Improving PCFG Chinese Parsing with Context-Dependent Probability Re-estimation

Yu-Ming Hsieh[1,2] **Ming-Hong Bai**[1,2] **Jason S. Chang**[2] **Keh-Jiann Chen**[1]
[1] Institute of Information Science, Academia Sinica, Taiwan
[2] Department of Computer Science, National Tsing-Hua University, Taiwan
morris@iis.sinica.edu.tw, mhbai@sinica.edu.tw,
jason.jschang@gmail.com, kchen@iis.sinica.edu.tw

Abstract

Selecting the best structure from several ambiguous structures produced by a syntactic parser is a challenging issue. The quality of the solution depends on the precision of the structure evaluation methods. In this paper, we propose a general model (context-dependent probability re-estimation model, CDM) to enhance the structure probabilities estimation. Compared with using rule probabilities only, the CDM has the advantage of an effective, flexible, and broader range of contexture-feature selection. We conduct experiments on the CDM parsing model by using Sinica Chinese Treebank. The results show that our proposed model significantly outperforms the baseline parser and the open source Berkeley statistical parser. More importantly, we demonstrate that the basic framework of the parsing model does not need to be changed, and the proposed re-estimation functions will adjust the probability estimation for every particular structure, and obtaining the better parsing results.

1 Introduction

Structure evaluation method is an important task in selecting the best structure from several ambiguous structures produced by a syntactic parser, particularly for Chinese. Since Chinese is an analytic language, words can play different grammatical functions without inflection. To implement a structure evaluation model, treebank is a necessary resource, since it provides useful statistical distributions regarding grammar rules, words, and part-of-speeches. Learning grammar rules and probabilities from treebanks is an effective way to improve parsing performance (Johnson, 1998). Unfortunately, sizes of treebanks are generally small; certain strategies of rule generalization and specialization have to be devised to improve the coverage and precision of the extracted grammar rules. However no matter how the grammar rules are refined, syntactic ambiguities are unavoidable. The ambiguous structures should be ranked according to their structural evaluation scores, which may be an accumulated score of rule probabilities and feature-based scores. In general, the evaluation functions are derived from very limited and biased resources, such as treebanks. Therefore we need to find a way to improve the evaluation functions under the constraint of very limited resources.

Suppose that the parsing environment is a model of probabilistic context-free grammar (PCFG). Several researchers are attaching many useful features to the grammar rules to improve the precision of the grammar rules (Johnson, 1998; Sun and Jurafsky, 2003; Klein and Manning, 2003; Hsieh et al., 2005). In this paper, we follow grammar representation in Hsieh et al. (2005), and propose a context-dependent probability re-estimation model (CDM) to enhance the performance of the original PCFG model. CDM combines rule probabilities and machine learning techniques in structure evaluation. Similar to other machine learning methods (Ratnaparkhi, 1999; Charniak, 2000; Wang et al., 2006), the CDM has the flexibility to adjust the features, and to obtain better re-estimated structure probabilities.

The remainder of this paper is organized as follows. Section 2 provides background on PCFG parsing with grammar rule representation. Section 3 describes the proposed CDM and our selected features. The experimental evaluation and results are in Section 4. The last section contains some concluding remarks.

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 216–221,
Tianjin, China, 20-21 DEC. 2012

2 Background

2.1 The baseline model, PCFG

PCFG-based parsing, a probabilistic context-free grammar parsing model that trains rule probabilities from treebank, is frequently used for parsing syntactic structures. Its parsing process is formulated as follows:

Given a sentence (*S*), a combination of words (*W*) and parts-of-speech (*POS*) sequences,

$$S = (W, POS) = (< w_1, ..., w_m >, < t_1, ..., t_m >),$$

a PCFG parser tries to find possible tree structures (*T*) of *S*. The parser then selects the best tree (T_{best}) according to the evaluation score of all possible trees:

$$T_{best} = \underset{T}{\operatorname{argmax}}\, Score(T, S)$$

Under the PCFG model, we divide a tree structure *T* into a set of sub-trees; that is, a set of grammar rules applied in *T*. If there are *n* context free grammar rules in a tree *T*, then:

$$Score(T, S) = \prod_{i=1 \in T}^{n} P(RHS_i \mid LHS_i)$$

Where *LHS* denotes the left-hand side of the grammar rule (e.g., non-terminal); *RHS* denotes the right-hand side of the grammar rule. To satisfy the probabilistic constraint, the following restriction is placed on the PCFG model:

$$\sum_{RHS \in R} P(RHS \mid LHS) = 1$$

We adopt logarithmic parsing probabilities in decoding; therefore, the cumulative product of probabilities *Score(T,S)* can be replaced by accumulation of logarithmic probabilities in formula 1.

$$\begin{aligned} Score(T, S) &= \sum_{i=1 \in T}^{n} \log(P(RHS_i \mid LHS_i)) \\ &= \sum_{i=1 \in T}^{n} RP_i \end{aligned} \quad (1)$$

where RP_i represents the logarithmic probabilities of the *i*-th grammar rule in the tree *T*.

2.2 F-PCFG - the feature-extended PCFG

We adopt a linguistically-motivated grammar generalization method (Hsieh et al., 2005) to obtain a binarized grammar, called F-PCFG, from original CFG rules extracted from treebank. The binarized F-PCFG grammars are produced by grammar generalization and grammar specialization processes. The grammar binarization process may produce generalized grammars with better coverage. However, such grammars may degrade the representational precision. Therefore, a grammar specialization process is needed to improve precision of the generalized grammars under the constraint of without much sacrificing grammar coverage.

A method of embedding useful features in phrasal categories is adopted. In the following we use an example shown in Figure 1 to illustrate the grammar generalization and specialization processes. See Hsieh et al. (2005) for details. In this tree structure, Nh is pronoun; VF is active Verb with VP object; VC is Active transitive verb; Na is Noun. For detail explanation of POS, please refer to CKIP (1993).

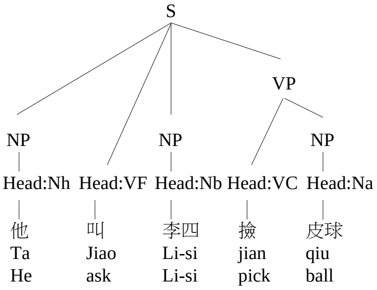

Figure 1. An example of a labeled syntactic tree structure in Treebank

Figure 2 shows the transformed tree representation by right-association binarization and feature embedding. We see that terminal nodes (i.e., $S_{\text{-NP-Head:VF}}$, $NP_{\text{-Head:Nh}}$) and intermediate nodes (i.e., $S'_{\text{-Head:VF-1}}$, $S'_{\text{-NP-0}}$, etc.). Both type of nodes attached the features of the left-most constituent of the RHS, phrasal category of parent-node, and existence of the phrasal head.

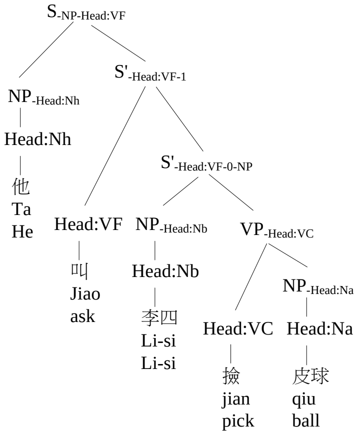

Figure 2. A transformed tree structure from original tree structure

We then use transformed binary trees to extract CFG and use maximum likelihood estimation to derive the rule probabilities from transformed Sinica Treebank (http://TreeBank.sinica.edu.tw).

3 Context-Dependent Probability Re-estimation Model

Many works try to improve rule probability estimation by using context-dependent probabilities in PCFG model, and show that rules with dependent context features perform better than PCFG alone (Ratnaparkhi, 1999; Charniak, 2000; Wang et al., 2006; Li et al., 2010). Charniak (2000) presented a maximum-entropy-inspired model to estimate probabilities in Markov grammar. The model uses a standard bottom-up best-first probabilistic chart parser to generate possible candidate parses in the first pass, and then evaluates the candidates with the proposed probabilistic model in the second pass. Therefore Charniak's method (2000) generates possible candidate parses first and then evaluates these candidates without early pruning. We adopt the maximum entropy method for structure evaluation, and integrate it into present PCFG model, called as CDM.

CDM integrates the original rule probabilities of PCFG and contextual probabilities as in the Formula 2:

$$Score(T,S)=\sum_{i=1\in T}^{n}\lambda\times RP_i+(1-\lambda)\times CDP_i\ , \qquad (2)$$

where CDP_i represents the logarithmic probabilityestimated according to the *i-th* rule and related lexical, grammatical and contextual features. We calculated CDP_i by using the maximum-entropy toolkit (Zhang, 2004). The advantage of using the maximum entropy model is that it hasthe flexibility to adjust features. To set a proper ratio for the probabilities estimated by the joint RP_i and CDP_i, we use the parameter λ in Formula 2. We use Collins' (1999) smoothing method during the estimation of the probabilities.

3.1 Feature Design

Feature selection is the most important step of any classifier and directly influences the parsing performance. Johnson (1998) observed that adding linguistic features (such as a parent node's category) improves accuracy of grammar rules; and Collins (1999) assessed the importance of head word and word bigram information in phrases. Sun and Jurafsky (2003) posited that the number of syllables in a word plays an important role in Chinese syntax. Hence, we try to include useful features for parsing Chinese. Suppose we need to calculate CDP_i based on the related features, while the *i-th* rule is applied for covering a span of words [L…R]. The used context and contextual features are as follows:

- **Lexical features** include word (*W*), parts-of-speech (*C*) and word sense (*V*) features. Our word sense feature uses the E-HowNet (will be discussed in Section 4) sense definition.
- **Grammar features**, which provide relevant information used in applying grammar rules, include features of the phrasal category of the LHS (*LHS Category*), the constituents of the right-hand-side of rule (*RHS*), and the attached features of the LHS (*LHS Feature*) in our F-PCFG.
- **Context features** include span words and immediately neighboring lexical units.

Table 1 shows the details of the feature templates. After feature selection phase, we train a CDM model by the maximum entropy method and apply it to re-estimate structure evaluation score in every parsing stage.

Feature template and description
The word *L*, *R* information. (LW_0, LC_0, LV_0, RW_0, RC_0, RV_0)
The LHS, RHS and features of each grammar rule. (*LHS Category*, *RHS*, *LHS Feature*)
The previous and next lexical unit of the word L,R (LW_{-1}, LC_{-1}, LW_1, LC_1, RW_{-1}, RC_{-1}, RW_1, RC_1)
The word bigram information of the RHS, including word， parts-of-speech and word sense combination. ($RhsW_1$&$RhsW_2$, $RhsC_1$&$RhsC_2$, $RhsV_1$&$RhsV_2$)
The combination of *L* or *R* with the previous lexical unit, or with the next lexical unit. (LW_{-1}&LW_0, LC_{-1}&LC_0, LW_0&LW_1, LC_0&LC_1, RW_0&RW_1, RC_0&RC_1, RW_{-1}&RW_0, RC_{-1}&RC_0)
The combination of *L* and *R*'s immediate neighboring lexical units (LW_0&RW_0, LC_0&RC_0, LW_{-1}&RW_1, LC_{-1}&RC_1)

Table 1. Feature templates for context-dependent estimation of partial tree structure while covering a span of words [L…R]

For instance, Figure 3 shows a partial parsing stage. We estimate the structure evaluation score P(S'$_{\text{-Head:VF+0+NP}}$ | features as shown in Table 1) for the non-terminal S'$_{\text{-Head:VF+0+NP}}$ which covers a span of words [李四 Li-si ... 皮球 ball] by the maximal entropy model. Some examples of contextual features are "*LW_0=李四, RW_0=皮球, LW_{-1}=叫, LW_1=撿, RW_{-1}=撿, RW_1=X, LW_{-1}&LW_0=叫&李四, LW_0&LW_1=李四&撿, RW_{-1}&RW_0=撿&皮球, RW_0&RW_1=皮球&X, $RhsW_1$=李四, $RhsW_2$=撿, $RhsC_1$=Nb, $RhsC_2$=VC, RHS=$NP_{-Head:Nb}$_$VP_{-Head:VC}$, …*", *etc*. Afterwards, we integrate and calculate the evaluation score by Formula 2.

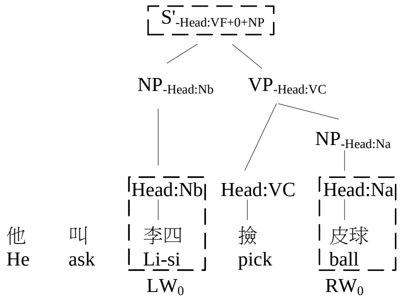

Figure 3. A partial tree of a parsing stage covered from "李四 Li-si" to "皮球 ball".

4 Experiments and Results

In this section, we describe the experiment design, and then evaluate the proposed models based on Sinica Treebank. We also analyze the results, and compare them with the results derived by the open source Berkeley statistical parser on the same test set.

4.1 Experimental Settings

Treebank: We employ Sinica Treebank as our experimental corpus. It contains 61,087 syntactic tree structures and 361,834 words. The syntactic theory of Sinica Treebank is based on the Head-Driven Principle (Huang et al., 2000); that is, a sentence or phrase is composed of a phrasal head and its arguments or adjuncts. We divide the treebank into four parts: the training data (55,888 sentences), the development set (1,068 sentences), the test data T06 (867 sentences), and the test data T07 (689 sentences). The test datasets (T06, T07) were used in CoNLL06 and CoNLL07 dependent parsing evaluation individually. The main difference between Sinica Treebank data and CoNLL data is that the CoNLL is in dependency format.

Word Sense: With regard to semantic features, we use the head senses of words expressed in E-HowNet (http://ehownet.iis.sinica.edu.tw/) as words' sense types. For example, the E-HowNet definition of 車輛 (Na), is {LandVehicle|車:quantity={mass|眾}}, and its head sense is "LandVehicle|車". For detailed description about E-HowNet, readers may refer to Huang et al. (2008).

Estimate Parsing Performance: To evaluate a model, we compare the parsing results with the gold standard. Black et al. (1991) proposed a structural evaluation system is called PARSEVAL. In all the experiments, we used the bracketed *f*-score (BF) as the parsing performance metric.

$$\text{Bracketed F-score (BF)} = \frac{\text{BP} * \text{BR} * 2}{\text{BP} + \text{BR}}$$

$$\text{Bracketed Precision (BP)} = \frac{\text{\# bracket correct consitituents in parser's parse of testing data}}{\text{\# bracket constituents in parser's of testing data}}$$

$$\text{Bracketed Recall (BR)} = \frac{\text{\# bracket correct consitituents in parser's parse of testing data}}{\text{\# bracket constituents in treebank's of testing data}}$$

For training *CDP* in CDM model, we extract relevant features from each parse tree in training data, in accordance with features setting in Table 1. Zhang (2004) provides a maximum entropy toolkit (MaxEnt) to help us training. We use option "-i 30 –gis –c 0" in MaxEnt training parameter. The training scale is 407 outcomes, 2438366 parameters and 1593985 predicates.

4.2 Results

Figure 4 shows the parsing performances on the developing data for different values of the parameter λ in Formula 2. The appropriate setting (λ=0.6) is learned and adopted for the future experiments.

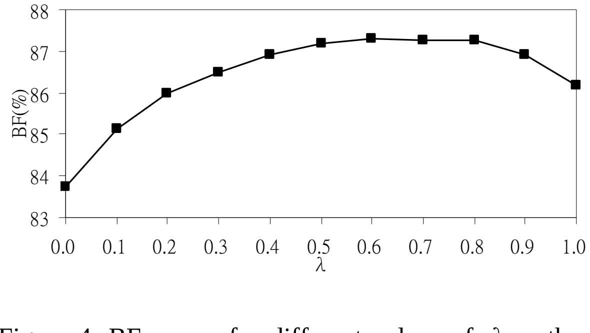

Figure 4. BF scores for different values of λ on the development data set

The results in Table 2 show that the integrated a general PCFG model with a CDM can improve the parsing performance. Implementing the integrated CDM on the T06 and T07 test datasets indicted improved the parsing performance by 1.45% and 1.53% respectively. The purpose in this research is to incorporate the rich contextual features to assist the constituent parsing. Results in Table 2 prove our method to be useful. As shown in the bracketed *f*-scores, about 20% of the errors are reduced. For instance, the ambiguous structures like "((Nh Nc) Nc)" and "(Nh (Nc Nc))" can be better resolved by our CDM model, since it can provide rich contextual features as additional information to help the parser making more precise evaluation scores in resolving ambiguous structures.

BF-Score (%)	T06	T07
PCFG	87.40	81.93
F-PCFG	88.56	83.96
CDM	90.01 (+1.45)	85.49 (+1.53)

Table 2. The bracketed *f*-score of the integrated CDM.

4.3 Comparison with the Berkeley Chinese parser

Berkeley parser[1] (Petrov et al., 2006) is used for comparison in our experiments because it appears to be the best PCFG parser for non-English languages. The parser has POS tagging and parsing functions; meanwhile, it takes word segmented data as input and outputs Penn Treebank style tree structures. We need to use pre-specified gold standard POS tags in our experiment, we transform our test data to "Berkeley CoNLL format" with word and POS. In addition, we need to transform our training data from Sinica Treebank style to Penn Treebank style (see Table 3) for Berkeley parser training model.

Tree style	Example
Sinica Treebank	S(NP(Head:Nh:他們)|Head:VC:散播 |NP(Head:Na:熱情))
Penn Treebank	((S (NP (Head:Nh (Nh 他們))) (Head:VC (VC 散播)) (NP (Head:Na (Na 熱情)))))

Table 3. Comparison of the Sinica and Penn Treebank styles

After re-training the Berkeley's parser with parameters, "*-treebank CHINESE –SMcycles 6 -useGoldPOS*", a new model is obtained. We parse the test dataset based on the gold standard word segmentation and POS tags. Then, we transform to Sinica Treebank style from the parsing results and evaluate by the same parsing performance metric. In our experiment, Berkeley's parser has best performance in using training model with 2th split-merge iterations. The bracketed *f*-score results of T06 and T07 test datasets are 88.58% and 83.56% respectively. The results of Berkeley's parser are closed to F-PCFG model in Table 2. Either Berkely's parser or F-PCFG represents the ceiling results of a general method, and they both outperform the naïve PCFG model.

4.4 Experiments for Task4 of CLP2012

Task 4 of CLP2012 includes two sub-tasks: sentence parsing and semantic role labeling task. For each sub-task, the testing data are complete Chinese sentence with gold standard word segmentation. Therefore, a pipeline process is needed to solve the POS tagging, syntactic parsing and semantic role assignment in our experiment. We adopt the context-rule tagger proposed by Tsai and Chen (2004) for the POS tagging. For syntactic parsing, we use the CDM parser with same training data in Section 4.1. For semantic role labeling, we follow You and Chen's (2004) method to assignment semantic role automatically. The detail parsing results of our systems on the test set can be found on the official evaluation report. Our system obtains acceptable results on both sentence parsing and semantic role labeling tasks.

F1-Score	Micro-Averaging	Macro-Averaging
Task 4-1	0.7287	0.7448
Task 4-2	0.6034	0.6249

Table 4. Official scores of sentence parsing (task4-1) and semantic role labeling (task4-2).

Table 4 shows the F1-score results are reported by the official organizer of the 2012 CIPS-SIGHAN bakeoff task. The result of the first sub-task (Task4-1) is about 0.7448. The POS tagging accuracy directly influences the sentential structure. Therefore, F1-score will be improved with better POS tagging accuracy. On the other hand, the result of the semantic role labeling (Task 4-2) is about 0.6249. Semantic role labeling is processed after sentence parsing. Our labeling system is based on different decision features, such as head-argument/modifier pairs, special cases, sentence structures, etc. These statistical information are extracted from training

[1] The version is "2009 1.1" and download from http://code.google.com/p/berkeleyparser/

data (see Section 4.1), and we use a backoff approach to decide the best semantic role. In future work, we will try using lexical semantic and context information to improve accuracy of semantic role labeling.

5 Conclusion

In this paper, we propose effective models to improve the performance of Chinese parsing. The models employ a broad range of features to integrate general statistical parsing and machine learning techniques to re-estimate structure score in module and incremental way. Our evaluations show that by adding CDM models, the parser outperforms the baseline PCFG model and an open source statistical parser.

We also consider a number of future research directions. In addition to the current treebank and lexical semantic information, more knowledge could be obtained from massive amounts of unlabeled data to make CDM more precise through auto-parsing and self-learning process. Our ultimate goal is to generate unlimited amounts of training data by parsing web corpus. As a result, we expect that the overall performance of our parser will be improved continually by the never ending self-learning process.

References

Adwait Ratnaparkhi. 1999. Learning to Parse Natural Language with Maximum Entropy Models. *Machine Language*, 34(1-3):151-175.

Chu-Ren Huang, Keh-Jiann Chen, Feng-Yi Chen, Keh-Jiann Chen, Zhao-Ming Gao and Kuang-Yu Chen. 2000. Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface. In *Proceedings of 2nd Chinese Language Processing Workshop*, pages 29-37.

CKIP. 1993. *Chinese Electronic Dictionary*. Technical Report, No. 93-05, Academia Sinica, Taiwan.

Dan Klein and Christopher Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the ACL 2003*, pages 423-430.

E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 306-311.

Eugene Charniak. 2000. A Maximum Entropy Inspired Parser. In *Proceedings of NAACL 2000*, pages 132-139.

Honglin Sun and Daniel Jurafsky. 2003. The effect of rhythm on structural disambiguation in Chinese. In *Proceedings of SIGHAN Workshop*.

Jia-Ming You, Keh-Jiann Chen. 2004. Automatic Semantic Role Assignment for a Tree Structure. In *Proceedings of SIGHAN Workshop*.

Junhji Li, Guodong Zhou, and Hwee Tou Ng. 2010. Joint Syntactic and Semantic Parsing of Chinese. In *Proceedings of ACL 2010*, pages 1108-1117.

Le Zhang. 2004. *Maximum Entropy Modeling Toolkit for Python and C++*. Reference Manual.

Mark Johnson. 1998. PCFG Models of Linguistics Tree Representations. *Computational Linguistics*, 24(4):613-632.

Mengqiu Wang, Kenji Sagae, and Teruko Mitamura. 2006. A Fast, Accurate Deterministic Parser for Chinese. In *Proceedings of COLING-ACL 2006*, pages 425-432.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Shu-Ling Huang, You-Shan Chung, and Keh-Jiann Chen. 2008. E-HowNet: the Expansion of HowNet. In *Proceedings of the First National HowNet Workshop*.

Slav Petrov, Leon Barrett, Romain Thibaux and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceesings of COLING/ACL 2006*.

Yu-Fang Tsai and Keh-Jiann Chen. 2004. Reliable and Cost-Effective Pos-Tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(1):83-96.

Yu-Ming Hsieh, Duen-Chi Yang, and Keh-Jiann Chen. 2005. Linguistically-Motivated Grammar Extraction, Generalization and Adaptation. In *Proceedings of IJCNLP 2005*, pages 177-187.