Tibetan Information Extraction Technology Integrated with Event Feature and Semantic Role Labelling

Fucheng Wan¹ Jianhua Xia²

¹ (Key lab of China's National Languages Information Technology, Northwest University for Nationalities, Lanzhou, China)

² College of Computer and Information, Hohai University, Nanjing, China

Abstract: we integrate with semantic information which is based on syntactic analysis for extracting the Tibetan information. For Tibetan language information extraction, through experiments analyzed, syntactic analysis model which is integrated with information of semantics, as well as the evaluation of program can be used in Tibetan language information extraction task successfully.

1 Introduction

Information extraction refers to automatic extraction of main information from the text, displayed in a structured form. Information extraction [1], as a key technology in the field of information processing, has been widely used in information retrieval [2], automatic question and answer [3], text mining and so on. Tibetan information extraction algorithm of syntactic information and semantic parsing can be used the information security system such as the Tibetan public opinion surveillance, Tibetan text link detection, Tibetan hot topic recognition and tracking, etc., with good academic value and broad application prospect. It has positive significance in automatic generation of Tibetan database and knowledge base, Tibetan question and answer system, Tibetan information retrieval research, etc.

2 Related Research

In the late 1950s, Zellig Hays put forward the related information extraction and indexing of the literature of science and technology [4], and since 2000, the ACE evaluation meeting and information extraction have developed through decades of years.

In foreign countries, the researches on information extraction have developed from particular field to the open field, enabling the extraction information to be more comprehensive. Data format has developed from regular standard text to irregular common text. Data type has developed from news report and scientific paper to various website texts, Weibo, etc. Method has developed from manual modeling to automatic modeling of machine learning. The information extraction has obtained more notable achievements [6]. Heng Ji et al. [7] studied the cross-document event extraction and tracking, as well as

its evaluation standard. Alan Ritter, et al. [8] studied the event extraction facing the open field such as Twitter, applying event extraction technology to real-time and irregular social media space, which makes event extraction technology further practical.

In China, Chinese information extraction technology was started relatively late, but it has made some research achievements. Yuan Yulin [9] studied to semantic resources construction of event extraction. Zhao Yanyan, Liu Ting et al. [10] s proposed event type recognition based on trigger word and binary classification technology. Xu Ronghua, Zhu Qiaoming et al. [11] defined an event fusion framework, TEFF, and made theme event displayed in the hierarchical form according to the role of all kinds of meta events in the theme events. Ding Xiao, Qin Bing, Liu Ting et al. [12] researched event extraction in the field of music, and applied research results in the platform of Harbin industrial university language technology (LTP) [13].

The research foundation on Tibetan information processing is weak, and the information extraction research still stays on the lexical analysis. Xu Tao et al. [14] Researched automatic parsing model by combining the rules of sentence analysis and MEM based on the research of Li Xiang, Liu Qun, et al. [15]. Jin Ming, et al. [16] conducted Tibetan named entity recognition through rules and dictionary method. Zhu Hongzhi et al. [17] used the co-occurrence frequency of named entities to calculate the relationship between Tibetan entities, but in the open field, there are a fewer researches on Tibetan event extraction at home and abroad.

3 Analysis of Tibetan functional semantics

3.1 Tibetan part-of-speech tagging and syntactic

rule

Tibetan sentence components include subject, predicate, object, attributive and adverbial. The sentence is mainly predicate sentence as the core with subject-predicate structure. Tibetan word order is as follows: subject, object (direct object, indirect object), predicate, belonging to SOV syntactic structure, which is different from Chinese. When adjectives, numerals, demonstrative pronoun are taken as modifier, they are put after the center word. When personal pronoun and noun are taken as modifiers, they need to add utterances in front of central word. Verb adjective modifiers are generally before the central word.

For example, the characteristic form of Tibetan of verb phrases includes V + UT + UM + UE (UT as tense auxiliary, UM as modal auxiliary, UE voice end utterances). The order and quantity of case-auxiliary word (grammatical function is equal to postpositive preposition); the connection relation between case-auxiliary word and front nouns reflects the grammar characteristics of the Tibetan language.

In order to construct syntax treebank of Tibetan phrase, the Tibetan notional word and function word are respectively refined and tagged, shown as in figure 1:

Table 1 collection of part-of-speech tagging of Tibetan phrase treebank

Common noun --NN	Name --NR	Noun of time --NT	Name of organization --NO	Name of place --NS	Rhetoric word --NE
Buddhist Sanskrit --NF	Numeral --CD	Quantifier --M	Adverb --AD	Conjunction --CC	Noun of locality --LC
Interjection --IJ	Personal pronoun --RR	Demonstrative pronoun --RZ	Interrogative pronoun --RY	Indefinite pronoun --RB	Closing speech -E
Mimetic word --ON	Idiom --I	Dialect --L	Abbreviation --J	Case-auxiliary word --P	Verb--VV

3.2 Examples of Tibetan phrase treebank integrated with semantic information

According to the labelling in table 1, with the use of labeling specification about phrasal level of Chinese tree

from university of Pennsylvania library (university of Pennsylvania, Chinese tree CTB), this paper constructs the Tibetan phrase syntax treebank for Tibetan event. The phrase syntax treebank is as shown in figure 1

Figure. 1 sample 1 of phrase syntax treebank for Tibetan event

In the current construction of Tibetan syntax treebank, there are more relying on the syntax, but what is interdependent on syntactic withdrawal is the word dependencies. This paper focuses on Tibetan phrase syntax, with the samples in figures 1 and 2 as the standard, to manually label the Tibetan content words and function words respectively. The corpus with 2000 sentences is taken the training corpus, to decrypt test corpus, and then it is assisted with artificial proofreading, adopting the method of bootstrapping to gradually expand the Tibetan phrase syntax treebank.

4 Sample of Tibetan information extraction integrated with semantic feature

4.1 Trigger word recognition

Information extracting must identify the trigger word in the sentence firstly. Tibetan verbs are generally at the end of the sentence, and the core of the whole sentence; from

the aspect of position, the Tibetan language identification is relatively easier than other components. Many verbs do not appear in the training corpus, and are non-record words. (respect for teachers, for example), assuming (respect teacher; assuming (respect) does not appear in the training corpus as trigger word, it is easier to identify this event. However, the meaning of (respect) and (esteem) is close; based on this, this paper semantic similarity calculation based on Tibetan verb dictionary, and trigger word dictionary, automatically extends event trigger word, and cover all types of event trigger word as much as possible. For Tibetan word's semantic similarity, the word vector space model and the Tibetan vocabulary context words are used to describe, and the context can provide abundant linguistic information for words. First, a group of feature words are selected, and then according to the frequency of this set of words in the corpus and TF-IDF value, the relevance between a group of words and every word is calculated to get a featured vector; finally, the similarity between vectors is used as the similarity of two words. Root trigger word and its information category form a binary set (trigger, type), for example, (chairman, Person/Respect), to build the binary comparison table of "Tibetan root trigger word - information category"; the binary table has two columns of event trigger words and information categories.

Table 2 comparison table of "root trigger word-information category"

Root trigger word	Information category
(Close down)	Business/Declare-Bankruptcy
(Resist)	Conflict/Attack
(Escape)	Movement/Transport
(Demonstrate)	Conflict/Demonstrate
...	...

The first column is root trigger word; the second column represents the information category of each trigger; each trigger word corresponds to the only one information category, and the trigger words in this table will cover the entire information category. The information category is labelled as the different symbols, with maximum likelihood probability algorithm to get parameters for identification model, and the state transition matrix parameter is

$$Ta_{ij} = \frac{TC_{i,j}}{\sum_{i=1}^N TC_{i,k}}, 1 \leq i, j \leq N,$$

$TC_{i,j}$ is the corresponding state of all symbols,

TS_i is the number of times to reach the state TS_j ., representing the frequency of root trigger word in the corpus to the information category.

$$Tb_j(V_k) = \frac{TE_j(V_k)}{\sum_{i=1}^N TE_j(V_i)}, 1 \leq j \leq N, 1 \leq i \leq M$$

$TE_j(V_k)$ is V_k sending frequency in the state of TS_i , $Tb_j(V_k)$ is the proportion of sum of sending probability. $Init(i)$ is the total number in the state of i .

$$TP_i = \frac{Init(i)}{\sum_{j=1}^N Init(j)}, 1 \leq i \leq N$$

4.2 Semantic role labelling

The frame semantics of C.J.Fillmore upholds that the case of proposition includes: (1) Agentive; (2) Instrumental; (3) Dative; (4) Factitive; (5) Locative; (6) Objective; (7) Benefactive; (8) Source; (9) Goal; (10) Commutative. On the basis of the Tibetan syntax rule, with the reference of the classification of frame semantics about case, this paper takes five Tibetan function semantics clues including agent, object, time, place, way as the study content.

Chinese Proposition Bank (CPB) of University of Pennsylvania is the resource of Chinese superficial semantic labelling based on Penn Chinese Treebank. The labelled data of Penn Chinese Treebank mainly comes from Xinhua news, Smorama news and magazine and Hong Kong news. CPB contains more than 20 semantic roles; the same semantic role has different semantic meanings for different target verbs. The core semantic role is from Arg0 to Arg5. Arg0 usually means the action of agent; Arg1 usually shows the effect of action and so on. The remained semantic roles are additional semantic role, with the prefix ArgM to express, followed by some additional tags to represent semantic category of these parameters, for example, ArgM2LOC represents location, and ArgM2TMP represents time, etc.

With the structure of "predicate -argument role", for each predicate in the sentence (verbal predicate or nominal predicate, etc.), the corresponding semantic roles of the predicate in the sentence are marked, and the main argument role is the agent AGR0-SUB, object ARG1-OBJ, auxiliary argument role, time ARGM - TEMP, place ARGM-LOC, way ARGM - MNR, as shown in table 3. This integrates semantic information into the syntactic structure tree, as shown in figure 1 and figure 2 at phrasal level; for training and syntax analyzer, Berkeley Parser is used.

Table 3 Corresponding table of semantic labeling

Agent	ARG0-SUB	object	AGR1-OBJ	Time	AGRM-TMP
Place	AGRM-LOC	Method	AGRM-MNR		

5 Experiment and assessment

5.1 Experiment setting

(1) corpus selection: Tibetan text categorization corpus, including 1200 texts of 6 major categories including history, literature, religion, rhetoric, philosophy, politics, etc. among which there are 1000 are the training texts, and 200 are test texts.

Table 4 experiment corpus setting (unit: piece)

Training corpus	Test corpus
1200	200

(2) Result assessment scheme

In this paper, the results of Tibetan information are compared with the standard library for analysis; F value of accuracy (P), and recall rate (R) is used for comprehensive evaluation.

$P = \text{number of correct sampling attribute} / \text{number of actual sampling attribute} \times 100\%$

$R = \text{number of correct sampling attribute} / \text{number of imposed sampling attribute} \times 100\%$

$$F = \frac{2 \times P \times R}{P + R} \times 100\%$$

5.2 Experimental method and results

This experiment compares the methods of part-of-speech tagging, trigger word recognition, part-of-speech tagging integrated with trigger word, and semantic role labelling integrated with trigger word.

(1) Part-of-speech tagging: the part-of-speech method of only using Tibetan words, for example, verb is labelled as vv, and noun is labelled as NN.

(2) Trigger word recognition: taking the results of model recognition as the input of information extraction model to get the results of information extraction

(3) Part-of-speech tagging integrated with trigger word: combining the former two methods, and comparing it with single method

(4) Semantic role labelling integrated with trigger word: namely, the method used in this paper, trigger word as event feature, semantic role information as the text feature of Tibetan, as shown in figure 1 and 1, and labeled in treebank.

The experiment results are shown in table 5

Table 5 result analysis of Tibetan information extraction

Method	P	R	F
Part-of-speech tagging	60.52	57.63	59.04
Trigger word recognition method	65.34	53.26	58.68
part-of-speech tagging+trigger word	68.74	62.34	65.38
trigger word+semantic role labeling	76.78	75.32	76.04

Comparison of the methods of Tibetan information extraction

5.3 Result analysis

From table 5, it can be seen that when only using part-of-speech tagging method (method 1), and only using

the trigger word recognition method (method 2), the F value is low; thus, only using dictionary information and part-of-speech tagging method have low contribution rate for Tibetan information extraction. The reason is that both dictionary and part-of-speech tagging method more focus on the objective knowledge, but not to integrate the Tibetan linguistics into the model.

F value of the former two methods have not significantly improved, which also proves the a priori knowledge or artificial model has no good effect on increasing efficiency.

Semantic role labeling information fusion trigger word, that is, the method used in this paper, has improved overall accuracy and recall rate for the Tibetan extraction, thus, integrating Tibetan syntactic and semantic information of linguistics also helps Tibetan information processing of tasks. In addition, information extraction also has a lot to do with the accuracy of the syntactic analysis. The Tibetan information extraction model based on semantic role labeling relies on the accuracy of syntax analysis model of Tibetan language phrase, especially syntactic analysis of long sentences.

6 Conclusion and following work

Tibetan information extraction model integrated with semantic information is integrating semantic information for training on the basis of Tibetan language phrase syntax, so as to provide support for Tibetan information extraction. This paper provides a solution for the Tibetan information extracting on the aspect of semantic information, including the Tibetan semantic information classification, Tibetan semantic information labelling and the syntax analysis of phrases in the Tibetan language, etc. It contributes to providing an information extraction solution for other languages, including Mongolian, Uyghur and other ethnic minority languages.

The next step is to strengthen research in the lexical syntactic level to study and Tibetan lexical and syntactic integration model, so as to realize Tibetan language information extraction model without, namely, Tibetan language syntax analysis model based on syllable sequences, serving the Tibetan information extraction.

Acknowledgement

This work is supported by "the National Natural Science Foundation of China (Grant: 61602387)", and "the Fundamental Research Funds for the Central Universities(2014B33014)"

References

1. Liang Han, Chen Qunxiu, Wu Pingbo, Information Extraction System based on Event Framework. Journal of Chinese Information Processing, 2006, Vol.20No.2:40-46.
2. Naomi Daniel, Dragomir Radev and Timothy Allison. Sub-event based Multi-document Summarization [C]. In: Proceedings of the

- HLT-NAACL Workshop on Text Summarization. 2003. pp9-16.
3. Jiang Jifa. Question and Answer Method of Chinese Knowledge in Open Field, computer engineering, 2005.06, 31(11):1-3.
 4. W. W. Cohen, A. McCallum. Information Extraction and Integration: an Overview[C]. KDD 2003 Tutorial, Washington DC, U.S.A, 2003:pp1-89.
 5. Tan Yehong. Research on Key Technology of Chinese Event Extraction. Harbin: Harbin Institute of Technology, 2008.06.
 6. S. G Soderland. Building a Machine Learning Based Text Understanding System[C]. Proceedings of workshop on Adaptive Text Extraction and Mining (ATEM-2001), USA, 2001: pp133-154.
 7. Heng Ji, Ralph Grishman, Zheng Chen and Prashant Gupta. 2009. Cross-document Event Extraction and Tracking: Task, Evaluation, Techniques and Challenges[C]. Proc. RANLP, 2009: pp162-172.
 8. Alan Ritter, Mausam, Oren Etzioni and Sam Clark. Open Domain Event Extraction from Twitter[C]. In SIGKDD'12. 2012.
 9. Yuan Yulin. Research on semantic knowledge resources of information extraction. Journal of Chinese Information Processing, 2002, 16(5): 8-14.
 10. Zhao Yanyan, Qin Bing, Che Wanxiang, Liu Ting. Research on Extraction Technology of Chinese Event. Journal of Chinese Information Processing, 2008, 22(1): 3-8.
 11. Xu Ronghua, Wu Gang, Li Peifeng, Zhu Qiaoming. Research on theme event fusion based on event framework. Application Research Of Computers, 2009, 12, 12(26):4542-4545.
 12. Ding Xiao, Song Fan, Qin Bing, Liu Ting, Research on extraction method of typical event in music field. 5th national youth computer language seminar (YWCL 2010). 2010, 7: 431-437.
 13. Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform[C]. In Proceedings of the Coling 2010: Demonstration s. 2010.08:pp13-16.
 14. Xu Tao, Jia Yangji, Ding Hongzhi. Automatic punctuation of Tibetan in combination of statistics and rule. Journal of Yunnan University (natural science), 2012Vol.34(6):653-657, 663.
 15. Li Xiang, Cai Zangtai, Jiang Wenbin, Lv Yajuan, Liu Qun. Tibetan sentence boundary recognition method in combination of maximum entropy and rule
 16. Jin Ming, Yang Huanhuan, Shan Guangrong. Research on Tibetan identity of named entity. Journal of Northwest University for Nationalities (natural science)
 17. Hongzhi Yu, Tao Jiang, Bing Zhang. Relation Discovery by Named Entity Recognition from Tibetan Websites[C]. IEEE SWS, 2009: pp177-179.