Exploiting Contextual Information in Hypothesis Selection for Grammar Refinement

Thanaruk Theeramunkong
Japan Advanced Institute of
Science and Technology
1-1 Asahidai Tatsunokuchi
Nomi Ishikawa Japan
ping@jaist.ac.jp

Yasunobu Kawaguchi
Japan Advanced Institute of
Science and Technology
1-1 Asahidai Tatsunokuchi
Nomi Ishikawa Japan
kawaguti@jaist.ac.jp

Manabu Okumura
Japan Advanced Institute of
Science and Technology
1-1 Asahidai Tatsunokuchi
Nomi Ishikawa Japan
oku@jaist.ac.jp

Abstract

In this paper, we propose a new framework of grammar development and some techniques for exploiting contextual information in a process of grammar refinement. The proposed framework involves two processes, *partial grammar acquisition* and *grammar refinement*. In the former process, a rough grammar is constructed from a bracketed corpus. The grammar is later refined by the latter process where a combination of rule-based and corpus-based approaches is applied. Since there may be more than one rules introduced as alternative hypotheses to recover the analysis of sentences which cannot be parsed by the current grammar, we propose a method to give priority to these hypotheses based on local contextual information. By experiments, our hypothesis selection is evaluated and its effectiveness is shown.

1 Introduction

One of the essential tasks to realize an efficient natural language processing system is to construct a broad-coverage and high-accurate grammar. In most of the currently working systems, such grammars have been derived manually by linguists or lexicographers. Unfortunately, this task requires time-consuming skilled effort and, in most cases, the obtained grammars may not be completely satisfactory and frequently fail to cover many unseen sentences. Toward these problems, there were several attempts developed for automatically learning grammars based on rule-based approach(Ootani and Nakagawa, 1995), corpus-based approach(Brill, 1992)(Mori and Nagao, 1995) or hybrid approach(Kiyono and Tsujii, 1994b)(Kiyono and Tsujii, 1994a).

Unlike previous works, we have introduced a new framework for grammar development, which is a combination of rule-based and corpus-based approaches where contextual information can be exploited. In this framework, a whole grammar is not acquired from scratch(Mori and Nagao, 1995) or an initial grammar does not need to be assumed(Kiyono and Tsujii, 1994a). Instead, a rough but effective grammar is learned, in the first place, from a large corpus based on a corpus-based method and then later refined by the way of the combination of rule-based and corpus-based methods. We call the former step of the framework *partial grammar acquisition* and the latter *grammar refinement*. For the partial grammar acquisition, in our previous works, we have proposed a mechanism to acquire a partial grammar automatically from a bracketed corpus based on local contextual information(Theeramunkong and Okumura, 1996) and have shown the effectiveness of the derived grammar(Theeramunkong and Okumura, 1997). Through some preliminary experiments, we found out that it seems difficult to learn grammar rules which are seldom used in the corpus. This causes by the fact that rarely used rules occupy too few events for us to catch their properties. Therefore in the first step, only grammar rules with relatively high occurrence are first learned.

In this paper, we focus on the second step, grammar refinement, where some new rules can be added to the current grammar in order to accept unparsable sentences. This task is achieved by two components: (1) the rule-based component, which detects incompleteness of the current grammar and generates a set of hypotheses of new rules and (2) the corpus-based component, which selects plausible hypotheses based on local contextual information. In addition, this paper also describes a stochastic parsing model which finds the most likely parse of a sentence and then evaluates the hypothesis selection based on the plausible parse.

In the rest, we give an explanation of our framework and then describe the grammar refinement process and hypothesis selection based on local contextual information. Next, a stochastic parsing model which exploits contextual information is described. Finally, the effectiveness of our approach is shown through some experiments investigating the correctness of selected hypotheses and parsing accuracy.

2 The Framework of Grammar Development

The proposed framework is composed of two phases: partial grammar acquisition and grammar refinement. The graphical representation of the framework is shown in Figure 1. In the process of grammar development, a partial grammar is automatically acquired in the first phase and then it is refined in the second phase. In the latter phase, the system generates new rules and ranks them in the order of priority before displaying a user a list of plausible rules as candidates for refining the grammar. Then the user can select the best one among these rules. Currently, the corpus used for grammar development in the framework is EDR corpus(EDR, 1994) where lexical tags and bracketings are assigned for words and phrase structures of sentences in the corpus respectively but no nonterminal labels are given.

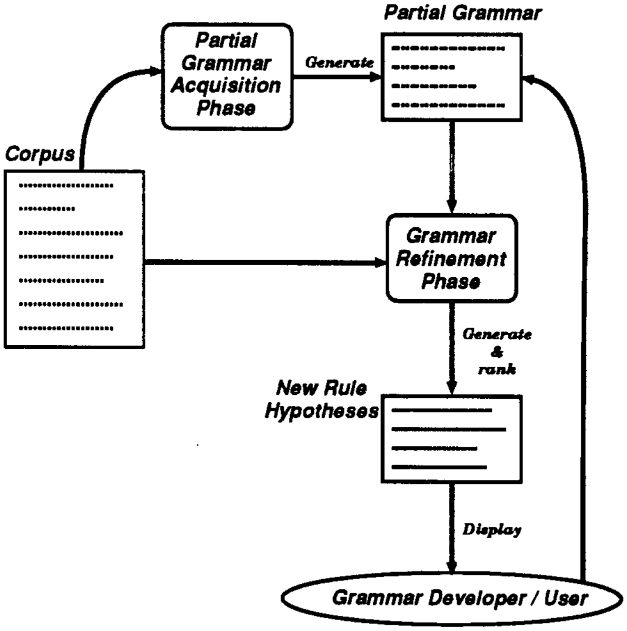

Figure 1: The overview of our grammar development framework

2.1 Partial Grammar Acquisition

In this section, we give a brief explanation for partial grammar acquisition. More detail can be found in (Theeramunkong and Okumura, 1996). In partial grammar acquisition, a rough grammar is constructed from the corpus based on clustering analysis. As mentioned above, the corpus used is a tagged corpus with phrase structures marked with brackets. At the first place, brackets covering a same sequence of categories, are assumed to have a same nonterminal label. We say they have the same bracket type. The basic idea is to group brackets (bracket types) in a corpus into a number of similar bracket groups. Then the corpus is automatically labeled with some nonterminal labels, and consequently a grammar is acquired. The similarity between any two bracket types is calculated based on divergence[1](Harris, 1951) by utilizing local contextual information which is defined as a pair of categories of words immediately before and after a bracket type. This approach was evaluated through some experiments and the obtained result was almost consistent with that given by human evaluators. However, in this approach, when the number of occurrences of a bracket type is low, the similarity between this bracket type and other bracket types is not so reliable. Due to this, only bracket types with relatively frequent occurrence are taken into account. To deal with rarely occurred bracket types, we develop the second phase where the system shows some candidates to grammar developers and then they can determine the best one among these candidates, as shown in the next section.

2.2 Grammar Refinement with Additional Hypothesis Rule

The grammar acquired in the previous phase is a partial one. It is insufficient for analyzing all sentences in the corpus and then the parser fails to produce any complete parse for some sentences. In order to deal with these unparsable sentences, we modify the conventional chart parser to keep record of all inactive edges as partial parsing results. Two processes are provoked to find the possible plausible interpretations of an unparsable sentence by hypothesizing some new rules and later to add them to the current grammar. These processes are (1) the rule-based process, which detects incompleteness of the current grammar and generates a set of hypotheses of new rules and (2) the corpus-based process, which selects plausible hypotheses based on local contextual information. In the rule-based process, the parser generates partial parses of a sentence as much as possible in bottom-up style under the grammar constraints. Utilizing these parses, the process detects a complete parse of a sentence by starting at top category (i.e., sentence) covering the sentence and then searching down, in top-down manner, to the part of the sentence that cannot form any parse. At this point, a rule is hypothesized. In many cases, there may be several possibilities for hypothesized rules. The corpus-based process, as the second process, uses the probability information from parsable sentences to rank these hypotheses. In this research, local contextual information is taken into account for this task.

[1] The effectiveness of divergence for detecting phrase structures in a sentence is also shown in (Brill, 1992).

3 Hypothesis Generation

When the parser fails to parse a sentence, there exists no inactive edge of category S (sentence) spanning the whole sentence in the parsing result. Then the hypothesis generation process is provoked to find all possible hypotheses in top-down manner by starting at a single hypothesis of the category S covering the whole sentence. This process uses the partial chart constructed during parsing the sentence. This hypothesis generation is similar to one applied in (Kiyono and Tsujii, 1994a).

[Hypothesis generation]

An inactive edge $[ie(A) : x_0, x_n]$ can be introduced from x_0 to x_n, with label A, for each of the hypotheses generated by the following two steps.

1. For each sequence of inactive edges, $[ie(B_1) : x_0, x_1], ..., [ie(B_n) : x_{n-1}, x_n]$, spanning from x_0 to x_n, generate a new rule, $A \rightarrow B_1, ..., B_n$, and propose a new inactive edge as a hypothesis, $[hypo(A) : x_0, x_n]$. (Figure 2(1))
2. For each existing rule $A \rightarrow A_1, ..., A_n$, find an incomplete sequence of inactive edges, $[ie(A_1) : x_0, x_1], ..., [ie(A_{i-1}) : x_{i-2}, x_{i-1}]$, $[ie(A_{i+1}) : x_i, x_{i+1}], ..., [ie(A_n) : x_{n-1}, x_n]$, and call this algorithm for $[ie(A_i) : x_{i-1}, x_i]$.(Figure 2(2))

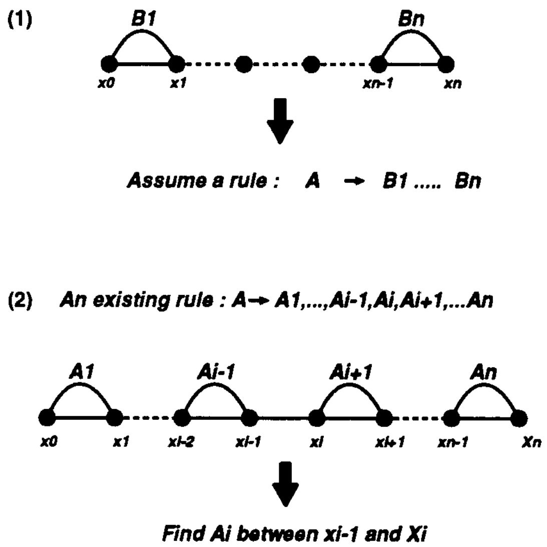
(1)
B1
Bn
x0
x1
xn-1
xn
Assume a rule : A → B1 ..... Bn
(2) An existing rule : A→ A1,...,Ai-1,Ai,Ai+1,...An
A1
Ai-1
Ai+1
An
x0
x1
xi-2
xi-1
xi
xi+1
xn-1
Xn
Find Ai between xi-1 and Xi

Figure 2: Hypothesis Rule Generation

By this process, all of possible single hypotheses (rules) which enable the parsing process to succeed, are generated. In general, among these rules, most of them may be linguistically unnatural. To filter out such unnatural hypotheses, some syntactical criteria are introduced. For example, (1) the maximum number of daughter constituents of a rule is limited to three, (2) a rule with one daughter is not preferred, (3) non-lexical categories are distinguished from lexical categories and then a rule with lexical categories as its mother is not generated. By these simple syntactical constraints, a lot of useless rules can be discarded.

4 Hypothesis Selection with Local Contextual Information

Hypothesis selection utilizes information from local context to rank the rule hypotheses generated in the previous phase. In the hypothesis generation, although we use some syntactical constraints to reduce the number of hypotheses of new rules that should be registered into the current grammar, there may still be several candidates remaining. At this point, a scoring mechanism is needed for ranking these candidates and then one can select the best one as the most plausible hypothesis.

This section describes a scoring mechanism which local contextual information can be exploited for this purpose. As mentioned in the previous section, local contextual information referred here is defined as a pair of categories of words immediately before and after the brackets. This information can be used as an environment for characterizing a nonterminal category. The basic idea in hypothesis selection is that the rules with a same nonterminal category as their mother tend to have similar environments. Local contextual information is gathered beforehand from the sentences in the corpus which the current grammar is enough for parsing.

When the parser faces with a sentence which cannot be analyzed by the current grammar, some new rule hypotheses are proposed by the hypothesis generator. Then the mother categories of these rules will be compared by checking the similarity with the local contextual information of categories gathered from the parsable sentences. Here, the most likely category is selected and that rule will be the most plausible candidate. The scoring function (probability p) for a rule hypothesis $Cat \rightarrow \alpha$ is defined as follows.

$$p(Cat \rightarrow \alpha) = p(Cat|l, r) = \frac{N(Cat, l, r)}{N(l, r)} \quad (1)$$

where $N(Cat, l, r)$ is the number of times that Cat is occurred in the environment (l, r). l is the category immediately before Cat and r is the lexical category of the word immediately after Cat. $N(l, r)$ is the number of times that l and r are occurred immediately before and after any categories. Note that because it is not possible for us to calculate the probability of $Cat \rightarrow \alpha$ in the environment of (l, r), we estimate this number by the probability that Cat occurs in the environment of (l, r). That is, how easy the category Cat appears under a certain environment (l, r).

5 The Stochastic Model

This section describes a statistical parsing model which finds the most plausible interpretation of a sentence when a hypothesis is introduced for recovering the parsing process of the sentence. In this problem, there are two components taken into account: a statistical model and parsing process. The model assigns a probability to every candidate parse tree for a sentence. Formally, given a sentence S and a tree T, the model estimates the conditional probability $P(T|S)$. The most likely parse under the model is $argmax_T P(T|S)$ and the parsing process is a method to find this parse. In general, a model of a simple probabilistic context free grammar (CFG) applies the probability of a parse which is defined as the multiplication of the probability of all applied rules. However, for the purposes of our model where left and right contexts of a constituent are taken into account, the model can be defined as follows.

$$P(T|S) = \prod_{(rl_i,l_i,r_i)\in T} p(rl_i, l_i, r_i) \qquad (2)$$

where rl_i is an application rule in the tree and l_i and r_i are respectively the left and right contexts at the place the rule is applied. In a parsing tree, there is a hypothesis rule for which we cannot calculate the probability because it does not exist in the current grammar. Thus we estimate its probability by using the formula (1) in section 4.

Similar to most probabilistic models, there is a problem of low-frequency events in this model. Although some statistical NL applications apply backing-off estimation techniques to handle low-frequency events, our model uses a simple interpolation estimation by adding a uniform probability to every events. Moreover, we make use of the geometric mean of the probability instead of the original probability in order to eliminate the effect of the number of rule applications as done in (Magerman and Marcus, 1991). The modified model is:

$$P(T|S) = \Big(\prod_{(rl_i,l_i,r_i)\in T} (\alpha * p(rl_i, l_i, r_i) + (1-\alpha) * \frac{1}{N_{rl}N_c}) \Big)^{\frac{1}{|T|}} \qquad (3)$$

Here, α is a balancing weight between the observed distribution and the uniform distribution. It is assigned with 0.8 in our experiments. N_{rl} is the number of rules and N_c is the number of possible contexts, i.e., the left and right categories. The applied parsing algorithm is a simple bottom-up chart parser whose scoring function is based on this model. A dynamic programming algorithm is used to find the Viterbi parse: if there are two proposed constituents which span the same set of words and have the same label, then the lower probability constituent can be safely discarded.

6 Experimental Evaluation

Some evaluation experiments and their results are described. For the experiments, we use texts from the EDR corpus, where bracketings are given. The subject is 48,100 sentences including around 510,000 words. Figure 3 shows some example sentences in the EDR corpus

```
(((ART,"a")((ADJ,"large")(NOUN,"festival")))
   ((VT,"held")(ADV,"biennially")))

((ADV,"again")((PRON,"he")((VT,"says")
   ((PRON,"he")((ADV,"completely")
   ((VT,"forgot")((PREP,"about")
   ((PRON,"his")(NOUN,"homework")))))))))
```

Figure 3: Some example sentences in the EDR corpus

The initial grammar is acquired from the same corpus using divergence shown in section 2.1. The number of rules is 272, the maximum length of rules is 4, and the numbers of terminal and nonterminal categories are 18 and 55 respectively. A part of the initial grammar is enumerated in Figure 4. In the grammar, *l1n1* is expected to be noun phrase with an article, *l1n2* is expected to be noun phrase without an article, and *l1n3* is expected to be verb phrase. Moreover, among 48,100 sentences, 5,083 sentences cannot be parsed by the grammar. We use these sentences for evaluating our hypothesis selection approach.

l1n1	→	adv, noun
l1n1	→	adv, l1n1
l1n1	→	adv, l1n2
l1n1	→	art, noun
....	......	....
l1n2	→	adj, noun
l1n2	→	adj, l1n1
l1n2	→	adj, l1n2
l1n2	→	adj, l1n8
....	......	....
l1n3	→	adv, l1n3
l1n3	→	aux, vt
l1n3	→	aux, l1n13
l1n3	→	l1n12, vt
....	......	....

Figure 4: A part of initial grammar

6.1 The Criterion

In the experiments, we use bracket crossing as a criterion for checking the correctness of the generated hypothesis. Each result hypothesis is compared with the brackets given in the EDR corpus. The correctness of a hypothesis is defined as follows.

- At least one of the derivations inside the hypothesis include the brackets which do not cross with those given in the corpus
- When the hypothesis is applied, it can be used to form a tree whose brackets do not cross with those given in the corpus.

6.2 Hypothesis Level Evaluation

From 5,083 unparsable sentences, the hypothesis generator can produce some hypotheses for 4,730 sentences (93.1%). After comparing them with the parses in the EDR corpus, the hypothesis sets of 3,127 sentences (61.5 %) include correct hypotheses. Then we consider the sentences for which some correct hypotheses can be generated (i.e., 3,127 sentences) and evaluate our scoring function in selecting the most plausible hypothesis. For each sentence, we rank the generated hypotheses by their preference score according to our scoring function. The result is shown in Table 1. From the table, even though only 12.3 % of the whole generated hypotheses are correct, our hypothesis selection can choose the correct hypothesis for 41.6 % of the whole sentences when the most plausible hypothesis is selected for each sentence. Moreover, 29.8 % of correct hypotheses are ordered at the ranks of 2-5, 24.3 % at the ranks of 6-10 and just only 6.2 % at the ranks of more than 50. This indicates that the hypothesis selection is influential for placing the correct hypotheses at the higher ranks. However, when we consider the top 10 hypotheses, we found out that the accuracy is (1362+3368+3134)/(3217+11288+12846) = 28.8 %. This indicates that there are a lot of hypotheses generated for a sentence. This suggests us to consider the correct hypothesis for each sentence instead of all hypotheses.

Ranking	whole (A) hypotheses	correct (B) hypotheses	A/B
1	3217	1340	41.6 %
2-5	11288	3368	29.8 %
6-10	12846	3134	24.3 %
11-20	22105	4300	19.4 %
21-30	17743	2673	15.0 %
31-50	27001	3033	11.2 %
51-	102015	6315	6.2 %
all	196214	24203	12.3 %

Table 1: Hypothesis Level Evaluation

6.3 Sentence Level Evaluation

In this section, we consider the accuracy of our hypothesis selection for each sentence. Table 2 displays the accuracy of hypothesis selection by changing the number of selected hypotheses.

From the table, the number of sentences whose best hypothesis is correct, is 1,340 (41.6%) and we can get up to 2,623 (81.5%) accuracy when the top 10 of the ordered hypotheses are considered. The result shows that our hypothesis selection is effective enough to place the correct hypothesis at the higher ranks.

Ranking	sentences with correct hypo.(A)	A/all
1	1340	41.6 %
2-5	1006	31.2 %
6-10	277	8.6 %
11-20	225	7.0 %
21-30	111	3.5 %
31-50	121	3.8 %
51-	136	4.2 %
all	3217	100.0 %

Table 2: Sentence Level Evaluation

6.4 Parsing Evaluation

Another experiment is also done for evaluating the parsing accuracy. The parsing model we consider here is one described in section 5. The chart parser outputs the best parse of the sentence. This parse is formed by using grammar rules and a single rule hypothesis. The result is shown in Table 3. In this evaluation, the PARSEVAL measures as defined in (Black and et al., 1991) are used:

Precision =

$$\frac{\textit{number of correct brackets in proposed parses}}{\textit{number of brackets in proposed parses}}$$

Recall =

$$\frac{\textit{number of correct brackets in proposed parses}}{\textit{number of brackets in corpus parses}}$$

From this result, we found out that the parser can succeed 57.3 % recall and 65.2 % precision for the short sentences (3-9 words). In this case, the averaged crossings are 1.87 per sentence and the number of sentences with less than 2 crossings is 69.2 % of the comparisons. For long sentences not so much advantage is obtained. However, our parser can achieve 51.4 % recall and 56.3 % precision for all unparsable sentences.

7 Discussion and Conclusion

In this paper, we proposed a framework for exploiting contextual information in a process of grammar refinement. In this framework, a rough grammar is first learned from a bracketed corpus and then the grammar is refined by the combination of rule-based and corpus-based methods. Unlike stochastic parsing such as (Magerman, 1995)(Collins, 1996), our approach can parse sentences which fall out the current grammar and suggest the plausible hypothesis rules and the best parses. The grammar is not acquired from scratch like the approaches shown in

Sent. Length	3-9	3-15	10-19	all length
Comparisons	1980	3864	2491	4730
Avg. Sent. Len.	6.9	9.5	13.4	10.8
Corpus Parses	5.15	7.65	11.47	8.95
System's Parses	5.78	8.27	12.07	9.57
Crossings/Sent.	1.87	3.32	5.69	4.18
Sent. cross.= 0	20.1%	10.6%	0.4%	8.9%
Sent. cross.≤ 1	43.9%	25.0%	3.9%	21.1%
Sent. cross.≤ 2	69.2%	41.7%	9.7%	35.1%
Recall	57.3%	53.2%	47.3%	51.4%
Precision	65.2%	58.7%	50.0%	56.3%

Table 3: Parsing Accuracy

(Pereira and Schabes, 1992)(Mori and Nagao, 1995). Through some experiments, our method can achieve effective hypothesis selection and parsing accuracy to some extent. As our further work, we are on the way to consider the correctness of the selected hypothesis of the most plausible parses proposed by the parser. Some improvements are needed to grade up the parsing accuracy. Another work is to use an existing grammar, instead of an automatically learned one, to investigate the effectiveness of contextual information. By providing a user interface, this method will be useful for grammar developers.

Acknowledgements

We would like to thank the EDR organization for permitting us to access the EDR corpus. Special thanks go to Dr. Ratana Rujiravanit, who helps me to keenly proofread a draft of this paper. We also wish to thank the members in Okumura laboratory at JAIST for their useful comments and their technical supports.

References

Black, E. and et al. 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proc. of the 1991 DARPA Speech and Natural Language Workshop*, pages 306-311.

Brill, Eric. 1992. Automatically acquiring phrase structure using distributional analysis. In *Proc. of Speech and Natural Language Workshop*, pages 155-159.

Collins, Michael John. 1996. A new statistical parser based on bigram lexical dependencies. In *Proc. of the 34th Annual Meeting of the ACL*, pages 184-191.

EDR: Japan Electronic Dictionary Research Institute, 1994. *EDR Electric Dictionary User's Manual (in Japanese)*, 2.1 edition.

Harris, Zellig. 1951. *Structural Linguistics*. Chicago: University of Chicago Press.

Kiyono, Masaki and Jun'ichi Tsujii. 1994a. Combination of symbolic and statistical approaches for grammatical knowledge acquisition. In *Proc. of 4th Conference on Applied Natural Language Processing (ANLP'94)*, pages 72-77.

Kiyono, Masaki and Jun'ichi Tsujii. 1994b. Hypothesis selection in grammar acquisition. In *COLING-94*, pages 837-841.

Magerman, D. M. and M. P. Marcus. 1991. Pearl: A probabilistic chart parser. In *Proceedings of the European ACL Conference*.

Magerman, David M. 1995. Statistical decision-tree models for parsing. In *Proceeding of 33rd Annual Meeting of the ACL*, pages 276-283.

Mori, Shinsuke and Makoto Nagao. 1995. Parsing without grammar. In *Proc. of the 4th International Workshop on Parsing Technologies*, pages 174-185.

Ootani, K. and S Nakagawa. 1995. A semiautomatic learning method of grammar rules for spontaneous speech. In *Proc. of Natural Language Processing Pacific Rim Symposium'95*, pages 514-519.

Pereira, F. and Y. Schabes. 1992. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of 30th Annual Meeting of the ACL*, pages 128-135.

Theeramunkong, Thanaruk and Manabu Okumura. 1996. Towards automatic grammar acquisition from a bracketed corpus. In *Proc. of the 4th International Workshop on Very Large Corpora*, pages 168-177.

Theeramunkong, Thanaruk and Manabu Okumura. 1997. Statistical parsing with a grammar acquired from a bracketed corpus based on clustering analysis. In *International Joint Conference on Artificial Intelligence (IJCAI-97), Poster Session*.