Automatic Identification of Predicate Heads in Chinese Sentences

Xiaona Ren[a] **Qiaoli Zhou**[a] **Chunyu Kit**[b] **Dongfeng Cai**[a]
Knowledge Engineering Research Center[a]
Shenyang Aerospace University
Department of Chinese, Translation and Linguistics[b]
City University of Hong Kong
rxn_nlp@163.com ctckit@cityu.edu.hk

Abstract

We propose an effective approach to automatically identify predicate heads in Chinese sentences based on statistical pre-processing and rule-based post-processing. In the pre-processing stage, the maximal noun phrases in a sentence are recognized and replaced by "NP" labels to simplify the sentence structure. Then a CRF model is trained to recognize the predicate heads of this simplified sentence. In the post-processing stage, a rule base is built according to the grammatical features of predicate heads. It is then utilized to correct the preliminary recognition results. Experimental results show that our approach is feasible and effective, and its accuracy achieves 89.14% on Tsinghua Chinese Treebank.

1 Introduction

It is an important issue to identify predicates in syntactic analysis. In general, a predicate is considered the head of a sentence. In Chinese, it usually organizes two parts into a well-formed sentence, one with a subject and its adjunct, and the other with an object and/or complement (Luo *et al*., 1994). Accurate identification of predicate head is thus critical in determining the syntactic structure of a sentence. Moreover, a predicate head splitting a long sentence into two shorter parts can alleviate the complexity of syntactic analysis to a certain degree. This is particularly useful when long dependency relations are involved. Without doubt, this is also a difficult task in Chinese dependency parsing (Cheng *et al*., 2005).

Predicate head identification also plays an important role in facilitating various tasks of natural language processing. For example, it enhances shallow parsing (Sun *et al*., 2000) and head-driven parsing (Collins, 1999), and also improves the precision of sentence similarity computation (Sui *et al*., 1998a). There is reason to expect it to be more widely applicable to other tasks, e.g. machine translation, information extraction, and question answering.

In this paper, we propose an effective approach to automatically recognize predicate heads of Chinese sentences based on a preprocessing step for maximal noun phrases [1](MNPs). MNPs usually appear in the location of subject and object in a sentence. The proper identification of them is thus expected to assist the analysis of sentence structure and/or improve the accuracy of predicate head recognition.

In the next section, we will first review some related works and discuss their limitations, followed by a detailed description of the task of recognizing predicate heads in Section 3. Section 4 illustrates our proposed approach and Section 5 presents experiments and results. Finally we conclude the paper in Section 6.

2 Related Works

There exist various approaches to identify predicate heads in Chinese sentences. Luo and Zheng (1994) and Tan (2000) presented two rule-based methods based on contextual features and part of speeches. A statistical approach was presented in Sui and Yu (1998b), which utilizes a decision tree model. Gong *et al*. (2003) presented their hybrid method combining both rules and statistics. These traditional approaches only make use of the static and dynamic grammatical features of the quasi-predicates to identify the predicate heads. On this basis, Li and Meng (2005) proposed a method to further utilize syntactic relations between the subject and the predicate in a sentence. Besides the above monolingual proposals, Sui and Yu (1998a) discussed a bilingual strategy to recognize predicate heads in Chinese

[1] Maximal noun phrase is the noun phrase which is not contained by any other noun phrases.

sentences with reference to those in their counterpart English sentences.

Nevertheless, these methods have their own limitations. The rule-based methods require effective linguistic rules to be formulated by linguists according to their own experience. Certainly, this is impossible to cover all linguistic situations concerned, due to the complexity of language and the limitations of human observation. In practice, we also should not underestimate the complexity of feature application, the computing power demanded and the difficulties in handing irregular sentence patterns. For instance, a sentence without subject may lead to an incorrect recognition of predicate head. For corpus-based approaches, they rely on language data in huge size but the available data may not be adequate. Those bilingual methods may first encounter the difficulty of determining correct sentence alignment in the case that the parallel data consist of much free translation.

Our method proposed here focuses on a simple but effective means to help identify predicate heads, i.e., MNP pre-processing. At present, there has some substantial progress in automatic recognition of MNP. Zhou *et al*. (2000) proposed an efficient algorithm for identifying Chinese MNPs by using their structure combination, achieving an 85% precision and an 82% recall. Dai *et al*. (2008) presented another method based on statistics and rules, reaching a 90% F-score on HIT Chinese Treebank. Jian *et al*. (2009) employed both left-right and right-left sequential labeling and developed a novel "fork position" based probabilistic algorithm to fuse bidirectional results, obtaining an 86% F-score on the Penn Chinese Treebank. Based on these previous works, we have developed an approach that first identifies the MNPs in a sentence, which are then used in determining the predicate heads in the next stage.

3 Task Description

The challenge of accurate identification of predicate heads is to resolve the problem of quasi-predicate heads in a sentence. On the one hand, the typical POSs of predicate heads in Chinese sentences are verbs, adjectives and descriptive words[2]. Each of them may have multiple instances in a sentence. On the other hand, while a simple sentence has only one predicate head, a complex sentence may have multiple ones. The latter constitutes 8.25% in our corpus. Thus, the real difficulty lies in how to recognize the true predicate head of a sentence among so many possibilities.

Take a simple sentence as example:

这/rN 种/qN 有/v 特大/a 翅膀/n 的/uJDE 大/a 鸟/n 没有/v 足够/aD 的/uJDE 支撑/v 力/n 和/cC 前进/v 力/n 。/wE

The quasi-predicate heads (verbs and adjectives) include 有/v, 特大/a, 大/a, 没有/v, 支撑/v, and 前进/v. However, there are two MNPs in this sentence, namely, "这/rN 种/qN 有/v 特大/a 翅膀/n 的/uJDE 大/a 鸟/n" and "足够/aD 的/uJDE 支撑/v 力/n 和/cC 前进/v 力/n". These two MNPs cover most quasi-predicate heads in the sentence, except 没有/v, the true predicate head that we want.

An MNP is a complete semantic unit, and its internal structure may include different kinds of constituents (Jian *et al*., 2009). Therefore, the fundamental structure of a sentence can be made clear after recognizing its MNPs. This can help filter out those wrong quasi-predicates for a better shortlist of good candidates for the true predicate head in a sentence.

In practice, the identification of predicate head begins with recognizing MNPs in the same sentence. It turns the above example sentence into:

[这/rN 种/qN 有/v 特大/a 翅膀/n 的/uJDE 大/a 鸟/n] 没有/v [足够/aD 的/uJDE 支撑/v 力/n 和/cC 前进/v 力/n] 。/wE

These MNPs are then replaced with the conventional label "NP" for noun phrase, resulting in a simplified sentence structure as follows.

NP/NP 没有/v NP/NP 。/wE

This basic sentence structure can largely alleviates the complexity of the original sentence and narrows down the selection scope of quasi-predicates for the true head. In this particular example, the only verb left in the sentence after MNP recognition is the true predicate head.

4 Predicate Head Identification

This section describes the process of identifying predicate heads in sentences. As illustrated in Figure 1 below, it can be divided into three steps:

Step 1: recognize the MNPs in a sentence and replace the MNPs with "NP" label to simplify the sentence structure.

Step 2: recognize the predicate heads in the resulted simplified structure.

[2] We only focus on Verbs and adjectives in this work.

Step 3: post-process the preliminary results to correct the wrong predicate heads according to heuristics in a rule base.

4.1 MNP Recognition

The MNP recognition is performed via a trained CRF model on unlabeled data. We adopt the method in Dai *et al*. (2008), with modified templates for the different corpus. Each feature is composed of the words and POS tags surrounding the current word i, as well as different combination of them. The context window of template is set to size 3. Table 1 shows the feature template we use.

Type	**Features**	
Unigram	$Word_i$	Pos_i
Bigram	$Word_i/Pos_i$	
Surrounding	$Word_{i-1}/Word_i$	Pos_{i-1}/Pos_i
	$Word_i/Word_{i+1}$	Pos_i/Pos_{i+1}
	$Word_{i-2}/Pos_{i-2}$	Pos_{i-2}/Pos_{i-1}
	$Pos_{i-2}/Pos_{i-1}/Pos_i$	Pos_{i-3}/Pos_{i-2}
	$Pos_{i-1}/Pos_i/Pos_{i+1}$	$Word_{i+3}/Pos_{i+3}$
	$Pos_{i+1}/Pos_{i+2}/Pos_{i+3}$	$Word_{i+2}/Word_{i+3}$

Table 1: Feature Template

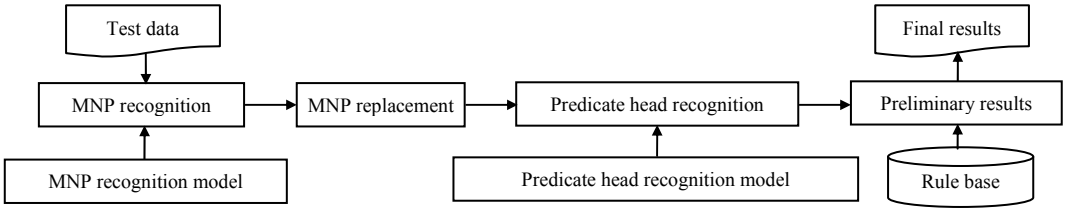

Figure 1: Flow Chart of Predicate Head Identification

The main effective factors for MNPs recognition are the lengths of MNPs and the complexity of sentence in question. We analyze the length distribution of MNPs in TCT[3] corpus, finding that their average length is 6.24 words and the longest length is 119 words. Table 2 presents this distribution in detail.

Length of MNP	**Occurrences**	**Percentage (%)**
len＜5	3260	48.82
5≤len＜10	2348	35.17
len≥10	1069	16.01

Table 2: Length Distribution of MNPs in TCT Corpus

The MNPs longer than 5 words cover 50% of total occurrences, indicating the relatively high complexity of sentences. We trained a CRF model using this data set, which achieves an F-score of 83.7% on MNP recognition.

4.2 Predicate Head Identification

After the MNPs in a sentence are recognized, they are replaced by "NP" label to rebuild a simplified sentence structure. It largely reduces the difficulty in identifying predicate heads from this simplified structure.

We evaluate our models by their precision in the test set, which is formulated as

$$Precision = \frac{right\_sentences}{Sum\_sentences} * 100\% \quad (1)$$

The *right_sentences* refer to the number of sentences whose predicate heads are successfully identified, and the *sum_sentences* to the total number of sentences in the test set. We count a sentence as *right_sentence* if and only if *all* its predicate heads are successfully identified, including those with multiple predicate heads.

For each predicate head, we need an appropriate feature representation *f (i, j)*. We test the model performance with different context window sizes of template. The results are shown in Table 3 as follows.

Template	**Context window size**	**Precision (%)**
Temp1	2	79.27
Temp2	3	82.59
Temp3	4	81.37

Table 3: Precisions of Predicate Heads Recognition under Different Context Window Sizes

It shows that the window size of 3 words gives the highest precision (82.59%). Therefore we apply this window size, together with other features in our CRF model, including words, POSs, phrase tags and their combinations. There are 24 template types in total.

4.3 Post-processing

The post-processing stage is intended to correct errors in the preliminary identification results of

[3] Tsinghua Chinese Treebank ver1.0.

predicate heads, by applying linguistic rules formulated heuristically. We test each rule to see if it improves the recognition accuracy, so as to retrieve a validated rule base. The labeling of predicate heads follows the standard of TCT and a wrong labeling is treated as an error.

There are three main types of error, according to our observation. The first is that no predicate head is identified. The second is that the whole sentence is recognized as an MNP, such that no predicate head is recognized. The third is that the predicate head is incorrectly identified, such as "是" in the expression "认为…是…", where the correct answer is "认为" according to the TCT standard.

Error types	Percentage	Improved percentage
No predicate head	17.50%	2.44%
a sentence as an MNP	10.63%	1.11%
"认为…是…"	8.75%	0.56%
Others	63.12%	2.77%

Table 4: Types of Error

Table 4 lists different types of error, together with their percentage in all sentences whose predicate heads have been mistakenly identified, and the improvement in percentage after the post-processing. To correct these errors, a number of rules for post-processing are formulated. The main rules are the followings:

- If no predicate head is recognized in a sentence, we label the first verb as the predicate head.

  Error sample：自/p［ 1 8 4 0/m 年/qT 鸦片战争/nR ］后/f ，/wP［ 中国/nS 逐步/d 沦为/v 半殖民地/b 半封建/b 社会/n ］。/wE

  Corrected：自/p［ 1 8 4 0/m 年/qT 鸦片战争/nR ］后/f ，/wP［ 中国/nS 逐步/d **沦为/v** 半殖民地/b 半封建/b 社会/n ］。/wE

- If the whole sentence is recognized as an MNP, such that no predicate head is identified, we label the first verb as the predicate head.

  Error sample：［ 针灸/n 包括/v 针/n 和/cC 灸/n 两/m 部分/n ］。/wE

  Corrected：［ 针灸/n **包括/v** 针/n 和/cC 灸/n 两/m 部分/n ］。/wE

- For expression "认为…是…", we label "认为" as the predicate head.

  Error sample：［ 另/rB 一/m 种/qN 观点/n ］认为/v 档案学/n 是/vC［ 兼/d 有/v 社会科学/n 和/cC 自然科学/n 性质/n 的/uJDE 综合性/b 科学/n ］。/wE

  Corrected：［ 另/rB 一/m 种/qN 观点/n ］**认为/v** 档案学/n 是/vC［ 兼/d 有/v 社会科学/n 和/cC 自然科学/n 性质/n 的/uJDE 综合性/b 科学/n ］。/wE

There are also other rules in the rule base besides the above ones. For example, if the first word of a sentences is "如" or "诸如", it is labeled as the predicate head.

5 Experiments

5.1 Data Sets

Our experiments are carried out on the Tsinghua Chinese Treebank (TCT). Every constituent of a sentence in TCT is labeled by human expert. We randomly extract 5000 sentences from TCT and remove those sentences that do not have predicate head. Finally, our data set contains 4613 sentences, in which 3711 sentences are randomly chosen as training data and 902 sentences as testing data. The average length of these sentences in training set is 20 words.

The number of quasi-predicate heads in a sentence is a critical factor to determine the performance of predicate head recognition. Reducing the number of quasi-predicate heads can improve the recognition precision. Table 5 shows the percentage of quasi-predicate heads in training data before and after MNP replacement.

Number of quasi-predicates	Percentage before MNP replacement(%)	Percentage after MNP replacement(%)
1	12.50	49.69
2	19.62	27.22
3	20.37	12.37
>3	47.51	10.72

Table 5: The Percentage of Quasi-predicate Heads Before and After MNP Replacement

From Table 5, we can see that almost half sentences contain more than three quasi-predicate heads. Only 12.5% of sentences have only one quasi-predicate head before MNP replacement. However, after MNPs are replaced with the "NP" label, only 10.72% contain more than three quasi-predicate heads and nearly 50% contain only one quasi-predicate head. We have evidence that MNP pre-processing can reduce the number

of quasi-predicate heads and lower the complexity of sentence structures.

5.2 Results and Discussion

For comparison purpose, we developed four different models for predicate head recognition. Models 1 and 2 are CRF models, the former recognizing predicate heads directly and the later recognizing MNPs at the same time. Model 3 recognizes predicate heads based on MNP preprocessing. Model 4 is based on model 3, including the post-processing stage. Table 6 shows the recognition performance of each model using the best context window size.

Model	Context window size	Number of correct sentences	Precision(%)
model 1	4	680	75.39
model 2	4	687	76.16
model 3	3	745	82.59
model 4	3	804	**89.14**

Table 6: Performance of Different Models

Comparing these models, we can see that the additional feature in model 2 leads to 1% improvement in precision over model 1. Moreover, the MNP pre-processing in model 3 results in a large increase in accuracy, compared to model 1. It indicates that the MNP pre-processing does improve the precision of recognition. Compared with model 3, model 4 achieves a precision even 6.55% higher, indicating that the post-processing is also an effective step for recognition.

As shown, the performance is affected by the effect of MNP recognition. There are three kinds of relation between the predicate heads and the types of MNP recognition error:

Relation 1: The whole sentence is recognized as an MNP.

Relation 2: The boundaries of an MNP are incorrectly recognized and the MNP does not contain the predicate head.

Relation 3: The boundaries of an MNP are incorrectly recognized and the MNP contains the predicate head. Table 7 shows the distribution of these three relations in the recognition errors.

Relation	Number of sentences	Percentage(%)
Relation 1	17	5.47
Relation 2	281	90.35
Relation 3	13	4.18

Table 7: Distribution of the Three Relations in Recognition Errors

In our approach, the errors of relation 1 and relation 3 can be solved by the post-processing, as presented in Section 4.3. Relation 2 holds the largest proportion among the three. But the error rate of predicate head recognition only reaches 31.67% in this case. That is to say, although the MNP boundaries are incorrectly recognized, the accuracy of predicate head recognition can still reach 68.33%.

Chen (2007) proposed a probabilistic model (model 5) for recognizing predicate heads in Chinese sentences. The probabilities of quasi-predicates are estimated by maximum likelihood estimation. A discounted model is used to smooth parameters. We compare his model with our model 3 using different contextual features on TCT corpus. Table 8 shows the comparison results.

The highest precision of model 3 is 82.59% when the context window size is set to 3. For model 5, it is 70.62% at a context window size of 4. Experimental results show that the precision of our method is about 12% higher than Chen's.

Context window size	Model	Precision (%)
2	model 5	69.18
	model 3	79.27
3	model 5	70.18
	model 3	**82.59**
4	model 5	**70.62**
	model 3	81.37

Table 8: Comparison between model 3 and Chen's model

Beside Chen's method, the Stanford Parser can also recognize the predicate heads in simple Chinese sentences. The root node of dependency tree is the predicate head. For a comparison, we randomly extract two hundred simple sentences in our test data to compare it with the outputs of our model 3. We also train a model of predicate head recognition (model 6), which assumes that all MNPs are successfully identified. The comparison is shown in Table 9. We can see that the precision of model 6 is 8.35% higher than model 3. This means that our method still has a certain room for further improvement.

Stanford Parser	model 3	model6
78.17%	83.15%	91.5%

Table 9: Comparison between model 3 and Stanford Parser

5.3 Error Analysis

As shown above, the post-processing can correct most errors in the recognition of predicate heads. But we also observe some errors that cannot be corrected this way. For example,

地理学/n以/p 描述性/n 记载/v [地理/n 知识/n] 为主/v 。/wE

The predicate head here is "为主", but usually "记载" is recognized as the predicate head. This is because "记载" can be used either as a verb or a noun. There are many verbs of this kind in Chinese, such as "主张" and "应用". Mistakes caused by the flexibility of Chinese verb and the ambiguity of sentence structure appear to deserve more of our effort. Meanwhile, there are also some other unusual cases that cannot be properly solved with statistical methods.

6 Conclusion

Identification of predicate heads is important to syntactic parsing. In this paper, we have presented a novel method that combines both statistical and rule-based approaches to identify predicate heads based on MNP pre-processing and rule-based post-processing. We have had a series of experiments to show that this method achieves a significant improvement over some state-of-the-art approaches. Furthermore, it also provides a simple structure of sentence that can be utilized for parsing.

In the future, we will study how semantic information can be applied to further improve the precision of MNP recognition and predicate head identification. It is also very interesting to explore how this approach can facilitate parsing, including shallow parsing.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments and suggestions. We also thank Billy Wong of City University of Hong Kong for his much-appreciated input during the writing process.

References

Zhiqun Chen. 2007. Study on recognizing predicate of Chinese sentences. *Computer Engineering and Applications,* 43(17): 176-178.

Yuchang Cheng, Asahara Masayuki, and Matsumoto Yuji. 2005. Chinese deterministic dependency analyzer: examining effects of global features and root node finder. *In Proceedings of the Fourth SIGHAN Wordshop on Chinese Language Processing,* pp. 17-24.

Cui Dai, Qiaoli Zhou, and Dongfeng Cai. 2008. Automatic recognition of Chinese maximal-length noun phrase based on statistics and rules. *Journal of Chinese Information Processing,* 22(6): 110-115.

Xiaojin Gong, Zhensheng Luo, and Weihua Luo. 2003. Recognizing the predicate head of Chinese sentences. *Journal of Chinese Information Processing,* 17(2): 7-13.

Ping Jian, and Chengqing Zong. 2009. A new approach to identifying Chinese maximal-length phrase using bidirectional labeling. *CAAI Transactions on Intelligent Systems,* 4(5): 406-413.

Guochen Li, and Jing Meng. 2005. A method of identifying the predicate head based on the correpondence between the subject and the predicate. *Journal of Chinese Information Processing,* 19(1): 1-7.

Zhensheng Luo, and Bixia Zheng. 1994. An approach to the automatic analysis and frequency statistics of Chinese sentence patterns. *Journal of Chinese Information Processing,* 8(2): 1-9.

Zhifang Sui, and Shiwen Yu. 1998a. The research on recognizing the predicate head of a Chinese simple sentence in EBMT. *Journal of Chinese Information Processing,* 12(4): 39-46.

Zhifang Sui, and Shiwen Yu. 1998b. The acquisition and application of the knowledge for recognizing the predicate head of a Chinese simple sentence. *Journal of Peking University (Science Edition),* 34(2-3): 221-229.

Honglin Sun, and Shiwen Yu. 2000. Shallow parsing: an overview. *Contemporary Linguistics,* 2(2): 74-83.

Hui Tan. 2000. Center predicate recognization for scientific article. *Journal of WuHan University (Natural Science Edition),* 46(3): 1-3.

Qiang Zhou, Maosong Sun, and Changning Huang. 2000. Automatically identify Chinese maximal noun phrase. *Journal of Software,* 11(2): 195-201.

Michael Collins. 1999. *Head-driven statistical models for natural language parsing.* Ph. D. Thesis, University of Pennsylvania.