CRF tagging for head recognition based on Stanford parser

Yong Cheng, Chengjie Sun, Bingquan Liu, Lei Lin
Harbin Institute of Technology
{ycheng, cjsun, linl,liubq}@insun.hit.edu.cn

Abstract

Chinese parsing has received more and more attention, and in this paper, we use toolkit to perform parsing on the data of Tsinghua Chinese Treebank (TCT) used in CIPS, and we use Conditional Random Fields (CRFs) to train specific model for the head recognition. At last, we compare different results on different POS results.

1 Introduction

In the past decade, Chinese parsing has received more and more attention, it is the core of Chinese information processing technology, and it is also the cornerstone for deep understanding of Chinese.

Parsing is to identify automatically syntactic units in the sentence and give the relationship between these units. It is based on a given grammar. The results of parsing are usually structured syntax tree. For example, the parsing result of sentence "中国是多民族国家" is as following.

```
(ROOT
  (dj (nS 中国)
    (vp (v 是)
      (np
        (np (m 多) (n 民族))
          (n 国家)))))
```

With the development of Chinese economy, Chinese information processing has become a worldwide hot spot, and parsing is an essential task. However, parsing is a recognized research problem, and it is so difficult to meet the urgent needs of industrial applications in accuracy, robustness, speed. So the study of Chinese grammar and syntax analysis algorithm are still the focus of Chinese information processing.

In all the parsing technology research, English parsing research is the most in-depth, and there are three main aspects of research in statistical parsing, they are parsing model, parsing algorithm, and corpus construction. As for the parsing model, currently there are four commonly used parsing models, PCFG model [1], the model based on historical, Hierarchical model of progressive, head-driven model [2].

Since parsing is mostly a data driven process, its performance is determined by the amount of data in a Treebank on which a parser is trained. Much more data for English than for any other languages have been available so far. Thus most researches on parsing are concentrated on English. It is unrealistic to directly apply any existing parser trained on an English Treebank for Chinese sentences. But the methodology is, without doubt, highly applicable. Even for those corpora with special format and information integrated some modification and enhancement on a well-performed parser to fit the special structure for the data could help to obtain a good performance.

This paper presents our solution for the shared Task 2 of CIPS2010-Chinese Parsing. We exploit an existing powerful parser, Stanford parser, which has showed its effectiveness on English, with necessary modifications for parsing Chinese for the shared task. Since the corpus used in CIPS is from TCT, and the sentence contains the head-word information, but for the Stanford parser, it can't recognize the head constituents. So we apply a sequence tagging method to label head constituents based on the data extracted from the TCT corpus, In section 2 and section 3, we will present the

Table 1. Training data with different formats

Parsing model	1.(ROOT (np-0-2 (n 货币学派) (cC 及其) (np-0-1 (n 政策) (n 主张)))) 2.(ROOT (vp-1 (pp-1 (p 对) (np-0-2 (np-1 (n 金融) (n 政策)) (cC 以及) (np-2 (a 类似) (uJDE 的) (np-1 (n 宏观) (np-1 (n 经济) (n 政策))))))) (vp-1 (d 必须) (vp-1 (d 重新) (v 估价)))))
POS model	1. 中国/nS 传统/a 医学/n 2.中国/nS 是/vC 多/a 民族/n 国家/n ，/wP 中华/nR 民族/n 是/vC ５０/m 多/m 个/qN 民族/n 的/uJDE 总称/n ◦/wE
Head-recognition model	a O n np 0 n a O np 1 nS O np np 0 np nS O np 1

details of our approach, and In section 4, we present the details of experiment.

2 Parsing

Since English parsing has made many achievements, so we investigated some statistical parsing models designed for English. There are three open source constituent parsers, Stanford parser [3], Berkeley parser [4] and Bikel's parser [5]. Bikel's parser is an implementation of Collins' head-driven statistical model [6]. The Stanford parser is based on the factored model described in [7]. Berkeley parser is based on unlexicalized parsing model, as described in [8].

All the three parsers are claimed to be multilingual parsers but only accept training data in UPenn Treebank format. To adapt these parsers to Tsinghua Chinese Treebank (TCT) used in CIP, we firstly transform the TCT training data into UPenn format. Then, some slight modifications have been made to the three parsers. So that they could fulfill the needs in our task.

In our work, we use Stanford parser to train our model by change the training data to three parts with different formats, one for training parsing model, one for training POS model, and the last for training head-recognition model. Table 1 shows the three different forms.

3 Head recognition

Head recognition is to find the head word in a clause, for example, 'np-1' express that in the clause, the word with index '1' is the key word.

To recognize the head constituents, and extra step is needed since Stanford parsing could not provide a straight forward way for this. Consider that head constituents are always determined by their syntactic symbol and their neighbors, whose order and relations strongly affects the head labeling. Like chunking [9], it is natural to apply a sequence labeling strategy to tackle this problem. We adopt the linear-chain CRF [10], one of the most successful sequence labeling framework so far, for the head recognition is this stage.

4 Experiment

4.1 Data

The training data is from Tsinghua Chinese Treebank (TCT), and our task is to perform full parsing on them. There are 37218 lines in official released training data, As the Table 1 show; we change the data into three parts for different models.

The testing data doesn’t contain POS labels, and there are 1000 lines in official released testing data.

Table 2. Different POS tagging results

	original	new
pos accuracy	80.40	94.82

4.2 Models training

4.2.1 Parsing model training

As for training parsing model with Stanford parser, since there are little parameters need to set, so we directly use the Stanford parser to train a model without any parameter setting.

4.2.2 POS model training

In this session of the evaluation, POS tagging is no longer as a separate task, so we have to train our own POS tagging model. In the evaluation process, we didn't fully consider the POS tagging results' impact on the overall results, so we didn't train the POS model specially, we directly use the POS function in Stanford parser toolkit. This has led to relatively poor results in POS tagging, and it also affects the overall parsing result. After the evaluation, we train a specific model to improve the POS tagging results. As the table 1 shows, we extract training data from the original corpus and adopt the linear-chain CRF to train a POS tagging model. Table 2 shows the original POS tagging results and new results.

4.2.3 Head recognition model training

As the table 1 shows, we extract specific training data from original corpus.

Table 3. Training data formats for Head-recognition

original corpus	1.[vp-0 减少/v [np-1 财政/n 收入/n]]
temp corpus	1.[np-1 财政/n 收入/n] 2.[vp-0 减少/v [np-1 财政/n 收入/n]]
final corpus	n O n np 0 n n O np 1 v O np vp 1 np v O vp 0

Table 4. Statistics the frequency of the words in each clause

number of word	statistics number
< 1	160
2	50834
3	12592
4	56
5	664
>5	360

And for head-word recognition, since the adjacent clause has little effect on the recognition of head-word, so we set the clause as the smallest unit. We chose CRF to train our model. However, for getting the proper format of data for training in CRF, We have to do further processing on the data. As the table 3 shows, the final data set word as the unit.

For example, the line 'n O np vp 1', the meaning from beginning to end is POS or clause mark of current word or clause, POS or clause mark of previous word, POS or clause mark of latter word, the clause mark of current word, and the last mean that if current word or clause is headword 1 represents YES, 0 represents NO.

4.4 Result and Conclusion

As we mention before, in evaluation, we didn't train specific POS tagging model, So we re-train our pos model, and the new results is shown in table 6, it can be seen that, with the increase of POS result, there is a corresponding increase in the overall results.

Table 5. Performance of head recognition and the template for model training

Boundary + Constituent	70.58
Boundary + Constituent + Head	66.97
template	U00:%x[0,0] U01:%x[-1,0] U02:%x[1,0] U04:%x[0,0]/%x[-1,0] U05:%x[0,0]/%x[1,0] U06:%x[-1,0]/%x[1,0]

Table 6. Overall results on different POS results

	POS	Boundary + Constituent
original	80.40	67.00
new	94.82	74.28

Through our evaluation results, we can see that it is not appropriate to directly use English parser toolkit to process Chinese. And it is urgent to development parsing model based on the characteristics of Chinese.

References

[1] T. L. Booth and R. A. Thompson. Applying Probability Measures to Abstract Languages. IEEE Transactions on Computers, 1973, C-22(5):422-450.

[2] M. Collins. Three Generative, Lexicalised Models for Statistical Parsing. In Proceedings of the 35th annual meeting of the association for computational linguistics.

[3] http://nlp.stanford.edu/software/lex-parser.html

[4] http://code.google.com/p/berkeleyparser

[5] http://www.cis.upenn.edu/~dbikel/download

[6] Michael Collins. 1999. Head-Driven Statistical Models for Natural Language Parsing. Ph.D. thesis. University of Pennsylvania.

[7] Dan Klein and Christopher D. Manning Accurate unlixicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics.

[8] S Petrov and D Klein. Improved inference for unlexicalized parsing. In Proceedings of NAACL HLT 2007.

[9] Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In Proceedings of HLT-NAACL 2003, pages 213-220, Edmonton. Canada.

[10] John Lafferty. Andrew McCallum. And Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML 2001, pages 282-289, Williams College, Williamstown, MA, USA.