A CRF Method of Identifying Prepositional Phrases in Chinese Patent Texts

Hongzheng Li and Yaohong Jin
Institute of Chinese Information Processing, Beijing Normal University,
Beijing, 100875, China
CPIC-BNU Joint Laboratory of Machine Translation, Beijing Normal University,
Beijing, 100875, China
lihongzheng@mail.bnu.edu.cn,jinyaohong@bnu.edu.cn

Abstract

This paper presents a Conditional Random Field (CRF) method of identifying prepositional phrases (PP) in Chinese patent documents. By using the CRF model, the identification process can be recognized as sequence labelling issue. After analyzing the characteristics of PP chunks in large scale corpus, we design several essential and helpful features and feature templates for recognizing PP chunks, and then use a CRF toolkit to train the model to identify PPs. At last, some experiments are conducted to justify the effects of the model, both the precision and recall rates are over 92%, higher than the baseline, indicating the method is reasonable and effective.

1 Introduction

Prepositional phrases (PP), as a traditional important phrase type, are widely distributed in Chinese patent documents. According to (Li, et al., 2014), in 500 randomly extracted sample patent sentences, 226 sentences contained PP chunks, accounting for 45.2% of the sample. Compared with other Chinese domain texts, PP chunks in patent documents tend to have following more specific features.

To begin with, they usually have more complex and longer structures with more words, they can be composed of prepositions (prep.) and noun phrases (NP), verb phrases (VP) or even clauses. Second, some preposition in PP are multi-category words, the preposition may also serve as a noun, verb, conjunction etc. in various contexts. Last but not least, there also exists many parallel and nested PPs. While coordinate PPs means several PPs appear together in a sentence, nested refer to those PPs composed of another PP and other ingredients. Following is an example in patent texts:

该真空工具[PP1 通过[PP3 在控制器中]连接这些网络环片段][PP2 为实验装置]提供一个低温泵。(The vacuum tool can provide a pump for the experiment instrument by connecting the network ring parts in the controller.)

As shown, the example contains three PPs, in which PP1 and PP2 are parallel, and in the long nested PP1 chunk "通过……片段", there exists another PP3 "在控制器中"(in the controller).

All these features result in more difficulties in identifying PPs. However, it is worth noting that, recognizing the PPs properly plays positive roles in various application fields of Natural Language Processing (NLP).

Assuming in the Chinese sentence S=W_1, W_2, W_3……W_n, string W_i, W_{i+1}……W_j is the supposed PP, the main task of PP identification is actually to identify W_i as left boundary word(LBW)and W_j as right boundary word(RBW) of the PP and then recognize the whole string as PP chunk. More specifically, since the LBW is the preposition itself, how to identify the RWB correctly is a key issue in the whole identification process.

Considering the wide distribution of PPs in patent documents and its important impacts on processing modules such as chunking and parsing in NLP, in this paper, we tried to apply the Conditional Random Field (CRF) model to PP identification in patent texts. By designing some features and labelling the PP sequences in corpus first and then training the features with the CRF toolkit, PP chunks can be identified. We also conducted experiments to justify the effects of the method, and the experimental results showed the proposed approaches can improve the performance of identifying Chinese PPs significantly.

The rest of this paper are organized as follow. Section 2 discusses some related work, section 3 presents the CRF-based identification method, section 4 conducts some experiments and gives

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 86–90,
Beijing, China, July 30-31, 2015.

related analysis, and the last section discusses the conclusion and future work.

2 Related Work

As a powerful statistical sequence modeling framework that combines the advantages of both the generative model and the classification model, CRF was first introduced into language processing in (Lafferty, et al., 2001). Since then, the model has been 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), gaining great achievement. And CRF has become increasingly popular in recent years.

PP structures in sentences has been studied for long decades. However, differences in syntactic structures between Chinese and English have resulted in various research strategies: for English PP, researchers mainly focus on PP attachment disambiguation based on statistic and corpus methods (Brill, et al., 1994; Pantel and Lin, 1998; Briscoe and Carroll, 1995; Schwartz, et al., 2003; McLauchlan, 2004).

On the other hand, for Chinese PP, mainly focus on identifying and parsing the PP chunks by using rule-based method (Liang, et al. 2013, Hu, 2015) and popular statistical models, including HMM (Xi and Luo, 2007; Zhang, et al., 2011), SVM (Wen and Wu, 2009), Maximum Entropy (ME) Model (Lu, et al., 2010), and CRF models (Tan et al., 2005; Hu, 2008; Zhang, 2013). (Chen, et al.)(2005) proposed four models (SVMs, CRFs, TBL and MBL) to describe an empirical study of Chinese chunking on a corpus extracted from UPENN Chinese Treebank-4 (CTB4). Some others (Fu and Li, 2010; Zan, et al., 2013) also presented hybrid methods to recognize PPs by combining rules with statistic methods. Generally, recognizing Chinese PPs belongs to the category of shallow parsing in NLP.

While the CRF method has been usually applied to identifying Chinese PPs in common newswire texts, there exists few research on other specific domains. Thus, we decide to apply the method in patent documents.

3 CRF Identification Model

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

3.1 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. The probability assigned to a labelled sequence for a particular sequence of characters by a CRF model can be defined as follow:

$$P(Y|X) = \frac{1}{Z(X)} exp(\sum_k \lambda_k f_k) \quad (1)$$

Where $Z(X)$ is the normalization factor, f_k is a set of features, and λ_k is the corresponding weight.

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

Here is an example in patent text:

本发明*通过采用先进技术*而提高生产力。
(The invention improves the productivity by adopting advanced technology.)

The italic string "通过……技术"is the PP chunk. After word segmentation processing, the sentence can be labelled as:

本发明 O 通过 B 采用 I 先进 I 技术 E 而 O 提高 O 生产力 O 。O

Thus, Input X = {本发明 通过 采用 先进 技术 而 提高 生产力 。}

Correspondingly, Output Y = {O B I I E O O O O}

3.2 Features

Features play a very important role in the CRF model. Although CRF can define features indefinitely, the more features don't always means the better training result. After analyzing the structural and linguistic features of patent sentences in large scale corpus, we defined following five effective and representative features for the model. Each feature is composed of feature name and its value.

Feature	Value
Word	Each word itself in the sentence.
POS	POS of each word and punctuations (marked as "punc") in the sentence.
Candidate left boundary (LB)	From the current word, find forward to find the prep. If the preposition exists, the value is the preposition itself; otherwise "N".

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

Candidate right boundary (RB)	If current word can be RBW of PP, mark "Y"; otherwise "N".
Candidate LW	The word behind the RB, which is also helpful in the identification, is defined as last word (LW). If current word is LW, then mark "Y"; otherwise "N".

Table 1. Feature Sets of the Model

After word segmentation, we manually label each patent sentence that includes PP chunks with above features.

Table 2 shows a tagged sequence example in part 3.1.

Words	POS	Candidate LBW	Candidate RBW	Candidate LW	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 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.

3.3 Feature Templates

We also design essential feature templates for the model according to the defined feature sets. The model generates numerous feature functions, which will directly affect the performance of the model in turn.

CRF models generally include atomic and composite feature templates. Since alone atomic feature templates only show feature information of single locations, which is likely to cause greater deviations between expectations and actual results, leading to inaccurate estimation parameters. Therefore, in our paper, the atomic features are combined to form composite feature templates to describe dependencies between the characteristics of labelled units and contexts by defining window of each feature.

The size of window in the sequences is defined as two. That means, we consider the features of current word (W_0), next word (W_1), second character back W_0 (W_2), previous character (W_{-1}) and second character before W_0 (W_{-2}). All the templates are in the form of Unigram in the toolkit to train the data, and no Bigram templates are used.

3.4 Architecture

Here's the basic architecture of the CRF model for identifying the PP chunks.

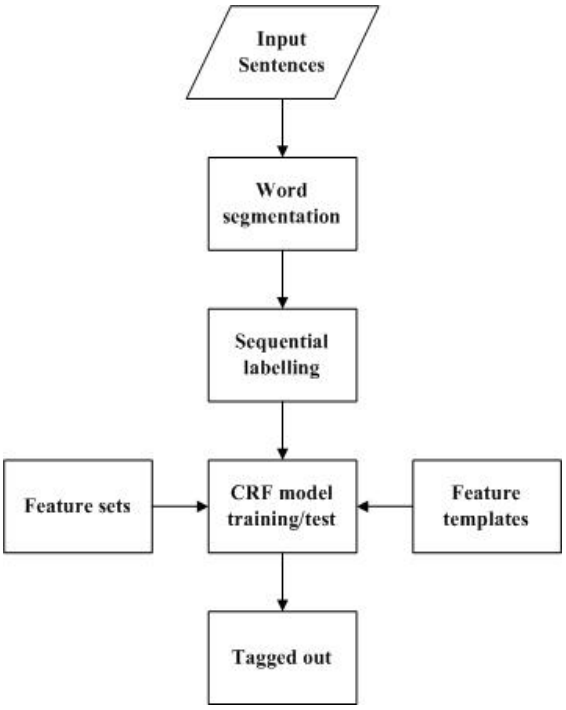

Figure 1. CRF Model Architecture

4 Experiment

After training the model, in this part, we continue use the toolkit to test the identification effects. Precision rate (P), Recall rate (R) and F1, defined as follows, are three evaluation metrics of the experiment.

$$P = \frac{N2}{N1} \times 100\% \quad (2)$$

$$R = \frac{N2}{N} \times 100\% \quad (3)$$

$$F1 = \frac{2 \times P \times R}{P+R} \times 100\% \quad (4)$$

Where N refers to the total number of PP chunks in the test set, $N1$ refers to the identified number of PP by the model, and $N2$ refers to the correctly identified number of PP.

4.1 Test Results

We manually extracted 1017 sentences containing PP chunks as the final test set from the developing set of patent MT subtask in the NTCIR-9 workshop[2] , which is composed of 2000 parallel Chinese-English sentences.

The experiment adopted the 5-fold cross validation method: the whole set was divided into five equal parts, in which four parts were used as training sets, and the other one as test set. Thus, we totally conducted five experiments, and the average data of the five experiments were considered as final results. Then, we compared the results with the baseline (Hu, 2015), which used the same test set and tested with a rule-based system (Zhu and Jin, 2012).

Performances of the five experiments and comparison are shown in the following tables.

Test	P (%)	R (%)	F1 (%)
Test1	94.36	91.09	92.70
Test2	92.41	91.77	92.97
Test3	93.10	95.30	94.19
Test4	93.83	92.12	93.51
Test5	91.68	93.22	92.44
Average	**93.08**	**92.71**	**93.16**

Table 3. Performances of the experiments

	P (%)	R (%)
Baseline	90.81	86.64
CRF	93.08	92.71
Gain	**+2.27**	**+6.07**

Table 4. Comparison of Baseline and CRF

4.2 Discussion

In the experiments, the final metrics were all over 92%, and were higher than baseline, clearly indicating that the method performed well in identifying the PP chunks. Different from other three tests, the reason why the recall rates in test 3 and test 5 were higher than the precision rates lied in that the identified numbers of PP were more than the total numbers of PP in the two tests.

Since most experiments in previous related works employed newswire corpus as test set, totally different from the patent texts, thus we suppose that there may exist no necessary comparisons between our results with previous works.

After analyzing the results, we also concluded several following reasons accounting for error identifications:

First, some prepositions almost did not appear in the training test, as a result, the model could not obtain their features, and consequently, while they appeared in the test set, they were more difficult to be correctly identified.

Second, some PP chunks were ambiguous. Under this condition, it was not easy to determine the right boundaries of the 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 the proper boundary of the PP chunk. If the two nouns represent a compound noun, then the boundary should be the second noun; on the contrary, 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.

Last, 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.

5 Conclusion and Future Work

In this paper, we presented a CRF-based method for identifying the Chinese PP chunks in patent texts. Based on analysis of large scale patent texts, we designed several essential features for the model according to various characteristics of Chinese PPs, after labelling the sequences and training the model by using a CRF toolkit, we conducted some experiments to justify the performance of the method. Both final precision and recall rates were over 92%, and higher than the baseline, indicating the CRF-based method is effective and performs well in identifying PPs, although there still existed some error identifications.

In the future, we will pay more attention to the ambiguous PP chunks, consider more useful and effective features into the model, and continue to expand the size of patent corpus to be labelled, hoping to further improve the identification effects of PP chunks.

Acknowledgements

This work was supported by the National Hi-Tech Research and Development Program of China (2012AA011104).

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

Reference

Brill E. and Resnik P. 1994. A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation. In *Proceedings of the 15th Conference on Computational Linguistics*, 1198-1204.

Chaohua Lu, Guangjun Huang and Zhibing Guo. 2010. Identification of Chinese Prepositional Phrase Based on Maximum Entropy. *Communications Technology*, 43(5):181-183,186.

Edward Briscoe and John Carroll. 1995. Developing and Evaluating a Probabilistic lr Parser of Part-of-Speech and Punctuation Labels. In *Proceedings of the IWPT*, 48–58.

Goto, I., Lu, B., Chow, K. P., Sumita, E., and Tsou, B. K. 2011. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In *Proceedings of NTCIR9*, 559-578.

Hefang Fu and Zhaoxia Li. 2010. Discussion on the Integration of Statistical Learning Method and Artificial Rule Method for Prepositional Phrase Recognition. *Modern Computers*, 11:17-20.

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

Hongying Zan, Tengfei Zhang and Kunli Zhang. 2013. Automatic Recognition Research on Preposition's Usages Based on Combination of Rules and Statistics. *Computer Engineering and Design*, 34(6):2152-2157.

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.

Jenny Rose Finkel, Alex Kleeman and Christopher D. Manning. 2008. Efficient, Feature-based, Conditional Random Field Parsing. In *Proceedings of ACL*, 959-967.

Kunli Zhang, Yingjie Han, Hongying Zan and Yingcheng Yuan. 2011. Prepositional Phrase Boundary Identification Based on Statistical Models. *Journal of Henan Normal University (Natural Science Edition)*, 41(6): 636-640.

Lafferty, John, A. McCallum, and F. Pereira. 2001. Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning 2001*, 282-289.

Mark McLauchlan. 2004. Thesauruses for Prepositional Phrase Attachment. In *Proceedings of CoNLL*, 73-80.

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.

Miaomiao Wen and Yunfang Wu. 2009. Feature-rich Prepositional Phrase Boundary Identification Based on SVM. *Journal of Chinese Information Processing*, 23(5):19-24.

Pantel P, Lin D. 1998. An Unsupervised Approach to Prepositional Phrase Attachment Using Contextually Similar Words. In *Proceedings of Association for Computational Linguistics*, 101-108.

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

Schwartz L, Aikawa T, Quirk C. 2003. Disambiguation of English PP Attachment Using Multilingual Aligned Data. In *Proceedings of MT Summit IX*.

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.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. 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 Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, 97–104.

Yongmei Tan, Tianshun Yao, Qing Chen, and Jingbo Zhu. 2005. Applying Conditional Random Fields to Chinese Shallow Parsing. In *Proceedings of CICLing-2005*, 167–176.

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 ACL*, 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*, 140-146.