Construction of Semantic Collocation Bank Based on Semantic Dependency Parsing

Liu Shijun[1], Shao Yanqiu[1*]Zheng Lijuan[1], Ding Yu[2]
[1]Information Science School, Beijing Language and Culture University, Beijing, China
[2]Computer Science and Technology School,Harbin Institute of Technology, Harbin, China

Abstract

This paper extracts collocationbasing on semantic dependency parsing, and then constructs a collocation bankwith two levels according to frequency: the instance-level and semantic level. Compared with conventional extracting ways, the collocationsextracted in this paper have closer relationship and higher quality both on the lexical structure and semantic structure.

1 Introduction

Collocation has always been an important issue in language research, especially in Chinese language researches. Chinese is an isolated language, which lacks morphological changes.Establishing a relatively complete dictionary of Chinese collocation will be a great contribution to Chinese study and research.

Collocation plays a significant supporting role in many fields of NLP, such as information retrieval, machine translation, information extraction, and so on. Ding and Bai proposed a method of query expansion based on local co-occurrence[1]; Lin put relationship ofcollocation into language model for query expansion, which got over the deficiency of insufficient relationShips caused by lacking context in tradition query[2]. In the basic research field of NLP, such as syntax, semantics, etc., collocation also plays an important role.Based on the comparison of different patterns in adjective collocation between the Chinese English learners and native speakers, Zhang analyzed the typical characteristics of different learners when using adjective collocations[3]; Xingemphasized on the importance of collocation in the second language learning[4].

The early research of automatic collocation extraction was made by Choueka, Klein and Neuwtiz,they defined collocation as adjacent words, and used co-occurrence frequency to extract collocation[6];Church and Hanks improved the automatic extraction technology and put forward mutual information as the index ofcollocationevaluating[7].By proposing a formula for calculating strengthbetweencollocation,introducing dispersion formula,as well as integrating with the automatic speech tagging technology, the Xtract system of Smadja improved the extraction accuracy rate of collocation extraction up to 80%[8]; Lin extracted collocation based on shallow syntactic parsing[9];Shouxun YANG applied the method of decision tree to extract collocation by integrating frequency, likelihood ratio, point mutual information, variance and other statistical indicators[10].

In China, there werea number of outstanding dictionaries had been published,

[*]Correspondent Author

29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 232 - 240
Shanghai, China, October 30 - November 1, 2015

for example,*Modern Chinese Notional Words Collocation Dictionary* composed by Lin and *Modern Chinese Collocation Dictionary* by Mei.Recently, based on the foreign research results, Sunstarted a researchon automatic extraction of collocation and proposed three statistical indicators namely strength, dispersion and spike[11];Sun used the rule-based method to identify the verb object structure[12]; Quproposed the framework-based method to extract collocation[13];Cheappliedfrequency, distance, and variance, using improved t-test method to get the value of "collocation intensity coefficient", which was used to measure relationship of collocation[14].

This paper adopted method based on semantic dependency parsing (SDP) to extract collocation, combining with the semantic information from "*HowNet"* to make a semantic classification of collocation, which is a new perspective of collocation study.

2 Redefinition of Collocation on the Basis ofSemantics

From the perspective of linguistics, the research done by Benson has been most influential in the field of collocation. This is the definition of collocation given in his famous book *BBI Combinatory Dictionary of English*: collocation is the combination of words with arbitrary and frequent co-occurrence. At present, most of the research on collocation is based on the above definition, introducing different statistics to express different features of collocation.For example: "frequent co-occurrence" can be calculated by "word frequency", and "arbitrary" can be calculated by "mutual information". But extracting by these statistical methods will lose important language information. Choueka extracted the adjacent words as collocation, missing the non-adjacent words such as "make…decision"; Church adopted MI as the feature to extract collocations, then words which is closely related to each other but has no grammatical relations would interfere the results, such as "doctor-nurse". Linguistic symbolis the combination of sound and semantic,referring to the psychological reality, which reflects the objective reality. Therefore, any two of the words with semantic relationship can express certain objective reality. Here are two types of special collocation need to be explained.

One is the so-called unusual collocation. Such as "土豪(tyrant)+金(gold)"(Chinese word, referring to gold iPhone 5S), originallythis combination was not a collocation, but in recent years, with the popularity of Apple's mobile phone, this combination has become a common collocation. This is the evolution of language, which involves the principle of "established by usage". The principle stipulates if the language phenomenon is widely accepted by language users, we should keep it as a common usage. Therefore, this paper choose the frequency to represent the principle and defined "土豪+金", such a kind of frequent words, as collocation.

The other is free combination. Unlike constraint combination, the free combination is not combined in a relatively specific way; they can also be combined with other words, according to Benson. This kind of combination had been abandoned by Lin, when he was composing*Modern Chinese Collocation Dictionary*as they go against the principle of "less but better"[5]. For instance, the constraint combination such as "现代(modern)+词语(words)，古代(ancient)+词语(words)" had all been collected in thisdictionary, but the free combination such as "好(good)+词语，坏(bad)+词语"hadnot. This paper regarded them as a part of language, tried to find out their common semantic features, and generalized these combinations into the form of "word + semantic category".

The collocation this paper defined has the following characteristics: 1.there must be semantic dependency relations between words. 2. Reaching the threshold of frequency of co-occurrence. In this paper, there are two kinds of collocationforms: “word + word” and “word + Semantic Category”.

3 Corpus and Semantic Dictionary

Our research of collocation is based on SDP and semantic dictionary. Before the introduction of the bank, we should make a brief introduction to the SDP corpus we built and the semantic dictionary *“HowNet”* applied in this paper.

3.1 Corpus of Semantic Dependency Graph(SDG)

Chinese is a parataxis languagewith flexible word order and diverse function of word class. In real language context Chinese word often depends on several words simultaneously [12], which means in the same sentence one word can be semantically related to several other words. It also may exist in the non-projection phenomenon of crossed arcs. These phenomenon cannot be explained by the traditional dependency trees[15]. In order to express these phenomenon and also take the advantages of dependency expression, this paper break through the limitation of the dependency tree and express the semantics of the sentence by using the dependency graph, namely we connected two words to a dependency arc as long as there is a semantic relationship between them, which means the situation that a word with multiple parent nodes and crossed arcs will be reasonable. For example, “她(she) 眼睛(eyes) 哭(cry) 肿(swollen) 了(already)”, its dependency graph is shown in Fig. 1.

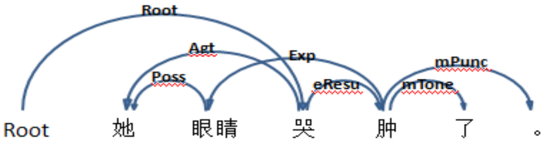

Fig. 1. An example of dependency graph

As shown in Fig. 1, the node “她(she)” has semantic relations with both “哭(cry)” and “眼睛(eyes)”, which means that there are two heads for “她(she)”: “哭(cry)” and “眼睛(eyes)” , separately indicates Agent of “哭(cry)” and Possessor of “眼睛(eyes)”Meanwhile, arcs (哭(cry)->她 (she), Agt) and (肿(swollen)->眼睛(eyes), Exp) cross. Such a multi father node and the crossed arcs express the true meaning of sentence.In addition, the meaning cannot be comprehensively expressed by dependency tree. “哭(cry)” is the core word of whole sentence, and the result of dependency tree parsing is as follows: （哭->她，Agt),（她->眼睛，Bleg),（哭->肿，eResu),（肿->了，mTone）,so as to lose the semantic relationship between "眼睛(eyes)" and "肿(swollen)". As in Fig. 2:

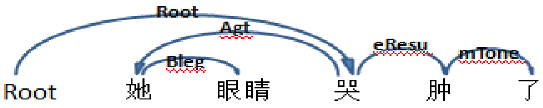

Fig. 2. An example of dependency tree

A set of semantic system has been constructed, and this paper will make a brief introduction.

On the basis of this system, we built a semantic dependency graph corpus, which contains 30,000 sentences.We have completed correcting 10,038 sentences. These data are from different areas, including the news corpus (10,068), Chinese textbooks (10,038), Sina Weibo corpus (5,000) and corpus for machine translation (4,900).This semantic dependency graph database aims at solving some Chinese phenomenon perfectly by introducing the non-projective phenomenon, as well as improving the automatic semantic dependency tagging.

3.2 Corpus for Collocation Extraction

In order to reflect the truth of language more accurately, we chose a 4G news corpus to extract collocation. After carryingout the word segmentation and POS tagging to the corpus, we conducted the automatic SDP to the sentences.The results of the analysis are represented in the form of CoNLL data format [16], which is shown in Table 1.

The meaning of the semantic tags in Table 1 are:Agt-agent ， Poss-possessor ， Exp-experiencer，Root-root（core word）, eResu-result，mTone-tone mark.

Table 1. Table form of dependency graph

Word index	word	Part of speech	Head index	Head word	Semantic Role
1	她(she)	PN	2	眼睛(eys)	Poss
1	她(she)	PN	3	哭(cry)	Agt
2	眼睛(eyes)	NN	4	肿(swollen)	Exp
3	哭(cry)	VV	Root	Root(root)	Root
4	肿(swollen)	VA	3	哭(cry)	eResu
5	了(already)	AS	4	肿(swollen)	mTone

3.3 HowNet

HowNet, built by Dong, is a knowledge base of common sense, aiming at revealing the relationship between the concepts and between the attributes of concept.

The semantic knowledge dictionary is the basic file system of HowNet, and the concept and description of the word in this dictionary form a record. Each record consists of several items.Each item has two parts separated by "=". The left part of the "=" representsthe domain name of the data, the right is the value. Set the word "eye" as an example, the data recorded in HowNetis shown in Table 2.Each line in Table 2 represents a record item, the meanings of itemareas follows: word index(No.), word(W_C), part of speech and pronunciation(G_C), English explanation(W_E), part of speech in English(G_E), concept(DEF). The first position of the concept of DEF is the main characteristic, specified in HowNet, which is the most distinguishing characteristic from other words and usually expressed by the sememe. We extracted this main characteristic as the foundation for classifying semantic category. As shown in Table 2, we defined the semantic category of “眼睛(eyes)” as “part（部件）”.

In this paper, we mainly investigated the collocation of “VV + NN”structure. With various relationships of entity in HowNet, we made a further classification to the nouns

(“NN”) in the instance collocation bank, generalizing the collocation into the form of “VV + semantic category”, so as to establish the semantic collocation bank.

4 Construction of Collocation Bank

4.1 Framework of System

Table 2.An representation example of HowNet

NO.=131783
W_C=眼睛
G_C=N [yan3 jing1]
W_E=eye
G_E=N
DEF={part|部件:PartPosition={eye|眼},whole={AnimalHuman|动物}}

This paper takes the “VV+NN” as an example to build a collocation bank.There are two levels of bank; one is the instance collocation bank with collocation of “word + word”, which was extracted based on SDP, the other is semantic collocation bankwith collocation of “word + semantic category”, which is the result from generalization of the instance bank based on HowNet. The system framework is shown in Figure 3.The upper section of the framework is the training part of the SDP model.We used this model to do SDP to the original corpus and got the semantic dependency graph. Then we extracted the ordered pairs with semantic dependency relations as the candidate set of collocation; on the basis of the candidate set, we extract the collocation of “VV + NN”, and introduce HowNet to construct the semantic collocation bank. Finally, the results of these two types of collocation database can be fed back to train the SDP model, to improve the effectiveness of the semantic dependency parsing system.

4.2 Construction of Instance Collocation Bank

4.2.1 Principles for Extracting the Semantic tuples

The structures of “NN + VV” and “VV + NN”are different in traditional syntactic analysis. The former is a subject-predicate structure, while the latter is a verb-object structure. In semantic analysis, the parent node of the noun “NN”is verb“VV”insubject predicate structure and predicate object structure, and the direction of the dependency arcs is from verb to nouns, the dependency pairs are all “VV + NN” structure. For instance, “政府(government)打击(beat)盗版(piracy)” and “盗版(piracy)被(been)政府(government)打击(beat)”, in these two sentences, the extracted dependency pairs (打击(beat),政府(government),Agt) 和 （打击(beat), 盗版(piracy),Pat）, are both in“VV + NN” structure. This paper mainly researches on the predicate-object structure, which can better reflect the collocation relationship between verbs and nounscompared with the subject-predicate.

Besides, Semantic level and syntactic level are not one to one correspondence, such as “吃(eat)食堂(canteen)” and “在(at)食堂(canteen)吃(eat)”, the former is the predicate object relationship,and the latter is the place adverbial.Butafter semantic dependency parsing, the arcs between them are both from verb “吃(eat)” to noun “食堂(canteen)”, shown as thetuple (吃(eat)，食堂(canteen)，Loc), “Loc” refers to “location”.

According to the above analysis, we got rid of the ordered pairs which contains semantic subject roles, such as “Agt(agent), Poss(possessor), Aft(affection), Exp(experience), to obtain the candidateinstances of “VV+NN” collocations in the predicate object structure.

4.2.2 Steps for Extracting Collocation Instance

Step 1: According to the result of POS tagging and SDP, draw out thesets of tuple (W_{VV}, Role, W_{NN}). “W_{VV}” and “W_{NN}” respectively represents the words of verb and noun, “Role” represents the semantic relationship between these two dependency words.

Step 2: Get rid of the tuple of which the “Role” remarked with subject role such as “Agt, Poss, Aft, Exp”, to get tuple in the predicate object structure as candidate collocation. In the meantime, as collocation is not only to meet the requirements of the semantics, but also to achieve a certain frequency, we set a threshold value for the frequency of co-occurrence, the pair with co-occurrence frequency> 50 can be selected into our bank. Table 3 lists part of the collocations ranked in the top 20.

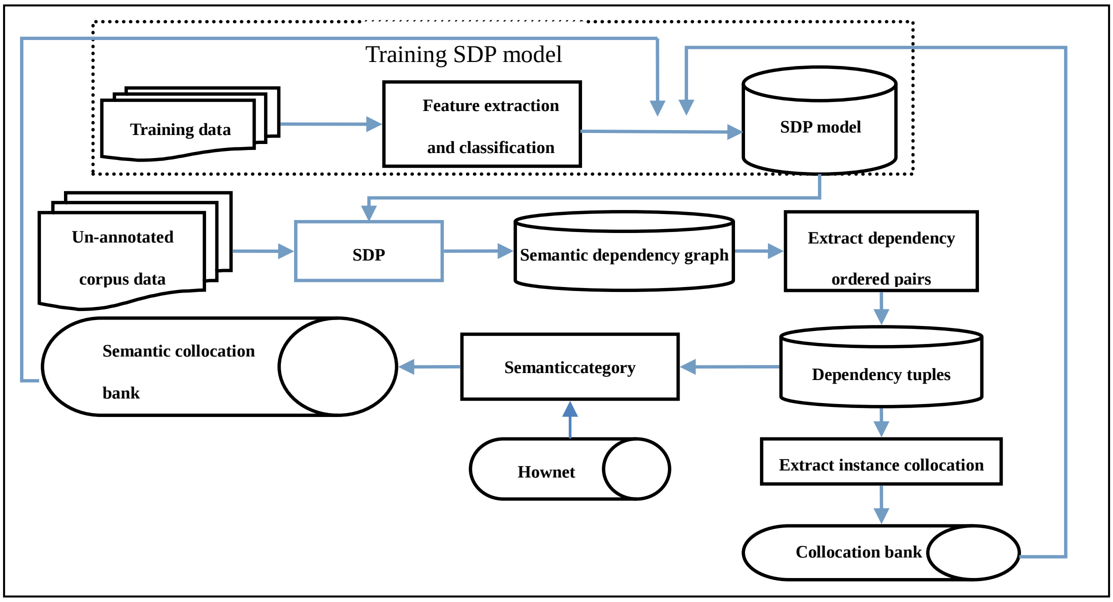
Training SDP model
Training data
Feature extraction and classification
SDP model
Un-annotated corpus data
SDP
Semantic dependency graph
Extract dependency ordered pairs
Semantic collocation bank
Semanticcategory
Dependency tuples
Hownet
Extract instance collocation
Collocation bank

Fig. 3.system framework of collocation extraction

Table 3: part of the collocation ranked in the top 20

Verb	Noun	Semantic role	Frequency of co-occurrency
采取(take)	措施(measure)	Cont	4669
解决(solve)	问题(problem)	Pat	4473
出席(attend)	会议(meeting)	Cont	3469
举行(hold)	会谈(interview)	Cont	2921
充满(be filled with)	信心(confidence)	Cont	2910
发挥(play)	作用(role)	Cont	2251
交换(exchange)	意见(opinion)	Cont	2075
拉开(pull)	帷幕(curtain)	Cont	1886
处于(be)	状态(status)	Loc	1662
赶到(arrive)	现场(spot)	Dir	1537

Step 3: the same combination can be labeled with different semantic role by automatic SDP, we adopted probability to distinguish same pairs with different semantic labels. As shown in Table 4, the frequency of “争执(dispute)”marked as “Cont” in “发生(happen) + 争执(dispute)” is 181, marked as “Prod” is 21.We calculated the probability on the basis of frequency, and formed thecollocation as the structure of "W_{VV}+ W_{NN}+ semantic role + probability".After the semantic role’s classification, the part of the "vocabulary + vocabulary" instance of collocation is shown in table 4.

In accordance with the above steps, this paper has a total of 67912 “VV+NN” structure as the collocation examples.

Table 4: instance bank of “word+word”

Semantic role	Verb	Noun	Probability
Prod	发生(happen)	争执(dispute)	0.10
Pat	限制(restrict)	自由(freedom)	0.26
	调查(investigate)	此案(case)	0.34
Datv	介绍(introduce)	总统(president)	1.00
	调查(investigate)	此案(case)	0.06
Cont	发生(happen)	争执(dispute)	0.90
	调集(assemble)	军队(army)	1.00
	举办(hold)	讲座(lecture)	1.00
	调查(investigate)	此案(case)	0.60
	限制	自由	0.74

4.3 Construction of Semantic Collocation Bank

To make up the deficiency of data sparse of instance bank, we generalized “VV + NN” collocation in instance bankto construct “Word + semantic category”bank. This kind of semantic bank represents semantic relations between verb and semantic category. For example, “吃(eat)+食物(edible)”, “食物(edible)” refers to semantic category, this semantic collocation labeled with “Pat”can cover all collocation consisting of all edible nouns and the verb“吃(eat)”. We need to applya semantic knowledge dictionary to determine whether a noun belongs to a certain semantic category. This paper introduced “HowNet”.The constructive algorithm of the semantic collocation bank is shown in Fig.3.

By algorithm 1, we generalized the semantic collocation from the candidate instance, and add the probability to each semantic collocation in accordance with the method used in the extraction of instance collocation bank.The result is shown in Table 5. We totally extracted 1446“word + semantic category” collocation.

5 Conclusions

On the basis of SDP, we extracted all the “VV + NN” ordered pairs from a large scale of news corpus, filtered collocation according to the set threshold and constructed the instance collocation bank. According to the semantic information provided by HowNet,we generalized the instance bank to conduct the semantic collocation bank. The relationship between the instance collocation bank and semantic collocation bank is complementary. The instance collocation bank cannot cover all the collocation in a language, so the semantic collocation is needed to be generalized, to enhance its robustness. For the extraction of collocation was based on SDP, the collocations are closely related to the syntactic structure and semantic structure. Therefore, the quality of the collocations is better than those extracted by the traditional way.

In the actual process, we find that not all nouns can be classified, for example, “吃(eat) 定心丸(assurance)”, this kind of collocation lacks similarity, main feature of “定心丸(assurance)” in HowNet is “Text”. It is obvious that the nouns in “Text” category almost cannot be matched with the verb “吃(eat)”. We put them into the instance collocation bank.

Two kinds of collocation bankhave been automatically established in this paper.The quality of the bank has not been evaluated. Besides, these two banks can be used to improve the accuracy of SDP,which will be our next work.

```
algorithm 1：Constructing semantic bank
    input 1：set of tuple with semantic relation，each of which represented as($W_{VV}$，$W_{NN}$，Role，Freq)，Role
        refers to the semantic labels of dependency pair ($W_{VV}$，$W_{NN}$)，Freq refers to frequency of dependency
        pair with semantic label “Role”.
    Input 2：Hownet（as shown in Table 2）.
    Output：semantic collocation bank. Every item is in represented by the form of ($W_{VV}$，SemC，Role, Prob)，
                SemC refers to semantic category of nouns，Role refers to semantic role.
    Process:
        For each tuple in tuple set, Do{
        If    $W_{NN}$in the tuple is in the Hownet：
            Add SemC represented by $W_{NN}$ to the Four-dimensional tuple（$W_{VV}$，$W_{NN}$，Role，Freq），extending
            the tuple into five-dimension（$W_{VV}$，$W_{NN}$，Role，Freq，SemC）
        If    Freq>= 4
            Replace the （$W_{VV}$，$W_{NN}$，Role，Freq，SemC） with （$W_{VV}$，SemC，Role, Prob）}
```

Fig. 3.algorithm 1

Table 5: semantic collocation bank

Verbs	SemC	Role	Prob	Verbs	SemC	Role	Prob
预防(prevent)	disease	Cont	1.000	去(go)	place	Dir	1.000
返回(return)	place	Dir	1.000	拘留(detain)	human	Pat	1.000
抵达(reach)	place	Cont	0.667	发展(develop)	Ability	Cont	1.000
带来(bring)	mishap	Cont	1.000	帮助(help)	human	Datv	1.000
看(look)	text	Cont	1.000	举办(hold)	place	Loc	1.000
	information	Cont	1.000	迎接(greet)	human	Datv	0.821
	shows	Cont	1.000	学习(study)	language	Cont	1.000
吃(eat)	vegetable	Pat	1.000	逮捕(arrest)	human	Pat	0.595
	edible	Pat	1.000	发出(emit)	sound	Cont	1.000
	fruit	Pat	1.000	公布(publish)	plans	Cont	1.000

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 (501321303).

References

[1] Ding Guodong, Baishuo, Wang Bin. Local Co-occurrence Based Query Expansion for Information Retrieval. Journal of Chinese Information, 2006, 20(3): 84-91.

[2] Lin Jianfang. Research on Collocation Extraction and Its Application in Information Retrieval. Harbin: Harbin Institute of Technology, 2010.

[3] Sun Haiyan. A Corpus-Based Study of Semantic Characteristic of Adjective Collocations in CLEC. Modern Foreign Language, 2004 (4).

[4] Xing Hongbin. Collocation Knowledge and Second Language Lexical Acquisition. Applied Linguistics, 2013(4).

[5] Lin Xingguang. Research on Collocation. Chinese Teaching & Studies, 1994(4).

[6]Choueka, YandKlein, TandNeuwitz, E. Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in A Large Corpus. Journal of Literary and Linguistic Computing, 1983,4.

[7]K.church, P.Hanks. Word Association Norms.Mutual information and Lexicography. Computational Linguistics, 1990, 16 (1) : 22-29.

[8]Smadja,F. Retrieving Collocation from Text : Xtract. Computational Linguistic, 1993, 19(1) : 143-177.

[9]Lin D. Extracting Collocations from Text Corpora[C]. In First Workshop on Computational Terminology, Montreal, Canada, 1998: 8-12.

[10]Yang S. Machine Learning for Collocation Identification. In 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 03). Beijing, 2003: 315-320.

[11] Sun Maosong. Preliminary study on quantitative analysis of Chinese Collocation. Studies of The Chinese Language, 1997(1).

[12] Sun Honglin. Generalizing Grammar Rules from Annotated Corpus: analysis of “V + N”. The Forth China National Conference on Computational Linguistics, 1997. Beijing: TsinghuaUniversity Press.

[13] Qu Weiguang, Chen Xiaohe, Ji Genlin. A Frame-based Approach to Chinese Collocation Automatic Extracting. Computer Engineering, 2004.12.

[14] CheWanxiang, Liu Ting, Qin Bing. A Method to Fetch Collocations Orienting Dependency Grammar. The sixth China National Conference on Computational Linguistics, 2001. Beijing: TsinghuaUniversity Press.

[15] Zheng Lijuan, Shao Yanqiu, Yang Erhong. Analysis of the Non-projective Phenomenon in Chinese Semantic Dependency Graph. Journal of Chinese Information Processing, 2014.11.

[16] Ding Yu, Shao Yanqiu, CheWanxiang, Liu Ting. Dependency Graph based Chinese Semantic Parsing. Harbin: Harbin Institute of Technology, 2014.