ICIC Express Letters
Part B: Applications
Volume **7**, Number **5**, May **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **1063–1069**

THE ANALYSIS OF SENTENCES CONTAINING WORDS WITH MULTIPLE HEADS BASED ON CHINESE SEMANTIC DEPENDENCY GRAPH

Yanqiu Shao and Lijuan Zheng

School of Information Science
Beijing Language and Culture University
No. 15, Xueyuan Road, Haidian District, Beijing 100083, P. R. China
yqshao163@163.com; lijuanzhengzai@126.com

Received October 2015; accepted January 2016

Abstract. *Semantic analysis is one of the key problems both in the fields of NLP and linguistics. However, Chinese is a kind of paratactic language. Many Chinese sentences could not be well expressed the semantics only by traditional dependency theory. In this paper, the semantic dependency tree (SDT) is extended to semantic dependency graph (SDG). SDG expression breaks two limitations: the first is that one word could only have single head node and the other is that the dependency arcs could not cross in SDT. By statistics, the ratio of the former situation in real corpus is about 21.96%, and the latter situation which is so called non-projective situation is about 17.4%. The paper mainly discussed the former situation, i.e., the types of sentences which contain words with multiple heads. The paper also introduced the corpus and annotation scheme of SDG. Besides, some statistics of sentences containing words with multiple heads in Chinese is also provided in the paper.*
Keywords: Semantic analysis, Semantic dependency graph, Dependency trees, Multiple heads, Semantic role

1. **Introduction.** Semantic analysis is one of the key problems both in the fields of NLP and linguistics. The study of semantic analysis has achieved some accomplishments. For English, many representative semantic resources have been built: the FrameNet [1] of University of California-Berkeley, the PropBank [2] of University of Pennsylvania and the NomBank [3] of New York University. For Chinese, there are also some representatives though the systems are not mature as the systems for English, such as Chinese PropBank (CPB) [4], Chinese FrameNet [5] and Chinese NomBank [6].

There are two different levels of semantic analysis. One is shallow semantic analysis and the other is deep semantic analysis. Up to now, semantic analysis on sentences is mainly focused on semantic role labeling (SRL) which is one of the concrete methods to realize the shallow semantic analysis [7]. However, compared with deep semantic analysis, SRL is not complete semantic analysis. SRL only finds the arguments related to the predicate in a sentence, and then labels the semantic role of every argument, without analyzing the internal semantic relations among different parts of the argument [4,8]. In addition, in SRL system, the same semantic role label such as Arg0, Arg1, Arg2, has different meaning for different words. Thus, the real semantic meaning could not be understood clearly only by the role marks without searching for verb framesets.

Deep semantic analysis, such as semantic dependency analysis, always tries to do the complete semantic parsing. It is such a method that analyzes the semantic role of every word in a sentence. For example, a long noun phrase which could be one argument in SRL system will not be more analyzed the roles of the internal parts of the phrase, but deep semantic analysis will give the roles of each word in the phrase. Generally, the meaning of the same role for different verbs is the same. For syntactic analysis, dependency grammar

has been proven to be useful in some applied fields [9]. However, semantic dependency is seldom studied until the shared tasks on the SemEval 2012 [10] and SemEval 2014 [11]. For Chinese, Li's work on semantic dependency relations is representative [12]. They constructed a Chinese semantic dependency tree (SDT) corpus which is based on dependency grammar. The corpus consists of 132,398 sentences; and the semantic roles set is based on HowNet [13].

However, SDT is based on Robinson's dependency grammar and it rules that the dependency structure must be single headed, connective, acyclic and projective which could guarantee that the dependency parsing result is a tree with single root. As a paratactic language, Chinese is different from hypotactic language such as English. Chinese organizes sentences based on logical connections. A lot of sentences with informal syntactic structures are allowed. In fact, for many sentences only based on the dependency tree, it cannot completely describe the semantic relations between the words in a sentence. For example, the Chinese sentence "他 (he) 有 (has) 个 (a) 妹妹 (sister) 很 (very) 能干 (competent)", when it is expressed by SDT, the word "妹妹 (sister)" could only have one single head. However, it is known that both of the words "有 (has)" and "能干 (competent)" should be the heads of the word "妹妹 (sister)". Therefore, sometimes, dependency trees could not express the sentence semantics clearly.

Taking into account the flaws listed ahead, we develop the semantic dependency tree to semantic dependency graph (SDG). SDG could comprehensively reflect the relations of the words in a sentence. Though Wang and Ji [14] and Sun et al. [15] have already done some work related to dependency graph for Chinese, there are some great differences among our work and their work. Sun's work is mainly related to syntactic analysis [15], while our work is mainly on semantics. Compared with Ji's dependency graph [14], we have different definitions of semantic relations and a larger number of semantic labels. In addition, Ji's graph is an undirected graph while our SDG is directed graph. Furthermore, the theories of Ji's and our graph are different; our SDG is on the basis of dependency grammar [16], parataxis net by Lu [17] and semantic relations set referenced to Lu [17], HowNet [13] and Yuan [18]; while Ji's graph is on the basis of feature structure.

The rest of this paper is organized as follows: Section 2 makes a brief introduction on our SDG corpus and the annotation scheme. Section 3 divides the sentences containing words with multiple heads into 4 classes based on linguistic analysis. Section 4 gives the statistics of these 4 different types of sentences in real corpus. Section 5 concludes this paper.

2. Chinese SDG Corpus and Annotation Scheme.

2.1. Comparison of SDT and SDG.

Semantic dependency graph is constructed based on extended dependency grammar theory [16]. Thus, SDG partially satisfies the condition of dependency grammar. Both of the dependency structure of SDG and SDT should be: 1) Only one word in a sentence is independent, which is usually the predicate of the sentence. In most cases, the predicate is a verb or adjective; 2) Except for the predicate, each word depends directly on another word by dependency arc. There are two words connected by directed dependency arc. One is the modifier or dependent word which the arc arrow points to, and the other one is the head, father or governor word that the arc comes out from.

On the other hand, SDG is different from SDT in allowing more than one head on certain word and crossing of arcs, which means that in SDG, except for the root word, any other word could have more than two arc arrows pointing to it and the existence of crossing arcs is also allowed. In fact, by statistics of our corpus, the ratio of the former situation in real corpus is about 21.96% and the ratio of the latter situation is about 17.4%. SDG system aims to find all the word pairs with real semantic relations and link

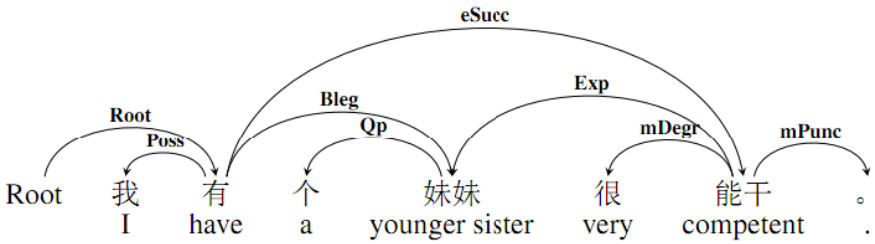


FIGURE 1. A sample sentence annotated with multiple heads

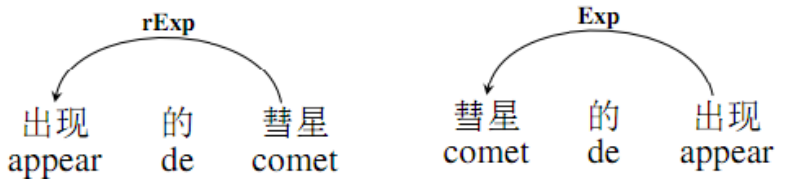
up each word pair with a dependency arc with a semantic label on it. Figure 1 shows an example of analysis result by using SDG. As shown in the figure, the word "妹妹 (sister)" has semantic relations with both "有 (has)" and "能干 (competent)", which means that "妹妹 (sister)" has two heads, which does not consist with dependency grammar. The semantic role of the word pair (有 (have), 妹妹 (sister)) is belongings (Belg) and the semantic role of the word pair (能干 (competent), 妹妹 (sister)) is experiencer (Exp). It can be seen that there are two arrows pointing to the word "妹妹 (sister)".

2.2. **Corpus.** Our corpus contains more than 30,000 sentences. The sentences are from newspapers, the textbooks of primary and junior school, Sina microblog and sentences for machine translation. We have already finished the annotation of newspapers (10,068), textbooks (10,038), Sina microblog (5,000) and sentences for machine translation (4900). All of the sentences are annotated by 4 master students who all major in linguistics. To evaluate the agreement of their annotation, we employed three of them to annotate the same small corpus blindly. The small corpus includes 422 randomly selected sentences from 30,000 sentences. We evaluate agreements on the level of dependency arcs and both arcs and relations respectively. The average agreements among three pairs of annotators are 88.78% (arcs only) and 72.15% (arcs and relations). The semantic labeling is more difficult than many other corpus annotations. The agreement ratio could be accepted.

2.3. **Annotation scheme.** We combined the system of parataxis network [17] with some concept of HowNet [13], and then defined a set of semantic labels. By revising the set of semantic labels constantly based on the practice of annotation, finally, we have defined a set consisting of 127 semantic labels. The set can be divided into 5 parts. They are semantic roles (32), reverse relations (29), nested relations (30), event relations (19) and syntactic marks (17). Event relations refer to the syntactic relations between multiple events in compound and contracted sentences, such as supposition (eSupp), progression (eProg), adversative (eAdvt), while syntactic marks refer to the words with grammatical meaning and no lexical meaning, such as conjunction (mConj), modal (mMod), preposition (mPrep).

Here it is necessary to introduce the definition of two special situations: reverse relations and nested relations. A reverse relation is marked when the modifier in a noun phrase is a verb. For example, for the two noun phrases "出现 (appear) 的 (de) 彗星 (comet)" and "彗星 (comet) 的 (de) 出现 (appear)", apart from the opposite direction of arcs, the semantic relation between "出现 (appear)" and "彗星 (comet)" is the same. Both refer to an experience relation. The head word of Figure 2(a) is "彗星 (comet)", and the modifier is the verb "出现 (appear)". Figure 2(b) is quite opposite. To avoid the influences of syntactic structure on semantic analysis, the semantic relation between the head word "彗星 (comet)" and modifer "出现 (appear)" in Figure 2(a) is reverse experiencer (r-Exp).

When one event is degraded as a constituent of another event, a nested relation is marked on the dependency arc. For example, in the sentence "爷爷 (grandfather) 看见 (see) 小 (little) 孙女 (granddaughter) 在 (is) 玩 (play) 计算机 (computer)." The underlined part is degraded as the object of "看见 (see)". A tag "d-Cont" which means



(a) Sample of reverse relation (b) Sample of common relation

FIGURE 2. Comparison of reverse relation and common relation

degraded content role is labeled between the predicate "看见 (see)" and another verb "玩 (play)". "玩 (play)" is the predicate of the degraded event.

3. **Analysis of Sentences Containing Words with Multiple Heads.** Different from dependency tree, there is no existing algorithm in SDG analysis. For improving the performance of automatic semantic dependency parsing, it is necessary to analyze the situations in which some special characteristics appear and make rules for them. As listed ahead, the SDG has its own characteristics. Here, we analyze one kind of situation, which is the sentences containing words with multiple heads. In this section, based on the 10,381 golden sentences of the corpus, we subdivide the situation of the sentences with multiple heads into 4 types. These 4 types will be respectively explained as follows.

3.1. **The sentences with pronouns.** In a context, every pronoun has its antecedent. No matter how long the distance between them is, we can find the antecedent of every pronoun. Because the basic unit of our semantic analysis is sentence, we cannot find the antecedent of every pronoun. We can only annotate the antecedent in the same sentence. Figure 3 shows an example with multiple heads. The appearance of multiple heads is because there is a pronoun in the sentence. In this sentence, "工程师 (engineer)" has two fathers (heads). One father is "走来 (walk to)", with the role of agent (Agt). The other one is "他 (he)", with the role of equivalence (eEqu). In dependency trees, the arc between "他 (he)" and "工程师 (engineer)" is lost. Thus, the semantic relation between them is ignored.

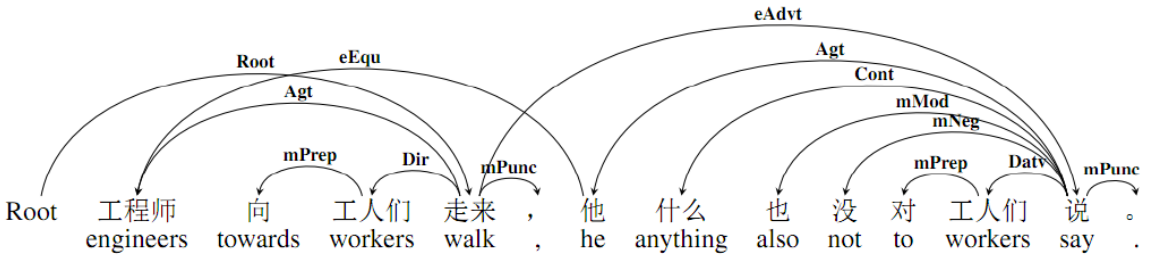


FIGURE 3. A sample sentence with pronouns

3.2. **The sentences with serial verb constructions.** In serial verb constructions, there are at least two verbs. These verbs usually have the same subject. Certainly, the conformation of serial verb constructions must have its own prerequisite. That is the two verbs must have logical relations; cause-effect, method-purpose, chronological order and so on. When we analyze the syntactic structure of the sentences with serial verb constructions, the omission of one dependency arc will have no effect on syntactic analysis. While the semantic analysis is different since the relations between the two word pairs may be different. As the example shown in Figure 4, the relationship between "想 (think)" and " 我 (I)" is affection (Aft), while the relationship between "通知 (inform)" and "我 (I)" is agent (Agt). Moreover, the discrimination between agent and affection in semantic

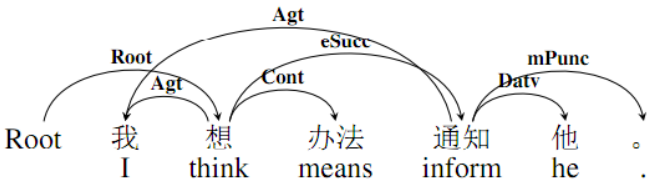


FIGURE 4. A sample sentence with serial verb constructions

analysis is very important. So we decided to add one arc even though it affects the harmony of the whole picture.

3.3. **The sentences with pivotal constructions.** Pivotal construction is such a construction that can be described by the formula "V_1+N+V_2". Usually, "N" is the object of "V_1" and the subject of "V_2". Compared with traditional dependency trees, SDG can display the relations between "N" and "V_2". Sometimes the omission of "N" and "V_2" may even cause the misunderstanding of the sentence. Figure 1 is an example. The word "妹妹 (sister)" has two heads, "有 (have)" and "能干 (competent)".

3.4. **The compound complex sentences with different subjects.** In the compound complex sentences with different subjects, one of subjects of the clauses is omitted, but the omitted subject is usually the attribute of the other subject. To describe more clearly, see the example in Figure 5. The two independent sentences in the compound complex sentence are "我 (I) 头 (head) 疼 (aches) 得 (de) 厉害 (serious)." and "我 (I) 还 (still) 流 (flow) 鼻涕 (snot)". Thus, the subject of each sentence is obvious. The subject of "疼 (ache)" is "头 (head)", while the subject of "流 (flow)" is "我 (I)". Actually, the two pairs of subject and predicate have direct semantic relations, with referring to experience. Chinese is a paratactic language; as long as the expression is logically consistent, all forms of sentences are available. So the two independent sentences compose one compound complex sentence. In the compound complex sentence, "我 (I)" is the subject of "流 (flow)", with referring to experiencer, and "我 (I)" is the modifier of " 头 (head)","头 (head)" is the subject of "疼 (ache)", with referring to experiencer. So "我 (I)" has relations with both "头 (head)" and "流 (flow)".

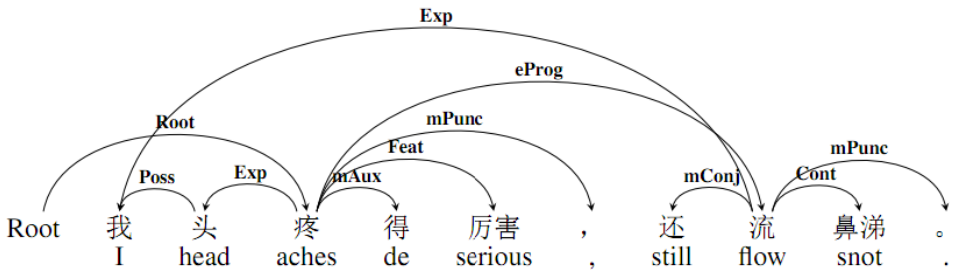


FIGURE 5. A sample compound complex sentence with different subjects

4. **Statistics of Different Types of Sentences.** In real natural language, the probability of the appearance of the above 4 situations is different. Some situations occur at a very high frequency, while the others seldom appear. In addition, in some sentences, the appearance of multiple heads is because one of the 4 situations exists in the sentence, it is also possible that 2 or 3 situations occur at the same time. Based on the 10,381 golden annotated sentences, we did some statistical work. Table 1 and Table 2 demonstrate the statistical results.

We can see that sentences with multiple heads in Chinese account for 21.96 percent of all the sentences, which means it is impossible to reveal the whole information of a

TABLE 1. The ratios of the sentences containing words with multiple heads in Chinese

Type	The number of occurrence	Percentage (%)
Sentences with multiple heads	2280	21.96
Sentences with 1 situation	1792	78.60
Sentences with at least 2 situations	488	21.40

TABLE 2. The number and ratio of each situation

Type	The number of occurrence	Percentage (%)
The sentences with pronouns	203	11.33
The sentences with serial verb constructions	1092	60.94
The sentences with pivotal constructions	459	25.61
The compound complex sentences with different subjects	38	2.12

sentence if we only use dependency trees which rules that one node could only have one head. In other words, there will be the disappearance of information because we analyze the semantics of sentences by trees, other than graphs. Furthermore, in most cases, if it is a sentence with multiple heads, the percentage that the multiple heads is caused by the existence of 1 situation in a sentence is 78.60%. The co-occurrence of several situations is not so common as the appearance of only 1 situation.

We calculate the number of the occurrence of each situation and its ratio in detail in Table 2. Besides, that the number of occurrence of each situation is divided by the number of occurrence of sentences with 1 situation (1792) is the percentage.

From Table 2, we can easily make a conclusion that the omission of the same subject is the main reason of the appearance of sentences with multiple heads. This is because Chinese is a kind of paratactic language, although some of the components of the syntactic structure are omitted, we can still understand the meaning as long as new information exists.

Through the analysis and the statistics in the paper, it will be helpful to establish the automatic semantic dependency graph parsing system by using these different types of sentences containing words with multiple heads as discriminate features.

5. **Conclusions.** In this paper, the construction of semantic dependency graph combined the advantages of dependency grammar with the characteristics of Chinese. SDG extends the SDT expression. Apart from their similarities, SDG breaks some of the limitations of dependency grammar and it has its own characteristics. One is the permission of one word with multiple heads, and the other is the appearance of crossing of arcs. In fact, the ratio of multiple heads is about 21.96%, and the ratio of crossing arcs is about 17.4%. All of these sentences could not be well expressed semantics only by SDT. In this paper, we discussed those sentences that contain words with multiple heads, The conditions are divided into 4 situations: 1) the sentences with pronouns; 2) the sentences with serial verb constructions; 3) the sentences with pivotal constructions; 4) the compound complex sentences with different subjects. The paper also introduced the corpus and annotation scheme of SDG. Besides, we proved the necessity of expanding dependency trees to dependency graph by the statistics.

The type classification will help us to better establish the automatic SDG parsing system. In the future, we will annotate more sentences and build the automatic SDG analysis platform.

Acknowledgement. We appreciatively acknowledge the support of the National Natural Science Foundation of China (NSFC) via Grant 61170144, Major Program of China's National Linguistics Work Committee during the twelfth five-year plan (ZDI125-41), important special fund of Beijing Language and Culture University (13ZDY03) and young and middle aged academic cadre support plan of Beijing Language and Culture University.

REFERENCES

[1] C. J. Fillmore and C. F. Baker, FrameNet: Frame semantics meets the corpus, *The 74th Annual Meeting of the Linguistics Society of America*, 2000.

[2] M. Palmer, D. Gildea and P. Kingsbury, The proposition bank: A corpus annotated with semantic roles, *Computational Linguistics*, vol.31, no.1, pp.71-105, 2005.

[3] C. Liu and H. T. Ng, Learning predictive structures for semantic role labeling of NomBank, *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, pp.208-215, 2007.

[4] N. Xue and M. Palmer, Annotating the propositions in the Penn Chinese Treebank, *Proc. of the 2nd SIGHAN Workshop on Chinese Language Processing*, pp.47-54, 2003.

[5] L. You and K. Liu, Building Chinese FrameNet Database, *Natural Language Processing and Knowledge Engineering*, pp.301-306, 2005.

[6] N. Xue, Annotating the predicate-argument structure of Chinese nominalizations, *Proc. of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

[7] M. Zhang, W. Che, G. Zhou, A. Aw, C. Tan, T. Liu and S. Li, Semantic role labeling using a grammar-driven convolution tree kemel, *IEEE Trans. Audio Speech and Language Processing*, vol.16, no.7, pp.1315-1329, 2008.

[8] W. Ding and B. Chang, Chinese semantic role labeling based on semantic chunking, *Journal of Chinese Information Procession*, vol.23, no.5, pp.53-61, 2009.

[9] L. Phuong, X. Phan and X. Nguyen, Using dependency analysis to improve question classification, *Knowledge and Systems Engineering*, vol.326, pp.653-665, 2015.

[10] W. Che, M. Zhang, Y. Shao and T. Liu, SemEval-2012 Task 5: Chinese semantic dependency parsing, *Proc. of the 1st Joint Conference on Lexical and Computational Semantics (SemEval 2012)*, Montreal, Canada, pp.385-393, 2012.

[11] S. Oepen, M. Kuhlmann, Y. Miyao, D. Zeman, D. Flickinger, J. Hajič, A. Ivanova and Y. Zhang, SemEval 2014 Task 8: Broad-coverage semantic dependency parsing, *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 2014.

[12] F. You, J. Li and Z. Wang, On construction of a Chinese corpus based on semantic dependency relations, *Journal of Chinese Information Procession*, vol.17, no.1, pp.46-53, 2003.

[13] Q. Dong and Z. Dong, *Hownet and Computation of Meaning*, World Scientific Publishing Company, 2006.

[14] Y. Wang and D. Ji, A study on the construction of Chinese Dependency Graph-bank, *Proc. of the 7th International Conference on Chinese Computing*, pp.251-256, 2007 (in Chinese).

[15] W. Sun, Y. Du and X. Kou, Grammatical relations in Chinese: GB-ground extraction and data-driven parsing, *Proc. of the Association for Computational Linguistics*, pp.446-456, 2014.

[16] J. Robinson, Dependency structures and transformation rules, *Language*, vol.46, no.2, pp.259-285, 1970.

[17] C. Lu, *The Parataxis Network of the Chinese Grammar*, The Commercial Printing House, 2001 (in Chinese).

[18] Y. Yuan, *The study of Chinese Computational Linguistics Based on Cognition*, Peking University Press, 2008 (in Chinese).