Text Meaning Representation for Chinese

Wanying Jin
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003-8001
U.S.A.
wanying@crl.nmsu.edu

Abstract

This paper describes text meaning representation for Chinese. Text meaning representation is composed of a set of ontological concept instances along with ontological links among them. It integrates lexical, textual and world knowledge into a single hierarchical framework. In NLP application it serves as an interlingual representation for various processing. The methodology and implementation of text meaning representation is discussed in detail. [1]

1 Introduction

In natural language text processing, it becomes inevitable to require a system having capability to automatically extract and represent the information conveyed in a given text. The theory and methodology of text meaning representation (TMR) has been studied in the past decade (Nirenburg, 1991; Mahesh and Nirenburg, 1996; Beale *et al.*, 1995; Onyshkevych 1997) and its application has been presented in machine translation systems that employ interlingual approach. An ideal text meaning representation will be a language-neutral description of the linguistic information conveyed in a natural language text. TMR captures the meanings of words in the text and represents them in a set of ontological concepts interconnected through ontological relations. In addition, TMR provides information about the lexicon-semantic dependencies as well as stylistic factors. Based on this study, two supportive resources to compose TMR are ontology and semantic lexicon. An ontology is a set of knowledge concepts about the world. It is composed of thousands of concepts organized into a particular hierarchy so that each concept is related to other concepts through semantic links (Carlson *et al.*, 1990; Bateman, 1993; Dowell *et al.* 1995; Nirenburg *et al.*, 1995; Bouaud, *et al*, 1995; Mahesh *et al.* 1995;). Semantic lexicon represents the senses of the words. Each lexicon entry is a framework that maps a word sense to an ontological concept. It also includes information about syntax, semantics, morphology and pragmatics, as well as annotation that keep the record of data management. Over the years, various methodologies have been investigated carrying out the structure of a computational lexicon entry within a knowledge-based framework(Onyshkevych *et al*, 1995; Viegas and Raskin, 1998; Viegas, 1999; Viegas, *et al*, 1998a, 1998b). A semantic parser uses information in the semantic lexicon and makes a decision on word sense disambiguation based on the strategy proposed by Beale *et al.* (1995) and Beale(1997).

This paper presents the development and application of text meaning representation for Chinese. Detailed discussion about the principle for building a computational semantic Chinese lexicon is illustrated in Jin (1999). The methodology and implementation of word sense disambiguation is fully discussed in Viegas, Jin and Beale(1999a, 1999b, 1999c).

2 Overview of Ontology

An ontology is a body of knowledge about the world. It is a repository of concepts used in meaning representations. All concepts are organized in a tangled subsumption hierarchy and further interconnected using a system of semantic relations defined among the concepts. The ontology is put into well-defined

[1] This work has been supported by the Department of Defense of the United States under contract number MDA-904-92-C-5189.

relationships with knowledge sources in the system. In an NLP application the ontology supplies world knowledge to lexical, syntactic, semantic and pragmatic processes.

In the MikroKosmos project,[2] the ontological concepts consist of: OBJECT, the static things existing in the world; EVENT, any activities happening in the world, and PROPERTY, the properties of OBJECTs and EVENTs. The ontology organizes terminological nouns into a taxonomy of objects, verbs into a taxonomy of events, and adjectives into a taxonomy of attributes. It further includes many ontological relations between objects and events to support a variety of disambiguation tasks. Currently, the ontology in the MikroKosmos project contains about 5,000 concepts covering a wide range of categories in the world. Each concept, on average, has 14 relation links. An example below presents the top three level of the ontology, which differentiates between OBJECT; EVENT and PROPERTY.

```
+ ALL
  + EVENT
    + MENTAL-EVENT
    + PHYSICAL-EVENT
    + SOCIAL-EVENT
  + OBJECT
    + MENTAL-OBJECT
    + INTANGIBLE-OBJECT
    + PHYSICAL-OBJECT
    + SOCIAL-OBJECT
  + PROPERTY
    + ATTRIBUTE
    + RELATION
    + ONTOLOGY-SLOT
```

Each concept is a frame in which a collection of ontological slots such as DEFINITION, IS-A, SUBCLASSES, INVERSE; case roles such as AGENT, THEME and properties such as HEADED-BY, HAS-MEMBER, etc. link one concept to other concepts in the ontology. Below is an example of a concept frame of GOVERNMENT-ACTIVITY:

```
Concept      GOVERNMENT-ACTIVITY
DEFINITION:  an activity commonly
             carried out by a
             government.
IS-A:        POLITICAL-EVENT
AGENT:       HUMAN
THEME:       EVENT OBJECT
ACCOMPANIER: HUMAN
LOCATION:    PLACE
```

This example indicates that the concept GOVERNMENT-ACTIVITY is a subclass of the concept POLITICAL-EVENT; its case role AGENT requires its semantic value as HUMAN and its THEME requires its value as either OBJECT or EVENT; the GOVERNMENT-ACTIVITY can also have case role ACCOMPANIER with value as HUMAN and LOCATION with value as PLACE. Any lexicon entry mapping to the concept GOVERNMENT-ACTIVITY gets extended information through this frame.

3 Semantic Lexicon

Semantic lexicon is another knowledge source for text meaning representation. In the MikroKosmos project each lexicon entry is designed as a frame with 11 zones corresponding to information relevant to orthography, morphology, syntax, semantics, syntax-semantic linking, stylistics, and database type management record, etc. The core of the lexicon frame is syntactic zone SYN-STRUC, semantic zone SEM-STRUC and their link SYNSEM zone. Syntax particular to a given language is described in the syntactical zone. The semantic zone maps a sense into an ontological concept in the case of single sense, or to several concepts in the case of multiple senses. Through the syntactic-semantic link zone the information of each word in the text can be extracted directly from lexicon database and its relevant world knowledge also can be retrieved. The general template of semantic lexicon entry is shown as follows. For a detailed description see Viegas and Raskin (1998).

```
Entry Elements =
[FORM:       Word Form
 CAT:        Part of speech
 MORPH:      Morphology
 ANNO:       Annotation
 TRANS:      Translation
```

[2]The MikroKosmos project is a knowledge-based machine translation system using an interlingual approach. The source languages are Spanish, Chinese and Japanese. The target language is English.

```
SYNSEM:      Syntax-Semantic link
SYN-STRUC:   Syntactic structure
SEM-STRUC:   Semantic structure
LEX-RULES:   Lexical rules
PRAGM:       Pragmatic information
STYL:        Stylistic information]
```

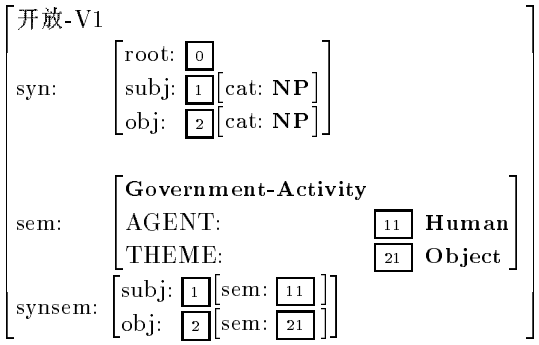

Figure 1: Sense for the Chinese lexicon 开放.

Figure 1. is a simplified structure of a Chinse lexicon entry 开放 in the sense of {GOVERNMENT-ACTIVITY. The SYN zone indicates that when parsing a sentence containing this entry, subcategories SUBJ and OBJ are required. The SEM zone presents the semantic value of each case role, i.e. AGENT with value HUMAN, and THEME with value EVENT or OBJECT. The SYNSEM zone provides information about the syntax-semantic linking. That is, SUBJ [1] is linked to AGENT [11] with value HUMAN and OBJ [2] is linked to THEME [21] with value EVENT or OBJECT.

Due to the lack of morphological information in Chinese, it is often the case that the same Chinese word form can be mapped to a different part of speech and has multiple senses, such as the word 开放:

- in the context 鲜花 开放 *flowers bloom*, can be an intransitive verb mapping to a concept BLOOM with the definition *to produce flower.*
- in the context 政府 开放 对外 贸易 政策 *the government opens the foreign trade policy*, can be a transitive verb mapping to a concept GOVERNMENT-ACTIVITY with the definition *an activity that is commonly carried out by a government at any level.*
- in the context 政府 实行 开放 政策 *the government carries open policy*, can be an adjective mapping to OPEN-TO-PUBLIC with the definition *to be available to the public.*
- in the context 图书馆 开放 *the library is open*, can be an intransitive verb with the same concept OPEN-TO-PUBLIC.

Using the ontological concepts as the value of semantic variables and linking them to syntactic variables makes the lexicon very informative. Figure 2. and Figure 3. present each sense and POS of the lexicon entries for the Chinese word 开放.

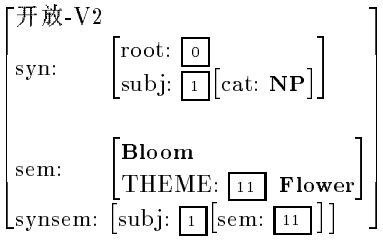

Figure 2: Sense for the Chinese lexicon 开放.

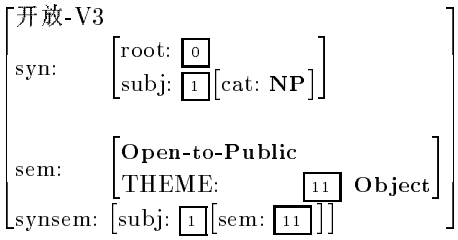

Figure 3: Sense for the Chinese lexicon 开放.

4 Semantic Analysis for Word Sense Disambiguation

The task of a semantic analyzer is to combine the knowledge contained in the ontology and lexicon and apply it to the input text to produce text meaning representation output. The central tasks involved are to retrieve the appropriate semantic constraints for each possible word sense, test each sense in context, and construct the output TMRs by instantiating the concepts in semantic zones of the word senses that best satisfy the combination of constraints. Figure 4. illustrates the process of text meaning representation. Below illustrates the process through a sentence 中国

政府 开放 了 对外 贸易 政策. *The Chinese government has opened foreign trade policy.* The syntactic analysis gives the following output:

```
((ROOT 开放)(CAT V)(TRANS open)
  (SUBJ
   (MODS 中国)(CAT N)(TRANS China)
   (ROOT 政府)(CAT N)(TRANS government))
  (OBJ
   (MODS
    (MODS 对外)(CAT ADJ)(TRANS foreign)
    (ROOT 贸易)(CAT N)(TRNAS trade))
   (ROOT 政策)(CAT N)(TRANS policy)  ))
```

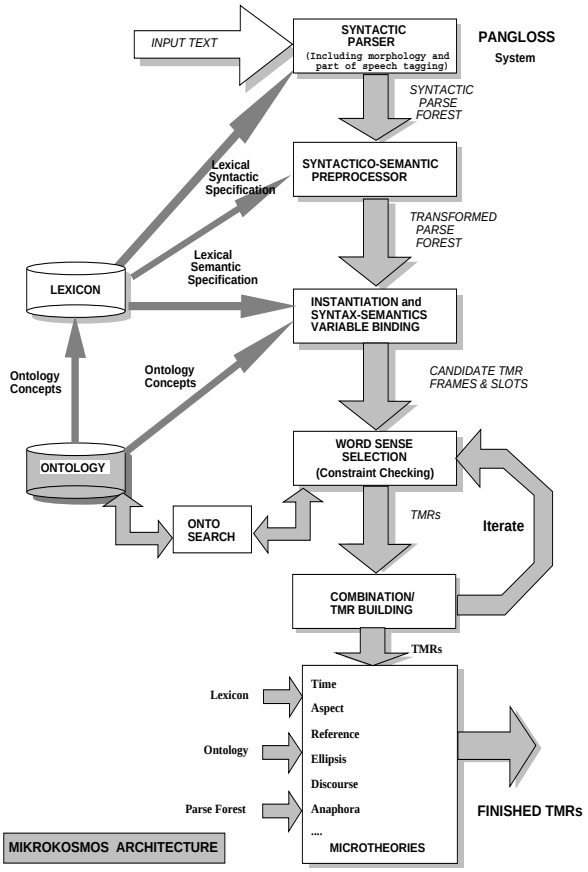

Figure 4: The Architecture of TMR.

The semantic analysis process takes the following steps:

- to gather all of the possible lexicon for each of the words with instantiated each concepts.

```
中国  N    CHINA-6
政府  N    FEDERATION-1
开放  V1   BLOOM-2-1
      V2   GOVERNMENT-ACTIVITY-2-2
      V3   OPEN-TO-PUBLIC-2-3
      ADJ  OPEN-TO-PUBLIC-2-4
对外  ADJ1 INTERNATIONAL-ATTRIBUTE-5-1
      ADV1 OUTWARD-5-2
贸易  N    COMMERCE-EVENT-4-1
      V    COMMERCE-EVENT-4-2
政策  N    LAW-3
```

Syntactic variables are bound to one another using the syntactic patterns in the lexical entries to establish syntactic dependencies. In addition, ontological concepts referred to the semantic zones of the lexical entries are instantiated and linked through ontological relations to establish semantic dependencies. For example, the syntactic structure of the text requires 开放 to be a verb. Thus, the ADJ category with sense OPEN-TO-PUBLIC-2-4 is rejected. From Figure 2. and Figure 3. both SYN zones indicate an intransitive verb that violates the required syntax. Therefore, the concepts BLOOM and OPEN-TO-PUBLIC are also rejected. In the same way, the adverb 对外 with sense OUTWARD-5-2 and the verb 贸易 with sense COMMERCE-EVENT-4-2 are also rejected because of the violation of required POS. Finally the ADJ 对外 with the sense INTERNATIONAL-ATTRIBUTE-5-1 and the NOUN 贸易 with COMMERCE-EVENT-4-1 are selected. After all senses are determined, SYN-SEM zone binds all syntactic variables with semantic variables, i.e. SUBJ FEDERATION-1 is bound to the AGENT of GOVERNMENT-ACTIVITY-2-2, OBJ LAW-3 is bound to the THEME of GOVERNMENT-ACTIVITY-2-2.

- In the next step, selectional constraints are retrieved from the ontology. Individual selectional constraints are checked. In the example, the concept GOVERNMENT-ACTIVITY requires AGENT to be HUMAN and THEME to be EVENT or OBJECT. The lexical information indicates that the SUBJ 政府 with sense FEDERATION must satisfy the AGENT of GOVERNMENT-ACTIVITY with value HUMAN. An inference rule described below checks the satisfaction. The OBJ 政策 with the sense LAW satisfies THEME of GOVERNMENT-

ACTIVITY with value OBJECT. Through IS-A links it is found that LAW is a descendant of OBJECT. Therefore, the semantic constraints are satisfied.

- Seeking satisfaction through inference rules, the semantic analyzer does more than match selectional constraints or find the distance along IS-A links. The search inside the ontology also involves looking for metonymic type links, such as FEDERATION in a metonymic relation with HUMAN through the property HAS-REPRESENTATIVE:

```
Concept:   HAS-REPRESENTATIVE
IS-A:      ORGANIZATION-RELATION
DOMAIN:    ORGANIZATION
RANGE:     HUMAN BUSINESS-ROLE
           GOVERNMENTAL-ROLE
INVERSE:   REPRESENTATIVE-OF
```

  in which DOMAIN is ORGANIZATION that has subclass FEDERATION and RANGE is HUMAN. Thus, the constraint of AGENT of GOVERNMENT-ACTIVITY to be HUMAN is satisfied.

- In case multiple senses all satisfy the constraints, the concept with the shortest path is selected as the best choice. An ontological search program, Onto-Search, is presented in Onyshkevych (1997). The resulting preference values for each constraint are combined in an efficient control and search algorithm called Hunter-Gatherer that combines constraint satisfaction, branch and bound, and solution synthesis techniques to pick the best combination of word senses of the entire sentence in near linear time, as described in Beale (1997).

- Chosen word senses are assembled into TMR frames.

5 Text Meaning Representation

A text meaning representation(TMR) is a language-neutral description of the meaning conveyed in a text. It is derived by syntactic and semantic analysis. TMR captures not only the meaning of individual words in the text, but also the relation between those words. It provides information about the lexicon-semantic dependencies. In addition, it also represents stylistic and other factors presented in the text. From the result of word sense disambiguation, TMR integrates lexical, ontological and textual information into a single hierarchical framework. Below is a TMR for the example sentence 中国 政府 开放 了 对外 贸易 政策. *Chinese government has opened foreign trade policy.*

```
(FEDERATION-1
 (AGENT-OF
   (VALUE GOVERNMENT-ACTIVITY-2))
 (RELATION (VALUE CHINA-6))
 (INSTANCE-OF (VALUE FEDERATION)))
(GOVERNMENT-ACTIVITY-2
 (AGENT (VALUE FEDERATION-1))
 (THEME (VALUE LAW-3))
 (INSTANCE-OF
   (VALUE GOVERNMENT-ACTIVITY)))
(LAW-3
 (THEME-OF
   (VALUE GOVERNMENT-ACTIVITY-2))
 (INSTANCE-OF (VALUE LAW)))
(COMMERCE-EVENT-4
 (EVENT-OBJECT-RELATION (VALUE LAW-3))
 (INSTANCE-OF (VALUE COMMERCE-EVENT)))
(INTERNATIONAL-ATTRIBUTE-5
 (DOMAIN (VALUE COMMERCE-EVENT-4))
 (RANGE (VALUE INTERNATIONAL))
 (INSTANCE-OF
   (VALUE INTERNATIONAL-ATTRIBUTE)))
(CHINA-6
 (INSTANCE-OF (VALUE CHINA)))
(ASPECT-7
 (SCOPE (VALUE GOVERNMENT-ACTIVITY-2))
 (TELIC (VALUE YES)))
(TIME-8
 (DOMAIN (VALUE GOVERNMENT-ACTIVITY-2))
 (RANGE (VALUE *speaker-time*)))
(INFERENCE (TYPE METONYMY)
  (HUMAN
    (REPRESENTATIVE-OF
      (VALUE FEDERATION-1))))
```

After semantic analysis, a variety of microtheories are applied to further analyze elements of text meaning such as time, aspect, propositions, sets, co-reference, and so on, to produce a complete TMR. In the example, ASPECT-7 is applied within the scope of GOVERNMENT-ACTIVITY in which TELIC with value YES indicates the GOVERNMENT-ACTIVITY is complete that means the action

of opening foreign trade policy is done. TIME-8 indicates the GOVERNMENT-ACTIVITY happens at the time the speaker make the utterance. Thus, the meaning of the Chinese sentence 中国 政 府 开放 了 对外 贸易 政策 is completely represented in the TMR.

6 Discussion

A knowledge-based machine translation can be viewed as extracting and representing the meaning of a text and generating a text in target language based on the meaning presented. Thus, text meaning representation plays the key role in an interlingual approach to machine translation. The approach described in this article enables integrating linguistic knowledge of source languages with general world knowledge to reach high quality translation. It is because TMR represents meaning deeper and broader than what the context presents. For example, in the context 提供 服务 *provide service*, the linguistic information indicates syntactic-semantic dependency as SUBJ*(human)* -V*(service-event)*-OBJ*(event)*. In an ontology, the SERVICE-EVENT concept frame contains information about AGENT and THEME as well as BENEFICIARY, ACCOMPANIER, INSTRUMENT, LOCATION, etc. which extends the meaning of the given text 提供 服务 to 某人 在 某处 为 某人 提供 服务 *Someone provides service to someone else at some place.* If the extended information is not explicitly presented in the text, the default value provides the assumption based on the world knowledge. In machine translation, it enables the generation of high quality text, especially in the case where the syntax in source language is different from that in target language or in the case where ellipsis is allowed in one language such as *He plays Bach* in English, but is not allowed in other language such as in Chinese, where one must say 他 演奏 巴哈 的 作品 *He plays Bach's work.*

With the rapid development of internet, information retrieval plays the key role in search engin. The extended information about the world knowledge allows to retrieve relevant data through the ontology that is implicitly specified in the query. As result, more broad and deep information can be extracted. It is extremely valuable to the development of next generation of internet search engin. All in all, using ontology strengthens all NLP systems.

References

Bateman, J. A. 1993. Ontology Construction and Natural Language. *Proc. International Workshop on Formal Ontology*, Padua, Italy.

Beale, S. 1997. *Hunter-Gatherer: Applying Constraint Satisfaction, Branch-and-Bound and Solution Synthesis to Computational Semantics.* Ph.D. Diss., Carnegie Mellon University.

Beale, S., Nirenburg, S. and K. Mahesh. 1995. Semantic Analysis in the MikroKosmos Machine Translation Project. *Proc. of the 2nd SNLP-95*, Bangkok, Thailand.

Bouaud, J., Bachimont, B., Charlet, J., and Zweigenbaum, P. 1995. Methodological Principles for Structuring an ONTOLOGY. *Proc. the Workshop on Basic Ontological Issues in Knowledge Sharing*, International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada.

Carlson, L. and Nirenburg, S. 1990. *World Modeling for NLP.* Technical Report CMU-CMT-90-121, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, PA.

Dowell, M., Stephen, L., and Bonnell, R. 1995. Using a Domain Knowledge Ontology as a Semantic Gateway among Database. *Proc. the Workshop on Basic Ontological Issues in Knowledge Sharing*, International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada.

Jin, W., Viegas, E. and Beale, S. 1999. Building a Chinese Computational Semantic Lexicon *Proc. of International Symposium on Machine Translation and Computer Language Information Processing-1999 (ISMT & CLIP)*, Beijing, China.

Mahesh, K. and Nirenburg, S. 1995c. Semantic Classification for Practical Natural Language Processing. *Proc. Sixth ASIS SIGICR Classification Research Workshop: An Interdisciplinary Meeting.* Chicago, IL.

Mahesh, K and Nirenburg, S. 1996. Meaning Representation for Knowledge Sharing in Practical Machine Translation. *Proc. the FLAIRS-96 Track on Information Interchange.* Florida AI Research Symposium.

Mahesh, K. 1996. *Ontology Development for Machine Translation: Ideology and Method-*

ology. MCCS-96-292, Computing Research Laboratory, New Mexico State University.

Nirenburg, S., Raskin, V. and Onyshkevych, B. 1995. Apologiae Ontologiae. *Proc. of The Conference on TMI*, Leuven, Belgium.

Nirenburg, S. 1991. Application-Oriented Computational Semantics. *Computational Linguistics and Formal Semantics*, R. Johnson and M. Rosner (eds.)

Onyshkevych, B. 1997. *An Ontological-Semantic Framework for Text Analysis.* Ph.D. Diss., School of Computer Science, Carnegie Mellon University.

Onyshkevych, B. and Nirenburg, S. 1995. A Lexicon for Knowledge-Based MT. *Machine Translation Issue on Building Lexicon for MT*, B. Dorr and J. Klavens (eds.) 10:1-2, 5-57.

Viegas, E. and Raskin. V. 1998. *Computational Semantic Lexicon Acquisition: Methodology and Guidelines.* MCCS-98-315. Computing Research Laboratory, New Mexico State University.

Viegas, E. 1999. An Overt Semantics with a Machine-guided Approach for Robust LKBs. In Proc. of *SIGLEX99 Standardizing Lexical Resources*, University of Maryland.

Viegas, E., Beale, S. and S. Nirenburg. 1998. The Computational Lexical Semantics of Syntagmatic Relations. *Proc. of the 36th ACL and the 17th COLING*, Montréal, Québec, Canada.

Viegas, E., Jin, W. and Beale, S. 1999. A Knowledge-based Approach for Chinese-English Translations *Proc. of 5th Natural Language Processing Pacific Rim Symposium (NLPRS-99)*, Beijing, China.

Viegas, E., Jin, W. and Beale, S. 1999. Long Time No See: Overt Semantics for Machine Translation *Proc. of Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, England.

Viegas, E., Jin, W. and Beale, S. 1999. Using Computational Semantics for Chinese Translation *Proc. of Machine Translation Summit-99*, Singapore.