability of state sequence which the parameter output could be:

\[ P^*_\Delta = \langle S|O \rangle = \frac{1}{Z_0} \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t)) \]  
(4)

Among it, \( f_k(s_{t-1}, s_t, o, t) \) is an arbitrary characteristic function, and \( \lambda_k \) is the weight for each characteristic function. \( Z_0 \) is the normalized factor, and it is defined as in formula (5):

\[ Z_0 = \sum_s \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t)) \]  
(5)

3.3 A hybrid approach to Tibetan person name identification using maximum entropy models and conditional random fields

Many problems in Natural language processing can be considered linguistic classification issues as can name recognition. For those candidate words in a corpus, we can determine whether it is a name according to the label information of the border words.

Wherein the Characteristic function \( f(x,y) \) is a binary function, it is one of the representations of acquired characteristics. For the feature \((x_i, y_i)\) among Tibetan name recognition, Characteristic function is defined as in equation 6:

\[ f(x,y) = \begin{cases} 1 & \text{if } x = x_i \text{ and } y = y_i \\ 0 & \text{Otherwise} \end{cases} \]  
(6)
Where “y” is the resulting output of names entity, which represents that the central word $W_0$ equals to $Y$ (is a name) or $N$ (not a name) under the condition $x$, and $x$ is the boundary feature information corresponding to.

### 3.3.1 Characterization

1) Boundary feature set

**Target Words**

Boundary word window size of the target word $W_0$ will take plus or minus 1, boundary characteristics is constituted by the left boundary characteristics and right boundary characteristics.

①Names left boundary feature set:

When $W_{-1}$ appears on the Left Edge Word Table (ZNRL), as ZNR ($W_{-1}$) = true, then the left boundary Characteristic is satisfied, as the formula 7 shows:

$$
znr(x, y) = \begin{cases} 
1 & \text{if } znrf \left( W_{-1} \right) = true \text{ and } y = person \\
0 & \text{Otherwise} 
\end{cases}$$

(7)

②Names right boundary feature set: When $W_{1}$ appears on the Right Edge Word Table (YNRL), as YNR ($W_{1}$) = true, then the left boundary Characteristic is satisfied, as the formula 8 shows:

$$
ynr(x, y) = \begin{cases} 
1 & \text{if } ynr \left( W_{-1} \right) = true \text{ and } y = person \\
0 & \text{Otherwise} 
\end{cases}$$

(8)

2) Template feature set

Binding information $W_{-1}$ and $W_{1}$, unite them as a border template feature.

3) Names Dictionary Feature Set

A names dictionary is an important resource in Name Recognition. We established a common Tibetan names dictionary by choosing 2058 common names from “dictionary of common Tibetan personal and place names” (Chen 2004).

4) Word feature set of translated names

According to the statistics, words using of translated names exhibit some patterns. Most of them are not commonly used Tibetan word. We collected 273 translated names like “ཀྲང་” (张), “ཧའོ་” (郝), “ལི་” (李), “ʆན་” (沈) and”ȣང” (董).

### 3.3.2 Feature extraction

Feature selection and extraction is an important step in establishing the model for Tibetan names recognition. Tibetan names characteristics can be seen by analyzing that the most important information included names is the boundary information except wording information. Boundary information refers to boundary words place before or after names. For example: “ཉི་མ་ཚǃ་རིང་” is the target name inside the sentence “གསར་འགོད་པ་ཉི་མ་ཚǃ་རིང་གིས་བ­iར།”, and “གསར་འགོད་པ་”, “གིས་” are the left and right boundary word respectively.

When considering to take plus or minus 1 to boundary word window size of the target word $W_0$, there are three pieces of information that can be extracted: 1) left boundary word $W_{-1}$, 2) target
word $W_0$, 3) right boundary word $W_1$. After using this information to build boundaries and names wording template, we can extract a lot of features from the training corpus.

3.3.3 Merging method

The maximum entropy model kept a better recall rate in experiments because it has more flexible feature selection and stronger portability when applied to different areas and is highly robust on its own. CRF can solve the label bias by maximum entropy models, so it can greatly improve the accuracy rate. For the merits of the two models, we propose a method to identify the integration of the two models. The evaluation function is defined as:

$$Total = (\lambda \times CRF) + ((1 - \lambda) \times Maxent) \quad (9)$$

Among them, $\lambda$ represents the weight, we can get the best result by adjusting the value of $\lambda$. CRF and Maxent are present testing result on name recognization by the CRF method and the maximum entropy method.

We see from the experimental data fusion that we can overcome their respective weaknesses and effectively solve the Tibetan name recognition problems by integrating these two models.

4 Results and error analysis

The experiment utilized 2007 Januarys corpus (about 3.5MB) of the Tibet Daily Newspaper as the training corpus. The part from February 1$^{st}$ to 20$^{th}$ corpus (about 2.1MB) are taken as an open test corpus. We took three evaluation indicators in the test:

(1) Precision rate

$$P = \frac{\text{correct recognized name numbers}}{\text{whole recognized name numbers}} \times 100\% \quad (10)$$

(2) Recall rate

$$P = \frac{\text{correct recognized name numbers}}{\text{whole names among the test corpus}} \times 100\% \quad (11)$$

(3) F-Measure

$$F = \frac{2 \times P \times R}{P + R} \quad (12)$$

In the experiment, we first tested the performance of Tibetan name recognition using the maximum entropy method and the CRF method. The results are shown in Table 3:
As can easily be seen in Table 3, *Maxent* produces a better recall rate but worse precision rate. CRF method even could produce a better precision rate, but names can’t be recalled. The system recall rate has a significant reduction.

After combining the two models by using the formula fusion 9, the results are as follows:

<table>
<thead>
<tr>
<th>Methods</th>
<th>True name numbers</th>
<th>System label name numbers</th>
<th>Correct name numbers</th>
<th>Precision rate%</th>
<th>Recall rate%</th>
<th>F-Measure %</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>773</td>
<td>697</td>
<td>679</td>
<td>97.42</td>
<td>87.83</td>
<td>92.38</td>
</tr>
<tr>
<td>Max-ent</td>
<td>773</td>
<td>741</td>
<td>693</td>
<td>93.52</td>
<td>89.65</td>
<td>91.54</td>
</tr>
</tbody>
</table>

Table 3. Recognition result

When the value is 0.58, the recall rate increased by 0.39% compared with Maxent. F-Measure of the system is changing by taking 0.61 to $\lambda$ as a reference, when $\lambda \leq 0.61$, F-Measure value increases as $\lambda$ increased. When $\lambda \geq 0.61$, the F-Measure decreases as $\lambda$ decreases. When $\lambda$ equals to 0.61, F-Measure value of system will improve 1.53% than Maxent and improve 0.7% than CRF. Experimental results show that the integration of the maximum entropy method and conditions method is very effective to Tibetan name recognition.

As shown in Table 5, we found the following four categories of the more typical errors in the experiment.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Precision rate%</th>
<th>Recall rate%</th>
<th>F-Measure%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>97.42</td>
<td>87.83</td>
<td>92.38</td>
</tr>
<tr>
<td>Max-ent</td>
<td>93.52</td>
<td>89.65</td>
<td>91.54</td>
</tr>
</tbody>
</table>

Table 4. Recognize result
We can see the typical errors in Table 5 that even the method could have a nice recognition result, but as the identification method has strong dependence on a names dictionary and boundary information, it lead to the first two wrong categories in Table 5. The method can't handle the circumstances of having conflict with general term and blurry boundary characteristics, only in order to know accurately by obtaining more syntactic structure and context information.

### 5 Conclusion and outlook

From the Tibetan name characteristics, we presents a hybrid approach to Tibetan person name identification using maximum entropy models and conditional random fields by analyzing the Tibetan names of naming rules, names wording features, boundary information and other features. This method blends the advantages of both maximum entropy and CRFs. We also have auxiliary implements for Tibetan automatic name recognition such as boundary word table and translated name table.

Experimental results show that the method can achieve better recognition results.
According to the experimental analysis, issues like sparse data and person name and common word confictions still exist for current Tibetan name recognition methods. Therefore, we will further improve the quality of the corpus and expand the size of the training corpus in the following study. Simultaneously, we will optimize the boundary repositories, expand the feature templates and test on other possible recognition models like SVM (support vector machine) to improve the accuracy of name recognition.

References


Qian, Jing; Zhang, Yuejie; and Zhang, Tao. 2006. “Research on Chinese person name and location name recognition based on maximum entropy models”. Mini-Micro Systems 27.9: 1701-1765. (钱晶，张明杰，张涛. 基于最大熵的汉语人名地名研究[J]. 小型微型计算机系统，2006，27(9): 1701-1765.)


Yangji Jia
jyangji@163.com