Build a Situation-based Language Knowledge Base

Qiang ZHOU, Zushun CHEN

State Key Laboratory of Intelligent Technology and Systems
Dept. of Computer Science and Technology
Tsinghua University, Beijing 100084, P. R. China
zq-lxd@tsinghua.edu.cn

Abstract. Language resources are very important for natural language processing research and applications. This paper will introduce our ongoing research work to build a situation-based language knowledge base for the Chinese language, based on two basic language resources: three Chinese semantic lexicons and a large scale Chinese treebank. We developed a supporting platform to make full use of the abundant information contained in current Chinese semantic lexicons so as to gradually summarize the complete situation descriptions, organize them as situation network and build corresponding descriptive definition dictionary for different concepts. We explored an efficient algorithm to link from syntax to semantics so as to introduce suitable semantic explanations into current Chinese treebank and gradually build a situation-based semantically-annotated corpus. All these research work will lay a good foundation for the computational infrastructure in Chinese natural language processing.

1 Introduction

Language resources are very important for natural language processing research and applications. Therefore, in recent years, many researchers have devoted themselves into the construction of large scale language resources. Nowadays, there are two types of commonly used language resources. One is the syntactically annotated corpus. Some typical examples include the Penn Treebank for English [9], the Prague Dependency Treebank for Czech [6] and the TIGER treebank for German [3]. The other is the semantic lexicons. Most of them are manually compiled by linguists or lexicographers. Some typical examples are the WordNet [10] and Levin's English verb classes [8]. The key issue is how to integrate these two types of language resource so as to build the linking bridge between syntax and semantics. The Proposition Bank (PropBank) [7] and FrameNet [1] projects have made some tentative explorations in these respects.

Unlike above research projects, we proposed a new situation-based language knowledge description framework. Under this framework, we use a situation as a mathematical model to describe a cognition scheme and try to define a concept under its generating situation. Therefore, the situation theory can serve as a unified theoretical framework for constructing the lexical semantics and the natural language knowledge infrastructure built upon it. This paper will introduce our ongoing

research work to build suit a situation-based language knowledge base for the Chinese language, based on two basic language resources: three Chinese semantic lexicons and a large scale Chinese treebank.

2 The situation-based knowledge framework

In our opinion, a concept is generated in a peculiar cognition scheme, which will be called its generating scheme. The new concept absorbs and condenses the new knowledge contained in the scheme to gradually form a relatively stable individual that can be independently quoted in furthermore cognitive activities. At that moment, we can coin a new word (or phrase) to name the concept so that it can be easily used or quoted in common communication and conceptual thought. So the word becomes the symbolic embodiment of a concept. The basic and the most important attributes of the concept are issued in its generating scheme. We cannot describe and define the concept clearly unless we put it into its generating scheme.

We proposed to use the situation [2] as a mathematical model to describe a cognition scheme. Therefore, the situation theory [2] can serve as a unified theoretical framework for constructing the lexical semantics and the natural language knowledge infrastructure built upon it. Under this framework, many new issues should be explored, including: (1) how to use a situation to express a scheme and use a situation to describe a concept; (2) how to formulate the situation algebra for describing the relations, transformations, and operations among situations so as to simulate conceptual thinking by means of algebraic calculation; (3) how to construct a situation network to implement a scheme structure and conceptual structure, where the key point is the constitution and organization of a semantic dictionary. Chen and Zhou (2002) discussed more detailed questions about them.

3 Supporting platform for situation development

To build a large-scale situation-based language knowledge base for the Chinese language, we proposed a two-stage approach method.

At the first stage, we manually summarize some commonly-used typical cognition schemes under intuitional thought on several semantic lexicons. Then, we construct rough situations to express these cognition schemes so as to reflect the key information among them. These situations can be organized into an initial situation network, based on the basic ontological classifications under four main domains: physical world, mental world, symbolic world and human world.

At the second stage, we search and extract a group of word entries or concepts with coherent relations for a special situation and try to define or describe these conceptual meanings by using this situation. In the process of defining concept, the situation description can be refined to reflect more detailed cognitive contents, and a new semantic dictionary can be built. In the dictionary each word entry can be assigned a suitable situation-based definition for its conceptual meaning. So it is an interdependent and interaction process for constructing both situation descriptions and semantic dictionary.

To make the above construction method more feasible, we developed a supporting platform, under which three Chinese semantic lexicons were merged to form the basic semantic resources and many useful tools were developed to make full use of the semantic knowledge defined among them.

Based on the platform, we have summarized about 50 situations now. Some of them are basic situations for further descriptions, such as 'existence', 'maintain', 'transfer', 'destroy', etc. Some are detailed situations for organization name identification, such as 'transaction', 'transport', 'manufacture', etc. All of them are related with hundreds of words and concepts. We hope to build a small size situation network based on them at the end of this year.

4 Bridge the gap between syntax and semantics

Apart from the above semantic lexicons, another useful language resource for situation development is the large-scale annotated corpus, where different kinds of language performance phenomena will bring in the paraphrase problems that a semantic content may be expressed by a variety of ways. If we can anchor the parameter of a situation to the suitable referring expressions in real world sentences, we will obtain a new viewpoint to study the implementation of a concept in the procedure that contrasts, restores, and refers to its generating situation in a special contextual environment.

Nowadays, we have built a large-scale syntactically-annotated Chinese corpus: the Tsinghua Chinese Treebank (TCT) [11], which contains about 1,000,000 Chinese words of texts drawn from a balanced collection of texts published in 1990s. Here the key issue is how to bridge the gap between syntax and semantics so as to introduce situation-based description information in current TCT.

Our current strategy is to select a suitable semantic representation similar with the predicate-argument structure used in PropBank project, and focus on the research of syntax and semantics linking to bridge the largest gap between current syntactic annotations in TCT and the new semantic representation. Here, the function of syntax and semantics linking is twofold. Through syntax to semantics linking, we can assign suitable semantic explanation for each syntactic template in current treebank. The combination of syntactic and semantic representations will give us enough information for deep conceptual understanding. Through semantic to syntax linking, we can assign useful syntactic distributions for different sense descriptions. They are very important for natural language generation.

Now, we have developed an efficient Chinese 'syntax→semantic' linking algorithm [5], whose accuracy is about 83%. Based on it, we can extract and build a large scale Chinese verb knowledge base from current TCT, where each verb entry is related with its syntactic templates, semantic role frames and detailed annotated examples. It will bring strong supports to refine current situation network.

5 Conclusions

We take situation as a suitable framework for organizing and positioning lexical semantic knowledge. This paper introduced our current research work to build a situation-based Chinese language knowledge base, whose basic language resources are a large scale Chinese treebank and three Chinese semantic lexicons. The research of syntax and semantics linking algorithm build the bridge between these two basic language resources through the assignation of suitable semantic explanations for syntactic constructions in current treebank. And the development of a language resource supporting platform makes full use of the abundant syntactic and semantic knowledge to build a situation-based Chinese computational infrastructure.

In the future research work, we hope to refine the language knowledge base in the following respects: (1) Improve the syntax and semantics linking algorithm to obtain more accuracy linking results; (2) Develop a semi-automatic tool to summarize rough situations based on current syntactic and semantic knowledge.

This work was supported by the Chinese National Science Foundation (Grant No. 60173008), National 973 Foundation (Grant No. 1998030507) and National 863 plan (Grant No. 2001AA114040).

References

1. Baker, C.F., Fillmore, C.J., and Lowe, J.B. (1998). The Berkeley FrameNet project. In *Proceedings of the COLING-ACL'98, Montreal, Canada*, p86-90.
2. Barwise, J.; Perry, J. (1983) *Situations & Attitude*, MIT Press. Re-issued by CSLI Publications, 1999.
3. Brants, S., & Hansen, S. (2002). Developments in the TIGER annotation scheme and their realization in the corpus. In *Proc. of the Third Conference on Language Resources and Evaluation LREC-02*, Las Palmas, Spain. p.1643-1649.
4. Chen Zushun and Zhou Qiang (2002) Situation – A suitable framework to organize and position lexical semantic knowledge. *Computational Linguistics and Chinese Language Processing, 7(2)*, p1-36
5. Dang Zhengfa, Zhou Qiang (2004). Link Syntactic Tags with Semantic Roles. *Technical report 04-04, State Key Lab. of Intelligence Tech. and Systems, Tsinghua University.*
6. Hajic, J. (1999). Building a syntactically annotated corpus: The Prague Dependency Treebank. In E. Hajicova (Ed.), *Issues of valency and meaning. Studies in honour of Jarmila Panevova.* Prague, Czech Republic: Charles University Press.
7. Kingsbury, P.; Martha Palmer, and Marcus, M. (2002). Adding Semantic Annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference, San Diego, California*.
8. Levin, B. (1993). English Verb Classes and Alternations A Preliminary Investigation. MIT Press.
9. Marcus, M.P., Marcinkiewicz, M.A. and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313-330
10. Miller, George A., & Fellbaum, C. (1991). "Semantic Network of English", In *Beth Levin and Steven Pinker (Eds.) Lexical & Conceptual Semantics*. Elsevier Science Publishers, B. V., Amsterdam, The Netherlands.
11. Zhou Qiang (2003) Build a Large-Scale Syntactically Annotated Chinese Corpus. In *Proc. of 6th International Conference of Text, Speech and Dialogue (TSD2003)*, Czech Republic, Sept. 9 –12. Springer LNAI 2807. p106-113.