Improving Chinese Parsing with Special-Case Probability Re-estimation

Yu-Ming Hsieh[1,2], Su-Chu Lin[1], Jason S. Chang[2], and Keh-Jiann Chen[1]
[1] Institute of Information Science
Academia Sinica, Taipei, Taiwan
e-mail: {morris, jess, kchen}@iis.sinica.edu.tw
[2] Department of Computer Science,
National Tsing-Hua University, Hsinchu, Taiwan
e-mail: jason.jschang@gmail.com

***Abstract*—Syntactic patterns which are hard to be expressed by binary dependent relations need special treatments, since structure evaluations of such constructions are different from general parsing framework. Moreover, these different syntactic patterns (special cases) should be handled with distinct estimated model other than the general one. In this paper, we present a special-case probability re-estimation model (SCM), integrating the general model with an adoptable estimated model in special cases. The SCM model can estimate evaluation scores in specific syntactic constructions more accurately, and is able for adopting different features in different cases. Experiment results show that our proposed model has better performance than the state-of-the-art parser in Chinese.**

Keywords-Syntactic Parsing; PCFG; Chinese Parser; Structural Evaluation.

I. INTRODUCTION

Many syntactic constructions exhibit idiosyncratic syntactic behaviors so they are hard to be parsed with general models. These constructions should be handled with different parsing models. Previous researches separate well-known constructions from general sentences for further procedure show better performance than those which only use one general model [1], [5], [11], [13], [15], [16]. [13] proposed a divide-and-conquer approach to improve the precision of the conjunctive boundary detection. [1] showed that semantic classes help to obtain significant improvement in both parsing and PP attachment tasks. [16] analyzed Maximal Noun Phrase (MNPs) on pre-processing stage and brought MNP structures and each probability to full parsing. With a web-scale corpus, [11] produced various bracketed structures of the input sentences containing NP and selected the best candidate.

We have observed some common parsing errors caused by the general PCFG model, including 的-DE construction, conjunction, prepositional phrases, postpositional phrases, and two-part fixed expression (i.e., D-Case and Cb-case), as shown in Table I. Those parsing errors are usually constructions with long distance dependency structures and co-occurring words (i.e., collocation). In the sentence "他 比 那些 評論家 還 要 更 謙遜"('He is more humble than those critics'), "比…還"('more …than…') is a long distance dependency and the words co-occur with each other. General models, such as Collin's parser [4], focus on the head word and statistics on its right and left elements. As a result, the non-head word "還" in the previous sentence is not handled effectively.

In this paper we propose a special-case probability re-estimation model (SCM) to resolve these problems. SCM can calculate more accurate probabilities in parsing stage with special case probabilities learned from treebank. Among common parsing errors in Table I, we find that P-Case, D-Case and Cb-Case have explicit structure construction meanings with co-occurring words (x, y). Therefore we aim to solve these construction cases in our experiment. Learning probabilities of each case is straightforward. First we use (x, y) pairs to anchor the construction pattern for each case. Then we find out possible bracketed information (i.e., cases corresponding to the phrasal scope or grammar rules) and calculate probabilities for each case in treebank automatically. Finally we incorporate the case specific probabilities in parsing model. Experiments show that our proposed model has better performance than the state-of-the-art parser in Chinese. More importantly we demonstrate that the basic framework of the parsing model does not need to be changed.

TABLE I. COMMON PARSING ERRORS.

Type	Description
DE-Case	的-DE construction: errors occur on the phrase boundaries of the pre-argument of DE.
Ng-Case	Postposition Ng: errors occur on the boundaries of the pre-argument of Ng.
P-Case	Prepositional phrases: errors occur on the phrase boundaries of the argument of P.
D-Case	Construction with two adverb parts
Caa-Case	Conjunction: errors occur on the left and right constituents of conjunction Caa.
Cb-Case	Construction with two correlative construction

The remainder of this paper is organized as follows: Section 2 provides a construction method on special case; Section 3 presents our proposed model in Chinese texts;

Section 4 presents the experimental results; finally, Section 5 is the conclusion and future works.

II. Extracting and Learning Special-Cases

Learning more knowledge from large-scale corpus is a research trend [3], [9]. However a parser in general model has poor performance in special-case structure, no matter how huge the data is. Therefore we aim to handle these special-case structures from treebank automatically. But due to the data scarcity, treebank is incapable of expressing realistic co-occurring words. As a result, we try to calculate word relativity from a large corpus and filter out the irrelevant. We use bigram of (x, y) to anchor the constructions with each PoS tag. Table II shows the syntactic constructions and examples of the target structure, which are consisted with the PoS tag of P, D, and Cb.

TABLE II. ERROR CONSTRUCTION TYPES AND EXAMPLES OF WORD PAIRS IN L_{SET}.

Type	Constraints on (x,y) Pairs and Examples

P-Case[a]	x's PoS is a preposition (P) and y's PoS is not restricted
	(在:P, 下:Ng) / under; (朝:P, 邁進:VCL) / toward;
	(如同:P, 一般:Ng) / be same as;
	(比:P, 還:Dfa) / more..than … etc.

D-Case	x's PoS is an adverb (D) and y's PoS is an adverb (D) (越:D, 越:D) / the more..; (忽:D, 忽:D) / and; (愈:D, 愈:D); (一邊:D, 一邊 D) / while …etc.
Cb-Case	x's PoS is a correlative conjunction (Cb) and y's PoS is not restricted (連:Cb, 都:D) / even; (既:Cb, 又:Ca) / both..and..;(凡:Cb, 皆:D) / all … etc.

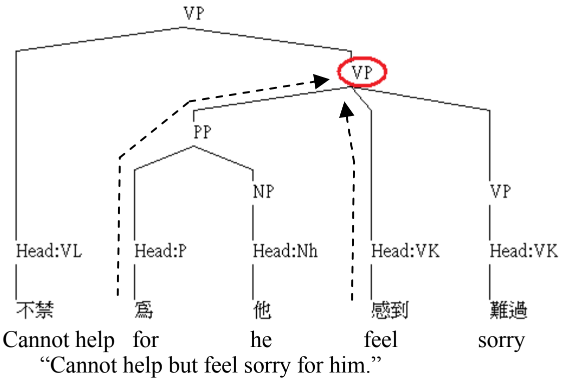
a. P is preposition; D is an adverb; Cb is correlative conjunction; Dfa is a degree adverb. For detail explanation of PoS tag, please refer [7].

Firstly we find the constrained bigram of (x, y) by using the pairwise mutual information [6] between all pairs of words in the sentence. We take the bigram (or collocation) within the distance m from a large corpus; then we calculate all PMI (1) and filter with a threshold.

$$\mathrm{PMI}(x,y)=\log\frac{p(\mathrm{x},\mathrm{y})}{p(\mathrm{x})p(\mathrm{y})} \tag{1}$$

Consequently we collect a set, L_{set} , of word pairs that need further processing, such as (在:P, 下:Ng), (比:P, 還:Dfa) in 'P-Case'; (愈:D, 愈:D), (一邊:D, 一邊 D) in 'D-Case'. Other detail instances are shown in Table II.

Secondly we obtain related bracketed types and features of each word pair in Lset extracted from treebank, and we train a bracketed classifier for these cases (e.g. P-Case, D-Case, and Cb-Case). For example, Figure 1 shows a sample tree in focus, (為/for, 感到/feel). We find the common ancestor node "VP(PP(Head:P:為…)|Head:VK:感到…)", a partial structure, which covers the nodes "為/for" and "感到/feel". Next we convert the partial structure into the bracketed type "(為:P …) 感到:VK". Finally we train a bracketed classifier for each case based on selected features and bracketed type of specific cases from the training data.



"Cannot help but feel sorry for him."

VP(Head:VL: 不 禁 |VP(PP(Head:P: 為 |NP(Head:Nh: 他))|Head:VK:感到|VP(Head:VK:難過)))[1].

Figure 1. An example of finding a common ancestor node in a tree.

A. Features Design

Suppose that we want to build a P-Case classifier to decide the bracketed type, the feature set we need is shown in Table III. Symbols of L and R denote the candidates x and y respectively (i.e., L means "為:P / for" and R means "感到:VK / feel"). For the unigram feature, the symbol LW_0 indicates that the word of x is "為/for," and its left and right neighboring Words and PoSs in this training sample are "LW_{-1}/LC_{-1}=不禁/VL" and "LW_1/LC_1=他/Nh" respectively. Similarly, the symbol RW_0 indicates that the word of y is "感到/feel," and its left and right neighboring Words and PoSs are "RW_{-1}/RC_{-1}= 他 /Nh" and "RW_1/RC_1= 難 過 /VK" respectively. For the bigram feature, we choose a combination of neighboring Words and PoSs (i.e., "LW_{-1}/LW_0= 不禁/為" and "LC_{-1}/LC_0=VL/P"). The rest of the features in Table III are derived similarly.

TABLE III. FEATURE TEMPLATES FOR BRACKETED TYPE CLASSIFIER.

Type	Feature Templates
unigram feature	LW_{-1}, $\boldsymbol{LW_0}$, LW_1, LC_{-1}, $\boldsymbol{LC_0}$, LC_1, RW_{-1}, $\boldsymbol{RW_0}$, RW_1, RC_{-1}, $\boldsymbol{RC_0}$, RC_1
bigram feature	$LW_{-1}/\boldsymbol{LW_0}$, $LC_{-1}/\boldsymbol{LC_0}$, $\boldsymbol{LW_0}/LW_1$, $\boldsymbol{LC_0}/LC_1$, $RW_{-1}/\boldsymbol{RW_0}$, $RC_{-1}/\boldsymbol{RC_0}$, $\boldsymbol{RW_0}/RW_1$, $\boldsymbol{RC_0}/RC_1$, LW_{-1}/RW_1, LC_{-1}/RC_1, $\boldsymbol{LW_0}/\boldsymbol{RW_0}$, $\boldsymbol{LC_0}/\boldsymbol{RC_0}$

III. The Special-Case Probability Re-estimation Model, SCM

After constructing classifiers, we propose a special-case probability re-estimation model (SCM), integrating special-

[1] The PoS of VL is stative verb with predicative object; Nh is pronoun; VK is stative verb with sentential object.

case classifier into general parsing model. The SCM tries to obtain more accurate probabilities by building classifiers for specific syntactic constructions. The proposed model has the advantage that the probabilities produced by a general parsing model can be adjusted according to SCM in every parsing stage, if and only if the (x, y) pair is a related construction pattern. Another advantage is that the model is incremental, easy to add other case and train classifiers. As a result, the pruning process and structure evaluation is based on the more reliable adjusted scores produced by aggregating the scores of the SCM.

A. Parsing with SCM

We adopt a Chinese PCFG Parser [8] in our experiment to enhance the structure probabilities estimation. Compared with using rule probabilities only, the CKIP Parser has the advantage of an effective, flexible, and broader range of contexture-feature selection. Results show that their model significantly outperforms the open source Berkeley statistical parser [12]. And the parser achieves the best score in Traditional Chinese Parsing task of SIGHAN Bake-offs 2012. The detail description of bake-off tasks is on the web site (http://www.cipsc.org.cn/clp2012/).

It is well-known that a PCFG parser tries to find possible tree structures (*T*) of a given sentence (*S*). The parser then selects the best tree according to the evaluation score *Score(T,S)* of all possible trees. If there are *n* context free grammar rules in a tree *T*, the *Score(T,S)* is the accumulation of logarithmic probabilities of the *i*-th grammar rule (RP_i). Equation (2) shows the baseline PCFG parser. Equation (3) is the proposed CKIP parser [8] evaluation model, joining RP_i and CDP_i with the parameter λ. CDP_i represents the logarithmic probability estimated according to the lexical, grammatical and contextual features in parsing *i*-th stage. To differentiate from general model and to focus on special issue, we modify the structural evaluation function in (3) by adding SCP_i to generate a new score, as shown in (4).

$$Score(T,S) = \sum_{i=1}^{n} RP_i \quad (2)$$

$$Score(T,S) = \sum_{i=1}^{n} (\lambda \times RP_i + (1-\lambda) \times CDP_i) \qquad (3)$$

$$Score(T,S) = \sum_{i=1}^{n} \rho \times (\lambda \times RP_i + (1-\lambda) \times CDP_i) + (1-\rho) \times SCP_i \qquad (4)$$

We calculated SCP_i by using the maximum entropy toolkit [14] and default training parameters. When the related SCM classifier is triggered, the value of SCP_i is calculated according to the lowest-level non-terminal node in *i*-th rule (applied in chart *i*) covering a span of words (x, y); otherwise it is not processed (i.e., the value of SCP_i is 0).

$$SCP_i = \begin{cases} \mathrm{SCP_i} & \text{if chart}_\mathrm{i} \text{ is covering the case (x, y)} \\ 0 & \text{if chart}_\mathrm{i} \text{ is not covering the case (x, y)} \end{cases}$$

As in Figure 1, the VP grammar rule of chart *i* in P-case is covering a span of words from “為/for” to “感到/feel” and is the lowest-level non-terminal node. We calculate the SCP_i probabilities according to the bracketed type “(為:P …) 感到:VK” and the feature set in P-case classifier. In the parsing stage, we only calculate each ambiguous partial tree once and the calculation will not be duplicated, which is a fair adjustment. In addition, the SCM will not affect the present model if the related construction pattern (x, y) does not occur.

IV. EXPERIMENTS AND RESULTS

A. Experimental Settings

Treebank: We use the same dataset in [8], which divides the data from Sinica Treebank [10] into four parts: training data (55,888 sentences), development set (1,068 sentences), test data T06 (867 sentences), and test data T07 (689 sentences).

Large Corpus: We employ the ten-million-word Sinica Corpus[2], a balanced modern Chinese Corpus with PoS tag. In order to search for high frequency and relevant collocation bigram, Chinese word bigram frequencies are obtained from this corpus.

Estimate Parsing Performance: In evaluation, we compare the parsing results with the gold standard. [2] proposed a structural evaluation system called PARSEVAL. Throughout the experiment, we use bracketed *f*-score (BF) from PARSEVAL as the parsing performance metric.

B. Results

The value of the parameter λ in (3) is 0.6[3]. We use gold-standard segmentation and PoS tags in our parsing experiments. The results in Table IV show that CKIP parser has better parsing performance than PCFG parser [8], which increases 2.61% and 3.56% in BF-score in the T06 and T07 respectively.

TABLE IV. THE BRACKETED *F*-SCORE OF BASELINE PCFGPARSER AND CKIP PARSER.

BF-Score (%)	T06	T07
PCFG	87.40	81.93
CKIP Parser	90.01 (+2.61)	85.49 (+3.56)

To construct a SCM, we first obtain the average sentence length to help us determine the distance of bi-gram collocations (x, y). As shown in Table V, the average sentence length in treebank is about 6 words, so the distance *m* of our bigram (x, y) is restricted to the range from 2 to 6.

[2] http://app.sinica.edu.tw/kiwi/mkiwi/

[3] The value of weight is suggested from [8].

TABLE V. THE SENTENCE LENGTH DISTRIBUTION OF THE EXPERIMENTAL DATA SET.

Sentence Numbers	Train	Dev	T06	T07
$2 \le len \le 5$	25,496	486	401	238
$6 \le len \le 10$	21,806	430	343	324
$11 \le len$	5,206	121	121	127
average length	6.23	6.25	5.79	7.49

Then we filter out 2,746 instances (x, y) with distance *m* is between 2 and 6 and PMI(x,y)>1.0 for the specific cases in L_{set}. We train a classifier to re-estimate probability and use $\rho = 0.5$ to integrate SCP_i into parser in (4). This means that the SCM and the original model have the same weight. Table VI shows the evaluation results on the test dataset. Compared with CKIP Parser, implementing SCM on T06 and T07 test dataset improves the parsing performance by 0.60% and 0.44% respectively. Results show that our proposed model is feasible and it obtains better performance than PCFG and CKIP Parser.

TABLE VI. PARSING ACCURACY OF DIFFERENT PARSING MODELS ON THE TEST DATASET.

BF-Score (%)	T06	T07
PCFG	87.40	81.93
CKIP Parser	90.01	85.49
SCM	90.61 (+0.60)	85.93 (+0.44)

If we focus on sentences containing specific cases, there are 105 sentences in T06 and 104 sentences in T07. Results of the partial sentences in Table VII indicate that SCM improves parsing performance by 1.60% and 1.10% on T06 and T07 test dataset respectively.

TABLE VII. THE EVALUATION RESULTS ON INCLUDING THE SPECIFIC CASE SENTENCES ONLY.

BF-Score (%)	T06	T07
CKIP Parser	85.39	86.00
SCM	86.99 (+1.60)	87.10 (+1.10)

We use Berkeley parser [12], the best PCFG parser for non-English language, to observe the problem of special-case structures in general parser. From our experiments, these cases cannot be handled well by Berkley parser either. For example, the D-Case as (一:D, 一:D); P-Cases as (在:P, 內:Ncd), (在:P, 之外:Ng), (和:P, 不同:VH); Cb-Cases as (連:Cb, 都:D), (但:Cb, 可能:D) and etc. Therefore, we believe the special-case problem is a common problem in general model parser.

V. CONCLUSION AND FUTURE WORKS

In this paper we propose effective models for Chinese parsing on special-case processing. Experimental results show the proposed model improved parsing accuracy.

We are considering a number of future research avenues. In order to expand SCM model and make it more robust, we will deal with more error-prone cases. Moreover we will modify the representation of special-case pattern to obtain better accuracy. We will further enhance selection method to extract and filter more specific and useful co-occurring words on large corpus. As a result, we expect that the overall performance of our parser will improve continually.

REFERENCES

[1] E. Agirre, T. Baldwin, and D. Martinez, “Improving Parsing and PP attachment Performance with Sense Information,” in *Proceedings of ACL-08: HLT.* Association for Computational Linguistics, 2008, pp. 317-325.

[2] 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, “A procedure for quantitatively comparing the syntactic coverage of English grammars,” in *Proceedings of the workshop on Speech and Natural Language,* 1991, pp. 306-311.

[3] W. Chen, J. Kazama, K. Uchimoto, and K. Torisaw, ”Improving Dependency Parsing with Subtrees from Auto-Parsed Data,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2009, pp. 570-579.

[4] M. Collins, “Head-Driven Statistical Models for Natural Language parsing,” Ph.D. thesis, University of Pennsylvania, 1999.

[5] E. Charniak and M. Johnson, “Coarse-to-fine n-best Parsing and MaxEnt Discriminative Reranking,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL),* 2005, pp. 173-180.

[6] K. Ward Church and P. Hanks, “Word Association Norms, Mutual Information, and Lexicography,” *Journal of Computational Linguistics*, vol. 16, no. 1, March 1990, pp. 22-29.

[7] CKIP. 1993. Chinese Electronic Dictionary. *Technical Report,* no. 93-05, Academia Sinica, Taiwan.

[8] Y.-M. Hsieh, M.-H. Bai, J.-S. Chang and K.-J. Chen, “Improving PCFG Chinese Parsing with Context-Dependent Probability Re-estimation,” in *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing,* 2012, pp. 216–221.

[9] Y.-M. Hsieh, D.-Chi Yang and K.-J. Chen, “Improve Parsing Performance by Self-Learning,” *Computational Linguistics and Chinese Language Processing,* vol.19, no.2, pp.195-216, June 2007.

[10] C.-R. Huang, K.-J. Chen, F.-Y. Chen, Z.-M. Gao and K.-Y. Chen, “Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface,” in *Proceedings of 2nd Chinese Language Processing Workshop*, 2000, pp. 29-37.

[11] E. Pitler, S. Bergsma, D. Lin, and K. Church, “Using Web-scale N-grams to Improve Base NP Parsing Performance,” In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING),* 2010, pp. 886-894.

[12] S. Petrov, L. Barrett, R. Thibaux and D. Klein, “Learning Accurate, Compact, and In-terpretable Tree Annotation,” in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006, pp. 433-440.

[13] D.-C. Yang, Y.-M. Hsieh, and K.-J. Chen, “Resolving Ambiguities of Chinese Conjunctive Structures by Divide-and-Conquer Approaches,” in *Proceedings of the third International Joint Conference on Natural Language Processing (IJCNLP),* 2008, pp. 715-720.

[14] L. Zhang. 2004. Maximum Entropy Modeling Toolkit for Python and C++. *Reference Manual.*

[15] Y. Zhang and S. Clark, “Transition-Based Parsing of the Chinese Treebank using a Global Discriminative Model,” in *Proceedings of the 11th International Conference on Parsing Technologies,* Association for Computational Linguistics, 2009, pp. 162-171.

[16] Q. Zhou, X. Liu, X. Ren, W. Lang, and D. Cai, “Statistical Parsing Based on Maximal Noun Phrase Pre-processing,” in *Proceedings of CIPS-ParsEval-2009.* 2009, pp. 1-7.