A Semi-supervised Learning Approach for Person Name Recognition in Tibetan

Zhijuan Wang[1,2], **Fuxian Li**[3]
Minzu University of China[1], Alibaba[3]
National Language Resource Monitoring and Research Minority Language Center[2]
No.27, South Street of Zhongguancun, Haidian District, Beijing, China[1,2]
No.8, Haidian Street, Haidian District, Beijing, China[3]
{wangzj.muc, fushine.lee}@gmail.com

Abstract

Massive labelled data is important for Named Entity Recognition(NER). For Low Resource Languages(LRL), massive labelled data means more labor, more time and more cost. A semi-supervised learning (SSL) that need fewer labelled data is proposed to recognize person name in Tibetan texts. Based on Conditional Random Fields (CRFs) and Radial Basis Function (RBF), this method use 5-element feature matrix to propagate information from few labeled data to massive unlabelled data. Experiments demonstrate that its F-measure can achieve 84% using only 100 documents as seeds, whereas about 800 labeled documents are required for a supervised learning based on pure CRFs.

Keywords: Low Resource Languages, Person Name Recognition, Semi-supervised Learning

1. Introduction

Named Entity Recognition (NER), whose main task is to recognize the names of persons, locations and organizations from texts in different languages texts, is an important task for information extraction (IE), Information Retrieval (IR)Information Retial. As mentioned in (David and Satoshi, 2007) (Chung et al., 2003) (Popov et al., 2004) (Benajiba et al., 2007) (Seker and Eryigit, 2012), NER research had covered many languages such as English, German, Spanish, Chinese, Japanese, Korean, Russian, Arabic, Turkish and so on.

Machine learning approaches such as CRFs, HMM often be used in NER. According to the size of labelled data, machine learning approaches can be divided into supervised learning(need massive labelled data), semi-supervised learning(need a small amount of labelled data) and unsupervised learning(no labelled data). Supervised learning has the better performance in NER. In order to make up for the deficiency of the labelled data,some semi-supervised methods (Nadeau et al., 2007) and unsupervised methods(Michael et al., 2005) are used.

Tibetan is a low resource language which is a cluster of Sino-Tibetan languages and spoken primarily by Tibetan peoples, who live across a wide area of eastern Central Asia. There are some research focused on Tibetan NER, especially on person name recognition. These methods are all based on rules or supervised learning approaches. Very few efforts have been made to develop semi-supervised learning or unsupervised learning for Tibetan NER.

A Semi-supervised Learning Approach is proposed to recognize Person Name in Tibetan.The remainder of this paper is structured as follows: Section 2 includes background information about the features of Tibetan person name and recent work of Tibetan person name recognition. Section 3 illustrates the methodology of the proposed algorithm. The data used in experiment and the evaluation results are reported and discussed in section 4. Finally, we present the conclusion and future work.

2. Background

There is little introduction about Tibetan person names in English. So, we give a brief introduction of Tibetan person name firstly.

2.1. Introduction of person name in Tibetan

Tibetan is an alphabetic writing language, which has 30 consonants and four vowel signs. Its smallest grammar unit is syllable. "་" is the mark of syllable. One or more alphabets compose a syllable and one or more syllables can compose a word. Fig.1 is an example of Tibetan sentence with named entities. We can see that there is no white space between Tibetan words and there is no obvious feature such as capitalization of first letter to identify the person name in Tibetan.

Person names in Tibetan are complex. There are four kinds person name in Tibetan text.

(1) Tibetan first name

Most Tibetan people' name only have first names, which length often range from two syllables to five syllables. For example, "དཔལ་བཟང་ " (Passang), "ལང་མ་འཚོ" (Pematso), "བློ་བཟང་ཆོས་ར་ " (Lobsang Champa). First names of Tibetan people often come from Buddhism or other good wish. Therefore, some first names often be used, which are called high-frequency syllables of Tibetan people'name. For example, "སྒྲོལ་མ་ " (Dolma), "བཀྲ་ཤི་" (Tashi).

(2) Chinese surname name + Chines first name

There are a large number of Transliteration of Chinese person names in the Tibetan text, which may come from Tibetan people(some Tibetan people use Chinese person name) or Chinese people. For example, "ལས་ཅ་ལྷིང་" (Li Ka-shing).

(3)Chinese surname name + Tibetan first name

Some Tibetan people's names not only have Tibetan first name, but also Chinese surname. For example, "ལི་སྒྲོན་མ་ " (Li Dolma) is Tibetan people name. "ལི་" (Li) is Chinese Surname, "སྒྲོན་མ་" (Dolma) is Tibetan first name.

(4) Tibetan surname name + Tibetan first name

Generally, only Tibetan nobility have Tibetan surname. For example, "ང་ཕོད་དགུ་དབང་འཇིགས་མེད་" (Ngapoi Ngawang Jigme) is Tibetan people name. "ང་ཕོད་དགུ་" (Ngapoi) is Tibetan Surname, "དབང་འཇིགས་མེད་" (Ngawang Jigme) is Tibetan first name.

Since the latter two kinds of person names share a small proportion in the Tibetan text, we focus on how to identify the first two kinds person names in the Tibetan text.

2.2. Related Work of NER in Tibetan

The research of Tibetan NER are focused on two approaches.

Rules: (Yu and Jiang, 2010) utilized a rule-based model on case-auxiliary words, lexicon and boundary information list to recognize Tibetan named entity; also, (Sun et al., 2010) used multi-features such as internal features, contextual features and boundary features for recognition task.

Supervised learning: (Jin et al., 2010) uses rules and Hidden Markov model(HMM) to Tibetan NER.(Jia et al., 2014) combines Maximum Entropy (MaxEnt) and Conditional Random Fields (CRFs) to identify Tibetan person names. (Hua et al., 2015) proposed a Perception Training Model based on Tibetan syllable features to identify Tibetan NER. (Kang et al., 2015) and (zhu et al., 2005) used CRFs to recognize Tibetan person names.

The above approaches based on rules and supervised learning require Tibetan linguists construct rules or native speakers to annotate a lot of training data. (Jia et al., 2014)'s approach based on CRFs and MaxEnt needs 3.5 MB training data; (Hua et al., 2015)'s training data contain 15,000 sentences; (Kang et al., 2015)'s CRF-based method takes 40,000 words as training data.

In this paper, we propose a semi-supervised learning approach to recognize the person name in Tibetan to reduce the human labor and budgets.

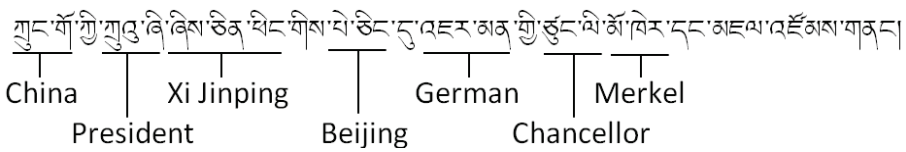

Figure 1: An example of named entities in Tibetan

3. Methodology

There are some often-used semi-supervised learning methods, including: EM with generative mixture models, self-training, co-training, transductive support vector machines, and graph-based methods(Zhu, 2005). And some supervised learning algorithm can also be used in semi-supervised learning algorithm ((Jiao et al., 2006),(Mann and McCallum, 2010),(Liu et al., 2011)). Here, we propose a semi-supervised learning based on Conditional Random Fields (CRFs) and Radial Basis Function (RBF) to recognize person name in Tibetan.

There are two reasons that we adopt CRFs to realize our semi-supervised learning approach. Firstly, CRFs(Lafferty et al., 2001) are a type of discriminative undirected probabilistic graphical model. Because the model is conditional, dependencies among the input variables do not need to be explicitly represented, affording the use of rich, global features of the input (Sutton and McCallum, 2006), CRFs are often applied in named entity recognition (McCallum and Li, 2003) (Settles, 2004). Secondly, we can train CRFs model based on Tibetan syllable, which need not word segmentation of Tibetan.

3.1. Methods

We assume that there are l labelled points $(x_1, y_1), ..., (x_l, y_l)$ and u unlabelled points $x_{l+1}, ..., x_{l+u}$; typically $l << u$. Using L and U as labelled points set and unlabelled points set separately. We suppose the labels are binary.

$$y_L = \begin{cases} 1 & \text{Person names} \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

The semi-supervised learning algorithm based on CRFs and RBF is shown in Algorithm 1.

Algorithm 1:

Given:

L: a small set of labelled training data. There are m annotated person names $PER_j, j \in m$.

U: a lot of unlabelled training data.

Training Model

Step1: train a CRFs model M_L based on L.

Step2: use M_L to classify unlabelled data U and get n labelled person names $MPER_i, i \in n$.

Step3: extract 5-element feature matrixes of PER_j and $MPER_i$, and calculate their similarities based on RBF. Afterwards, select the biggest K similarity values for every annotated entity $MPER_i$ using k-Nearest Neighbors (KNN) algorithm. Then, calculate the mean $Sim(MPER)_i$ of the K similarity values as the similarity of $MPER_i$ to PER_j.

Step4: extract the data which have the most similarity to PER_j and add this data into seeds set L. Meanwhile, remove them from unlabeled data U.

Step5: If the algorithm is converged or the number of loops reached the max iteration, then end this algorithm, else go to step 1.

3.2. Seeds selection

For semi-supervised learning, a small amount of labelled data, which can be called gold seeds, are very important. For person names recognition in Tibetan, in order to ensure the precision and efficiency of model, the gold seeds should cover the important features, such as:

(1) Tibetan person names, transliteration names from Chinese and other foreign countries.

(2) Tibetan person named with titles and case-auxiliary words, Tibetan person named without titles and case-auxiliary words.

(3) Some words which can be used as person name as well as ordinary nouns. For example, "ཉི་མ" (Nyima) can be a Tibetan people name or ordinary nouns "Moon".
The golden seeds in our experiment based on this requirement.

3.3. Feature selection for person name recognition in Tibetan

Seeds propagation is crucial for semi-supervised learning. We use feature matrix of person name in Tibetan to realize the seed propagation. Therefore, we will discuss the feature selection of the person name in Tibetan firstly.
Here, two kinds features are used: initial feature and context feature.
(1) Initial features
The initial features are often used in NER in many languages. For example, capitalization and family names are used in the NER of English and Chinese. Tibetan people names and transliteration of Chines person have different initial features. For Tibetan person names, high-frequency syllables can be initial features. Meanwhile, Chinese Surname can be initial feature for transliteration of Chinese person names.
Using 10,460 (41,755 syllables) Tibetan person names, we select some Tibetan person names that their frequencies are exceed 1% as high-frequency syllables. The top 5 examples are shown in Tab.1.

Tibetan	English
བཀྲ་ཤིས་	Tashi
ཚེ་རིང་	Tsering
བསྟན་འཛིན་	Tenzin
སྒྲོལ་མ་	Dolma
ཉི་མ་	Nyima

Table 1: High-frequency syllables of person names of Tibetan people.

For 504 Hundred Family Surnames, 444 are single-character surnames and 60 are double-character surnames. Because some Chinese Family Surnames have same pronunciation, the 504 Hundred Family Surnames can be translated in 291 tibetan syllables. The example is shown Tab.2.

Tibetan	English
ཝང་	Wang
ལི་	Li
ཛང་	Zhang
ཡན་	Yan
ཨུའུ་	Wu
ཀུང་	Gong

Table 2: Some Chinese surname in Tibetan

(2) Context features: Two features, case-auxiliary word and title, are used in this paper as context features.
Case-auxiliary word is one of the most important components for Tibetan. Among eight kinds of Case-auxiliary, two of them are used as features to recognize Tibetan person name, do-case-auxiliary words and belong-case-auxiliary words, because they are often appeared after or before person name in Tibetan.
do-case-auxiliary words:
གིས་, ཀྱིས་, བྱིས་, འིས་, ཡིས་.
belong-case-auxiliary words:
གི་, ཀྱི་, བྱི་, འི་, ཡི་.
Tile is an important feature for NER task. For person names recognition in Tibetan, two kinds title are used. The first is traditional title such as president, chairman. The other is special titles that are unique for Tibetan. The example is shown Tab.3.
The position of **title** in Tibetan is different from the position in other languages(English, for example) since it can be inserted before or after a person name.
Therefore, in this paper, we use five features to extract Tibetan person name: high-frequency syllables (include high-frequency syllables of Tibetan people's names and Chinese family names); left/right title (the title appear before or after person names); left/right case-auxiliary

Tibetan	English
རྒྱལ་ཕྱག་	President
རྒྱུན་ཁྲིམས་	Chairman
རྒྱལ་ཞབས་	Minister
དབུ་ཚེན་བླ་མ་	Panchen Lama
རིན་དོ་ཆེ་	Rinpoche
སྤྲུལ་སྐུ་	Tulku

Table 3: The example of Tibetan Title

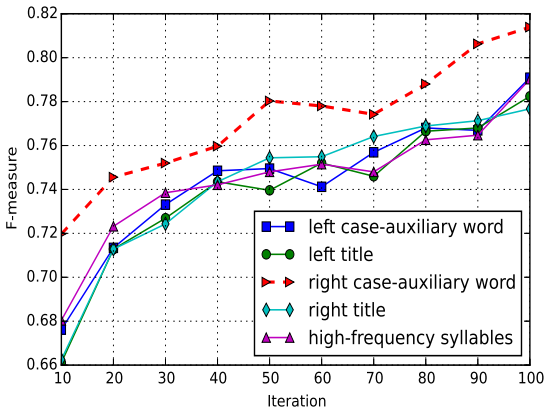

Figure 2: The influence of features to recognize Tibetan person names

word (the case-auxiliary word appear before or after person names).
The influences of five features on person name recognition in Tibetan is shown in Fig.2. We can see that right case-auxiliary word has the biggest influence comparing with the influences of other features.

3.4. Seeds selection

For semi-supervised learning, the annotated data, which can be called gold seeds, are very important. For Tibetan person names recognition task, in order to ensure the precision and efficiency of model, the gold seeds should cover the important features, such as:
(1) Tibetan person names, transliteration names from Chinese and other foreign countries.
(2) Tibetan person named with titles and case-auxiliary words, Tibetan person named without titles and case-auxiliary words.
(3) Some words which can be used as person name as well as ordinary nouns. The golden seeds in our experiment based on this requirement.

3.5. Seed propagation

Using M_L gotten by CRFs, the unlabeled data U can be annotated. Take annotated entities PER_j and new labeled entities $MPER_i$ as nodes V to construct graph $G = (V, E)$. E are edges. We assume an $i * j$ symmetric matrix W on the edges of the graph is given. Then, the similarities of $MPER_i$ andPER_j, w_{ij}, can be calculated by RBF (Zhu et al., 2003)in Formula 2.

$$w_{ij} = exp(-\sum_{d=1}^{5} \frac{\beta_d \cdot (x_{id} - x_{jd})^2}{\sigma^2}) \quad (2)$$

Where x_{jd} and x_{id} is the $d-th$ feature of PER_j and $MPER_i$ respectively.
As shown in Fig.2, the influences of different features on Tibetan person name recognition are different. So, we give 5 features with different weights. β_d is feature weight.
Then, we can calculated $Sim(MPER)_i$ using formula (2) and k-NN graph (k=5)(Zhou et al., 2004) in Formula (3).

$$Sim(MPER_i) = \frac{1}{K} \sum_{x_j} w_{ij}, x_j \in KNN(x_i) \quad (3)$$

4. Evaluation

We used 1100 documents from websites (tibet.people.com.cn, from 2015-2017) as experiment data.
In our experiment, some documents are selected as gold seeds according to the principle of seeds selection. 100 documents are selected as test data. The reminder documents are unlabelled data.

Labelled Data(Documents)	F-measure%
100	45.23
200	59.78
300	66.21
400	73.75
500	75.93
600	78.58
700	82.77
800	84.12
900	86.73
1000	90.31

Table 4: F-measure of Tibetan person name based on CRFs.

4.1. The baseline

We train a baseline model based on (CRF++-0.58) using 1000 annotated documents.Its precision, recall and F-measure are shown in Table 2.
(Kang et al., 2015)'s F-measure is 94.31% of, which use CRFs and some features to recognize Tibetan person names. Our baseline use CRFs and less training data. So, the F-measure (90.31%) of baseline is acceptable.

4.2. The influence of multi-feature

In the process of seed propagation, 5 features were given with different weights. The influence of features on F-measure based on 100 gold seeds are shown in Fig.3. We can see that the performance of multi-feature with different weights is better at least 3% than the performances of other single features.

4.3. The influence of seeds

Fig.4 shows the relationship of F-measure and iterations of 50, 60, 70, 80, 90, 100 seeds. As the number of seeds increases, the F-measure increases. The highest F-measure of Tibetan person names recognition can reach about 84% when 90 or 100 seeds iterate 100 times.
For semi-supervised learning, less seeds means less annotated data and low cost of money and time. Therefore, we can use about 100 documents as golden seeds to extract person name in Tibetan.

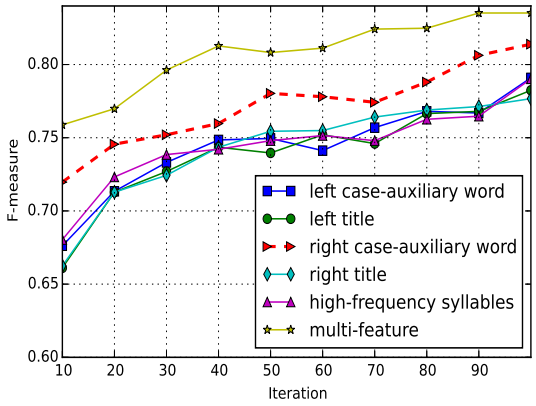

Figure 3: the influence of feature on F-measure

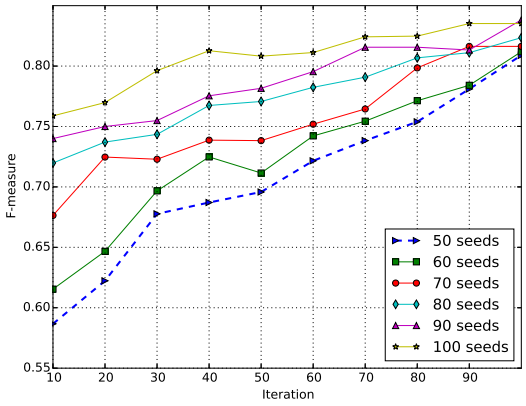

Figure 4: the influence of seeds on F-measure

4.4. The comparatione of semi-supervised learning and supervised learning

Here, we will compare the semi-supervised leaning based on CRFs and RBF to supervised learning method based on CRFs.
As shown in Table 3, the semi-supervised approach achieves much better results than supervised approach when the same amount labelled documents are used. Using only 100 annotated documents, the F-measure of semi-supervised

Labelled data	F-measure
supervised (100 documents)	24.24
supervised (800 documents)	84.12
semi-supervised (100 documents)	83.82

Table 5: The comparison of semi-supervised learning and supervised learning.

learning approach can reach 84%, whereas about more than 800 labelled documents are required for a supervised learning approach based on pure CRFs.

5. Conclusion

For low resource language such as Tibetan, the methods of person name recognition based on supervised learning need a lot of annotated data, which means more human labor, higher budget, and more time. We propose a semi-supervised learning (SSL) approach based on Conditional Random Fields (CRFs) and Radial Basis Function (RBF) to recognize Tibetan person names. And Five feature (high-frequency syllables, left/right title and left/right case-auxiliary words) are used to propagate information from labelled documents to unlabelled data. Experiments demonstrate that this method can recognize person name in Tibetan at low cost with an acceptable performance.

In the future, we will try to construct a common system to extract person name in other low resource language, which based on CRFs, RBF and feature matrix at three level (word level, context level and sentence level). Moreover, we will try to improve the efficiency of propagation.

6. Acknowledgements

This work was supported by the China National Natural Science Foundation No. (61331013).

7. References

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing*, pages 143–153. Springer.

Euisok Chung, Yi-Gyu Hwang, and Myung-Gil Jang. 2003. Korean named entity recognition using hmm and cotraining model. In *Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11*, pages 161–167. Association for Computational Linguistics.

Nadeau David and Sekine Satoshi. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.

Quecairang Hua, Wenbin Jiang, Haixing Zhao, and Qun Liu. 2015. Tibetan name entity recognition with perceptron model. *Computer Engineering and Application*, 50(15):172–176.

Yangji Jia, Yachao Li, Chengqing Zong, and Hongzhi Yu. 2014. A hybrid approach to tibetan person name identification by maximum entropy model and conditional random fields. *Journal of Chinese Information Processing*, 28(1):107–112.

Feng Jiao, Shaojun Wang, Chi-Hoon Lee, Russell Greiner, and Dale Schuurmans. 2006. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 209–216. Association for Computational Linguistics.

Ming Jin, Huanhuan Yang, and Guangrong Shan. 2010. The studies of named entity recognition for tibetan. *Journal of Northwest University for Nationalities(Natural Science)*, 31(3):49–52.

Caijun Kang, Cunjun Long, and Di Jiang. 2015. Tibetan name recognition research based on crf. *Computer Engineering and Application*, 51(3):109–111.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual*

Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 359–367. Association for Computational Linguistics.

Gideon S Mann and Andrew McCallum. 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *The Journal of Machine Learning Research*, 11:955–984.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

Borislav Popov, Angel Kirilov, Diana Maynard, and Dimitar Manov. 2004. Creation of reusable components and language resources for named entity recognition in russian. In *LREC*.

Gökhan Akin Seker and Gülsen Eryigit. 2012. Initial explorations on using crfs for turkish named entity recognition. In *COLING*, pages 2459–2474.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics.

Yuan Sun, Xiaodong Yan, Xiaobing Zhao, and Guosheng Yang. 2010. Reseach on automatic recognition of tibetan personal names based on multi-features. In *2010 international conference of natural language processing and knowledge engineering*.

Charles Sutton and Andrew McCallum. 2006. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, pages 93–128.

Hongzhi Yu and Ning Jiang, Tao anf Ma. 2010. Named entity recognition for tibetan texts using case-auxiliary grammars. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with local and global consistency. *Advances in neural information processing systems*, 16(16):321–328.

Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pages 912–919.

Xiaojin Zhu. 2005. Semi-supervised learning literature survey.

OrenEtzioni Michael Cafarella and etc. 2005. Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 175(1):91–134.

Nadeau, David. 2007. Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision. University of Ottawa.

Zhu Jie,Li Tianrui,Liu Shengjiu. 2016. Research on Tibetan name recognition technology under CRF *Journal of Nanjing University(Natural Sciences)* , 52(2):289–299.