Empirical Support for Probabilistic GLR Parsing

Virach Sornlertlamvanich and **Kentaro Inui** and **Hozumi Tanaka** and **Takenobu Tokunaga**
Department of Computer Science, Tokyo Institute of Technology
2-12-1, Oookayama, Meguro-ku, Tokyo 152
{virach,inui,tanaka,take}@cs.titech.ac.jp

Toshiyuki Takezawa
ATR Interpreting Telecommunications Research Laboratories
takezawa@itl.atr.co.jp

Abstract

This paper discusses the effectiveness of a new probabilistic generalized LR model (PGLR) in word-based parsing (morphological and syntactic analysis) tasks, in which we have to consider the word segmentation and multiple part-of-speech problems. Parsing a sentence from the morphological level makes the task much more complex because of the increase of parse ambiguity stemming from word segmentation ambiguities and multiple corresponding sequences of parts-of-speech. The experiments show that the PGLR model yields the best results comparing with the existing Briscoe and Carroll model (B&C) for GLR parsing, and "two-level PCFG", on experimentation on the ATR Japanese corpus.

1 Introduction

The PGLR model (Inui et al., 1997) has previously been proven better than the existing models, namely the model proposed by Briscoe and Carroll (1993) and the baseline model using a probabilistic context-free grammar (PCFG), in parsing strings of parts-of-speech (non-word-based parsing) (Sornlertlamvanich et al., 1997). Parsing a sentence from the morphological level makes the task much more complex because of the increase of parse ambiguity stemming from word segmentation ambiguities and multiple corresponding sequences of parts-of-speech. In this paper, we empirically evaluate the preciseness of a probabilistic model for PGLR against one for Briscoe and Carroll's model (B&C), which is based on the same GLR parsing framework. We also examine the benefits of context-sensitivity in GLR parsing, of the PGLR model against the "two-level PCFG" model (Chitrao and Grishman, 1990) or "pseudo context-sensitive grammar" model (PCSG)—recently presented in (Charniak and Carroll, 1994)—which has been shown to capture greater context-sensitivity than the original PCFG model, by empirical results and qualitative analysis.

Sornlertlamvanich et al. (1997) demonstrated the superior performance of PGLR over B&C and PCFG in a syntactic analysis task involving a determined sequence of parts-of-speech as input—non-word-based parsing. Most syntactic parsing model evaluation takes a string of parts-of-speech as input and leaves the problems of word segmentation and part-of-speech determination to other morphological analysis modules, such as part-of-speech taggers. Since GLR parsing has the ability to integrate morphological and syntactic analysis (Tanaka et al., 1996), we can easily realize PGLR parsing for morphosyntactic analysis, by adding lexical probabilities. To prove that the local n-gram constraints in PGLR are effective in morphological parsing, we conducted a word-based experiment on the ATR Japanese corpus, and compared the resulting performance with the existing B&C model and two-level PCFG, an extension of PCFG which is claimed to yield a significant performance advantage over the original PCFG framework. Since no spaces are placed between words in Japanese sentences, the models can in this way be evaluated in terms of both morphological and syntactic analysis.

Section 2 briefly reviews the various probabilis-

tic models, namely B&C, two-level PCFG and PGLR, which are evaluated through word-based parsing on the ATR Japanese corpus. Section 3 shows the results of experiments carried out on the three models and the baseline model of PCFG. We discuss the empirical results and give case analyses in Section 4.

2 Probabilistic Models

In this section, we briefly describe the existing probabilistic models, namely B&C and two-level PCFG, which are to be evaluated against the PGLR model. B&C is a probabilistic model proposed for the GLR parsing framework and has a significantly high performance, as presented in (Briscoe and Carroll, 1993). Two-level PCFG is an extended PCFG model for yielding greater context-sensitivity than the original paradigm. It was recently explored more thoroughly by Charniak and Carroll (1994), using the terminology "pseudo context-sensitive grammar" (PCSG), showing the improvement in per-word cross-entropy over the original PCFG model. Our motivation in selecting B&C and two-level PCFG for the comparative evaluation of PGLR is to examine the effectiveness of LR table context-sensitivity (global context over CFG-derived structures and local n-gram context from adjoining preterminal constraints), and the appropriateness of PGLR for GLR parsing.

In word-based parsing, given a string of characters $C = c_1, \ldots, c_n$ as an input, the joint probability of parse tree (T) and word sequence (W) is:

$$P(T,W|C) = \frac{P(T)\cdot P(W|T)\cdot P(C|W,T)}{P(C)} \quad (1)$$

$$\approx P(T)\cdot P(W|T) \quad (2)$$

The term $P(C|W,T)$ becomes one when word sequence W is determined, and $P(C)$ is a constant scaling factor, independent of T and W, which is not worthy of consideration in ranking parse trees and word sequences.

Probabilistic models allow us to estimate parse tree probabilities ($P(T)$). For the lexical probability $P(W|T)$, in our evaluation, we naively assume that word w_i in word sequence $W = w_1, \ldots, w_m$ depends only on its part-of-speech (l_i). Therefore,

$$P(W|T) \approx \prod_{i=1}^{m} P(w_i|l_i) \quad (3)$$

The estimation of lexical probability is applied identically in all models.

2.1 Briscoe and Carroll's Model (B&C)

Briscoe and Carroll (1993) introduced probability to the GLR parsing algorithm in the light of the fact that LR tables do provide appropriate contextual information for solving the context-sensitivity problems observable in real world NL applications. They pointed out that an LR parse state encodes information about the left and right context of the parse. This results in distinguishability of context for an identical rule reapplied in different ways across different derivations. Briscoe and Carroll's method allows us to associate probabilities with an LR table directly, rather than simply with the rules of the grammar.

They consider the LR table as a nondeterministic finite-state automaton. Each row of the LR table corresponds to the possible transitions out of the state represented by that row, and each transition is associated with a particular lookahead item and a parsing action. Nondeterminism arises when more than one action is possible given a particular input symbol. The following is a review of B&C in terms of our formalization.

Briscoe and Carroll regard a parse derivation as a sequence of state transitions (T):-

$$s_0 \overset{l_1,a_1}{\Longrightarrow} s_1 \overset{l_2,a_2}{\Longrightarrow} \ldots \overset{l_{n-1},a_{n-1}}{\Longrightarrow} s_{n-1} \overset{l_n,a_n}{\Longrightarrow} s_n \quad (4)$$

where a_i is an action, l_i is an input symbol and s_i is the state at time t_i. The probability of the parse derivation T is estimated by equation (5):-

$$P(T) \approx \prod_{i=1}^{n} P(l_i, a_i, s_i|s_{i-1}) = \prod_{i=1}^{n} P(l_i, a_i|s_{i-1})\cdot P(s_i|s_{i-1}, l_i, a_i) \quad (5)$$

Based on B&C, the following is a summary of the scheme for deriving the action probabilities ($p(a)$) from the count of state transitions resulting from parsing a training set.

1. The probability of an action given an input symbol is conditioned by the state it originated from. The probabilities assigned to each action at a given state must sum to one. Therefore,

$$\sum_{l\in La(s)} \sum_{a\in Act(s,l)} p(a) = 1 \quad (\text{for } \forall s \in \boldsymbol{S}) \quad (6)$$

where $La(s)$ is the set of input symbols at state s, $Act(s,l)$ is the set of actions given a pair of state s and input symbol l, and $\boldsymbol{S}$ is the set of all states of the LR table. This

means that the actions in the LR table are normalized within each state.

2. In the case of a shift action ($\boldsymbol{A}_s$), $P(s_i|s_{i-1}, l_i, a_i)$ in equation (5) is equal to one because shift conflict never occurs in an LR table. Therefore,

$$p(a) = P(l_i, a_i|s_{i-1}) \quad (\text{for } a_i \in \boldsymbol{A}_s) \qquad (7)$$

3. In the case of a reduce action ($\boldsymbol{A}_r$), the probability is subdivided according to the state reached after applying the reduce action. The reason for this is that Briscoe and Carroll associate probabilities with transitions in the automaton rather than with actions in the action part of the LR table. In this case $P(s_i|s_{i-1}, l_i, a_i)$ in equation (6) is not one. Therefore,

$$p(a) = P(l_i, a_i|s_{i-1}) \cdot P(s_i|s_{i-1}, l_i, a_i) \quad (\text{for } a_i \in \boldsymbol{A}_r) \qquad (8)$$

The probability of a parse derivation in B&C is the geometric mean of the probabilities of the actions for state transitions across the whole parse derivation:-

$$P(T) = (\prod_{i=1}^{n} p(a_i))^{1/n} \qquad (9)$$

2.2 Two-level Probabilistic Context-Free Grammar (two-level PCFG)

Two-level PCFG is an extended version of PCFG, deriving from the idea of providing context-sensitivity for a context-free grammar.

In the original PCFG model, the probability of a parse derivation (T) is regarded as the product of probabilities of the rules which are employed for deriving that parse derivation. Each production rule of the grammar (r_i) is of the form $< A \rightarrow \alpha, P(r_i) >$ where $P(r_i)$ is the associated probability, and the probabilities associated with all rules with a given nonterminal A on the left-hand side must sum to one. Therefore,

$$\sum_{\alpha} P(\alpha|A) = 1 \qquad (10)$$

$$P(T) = \prod_{i} P(r_i) \qquad (11)$$

Two-level PCFG utilizes extra information provided by the parent of nonterminals in expanding rules (r_i) through assignment of rule probabilities. Thus, the rule probability in equation (10) can be rewritten as:

$$\sum_{\alpha} P(\alpha|\rho(A)) = 1 \qquad (12)$$

where $\rho(A)$ is the nonterminal that immediately dominates A (i.e. its parent).

2.3 Probabilistic Generalized LR (PGLR)

Inui et al. (1997) recently proposed a new formalization of a probabilistic model for GLR parsing. Unlike B&C, a parse derivation is regarded as a sequence of transitions between LR parse stacks (T) as shown in (13), where σ_i is the stack at time t_i, a_i is an action, and l_i is an input symbol. Schema (13) shows the inherent diversion from B&C in the definition of parse derivation.

$$\sigma_0 \overset{l_1,a_1}{\Longrightarrow} \sigma_1 \overset{l_2,a_2}{\Longrightarrow} \cdots \overset{l_{n-1},a_{n-1}}{\Longrightarrow} \sigma_{n-1} \overset{l_n,a_n}{\Longrightarrow} \sigma_n \qquad (13)$$

Based on the above definition, the probability of a complete stack transition sequence T can be represented with equation (14), by assuming that σ_i contains all the information of its preceding parse derivation:-

$$\begin{aligned} P(T) &= \prod_{i=1}^{n} P(l_i, a_i, \sigma_i|\sigma_{i-1}) \\ &= \prod_{i=1}^{n} P(l_i, a_i|\sigma_{i-1}) \cdot P(\sigma_i|\sigma_{i-1}, l_i, a_i) \end{aligned} \qquad (14)$$

Due to the two different types of actions in the LR table, namely *shift* and *reduce* actions (*accept* is an additional special dummy action to successfully end the parsing process), **states are treated differently according to the type of action applied to reach that state**. That is, states reached after the application of a reduce action have the same input symbol as the former state, whereas states reached after the application of a shift action read in a new input symbol. As a result, states are classified into two classes and the probabilities of actions are estimated differently corresponding to the class they belong to. By assuming that the stack-top state contains sufficient information of the current stack, the probability of the current action a_i is estimated from the state s_i on top of the current stack, instead of the full stack σ_i. Therefore,

$$P(l_i, a_i, \sigma_i|\sigma_{i-1}) \approx \begin{cases} P(l_i, a_i|s_{i-1}) & (s_{i-1} \in \boldsymbol{S}_s) \\ P(a_i|s_{i-1}, l_i) & (s_{i-1} \in \boldsymbol{S}_r) \end{cases} \qquad (15)$$

such that:

$$\sum_{l \in La(s)} \sum_{a \in Act(s,l)} p(a) = 1 \quad (\text{for } s \in \boldsymbol{S}_s) \qquad (16)$$

$$\sum_{a \in Act(s,l)} p(a) \quad = \quad 1 \quad (\text{for } s \in \boldsymbol{S}_r) \quad (17)$$

where $p(a)$ is the probability of an action a, $\boldsymbol{S}_s$ is the class of states reached after applying a shift action, including the initial state, and $\boldsymbol{S}_r$ is the class of states reached after applying a reduce action.

The probability of a parse derivation is the product of the probabilities of the actions for stack transitions across the whole parse derivation:-

$$P(T) = \prod_{i=1}^{n} p(a_i) \quad (18)$$

3 Experimental Results

We evaluated the various probabilistic models (i.e. B&C, two-level PCFG and PGLR) on a portion of the ATR Japanese corpus, called Spoken Language Database (SLDB) (Takezawa, 1997). Given an input string of Japanese characters, each model produces probabilistically ranked parses together with the associated parse probabilities, computed as described in Section 2.

As the dictionary to generate word candidates and their corresponding parts-of-speech, we collected all the words used in the corpus. Each word in the dictionary retains lexical probability $P(w|l)$, which is the probability of generating word 'w' from an arbitrary part-of-speech 'l'.

Each model was trained equally with the same hand-annotated training set. For unseen events, we simply added part of a count to smooth the model probabilities. Evaluation of the smoothing method is beyond the scope of this paper. We additionally present the results of the original PCFG framework as the baseline for the evaluation.

3.1 ATR Corpus and Grammar

The "Spoken Language Database" (SLDB) is a treebank (a collection of trees annotated with a syntactic analysis, or "trees" for short) developed by ATR based on Japanese dialogue. A portion of the corpus has been revised through application of a more detailed phrasal grammar developed by Tanaka et al. (1997). We randomly selected about 5% of the revised corpus to use as a test set and trained each parsing model with the remaining approximately 10,000 trees. Table 1 describes a breakdown of the corpus. The range and average of sentence length in both the training and test sets are very close, from which it is plain that the test set was appropriately selected from the corpus.

We implemented all the models using a GLR parser. We generated an LALR table from the treebank governing context-free grammar of 762 production rules, comprised of 137 non-terminal symbols and 407 terminal symbols. The generated LALR table contained 856 states.

Table 1: ATR Corpus

		# of Morphemes		# of Characters	
ATR Corpus	# of Sent.	Range	Ave.	Range	Ave.
Training set	10,361	1-34	6.69	2-58	12.57
Test set	534	1-22	6.14	2-42	11.74

3.2 Parsing the ATR Corpus

We used PARSEVAL measures (Black et al., 1991) to compare the performance of the top-ranked parses for each model.

Table 2 shows that the PGLR model outperformed the other models in every metric. The average parse base[1] (APB) of the test set is as high as 1.341 in the character-based measure and 2.067 in the word-based measure. This is comparable with the SUSANNE corpus (1.256) and SEC corpus (1.239), as reported in (Briscoe and Carroll, 1995) Therefore, the performance of word-based parsing in this test mainly depends on the accuracy of selecting words and their corresponding parts-of-speech. This means that a model that could provide local context in addition to the global context would result in higher performance.

As expected, the models which make effective use of the local context modeling nature of GLR parsing, namely B&C and PGLR, returned significantly higher results than PCFG-based parsing. Although the PCFG rule context in two-level PCFG extends to a step higher (i.e. to the parent of the reduced rule), the model still failed to include appropriate context in some cases. One such case is shown in Section 4.

Parse accuracy (PA) shows the percentage of correct parses that are ranked topmost according to the model probability. By this measure, PGLR maintained the highest accuracy in ranking parses, while the PCFG-based models dropped down to slightly higher than 50%. Since the corpus is a kind of spoken language database, there are a lot of short response utterances i.e. "yes", "no" and "take care". The lower table in Table 2 is added to show the performance on sentences ranging from 14 to 42 characters. The difference in performance becomes obvious in parsing longer sentences.

[1] Briscoe and Carroll (1995) defined APB as the measure of ambiguity for a given corpus. It is the geometric mean over all sentences in the corpus of $\sqrt[n]{p}$, where n is the number of words in a sentence, and p is the number of parses for that sentence.

Table 2: Performance on the ATR Corpus. **PA** is the parse accuracy and indicates the percentage of top-ranked parses that match standard parses. **LP/BR** are label precision/recall. **BP/BR** are bracket precision/recall. **0-CB** and **m-CB** are zero crossing brackets and mean crossing brackets per sentence, respectively.

	2-42 Characters (534 sentences)						
Models	**PA**	**LP**	**LR**	**BP**	**BR**	**0-CB**	**m-CB**
B&C	89.33	97.79	97.54	98.53	98.06	94.57	0.11
Two-leval PCFG	62.55	96.31	95.31	98.66	97.38	95.32	0.09
PCFG	53.93	95.64	94.48	98.77	97.31	94.76	0.08
PGLR	95.32	99.06	98.47	99.53	98.73	98.50	0.03

Parse performance partly depends on the grammar and the corpus. According to the report of an experiment on the SUSANNE English corpus by Carroll (1997), the difference between the performance of PGLR and B&C was not significantly observed. However, the input test set was a set of part-of-speech sequences, excluding ambiguity in word and part-of-speech selection. Even here, though PGLR returned the best result in terms of the m-CB metric.

4 Discussion

It is obvious that two-level PCFG shows the benefits of context-sensitivity and yields significant gains over the original PCFG model. However, the results are still far below those for the probabilistic GLR-based parsing models. One reason would be the advantages of local context, i.e. pre-terminal n-gram constraints encoded in the LR table. The n-gram constraints are distributed over the actions of the table. Therefore, the parse trees generated by probabilistic GLR-based parsers include pre-terminal n-gram constraints in the parse probabilities.

The exemplified case below shows that probabilistic GLR-based parsing can successfully exploit the advantages of pre-terminal n-gram constraints, and assign parse probabilities in a more accurate manner. Based on Grammar-1, the three parse tree types in Figure 1 can be generated. Supposing that (S1) and (S2) are found one and two times respectively in our training set, but (S3) does not occur. (S3) can be found very rarely, or alternatively never occur because it may have no obvious meaning. This actually happens for most wide-coverage grammars.

The case shown in Figure 1 was indeed found in our test when 'b' is a sentence-ending terminal symbol and 'a' usually occurs with 'c'. Especially in word-based parsing where terminal symbols are not fixed in the input strings, the parser must allocate appropriate preferences across all possible parses. In this case, a string can have a part-of-speech of 'b', or be broken into two words with parts-of-speech of 'a' and 'c'.

Grammar-1:-

1. X → U c
2. X → U
3. U → a
4. U → b

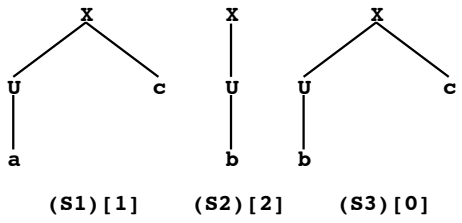

Figure 1: Parse trees, with the frequency in the training set shown in brackets

Probability-1:-

1. S ; X → U c (1/3)
2. S ; X → U (2/3)
3. X ; U → a (1/3)
4. X ; U → b (2/3)

The bracketed values given for Probability-1 are the rule probabilities estimated according to the two-level PCFG model from the training set in Figure 1. In fact, they are the same as for PCFG because the parents of rules (1) and (2) are not different, and neither are the parents of rules (3) and (4). This means that the extended context in two-level PCFG does not have any effect if direct parents are the same. We need more information to distinguish the cases. Unfortunately, however, there are no other parent nodes in this case.

Table 3 is an LALR table generated from Grammar-1. The associated probabilities below each action are estimated according to B&C and PGLR, indicated in the first and second lines of each state row, respectively. For the sake of brevity, we do not consider any smoothing technique in this table, although smoothing was performed in the experiments described in Section 3.

Table 3: LALR table with its associated probabilities. Probabilities in the first line of each state row are those estimated by B&C and the bracketed values in the second line are those estimated by PGLR

	Action				Goto	
State	a	b	c	$	U	X
0	s3 1/3 (1/3)	s2 2/3 (2/3)			1	4
1			sh5 1/2 (1)	re2 1/2 (1)		
2			re4	re4 1 (1)		
3			re3 1 (1)	re3		
4				acc 1 (1)		
5				re1 1 (1)		

Applying the probabilities prepared in Probability-1 for two-level PCFG (as well as PCFG), and Table 3 for B&C and PGLR, to estimate the parse probabilities of (S1), (S2) and (S3) in Figure 1, we obtain the results shown in Table 4. Two-level PCFG (and PCFG) wrongly assigned preference to (S3) over (S1), whereas (S3) never occurs in the training set. Although B&C yields correct preference, the probabilities are smaller than what they should be. In this case, there is no difference between B&C and PGLR in ranking the parses. The side-effects of inappropriate normalization of probabilities in B&C has already been explored in (Inui et al., 1997) and empirically confirmed in the evaluation in Section 3.

5 Conclusion

The results of our experiments clearly showed that the PGLR model is able to make effective use of both global and local context within the GLR parsing schema. As a result, our model outperformed both Briscoe and Carroll's model and the two-level PCFG model in all tests. In addition, PGLR needs only the probability for each action in the LR table to compute the overall probability of each parse. It is thus tractable to training, with the degree of free parameters as small as the number of distinct actions, and associates a probability directly to each action.

Table 4: Probabilities of parse trees, (S1), (S2) and (S3), estimated with each model

Models	(S1)	(S2)	(S3)
PCFG	1/9	4/9	2/9
Two-level PCFG	1/9	4/9	2/9
B&C	1/6	1/3	0
PGLR	1/3	2/3	0

References

A. Aho, R. Sethi, and J. Ullman. 1986. *Compilers: Principles, Techniques, and Tools.* Addison-Wesley.

E. Black et al. 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.

T. Briscoe and J. Carroll. 1993. Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars. *Computational Linguistics*, 19(1):25–59.

T. Briscoe and J. Carroll. 1995. Developing and Evaluating a Probabilistic LR Parser of Part-Of-Speech and Punctuation Labels. In *Proceedings of the 4th International Workshop on Parsing Technologies*, pages 48–58.

J. Carroll. 1997. Report of Visit to Tanaka Laboratory, 1997/11/8 - 1997/12/8. (unpublished).

E. Charniak and G. Carroll. 1994. Context-Sensitive Statistics for Improved Grammatical Language Models. In *Proceedings of AAAI-94*, pages 728–733.

M. Chitrao and R. Grishman. 1990. Statistical Parsing of Messages. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 263–266.

K. Inui, V. Sornlertlamvanich, H. Tanaka, and T. Tokunaga. 1997. A New Formalization of Probabilistic GLR Parsing. In *Proceedings of the 5th International Workshop on Parsing Technologies.*

V. Sornlertlamvanich, K. Inui, K. Shirai, H. Tanaka, T. Tokunaga, and T. Takezawa. 1997. Empirical Evaluation of Probabilistic GLR Parsing. In *Proceedings of NLPRS-97*, pages 169–174.

T. Takezawa. 1997. ATR Japanese Syntactic Structure Database and the Grammar. Technical report, ATR Interpreting Telecommunications Research Laboratories, April. (in Japanese).

H. Tanaka, T. Tokunaga, and M. Izawa. 1996. Integration of Morphological and Syntactic Analysis Based on GLR Parsing. In H. Bunt and M. Tomita, editors, *Recent Advances in Parsing Technology*, pages 325–342. Kluwer Academic Publishers.

H. Tanaka, T. Takezawa, and J. Etoh. 1997. Japanese Grammar for Speech Recognition Considering the MSLR Method. Technical Report 97-SLP-15-25, Information Processing Society of Japan. (in Japanese).