Semantic Analysis of Chinese Prepositional Phrases for Patent Machine Translation

Renfen Hu, Yun Zhu, Yaohong Jin

Institute of Chinese Information Processing, Beijing Normal University, Beijing, China
bnuhurenfen@126.com
diana_zhupier@hotmail.com
jinyaohong@bnu.edu.cn

Abstract. In Chinese patent texts, prepositional phrases(PP) are quite long with complicated structures. The correct identification of PP is very important for sentences parsing and reordering in machine translation. However, existing statistical and rule-based methods perform poorly in identifying these phrases because of their unobvious boundaries and special structures. Therefore, we present a method based on semantic analysis. Chinese prepositions are divided into two categories due to their semantic functions, and more contextual features are employed to identify the phrase boundaries and syntax levels. After integrating into a patent MT system, our method has effectively improved the parsing result of source language.

Keywords: Machine translation; Patent; Prepositional phrase; Identification; Syntactic analysis

1 Introduction

Patent machine translation (MT) is one of the major application fields of MT [1].However, sentences in Chinese patent texts are known for their complicated structures with multiple verbs and prepositions. It is necessary to make syntactic analysis of source language to deal with the long distance reordering and translation, and the identification of Chinese prepositional phrases (PP) plays an important role in this analysis.

After analyzing sentences of 500 Chinese patent texts, we find that a patent sentence contains approximately 1.9 prepositional phrases (PP) in average, and the average length of each PP is 12.3 Chinese characters, while in news corpus PP contains only 4.9 characters in average [2]. Therefore, the identification of PP in patent texts is concerned in this paper, and we will present a method based on semantic analysis. After fully considering the functions and contextual information of PPs, we classify all prepositions into two semantic categories, and define some contextual features for each. In the analysis, phrase boundaries and syntax level are successively determined based on search algorithm and semantic rules. As a result, the correct identification of

adfa, p. 1, 2011.

PPs can help to achieve better results in predicate identification and syntactic reordering.

After integrating our method into an online patent MT system running in SIPO (State Intellectual Property Office of People's Republic of China)[1], we take a closed test and an open test. The result shows that our method has improved the performance of patent translation.

After a discussion of related work in section 2, we will introduce the semantic features in section 3. Section 4 presents the semantic analysis of PP and the processing steps, and section 5 gives the experiment and evaluation. Finally we draw some conclusions in section 6.

2 Related Work

Chinese prepositional phrase differs a lot from English in locations and functions. For this reason, the identification of PP becomes a procedure of crucial importance in Chinese-English machine translation.

Researchers find that PP mainly served as attribute, adverbial, complement or other adjuncts, and verbs in PP cannot be core predicates. Thus with the identification of PPs, we can narrow down the list of probable predicates, and the syntactic analysis can also be greatly simplified [2][3].

In recent years, a number of statistical methods have been proposed to make text chunking, an intermediate step towards full parsing. PPs, as well as other type of phrases, are identified in this analysis with statistical models such as HMM, maximum entropy, SVM and so on [4]. In order to deal specifically with the identification of PPs, linguistic rules are integrated into the statistical methods, which have greatly improved the identification result [5].

However, existing methods perform poorly in identifying PPs in patent sentences. As mentioned above, PPs in Chinese patent texts are often quite long with complicated structures, which may contain nested phrases, or even clauses. In addition, phrase boundaries are often omitted. While in most statistical systems, phrase boundaries are determined by probabilities depending on features of no more than 5 words. Moreover, in both statistical models and linguistic rules, words' features and contextual information are very limited, including only word collocations and part of speeches. On account of data sparseness and limited features, it turns out to be extremely difficult for existing systems to perform well in PP identification in patent corpus.

To solve this problem, we will describe an approach based on semantic analysis, which employs more contextual information and features of prepositions, including their semantic categories, functions, positions, collocations, ambiguities and so on. With the identification of phrase boundaries and their syntax levels, we can parse and reorder a sentence more explicitly.

[1] http://c2e.cnpat.com.cn/sesame.aspx

3 Semantic Features

3.1 Semantic Categories

One of the important differences between Chinese and English is the function of prepositions. In this part, we will define two semantic categories to draw a clear distinction of them.

In the view of semantics, sentences are composed of propositions and arguments, rather than phrases in syntactic structures. For example, semantic roles are used to make shallow semantic analysis in Proposition Bank [6]. Sentences are annotated with two types of roles, thematic and adjunct. Thematic roles mainly refer to the action or state described by a sentence's predicate, such as agent, patient and experiencer, while adjunct roles represent auxiliary information which is not structurally dispensable in a sentence, such as time, location and manner.

English prepositions mainly introduce adjunct roles, while Chinese prepositions introduce both two types of roles. As shown in table 1, Chinese prepositions can be classified into two categories according to the semantic roles[2] they introduce.

Table 1. Semantic categories of Chinese prepositions

Semantic Category	Introduced Roles	Example Prepositions
SC0	Thematic roles, such as agent, patient, theme, etc.	*把, 将, 对, 由*
SC1	Adjunct roles, such as time, location, manner, etc.	*在, 通过, 除了, 根据*

SC is the abbreviation of Semantic Category. In our knowledge base, 15 Chinese prepositions are labeled as SC[0], and 110 prepositions as SC[1].

3.2 Word Collocations

To identify a phase in a sentence, we need to determine the left and right boundaries. As to Chinese prepositional phrase, the left boundary is the preposition, while the right boundary strongly depends on word collocations. We note that SC0 and SC1 are collocated with different components in sentences, and these are important features for the identification of PPs.

SC0s are special Chinese prepositions which are used to emphasize a part of the sentence, or to make nuance of the meaning by changing the word order. Each SC0 must appear together with a predicate. As shown in sentence 1, *由* and *把* are two Chinese SC0 prepositions. *由* is collocated with *激活*, while *把* is collocated with *固定*. Thus the predicates in a sentence can help to determine the right boundary of PPs.

[2] Semantic roles mentioned in this paper are from PropBank, a corpus of text annotated with information about basic semantic propositions. http://verbs.colorado.edu/~mpalmer/projects/ace.html

Sentence 1 *一种由紫外线激活的粘合剂把传感器壳体固定在中支架上。(An adhesive activated by ultraviolet secures the sensor housing to the middle bracket.)*

Similar to English prepositions, SC1 don't collocate with predicates. They are used independently or in collocation with postpositions. The prepositional phases beginning with SC1 can be modifiers of either predicates or noun phrases. In sentence 2, *在下面的酰化纤维素树脂中* and *按照程序* are two PPs beginning with SC1s. We can note that *在* is collocated with postposition *中*, while *按照* occurs independently. In sentence 3, there is no postposition after *通过*, but the conjunction *来* can suggest the right boundary.

Sentence 2 *在下面的酰化纤维素树脂中，按照程序详细地描述适用于本发明的处理酰化纤维素膜的方法等。(In the following cellulose acylate resins, methods for processing a cellulose acylate film, etc. suitably used for the present invention will be described in detail following the procedures.)*

Sentence 3 *所述共享可以通过扩展频谱数字调制来实现。(Such sharing can be achieved through spread spectrum digital modulation.)*

3.3 Verb Valency

Many linguistic theories proposed that a verbal predicate and its arguments can form a predicate-argument structure, in which the arguments help to complete the meaning of the predicate [7][8]. Verb valency(VV) refers to the number of arguments in the structure, and it is an important feature to help us identify SC0 prepositions in sentences.

Sentence 4 *Jane sent me a letter.*

Sentence 5 *Tom hits Bob.*

In sentence 4, *sent* is a predicate with 3 arguments (*Jane, me, letter*), so its valency is 3. In sentence 5, *hit* has only 2 arguments(*Tom, Bob*), so its valency is 2.

In the knowledge base, we label the verb valency for each verb as VV[1], VV[2] or VV[3]. This value has played an important role in the identification of PPs, which will be discussed in detail in the following section.

3.4 Syntax Level

We have stated that in patent texts, sentences often contain nested phrases. In this case, the syntax levels of PPs must be distinguished so as to find the correct right boundary for each preposition. Here we define a LEVEL value for PPs according to their node locations in the syntax tree. In our method, the syntax level of a PP is as same as its preposition's. However, the LEVEL values of SC0 and SC1 depend on different factors.

As to PPs beginning with SC0 prepositions, the LEVEL value is determined by the parent node of the PP. we define a PP as LEVEL[1] if it is a child node of S(sentence), as LEVEL[2] if it is a child node of NP.

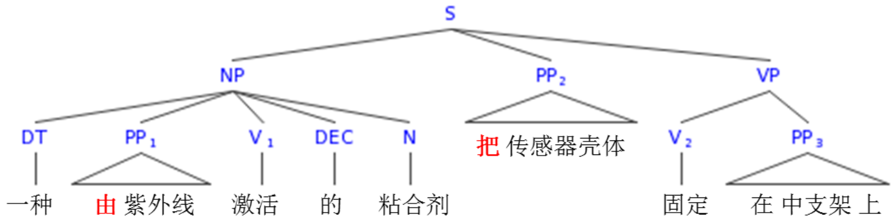

Fig. 1. The syntax tree of Sentence 2

Fig. 1 presents a syntax tree of sentence 1. We can note that PP$_2$(把传感器壳体) appears independently in the sentence, while PP$_1$(由紫外线激活) is nested in a NP(noun phrase). Base on this definition, we can give a syntax level analysis shown in table 2.

Table 2. Syntax level analysis of PPs beginning with SC0s

LEVEL	PP	Parent Node	LB*	RB* Information
[1]	*把传感器壳体*	S	*把*/SC0	V
[2]	*由紫外线*	NP	*由*/SC0	V+*的*

*LB: left boundary; RB: right boundary

As to PPs beginning with SC1 prepositions, we define the LEVEL value as follows. Given two PPs, PPi and PPj, if PPi is nested in PPj, then PPi is LEVEL[2], PPj is LEVEL[1].

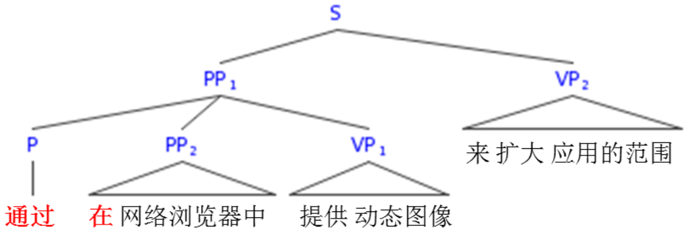

Fig. 2. The syntax tree of a patent sentence

Fig. 2 presents a sentence with two SC1 prepositions. We can also give a syntax level analysis of the PPs in table 3.

Table 3. Syntax level analysis of PPs beginning with SC1s

LEVEL	PP	LB	RB Information
[1]	*通过在网络浏览器中提供动态图像*	*通过*/SC1	*来*
[2]	*在网络浏览器中*	*在*/SC1	*中*

4 Semantic Analysis of Chinese Prepositional Phrases

Given a patent sentence S=W_1, W_2, W_3 … W_{n-2}, W_{n-1}, W_n, let W_i be a preposition, i.e. the LB(left boundary), and W_j be the RB(right boundary). Therefore, to identify a PP, we need to determine three parameters, W_i, W_j and the syntax level of the phrase. In this section, we will discuss the semantic analysis of PPs beginning with SC0 and SC1 separately.

4.1 The identification of PPs beginning with SC0

Chinese PP beginning with SC0 has no obvious right boundary(RB). However, as we mentioned above, each SC0 must appear together with a predicate, so the valency and location of verbs have played important roles in the identification.

Table 4. Basic collocations of SC0s and verbs

W_i(LB)	W_j(RB)	W_{j+1}	W_{j+2}	LEVEL	Example Sentence
SC0	-	Verb & VV[2]	PU*	1	*硬件结构也仅由一块半导体芯片实现。*
SC0	-	Verb & VV[2]	*在,到,给,成,为,至,于*	1	*通信模块将数据发送到计算机系统。*
SC0	-	Verb & VV[2]	*的*	2	*一种与抗原蛋白靶位互补的肽*
SC0	-	Verb & VV[3]	!PU&!的	1	*工作人员把药物注入容器。*

* PU refers to Chinese punctuations, such as ，, ；, and 。.

As shown in table 4, W_i can be determined by the collocation of SC0 and a verb(W_{j+1}), and there are four basic types of collocations, in which the right boundary(RB) and syntax level of a PP are dependent on verb valency of W_{j+1} and its location. It is important that some collocations are not applicable to all SC0s. For example, *与* cannot collocate with a structure of *Verb&VV[2]+在,到,给,成,为,至,于*. These details are considered in our semantic rules.

After fully considering the contextual information, we have made 43 rules for the identification, including 2 steps. In step 1, PPs are identified as LEVEL 1. In step 2, the phrases are identified as LEVEL 2. In our model, rules are circularly matched until the system has nothing new to output. If a phrase is given multiple LEVEL values, take the last one. Here Sentence 1 is taken as an example to illustrate the identification process.

Sentence 1 *一种由紫外线激活的粘合剂把传感器壳体固定在中支架上。(An adhesive activated by ultraviolet secures the sensor housing to the middle bracket.)*

Step 1 *由* and *把* are both identified as LEVEL[1] by matching Rule 1.

Step 2 *由* is identified as LEVEL[2] by matching Rule 2.

Rule 1 *(0)SC[0]+(f){(m)Verb&VV[2]}+(m+1)CHN[在,到,给,成,为,至,于]=> LB(0)+RB(m-1)+PUT(LEVEL,1)*

Rule 2 *(0)SC[0]+(f){(m)Verb&VV[2]}+(m+1)CHN[的] => LB(0)+RB(m-1)+ PUT(LEVEL,2)*

After the analysis module, *由紫外线* is identified as a PP in LEVEL[2], and *把传感器壳体* is identified as PP in LEVEL[1].

In addition to the collocations of SC0s and verbs, we also find other information that can help to determine the level of SC0. For example, if a SC0 appears behind a SC1 preposition, numeral, quantifier or pronoun, we can put it in LEVEL[2].

4.2 The identification of PPs beginning with SC1s

As we discussed in section 3, SC1 prepositions mainly introduce adjunct roles, thus in most cases the PPs beginning with SC1s are modifiers of predicates or NPs. With SC1 as a certain left boundary, we need to determine the right boundary and syntax level of the PP. Note that the LEVEL value is given only when PPs are nested, so not all PPs has LEVEL values. After analyzing 500 Chinese patent texts, we have found different contextual information and collocations for SC1 prepositions. Table 5 shows some basic identification patterns, and SC1 varies in different collocations.

Table 5. Basic patterns for identification of PPs beginning with SC1s

W_i(LB)	W_j(RB)	W_{j+1}	Example Sentence
SC1	Postposition	-	*对于HTTP摘要而言将是这种情况。*
SC1	-	*以, 来, 而*	*可使用DNS来供给任意网络服务。*
SC1	-	PU	*为了找到突发脉冲的最优定时，*
SC1	-	Predicate	*所捕捉的图像对于该设备呈现白色。*
SC1	-	SC0	*其根据场景光源对数据进行处理。*

Based on above conclusions, we have developed a 3-Step identification model. In Step 1, LBs and RBs are generated in the positions of SC1s and postpositions. In Step 2, we check if the LB or RB of current node is nested in another PP with search algorithm, and give LEVEL value to the LB or RB of nested phrase through 12 semantic rules. In Step 3, all PPs are generated and LEVEL values are given to the nested phrases. Here we take sentence 6 as an example to illustrate the semantic analysis.

Sentence 6 *根据本发明的示例性实施例，可通过在网络浏览器中提供动态图像来扩大UI显示方法的应用范围。(According to exemplary embodiments of the present invention, by providing a dynamic image in a web browser, the range of applications of the UI display method can be enlarged.)*

Step 1 *根据, 通过, 在* are identified as LBs, and *中* is identified as a RB.

Step 2 *在* is identified as LB in LEVEL[2], *中* is identified as RB in LEVEL[2].

Step 3 By matching the following 3 rules, the system generates three PPs, PP1(*根据本发明的示例性实施例*)，PP2(*通过在网络浏览器中提供动态图像*) and PP3(*在网络浏览器中*). PP3 is given LEVEL[2].

Rule 3 *(0)LB&CHN[根据]+(f)(m)CHN[，]+(f)(0,m)!Verb=>RB(m-1)+PP(0,m-1)*

Rule 4 *(0)LB&CHN[通过,利用,采用,使用,用]+(f)(m)CHN[以,而,来]=>RB(m-1)+PP(0,m-1)*

Rule 5 *(0)LB&LEVEL[2]}+(f)(m)RB&LEVEL[2] =>PP(0,m)+PUT(LEVEL,2)*

5 Experiment and Evaluation

The experiment takes 500 authentic patent texts provided by SIPO (State Intellectual Property Office of China) as the training set. The evaluation will use the development data for the NTCIR-9 Patent Machine Translation Pilot Task[3], containing 2,000 bilingual Chinese-English sentence pairs.

After integrating the method into a Chinese-English patent machine translation system [9], we take a closed test on training set, and an open test on evaluation set. The precision and recall are calculated for both two tests to evaluate the identification of PPs after semantic analysis. Necessarily, only when the LB, RB and syntax level of a PP are all correctly identified, we count it as a correct identification. In the open test, BLEU score[10] is also employed to evaluate the translation performance. Table 6 shows the result of the closed test.

Table 6. Experiment Result on the Training Set

	Precision(%)	Recall(%)
PP (SC0)	90.91	84.51
PP (SC1)	90.71	88.77

We can note that the recall is lower than precision for both two types of PPs. Two reasons can account for this phenomenon. (1) The preposition is not recognized as a left boundary due to mistakes of segmentation and word sense disambiguation. For example, *对调焦误差信号* is a PP beginning with *对*(SC0), but *对调* is segmented as a word. In the sentence *顾客将编码游戏卡插入其内*, *将* is identified as a verb modifier, not a preposition. (2) In this system, we make strict conditions for the generation of PPs, which might also result in a lower recall.

In the open test, comparison is made as shown in table 7. RB-MT is the baseline system running on SIPO. HYBRID-MT is the system integrated with our semantic analysis. Google is an online statistical MT system, the identification result of which is inferred from its translation result, so we count its identification as correct when the LB and RB are identified correctly, regardless of the syntax level and reordering result.

Table 7. Compared result of PP identification in the open test

	Precision (%)		Recall (%)		F-score (%)	
	PP(SC0)	PP(SC1)	PP(SC0)	PP(SC1)	PP(SC0)	PP(SC1)
RB-MT	71.23	82.51	62.02	74.30	66.31	78.19
HYBRID-MT	88.11	94.09	75.90	89.25	81.55	91.61
GOOGLE	60.71	76.44	51.20	68.22	55.56	72.10

[3] http://research.nii.ac.jp/ntcir/ntcir-9/data.html

The result of the open test shows that the semantic analysis has effectively improved the identification result of Chinese PPs, and Google performs poorly in this test. It is mainly because statistical methods face difficulties in determining the RB of a long phrase, and technical texts(including patent texts) account for a fairly low proportion in the training bilingual corpus. Thus, our method is advantageous in processing technical texts with long and complicated sentences. In addition, we find that the identification result of PPs with SC1 is generally better than PPs with SC0. According to statistics, about 40% PPs with SC1 have postpositions as the certain right boundaries, while PPs with SC0 does not have any obvious right boundaries, the identification of which mainly depends on contextual information. After calculating the precision and recall, we give the BLEU-4 score of the three systems shown in table 8.

Table 8. The BLEU scores of three MT system

System	BLEU-4
RB-MT	0.1997
HYBRID-MT	0.2233
GOOGLE	0.3076

From table 8, we can see that after integrating the semantic analysis, the BLEU score has increased by 11.82% from 0.1997 to 0.2233. However, the BLEU scores of three systems all not very high, the highest is 0.3076 of Google. It is mainly because the corpus domain is not limited, unknown terms or entities may result in a bad translation performance, and in BLEU-4 evaluation, sentence will be given a score of 0 if it does not have at least one 4-gram match. Besides, we need to note that Google performs better in word selection, so our system needs to improve this module urgently.

After the experiment, we also make analysis of the identification errors and summarize 5 problems that need to be solved in the future. (1) A sentence may contain multiple verbs, which would interfere with the semantic rules; (2) PPs across comma and PPs of nesting level ≥3 have not been considered yet; (3) There are labeling mistakes in the knowledge base that needs a careful review; (4) Preprocessing module (word segmentation and sense disambiguation) as we mentioned above also needs to be improved; (5) Our method is strongly dependent on the completeness of rules, which still need to be complemented.

6 Conclusion

To deal with the identification of Chinese prepositional phrases in patent sentences, we present a method based on semantic analysis, and integrate it into a source language parser for patent machine translation.

By identifying PPs of two semantic categories, our method has enhanced the performance of patent machine translation. In the future, the rule set and knowledge base need to be improved, as well as the other analysis modules in the MT system. Furthermore, our identification method can be extended to language parsing of technical texts in other fields.

ACKNOWLEDGMENT

The authors are grateful to *Multi-level Knowledge Representation of Mass Texts and Chinese Language Understanding System* (National 863 Program, No.2012AA011104) and the Fundamental Research Funds of Central Universities for financial support.

Reference

1. Jin, Y., Liu, Z.: Improving Chinese-English Patent Machine Translation Using Sentence Segmentation. In: 7th International Conference on Natural Language Processing and Knowledge Engineering, pp. 620-625, Tokushima (2011)
2. Gan, J., Huang, D.: Automatic Identification of Chinese Prepositional Phrase (in Chinese). J. Chinese Information. 19, 17-23 (2005)
3. Yin, L., Yao, T., Zhang, D., Li, F.: A Hybrid Approach of Chinese Syntactic and Semantic Analysis (in Chinese). J. Chinese Information. 04, 45-51 (2002)
4. Yu, J.: The Automatic Identification of Chinese Prepositional Phrase based on Maximum Entropy (in Chinese). Dalian University of Technology (2006)
5. Lu, C., Xu, H., Wang, Y.: Research on the Identification of Chinese Prepositional Phrase base on Semantic Analysis (in Chinese). J. Computers and Telecommunications. 03, 46-48(2012)
6. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles. J. Computational Linguistics. 31, 71-106(2005)
7. Gildea, D., Palmer, M.: The Necessity of Parsing for Predicate Argument Recognition. In: Proceedings of the 40th Meeting of the Association for Computational Linguistics, pp. 239-246, Philadelphia(2002)
8. Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Schasberger, B.: The Penn Treebank: annotating predicate argument structure. In: Proceedings of the Workshop on Human Language Technology, pp. 114-119, Plainsboro(1994)
9. Wang, D.: Chinese to English automatic patent machine translation at SIPO. J. World Patent Information. 31, 137-139(2009)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176 (W0109-022), IBM Research Report.