Active Learning for Tibetan Named Entity Recognition based on CRF

Fei-Fei Liu, Zhi-Juan Wang*
Minzu University of China, National Language Resource Monitoring & Research Center of Minority Languges ; Minzu University of China National Language Resource Monitoring & Research Center of Minority Languges
Beijing China, Beijing China,
Liufeifei_muc@163. com, wangzj. muc@gmail. com

Abstract
Named entity recognition (NER) is a major subtask of information extraction. Previous research tent to use huge amount of labeled data to train a classifier. But it is expensive for low resource languages One of the dominant problems facing Tibetan named entity recognition is the lack of training data. Active learning is a supervised machine learning algorithm which can achieve greater accuracy with fewer training labels. Active learning has been successfully applied to a number of natural language processing tasks, such as, information extraction, named entity recognition, text categorization, part-of-speech tagging, parsing, and word sense disambiguation. In this paper, we apply active learning based on Conditional Random Field (CRF) for Tibetan named entity recognition to minimize labeling effort by selecting the most informative instances to label. This paper proposes two kinds of query strategies, including Confidence, and Named Entity features. We compare the query strategies with the random method, and show that considerable performance improvements in reduce the human effort.

Keywords: Active learning, Tibetan Named Entity Recognition, Query Strategy,CRF

1. Introduction

Named entity recognition (NER) is one of the most elementary and core problems in natural language processing (NLP). There are supervised learning (SL), semi-supervised learning (SSL), unsupervised learning(UL) for named entity recognition. At present, supervised machine-learning methods in the task are in the leading position, such as Hidden Markov Model (HMM), Conditional Random Field (CRF), Support Vector Machine (SVM). The obstacle of supervised machine-learning methods is the great requirement of the annotated training data which is essential for achieving good performance. Building a high quality annotated corpora by hand is time-consuming and expensive. Because of the lack of corpora and person who understand those languages, creating training corpora for resource-scarce languages is particularly expensive.

Nowadays, named entity recognition had achieved good results in various languages, such as English. State-of-the-art NER systems for English produce nearhuman performance. For example, the best system entering MUC [1] -7. scored 93.39% of F-measure while human annotators scored 97.60% and 96.95%. However, Tibetan NER started late. It has yielded a great number of positive results， but it is still a new study field in which there are series of problems, such as the conflicts between Tibetan names and ordinary words, the misinterpretation of translations, and the difficulties in identifying Tibetan NE boundaries. The biggest reason for these problems is the lack of high quality annotated corpora.

There is a way, active learning, to solve this problem. Active learning is a special case of semi-supervised machine learning in which a learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points(Settles B, 2010 ; Rubens N et al., 2015). We can use it to select the most informative samples for training. In this way, we will undoubtedly enhance the performance under cutting annotating cost.

[1] Message Understanding Conference

Many existing researches show that active learning can be effectively reduce the training quantity in NLP task(Olsson F, 2009), such as information extraction(Culotta A et al., 2006), text categorization (Dasgupta S and Hsu D , 2006), named entity recognition(Kim S et al., 2006 ; Tomanek K and Hahn U, 2009 ; Ekbal A et al., 2016 ; Saha S, 2012). However, the majority of this literature focuses on biomedical domain or the official languages of a certain country. Those research showed that active learning can select the useful data from a huge pool of unlabeled documents. We can use this method in Tibetan named entity recognition.

In this paper, we propose an alternative active learning strategy for the Tibetan NER task. Without large-scale labeled data, the proposed method greatly reduces the training time and annotating cost. Two methods are presented, the first method is based on the confidence, the second is mixed the tags information.

There are two kinds of query strategies to Tibetan named entity recognition. One is by the degree of confidence measure, we choose the lowest part of the degree as it hard to process. Another would calculate uncertainty for the tag. Last, we get the most likely annotation results, for each result, the confidence scores can be calculated. And the uncertainty is quantified by the difference of confidence.

The organization of the paper is as follows. Following the introduction in Section 1. Section 2 presents the background ,including Tibetan named entity recognition and active learing. Section 3 presents active learing for Tibetan NER based on CRF. Section 4 shows oue experiment and discussion. Finally, we summarize in section 5.

2. Back ground

2.1 Tibetan Tibetan Named Entity Recognition

2.1.1 Introduction to Tibetan

Tibetan(ཟང་ཡིག) refers to the use of Tibetan language Tibetan. The glyph structure is a letter as the core, the rest of the letters are based on this before and after the additional and overlapping from top to bottom, combined into a complete word table structure. Writing habits from

left to right. The font is divided into "head" and "headless" two categories.

Tibetan is a phonetic alphabet, with 30 consonants and 4 vowels. One Tibetan syllable can have 1 to 7 basic characters, if you consider Sanskrit, characters may be more. The seven basic characters have a base character and a vowel , the other characters were added to the base word, the up, down, front, back, and then back.

There are fewer types of punctuation in Tibetan . Tibetan various syllables separate with a small point, this point named the syllable node (·). In addition to the syllable node, the most common punctuation is a single vertical line (།), as a full stop, colon and other situations. And the paragraph ends with a double vertical line (།།).

2.1.2 Methods of Tibetan NER

The methods of Tibetan NER can be divided into rule-based methods and based on supervised machine learning methods.

Rule-based methods

In the early days, the study of Tibetan NER was based on a rule-based approach. Yu et al. used a rule-based model based on case-auxiliary word and lexicon, and also adapt boundary information list static from large corpus to improve recognition(Yu HZ et al.,2010). And experiments shows that recall rate and precision are respectively 90.13% and 94.02% in the newspaper corpus, 85.67% and 88.20% in the website text. Sun et al. used the internal features of names, contextual features and boundary features of names, and establishes the dictionary and feature base of Tibetan names(Sun Y et al.,2010). The results prove the algorithm is effective with 0.8391 F-score. Dou et al. used the Statistical Method of Mutual Information to, combining the rules of lattice auxiliary and the dictionary of person names, F value in the test can be up to 93.55%(Dou R et al.,2010).

Supervised machine learning methods

After 2014, supervised machine learning methods are increasingly applied to Tibetan NER. Jia et al. came up with Maximum entropy(ME) and conditional random field(CRF), and the F-score of the recognition of names can be 92.08%(Jia et al.,2014). Hua et al. proposed a syllable features with Perceptron training model to identify Tibetan name entity with detail analysis NE structure rule and word segmentation ambiguity(Hua et al.,2014). The F-score of NE identification is 86.03% for the test set. Kang et al. defined a feature tag set to fit in with the characters of Tibetan names, used CRF as tagging model to train and test corpus data(Kang et al.,2015). The highest F-score obtained in the experiment can reach 94.31%. Zhu et al. studied Tibetan name recognition technology using conditional random fields (CRF) principle， focuses on analysis of the internal structure of the Tibetan names, contextual features, feature selection and data preprocessing, etc. and evaluated the effectiveness of different features through experiments(Zhu et al.,2016). The recognition rate of Tibetan names can reach 80% of F-score.

2.1.3 Difficulties in Tibetan NER

Tibetan belongs to the Sino-Tibetan language family. In theory, the natural language processing methods used in Chinese can be used in Tibetan information processing. But in practice, it must be considered in the specific problems. The main difficulties in Tibetan NER are as follows:

Tibetan is a complex system of phonetic logic. The basic unit of the sentence is syllable. Syllables are separated by syllable node. One syllable or more syllables constitute words. There is no obvious mark between the word and next word. The boundaries of named entities are difficult to determine. And too few punctuation types, just single vertical line (།) and double vertical line (།།), will make the too long analysis object length, increasing the difficulty of recognition algorithm.

There is no morphological difference between named entities and unnamed entities in Tibetan. Unlike English, the person names, location names and organization names in English with the capitalized first letter, are easy to extract. And compared to Chinese person name, most of the Tibetans do not have the family name and the length of the name which can be from single syllable to twenty-six syllables.

The name dictionary, the labeled corpus and other related resources is insufficient. Nowadays, the main method of Tibetan Named Entity Recognition is supervised learning algorithms which require large-scale of labeled corpus. But Tibetan resource is not easy to obtain.

The biggest reason for these difficulties is the lack of Tibetan labeled corpus. We propose active learning to solve the problem.

2.2 Active Learning

In traditional supervised machine-learning, unlabeled data is selected for annotation at random under the huge amount of labeled data demand. Differently, the most useful data for the classifier are seriously selected in active learning.

2.2.1 Active Learning Examples

A learner may begin with a small number of instances in the labeled training set L, request labels for one or more carefully selected instances, learn from the query results, and then leverage its new knowledge to choose which instances to query next (Settles B, 2010). There is a Fig. 1 to indicate the typical pattern.

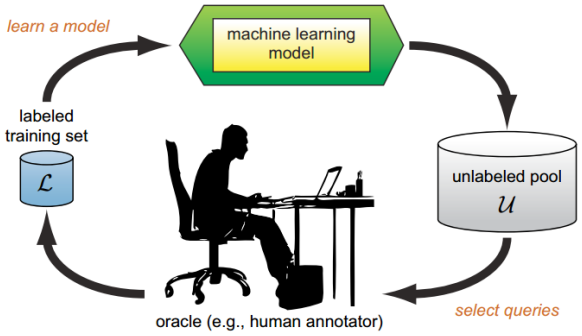

Figure 1: The typical pattern of active learning

And our frame of named entity recognition is by the following procedure. L is the labeled training set; U is the unlabeled data set; Q is the query strategy; C is the classifier for named entity recognition, in our work; N is the number of iterations.

Input：L, U, C, Q, N

Begin

For i from 1 to N

M=Train(C, L)

/ Train classifier on L, get model M*/*

T=Test(C,M,U)

*/*with M，test U by C */*

T'=Select (Q, U|T)

*/*select useful by Q*/*

Label (T')

*/*query the human annotator for labeling*/*

L=L+ T';

*/*Add T' to L */*

U=U- T'; / Delete T' from U*/*

END

2.2.2 Query Strategy

There have been many proposed ways of formulating such query strategies.

Uncertainty Sampling

Perhaps the simplest and most commonly used query framework is uncertainty sampling (Lewis and Gale, 1994; Settles B, 2010). In this method, system queries the sentences about which it is least certain how to label the corpus, the criterion for the least confident strategy only considers information about the most probable label. It is straightforward and entropy is often used as an uncertainty measure.

Query-By-Committee

Query-by-committee (QBC) algorithm (Seung et al., 1992) as the more theoretically-motivated query selection framework is a good way to minimize the vision space. Each committee member is then allowed to vote on the labelings of query candidates. The most informative query is considered to be the instance about which they most disagree(Settles B, 2010).

Both of the above options are usual. There are other query strategies, for example, Expected Model Change, Expected Error Reduction, Variance Reduction, Density-Weighted Methods, etc.

In this paper, we propose two kinds of query strategies baded on the uncertainty sampling, including Confidence and Named Entity features. These query strategies are described in more detail in the subsequent sections.

3. Active Learning for Tibetan Named Entity Recognition based on CRF

3.1 Conditional Random Field

CRFs are a type of discriminative undirected probabilistic graphical model. It is used to encode known relationships between observations and construct consistent interpretations. It is often used for labeling or parsing of sequential data, such as natural language text. Specifically, CRFs find applications in shallow parsing, named entity recognition.

Lafferty, McCallum and Pereira define a CRF on observations X and random variables Y as follows:

Let G = (V , E) be a graph such that $Y=(Y_v)$ $v \in V$, so that Y is indexed by the vertices of G. Then (X,Y) is a conditional random field when the random variables Y_v , conditioned on X, obey the Markov property with respect to the graph: $p(Y_v |X, Y_w , w \neq v)=p(Y_v |X, Y_w , w\sim v)$, where w~v means that w and v are neighbors in G.

What this means is that a CRF is an undirected graphical model whose nodes can be divided into exactly two disjoint sets X and Y, the observed and output variables, respectively; the conditional distribution p(Y|X) is then modeled.

By now, CRF has become a widely used technique which is applied in named entity recognition on low resource language, such as Hindi, Bengali, Tamil, and Telugu.

3.2 Tibetan Named Entity Recognition based on CRF

Tibetan NER can be defined as a sequence labeling problem for determining whether a observations belongs to a labeled set of markers. Suppose that a given marker sequence $y= (y_1, y_2 , \ldots, y_n)$ is labeled, n is the length of the sequence. The sequence of Tibetan NE is represented as $w= (w_1, w_2, \cdots, w_m)$, m is the length of the NE. The model of CRF is defined as follows:

$$p(y|w) = \frac{1}{Z(w)} \exp\left(\sum_i \sum_k \lambda_k f_k(y_i, y_{i-1}, w)\right)$$

Z(w) is normalization factor, determined by the observation sequence.

$$Z(w) = \sum_y \exp\left(\sum_k \lambda_k f_k(y_i, y_{i-1}, w)\right)$$

λ_k is the weight of the k-th function, f_k $(y_i, y_{(i-1)}, w)$ is a characteristic function.

$$f_k(y_i, y_{i-1}, w) = \begin{cases} 1, \text{if } y_i = u \text{ and } y_{i-1} = v \\ 0, \text{otherwise} \end{cases}$$

3.3 Active Learning for Tibetan Named Entity Recognition

To solve the lack of Tibetan traing data,we present two kinds of query strategies in active learing .

3.3.1 Query Strategy based on Confidence

In **confidence**, we believe that the lower the confidence score of the sentence, the more difficult to identify for the classifier. This kind sentences need to manually participate in the annotation. And the confidence score can be calculated by Conditional probability. Give an input sequence x, in the situation that we have gotten the module, the P(y|x) is the Conditional probability that x corresponds to the tag sequence y. This probability can be regarded as confidence measure.

By using the equation for CRF (Lafferty et al. 2001) module, we can calculate the probability of any possible state sequence s given an input sequence. It is defined to be:

$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \varphi_t (y_t, y_{t-1}, x_t)$$

$$\varphi_t(y_t, y_{t-1}, x_t) = \exp\left\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right\}$$

To get the best sequence, we used the Viterbi algorithm, a dynamic programming algorithm for finding the most likely sequence of hidden states . So, the confidence of the input sequence x we used is defined as :

$$\textbf{Confidence}(\mathbf{x}) = \text{argmax } p(y|x)$$

3.3.2 Query Strategy based on NE features

NE features means Named Entity features. The main features for the NER task are identified based on the different possible combinations of available word and tag contexts. We use the following set of specific features, which is conducive to the improvement of the Tibetan named entity recognition performance.

a. Feature of tibetan person names

Tibetan person names could be divided into three catalogues: translation names and common names. The translation has special syllables and common tibetan person name has frequently used syllables. We collect 237 han surname as the feature of translation names. In common tibetan person name, we calculated the frequency based on 10460 Tibetan person name, and selected the top 97 as High frequency syllables .

In Tibetan, many words can indicate the boundary of names, such as ཀྲུའུ་ཞི། (chairman), དགེ་རྒན(teacher), ཟླ་མ(lamaism). 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.

b. Feature of Tibetan location names

Location names usually has particular syllables, such as རྫོང་(county), རི།(mountain). We collect 20 words as the feature of Tibetan location names.

c. Feature of Tibetan orgnization names

The feature of orgnization names and location names is practically identical. We collect 24 words as the feature of Tibetan orgnization names, inculding སློབ་གྲྭ(school), དབུལ་ཁང་།(bank).

4. Experiment and Discussion

4.1 Experiment design

In iterative development cycles, we select Top 10 sentences in each iteration. We test three different active learning methods: Random selection, Confidence-based Query Strategy, NE feature-based Query Strategy.

The result of random selection is the baseline in our experiment.

We use F1 to evaluate the performance of each graininess, which are very common in NLP evaluation.

P= (number of correctly identified NE)/(number of identified NE)

R= (number of correctly identified NE)/(number of all NE)

F1=(2*P*R)/(P+R)

4.2 Experiment data

We conducted our active learning experiments under Tibetan language. For our empirical evaluation, we used the training data and test data from four sites, include People's Network （Tibetan version）, Aba News Network, Tibet News Network, The Voice of America(Tibetan version). We marked the person name(PER), location names(LOC) and orgnization names (ORG) with a part of data,as labelled train data set and test data set, the remain corpus as unlabeled data set. There are about 7,000 sentences. Some statistics of training, development and test data are presented in Table 1.

	sentences	PER	LOC	ORG
Labelled train data set	249	231	164	76
Unlabeled data set	7269	-	-	-
Test data set	246	112	147	165

Table 1: Data source

4.3 Results and Analyses

The initial NER module gets an F-score of 10.7, while the train set contains only 249 sentences. We plotted the learning curves for the different query strategies.

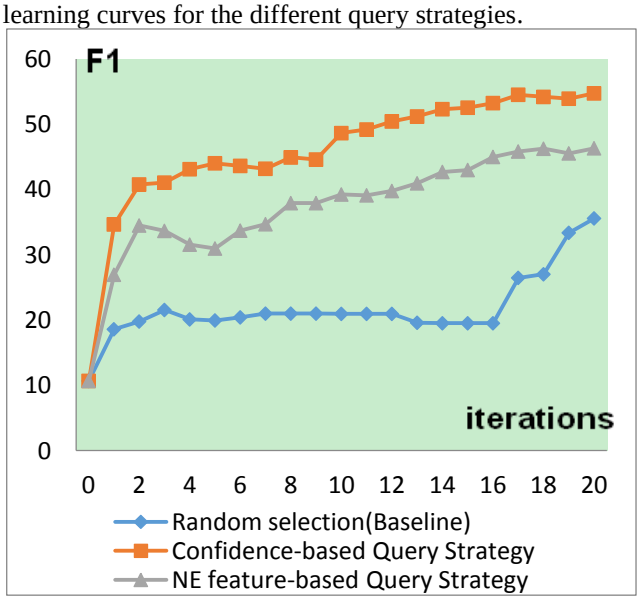

Figure 2: Comparison of active methods

The curves in Figure 2 show the relative performance. The F1 increases along with the number of selected sentences. The results suggest that both active learning methods consistently outperform the random selection.

The amplitude of variation in Random selection is irregular. After 16 iterations, the F1 is increasing. It says it is impossible to be sure about the value of the sentence we choose in Random selection.

The confidence-based query strategy has improved performance after each iteration. By comparison test, the strategy is better than random selection. The F1 has shot up by far more than the other methods. Under the same iterative numbers, the F1 increases from 10.7 to more than 54. The effect for the first two iterations is notable. After iterations, F1 is higher than the Random selection.

The NE feature-based query strategy also shows better result than Random selection. Although it is not as good as the confidence-based query strategy. Its dominance looks shaky. We think the reason for this is that named entity features we have collected are not enough, and there are still some exceptional circumstances.

5. Conclusion

Nowadays, the biggest cause for Tibetan Named Entity Recognition is the lack of training data. Because of the high cost and long time-consuming of tagging data, to get a lot of labeled data has been very difficult and expencive, and on the other hand, it is relatively easy to get a lot of unlabeled data. In this paper, we use active learning based on CRF to select useful data from a large number of unlabled corpus.The experiment shows that we can achieve better F1 by our mothods in the same iterative. We compared different active learning algorithms for Tibetan named entity recognition. Our results showed that active learning algorithms considerable performance improvements in reduced savings of annotation. In future research , we will investigate some new query strategies to get better effect.

6. Acknowledgement

The project was supported by Key Program of National Natural Science Foundation of China (Grant No. 61331013), Minzu University of China Tier-1 Universities and Tier-1 Subjects Graduate independent research projects (Grant No. 10301-0170040601-184).

7. Bibliographical References

Arkin M, Mahmut A, Hamdulla A. Person Name Recognition for Uyghur Using Condi-tional Random Fields[J]. International Journal of Computer Science Issues, 2013.

Chen Y, Lasko T A, Mei Q, et al. A study of active learning methods for named entity rec-ognition in clinical text[J]. Journal of biomedical informatics, 2015, 58: 11-18.

Culotta A, Kristjansson T, Mccallum A, et al. Corrective feedback and persistent learning for information extraction[J]. Artificial Intelligence, 2006, 170(14):1101-1122.

Cui B, Lin H, Yang Z. Uncertainty sampling-based active learning for protein–protein in-teraction extraction from biomedical literature[J]. Expert systems with Applications, 2009, 36(7): 10344-10350.

Dasgupta S, Hsu D. Hierarchical sampling for active learning[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 208-215.

Dou R, Jia YJ, Huang W.Automatic recognition of tibetan name with the combination of statistics and regular. Journal of Changchun Institute of Technology (Social Science Edition), 11 (2) 113-115.,2010.2:113-115.

Ekbal A, Saha S, Sikdar U K. On active annotation for named entity recognition[J]. Inter-national Journal of Machine Learning and Cybernetics, 2016, 7(4): 623-640.

Hua Q, Jiang W, Zhao H, et al. Tibetan name entity recognition with perceptron model[J]. Computer Engineering & Applications, 2014, 50(15):172-176.

Jia Y, Li Y, Zong C, et al. A Hybrid Approach Using Maximum Entropy Model and Con-ditional Random Fields to Identify Tibetan Person Names[J]. Himalayan Linguistics, 2016, 15(1).

Jia YJ, Yachao L I, Zong C, et al. A Hybrid Approach to Tibetan Person Name Identification by Maximum Entropy Model and Conditional Random Fields[J]. Journal of Chinese Information Processing, 2014.

Kang CJ, Long C, Jiang D. Tibetan names recognition research based on CRF[J]. Computer Engineering and Applications, 2015.

Kim S, Song Y, Kim K, et al. Mmr-based active machine learning for bio named entity recognition[C] // Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. Association for Computational Linguistics, 2006: 69-72.

Mo H M, Nwet K T, Soe K M. CRF-Based Named Entity Recognition for Myanmar Lan-guage[J]. 2016.

Olsson F. A literature survey of active machine learning in the context of natural language processing[J]. 2009.

Rubens N, Elahi M, Sugiyama M, et al. Active learning in recommender sys-tems[M]//Recommender systems handbook. Springer US, 2015: 809-846.

Saha S, Ekbal A, Verma M, et al. Active learning technique for biomedical named entity extraction[C]//Proceedings of the International Conference on Advances in Computing, Communications and Informatics. ACM, 2012: 835-841.

Settles B. Active learning literature survey[J]. University of Wisconsin, Madison, 2010, 52(55-66): 11.

Sun Y, Yan X, Zhao X, et al. Research on automatic recognition of Tibetan personal names based on multi-features[C]// International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2010:1-5.

Tomanek K, Hahn U. Reducing class imbalance during active learning for named entity annotation[C]// Proceedings of the fifth international conference on Knowledge capture. ACM, 2009: 105-112.

Tomanek K, Laws F, Hahn U, et al. On proper unit selection in active learning: co-selection effects for named entity recognition[C]//Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing. Association for Computational Linguistics, 2009: 9-17.

Yao L, Sun C, Li S, et al. CRF-based active learning for Chinese named entity recogni-tion[C]//Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on. IEEE, 2009: 1557-1561.

Yu HZ, Jiang T, Ma N.Named Entity Recognition for Tibetan Texts[J].Lecture Notes in Engineering and Computer Science,2010,2180.

Zhu J, Li T, Liu S. Research on Tibetan name recognition technology under CRF[J]. Journal of Nanjing University, 2016.