GLR PARSER WITH CONDITIONAL ACTION MODEL USING SURFACE PHRASAL TYPES FOR KOREAN

Yong-Jae Kwak, So-Young Park, and Hae-Chang Rim
NLP Lab., Dept. of CSE, Korea University, Seoul, Korea
{yjkwak, ssoya, rim}@nlp.korea.ac.kr

Abstract

In this paper, we propose a new probabilistic GLR parsing method that can solve the problems of conventional methods. Our proposed Conditional Action Model uses Surface Phrasal Types (SPTs) encoding the *functional word sequences* of the sub-trees for describing structural characteristics of the partial parse. And, the proposed GLR model outperforms the previous methods by about 6~8%.

1 Introduction

Since the first approach [Wright and Wrigley 1991]of combining a probabilistic method into the GLR technique was published, Some probabilistic GLR parsers also have been implemented in which probabilities are assigned to actions of LR parsing tables by using lookaheads or LR states as simple context information of [Briscoe and Carroll 1993], [Kentaro et al. 1998], and [Ruland, 2000] which does not use the stack information of the GLR parser effectively, because of highly complex internal GLR stack. As a result, they have used relatively limited contextual information for disambiguation. [Kwak et al., 2001] have proposed a conditional action model that uses the partially constructed parse represented by the graph-structured stack as the additional context. However, this method inappropriately defined sub-tree structure. Our proposed model uses Surface Phrasal Types representing the structural characteristics of the sub-trees for its additional contextual information.

2 Conditional Action Model(CAM) using Surface Phrasal Type (SPT)

CAM is devised based on the hypothesis that this model can actively use rich information provided by the partially constructed parse built on the graph-structured stack, and thus estimate the probability of the shift/reduce actions more precisely [Kwak et al., 2001].

Surface Phrasal Type (SPT) is represented by a sequence of the primitive mnemonics which describes the specific types of phrases based on their terminal nodes. In this work, we use *functional words* for mnemonics in SPT. In Korean, the functional word system is highly developed in the morpheme level. Therefore, this kind of phrasal description is meaningful way of representing the parse structure without considering the internal relation of the parse forest. Moreover, this scheme can avoid the overhead of taking care of the packed node with a local ambiguity. We represent SPTs as the corresponding mnemonic sequence(in backward order) as shown in Figure 1. We have defined mnemonic sets of SPT combination for the production of the noun phases and verb phrases, respectively. Example mnemonic sets for the both production forms are shown in Table 1.Elements in each mnemonic set consist of representatives of part-of-speeches (POSs) with the same syntactic (structural) function.

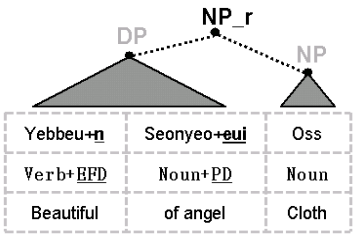

Figure 1. representations of SPT. Functional words are underlined.

For probabilistic model, we define the entire parse of the given input sentence as the sequence of actions taken until the parser reaches the accept state. Thus, the probability of the *i*-th action and parse probability are calculated by the following formula:

$$P(action_i) = \begin{cases} P(a_i \mid TY_0,...TY_{n-1}, s_{i-1}, l_i) & (\text{NP \& VP case}) \\ P(a_i \mid s_{i-1}, l_i) & (\text{other case}) \end{cases}, \quad P(T \mid S) = \prod_{i=1} P(action_i)$$

Here, *a* stands for either shift or reduce action. $TY_0,\ldots,TY_{n-1}$ indicates a sequence of sub-SPTs for the

sub-trees along the reduce route. s_{i-1} indicates the number of the state nodes at the top of the stack, and l_i is the lookahead symbol (POS) read by the parser., and a_i represents the i-th action. Then, the probability of a parse tree can be calculated by the product of all action probabilities. To cope with the sparse data problem when using our probabilistic model, we use a deleted interpolation method with the backing-off strategy similar to [Collins, 1999].

3 Experimental Results

We have experimented on the Korean treebank which consists of 12,084 sentences tagged with Korean grammar scheme with 56 CFG rules of [Park et al. 1999]. The distribution of sentence length over the corpus is shown in Table 3. We have used 10,906 sentences for the training data and 1,178 sentences for the test data. Average morpheme length is 22.5. For CAM, because of the sparse data problem, we have restricted the maximum continuous repetition count of the same mnemonic and the maximum length of one SPT to 1 and 3, respectively (empirically optimal value) Our GLR parser uses the canonical SLR(1) parsing table constructed from the binary CFG entries provided by the CFG grammar.

Table 1: SPT mnemonic codes (partial) for NP

code	property of the produced NP	syntactic structure	POS
ED	modified by clause	verb+**ED** +noun	EFD
EN	transformed by ending	verb+**EN**	EFN
PD	genitive noun	noun+**PD**+noun	PD
…			

Table 2: Parsing Accuracy

%	B&C 1993	Kentaro 1998	Kwak 2001	Proposed Model
L R	72.02	74.29	77.23	**83.64**
	71.22	74.27	76.01	**82.18**
E M	2.13	3.81	6.99	**12.94**
	1.70	3.77	6.04	**10.36**

As shown in the experimental results of Table 2, our proposed model outperforms previous models by about 6~8 %(Upper and lower parts show the results for training data and test data, respectively). Furthermore, the performance of our parser could be improved if it is integrated with the properly lexicalized information. The results show that functional category is an effective way of describing structural aspects of a phrase and can be used as contextual information in GLR parsing.

References

[Briscoe and Carroll 1993] Ted Briscoe and John Carroll. 1993. Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars. In *Computational Linguistics,*19(1):pages 25-59.

[Collins 1999] Michael Collins. 1999. *Head-Driven Models for Natural Language Parsing.* PH.D thesis, Dept. of Computer and Information Science, University of Pennsylvania.

[Kentaro et al. 1998] Inui Kentaro, Virach Sornlertlamvanich, Tanaka Hozumi and Tokunaga Takenobu. 1998. Probabilistic GLR parsing: a new formalization and its impact on parsing performance. In *Journal of Natural Language Processing*, Vol. 5, No. 3, pages 33-52.

[Kwak et al. 2001] Yong-Jae Kwak. So-Young Park, Hoojung Chung, Young-Sook Hwang, Sang-Zoo Lee, Hae-Chang Rim, 2001. GLR Parser with Conditional Action Model (CAM). In *proceedings of 6th Natural Language Processing Pacific Rim Symposium*, pages 359-366.

[Park et al. 1999] So-Young Park, Young-Sook Hwang, Hoojung Chung, Yong-Jae Kwak, and Hae-Chang Rim. 1999. A Feature-based Grammar for Korean Parsing. In *proceedings of 5th Natural Language Processing Pacific Rim Symposium*, pages 167-171

[Ruland 2000] Tobias Ruland, 2000. A Context-Sensitive Model for Probabilistic LR Parsing of Spoken Language with Transformation-Based Postprocessing. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 677-683.

[Wright and Wrigley 199] J. H. Wright and E. N. Wrigley. 1991. GLR Parsing with Probability. In *Generalized LR Parsing*. Kluwer Academic Publishers.