Incorporating Prepositional Phrase Classification Knowledge in Prepositional Phrase Identification

Qiaoli Zhou(✉), Ling Zhang, Na Ye, and Dongfeng Cai

Knowledge Engineering Research Center, Shenyang Aerospace University,
Shenyang 110136, China
zhou_qiao_li@hotmail.com, 710138892@QQ.com, yena_1@126.com,
caidf@vip.163.com

Abstract. This paper proposes a method of prepositional phrase (PP) identification by incorporating PP classification knowledge. When PPs act as different syntactic constituents, they have different characteristics in terms of location and context. In this paper, PPs are classified based on the context in which they appear. We select features based on the category of PPs to train multiple machine learning models for PP identification, and recombine these identification results. In this way, we can make full use of the complementary advantage of multiple models.

Keywords: Chinese information processing · Prepositional phrase identification · Multi-model · Prepositional phrase classification

1 Introduction

Preposition belongs to function word and is a closed set. There is a preposition list in *the Contemporary Chinese Grammatical Knowledge Base* [1] which has 85 prepositions. PP consists of two parts: the front part is the preposition, and the latter part is a word or phrase that is the attachment of preposition. In the process of PP identification, the first word of PP is preposition, so the first word can be identified based on the part of speech, and the identification of tail word of PP is our main task. In a sentence, PP always plays the role of attributive, adverbial and complement component. Therefore the PP identification result can help sentence framework (subject, predicate and object) identification, and make the parsing easier in the next step. There are significant differences in the words adjacent to the tail word of PP, due to the different components of PP in a sentence (Detailed analysis see section 3). In this paper PP was classified, and then we select features based on the category of PPs to train models of machine learning for PP identification, and recombine these results. In this way, we can make full use of the complementary advantage of multi-model.

This work is supported by Humanities and Social Sciences Foundation for the Youth Scholars of Ministry of Education of China (№-14YJC740126) and National Natural Science Foundation of China (№-61402299).

Q. Lu and H.H. Gao (Eds.): CLSW 2015, LNAI 9332, pp. 566–576, 2015.
DOI: 10.1007/978-3-319-27194-1_57

2 Related Work

State-of-the-art PP identification methods can be classified into two types. One is based on statistical method, the other is a combination of statistical and rule method. The statistical method is based on machine learning. For instance, Support Vector Machine (SVM) [2], Conditional Random Fields (CRFs) [3] model and Maximum Entropy (ME) [4] model were applied to identify PP. Zhang Kunli et al. [5] use the "People's Daily" as corpus to identify PP based on SVM, ME and CRFs models respectively and the result shows that CRFs model obtain the best result. In the paper of Dongfeng Cai et al. [6] PP identification is transformed into the collocation identification of preposition itself and the tail word of PP. The Cascaded Conditional Random Fields (CCRFs) is used in this approach. This approach obtains breakthrough in this specific field as the current F1 is about 8.6% higher than any publicly published paper at that time.

For the combination of rule and statistical methods, the rules were used as a post-processing method to correct errors of the statistical recognition results. Xi Jianqing [7] proposed a model based on Hidden Markov Model (HMM) for identifying the boundaries of PP. They correct the mistakes of HMM recognition by using dependence grammar. Lu Zhaohua et al. [8,9] proposed a model based on ME for identifying the boundaries of PP. They also correct the mistakes in the identification results of ME model by using dependence grammar knowledge. Hu Silei and Huang Degen [10] proposed a two-layer PP identification approach based on CRFs. Their experiment involved many features for statistical decision and 22 rules to correct the result. However, the 22 rules are extracted from the experiment result and the error-prone PPs in test set, therefore these rules are perhaps not applicable to other corpora. Besides, this method can only work for the sentences that have less than two nested PPs. Song Guizhe et al. [11] use dual-layer CRFs to identify PP, and the PP rules are used for post-processing.

When the method of machine learning (ML) was used to identify PP, feature selection is very important. Experimental results show the same ML model combining different features will get different recognition results. In these papers [6,12] experimental results show that the same ML method combined different features will create nearly 9% precision gap. We do deep analysis on the PP, and discover the classification knowledge of PP. Therefore this paper proposes the strategy of multi-model advantage complementation to identify PP based on the classification knowledge of PP.

3 Classification of PP

Preposition is a closed set, and the absolute number is limited, but in Chinese grammar system the preposition played an important role. Preposition definition is different from other parts of speech. PP was formed by preposition combinations with other word, and act as modifiers in Chinese sentences, which can be used as adverbial, attributive, complement, or other components [13]. PPs serve as different sentence constituents, so the locations in the sentences will be different, and have different features of context. The main syntactic constituent PP serves as is the adverbial. When PP serves as the adverbial, PP appears in different positions in the sentence. Some may be around the subject, some only appear between subject and predicate, and some only appear before the subject [14]. The Category Description of PP is as follows:

- When PP serves as the adverbial before the subject, a pause usually appears between the adverbial and main sentence. Examples are shown in Table 1.

Table 1. PP Serves as Adverbial before the Subject

Adverbial of time	PP【 在国家队效力的时候】，他打进了 5 4 个球。 *(When he worked for the national team, he scored 54 goals.)*
Adverbial of place	PP【在国际关系中】，两国确立发展友好合作关系。 *(In international relations, the two countries have established friendly relations and cooperation.)*
Adverbial of action	PP【对于公司的决定】，与会领导给予了高度评价。 *(For the decision of the company, the leaders gave a high degree of evaluation.)*
Adverbial of scope	PP【 除了节省时间外】，它还可降低费用。 *(In addition to saving time, it also reduces the cost.)*
Adverbial of reason	PP【 由于历史的原因】，铁路的国际联运曾两度中断。 *(Due to historical reasons, the international intermodal rail was twice interrupted.)*

- PP serves as the adverbial and appears between subject and predicate. As we all know agent, patient and beneficiaries have a close relationship with predicate. When a preposition is combined with agent, patient or beneficiaries to form PP, this kind of PP has a significant characteristic that the tail word of PP is most close to predicate. Examples are shown in Table 2.

Table 2. PP Serves as Adverbial between the Subject and Predicate

Adverbial of patient	他 PP【把/p 桌子】擦干净了*(He cleaned the table.)*
Adverbial of agent	杯子 PP【被/p 我】打破了*(The cup was broken by me)*
Adverbial of beneficiary	他 PP【替/p 妈妈】干活*(He works for his mother.)*
Adverbial of recipient	我们 PP【同/p 他】讨论合作的事(*We discuss cooperation with him.*)
Adverbial of action direction and source	我 PP【从/p 图书馆】回来(*I come back from the library.*)
Adverbial of tool	他 PP【用/p 那把刀】切菜(*He used the knife for cutting.*)

- There are some prepositional phrases that can act as attributive. Such as “关于”, “对于”, “随着”, “在” and “为了”. Examples are shown as following.

“PP【关于/P 事件/NN 全貌/NN 问题/NN】 的/DEG 诉讼/NN”
(*Litigation concerning the outline of the problem*)
“PP【对于/P 民间/NN 投资/NN 】 的/DEG 拉动/NN 作用/NN”
(*Stimulating role for private investment*)
“PP【 在/P 中国/NR 高速/JJ 公路网/NN 中/LC】 的/DEG 枢纽/NN 地位/NN”。
(*The position of the hub in the Chinese highway network*)

Based on the above analysis and examples PP usually combining with “的” (*de*) act as attributive. That is to say, when PP acts as attributive, the adjacent word of tail word is “的” (*de*).

- When PP acts as complement, the sentence structure is “V + preposition + noun phrase”. Such as “来/v 自/p 北京/n” (*came from Beijing*), “写/v 于/p 上海/n” (*written in Shanghai*).

From the above 4 classification of PP, we could conclude that adjacent words of PP tail word are associated with the sentence constituent PP acted as and the position PP appeared. Adjacent sign of PP tail word are shown in table 3.

Table 3. Adjacent Sign of PP Tail Word

Sentence Constituent of PP	Adjacent sign of PP tail word
Adverbial (before the subject)	comma
Adverbial (between subject and predicate)	predicate
Attributive	“的” (*de*)

4 Identification of PP Based on Multi-model

This paper uses CRFs model to identify PP, so we need to choose the feature of Model according to the training corpus. Identification Model was classified according to the characteristics of PP, so features are needed to correspond to Model. This paper therefore trains multi-model as the same time and use the advantage of multi-model to identify PP.

4.1 Identification Model Category

According to section 3 of this paper PP was summarized and analyzed, and identification Model of PP was divided into three categories.

- There is a collocation between preposition and the tail word of PP. For example, “ 在……上” (*on*), “除了……外” (e*xcept*), “对于……而言” (*for*).
- The adjacent words of the tail word of PP belong to some kind of word or sign due to the syntactic constituent PP served as in sentence as shown in table 3.
- The internal structure of the PP is simple and the tail word of PP is easy to be identified by CRFs Model. For example, “在/p 上海/n” (*at Shanghai*), ”自/p 去年/n” (*last year*).

4.2 Feature Selection Based on the Category

This paper uses CRFs model to identify PP, so we need to select feature sets respectively for every category from section 4.1. Through the experiments on the development set this paper selects feature sets as table 4.

Table 4. Features of Different Models

	Model	Feature
1	Model1	O|W, W, P, O
2	Model2	O|W, W, P, O, F|B|W, |B|P
3	Model3	W, P, O

'W' represents the current word, 'P' represents the part-of-speech of the current word, 'O' represents the preposition, 'F' represents the left adjacent word of preposition that appears first in the left position of current word, and 'B' represents the right adjacent word of current word. Because the first word and the tail word is a collocation in some PP, we treat the first word and tail word of PP as a collocation. Therefore the combination of features is used. "o|w" represents the combination of the first word and the candidate tail word of PP.

4.3 Tagging Sets

This paper transforms the PP identification into collocation identification of PP as Dongfeng Cai [6]. In this paper the tagging sets were composed of "OIEN", wherein 'O' represents the collocation before the PP; 'I' represents the inner collocation of PP; 'E' represents the collocation that was formed by the first word with tail word of PP, and 'N' represents behind collocation of PP. Detailed introduction is given in the following sentence. 李鹏/NR PP[对/P 韦奇立/NR 再次/AD 来访/VV] 表示/VV 欢迎/NN 。/PU (*Li Peng expressed welcome at the periodic visit of Viera.*) For this sentence, the tagging set of "OIEN" in Table 5.

Table 5. Tagging Sets

Tagging set of CRF		
O|W	P	Tag
|李鹏(|*Li Peng*)	NR	O
|对(|*at*)	P	O
对|韦奇立(*at*| *Viera*)	NR	I
对|再次(*at*| *the periodic*)	AD	I
对|来访(*at*| *visit*)	VV	E
对|表示(*at*| *expressed*)	VV	N
对|欢迎(*at*| *welcome*)	NN	N
对|。(*at*|.)	PU	N

In table 5 'w' represents the current word, which is the candidate tail word of PP. 'P' represents the part-of-speech of the current word. The 'tag' shows the tagging sets. "*|李鹏" (*|*Li Peng*) shows that there is no preposition before the current word "李鹏" (*Li Peng*), the preposition is represented by "*"; "对|韦奇力" (*at*| *Viera*) represents that the first word of PP is "对" (*at*), "韦奇力" (*Viera*) is the candidate tail word that may form collocation, at the same time “对” (*at*) appears before “韦奇力” (*Viera*); “*|对” (*|*at*)means that when the preposition “对” (*at*) itself appears as a candidate tail word, the first word of PP is replaced by “*”.

4.4 Experiments on Three Models

We perform the experiments on the Penn Chinese Treebank 4. The corpus contains 1064 files and 15165 sentences. We test on sentences 14126-15162, train on sentences 0-13074 and develop on sentences 13075-14125. The experiment result is tested by recall rate, precision rate and F1. The one with highest F1 has the best performance.

The method that identifies PP of nested structure uses the bottom-up strategy. The identification process is as follows:

1) Scan the sentence from right to left to check for the first preposition. If there is preposition in the sentence, go to step 2, otherwise output the result and finish the algorithm.
2) Use CRFs model to identify the PP.
3) Delete the PP that has been identified in original sentence and generate a new sentence.
4) Repeat step 1 for the new sentence.

Table 6 shows the results of different models on the identification of PP.

Table 6. Comparison of Models

	Model1	Model2	Model3
Development Set	83.59%	84.40%	80.95%
Test Set	85.53%	87.22%	85.53%

As shown in table 6 Model2 is the best Model among the three Models.

4.5 Multi-model Complementary Advantage

Multi-model complementary advantage (MCA) this paper proposed is the fusion results of multiple models based on the multi-model complementary advantage table (MCAT). Based on a combination of the features mentioned in table 4 we trained three models, and used the three models to identify PP on development sets. After analyzing the different identification results of three models, we obtained MCAT as shown in Table 7.

Parts of prepositions in the MCAT are shown in table 7. The first word of PP always is a preposition, so in this test PPs are divided into categories according to the preposition. Table 7 shows the comparative experiments of three models on the same kind of PP. As shown from Table 7, the same preposition has different experimental results based on different models. For instance, PP that was conducted by the preposition of "向"(*xiang*) gained the best identification results on the Model3, so the final Model of "向"(*xiang*) is Model3; PP that was conducted by the preposition "透过 "(*touguo*) gained the best identification results on the three models.

Table 7. Comparison of Models on the Same Kind of PP

P	Count	Model1	Model2	Model3	Best Model	Final Model
以(*yi*)	78	92.31%	93.59%	88.46%	2	2
因(*yin*)	26	65.38%	69.23%	61.54%	2	2
向(*xiang*)	17	82.35%	82.35%	88.24%	3	3
由(*you*)	41	90.24%	90.24%	87.80%	1,2	2
在(*zai*)	284	90.49%	91.20%	89.44%	2	2
透过(*touguo*)	8	87.50%	87.50%	87.50%	1,2,3	2
当(*dang*)	10	100%	90%	90%	1	1
与(*yu*)	39	94.87%	94.87%	92.31%	1,2	2
从(*cong*)	35	82.86%	82.86%	85.71%	3	3

When some kind of PP did not appear in the complementary table because it was not included in the development sets, the priority of models was set according to the table 6. Based on the overall recognition performance of the three models shown in table 6, Model1 is the highest-priority, Model2 is the second and Model3 is the lowest-priority. For instance, PP that was conducted by the preposition of "与"(*yu*) gained the best identification results on Model1 and Model2, so the final Model corresponding to "与"(*yu*) is Model2 on the priority of models. For preposition that did not appear in development set, the default choice is Model2, which is the optimal model corresponding to the priority of Model.

Each kind of PP has an optimal model corresponding to MCAT, so each kind of PP can obtain the best identification results from the three models based on MCAT. According to the above mentioned MCAT was shown in table 8.

Table 8. Part of MCAT

P	Final Model	P	Final Model	P	Final Model
以(*yi*)	2	透过(*touguo*)	2	在(*zai*)	2
因(*yin*)	2	当(*dang*)	1	由(*you*)	2
向(*xiang*)	3	与(*yu*)	2	从(*cong*)	3

5 Experiments

5.1 Methods

Our PP identification system design is shown in Figure 1. Every Model represents applying the CRFs sequence-labeling method to the extracted feature vectors to train the identification PP model. First, we identify PP on development set by use three Models respectively and obtain three results of each PP. Every kind of PP is fed to the

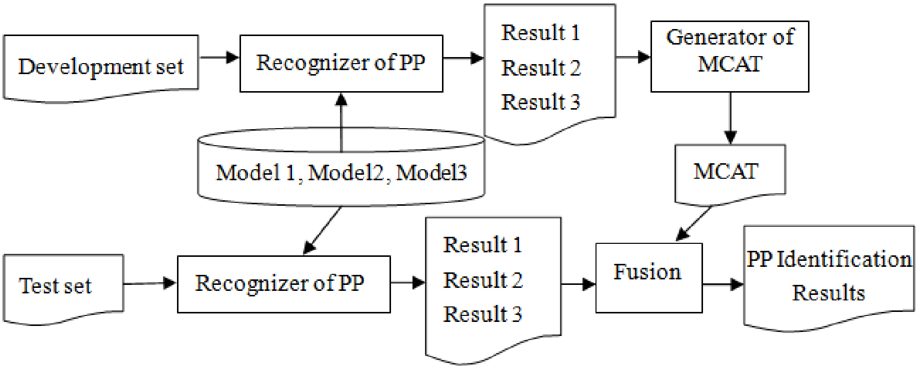

Fig. 1. System Design for PP Identification.

selection process of optimal model, which is generator of MCAT, to build MCAT. Finally, we apply MCAT to choose the PP that was identified by using three Models on the test set and obtained the final results.

5.2 Dataset

We perform the experiments on the Penn Chinese Treebank 4. The corpus contains 1064 files and 15165 sentences. We test on sentences 14126-15162, train on sentences 0-13074 and develop on sentences 13075-14125. We make statistic on corpus of CTB4, and got the average length and count in term of PPs and sentences we show in table 9. Al represents the average of length in table 9.

Table 9. Statistic Data of Corpus

	Training set	Development set	Test set
Count of Sentences	13077	1051	1037
Count of Words	343457	30847	30455
Count of PP	12578	987	1009
AL of Sentence	26.3	29.4	29.4
AL of PP	5.3	5.0	4.7

5.3 PP Identification

In order to enhance the accuracy and authenticity of the experiment and to avoid over fitting phenomenon, Models were tested using ten-fold cross validation method. Ten-fold cross-validation is to collect all the samples and the samples were divided into 10 parts, each subset of the data makes a test set, the remaining 9 parts make the training data set. So it would receive 10 models, and the average accuracy of 10 tests set as evaluation of the entire-system.

The whole corpus is divided into ten parts, the sentences and count of PP are shown in Table 10. Then one of them acts as a test set, and the other nine merge into a training set, which is repeated ten times. The test results are shown in Table 11.

Table 10. Statistic Data of Test Sets

Test set	Range of Sentence	Count of Sentence	Count of PP
test1	8325-13077	1551	1374
test2	9876-11608	1733	2060
test3	11609-13483	1875	1906
test4	13484-15162	1679	1628
test5	1-1370	1371	1338
test6	1371-2755	1385	1260
test7	2756-4135	1380	1245
test8	4136-5514	1379	1318
test9	5515-6927	1413	1245
test10	6928-8324	1396	1200

Table 11. Ten-fold Cross Validation

Test set	Model1	Model2	Model3	MCA
test1	90.9025%	92.2125%	90.393%	91.6303%
test2	86.9417%	87.3301%	86.1650%	87.2333%
test3	84.5226%	85.5194%	84.1028%	85.0472%
test4	85.5037%	86.6708%	83.7224%	87.2236%
test5	90.8072%	91.3303%	89.4619%	91.4798%
test6	91.3492%	91.9048%	88.8095%	92.4603%
test7	91.8876%	91.9679%	89.6386%	92.5301%
test8	89.2261%	90.2124%	88.6191%	90.0607%
test9	90.2811%	90.8434%	88.9960%	90.6024%
test10	90.2500%	91.6667%	88.1667%	91.0833%
Average	89.1671%	**89.9658%**	87.8075%	**90.1351%**

After a ten-fold cross-validation, as can be seen from Table 11, the recognition results of PP based on MCA increased by about 0.2 percentage points over the best single model that is Model2 , 1 percentage point higher than Model1 and 2.3 percentage points higher than Model3. It can be concluded that the MCA is efficient.

As can be seen from Table 11, not every test set of MCA results are the best, which is caused by the following reasons: First, There are great differences in the type and quantity of PP in the ten test sets. Second, types of preposition in development set are unevenly distributed, so development set cannot cover all types of preposition. Third, there are certain limitations on MCAT that was generated only based on the development set. These above issues will also be the focus of further work.

6 Conclusions

Based on the classification knowledge of PP, MCA was proposed for the recognition of PP. Ten-fold cross-validation and comparative experiments also demonstrate the effectiveness and applicability of MCA we proposed. For Chinese sentences, what makes PP different from other kind of phrase is that preposition is closed set and is the first word of PP, so we classify PP precisely according to preposition. Multi Models can be trained according to the type of PP, and we take advantage of multi Models to improve the recognition accuracy of PP.

References

1. Yu, S.: Grammatical Knowledge-base of Contemporary. Tsinghua University Press, Chinese Beijing (1998). (in Chinese)
2. Wen, M., Wu, Y.: Feature-rich Prepositional Phrase Boundary Identification based on SVM. Journal Of Chinese Information Processing **23**(5), 19–24 (2008). (in Chinese)
3. Zhu, D., Wang, D., Xie, J.: Automatic Identification of Propositional Phrase Based on Conditional Random Field. New Technology of Library and Information Service **26**(7/8), 79–83 (2010). (in Chinese)
4. Yu, J., Huang, D.: Automatic Identification of Chinese Prepositional Phrase Based on Maximum Entropy (Master’s degree thesis) Dalian University of technology (2011). (in Chinese)
5. Zhang, K., Han, Y., Zan, H.: Prepositional Phrase Boundary Identification Based on Statistical Models. Journal of Henan University (Natural Science) **41**(6), 636–640 (2011). (in Chinese)
6. Cai, D., Zhang, L., Zhou, Q., Zhao, Y.: A collocation based approach for preposition phrase identification. In: Processing of the 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2011), Tokushima, Japan, November 27–29, 2011 (in Chinese)
7. Xi, J., Luo, Q.: Research on Automatic Identification for Chinese Prepositional Phrase Based on HMM. Computer Engineering **33**(3), 172–182 (2007). (in Chinese)
8. Lu, Z., Huang, G., Guo, Z.: Identification of Chinese Prepositional Phrase. Communications Technology **43**(5), 181–183 (2010). 186 (in Chinese)

9. Lu, Z., Xu, H., Wang, Y.: Re-search on Identification Method of Chinese Prepositional Phrase Based on Semantic Analysis. Computer and Telecommunication **3**, 46–50 (2012). (in Chinese)
10. Hu, S.: Automatic Identification of Chinese Prepositional Phrase Based on CRF (Master's de-gree thesis) Dalian University of technology (2008). (in Chinese)
11. Song, G., Huang, D.: Recognition of Chinese Propositional Phrase (Master's degree thesis). Dalian University of technology (2011). (in Chinese)
12. Zhang, L.: Research on Chinese Preposition Phrase Identification Based on Cascaded Conditional Random Fields (Master's degree thesis). Shenyang Aerospace University (2013). (in Chinese)
13. Zhang, B.: Function Words In Modern Chinese. East China Normal University Press, Shanghai (2000). (in Chinese)
14. Shi, Y.: Chinese Grammatical. The Commercial Press, Beijing (2010). (in Chinese)