International Journal of Computer Processing of Oriental Languages

Building a Chinese Semantic Resource Based on Feature Structure

BO CHEN[a]

Department of Chinese Language, Hubei University of Art & Science, Xiangyang, Hubei, China
cb9928@gmail.com

DONGHONG JI

School of Computer, Wuhan University, Wuhan, Hubei, China
donghong_ji2000@yahoo.com.cn

CHEN LV

School of Computer, Wuhan University, Wuhan, Hubei, China
lvchen1989@ gmail.com

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Building large-scale semantic resource is one of the major tasks in Language Information Processing. We propose Feature Structure Theory, and apply this theory in building a large-scale Chinese semantic resource based on Penn Chinese Treebank corpus. The feature structure theory aims at addressing annotation problems from special sentence patterns, flexible word order, and serial noun phrase, etc., which are universal in Chinese. Annotation based on feature structure theory describes more semantic information than traditional approaches, and achieves higher annotating efficiency and higher accuracy.

Keywords: Semantic parsing; Feature Structure Theory; Chinese semantic resource.

1. Introduction

a Address: Department of Chinese Language, Hubei University of Art & Science, Xiangyang, Hubei, China. Postcode: 441053.

Currently, Semantic parsing and semantic understanding are two of the popular research approaches in Natural Language Processing(NLP). Semantic parsing plays an important role in the fields of Question Answering, Information Retrieval, Information Extract, and Machine Translate. Building a large-scale language resource with abundant semantic annotation information will improve the accuracy and efficiency of semantic parsing.

"Feature" is a term widely used in modern fields of linguistics and NLP. Initially, the term "feature" is used to distinguish between syllables in Generative Phonetics, and to distinguish and describe syntactic structure in formal syntactic theories. According to earlier work we reviewed, using feature structure theory in large-scale semantic descriptions and semantic parsing has not been attempted.

Syntactic structure and dependency structure are traditional annotation methods. Both of them achieve high performance in Indo-European language, such as English, French, Dutch, Spanish, Portuguese, etc. However, many problems are encountered in parsing Chinese sentences using these two methods, especially Chinese subject-predicate predicate sentence, Chinese serial verb sentence, Chinese pivotal sentences, and the serial noun phrase, etc. Examples are listed as follow:

(1) 肚子 笑 痛 了
dù zi xiào tòng le
Belly smile ache
To laugh so hard that his belly ached.

In (1), according to dependency structure, "肚子"(belly) depends on "笑" (smile), but there is no real conception relations between them; while "肚子" and "痛" (ache) have not any syntax relation between them, but have actual conception relation between them.

In summary, dependency structure is not satisfactory in Chinese to represent or derive the conception relations and the relation classes. It is necessary to research the characteristics of conception relations, find out its formal description method, and construct the annotation resources of Chinese phrases and sentences.

2. Feature structure theory

We have applied the Feature Structure Theory(FS) to label the semantic relations of Chinese sentences and build a large-scale annotation corpus with 30,000 Chinese sentences.

Generally, a phrase or sentence may be expressed as a collection of feature structure, and a feature structure is represented as a triple:

[entity, feature ,value]

(2)红 颜色 汽车
hóng yán sè qì chē

(3) 红 汽车
hóng qì chē

red color car red car

red car red car

In (2), "汽车"(car) is an entity, "颜色"(color) is the feature of "汽车"(car), "红"(car) is the value of the feature "颜色"(color). "颜色" (color)links to both "汽车"(car) and "红"(red). So "颜色" (color)can be viewed as a tag of the link between "汽车"(car) and "红"(red) .In (3), "汽车" (car)is the entity, "红" (red)is the value. We can express them as triples:

(2') [汽车, 颜色, 红] (3') [汽车, , 红]

Fig 1: the graph of [汽车, 颜色, 红] Fig2: the graph of [汽车, , 红]

In (2), all of the entity, feature and value appears in the phrase. But in (3), Only the entity and value appears while the feature "颜色" feature does not appear.

(4) 他 说 他 是 大学 教师。
tā shuō tā shì dà xué jiāo shī
He say he is university professor
He says he is university professor.

The triples are listed below: [说, , 他]; [说, , 他是大学教师]; [是, ,教师]; [教师, , 大学]; [是, , 他]. All feature structure of (4) can be represented as a graph, showed in Figure 3. As shown in this figure, feature and value can be used as entities, for example, "他"(he) is the value of "说"(say), and "他是大学教师"(he is university professor) is the other value of "说"(say). Here "他是大学教师" is viewed as an entity which has semantic relations with "说"(say). The value "他是大学教师" is also a complex feature structure which contains more 3 feature structure triples: "是"(to be) is the entity, "教师"(professor) is its value, "他" is "是"'s another value. Furthermore, the node of the value "大学教师"(university professor) is also a FS: "教师"(professor) is the entity, "大学"(university) is its value.

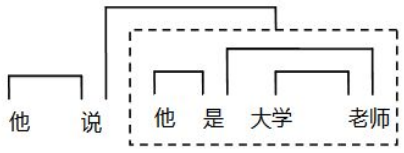


Fig 3: the graph of the sentence "他说他是大学教师。"

Formally, a triple can be seen as two nodes and the edge linking them, a node is an entity or a value , the edge is a feature. The feature must be the feature of one of the nodes, and the node serves as the owner of the feature, while another

node serves as the value. Thus, a feature structure can be seen as an undirected graph. Because the value can also be another feature structure, the feature structure may be represented as a recursive graph, in which a node can also be a graph. The graph of feature structure can be shown in Figure 4:

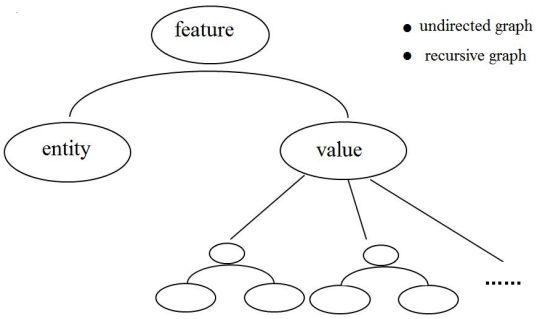


Fig 4: Feature structure graph

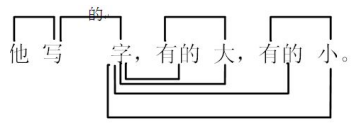
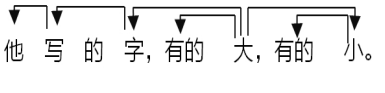
Analogous to dependency structure, feature structure mainly describes the relations among words, and do not need to define the syntactic category. However, in feature structure, include recursive feature structure, it is unnecessary to define the type of feature structure. In contrast to dependency structure, feature structure allows nesting and multiple correlations.

3. Application

In Chinese language processing, there are so many difficult problems, such as Chinese special sentence patterns, which include Chinese subject-predicate predicate sentence, Chinese Serial Verb Sentence, and the complex noun phrases etc. It is always difficult to label Chinese special sentence patterns. In Table 1, we used traditional dependency structure and feature structure to parse Chinese subject-predicate predicate sentence, such as (5).

(5) 他 写 的 字，有的 大，有的 小。
Tā xiě de zì yǒu de dà yǒu de xiǎo
he write character some big some small
The characters he wrote, some were big, some were small.

Table 1: comparing the semantic parsing of (5) with Feature Structure & Dependency Structure

Feature Structure	Dependency Structure
	

In (5), because traditional dependency structure emphasized "one head" in a sentence, it cannot represent the semantic relations of the phrase "有的大(some were big)" and "有的小(some were small)", nor can it represent the semantic relations among "字(character)", "有的 1" (some) and "有的 2" (some). In addition, it missed the semantic relations between "写"and "字", however, with feature structure these semantic relations can be represented as to the coordinate relations of "有的 1" and "有的 2" and the hidden relations of Phantom Reference.

4. Conclusion and Future Work

We put forward a new mechanism "Feature Structure" for the semantic representation. Compared to dependency structure, feature structure can represent more semantic relations and allows multiple links. Formally, feature structure is recursive graph and undirected graph. FS proves to be the superior mechanism for Chinese semantic parsing. According to the results, labeling feature structures is much more expeditious and effective than labeling syntactic structures and dependency structures.

In the application, our research can be used directly to Relation Extraction, Event Extraction, Automatic Question & Answering as well as the syntactic parsing in Machine Translation. Currently, FS has been used in the project of public sentiment analysis and monitoring.

It is an attempt to label the semantic relations of Chinese sentences using FS. We achieved good results. We have finished the sentence-level construction of Chinese Semantic Resources so far. The next step is to finish the construction of Chinese complex noun phrase Resources.

Acknowledgments

The work is funded by the project of NNSF (90820005/F020601, 61070243/F020606), the 985 project of Wuhan University foundation (985yk004).

References

[1] B.Chen and Y.J. Ma, Labeling Corpus with dependency relations, in *Proc. ICCC 2007*, Singapore: Chinese and Oriental Language Information Processing Society, 2008, pp. 205-208.

[2] B.Chen and D.H.Ji, Chinese Semantic Parsing Based on Dependency Graph and Feature Structure, in *Proc. EMEIT 2011*, Haerbin, 2011, pp. 1731-1734.

[3] B.Chen and C.Lv, Semantic Labeling of Chinese Subject-Predicate Predicate Sentence Based on Feature Structure, in *Proc. ISEEIP2012*, Shenyang,2012.

[4] D.H.Ji, Semantic Annotation of Chinese Phrases Using Recursive-graph, in *Proc. 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, 2000, pp.101-108.

[5] D.Y. Xiong, S.L. Li, Q. Liu, S.X. Lin, and Y.L. Qian, Parsing the Penn Chinese Treebank with Semantic Knowledge, in *Proc. 2th Int.Joint Conf. on Natural Language Processing (IJCNLP-05)*, Republic of Korea, 2005, pp. 70-81.

[6] J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S Riedel and D Yuret, The CoNLL 2007 Shared Task on Dependency Parsing , In *Proc.the CoNLL Shared Task Session of EMNLP-CoNLL*, Prague, 2007, pp. 915-932.

[7] L.D.Li, Modern Chinese Sentence Patterns, Beijing: the Commercial Press, 1986, pp.302-307.

[8] M.Zhou and C.N.Huang, Approach to the Chinese Dependency Formalism For the Tagging of Corpus, Chinese Information Processing, 1994(3), pp.35-53.

[9] N.W.Xue, F. Xia, F.D. Chiou, and M. Palmer, The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus, *Natural Language Engineering*, 2005,pp. 207-238.

[10] Q. Zhou, Annotation Scheme for Chinese Treebank, Chinese Information Processing, 2004 (4), pp.1-8.

[11] S.Buchholz and E. Marsi, CoNLL-X Shared Task on Multilingual Dependency Parsing, in *Proc. CoNLL-X(10th)*,NY,2006, pp. 149-164.

[12] S.Manabu, Linear-time Dependency Analysis for Japanese, in *Proc. the 20th int. conf. on Computational Linguistics*, Geneva, 2004, pp.8-es.

[13] Z.W.Feng, Machine Translation, China Translation and Publishing Corporation, 1998, pp.412-434.