Deep Knowledge Representation based on Compositional Semantics for Chinese Geography

Shengwu Xiong, Xiaodong Wang, Pengfei Duan*, Zhe Yu and Abdelghani Dahou
Computer Science and Technology, Wuhan University of Technology, 122 Luoshi Road, Wuhan, Hubei, China

**corresponding author*

Keywords: DAG Deep Knowledge Representation, Combinatory Categorial Grammar (CCG), Semantic Analysis.

Abstract: Elementary education resources for geography contain a wealth of knowledge that is a collection of information with various relationships. It is of vital importance to further develop human like intelligent technology for extracting deep semantic information to effectively understand the questions. In this paper, we propose a novel directed acyclic graph (DAG) deep knowledge representation built upon the theorem of combinational semantics. Knowledge is decomposed into nodes and edges which are then inserted into the ontology knowledge base. Experimental results demonstrate the superiority of the proposed method on question answering, especially when the syntax of question is complex, and its representation is fuzzy.

1 INTRODUCTION

In recent years, human like intelligence has been more pervasive worldwide, related research has become the focus of all countries. Elementary education resources represented by geography contain a wealth of knowledge, which has various test items and types, and put forward a huge challenge to human like intelligent question answering system understanding of the problem.

There are several discriminative methods that have been applied for the knowledge representation. For example, an analysis method CHILL based on deterministic shift-reduce parser is proposed in Zelle et al. (1993), that uses logical expression method of knowledge representation. A new knowledge representation method, dependency-based compositional semantics (DCS) that is used in Percy Liang et al. (2013), in which tree describes the knowledge representation of problems. There are some researches based on the automatic learning rules and templates. Shizhu He et al. (2014) proposed a learning-based method using Markov Logic Network. Also, predicate logic knowledge representation which uses first-order predicate in Bao. (2014) performs great result in describing attributes of entities, but exposes disadvantages that it has low accuracy and efficiency with complex relationships, especially with more entities.

Traditional semantic description methods mainly use the logical expression for the representational model with good computing properties, but in practice the lack of a direct and effective means of analysis and inferences. The existing systems use a lot of surface layer of the semantic analysis method, due to the lack of deep knowledge representation and the deep semantic analysis.

After decades of exploration concerning computational linguists, the four widely regarded mature deep grammatical paradigms are Combinatory Categorial Grammar (CCG), Lexical Functional Grammar (LFG), Head-driven Phrase-Structure Grammar (HPSG), and Lexicalized Tree Adjoining Grammar (LTAG). We think that the CCG proposed by Mark Steedman (2011) formally from University of Edinburgh, is an effective method to construct the semantic analysis of natural language. The advantage of CCG is that it could match a related combinatory semantic knowledge using logical expression, such as the λ expression, for each syntactic category of each entry. Therefore, results of parsing reflect the ones of semantic analysis. In other words, semantic knowledge would be stored on lexical items only, and also suitable to solve word sense disambiguation.

We aim to improve the performance of the DAG Deep Knowledge Representation (DAG) in complex fuzzy condition. In this paper, firstly, we analyse the features of the geographical college entrance examination questions. Then a pre-processing method of test questions based on template is proposed. And word2vec expands trigger words to

Xiong S., Wang X., Duan P., Yu Z. and Dahou A.
Deep Knowledge Representation based on Compositional Semantics for Chinese Geography.
DOI: 10.5220/0006108900170023
In *Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART 2017)*, pages 17-23
ISBN: 978-989-758-220-2

reflect templates. At last, we propose the DAG Deep Knowledge Representation according to the combinatorial semantics, and transform the exam text successfully into the DAG Deep Knowledge Representation combining with CCG and templates.

The remaining parts of this paper are organized as follows: a DAG Deep Knowledge Representation method is given in Section 3 which translates text into DAG deep knowledge representation. Then in Section 4, we demonstrate the effectiveness of our theory by applying it to ontology knowledge base, and evaluate the accuracy in the performance of different sentence patterns. Finally, we draw conclusions and mention areas for future work in Section 5.

2 RELATED WORK

In 2011, human like intelligent system — Watson developed by IBM in the quiz game 'Jeopardy!' beat the previous two human champion, caused a big stir in the field of artificial intelligence field. In 2013, under the premise of manual work in the text processing of some papers, Todai Robot developed by National Information Research Institute of Japan got a good grades among top 16% of examiner. Luke S. Zettle Moyer and Michael Collins proposed a semantic analyser based on CCG for the first time in 2005. This method uses syntactic and semantic information combined to form the basis of the analysis of the dictionary, analysing the natural language. But it has none perfect theoretical basis, so that it's hard to be applied in open domain. In particular, Tsinghua University research group cooperated with Microsoft Asia Research Institute for the trans-formation of CCG resources, and hold an international evaluation about the analysis of Chinese combinatory categorial grammar.

The combination semantic analysis based on Machine Translation thought has a strong operability. Relevant research results have won the best paper award (2007) and best paper nomination (2013) at Natural Language Processing's top conference ACL.

3 DAG DEEP KNOWLEDGE REPRESENTATION

We introduce the DAG Deep Knowledge Representation in this part. We pre-process the questions according to the templates and translate long sentences into simple phrases. In the meantime, trigger words of the templates are defined to reflect the questions into templates. Also, these trigger words are expanded by using word2vec toolkit. After comparing the general knowledge representation methods, we get the DAG from the pre-processing result combining with Combinatory Categorial Grammar. On this basis, templates are transformed into the structure of DAG, which could finally get the DAG Deep Knowledge Representation integrating with DAG of simple phrases.

3.1 Templates Pre-Processing

Due to the question usually consists of many long sentences which are complex, we apply a pre-processing method to this issue. Question templates shown in Table 1 are manually designed and oriented to problem-solving strategy. And the trigger word is defined as a word which can transform sentences into specified templates.

The trigger word is a mark of the problem domain, and it could reflect the specified templates. For example,

- Problem domain:

"我国主要入海河流年总输沙量变化可能是由于水土流失现象加剧" (The variation of annual sediment transportation volume of Chinese main rivers may be due to the intensification of soil and water loss)

- Question templates:

Causality (cause*, result*),

Tendency (entities, #aspects, #cause, changes),

Causality (Tendency (水土流失现象 , # , # , 加剧) , 我国主要入海河流年总输沙量变化)

Table 1: Question templates.

Location(entities, places)
Distribution(entities, places, feature)
Sort(aspects, sequence, List)
Tendency(entities, #aspects, #cause, change)
Influence(subject*, object, #result*)
Matching(entity 1, entity 2)
Measure(question, solution)
Comparation(entity 1, #entity 2, #aspect 1, #aspect 2, result)
Optimization(entities, #aspects, feature, #range)
Causality(cause*, result*)
Location(entities, places)

stands for not necessary , * stands for nesting

- Trigger words:

Causality —— “由于”

Tendency —— “加剧”

In this case, the trigger words “由于 (because)” and “加剧 (exacerbate)” reflect the problem templates “Causality” and “Tendency” respectively. Then there are two relatively simple sentences “水土流失现象 (soil and water loss)” and “我国主要入海河流年总输沙量变化 (the variation of annual sediment transportation volume of Chinese main rivers)”.

3.2 Trigger Word Expansion

It is very important to recognize the trigger words in the source sentences. As well different trigger word have the same meaning, so that both of them should be mapped to the same one template, Such as those trigger words —“位于”, “处于”, “坐落”, “位置” (locate) and so on. Hence, it’s necessary to cluster and expand trigger words to improve the accuracy of pre-processing.

Table 2: Hyper-parameters’ setting of Word2vec in training.

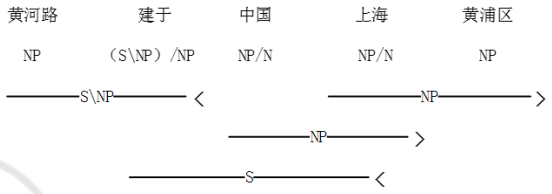
Parameter	Value	Meaning
-train	data.txt	corpus file need to train
-output	Vectors.bin	output file of word vector
-cbow	0	use ‘skip-gram’ framework
-size	200	vector dimension
-window	11	context window size
-negative	0	negative cases number
-hs	1	use hierarchy softmax
-sample	le^{-3}	sub-sampling threshold of high-frequency words
-threads	12	number of thread
-binary	1	binary file

The API of word embedding system ICTCLAS2014 developed by Chinese Academy of Sciences (CAS) is used towards questions of college entrance examination, which includes word embedding and POS tagging. Then the corpus is stored in a text file as the input of word2vec through filtering the irrelevant POS tagging and Splitting words.

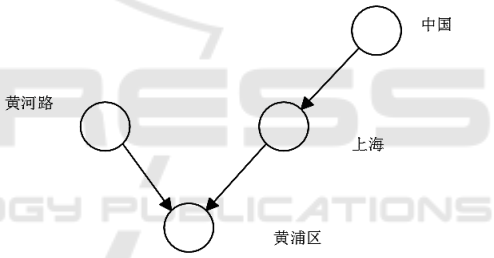
Therefore, when training word vector by word2vec, we can generate trigger words which have the same meaning and cluster them. As a result, those trigger words that hold the same meaning could be mapped to one template, this could increase the robustness of the system.

3.3 Deep Knowledge Representation

The traditional Knowledge Representation method based on first-order predicate logic could represent the knowledge accurately, at the same time has a common logical calculus method to ensure the completeness of the reasoning process. But in practice, this kind of representation requires high precision analysis method and a large number of accurate manual rules. And it is hard for us to expand to the open field. However, we found that there are natural graph structures in the ontology knowledge base, besides the DAG reflects the representation of knowledge and relevant rules with necessary constraints. Moreover, as a classical data structure, graph supports multiple operations, such as extraction, fusion, reasoning, etc. So, we treat DAG as the basic unit of deep knowledge representation.



(a) CCG-based knowledge representation.



(b) DAG deep knowledge representation.

Figure 1: Two kinds of knowledge representation. (a) CCG-based knowledge representation of sentence S using specific rules. (b) DAG deep knowledge representation, we delete some redundant nodes in CCG processing.

Figure 1 shows that CCG-based knowledge representation could be transformed into DAG, which could be inserted into the ontology knowledge base. Given this analysis above, we firstly apply CCG which bases on pruning algorithm and heuristic search to analyse natural language combinatorial semantics.

The technology roadmap of this paper as shown in Figure 2, we focus on the deep knowledge representation and Q&A for Chinese Geography, and aim to improve the performance of the DAG Deep Knowledge Representation in complex fuzzy condition.

Now we can search corresponding entities in the ontology knowledge base according to nodes information, then match predicate with edges

information to detect the semantically similar ones. Just like what shown in Figure 3, entities — “黄河路 (Huanghe Road)” and “黄浦区 (Huangpu District)” have predicate relation — “所在地” or “建于” (locate) in the ontology knowledge base. But, for entities — “中国 (China)” and “上海 (Shang Hai)”, there is no specific predicate relation existing in the base. So we appoint it as the default one. Just for now, we map the DAG (nodes and edges) into the base.

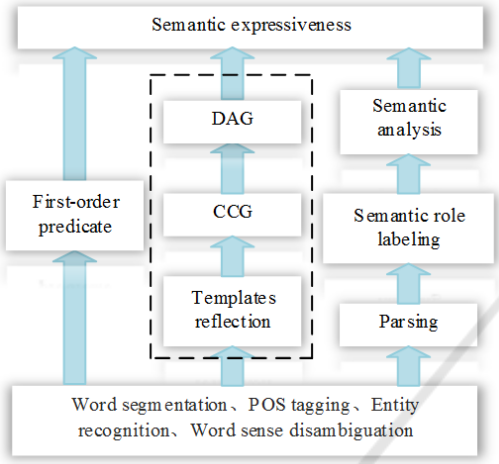


Figure 2: Technology Roadmap.

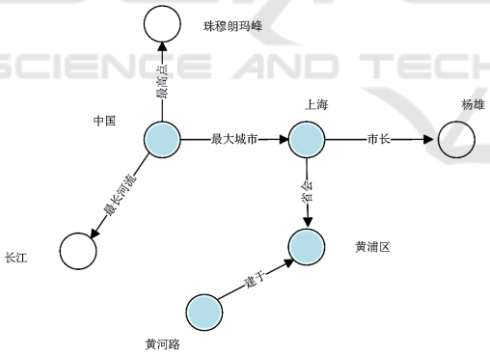


Figure 3: DAG in ontology knowledge base.

In this way, we make good use of the structured knowledge resources, and we could map the entities and relations to the base. The combination of semantics and knowledge base improve our ability to judge the correctness of results, furthermore, to carry on the more rigorous reasoning in the follow-up research.

3.4 Fuzzy Semantic Knowledge Extraction

However, when we extract some knowledge , the input is composed of incomplete or ambiguous statements. Thus after using DAG deep knowledge representation, the results are composed of multiple independent DAG subgraph. We cannot extract the corresponding knowledge from discrete structures, see Figure 4, but also cannot support the extraction, integration, reasoning and other follow-up operations towards knowledge.

Therefore, in this paper, we need to use the method of graph theory to calculate the relationship between multiple independent DAG sub graphs, and extract the approximate graph structure in the ontology knowledge base.

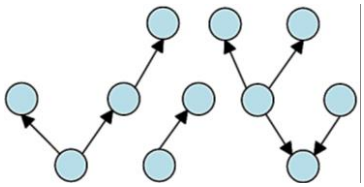


Figure 4: Incomplete questions’ DAG deep knowledge representation.

In the extraction of the approximation subgraph, like Figure 5, we will decompose subgraph into multiple DAG by using the recursive algorithm, until all the DAG decom-posed into local backbone graph; Specifically, for connected set of paths between two DAG subgraphs, we start to match from the shortest path recursively, looking for the backbone path connecting two DAG subgraphs, and extract subgraphs of the problem. Then the DAG set with matching degree is obtained according to the comparison between the extracted DAG and the problems. Select the DAG which has high matching degree as the final extracted knowledge, at the same time delete the sub graph in the cycle path, and generate DAG in Figure 6 that could be transformed into knowledge.

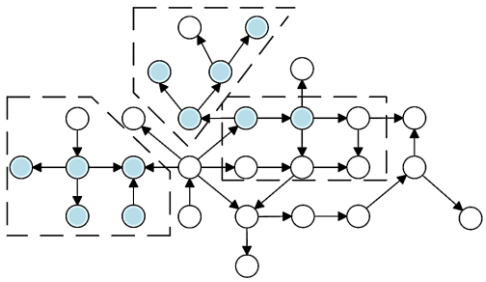


Figure 5: Sub-graph extraction of the most approximate input information.

Table 4: Recognition results of trigger words.

Serialization annotation	Recall	Precision	F-score
Location	0.972222	1	0.985915
Distribution	0.833333	0.862069	0.847458
Sort	0.722222	0.896552	0.8
Tendency	0.923077	0.972973	0.947368
Influence	0.689655	0.952381	0.8
Matching	0.593750	0.95	0.730769
Measure	0.972222	1	0.985915
Comparation	0.969697	1	0.984615
Optimization	0.833333	1	0.909091
Causality	0.656250	0.954545	0.777778
Average	0.816576	0.958852	0.882013

Table 5: Comparative analysis of knowledge representation.

knowledge representation methods	Accuracy of expression (%)		Accuracy of solving (%)	
	*	+	*	+
First-order predicate representation	93.0	74.5	72.5	50.5
DAG deep knowledge representation	91.5	80.5	79.0	72.5

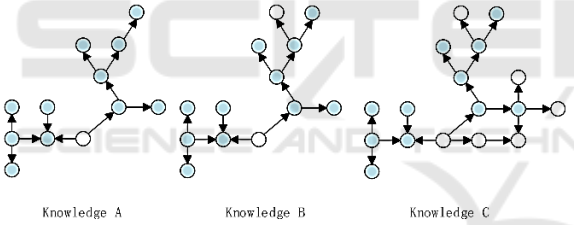


Figure 6: DAG local skeleton extraction.

4 EXPERIMENTAL EVALUATION

4.1 Data

In the experiment, we choose all of the corpus NLPCC (2015), which includes some extra related geography knowledge simulated by ourselves, as the training set and college entrance examination questions nearly ten years as the test set, the used test set is shown in Table 3.

Finally， there are nearly 600M data in the ontology knowledge base. Every record contains six columns (entity1 ID, predication ID, entity2 ID, entity1, predication, entity2).

Table 3: Experimental data.

Data set	Size
Training set	300,000 entities
Test set	200 groups

4.2 Results and Discussions

The experiment in this paper is based on word embedding, POS tagging, entity recognition and word sense disambiguation. And as shown in Table 4, we have achieved good results by mapping the text to the corresponding templates.

Given a set of question-answer pairs $\{Q_i, A_i^{rig}\}$ as the training set, we use the minimum error rate training to minimize the accumulated errors of the top-1 answer,

$$\widehat{\lambda}_i^M = \underset{\lambda_i^M}{\operatorname{argmin}} \sum_{i=1}^{N} \mathrm{Err}(A_i^{rig}, \widehat{A}_i; \lambda_i^M) \quad (1)$$

N is the number of questions in the training set, A_I^{rig} is the correct answers of the i^{th} question in the training set, $\widehat{A}_t$ is the top-1 answer of the i^{th} question in the training set.

Through experiment, we found that DAG deep knowledge representation could cover mostly geography questions in the last ten years. The recognition accuracy varies with the trigger words. What we summarized from the Table is that the DAG deep knowledge representation achieved a good performance with the accuracy that reaches 86%.

The main reason for the error is that the same word may be or not trigger word in different sentences. According to our statistics in Table 6, about 31% of the test questions don't include the trigger words, their structure is relatively simple. And applying the combinatory categorial grammar into DAG deep knowledge representation directly, the accuracy can reach more than 90%.

In Table 5, "*" is mainly composed of subject-predicate phrases, has relatively simple phrase structure and fewer words; "+" said long sentence structure which is relatively complex, has more modifiers (attributive and adverbial), parallel compositions.

Based on the experimental results in Table 5, we compare the proposed method with First-order predicate representation which is mainstream method in knowledge representation. There are also some representation methods based on tree-structure, but they are unsuitable for the experimental KB. The DAG deep knowledge representation shows its superiority especially when the question has complex sentences, it can effectively express complex

knowledge and extract the potential relationships between them.

Table 6: Knowledge Representation accuracy in different trigger words.

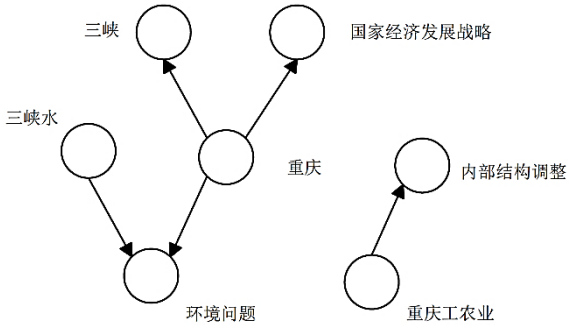
Serialization annotation	Accuracy
Location	0.96442
Distribution	0.84866
Sort	0.75123
Tendency	0.92282
Influence	0.82713
Matching	0.73463
Measure	0.94289
Comparation	0.92241
Optimization	0.90102
Causality	0.78452
Average	0.85141

For example, there is a question 14th in 2014 Tianjin college entrance examination. “山城重庆工业历史悠久，是大型综合性工业中心和西南地区综合交通枢纽。2014 年作为“长江经济带”的重要增长级，纳入国家经济发展战略。为了解决三峡水的环境问题，重庆工、农业内部结构应如何调整？(Mountain city Chongqing has a long history, and it’s a large-scale comprehensive industrial center and southwest integrated transport hub. In 2014 as the "Yangtze River Economic Zone" Chongqing was brought into the national economic development strategy. In order to solve the SanXia’s water environmental problems, how to adjust the internal structure of industry and agriculture in Chongqing?)” From the figure 7 (a), we can know that the results of the DAG deep knowledge representation is composed of two independent DAG, the two are not related in the structure. So they have no direct predicate relation after being inserted into the ontology knowledge base in figure 7 (b). Two independent DAG do not have connectivity. But we apply the algorithm finding the path between two DAG, and could get the answer to the question.

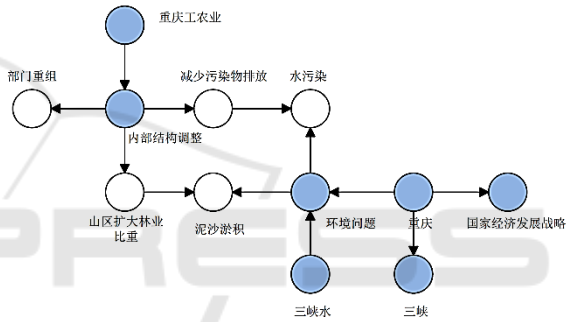
We also extract 200 groups of questions which are incomplete and has vague statements as the inputs and use the DAG knowledge representation carrying out the extraction experiment. Table 7 is the results. In 200 groups of questions, there are 188 groups can be extracted accurately using the subgraph algorithm mapped to the ontology knowledge base, while 162 groups can be extracted and find answers, about 20 groups miss because the error of trigger word recognition and syntax analysis caused by CCG.

Table 7. Accuracy in Fuzzy semantic knowledge extraction.

knowledge representation	Accuracy of expression (%)	Accuracy of solving (%)
DAG Deep Knowledge Representation	84.0	71.0



(a) DAG deep knowledge representation of the question.



(b) Insert DAG into the ontology knowledge base.

Figure 7: DAG knowledge representation and reasoning of the question.

5 CONCLUSIONS

In this paper, we applied the DAG deep knowledge representation to tackle the Chinese geographical knowledge entity relation extraction task with NLP technology. We have shown that under the condition of complex semantic, our deep knowledge representation showed a significant improvement in performance in practice, making this method more applicable to the fuzzy questions.

For future work, we will develop our method for constructing the corpus. We will expand CCG algorithm and refine DAG deep knowledge representation by carefully designing or automatic processing, aiming to capture more complex structures in the target domain. What’s more, we would try our best to link well with recent literature on NLP and sentiment analysis using convolutional

multiple kernel learning and deep convolutional neural networks.

ACKNOWLEDGEMENTS

This work was supported in part by the National High-tech R&D Program of China (863 Program) under Grant No.2015AA015403, Science & Technology Pillar Program of Hubei Province under Grant No.2014BAA146, Nature Science Foundation of Hubei Province under Grant No.2015CFA059, Science and Technology Open Cooperation Program of Henan Province under Grant No.152106000048, the Fundamental Research Funds for the Central Universities under Grant No.2016-zy-047 and CERNET Innovation Project under Grant No. NGII20150309.

REFERENCES

Zelle J M., Mooney R., 1993. Learning Semantic Grammars with Constructive Inductive Logic Programming. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*: 817-822.

Liang P., Jordan M. I., Klein D., 2013. Learning dependency-based compositional semantics. *Computational Linguistics*, 39(2): 389-446.

Shizhu H., Kang L., Yuanzhe Z., and et al, 2014 .Question answering over linked data using first-order logic. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ENMLP: 1092~1103.

Bao, J., Duan, N., Zhou, M., & Zhao, T., 2014. Knowledge-based question answering as machine translation. *Cell*, 2(6).

Steedman, M., & Baldridge, J., 2011. Combinatory categorial grammar. Non-transformational syntax, *ed. by Robert D. Borsley and Kersti Börjars*, 181–224.

word2vec[EB/OL].https://code.googlexom/p/word2vec.

Word Vectors[EB/OL].http://licstar.net/archives/328.

Mikolov T., Yih W., Zweig G., 2013. Linguistic regularities in continuous space word representations. *Proceedings of NAACL-HLT*: 746-751.

Mikolov T., Sutskever I., Chen K., 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*: 3111-3119.

NLPIR/ICTCLAS2014[EB/OL],http://ictclas.nlpir.org/(Accessec on Nov 20,2014).

Clark S., Curran J R., 2004. The importance of supertagging for wide-coverage CCG parsing. *Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics*: 282-288.

Klein D., Manning C D., 2003. A parsing: Fast exact Viterbi parse selection. *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HCT-NAACL03)*, Edmonton: 119-126.

Qiang Z., 2012. Evaluation Report of the third Chinese Parsing Evaluation: CIPS-SIGHAN-ParsEval-2012. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Somers H., 1999. Review article: Example-based machine translation. *Machine Translation* , 14(2): 113-157.

Wong Y W, Mooney R J., 2007. Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus. *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, June 23-30, Prague, Czech Republic.

Cai Q, Yates A., 2013. Large-scale Semantic Parsing via Schema Matching and Lexicon Extension. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers): 423-433.

Qiu X, Qian P, Yin L, et al., 2015. Overview of the NLPCC 2015 Shared Task: Chinese Word Segmentation and POS Tagging for Micro-blog Texts. *National CCF Conference on Natural Language Processing and Chinese Computing*. Springer International Publishing: 541-549.