Research on Sino-Tibetan Machine Translation Based on the Reusing of Domain Ontology

http://dx.doi.org/10.3991/ijoe.v9iS4.2720

Xiaodong Yan, Xiaobing Zhao
Minzu University of China, Beijing, China

***Abstract*—There are some problems on traditional machine translation and these problems have affected its application. The combination of ontology and machine translation can bring about change to machine translation. Machine translation based on the reusing of domain ontology can effectively solve the problems that exist in traditional machine translation. The domain ontology knowledge base of Tibetan folk culture is tried to be built and the Sino-Tibetan machine translation system based on the reusing of Tibetan folk culture domain ontology is tried to be constructed in our research. In this machine translation system, ambiguity problem which exits in traditional machine translation can be solved and the domain ontology can be constructed automatically. By that intelligent machine translation between Tibetan and Chinese can be achieved.**

***Index Terms*—ontology, reusing of ontology, sino-Tibetan machine translation, domain ontology,**

I. INTRODUCTION

Machine translation is that one language is translated into another language by computer program automatically [1-2]. Since GB Al Choonee, a French scientists had the idea of machine translation in the 1930s, after decades of research and exploration, there has been made a lot of progress on machine translation. Some dedicated translation system has been the effective promotion and application, and achieved good economic and social benefits. However, the evaluation on the machine translation business software is already far from people's expectations. According to statistics, the translation accuracy and readability of Machine translation is 70 % less than human translation [3].

On Sino-Tibetan machine translation, in 2010 Qinghai Normal University had the project "Research on the key technologies of Sino-Tibetan machine translation for public opinion", and in which a Pandita Chinese and Tibetan translation system is developed. Pandita Chinese and Tibetan translation system has achieved the sentence translation in which rule-based translation method is used [4-5]. However, there are also Chinese and other languages machine translation facing problems that is, the accuracy of the translation and readability are not in a high level. Key issues include:

1. There are not enough rich language data. The development of machine translation dictionary is a huge project and requires a lot of manpower and time. The size of the machine translation system dictionary cannot meet application needs, especially for specialized dictionaries in some translation subsystem, such as the case library Dictionary, Valence Analysis dictionary; all of them need further expansion.
2. We can store data and execute task by computer, but we cannot think, reasoning, judge by computer. And there are also lack of artificial translator's knowledge and long-term accumulated cultural knowledge in computer, so we cannot have a comprehensive understanding of the original solely by computer. We can only do one to one choice within defined range by computer.
3. Machine translation is usually literal translation. It has not combined literal translation, explanation translation, and other translation methods together. So it is difficult to achieve high accuracy.
4. Translation is to express the primitive language with the most appropriate and most natural language. Because of the diversity due to cultural differences, the understanding of machine is far from human translation.
5. There are "one term multi-domain ambiguity" and "one term multi-domain synonymous" phenomenon. So there are Lexical ambiguities or redundant which are difficult to eliminate.

So we always expect to find better ideas and methods to overcome the above problems, providing more practical of the Sino-Tibetan translation system. The emergence of ontology provided a new method to solve semantic problems. Ontology is "a clear specification of conceptual system [6-7], and appear in the 1990s with the development of the knowledge base technology, its main goal is to achieve knowledge sharing and reuse. So far, ontology is widely used in knowledge management, information retrieval, natural language processing, semantic web, knowledge-based machine translation (Knowledge-based Machine Translation KBMT) .Research on ontology abroad began from the 1990s.Combining ontology with machine translation can bring changes in three aspects:

1. It can realize the automatic processing of machine translation.
2. It can provide the concept for the semantic expression of dictionaries and machine translation.
3. It can provide semantic space, helping semantic computing.

In natural language understanding and processing, in recent years, researchers have become increasingly aware of the importance of ontology, some pioneer began to try to build the ontology which serve for natural language processing, and hope to use ontology to express the knowledge of the world on the basis of ontology. In such studies, some have had great impact in the research fields, such as FrameNet [8], WordNet [9], HowNet [10], HNC (conceptual level network) [11] and so on. The Establish-

ment of Sino-Tibetan Machine Translation System Based on Reusing Domain Ontology

Reference to the structure of the machine translation system, we intend to establish a Sino-Tibetan machine translation system based on reusing domain ontology as Figure 1. According to the statement for translating that user inputted, the Chinese dictionary and Tibetan dictionary will be called according to the rule and template library. On base of domain ontology library, concept instantiation, semantic analysis and semantic computing are all did and target statement is outputted. As the main part of the knowledge library, ontology stored expertise knowledge in knowledge library. Through the reuse of domain ontology and iterative, the domain ontology library is constantly enriched and expanded. So the domain ontology library can provide structured, formalized domain knowledge for semantic analysis improving machine translation semantic understanding ability and enhancing the accuracy of the Sino-Tibetan machine translation.

II. CONSTRUCTION OF TIBETAN CULTURE DOMAIN ONTOLOGY LIBRARY

Build a ontology library on the common knowledge of the world is a time-consuming and labor-intensive project, it need us spend many years. And this kind of ontology brings about ambiguity which is difficult to overcome. Therefore, we have chosen certain areas of ontology so that we can in the limited time and resources to complete this study.

Our chosen field is Tibetan folk culture field, including clothing, funeral customs, accommodation, catering and so on. All of these corpuses derived from some Tibetan website, http://www.tibetinfor.com/, http://www.bodrigs.com/. These sites include Chinese text, Tibetan text. The specific steps are as follows:

1. Determining the field and scope of the ontology and extracting ontology concept independent of the specific language, describing the links between concepts.
2. Establishing ontology frame. In order to generate a term dictionary based on domain ontology, the concept hierarchy HC and relation hierarchy HR should be established on the concepts and the hierarchy of the concepts.
3. Delimiting the relationship between the concepts. Selecting those concepts corresponding objects exist independently from the concepts established on step 1 and express them by term.
4. Coding and formalizing domain ontology. Encoding and formalizing the domain ontology with appropriate ontology description language. Common ontology languages include protégé, ontoGngua, loom, OWL, Powerloom.
5. Iteration of domain ontology. Users will give information and feedback timely to the system and this will help the ontology construction engineering staff make the appropriate modifications and adjustment of the ontology, enriching domain ontology library.

III. ONTOLOGY FRAMEWORK

Domain ontology contains a large number of concepts and relationships for specific fields. So ontology concepts

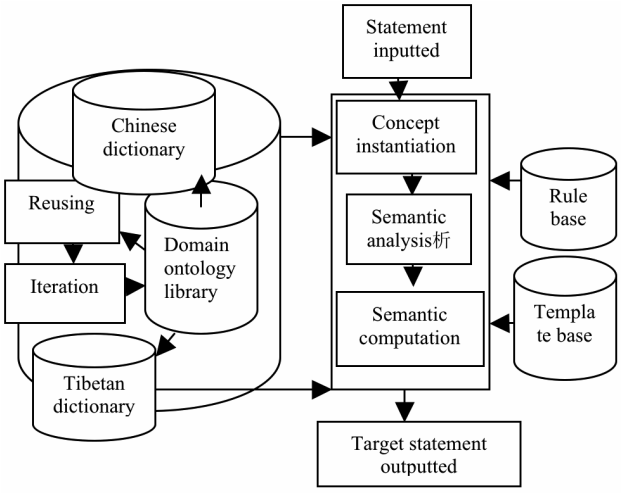

Figure 1. The structure of Sino-Tibetan machine translation system

are necessarily related to the specific subject area and it provides the relationship between the concepts. The relationship between these concepts, in turn, determines the framework of domain ontology. Usually a standard ontology head includes a set of XML namespace statement and the class concept of ontology will be defined in ontology body framework, which is mainly related to the extraction of concepts and the concept hierarchical processing. Although theoretically if there are head description of ontology and the domain ontology framework has been completed, the body of the domain ontology can be achieved by KDD technology expansion. In actual operation the relationships of ontology always are disordered.

(1) Extraction of concepts. The extraction of concept is generally after modeling and abstraction of concepts. In order to extract easy reused domain knowledge from the complex field knowledge system, conceptual modeling and abstraction are need. Extraction of concepts can be either based on clustering, in which the concepts are obtained by unsupervised learning from a large amount of web, text and database or based on a conceptual model which is from existing prior knowledge or experts in the field.

The Tibetan folk culture knowledge in our research is expressed by the usual method in the field, for example in the Tibetan folk culture, the concepts which are related to apparel are coral, agate, jade, Tibetan knives, Fire sickle, purse, snuff bottles, sewing kit, etc.. These concepts may be uncompleted but it can be used as the framework of this field concept ontology construction. Then in order to process these concepts it must be aware of the relationship between them.

(2) Hierarchical processing of the concept. The concept hierarchy is to classify and stratify the concept. Concept hierarchy is generally achieved through semantic hierarchy and it also can be carried out according to the experts in the field. For example: in the field of Tibetan folk culture, the concept classification includes clothing, food, shelter, transportation, tools, utensils, and burial customs. The concept tree is also used to represent the knowledge and hierarchy. The following is a part of the relationship tree of Tibetan folk culture propagation concept hierarchy (see Fig 2).

As can be seen from Figure 3, the folk culture character in Tibetan folk culture propagation ontology is stratified

by life function. The owl: Thing is the top-level concept, the life function is a special concept, and historical origins, routes of transmission, the regional and so on are all associated with the dissemination of Tibetan folk culture. The clothing, funeral and other concepts should be stratified according to the existing prior knowledge.

IV. ONTOLOGY KNOWLEDGE EXPANSION

KDD ontology expansion is to extend the ontology framework by KDD technology and gradually form a knowledgeable ontology. In order to speed up the comparison speed and convenience for RDFS / XML mapping, we first stratify the KDD rules in accordance with the existing concept of hierarchical mode. Meanwhile, the ontology extraction rule is also the same processing.

In order to achieve automatic extension, we did not use the conventional method to make KDD rules directly mapped to RDFS / XML document. We first make consistency maintenance on KDD rule concept and ontology rule concept based on the sample space and then mapped them to the ontology. Our research showed that if the KDD rules planned to add to ontology do not conflict with existing concepts in the ontology, the KDD rules mapping to ontology rules do not need the inference engine and this is help achieving automatically expansion of ontology. The specific implementation steps of automatic expansion of ontology are as following:

1. Concepts related with KDD rules are extracted from the ontology and then homogenization of them formed a priori rule set.
2. Concept hierarchy and homogenization of knowledge or rules obtained by KDD technology formed the new rule sets.
3. A certain amount of training samples are extracted and the consistency of the new rules which is planned to add to ontology and the priori rules are maintained in the sample space;
4. The new consistency rules or concepts will be mapped to the RDFS/XML documents;
5. Re-identifying of the new ontology and updating them (use protégé, the ontology building tool).

The above 4 steps have completed the automatic extension of the domain ontology, and step (5) just carried out the verification of the experimental results.

A. Testing and maintenance of the consistency of KDD rules and ontology rules

To ensure the uniqueness and consistency of domain field concept, there needs maintenance between KDD rules and the concept ontology. First, when the concept ontology which is planned to be expanded is only an empty frame, the consistency maintenance of it is mainly to ensure that the KDD rule itself has consistency. Second, if the concept ontology which is planned to be expanded has certain scale knowledge, the consistency maintenance of the ontology and KDD rules is important. In our research, we use a theoretical model of the sample space and the corresponding algorithm to check and eliminate conflict.

The priori rules based on the ontology include the meaning of its concept and rule, but the KDD rules do not include. Moreover KDD rule is proposed by certain support and all samples of it are limited. So in the process of consistency checking, maintenance of KDD rules and ontology concept, the ontology is always priori to others. It means that the new mined rules may not contradict to the priori rules in ontology, which is different from the consistency maintenance of KDD rules.

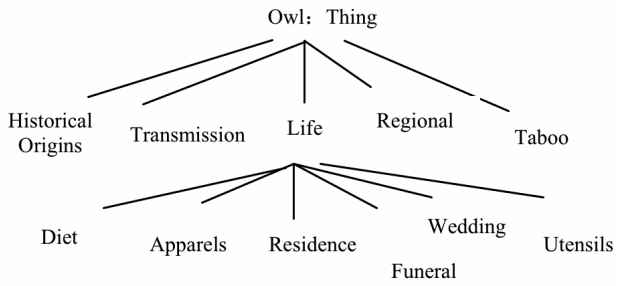

Figure 2. the relationship tree of Tibetan folk culture propagation concept hierarchy

B. Conversion and mapping between KDD rules and ontology rules

Conversion and mapping between the KDD rules and ontology rules includes the extraction of RDFS / XML concepts and rules and the mapping from KDD rules to RDFS / XML. Although this not means that these two kinds of rules are One-to-one correspondence. We will not discuss the concepts and rules which cannot be converted and mapped here. Because the new rules based on KDD rules are just supplementary of ontology priori rules. Moreover the extraction of ontology concepts will affect the accuracy and availability of the results. Steps of mapping from KDD rules to ontology are as following:

(1) Searching the sub-genus and seed-genus of KDD rules and removing the implication operator of KDD rules;

(2) Searching the sub-genus and seed-genus of existing ontology. That is, to find the most specific class, and add the latter transformation and mapping to the corresponding subclass description of ontology;

(3) If there is not Property in ontology, then we can create the corresponding attribute and transmit the Property of KDD rules to the Property of ontology and add it to the base class. Then we create the property according to the existing concept hierarchy principles;

(4) Adding the re-striation to the basic class.

V. ELIMINATION OF AMBIGUITY

Ambiguity problem is one of the most difficult to deal with problems of machine translation, so we will resolve this problem by ontology.

In the system, ambiguity is mainly from two aspects: syntactic analysis will produce a variety of possible syntactic parse tree and it is structure ambiguity; One vocabulary is corresponding multiple semantic concepts. It is lexical ambiguity.

We will eliminate the ambiguity with following methods:

(1) Using the characteristic inspection mechanism during the syntactic analysis process. In this method the syntactic information of vocabulary is important. When a syntax tree is generated, the structure of it is checked for whether it is compliance with the syntactic restriction. If

there is not object in a sentence, it will be judged that it is not comply with the syntax because of not meeting the dictionary definition. Ungrammatical syntactic analysis can be reduced by this means. Similar to English, Tibetan has a variety of morphological changes of the language and has enough syntactic information.

(2) Checking the semantic restrictions by use of the semantic fragment combination. As the sub-module of semantic analysis, the semantic fragment combination is not just some of the combined semantic description of fragments, but also the semantic of it must be checked. It is whether it is "acceptable". "Acceptable" is defined as following:

Assuming that the hierarchical relationships of 4 concepts A, B, C, D is shown in Figure 5, and if the semantic restrictions of agent x of word m in the dictionary is B, B itself and the concept A,D are all acceptable for x of word m. But if concept C is neither the super ordinate concepts of B nor the lower concept of B and there is no other defined relation and the associated concept of B, then C is unacceptable for agent x of word m. Here we also think that the lower concepts are comply with the all character of original concept and can certainly meet their semantic restrictions. The Super ordinate concepts are comply with some character of original concepts and can partially meet semantic restrictions.

(3) The weight of structure is defined by use of ontology and the ambiguity is eliminated by use of semantic description. In some cases, the difference of multiple semantics of one word is not manifest and it may also be able to meet semantic restrictions of other word. At this point, we can calculate the semantic weight to eliminate the ambiguity.

In Figure 3, for example, assuming that the semantic structure of the word M in the dictionary is (sem% G (agent var1 (sem% B), in the sentence word n is mapped as the agent of m. If the semantic concept of n is B or D, the weight of is 1.0. If the semantic concept of n is A, the weight of it is 0.9. If the semantic concept of n is x which has relation with B, then the weight of it is 0.8. Semantic restrictions also exist between the words and phrases in addition to between vocabulary and vocabulary.

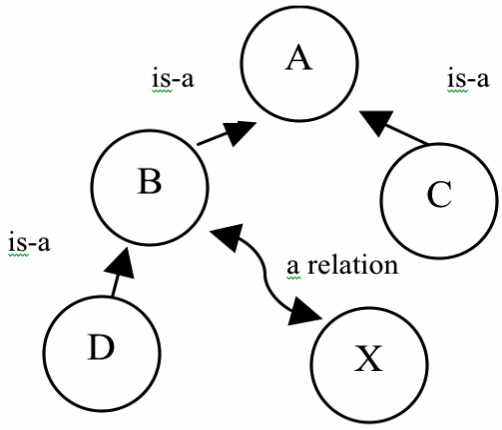

Figure 3. the relationship of semantic concepts

VI. CONCLUSION

We studied machine translation based on the reusing of domain ontology. We have built domain ontology knowledge base of Tibetan folk culture and constructed the Sino-Tibetan machine translation system based on the reusing of Tibetan folk culture domain ontology. In this machine translation system, ambiguity problem which exits in traditional machine translation can be solved and the domain ontology can be constructed automatically. By that intelligent machine translation can be achieved.

ACKNOWLEDGMENT

The work in this paper is supported by “the Funds of Minzu Universities Independent research projects”-research on emergency processing technology in some universities.

REFERENCES

[1] Kelsang. ChiMa, “Sino-Tibetan machine translation Exploration, ” *Southwest University of Science and Technology*, Vol.1, 1997,pp.84-88.

[2] Feng. Zhiwei, *New theory of natural language machine translation*, Languages Press, 1994.

[3] Wei. Yongpeng, Chen. Qun, “The design and realization of multi-language machine translation support environment,” in Proceedings of the Conference on Chinese Information Processing, Beijing, China, 7-8 July 2004, pp. 9-16.

[4] Cai. Zangtai, “Research on Syntax Analysis in the rules-based Sino-Tibetan machine translation system, ” in Proceedings of the Conference on the third ethnic minority youth Natural Language Information, Beijing, China, 22-24 July, 2010, pp. 128-130.

[5] Xiaodong. Yan, Xiaobing. Zhao, Guosheng. Yang, “Real-time building, enlarging and tracking Tibetan, Uighur sensitive word vocabulary, ” *Journal of convergence Information technology* (JCIT), vol. 7, Issue 2, 2012, pp 367-374.

[6] Zhaluo, Suonanrenqian. “Research on complex sentence translation Principle in Sino-Tibetan machine translation, ” *International Journal Human computer Studies*, vol. 43, Issue 6, 1995, pp. 907-928.

[7] Chi. Xiu, “The effect of Word Segmentation and Alignment on Statistical Machine Translation, ”*International Journal of Digital Content Technology and its Applications* (JDCTA), vol. 7, Issue 2, 2013, pp.164-171.

[8] International Computer Science Institute web portal (http://www.icsi.berkeley.edu/framenet/)

[9] WordNet Web Portal (http://wordnet.princeton.edu/)

[10] HowNet Web Portal (http://www.keenage.com/)

[11] HNC Web Portal (http://www.hncnlp.com/)

AUTHORS

Xiaodong.Yan. Author is with the Minzu University of China, Beijing, CO 68932421 CHINA (e-mail: yanxd3244@sohu.com).

Xiaobing. Zhao. Author, Jr., was with Minzu University of China, Beijing, CO 68932421 CHINA (e-mail: nmzxb@163.com).

This article is an extended and modified version of a paper presented at the International Conference on Mechanical Engineering, Automation and Material Science (MEAMS2012), held 22-23 December 2012, Wuhan, China. Manuscript received 26 January 2013. Published as resubmitted by the authors 01 May 2913..