International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.2 (2014), pp.45-62
http://dx.doi.org/10.14257/ijmue.2014.9.2.05

Verb Classification Using Bilingual Lexicon and Translation Information in Tibetan Language

Lirong Qiu

School of Information Engineering, Minzu University of China, Beijing, China
qiu_lirong@126.com

Abstract

Automatically acquiring semantic verb classes from corpora is a challenging task, especially with no existing treebank. Building a high-performing parser for a language is still crucially depends on the existence of large, in-domain texts as training data. While previous work has focused primarily on major languages, how to extend these results to other languages is the way to avoid working start from scratch. In general, a large monolingual corpus in a resource-rich source language labeled with lexico-syntactic information, and a very limited bilingual corpus are available. This paper addresses the problem of verb classification automatically in Tibetan using bilingual lexicon and translation information.

Keywords: *Verb Classification, Semantic Class Mining, Pattern Extraction*

1. Introduction

Verbs play an important role in conveying semantics of natural languages, and verb classification is a fundamental topic of computational linguistics research given its importance for understanding the role of verbs [1]. Additionally, generalization based on verb classification is central to many natural language applications, ranging from shallow semantic parsing to semantic search of information extraction. A substantial amount of research has been done in the area of automatic verb class identification, under a variety of different names and with a variety of different goals.

Currently, many languages have their extensive knowledge bases of verbs, such as VerbNet and FrameNet. Automatically acquiring verb class from corpora is a challenging task, especially with no existing treebank. Building a high-performing parser to acquire verb class for a language is still crucially depends on the existence of knowledge bases of verbs and large, in-domain texts as training data.

Note that learning algorithms for verb classification is rather complex and should be taken into account syntactic structures and should be enriched with lexical information to trig lexical preference. Semantic resources significantly reduced data sparseness. However, for many minority languages in China, we cannot acquire any knowledge base so far. For example, according to Tibetan language, all we got are Chinese-Tibetan dictionary and Chinese knowledge base named HowNet.

Considering the fact that, in general, we have a large monolingual corpus in a resource-rich source language labeled with lexico-syntactic information, and a very limited bilingual corpus, this paper addresses the problem of automatically verb classification using bilingual dictionary and translation information. From the viewpoint of verb extraction and collocation translation, a novel framework is proposed to automatically assign fine-grained semantic labels to verbs in Tibetan.

ISSN: 1975-0080 IJMUE

Consider the example Tibetan sentences shown in Figure 1, and suppose that we wish to parse the Tibetan side without access to any Tibetan knowledge base but a Tibetan-Chinese bilingual dictionary and translation information.

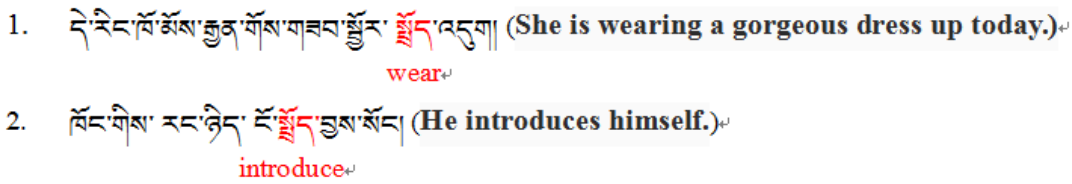

Figure 1. Sentences in Tibetan language both containing word སྐྱོད་ that has different meanings

In this work, we consider how to extend results about verb classification to Tibetan language to avoid working start from scratch. Our main approach is to map lexico-syntactic projection from a resource-rich source languages (English, Chinese) to a resource-poor target language (Tibetan) using a bilingual lexicon and translation information.

The remainder of this paper is organized as follows. Section 2 presents the related work in verb classification. Section 3 describes our verb classification method. Section 4 evaluates the proposed method and section 5 presents some analysis, including the candidate selection by semantic similarity and relatedness and the combination of methods. Section 6 draws the conclusions and mentions future work.

2. Related Work

Given the great deal of similar work in information extraction and a variety of methods have developed for automatic semantic verb class mining, but most of them rely on tagging relation from a large, in-domain corpus [2, 11]. How to transfer lexico-syntactic information from a resource-rich source language to a resource-poor target language with no existing treebank is still an open problem [3]. A key challenge in this setting is that it is hard to obtain sentence-level training data.

A number of extracting studies investigate the learning of semantic verb classes using two sources of information: lexicographic resources and distributional similarity [4, 10].

Lexicographic resources are manually-prepared knowledge bases containing semantic information on predicates, such as WordNet and VerbNet, which are commonly used. The handcrafted WordNet uses the hyperonymy or hyponymy relationship to structure the English verb lexicon into a semantic network. VerbNet is organized in a hierarchical structure of classes and sub-classes, each sub-class inheriting the full characterization of its super-class, aimed at achieving more coherent classes both semantically and syntactically [5].

Distributional similarity algorithms use large corpora to learn broader resources by assuming that semantically similar predicates appear with similar arguments [6, 9].

However, we cannot acquire any knowledge base in Tibetan language so far. We also tried to establish ontology in Tibetan language [7], but the ontology is mainly focus on the entities, mostly of which are nouns. Verbs play an important role in conveying semantics of natural languages, and our work is mainly focused on how to map the existing verb classification results to a resource-poor target language.

Greg Durrett considered the problem of using a bilingual dictionary to transfer lexico-syntactic information from English to Germany [3]. Hua Wu proposed a method that gets candidates of synonymous collocation pairs based on a monolingual corpora and a word

thesaurus, and selects the appropriate pairs from the candidates using translations in a second language [8].

Despite the relatively large related work in verb classification domain and syntactic transferring domain, few efforts have specifically addressed the topic of mapping verb classes from a resource-rich source language to a resource-poor target language using bilingual lexicon and translation information. Our focus is on identifying verb classes when one does not have access to a full-scale target language treebank, but one do have access to realistic auxiliary resources. To our knowledge, this is the first attempt for Tibetan language.

Our method for verb classification comprises of three steps, as shown in Figure 2: (1) establish the verb classes and extract verbs from monolingual corpora; (2) generate candidates of verb pairs with a bilingual dictionary; (3) classifying verbs using syntactic information with translation information.

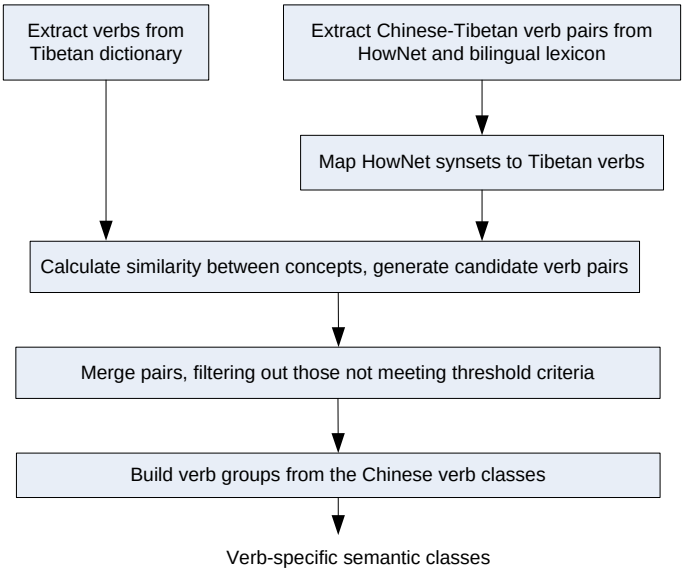

Figure 2. Approach for Building Semantic Verb Classes for Tibetan language

3. Corpus Used

The design of method for Tibetan verb classification requires the modeling of two different aspects: (1) the lexical similarity suitable for the task; (2) a classification for the Tibetan verbs.

3.1. Data Resources

Before we describe the details of our experiments, we sketch the data conditions under which we evaluate our method. As cited in Section 1, the most serious challenge so far to Tibetan language processing is lack of both knowledge base and the large training dataset. Both translation information provided by bitexts and bilingual dictionaries are more widely available than complete treebanks in terms of how word are used in practice, but for resource-poor languages, parallel text may only be available in small quantities, or be domain limited.

Inspired by the work of Greg Durrett [3], the first variable, we consider is whether we have access to a small number of verb classes and sentences annotated with verb semantic classes in Tibetan language or only pre-existing knowledge base in a number of other languages,

which can serve as a kind of proxy for transferring useful information. However, we cannot find any relationship between Tibetan and English, which has many existing treebanks.

Most of the existing bilingual corpora resources in Tibetan language are related to Chinese, such as Tibetan-Chinese bilingual dictionaries, Tibetan Verb dictionaries and Chinese-Tibetan translation sentences. How to utilize and extend results of verb classification to Tibetan language is the way to avoid working start from scratch, so we believe that this is an important consideration to connection from Tibetan to Chinese.

Chinese knowledge base HowNet is a semantic knowledge base for the English-Chinese words. It is generally used in the field of Chinese information processing in earlier studies. HowNet represents a word with the following format (Using Chinese word-"介绍" as an example), as shown in Table 1:

Table 1. Represent Format In HowNet

Description Flag	*Description*
W_C	介绍
W_E	introduce(acquaint, present)
DEF	{CauseToDo|使动：ResultEvent={know|知道}}

And we can extract the Tibetan words from the Chinese-Tibetan bilingual dictionary and establish a Tibetan word definition table, as shown in Table 2:

Table 2. Represent Format In HowNet with Tibetan words

Description Flag	*Description*
W_C	介绍
W_E	introduce(acquaint, present)
W_T	སྤྲོད་
DEF	{CauseToDo| འགྱུར་བར་བྱེད།：ResultEvent={know| རིག| }}

As our HowNet-based method uses the graph of the concepts and relations, we can easily compute the similarity between words from Chinese and Tibetan languages. For example, consider a Chinese-Tibetan pair like (介绍- སྤྲོད་), illustrated in Table 2. Given the Tibetan verb dictionary is included in HowNet, we can use HowNet as the common knowledge-base for the relations. We can compute the similarity and relatedness of Tibetan verbs, and the related method to extract Tibetan words and map to HowNet is introduced in our work [14].

3.2. Verb Classes in Tibetan Language

Our use of semantic classes of verbs is motivated by the work of Levin [12] and Di Jiang [13].

Levin (1993) groups verbs based on an analysis of their syntactic properties, especially their ability to be expressed in diathesis alternations; her approach reflects the assumption that the syntactic behavior of a verb is determined in large part by its meaning. The VerbNet lexical resource classifies English verbs into different classes based on the commonalities in their semantic and syntactic behavior on the basis of Levin's theory.

When it comes to Tibetan language, Di Jiang discusses the classification of Tibetan verbs according to verbal semantic types and syntactic types, which can describe the numbers of arguments and characters of components in sentences. In his work, he classified Tibetan verbs into 12 classes, which are stative verbs (*e.g.*, སྡུག་། 'ill'; ཤི། 'dead'), action verbs (*e.g.*, རྐྱག་པ་རྒྱག། 'kick'), cognition verbs (*e.g.*, ངོ་ཚ། 'shy'), perception verbs (*e.g.*, རེག། 'understand'), verbs of change (*e.g.*, འགྱུར། 'convert'), directional verbs (*e.g.*, ཡོང་། 'come'), narrate verbs (*e.g.*, བཤད།

'say'), copula (*e.g.*, ཨིན། 'am, is, are'), verbs of possession (*e.g.*, ཡོད། 'have'), existential verbs (*e.g.*, ཡོད། 'be at'), interrelation verbs (*e.g.*, འདྲ། 'be similar to'), causative verbs (*e.g.*, བཟོ། 'let sb. do'), based on both syntactic behavior and semantic criteria.

Di Jiang proposed that class distinctions in verbs should rely on the semantic and syntactic feature of the verb's arguments in narrowly defined verb-specific semantic classes. However, as cited by many researchers in other languages, many verbs belong to multiple classes, with each class membership corresponding roughly to a different sense of the verb.

The design of verb classification requires the modeling of extensive syntactic and semantic information for the verbs and the lexical similarity suitable to assign new verbs to verb class. We defined an effective mapping from the Tibetan verb classes to HowNet verb classes, therefore, we can transfer the corresponding words in HowNet to the twelve verb classes in Tibetan.

4. Verb Classification Method

Our work began with the simple idea of using an extremely specific pattern to extract semantic class members with high accuracy. With the aim of capturing syntactic features, we started from two different paradigms: verb pairs and similarity calculating of the resource-rich language.

4.1. Candidate Selection by Translation Information

In synonymous word extraction, the similarity of two verbs can be estimated based on the similarity of their contexts. However, this method depends on the large training dataset. And monolingual information is not enough to estimate the similarity of two verbs from different languages. Therefore, we use translation information to help find the candidate verbs.

In this paper, we select the synonymous verb pairs from the candidates in the following way. Firstly, we give some definitions as follows:

Definition 1. Suppose source sentence S and the target translation sentence T are defined as S= $\{S_1, S_2, \ldots, S_n\}$ and T= $\{T_1, T_2, \ldots, T_m\}$, in which $\forall S_i(1<=i<=n)$ and $\forall T_j(1<=j<=m)$ are the words in the sentences.

Definition 2. $A=<A_s, A_t> \subseteq S \times T$, in which $A_s \in S$, $A_t \in T$. Then we called set A is an alignment of Chinese sentence S and Tibetan sentence T. Triple <S, T, A> is a bilingual aligned sentence pair.

Algorithm 1. Acquire the alignment Chinese and Tibetan word pair Set A

Input: a Chinese sentence $S=S_1S_2 \ldots S_n$ and a Tibetan translation sentence $T=T_1T_2 \ldots T_m$ is translation sentences

Step1: initialize Set S=∅, set T=∅, set A=∅.

Step2: automatic segmentation of Chinese and Tibetan sentences and filter the auxiliary words, Set S and T are defined of words labeled with numbers.

Set S = $\{< S_1, 1>, < S_2, 2>, \ldots, < S_n, n>\}$

Set T= $\{< T_1, 1>, < T_2, 2>, \ldots, < T_m, m>\}$

Step3: for $\forall<S_i, i> \in$ Set S (1<=i<=n) and $\forall<T_j, j> \in$ Set T(1<=j<=m)

if *Sim*(S_i, T_j)> h (h is the threshold of value of word similarity)

then add < S_i, T_j > word pair to Set A

Step4: Output the word pair in set A and select the verb pair <S_v, T_v>.

4.2. Verb Classification by Translation Information

A verb's membership in different classes also depends on its meaning, making the inclusion of semantic features of translation words a possible benefit. As mentioned earlier, multiple class memberships usually correlate with different senses of the verb, making class disambiguation much like verb sense disambiguation task.

Three types of semantic features are used, all derived from the arguments of the target verb T_j: (1) extract the Chinese verb from the verb pair <S_i, T_j>; (2) synonyms of S_i as listed in HowNet; (3) hypernyms of the Chinese verb and their Tibetan translation.

Algorithm 2. Verb Classification Method

Input: *TV*, a set of Tibetan verbs;

A, a set of Tibetan verbs and Chinese translation words <S_v, T_v> pairs;

F, a set of (*word*, *frequency*) pairs;

D, a set of (*Verb_sense_id*, *def_verb*) pairs,

Step 1. Relates Tibetan verb senses that are defined in terms of the same verb

for $\forall T_j \in TV$, extract all (*verb_sense_id*, *def_verb*) pairs where $v = def\_verb$;

forall *v* that exist as *def_verb* in *D*, form $D_j \subseteq D$;

remove all D_j for which $|D_j| > 10$;

Step 2. Extracts (direct or extended) semantic relationships in HowNet

forall *v* that exist as *def_verb* in D, create <T_v, S_v>

return (*synset*, *related_synset*) pairs for all synsets directly related through hyponymy and entailment of S_{vi} in HowNet , $SimV_i$={S_{vi1},S_{vi2},…,S_{vin} }

forall $S_{vi} \in SimV$, measure the similarity between word pairs;

remove S_{vi} from *SimV*, $(S_{vi}, f) \in F$, in which $f = \sum_{i=1}^{k} w_i \square Sim(s_i, v_i), f > h$ where w_i is a constant value, *h* is the threshold of value.

Step 3. Relates Verb senses that map to the same HowNet synset

forall $S_{vi} \in SimV$, get the set A_{vi}={< S_{vi}, T_{vi} >}

return all combinations of two Tibetan verb senses mapped to the same HowNet synset

5. Experiments

Our solution is on classifying Tibetan language verbs that has no access to treebank. We tested our method on resources extracted from our storage bases. We carried out experiments of identifying different Tibetan verbs to different semantic verb class using bilingual semantic lexicon and translation information.

(1) From the Tibetan-Chinese dictionary, we extract 2651 Tibetan verbs with different meaning or with different usage from the perspective of grammar. (2) We extract the corresponding Tibetan-Chinese verb pairs, as one Tibetan verb has several semantic senses which corresponds several Chinese words, as shown in Figure 2. (3) The total number of the corresponding Chinese verbs comprised of hyponymy and entailment relations is 12450 extracted from HowNet. After calculating the similarities, the 8243 Chinese verbs are left. (4) For all the Tibetan verbs, the semantic verb class is defined.

0181	འཕྲོག་	1	知道，了解	realize
0182	འཕྲོག་	2	蜷缩，收缩	curl up
0183	འཕྲོག་	3	胆小，胆怯，畏缩	shrink
0184	འཕྲོག་	4	紧缩，减缩， 限制，拘束	retrench

Figure 2. Verb List of Tibetan language

Baselines were established for each target verb type by calculating the accuracy that would be achieved if all Chinese verbs were labeled with its most frequent same verb class.

6.. Conclusion and Future Work

Automatically acquiring semantic verb classes from corpora is a challenging task, which is still crucially depends on the existence of large, in-domain texts as training data. However, for Tibetan language, we cannot acquire any knowledge base so far. In general, a large monolingual corpus in a resource-rich source language labeled with lexico-syntactic information, and a very limited bilingual corpus are available. This paper addresses the problem of verb classification automatically in Tibetan using bilingual lexicon and translation information.

The results of this study suggest that automatic disambiguation of verb classes is a reasonable line of research, and a possible method for verb sense disambiguation. The method relies on lexical and syntactic feature acquired from the translation information. Experiments on a data set show that it can extract a very large high-quality semantic set of Tibetan verbs.

The work of this paper is a part of our ongoing research work, which aims to provide an open reusable verb semantic framesets for further Tibetan language processing progress. Various experiments and applications have been conducting in our current research. Future work includes how to acquire and verify semantic relations of verbs from Tibetan sentences, how to obtain verb patterns automatically and how to verify the semantic relations with self features and context features.

Acknowledgements

Our work is supported by the National nature science foundation of China (No. 61103161) and the Program for New Century Excellent Talents in University (NCET-12-0579).

References

[1] S. W. Brown, D. Dligach and M. Palmer, "Verbnet class assignment as a wsd task", In proceedings of the 9th International Conference on Computational Semantics (IWCS), **(2011)** January 12-14; Oxford, UK.
[2] D. Croce, R. Basili, A. Moschitti and M. Palmer, "Verb Classification using Distributional Similarity in Syntactic and Semantic Structures", In proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), **(2012)** July 7-14; Jeju, Korea.
[3] G. Durrett, A. Pauls and D. Klein, "Syntactic Transfer Using a Bilingual Lexion", In proceedings of the EMNLP-CoNLL, **(2012)** July 12-14; Jeju Island, Korea.
[4] D. Dligach and M. Palmer, "Novel Semantic features for verb sense disambiguation", In proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), **(2008)** June 15-20; Columbus, Ohio, USA.
[5] O. Abend, R. Reichart and A. Rappopart, "A supervised algorithm for verb disambiguation in VerbNet classes", In proceedings of the 22nd International Conference on Computational Linguistics, **(2008)** August 12-18; Manchester.
[6] R. Barziley and M. Lapata, "Modeling local coherence: An entity-based approach", Computational Linguistics, vol. 1, no. 34, **(2008)**, pp. 1-34.
[7] L. Qiu, Y. Weng and X. Zhao, "Acquisition method of hyponymy concepts based on patterns in Tibetan semantic ontology", Journal of Chinese Information Processing, vol. 25, **(2011)**, pp. 45-49.
[8] H. Wu and M. Zhou, "Synonymous Collocation Extraction Using Translation Information", In proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), **(2003)** July 7-12; Sapporo, Japan.
[9] B. Min, S. Shi, R. Grishman and C. -Y. Lin, "Ensemble Semantics for Large-Scale Unsupervised Relation Extraction", In proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Computational Natural language Learning, **(2012)** July 12-14; Jeju Island Korea.
[10] S. Schulte im walde, "Clustering verbs semantically according to their alternation behavior", In proceeding of 18th International Conference on Computational Linguistics, **(2000)** July 31- August 4; Saarbrücken, Germany.
[11] D. Albright, A. Lanfranchi, A. Fredriksen, W. F. Styler, C. Warner, J. D. Hwang, J. D. Choi, D. Dligach, R. D. Nielsen, J. Martin, W. Ward, M. Palmer and G. K. Savova, "Towards comprehensive syntactic and semantic annotations of the clinical narrative", Journal of the American Medical Informatics Association, vol. 5, no. 20, **(2013)**, pp. 922-930.
[12] L. Beth, "English Verb Classes and Alternations: A preliminary Investigation", University of Chicago Press, Chicago, **(1993)**.
[13] D. Jiang, "The Classification of Tibetan Verbs and Relative Patterns Based on Semantics and Syntax", Journal of Chinese Information Processing, vol. 1, no. 25, **(2006)**, pp. 37-43.
[14] X. Jiang and L. Qiu, "A Tibetan Ontology Concept Acquisition Method Based on HowNet and Chinese-Tibetan Dictionary", International Conference on Asian Language Processing, **(2013)** August 17-19; Urumqi, China.
[15] G. Fu and X. Wang, "Chinese sentence-level sentiment classification based on fuzzy sets", In proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), **(2010)** July 11-16; Uppsala, Sweden.

Author

Lirong Qiu

She received his M.Sc. in Computer Sciences (2004) and PhD in Information Sciences (2007) from Chinese Academy of Science. Now she is full professor of computer sciences at Information Engineering Department, Minzu University of China. Her current research interests include different aspects of natural language processing, artificial intelligence and distributed systems.

