Tibetan Information Extraction Technology Based on Functional Semantic Information

Fucheng Wan1,a*, Xiangzhen He2,a

(Key Laboratory of National Language Intelligent Processing , Gansu Province (Northwest University for Nationalities).730030, Lanzhou Gansu,)

awanfucheng@126.com; b306261663@qq.com

Keywords: Semantic Parsing; Tibetan semantic information; information extraction;

Abstract: We used Tibetan function semantic information to solve the problem of Tibetan information extraction. For Tibetan language information extraction, we also proposed a new evaluation program, and through experiments analyzed. The evaluation of program can be used in Tibetan language information extraction task successfully.

Introductions

Information extraction is the technology of extracting information from texts, and shows it with the structured form. It has been used widely in the field of information retrieval, question-answering and text mining. Tibetan information extraction can be applied to Tibetan public opinion research, Tibetan text tracking and hot topic detection.

Related Work

From the 1950s, Zellig Harris extracted the relevant information and indexed the scientific and technical literature, in 2000, ACE evaluation meeting was held, and information extraction research has gone through decades of years.

In abroad, domain research of information extraction was from oriented specific field to open field, makes it more useful; data format was from standard law text to general text without standard information; data type was from News report, scientific and technology papers to various page texts; method was from manual built to machine learning automatically built, information extraction has achieved many successful results. Heng Ji researched documents event detection and tracking, as well as its evaluation criteria; Alan Ritter research Twitter open domain-oriented event extraction, event extraction technology will be applied to irregularly in real time social media space, and event extraction technology to practical use.

Qin Bing, Liu Ting researched music-oriented events in the field of extraction and application of research results at the Harbin Institute of technology language technology platform (LTP).

Tibetan Function Semantic Parsing

Tibetan syntax rule and semantic classification

The elements of Tibetan Sentence contain subject, predicate, object, attribute and adverbial. The center element is Tibetan verbs or Tibetan predicate information. Tibetan sentence and word order:

subject, object (indirect object, direct object), predicate, syntactic structure is SOV type.

Based on the Tibetan syntactic rule, also according to the classification of frame semantics on the lattice framework, we divide the semantic information with the agent, patient, time, place and manner, five Tibetan feature semantics. CPB is annotated with shallow semantic annotation of Chinese resources based on Chinese Penn Treebank. CPB contained more than 20 more semantic information, the same semantic roles for different purposes have different semantic meanings of the verb. Core semantic role as Arg0 through Arg5 six, Arg0 is usually means the action of the agent, Arg1 is often said the effect of the action, and so on. Other semantic role such as additional semantic role, indicated by a prefix ArgM, followed by some additional markup to represent the semantic category of these parameters, such as ArgM2LOC, ArgM2TMP, and so on.

Tibetan syntax tree integrated with semantic cues

According to the annotation plan of CTB, we built the Tibetan syntax Treebank for Tibetan information extraction. Take the event bank “

(Peking is the capital of China)”for example，the structure of this sentence is as figure 1:

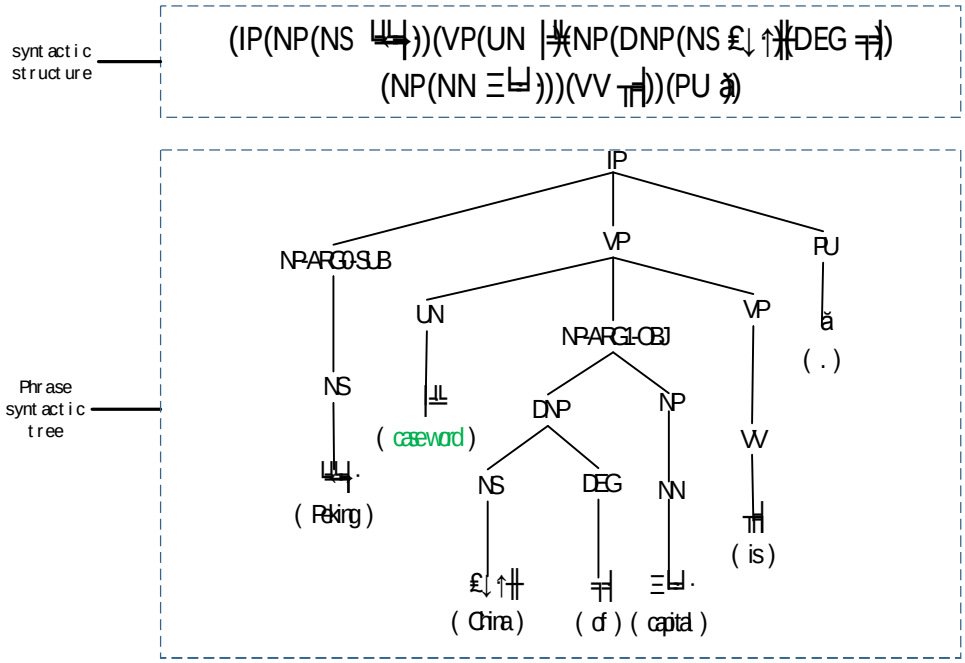

Fig 1 Tibetan syntax structure for information extraction

Tibetan Information Extraction Model Integrated with Semantic Feature

trigger words detection

Identifying the trigger word in a sentence is the first step of information extraction, but many verbs do not exist in the training corpus. For example: “respect teachers”. If "(respect)" do not appear as an event trigger words in the training corpus, it is not easy to recognize that it is an event. However, "(esteem)" and "(respect)" they are close in meaning, based on the dictionary of Tibetan verbs and words semantic similarity computation based on the dictionary of trigger words, automatically expand trigger words, covers a wide range of types of events trigger words as much as possible. Getting seeds information consists of a trigger Word tuple (trigger, type). Such as (President, Person/Respect), for the special case of a trigger word corresponds to more than one category, because the probability is small, we select most likely occurred in the training corpus that category, construction of "Tibetan trigger words-information classification" binary comparison table, binary comparison of two columns of trigger words and information for the event categories.

Table 1 “trigger word---information classification” comparison

Trigger word	Information classification
རྡོ་བ་ (daobi)	Business/Declare-Bankruptcy
འགོག་རྩོལ་ (kangzheng)	Conflict/Attack
རྒྱུ་སྐྱོལ་ (taoli)	Movement/Transport
ཚར་སྟོན་ (shiwei)	Conflict/Demonstrate
...	...

In table 1, the first column is the trigger word, the second column is triggered each word corresponding to the category of information, each trigger word corresponds to only one category of information, trigger word and the table covering all categories of information.

semantic role labeling

"The role of predicate-argument" form for each verb in the sentence (verb, such as verb or noun predicates) to mark the verb in a sentence appropriate semantic roles of constituents, including core semantic roles (agent, patient, etc) and the subsidiary body of semantic roles (such as where, when, how, why, and so on). Reference argument roles such as the main agent NP-SUB (ARG0), patient NP-OBJ (ARG1), and auxiliary arguments Orange such as, time ARGM-TEMP, ARGM-LOC, AGRM-MNR, and so on, which will, semantic information into a syntax tree, for training. Berkeley Parser is as the syntax training tool.

extraction algorithm

Parsing and decoding information from the Tibetan events, according to semantic information, output structured data.

First, using NP core Word extraction algorithm, extraction core semantic role, as agent (fozu), and patient (dizi), using VP core Word extraction algorithm extraction triggered word that predicate (jiang); and then, extraction subsidiary semantic information as locations (zangdi); last, on various semantic information for fusion get Tibetan event information, Figure 1 in the of phrases levels Shang of mark "%" representative child knot points in hid Han translation of when needs flip (reverse conversion grammar), so "(zangdi) (zai)" Should be translated as "zai zangdi".

Test and Evaluation

Experimental setup

(1) corpus selection:

Currently, also no special corpus for Tibetan information extraction, we use machine translation evaluation of bilingual parallel corpus library (and called central gold corpus), total 101629 sentences, selected 4001 sentence to 4800 sentence as training corpus, selected No. 4875 sentence to 5075 sentence, with 200 sentence as test corpus.

(2) results evaluation scheme

Tibetan information based on the extraction results into Chinese, and with the standard answer of Chinese sentences (Tibetan sentences sentence translation), extraction of Chinese sentence

similarity comparison of information for both, are made of BLEU to evaluation.

(3) experiment

Yang Jin corpus for Tibetan-Chinese bilingual sentence alignment corresponds to the Tibetan and Chinese translations in the training corpus sentence model of functional semantics information were added to the Tibetan-language parsing and syntactic analysis model for training, decoding by the Tibetan and Chinese grammar. Then decode the test corpora, decoded for Tibetan test corpora of grammar of the Tibetan language, Chinese decode grammar for Chinese test corpus.

test results

This Tibetan information taken artificially translation (machine translation, there are some mistakes), information, whole sentences in Chinese translation and Chinese extraction doing similar analysis results shown in table 2

Table 2 similar analysis

	sentences	BLEU
Chinese sentence extraction information	200	0.5832
Chinese translation of Tibetan sentence	200	0.2107

results analysis

We can conclude from table 4, Tibetan extraction information manual translated, similar with Chinese information extraction, but not many similar to the translation of the entire sentence, this is because extract information relative to the entire sentence, number is very small.

Due to the extraction of information (subject, predicate, object) relative to the entire sentence much less information, so lower similarity, but compared with other Chinese information extraction, the similarity is higher now BLEU Tibetan-Chinese machine translation values between 0.3 and 0.4, so after this extraction of Tibetan-language information is very accurate.

Conclusion and Future Work

Tibetan information integration semantic information extraction model, is based on the Tibetan phrase syntax, integration of the functional-semantic information for training, so as to support the Tibetan information extraction.

This Tibetan information extraction solution that provides a level of semantic information, including Tibetan semantic information classification and semantic information label, as well as Tibetan, Tibetan phrase parsing, and so on.

Contributions for other languages, including Mongolia, Uygur language and other minority languages provides a combination of information extraction program, as well as a comprehensive evaluation of machine translation programs, evaluation method is simple and practical.

This article is on the semantic level, next steps in the lexical syntactical level to enhance research work, studied Tibetan language lexical syntax integration model, which does not pass through Tibetan information extraction model for segmentation of the Tibetan-language parsing model based on syllable sequences, Tibetan information extraction service.

Acknowledgement

This research is supported by Key Laboratory of National Language Intelligent Processing , Gansu Province (Northwest University for Nationalities).

Reference：

[1] Naomi Daniel, Dragomir Radev and Timothy Allison. Sub-event based Multi-document Summarization [C]. In: Proceedings of the HLT-NAACL Workshop onText Summarization. 2003. pp9-16.

[2] W. W. Cohen, A. McCallum. Information Extraction and Integration: an Overview[C].KDD 2003 Tutorial, Washington DC, U.S.A, 2003:pp1-89.

[3] S. G. Soderland. Building a Machine Learning Based Text Understanding System[C]. Proceedings of workshop on Adaptive Text Extraction and Mining (ATEM-2001) at 17th International Joint Conference on Artificial Intelligence (IJCAI-2001), USA, 2001:pp133-154.

[4] 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.

[5] Alan Ritter, Mausam, Oren Etzioni and Sam Clark. Open Domain Event Extraction from Twitter[C]. In SIGKDD’12. 2012.

[6] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform[C]. In Proceedings of the Coling 2010:Demonstrations. 2010.08:pp13-16.