Identifying Prepositional Phrases in Chinese Patent Texts with Rule-based and CRF Methods

Hongzheng Li and Yaohong Jin
Institute of Chinese Information Processing, Beijing Normal University
19, Xinjiekou Wai St., Haidian District, Beijing, 100875, China
lihongzheng@mail.bnu.edu.cn, jinyaohong@bnu.edu.cn

Abstract

Identification of prepositional phrases (PP) has been an issue in the field of Natural Language Processing (NLP). In this paper, towards Chinese patent texts, we present a rule-based method and a CRF-based method to identify the PPs. In the rule-based method, according to the special features and expressions of PPs, we manually write targeted formal identification rules; in the CRF approach, after labelling the sentences with features, a typical CRF toolkit is exploited to train the model for identifying PPs. We then conduct some experiments to test the performance of the two methods, and final precision rates are over 90%, indicating the proposed methods are effective and feasible.

1 Introduction

In recent years, patent text information processing (such as patent machine translation) has gradually become an important application field of natural language processing (NLP), and has aroused widespread attention.

Prepositional phrases (PPs), as an important type of phrase, are widely distributed in Chinese patent text, in which the vast majority serve as adverbial components. According to (Li, et al., 2014), in a random sample of 500 Chinese patent sentences, the number of sentences containing PPs are 226, accounting for 45.2% of the total sample, indicating the high proportion of PPs.

In the sentence $S = W_1,W_2,W_3......W_n$, assuming the string W_i, $W_{i+1}......W_j$ is the PP to be identified, the main task of identifying PP is to recognize the word W_i and W_j as left and right boundaries of PP, and identify the whole string as PP chunk. Since W_i is the preposition itself, thus the key issue is to determine the position of W_j.

There exists some following difficulties in identifying PP of Chinese patent texts:

(1) Different with other domain texts, PPs in the patent texts are much longer, with more characters. According to (Gan, et al., 2005; Hu, 2015), the average length of PPs in news texts has 4.9 characters, while 12.3 characters in patent texts. On the other hand, PPs tend to have much more complex structures, which can be composed of prepositions and various kinds of phrases, or even clauses.

(2) Prepositions in Chinese are usually multi-category words, they can also serve as nouns, quantifiers, adjectives, conjunctions and verbs in different contexts.

(3) Several parallel or nested PPs can appear in the same sentence.

Here is an example sentence in the patent texts:
本发明[PP1 在条件允许的情况下][PP2 通过[PP3 为不同区域]提供预测信息]而提出了许多更加准确的结果。
(The invention has proposed more accurate results [PP1 under the permitted condition] {PP2 by providing forecast information [PP3 for various regions.]})

As shown, two parallel PP1 and PP2 appear together in the same sentence, where PP2 also includes a nested PP3.

29th Pacific Asia Conference on Language, Information and Computation pages 143 - 149
Shanghai, China, October 30 - November 1, 2015

Note that, correct identification of PPs is significant to many tasks and applications in NLP. Take patent machine translation for example, PPs have direct impacts on a plurality of processing modules such as source language parsing, transformation and word reordering.

Considering the wide distribution of PPs and significance of correct identification, we propose a rule-based method and Conditional Random Field (CRF) method to recognize the PPs. Although facing difficulties, patent text processing still have its advantages: from words to sentences, patent texts possess kinds of common and fixed structures and expressions, which are more suitable for rule-based approach to describe and process. That’s why we try to use the rule-based approach to identify the special PP chunks.

We test and compare the performances of the two approaches by designing some experiments. Final precision rates were over 90%, indicating that the approaches perform well in our task.

The rest of the paper are organized as follow: Section 2 discusses some related work. Section 3 introduces some structural and semantic analysis of PP in Chinese patent texts. Section 4 and 5 present the rule-based and CRF methods. Section 6 conducts some experiments and analysis, and the last section comes with the conclusion and future work.

2 Related Work

Identification of Chinese prepositional phrases has been an issue in the field of Chinese language processing. Many effective methods, including rule-based and statistical approaches, were proposed in past several years.

(Zhu, 2013; Hu, 2015) studied the identification of PPs towards Chinese-English patent machine translation by using a rule-based method. (Yu, 2006) applied the Maximum Entropy Model to the task of identifying PPs. Based on Hidden Markov Model, (Xi, et al., 2007) presented a novel method to identify PP chunks with dependency grammar, achieving good performance. (Jian, et al., 2009) tried to identify PP from two directions (left-right and right-left) by using the classical SVM classifier.

As a powerful sequence modeling framework that combines the advantages of both generative model and classification model, CRF was first introduced into language processing in (Lafferty, et al., 2001). Since then, the model has been successfully applied to various NLP tasks such as word segmentation (Tseng, et al., 2005), Semantic Role Labelling (Cohn and Blunsom, 2005) and parsing (Finkel, et al., 2008; Yoshimasa, et al., 2009).

(Hu, 2008; Song, 2011 and Zhang, 2013) proposed linear-CRF models to identify PPs in Chinese news corpus, aiming to identify the nested PPs.

Note that, most previous works focus on identifying PPs in news corpus, there exists few research in other domains. In this paper, we want to study some unique features of PP chunks in Chinese patent texts, and try to identify them with two different approaches.

3 Structural Analysis of PP

In this part, we need to introduce some structural and semantic analysis of PP in Chinese patent texts, which are the basis of the rule-based method in the following sections.

3.1 Types of Prepositions

After analyzing considerable Chinese patent texts, we divide the prepositions into two basic types. Some prepositions, such as “把(BA)”, “由(YOU)”, “将(JIANG)” and “被(BEI)”, usually introduce semantic components like agent, patient in the sentence, these can be marked as P0; Other prepositions which can lead the time, manner etc. are marked as P1, including “按/按照/根据 (according to)”, “通过 (by/through)” and so on. A significant difference between the two types is, components behind the P0 prepositions must be NPs, while components behind P1 are not just limited to NPs, and they can be other kinds of phrases or even clauses. Generally, the number of P1 is much more than that of P0.

3.2 Boundaries of PP

PP chunk has left and right boundary words, and the left boundary is preposition. Some right boundary words often appear together with some specific prepositions, forming fixed collocation structures. For example, in the strings “当……时 (when……)” and “在……中(in……)”, the word “时” is the collocation of preposition “当”, and the word “中” is the collocation of preposition “在”. Such PPs with collocation structures are called *explicit PP*. Clearly, prepositions in explicit PP usually belong to P1 type, correspondingly, the right boundary words can be marked as P1H. On the contrary, *implicit PP*, refer to those PPs whose right boundary words have no specific linguistic features and cannot form collocation with the prepositions. The number of implicit PPs are also much more than that of explicit ones.

3.3 Positions of PP

PP in Chinese usually located between the subject and core predicate, forming the “(NP) + PP + VP” format, which is the most common form. Meanwhile, in order to highlight the prepositional phrases, PP can also be separated from subject and predicate by commas, alone as an independent structural unit, forming “PP +, + (NP) + VP” format.

Both the two structures have something in common: Subjects in the sentences can sometimes be omitted; several parallel PPs can exist simultaneously; and the PPs can be either explicit or implicit. But the difference is that prepositions in first format can be either P0 or P1 type, while prepositions of the second format generally can only be P1 type, because PPs introduced by P0 type have much closer relationship with the predicate structures and cannot be separated from them.

3.4 Syntactic levels of PP

For the sake of parsing, it is necessary to distinguish the PPs according to their syntactic levels in the sentences. We define two levels: LEVEL1 and LEVEL2. From the point of syntax tree, the level of PP, whose upper node is the root node of sentence, should belong to LEVEL1, indicating that PPs are direct components of the sentences; and level of other PPs, whose nodes are non-terminals, should belong to LEVEL2. In the example sentence of section 1, for instance, the levels of PP1 and PP2 are LEVEL1, and PP3 belongs to LEVEL2.

4 Rule-based Method

Based on the Chinese patent corpus provided by *State Intelligent Property Office of China* (SIPO), we build a considerable knowledge base and artificially write numerous formal rules. In the knowledge base, all words extracted from the texts are labelled with several syntactic and semantic attributes. According to the P0 and P1 types of preposition, different rules are specially designed to identify the PPs. After integrating the knowledge base and rules into the system, the rules can use information shown in the knowledge base. We will discuss the identification progress by selecting some rules and examples.

4.1 Identifying PP Introduced by P0

As mentioned, PPs introduced by prepositions of P0 types have direct relationship with the predicate structure. We have found that such PPs always appear with two-valence or three-valence verbs. Thus in the rules, it is necessary to take the valence attributes of verb into consideration to help identify the PPs. The valence attributes have already been labelled in the knowledge base.

Rule1:
(0){CHN[与]}+(1)NP+(f){(2)Verb&Valence[2]
&END%}=>(PP,0,1)&PUT(PP,LEVEL,1)

Rule2:
(0){CHN[与]}+(1)NP+(f){(2)Verb&Valence[2]}
+(3)CHN[的]=>(PP,0,1)&PUT(PP,LEVEL,2)

The meaning of rule 1 is that, if there exists a two-valence verb behind the Chinese character (CHN) “与(with)”, and located at the end of the sentence (END%), then the string from node(0) to node (1) will be identified as PP, and its level should be LEVEL1.

Rule2 is similar to rule1, but since the verb is followed by the common auxiliary word “的(DE)”, the PP is just a modifier, and its level will be LEVEL2 instead of LEVEL1.

E.g.1:本发明的结果可以[PP *与*样本指数]*匹配*。(The results of the present invention can be matched [PP with the sample index].)

E.g.2: [NP[PP *与*样本指数]*匹配*的结果]表明了实验的有效性。(The results matched [PP with the sample index] has proved the effectiveness of the experiment.)

4.2 Identifying PP Introduced by P1

PPs introduced by P1 actually include explicit and implicit PPs. For explicit PPs, since the left and right boundary words are collocation, they can be labelled with special marks in the knowledge base and can be first identified. As a result, after identifying them as the boundary words of PP, the whole PP chunk will be recognized easily.

Rule3:

(0)CHN[当]+(f){(1)CHN[时]}=>(PP,0,1)$

The rule means that, if the character "时" is located behind the character "当" in the same sentence, then the string between the two characters will be identified as PP chunk.

E.g.3: [PP *当*产品的性能超过一定阈值*时*]可以出现下图所示的现象。(The phenomenon, as shown in the following figure, can occur [PP when the performance of products exceeds a certain threshold].)

For implicit PPs, since the right boundary words are not collocations of the preposition and have no specific features, it is much difficult to determine the proper positions of the right boundaries. However, we can employ other contextual information and expressions to help recognize them. For example, in many patent sentences, PPs are usually followed by some special conjunctions such as "以(Yi),来(Lai) and 而(Er)". In this case, the word in front of the conjunction will be identified as right boundary. In another case, as mentioned above, if the PP is separated by comma, then it is clearly that the comma can be used to identify the PP chunk.

Rule4:

(0){CHN[通过,经由,经过,基于,根据,藉由]}+(f)
{(2)CHN[以,而,来]}+(1)! CHK[，]=>(ABK,0,2]
&PUT(PP,LEVEL,1)

Rule5:

(0)P1+(f){(1)CHN[，]}=>(ABK,0,1]&PUT(PP,
LEVEL,1)

Rule4 indicates that if there exists Chinese conjunctions behind the prepositions at node 0, then the whole string before the conjunctions (not included) will be recognized as PP chunk (ABK). Rule5 means that the string, which begins with the preposition of P1 type and ends with the comma, will be recognized as PP chunk.

E.g.4: [PP1 根据本发明的实施例]，可[PP2 通过提供动态图像]*来*扩大方法的应用范围。([PP1 According to the embodiment of the present invention], the scope of application of the method can be expanded [PP2 by providing a dynamic image].)

Sum up, the identification rules try to take full advantages of the boundary words and contextual information around to identify PPs. The targeted rules only need to pay attention to local rather than global information in the sentence, thus they are more efficient and effective.

5 CRF Method

In this paper, we will use the CRF++ toolkit (V0.53)[1] to train the model for identifying the PP chunks and test the effects of the method.

5.1 Sequential Labelling

Chunking based on CRF method is usually recognized as sequential labelling issue. Input X is a data sequence to be labelled, and Output Y is a corresponding labelled sequence, which is taken from a specific tag set.

We adopt the B-I-E-O scheme as tag sets to label PP chunks in the sentence. B-I-E refers to Beginning, Intermediate and End elements of PP structure, and O for Outsides of the chunk.

5.2 Features

After analyzing the structural and linguistic features of patent sentences in the corpus, we defined following five effective and representative features for the model. Each feature, as shown below, is composed of feature name and its value.

Feature	Value
Token	Each token in the sentence.
POS	Marks only one proper POS of each word and punctuations (marked as "punc") according to context in the sentence.
Candidate left boundary (CLB)	From the current position of each word, find forward to find the preposition. If the preposition exists, the value is the preposition itself; otherwise marks "N".
Candidate right boundary (CRB)	If current word can be RBW of PP, marks "Y"; otherwise "N".
Candidate last word (CLW)	The word behind the RB, which is also helpful in the identification, is defined as last word (LW). If

[1] http://crfpp.googlecode.com/

	current word is LW, then marks "Y"; otherwise "N".

Table 1. Feature Sets of the CRF Model

After word segmentation, we manually label each sentence sequence including PP chunks with above features.Table2 shows a tagged sequence example.

Words	POS	CLB	CRB	CLW	Tag Set
本发明	n	N	N	N	O
通过	prep	通过	N	N	B
采用	v	通过	N	N	I
先进	a	通过	N	N	I
技术	n	通过	Y	N	E
而	conj	通过	N	Y	O
提高	v	通过	N	N	O
生产力	n	通过	N	N	O
。	punc	通过	N	N	O

Table 2. A Tagged Sentence Example

The first five columns are designed features, and the last column represents tag set of the sequences. According to the format of the CRF toolkit, each column is separated by a separator, and each sentence sequence is separated by a line break.

6 Experiments

In this section, we conducted some experiments to test the performance of the two methods mentioned above, and compared their results. Precision rate (P), Recall rate (R) and F1 are three evaluation metrics of the experiments.

6.1 Data

1000 sentences containing PPs, which were randomly selected from the patent corpus provided by SIPO, were considered as test set of the methods. In the CRF test, we chose another different 5000 sentences as training set from the same corpus to train the model in the toolkit.

6.2 Results

The experimental results of the two methods are shown in the following table.

	P (%)	R (%)	F1 (%)
Rule-based	96.86	74.67	84.33
CRF	92.65	90.07	91.33

Table 3. Experimental Results of the Two Methods

In order to observe the effects that the two methods identified different individual prepositions, we further tested identification precision and recall rates of 10 most frequently appeared prepositions in the test set. Following table and line chart showed the results.

No.	Prep.	RB Method		CRF Method	
		P (%)	R (%)	P (%)	R (%)
1	在(ZAI)	100	90.19	95.63	95.63
2	将(JIANG)	100	61.67	95.95	95.95
3	通过 (TONGGUO)	100	52.27	86.84	86.84
4	由(YOU)	90.67	68.00	69.57	66.67
5	从(CONG)	94.74	85.71	70.00	63.63
6	当(DANG)	100	90.48	87.50	87.50
7	与(YU)	92.6	25.00	88.89	88.89
8	对(DUI)	91.37	70.59	80.00	70.59
9	对于(DUIYU)	100	93.75	100	100
10	向(XIANG)	96.12	55.56	75.00	60.00

Table 4. Identification Results of 10 Most Frequently Appeared Prepositions (in descending order)

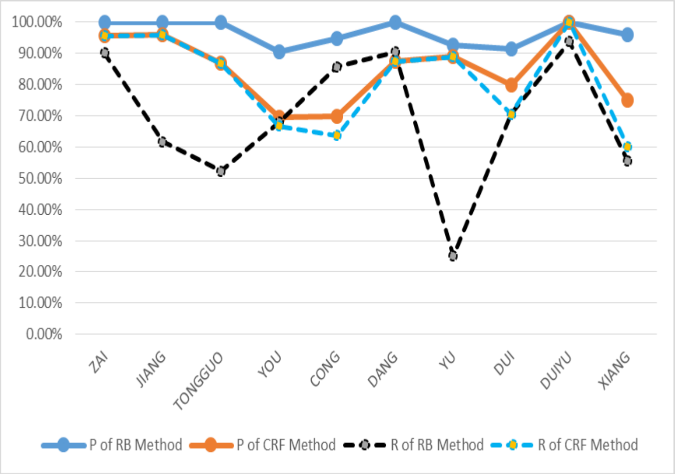

Figure 1. Line Chart of Identification Results

6.3 Analysis

As shown in Table 3, the overall precision rates of the two methods reached over 90%, indicating that the methods are feasible and effective for identifying prepositional phrases, showing a good performance.

Precision of rule-based method were higher than those of CRF in the overall test and identification of 10 prepositions. Identification precision of some individual prepositions even reached 100%, indicating that the rules can describe the linguistic information of PPs more accurately, especially for those PP chunks with long distance and collocations. However, recall rates of rule-based method were much lower than CRF, which were also clearly reflected in the line chart, there exists significant differences between the recall rates of various prepositions, what’s more, fluctuation ranges of recall rates of rule-based method were greater than CRF. From the results, we can come to the conclusion that, as a statistical approach, CRF method does have better stability and adaptability.

On the other hand, the recall rates were lower than precision rates in the two approaches. And, fluctuation ranges between precision and recall of rule-based method were greater than CRF. These are inevitable results of rule-based approaches in NLP.

Despite the methods performed well, we still found some reasons accounting for error identification after analyzing the experimental results.

For the rule-based method, the reasons included:

(1) Because of the performance of the current system itself, sometimes it has difficulties in processing sentences with much longer and complex structures.

(2) Word segmentation ambiguities resulted in error identification. For example, in the sentence “[PP ***将来自***前一步骤的溶液]加入到实验装置中。”(The solution from the previous step was added to the test device.), the word “来自(from)” was behind the preposition “将(Jiang)”, since the word “将来” is already in the word list, the system will first segment the word “将来” from the sentence, thus the monosyllabic word “将(Jiang)” cannot be identified as preposition, as a result, the PP chunk will not be identified at last.

(3) In some cases, it is harder for the system to recognize ambiguous strings caused by multi-category prepositions. For example, in the sentence “应用程序可以使用 SIM 工具包接口*与*移动设备通信” (The applications can use the SIM toolkit interface to communicate with mobile devices.), the preposition “与(YU, with)” can also serve as conjunction(equivalent to the word “and” in English) in Chinese. Thus, when chunking the sentence, the string “与移动设备(with mobile devices)” may not be identified as PP chunk, instead, the string “SIM 工具包接口*与*移动设备” is recognized as NP (the SIM toolkit interface *and* mobile devices).

For the CRF method, the possible reasons included:

(1) Some prepositions had little or no occurrences in the training set, and CRF model cannot study the features of these prepositions, thus it is difficult to identify them correctly when they appear in the test set.

(2) Some strings led by the prepositions were ambiguous. Under this condition, it was not easy to determine the right boundaries of PP chunks. For example, in the sentence “通过本发明的*墨水着色剂/*可以有效地使实验产品沉淀”, the italic noun “墨水(ink)” is followed by another noun “着色剂(colorants)”, it is not really clear which noun should actually be right boundary of the PP chunk. If the two nouns represent a compound noun, then the boundary should be the second noun; but if they are independent of each other, then the boundary should be the first noun, and the second noun will serve as subject of the sentence.

(3) The model is quite sensitive to features in the sequences, during the label process, error and improper manually tagged information is inevitable, which can also result in error identifications.

7 Conclusion and Future Work

In this paper, we proposed a rule-based and CRF method for identifying PP chunks in Chinese patent texts. In the rule-based method, we built the knowledge base and designed various targeted rules for different types of PPs, in the CRF method, we employed the effective CRF toolkit to train the identification models by labelling the sentences with several features. We also conducted several tests to justify the performance of the two approaches and compared the experimental results.

Which have proved the methods performed well in identifying the PPs, although there still existed some error identifications.

In the future, we will try to combine the two method together, and pay more attention to the reasons resulting in the error identification, hoping to improve the performance further.

Acknowledgments

This work was supported by the National Hi-Tech Research and Development Program of China (2012AA011104), we also want to express our sincere thanks to the anonymous reviewers of this article.

References

Guizhe Song. 2011. Research on Identification of Chinese Preposition Phrases.

Guodong Zhou, Jian Su and Tongguan Tey. 2000. Hybrid Text Chunking. In: *Proceeding of CoNLL2000 and LLL*, 163-165.

Hongqiao Li, Chang-Ning Huang, Jianfeng Gao and Xiaozhong Fan. 2004. Chinese Chunking with another Type of Spec. In: *Proceeding of 42nd Association for Computational Linguistics SIGHAN workshop*, 41-48.

Hongzheng Li, Yun Zhu, Yang Yang and Yaohong Jin. 2014. Reordering Adverbial Chunks in Chinese-English Patent Machine Translation. In: *Proceedings of 2014 3rd International Conference on Cloud Computing and Intelligence Systems*, 375-379.

Jenny Rose Finkel, Alex Kleeman and Christopher D. Manning. 2008. Efficient, Feature-based, Conditional Random Field Parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: HLT*, 959-967.

Jianqing Xi and Qiang Luo. 2009. Research on Automatic Identification for Chinese Prepositional Phrase Based on HMM. *Computer Engineering*, 33(3):172-173,182.

Jie Zhang. 2013. Research on Chinese Prepositional Phrase Identification based on Multi-Layer Conditional Random Fields.

Juntao Yu. 2006. Identification of Preposition Phrases Based on Maximum Entropy Model.

Junwei Gan and Degen Huang. 2005. Automatic Identification of Preposition Phrases in Chinese. *Journal of Chinese Information Processing*, 19(4):17-23.

Lafferty J., Mccallum A. and Pereira F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the International Conference on Machine Learning*. 282-289.

Mengjie Liang, Yu Song, Yingjie Han and Hongying Zan. 2013. Automatic Annotation Research on Preposition Usage Based on Sorting Rules. *Journal of Henan Normal University (Natural Science Edition)*, 41(3):152-155.

Ping Jian and Chengqing Zong. 2009. A New Approach to Identifying Chinese Maximal Phrases Using Bidirectional Labelling. *CAAI Transaction on Intelligence Systems*, 4(5):406-413.

Renfen Hu. 2015. on the Methods of Auto-Identifying Prepositional Phrases in Chinese-English Patent Machine Translation. *Applied Linguistics*, (1):136-144.

Silei Hu. 2008. Automatic Identification of Chinese Prepositional Phrase Based on CRF.

Trevor Cohn and Philip Blunsom. 2005. Semantic Role Labelling with Tree Conditional Random Fields. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)*, 169–172.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. 2005. A Conditional Random Field Word Segmenter for SIGHAN Bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.

Wenliang Chen, Yujie Zhang and Hitoshi Isahara. 2006. An Empirical Study of Chinese Chunking. In: *Proceeding of the COLING/ACL 2006*, 97–104.

Yoshimasa Tsuruoka, Jun'ichi Tsujii and Sophia Ananiadou. 2009. Fast Full Parsing by Linear-Chain Conditional Random Fields. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 790–798.

Yun Zhu and Yaohong Jin. 2012. A Chinese-English Patent Machine Translation System based on the Theory of Hierarchical Network of Concepts. *The Journal of China Universities of Posts and Telecommunications*, 19(Suppl. 2): 140-146.