International Journal of Knowledge and Language Processing
Volume 5, Number 2, **2014**

www.ijklp.org
KLP International ©2014 ISSN 2191-2734
pp.1–11

Studies on the Application of the Usages of Preposition "在[zai](*in*)" in Phrase Structure Syntactic Parsing

Mu Lingling, Feng Xiaobo, Wang Haoshi, Zan Hongying
School of Information Engineering
Zhengzhou University
Zhengzhou, Henan, China
iellmu@zzu.edu.cn

Selected Paper from Chinese Lexical Semantic Workshop 2014

ABSTRACT. *Chinese syntactic parsing is an important topic in the field of natural language processing. Compared to the development of English parser, the study of Chinese parser is still very weak. Aiming at the characteristics of Chinese, this paper presented a post-processing method that used the result of boundary identification of preposition phrases based on the usages of preposition "在[zai](in)" to improve the performance of PCFG parser. That is, in order to improve the precision and recall of Chinese parser, we designed a procedure to modify the syntactic trees that was produced by the PCFG parser with the result of the boundary identification of preposition phrases based on the features of prepositional usages. The experiments showed that the parsing result which used the features of usages was better than that did not use them.*

Keywords: syntactic parsing; prepositional usages; natural language processing; prepositional phrase boundary identification

1. Introduction. Syntactic parser automatically analyses the syntactic elements and their relationship of a sentence and maps the sentence to a structural syntactic parse tree according to the given grammar system[1]. Besides providing technical supporting for word sense disambiguation and semantic analysis[2], the result of syntactic parsing can be directly used in machine translation, question answering system and information extraction[3] to improve their performance.

Chinese words are divided into content words and function words. Content words are the words that have real meanings. The function words are the words that have no real meanings but have grammatical meanings or functional meanings. The composition of a Chinese sentence commonly relies on function words. The function words include prepositions, adverbs and conjunctions etc. Current syntactic parsers usually depend on the

tagging of part of speech (POS) without considering the usages of function words. That made the research of the Chinese syntactic parsing relatively poor. For example, the PCFG parser regarded the preposition phrase(PP) of the sentence "$_0$ 在 [zai](*in*)$_1$ 民 国 [minguo](*Minguo*)$_2$ 时 期 [shiqi](*period*)$_3$ 的 [de](*'s*)$_4$ 上 海 [shanghai](*Shanghai*)$_5$ 生 活 [shenghuo](*live*)$_6$。" (*Live in Shanghai in the period of Minguo.*) as "$_0$在[zai](*in*)$_1$民国 [minguo](*Minguo*)$_2$时期 [shiqi](*period*)$_3$" (*in the period of Minguo*). However, the correct preposition phrase of this sentence is "$_0$ 在[zai](*in*) $_1$ 民国[minguo](*Minguo*)$_2$ 时期 [shiqi](*period*)$_3$ 的[de](*'s*)$_4$ 上海[shanghai](*Shanghai*)$_5$" (*In Shanghai in the period of Minguo*). This kind of problems can be solved by using the usages of preposition "在 [zai](*in*)".

Preposition is an important word class in Chinese function words. According to the statistics of People's Daily corpus from February to June of 2000, the appearance frequency of prepositions is 227,495 times which is higher than that of the other function words. Meanwhile, the usages of the preposition are diversified and varied. The prepositions' syntactic meaning is hard to be understood and mastered. Because the prepositions' meanings, functions, usages and distributions are complex and special in Chinese syntactic system, the research of preposition always gains prominence in Chinese syntax research.

In this paper, the idea of syntactic parsing based on the usage of prepositions is to parse the sentence that includes prepositions according to the prepositions' usages and to generate its syntactic parsing tree. Because the appearance frequency of preposition "在[zai](*in*)" is about 28% in the prepositions according to the statics of People's Daily(2000.2-2000.6), which is the highest frequency in prepositions, the research of "在[zai](*in*)" plays an important role in the syntactic parsing. This paper will use the usages of preposition "在 [zai](*in*)" to improve the precision of syntactic parsing for the sentences that include the word "在[zai](*in*)".

The rest of the paper is organized as follows. In Section 2, we reviewed the related works of preposition "在[zai](*in*)" and the preposition phrase boundary identification based on the usages of preposition. In section 3, we gave the approach to modifying the parsing tree with the preposition phrase boundary identification based on the usages of preposition. In section 4, we gave the results of experiments on the method that was presented in this paper. At last, we drew a conclusion and listed further works.

2. **Related Work.** The foundation of our parsing work is the research on the recognition of the usages of prepositions and preposition phrase boundary identification based on the usages of prepositions.

2.1. **The Formal Description of the Usage of Preposition "在[zai](*in*)".** According to the statistics on the usage of prepositions in "Eight-Hundred Words of Contemporary Chinese"[4](Labeled as <b>)", "The Dictionary of contemporary Chinese"[5](Labeled as <h>), "The Dictionary of Function Words of contemporary Chinese"[6](Labeled as <x>)and "People's Daily"(Labeled as <r>), we compiled the human-oriented dictionary of the usages of prepositions[7]. In this dictionary, the number of proposition's usages differs

TABLE 1: ID, SENSE AND USAGE OF PREPOSITION“在[ZAI](*IN*)”[8]

Usage ID	Sense	Usage
p_zai4_1a	Express time.<x>	“~…”used before verb, adjective or subject.<b>
p_zai4_1b	Express time.<x>	“~…”used behind verb. The single syllable verbs are limited to “生(sheng born)|死(sidie)|定(ding determine)| 处 (chustay)| 改 (gaichange)| 放 (fang place)| 排 (paiarrange)” and so on, and the bisyllable verbs are limited to “出生(chushengborn)| 诞生(danshengborn)| 发生(fashenghappen)| 出现(chuxianappear)| 发现(faxianfind)| 布置(buzhiarrange)| 安排(anpaiarrange)| 确定(quedingdetermine)| 固定(gudingset)” etc..<b>
p_zai4_1c	Express time.<x>	Often followed by words “过程(guochengprocess)| 会议(huiyiconference)| 比赛(bisaigame)| 活动(huodongactivity)| 斗争(douzhengstruggle)| 接触(jiechucontact)" which can express duration time.<z>
p_zai4_2a	Express location. Refer to the place where action happens or things exist. <b>	“~…”used before verb, adjective or subject.<b>
p_zai4_2b	Express location. Refer to the birth, occurrence, produce, residence place or the place which the movement arrives. <b>	“~…”used behind verb.<b><z>
p_zai4_3a	Express scope.<b>	“~…”used before verb, adjective or subject.<b> Constituted prepositional phrases such as “~…方面(fangmianaspect)| 问题上(wentishangproblem)| 实践上(shjianshangpractice)| 生活中(shenghuozhonglife)| 生活上(shenghuoshanglife)|领域(lingyudomain)|工作上(gongzuoshangwork)” which express scope. <z>

p_zai4_3b	Express scope.<b>	“ ~… ” used behind verb. Constituted prepositional phrases such as “~…上(shangon)| 中 (zhongamong)| 以 内 (yineiwithin)| 以 外 (ciwaiexcept for)| 以 下 (yixiaunder)| 之 间 (zhijianamong)| 之 中 (zhizhongamong)| 之 外 (zhiwaiexcept for) ” which express scope. <x><z>
p_zai4_4	Express condition.<b>	Constituted prepositional phrase such as “~+gerund phrase+ 下 (xia under)”and used before verb or subject.<b>
p_zai4_5	Express subject of behavior. <b>	Often used before person noun or pronoun.<z>

from each other because the meaning and usage are different for each word. For preposition “在[zai](*in*)”, we compiled five meanings and nine usages. Table 1 showed its sense and usage.

The prepositions’ usage description for human is often ambiguous and subjective, and it is also not convenient to use them in automatic processing. In order to strictly and formally describe the prepositions’ usage, we formally described the prepositions’ usage rules for computer based on BNF rule[9]. The description of the usage rules of preposition “在[zai](*in*)” was showed in Figure1[8]. The capital letter means the context features, for example, F is head of sentence, M is left combination, L is left neighbor, R is right neighbor, N is right combination and E is end of sentence; the lower-case letter means lexicon; the Chinese words mean word forms.

```
$在
@<p_zai4_5>→N ^N→看来|来说|而言|说来|来看|来讲
@<p_zai4_3b>→L ^L→控制|限制|保持|维持|稳定|表现|体现
@<p_zai4_3a>→N ^N→方面|问题上|实践上|生活中|生活上|领域|工作上
@<p_zai4_4>→N ^N→(v|<vn>)<下/f>|(条件|前提|情况|情形|形势|背景|原则|努力)下|基础上
@<p_zai4_1c>→N ^N→过程中|活动中|活动上|会议中|会议上|会上|会中|赛中|赛上|斗争中|接触中|实践中
@<p_zai4_1a>→N ^N→(年|月|日|天|号|星期|世纪|期间|初|时|秒|之后|之前|之际|夜晚|同时|t)*v
@<p_zai4_1b>→LN ^L→v ^N→年|月|日|号|天|星期|世纪|期间|初|时|秒|之后|之前|之际|夜晚|t
@<p_zai4_2a>→N ^N→(<ns>|s)*v
@<p_zai4_2b>→LN ^L→v ^N→n|f
@<p_zai4_2a>→N ^N→(<ns>|s)
@<p_zai4_1a>→N ^N→(年|月|日|天|号|星期|世纪|期间|初|时|秒|之后|之前|之际|夜晚|t)
```

FIGURE 1. RULES OF “在[ZAI](IN)”

2.2. **Preposition phrase Boundary Identification Based on the Usages of Preposition “在[zai](*in*)”.** In order to meet the demand of natural language processing, Zhang[8-11] researched the automatic recognition of the usages of preposition and the preposition phrase boundary identification based on it.

In the research of automatic recognition of the usages of preposition "在[zai](*in*)", Zhang et al.[8-10] conducted the automatic recognition research of preposition "在[zai](*in*)" respectively based on rules, statistics and the combination of rules and statistics according to its formal description and the context exploration of different usages in real text corpus. When they used the People's Daily corpus(February, March, April of 2000) with word segmentation and POS tagging, the average precision of these three methods are separately 64.90%, 78.02% and 82.23%.. The best precision rate of automatic recognition of preposition's usages achieves 90.86%.

In the preposition phrase boundary identification research based on the usages of preposition "在[zai](*in*)", Zhang et al.[11] firstly recognized the semantic features of "在[zai](*in*)" and tagged it. Then they used these semantic and usage message to identify the boundary of the preposition phrase relevant to "在[zai](*in*)" with rule-based and statistic-based methods. The statistic-based method was studied on the model of ME, CRF and SVM. The average accuracy of these four methods are separately 57.39%, 96.85%, 97.26% and 90.58% for People's Daily corpus (January of 2000) with word segmentation and POS tagging. The average precision of the preposition phrase boundary identification based on the statistic method without the usages of "在[zai](*in*)" are respectively 95.61%, 96.12% and 89.17%. So, using the usages of "在[zai](*in*)" can help to get a better result in boundary identification.

3. **Syntactic Parsing Based on the Usage of Preposition "在[zai](*in*)".** When we looked through the parsing trees, we found that many errors were relevant to the boundary identification of preposition phrases. For example, the following sentence was parsed by PCFG parser [12] and the result is shown in Figure 2. The preposition phrase was tagged as PP node.

"$_0$ 在 [zai](*in*)$_1$ 民 国 [minguo](*Minguo*)$_2$ 时 期 [shiqi](*period*)$_3$ 的 [de](*'s*)$_4$ 上 海 [shanghai](*Shanghai*)$_5$ 生活[shenghuo](*live*)$_6$。" (*Live in Shanghai in the period of Minguo.*)

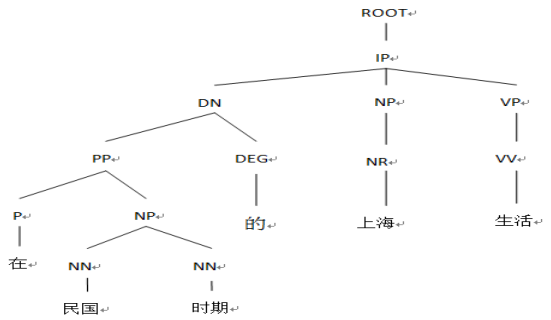

FIGURE 2. PARSING TREE GENERATED BY PCFG PARSER.

In this parsing tree, the boundary of preposition phrase preceded by "在[zai](*in*)" was PP(0,3):

“$_0$ 在[zai](*in*)$_1$ 民国[minguo](*Minguo*)$_2$ 时期[shiqi](*period*)$_3$” (*in the period of Minguo*). According to the context, the correct boundary should be PP(0,5): “$_0$ 在[zai](*in*)$_1$ 民国[minguo](*Minguo*)$_2$ 时 期 [shiqi](*period*)$_3$ 的 [de](‘s)$_4$ 上 海 [shanghai](*Shanghai*)$_5$ ” (*In Shanghai in the period of Minguo*). Figure 3 showed the correct parsing tree with the correct preposition phrase boundary.

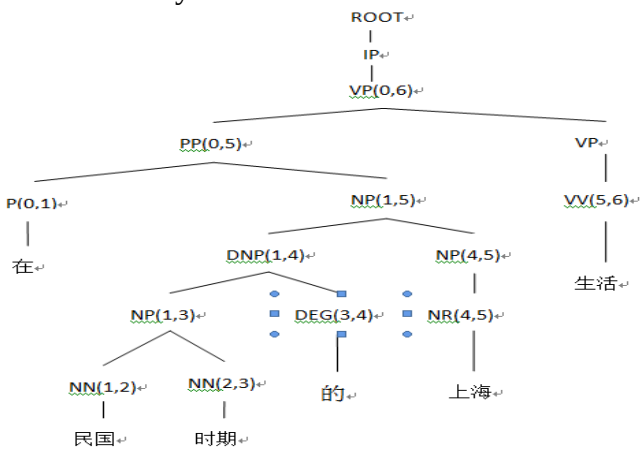

FIGURE 3. THE CORRECT PARSING TREE.

The reason for this error is identifying the preposition phrase boundary without considering the usages of preposition “在[zai](*in*)”. In PP(0,3), the object of preposition “在[zai](*in*)” is the time when the action happens and the usage of preposition “在[zai](*in*)” was labeled as p_zai4_1a; In PP(0,5), the object preposition “在[zai](*in*)” is the place where the action happens or the things exit, the usage was labeled as p_zai4_2a. If we firstly identify the boundary of preposition phrase based on the usage of “在[zai](*in*)”, and then use the correct boundary information to parse the sentence, we will get the correct parsing tree.

The parsing errors such as above occurred because of the boundary identification of preposition phrases. In this paper, we corrected this kind of parsing errors by post-processing the parsing trees according to the result of preposition phrase boundary identification based on the usage of preposition “在[zai](*in*)”.

3.1. **Post-processing for Parsing Trees.** The function of the prepositions in syntactic structure is mainly preceding some syntax elements that are usually combined with their previous prepositions to form preposition phrases to modify the predicate verbs. This paper mainly researched the PP nodes relevant to the preposition “在[zai](*in*)” in parsing trees and used the result of boundary identification based on the usage of preposition“在[zai](*in*)” to modify the PP nodes which included the preposition “在[zai](*in*)” in the parsing trees.

What we did in this paper is firstly parsing the sentences with PCFG parser, and then getting the boundary of preposition phrases based on the usages of preposition “在[zai](*in*)”.

Finally we modified the parsing trees with our post-processing procedure based on the boundary of preposition phrase. The operation flow in detail is showed in figure 4.

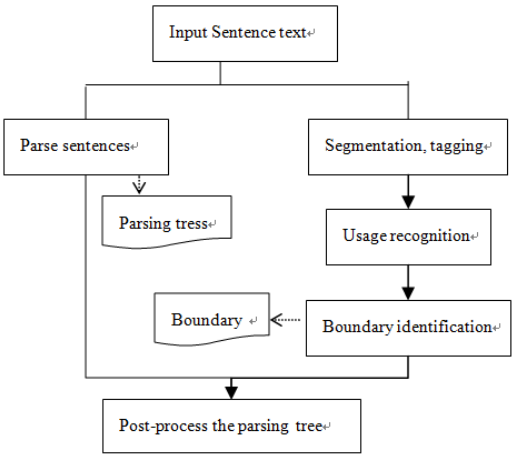

FIGURE 4. THE OPERATION STEPS FOR PARSING.

In figure 4, after parsing the sentence we will get a parse tree of this sentence. After identifying the boundary of the preposition phrase on the basis of the usages of "在[zai](*in*)", we will get the boundary of PP nodes in this sentence. In the post-processing flow we will compare the boundary of the preposition phrase in this sentence and its relevant PP node. According to the result of the comparison we will modify the parse tree according to the preposition phrase boundary.

The parsing tree of the sentence "$_0$ 在[zai](*in*) $_1$ 民国[minguo](*Minguo*)$_2$ 时期[shiqi](*period*)$_3$的[de](*'s*)$_4$上海 [shanghai](*Shanghai*)$_5$生活[shenghuo](*live*)$_6$。" (*Live in Shanghai in the period of Minguo.*) was shown in Figure 1. The PP node of this parse tree is PP(0,3). After this sentence was segmented, tagged and labeled the usage of preposition "在[zai](*in*)", the boundary of the relevant preposition phrase was identified as follows:

<PP>$_0$在$_1$_p<p_zai4_2a>民国$_2$_NN 时期$_3$_NN 的$_4$_DEG 上海$_5$_NN </PP>生活$_6$_VV

The label <PP> describes the front boundary and the label </PP> describes the hind boundary. The label <p_zai4_2a> describes the usage of "在[zai](*in*)" which means the place where the verb happens. So, the PP node of this sentence should be PP(0,5). Because the boundary of PP(0,3) and PP(0,5) are different, we will modify the parse tree with constraint of PP(0,5).

The post-processing procedure was implemented on the basis of PCFG parser as Figure 5 described.

Before conducting the algorithm based on PCFG, some pre-processing should be done. A self-defined product rule "PP'->_pp_" was added to the rule set and the probability value 1 was assigned to this rule. A self-defined terminal symbol "PP'" was added to the non-terminal symbol set. A self-defined terminal symbol "_pp_" was added to the terminal

symbol set. The words in the range of the preposition phrase's boundary in the sentence were replaced by terminal symbol "_pp_" and a relevant pseudo sentence was generated in this way.

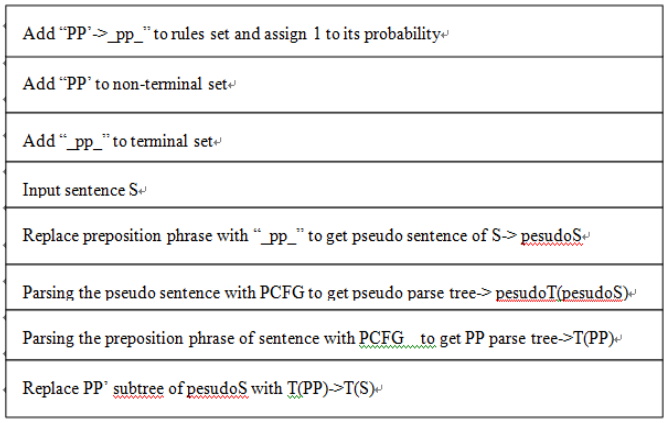

FIGURE 5. NS CHART OF POST-PROCESSING PROCEDURE.

Definition 3.1 A *Pesudo sentence is a string that is generated by replacing the preposition phrase in source sentence with terminal symbol _pp_ and is like a sentence. A pesudo sentence can be treated as a sentence.*

For example, with the identification of preposition phrase based on the usage of preposition "在[zai](*in*)", the preposition phrase was regarded as PP(0,5) for the sentence"$_0$在[zai](*in*)$_1$ 民国[minguo](*Minguo*)$_2$ 时期[shiqi](*period*)$_3$ 的[de]('*s*)$_4$ 上海[shanghai](*Shanghai*)$_5$生活[shenghuo](*live*)$_6$。" (*Live in Shanghai in the period of Minguo.*) The pseudo sentence of this sentence is "_pp_生活。".

After above operations, conduct the PCFG algorithm to parse the pseudo sentence and get a pseudo parsing tree such as the pseudo parsing tree in Figure 6.

Definition 3.2 A *Pesudo parsing tree is a tree structure which is mapped by a pseudo sentence and is like a parsing tree.*

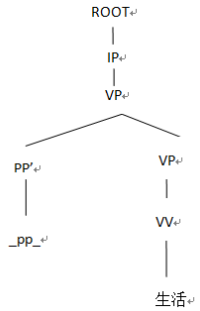

FIGURE 6. PESUDO PARSING TREE.

The preposition phrase in the sentence was separately parsed from PP nodes to get a PP parsing tree as is shown in Figure 7. Replace the PP' node in the pseudo parsing tree with

the PP parsing tree of the preposition phrase to get the real parse tree as is shown in Figure 2. We observed that the parsing tree which was got by PCFG with our post-processing method was the same as that we have predicted before.

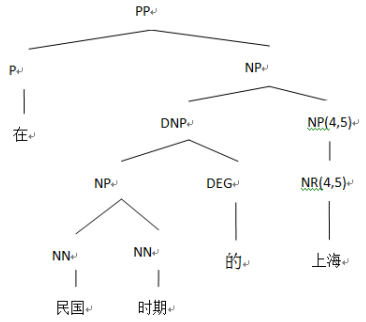

FIGURE 7. THE PARSING TREE OF PREPOSITION PHRASE.

3.2. **Experiments and Result Analysis.** In our experiments, we selected 126 Chinese sentences that include preposition "在[zai](*in*)" (less than 40 words) from Penn Treebank 5.1 as our test sentences. These sentences include 138 prepositions of "在[zai](*in*)".

In order to evaluate the performance of our approach, we used precision and recall as performance value. Precision is the fraction of correct number parsed and total number; Recall is the fraction of correct number parsed and the actual correct number. Because we only used the sentences that include "在[zai](*in*)" in our experiments, the precision is the same as recall. In the following text, we only discuss the precision.

The experiments were completed with the following steps.

Firstly, we used the PCFG parser of NLTK to parse the sentences and got the relevant parsing trees. We proofread these parsing trees especially the PP-subtrees which are relevant to preposition "在[zai](*in*)" and found that the precision is 48.4%. There were 65 incorrect parsing trees and there were 26 incorrect parsing trees (about 40 per cent of the incorrect parsing trees) whose incorrect reasons were relevant to PP nodes.

Then we used the preposition phrase boundary identification based on the usages of preposition "在[zai](*in*)" and the post-processing procedure to modify the parsing trees.

TABLE 2. PRECISIONS FOR TEST SENTENCES WITH DIFFERENT PARSING METHODS(%)

Methods	*PCFG*	*PCFG for pseudo sentences*	*PCFG for PP*	*PCFG with post-processing*
Precision	*48.4*	*72.6*	*96.6*	*70.8*

Table 2 is the experiment results. The number of test sentences that included preposition "在[zai](in) "was 126 in our experiments. The number of incorrect parse trees of PCFG was 61 and the number of incorrect parse trees that their incorrect reason was relevant to preposition "在[zai](*in*)" was 26.

The PP boundary was generated by preposition phrase boundary identification based on the usages of "在[zai](*in*)" with rule-based method. The PP boundary was used to modify

the parse trees that were generated by PCFG with the post-processing method that was presented in this paper. After post-processing, there were 11 parsing trees were modified (about 42.3% per cent) among 26 sentences whose former parsing trees were incorrect because of PP nodes.

The precision for our post-processing method that considers the preposition phrase boundary with the usages of preposition "在[zai](*in*)" was 70.8%. Compared with the former precision 48.4% that was gained by PCFG without using preposition phrase boundary, the precision was improved largely.

It can be viewed from Table 2 that the precision for pseudo sentences parsing is relatively high, and the precision of PCFG with post-processing is almost the same as that of pseudo parsing trees. So, we can draw a conclusion that the post-processing that based on the preposition phrase boundary identification with the usages of preposition "在[zai](*in*)" is effective.

4. **Conclusions.** We used the result of boundary identification which based on the usages of preposition "在[zai](*in*)" to post-process the syntax tree that got from PCFG. This method aimed at the shortcoming of preposition phrase boundary identification about preposition "在[zai](*in*)" in parsing trees and used the result of preposition phrase boundary identification which was based on the usages of preposition "在[zai](*in*)" in syntactic parsing to improve the precision and recall. The experiments showed that the parsing result with preposition phrase boundary identification based on the usages of preposition "在[zai](*in*)" was improved compared to the former without using it.

The post-processing method can also be used in other phrase boundary adjusting in syntactic parsing. Next we will try to automatically get phrase boundary of other function words on the basis of the usages of function words and use the post-processing that has been presented in this paper to modify the parsing trees to improve the performance of parser. Another method we will try is to improve the PCFG parser with usages of function words directly instead of the post-processing procedure.

Acknowledgment. This work was supported by the Natural Science Fund of China (No.60970083, No.61272221), the National Social Science Fund (No.14BYY096), 863 Projects of National High Technology Research and Development (No.2012AA011101), Science and Technology Key Project of Science and Technology Department of Henan Province(No.132102210407), Basic research project of Science and Technology Department of Henan Province(No.142300410231,No.142300410308) and Key Technology Project of the Education Department of Henan Province (No.12B520055, No.13B520381).

REFERENCES

[1] Liu ting and Ma Jinshan, The theory and method of Chinese automatic syntax parsing, *Contemporary Linguistics*,Vol.11, no.2, pp.100-112, 2009.[In Chinese]

[2] LIAN Le-xin, HU Ren-long, YANG Cu-il,i YUAN Chun-feng, Shallow semantic parsing based on Chinese Penn Treebank, *Application Research of Computer*, vol.25, no.3, pp.674-680, 2008.[In Chinese]

[3] LIANG Na, GENG Guo-hua, ZHOU Ming-quan, Mutual extraction between semantic relationships andlexical patterns in natural language processing, *Application Research of Computers*, vol.25, no.8, pp.2295-2298, 2008.[In Chinese]

[4] Zhang Bin, *Eight-Hundred Words of Contemporary Chinese*, The Commercial Press, Beijing, 2005.[In Chinese]

[5] Lv Shu-xiang, *The Dictionary of contemporary Chinese*, The Commercial Press, Beijing, 1980.[In Chinese]

[6] The editor office of Inst. of Linguistics of Chinese Academy of Science, *The Dictionary of Function Words of contemporary Chinese*, The Commercial Press, Beijing, 2007.[In Chinese]

[7] Zan, H.Y., Zhang K.L., Chai, Y. M., Yu, S. W. Studies on the Functional Word Knowledge Base of Modern Chinese. *Journal of Chinese Information Processing*. Vol.21, No.5, pp.107-111, 2007.[In Chinese]

[8] Zhang, K.L., Zan, H.Y., Han, Y.J., Zhang, T.F.Studies on Automatic Recognition of Contemporary Chinese Common Preposition Usage. *In: Proceedings of CLSW2012*,pp.219-229, Wuhan, China, 2012

[9] Hongying Zan, Kunli Zhang, Yumei Chai, and Shiwen Yu. Formal Descriptions on the Usages of Modern Chinese Adverbs, *In Proceedings of 8th Chinese Lexical Semantics Workshop (CLSW2007),* HongKong, China. 2007.

[10] Yuan Yingcheng, Zan Hongying, Zhang Kunli, Zhou Yihui, The automatic annotation algorithm design and system implementation Rule-base words' usages, *In Proceedings of 11st Chinese Lexical Semantics Workshop (CLSW2010),*SuZhou, China. 2010.

[11] Zhang Kun-li, Han Ying-jie, Zan Hong-ying, Yuan Ying-cheng, Prepositional Phrase Boundary Identification based on statistical. *Journal of Henan University(Natural Science).* Vol.41, no.6, pp.636-640.2011.[In Chinese]

[12] Steven Bird, Ewan Klein and Edward Loper, *Natural Language Processing with Python.* O'Reilly Media, Inc Press, USA.2009.