Tibetan Predicate Recognition Based on A Semi-supervised Model

Lin Li, Weina Zhao, Zewangkuanzhuo

Qinghai Normal University, Xining, China

Keywords: Tibetan Predicate Recognition, Semi-supervised model, CRFs, Word Embedding.

Abstract: The syntactic structure of a Tibetan sentence is largely determined by its predicate, thus whose recognition provides the basis for complete syntactic parsing and other NLP tasks. Tibetan predicate can be divided into two categories: verbal predicate and adjective predicate, both of which consist of a headword and other varied components. As a result, we focus on recognizing Tibetan predicate in this work. Previous approaches for this task often adopt rule-based or supervised machine learning methods. In this paper, we present a semi-supervised Tibetan predicate recognition model that adopts a Tibetan word embedding. Our semi-supervised model applies an unsupervised word embedding as extra-features into a supervised predicate recognition model. The strength of our model is that it uses a pre-trained word embedding and thus minimizes the need for prior knowledge. We use a near state-of-the-art baseline system that based on Conditional Random Fields (CRFs). On this base, we build up a supervised system combined with semantic features. Then, with a large-scale Tibetan corpus, we induce several Tibetan word embeddings. We evaluate these word embeddings on predicate recognition task. The results show varying degree improvements by using word embedding as features. And the F score of the semi-supervised reaches 88.58%, that is, the F score improves 7.92% and 3.10% compared with the baseline system and the supervised system respectively.

1. Introduction

The predicate is the core of a sentence, which reveals significant syntactic and semantic information. In Tibetan, the predicate is at the end of a sentence in most cases and composed of varying components [1]. The verbal predicate consists of a verb, an auxiliary verb, and an appearance-indication marker in the following example (cf. s1).

Tibetan ransliteration	ngas brog pa i nang la o ja [thung myong yod]{Predicate}	s1
English	I at a herdsmen's home milk tea [once drank]{Predicate}. I once drank milk tea at a herdsmen's home.	

Tibetan predicate identification and its related research have attracted increasingly research interests because the significance role of predicate recognition in Tibetan information process. By analyzing the structure of Tibetan predicate, two rule-based approaches are proposed to recognize verbal predicate [1] and adjective predicate [2]. A rule-based model [3] has been built up to identify the high frequency linking verb and existential verb in Tibetan. A semi-supervised approach has been presented to extract Tibetan trisyllabic verb phrase [4]. Tibetan phrase classification system has been systematically studies [5], and also comprehensively analyzes the verbal predicate part of speech tagging purpose. On this base, the linear CRFs model has performed well in Tibetan functional boundary identification [6-7]. These previous works study mainly adopt knowledge-based or supervised machine learning strategy, both of which rely on effective priori knowledge.

Semi-supervised approaches improve generalization accuracy by applying unlabeled dataset into labeled dataset in order to alleviate data sparsity. Unsupervised word representations have been used in English chunking task and the results proves that chunking models [8-9] can perform well with a small scale of training data when use a pre-trained word presentation.

First of all, a baseline system has been built up, which uses a simple and effective feature template [10]. Then, we build up a supervised system with a feature set combined with semantic information. A couple of word embeddings have been induced with different methods and

Copyright © (2019) Francis Academic Press, UK DOI: 10.25236/iwmecs.2019.021

parameters on large scale Tibetan corpus. In this paper, we present a semi-supervised Tibetan predicate recognition model and compare it with the baseline system and the supervised machine learning system. These word embeddings are used as feature to extend the baseline system. The results have demonstrated that the semi-supervised model performs better than the other two systems in our task.

2. Tibetan Predicate Structure

The basic word order of a Tibetan sentence is subjective objective and predicate (SOV), that is, the sentence ends with predicate. According to the part-of-speech of headword, Tibetan predicate can be divided into two types: verbal predicate (cf. sentence 2) and adjective predicate (cf. sentence 3).

Tibetan ransliteration	khyod kyis sgor sil ma [phyir bzlog mi dgos]{Predicate}/	s2
English	you the change [return not need]{Predicate} You can keep the change.	
Tibetan ransliteration	byang sne'i gnam gshis ha cang [grang]{Predicate}]/	s3
English	North Pole's weather very [cold]{Predicate} The weather in North Pole is very cold.	

Link and existential verbs are also very widely used in Tibetan, for instance, a judgment sentence (cf. s4), an existential sentence (s5), and a descriptive sentence (cf. s6). The predicates of above sentences are the link verb རེད་, the existential verb ཡོད་ and འདུག. The structure of these sentences is relatively simple because these verbs are not modified by other words normally.

Tibetan ransliteration	di bkra shis lags kyi gzim khang [red pas]{Predicate}]/	s4
English	This Tashi's bedroom [is]{Predicate}? Is this Tashi's bedroom?	
Tibetan ransliteration	kho la ba glang stong gnyis [yod]{Predicate}]/	s5
English	he cattle two thousand [has]{Predicate} He has two thousand cattle.	
Tibetan ransliteration	spus tshad ha cang ag po [dug] {Predicate}	s6
English	Very good quality [has]{Predicate} Has very good quality.	

The structure of Tibetan predicate is complex in many cases, which brings challenge to predicate recognition task. Tibetan predicate comprises of not only a headword but also other components [11] like mood words, adverbs, verb auxiliary, and appearance-indication marker, etc. Correspondingly, the predicate represents plentiful semantic meanings such as tense, mode, intentionality, and so forth.

The basic components of a Tibetan verbal predicate include verb, appearance indication marker, mood word, and adverb, etc. The comprehensive structure of Tibetan predicative phrase can be described as [1], which shows a general elements and order of a Tibetan predicate.

$$\text{Tibetan predicative phrase}=(\text{adverb})+\text{verb}+(\text{verb auxiliary})+(\text{appearance-indication marker})+(\text{mood word}) \quad (1)$$

3. Tibetan Word Embedding

Distributed word representation also called word embedding so far has achieved improvement in many NLP tasks. The shortage of traditional one-hot encoding is data sparse and semantic relation lost. Word embedding solves these problems by mapping words into dense, low-dimensional, and

continuous-valued vectors. Each dimension of word embedding contains the feature of a word and captures syntactic and semantic properties.

Word embedding originates in neural language model [12], which takes neural network as underlying predictive model [13]. The main idea of word embedding is to predict the occurrence of a word by its neighboring context. Word embedding inducing is an unsupervised process by which a word is represented as a k dimensional real number vector [14].

In this work we adopt two algorithms to induce Tibetan word embeddings, one is the continuous bag of words model (CBOW) and the other continuous skip-gram model (SKIP-GRAM) [15]. The training objective of both the CBOW and the SKIP-GRAM is to maximize the probability of P(C) in a given corpus.

where C denotes a token set of a given corpus, and context(w) is the local context of word w. The P(context(w)) aims to capture the relationship between context and word w, either by predicting w according to its context words, or predicting context words by w.

To a given corpus, the P(C) maximizing can be acquired by maximum likelihood estimate. As such the maximization of P(C) becomes:

when L is maximized and the P(C) is maximized as well.

The softmax is used to predict word occurrence for both CBOW and SKIP-GRAM:

where $v_{context(w)}$is the distributed representation of the local context, v_w is the distributed representation of word w, and V is the vocabulary of a given corpus which deletes the low-frequency words of C. Currently, hierarchical softmax and negative sampling to scale up are applied to word embedding inducing to reduce the computing complicated in large corpus.

In this paper, we introduce Tibetan word embedding into our recognition task. To evaluate the effect of word embedding, we acquire several word embeddings by adopting different algorithms and hyper-parameters and training on a fixed large-scale unlabeled corpus. Two key hyper-parameters of word embedding inducing are word embedding dimensionality and context window size. We train Tibetan word embeddings by word2vec [20] on the combined parameters:

Algorithm, $a \in \{CBOW, SKIP\text{-}GRAM\}$

Word embedding dimensionality, $d \in \{50,100,150\}$

Context window size, $m \in \{5,10\}$

4. Corpus

4.1 Corpus for Word Embedding Inducing

An unlabeled Tibetan corpus is used to induce word embedding, which contains Tibetan news, blogs, scripts, and bulletins. The corpus contains a large quantity of non-Tibetan characters because of the various sources. We adopt a syllable based POS tagging system [16] firstly. After word segmentation, the corpus contains about 114 million words and 7.3 million sentences.

The previous work [17] shows that the word embedding performs better when it is trained on a clean unlabeled corpus. Thus, we use a preprocess strategy to remove all sentences that are only composed of Arabic numerals, Chinese, or English characters. After this step, we acquire a clean corpus to induce Tibetan word embedding, which contain 61.8 million words and 3.9 million sentences.

4.2 Corpus for Predicate Recognition Model Training

A corpus with predicate annotation is essential for the supervised task, thus we build up a manually labeled corpus that comprises of 5401 sentences. These sentences are chosen from Tibetan textbooks, screenplays, news and stories. The first step is part-of-speech tagging [17], then we annotate predicate by native speakers. According to the predicate description in section 3, we annotate 5336 predicates in the corpus.

5. Tibetan Predicate Recognition Models

5.1 Conditional Random Fields, CRFs

A predicate recognition model based on CRFs has been built up as a baseline system. CRFs is a conditional probability model, which overcomes labeling bias problem and performs well in many NLP tasks.

The main idea of CRFs is that given a input sequence $X=(x_1, x_2,..., x_n)$, CRFs define conditional probability distribution $p(Y|X)$ of a label sequence $Y=(y_1, y_2,..., y_n)$ as

where

$$P_\lambda(Y \mid X) = \frac{exp(\lambda \cdot F(Y, X)}{Z_\lambda(X)}$$

$$Z_\lambda(X) = \sum_y exp\lambda \cdot F(y, x)$$

$$F(y, x) = \sum_{i=1}^{n} f(y, x, i)$$

is a normalization factor, F(Y,X) is a global vector, and f is a local feature vector, is the corresponding weight vector of f.

Consequently, the sequential label question turns into seeking for the optimal label sequence Y. The most probable label sequence for the input sequence X is

The key of CRFs model is feature extraction that directly influences recognition accuracy. Normally, an appropriate feature set plays significant role on recognition model. In this work, we choose a standard near-state-of-the-art feature template for the baseline system [10].

5.2 Recognition Models

Tibetan predicate recognition is solved as a sequential labeling task. Thus, we choose the outstanding sequence label model - CRFs model [18]. In this paper, three models are built up: a baseline system, a supervised model with semantic information, and a semi-supervised model with word embedding.

The baseline system just makes use of word and POS tag information. To improve the model, we integrate semantic information into the baseline template. The extended features are the syllable number of predicate, the headword of predicate, and the headword position [19] of predicate. To reduce the dependence of priori knowledge, we build up a semi-supervised model that take word embedding as a uniform feature into the CRFs model.

6. Experiments and Results

Table 1 shows the final recognition results of Tibetan predicate recognition. The results show that the recognition results can be improved by taking word embedding as features. Furthermore, different types of word embedding influences recognition accuracy.

Table 1. Tibetan Predicate Recognition Results.

ID	System	P/%	R/%	F/%
1	Baseline	83.72%	77.81%	80.66%
2	Baseline + Semantic feature	86.24%	84.72%	85.47%
3	Baseline+CBOW-50	88.80%	85.80%	87.27%
4	Baseline + SKIPGRAM-50	87.33%	87.87%	87.60%
5	**Baseline + CBOW-100**	**89.19%**	**87.97%**	**88.58%**
6	Baseline + SKIPGRAM-100	88.12%	88.60%	88.36%
7	Baseline + CBOW-150	88.19%	87.79%	87.99%
8	Baseline + SKIPGRAM-150-5	79.20%	89.80%	84.17%
9	**Baseline + Multi-feature set + CBOW-100**	**91.93%**	**88.96%**	**90.42%**
10	Baseline + Multi-feature set + SKIPGRAM-100	88.82%	88.26%	88.54%

F score of baseline system (experiment) reaches 80.66%, which proves that CRFs with simple information is an effective model for predication identifying. The F score of experiment 2 has been improved 4.81% by adding semantic information into the supervised CRFs model. It suggests that priori knowledge is an effective way to optimize the supervised machine-learning model. Compared with the result of baseline system, the F score has an obvious increment - 7.92% in experiment 5. Results suggest that word embeddings induced by CBOW perform better than SKIP-GRAM in our work. Furthermore, the model performs the best when we use the appropriate dimensionality 100. Compared with the result of semi-supervised model (experiment 5) and the supervised model (experiment 2), we find that the F score of semi-supervised model is higher 3.11% than the supervised model. The result proves our hypotheses that word embedding is effective pre-trained feature for supervised model. Experiment 9 adopts all effective features we got in this work and acquires a 1.84% in F score.

7. Conclusion

Word embedding can be acquired in advance and is task independent. A well-trained word embedding can be applied into many NLP tasks. In this work, we train a Tibetan word embedding firstly on a large-scale corpus. Then, the word embedding is used into predicate recognition task. The result shows that the semi-supervised model with word embedding feature performs better than baseline system. And like the priori knowledge, the word embedding is also effective feature for supervised machine learning method. Our future work will explore unsupervised methods, for instance, to apply neural network into Tibetan predicate recognition task.

Acknowledgement

This research was supported by the National Natural Science Foundation of China under Grand 61550004 and the Qinghai Natural Science Foundation under Grant 2016-ZJ-931Q.

References

[1] Di, J.: Recognition and Information Abstraction of Finite Verbs in Modern Tibetan. In: Advances in Computation of Oriental Languages--Proceedings of the International Conference on Computer Processing of Oriental Languages, Beijing (2003).

[2] Di, J., Yanhong, H.: The Construction and Identification Approaches of Adjectival Predicate in Modern Tibetan. Linguistics Study, vol. 5, pp. 115-122. (2005).

[3] Lin L., Congjun, L.: Recognition of Tibetan Linking Verb and Existential Verb. Journal of Chinese Information Processing, vol. 27(4), pp. 59–63. (2013).

[4] Weina, Z., Lin, L., Huidan, Liu.: Automatic extraction of trisyllabic verb phrases in Tibetan. Journal of Chinese Information Processing, vol. 29(3), pp. 196-200. (2015).

[5] Di, J.: The method and process of the modern Tibetan part-of-speech tagging by group strategy. Linguistics Study, vol. 4, pp. 30-39. (2003).

[6] Lin, L., Congjun, L., Di, J.: Tibetan functional chunks boundary detection. Journal of Chinese Information Processing, vol. 27 (6), pp. 165-169, (2013).

[7] Tianhang, W., Shumin S., Congjun, L.: Tibetan Chunking Based on Error-Driven Learning Strategy. Journal of Chinese Information Processing, vol. 28 (5), pp. 170-175, (2014).

[8] Joseph, T., Lev R., Yoshua B.: Word representations: a simple and general method for semi-supervised learning. In Proceedings of the48th Annual Meeting of the Association for Computational Linguistics, pp. 384–394, Uppsala, Sweden (2010).

[9] Lizhen, Qu., Ferraro, G., Liyuan, L.: Big Data Small Data, In Domain Out-of Domain, Known Word Unknown Word: The Impact of Word Representation on Sequence Labeling Tasks.

Computer Science, vol. 42(12), pp. 557-566, (2015).

[10] Fei, S., Pereira, F.: Shallow parsing with conditional random fields. In: Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Association for Computational Linguistics, pp. 134-141, (2003).

[11] Jewen, Z.: The grammar of Tibetan dialect in Lhasa dialect. Nationalities Publishing House, Beijing (2003).

[12] Bengio, Y., Ducharme, R., Vincent, P.: A neural probabilistic language model. Innovations in Machine Learning. Springer Berlin Heidelberg (2006).

[13] Bengio, Y.: Neural net language models. Scholarpedia, vol. 3(1), pp. 3881, (2008).

[14] Hinton, G.: Learning distributed representations of concepts. In: Eighth Conference of the Cognitive Science Society, (1986).

[15] Tomas, M., Kai C., Greg, C.: Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, (2003).

[16] Congjun, L., Huidan, L., Minghua, N.: Tibetan POS tagging based on syllable tagging. Journal of Chinese Information Processing. Vol. 29(5), pp. 211-216, (2015).

[17] Turian, J., Ratinov, L., Bengio, Y.: A preliminary evaluation of word representations for named-entity, (2009).

[18] Lafferty, J., McCallum, A., Pereira F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001). San Francisco: Morgan Kaufmann Publishers Inc. pp. 282-289, (2001).

[19] Changning, H., Hai, Z.: New method of Chinese word segmentation by word information. Advanced development of Chinese information processing, Beijing: Tsinghua University Press, (2006).

[20] http://word2vec.googlecode.com/svn/trunk/