A Statistical Grammar Acquisition Method Based on Clustering Analysis using a Bracketed Corpus

Thanaruk Theeramunkong
Department of Information Technology
Sirindhorn International Institute of Technology
Thammasat University, Pathumthani 12121, Thailand
ping@siit.tu.ac.th

ABSTRACT – This paper proposes a new method for learning a context-sensitive conditional probability context-free grammar from an unlabeled bracketed corpus based on clustering analysis and describes a natural language parsing model which uses a probability-based scoring function of the grammar to rank parses of a sentence. By grouping brackets in a corpus into a number of similar bracket groups based on their local contextual information, the corpus is automatically labeled with some nonterminal labels, and consequently a grammar with conditional probabilities is acquired. The statistical parsing model provides a framework for finding the most likely parse of a sentence based on these conditional probabilities. Experiments using Wall Street Journal data show that our approach achieves a relatively high accuracy: 88 % recall, 72 % precision and 0.7 crossing brackets per sentence for sentences shorter than 10 words, and 71 % recall, 51 % precision and 3.4 crossing brackets for sentences between 10-19 words. This result supports the assumption that local contextual statistics obtained from an unlabeled bracketed corpus are effective for learning a useful grammar and parsing.

KEY WORDS – Statistical Parsing, Grammar Acquisition, Clustering Analysis, Local Contextual Information

บทคัดย่อ – บทความนี้นำเสนอวิธีการเรียนกฎไวยากรณ์ที่มีค่าความน่าจะเป็นของการใช้กฎตามเนื้อความ(context) ที่อยู่รอบ การเรียนรู้นี้ใช้ชุดประโยค(corpus)ที่มีโครงสร้างของประโยคแต่ไม่มีการใส่ Label บอกว่าโครงสร้างนั้นคืออะไร นอกจากนี้บทความนี้ยังเสนอโมเดลการวิเคราะห์ประโยคเชิงโครงสร้าง (parsing) ที่ใช้กฎไวยากรณ์ที่คำนวณได้จากข้างต้น เพื่อทำการจัดลำดับความเป็นไปได้ของผลลัพธ์แต่ละอันที่ได้จากการวิเคราะห์โดยใช้ชุดประโยคภาษาอังกฤษขนาดใหญ่ของ Wall Street Journal มาใช้ในการทดลอง เราพบว่าวิธีการที่เสนอสามารถวิเคราะห์ประโยคได้อย่างถูกต้องสูงโดยมีระดับความถูกต้องอยู่ที่ 88 % recall 72 % precision และ 0.7 crossing brackets ต่อประโยค ในกรณีประโยคที่สั้นกว่า 10 คำ ส่วนสำหรับประโยคที่ยาว 10-19 คำ จะมีความถูกต้องอยู่ที่ระดับ 71 % recall, 51 % precision และ 3.4 crossing brackets ต่อประโยค ผลที่ได้นี้สนับสนุนสมมุติฐานที่ว่าเนื้อความใกล้ที่อยู่รอบ (local context) มีประโยชน์และประสิทธิผลในการช่วยให้สามารถเรียนรู้กฎไวยากรณ์อันจะเป็นประโยชน์ในการวิเคราะห์ประโยคด้วย

คำสำคัญ – การวิเคราะห์ประโยชน์ตามสถิติ, การเรียนรู้กฎไวยากรณ์, การวิเคราะห์โดยแบ่งกลุ่ม, ข้อมูลเนื้อความใกล้ที่อยู่รอบ

1. Introduction

Most natural language processing systems utilize grammars for parsing sentences in order to recognize their structure and finally to understand their meaning. Due to the difficulty and complexity of constructing a grammar by hand, there were several approaches developed for automatically training grammars from a large corpus with some probabilistic models. These methods can be characterized by properties of the corpus they used,

such as whether it includes information of brackets, lexical labels, nonterminal labels and so on.

Recently several parsed corpora which include full bracketing, tagging and nonterminal labels have been available for researchers to use for constructing a probabilistic grammar Black, E., Jelinek, F., Lafferty, J., Magerman, D.M., Mercer, R. and Roukos, S., "Towards history-based grammars: Using richer models for probabilistic parsing", *Proceedings of 1992 DARPA Speech and Natural Language Workshop*, 1992, pp. 134-139Collins, M.J., "A new statistical parser based on bigram lexical dependencies", *Proceedings of 34th Annual Meeting of the ACL* , 1996, pp. 184-191Magerman, D.M. and Marcus, M.P., "Pearl: a probabilistic chart parser", *Proceedings of the European ACL Conference*, 1991Magerman, D.M., "Statistical decision-tree models for parsing", *Proceedings of 33rd Annual Meeting of the ACL* , 1995, pp. 276-283. Most researches on these grammars calculate statistics of a grammar from a fully-parsed corpus with nonterminal labels and apply them to rank the possible parses of a sentence. While these researches report some promising results, it seems a hard task for a corpus builder to determine nonterminal labels for a corpus in comparison with annotating brackets and lexical labels, and the way to assign a nonterminal label to each constituent in the parsed sentence is usually ad hoc and arbitrary. From this point, it seems worth inferring a grammar from corpora without nonterminal labels. Moreover, compared with corpora including nonterminal labels, there are more existing corpora which include bracketings without nonterminal labels such as EDR corpus EDR, "EDR Electronic Dictionary User's Manual" (in Japanese), *Japan Electronic Dictionary Research Institute*, 1994, Ed. 2. and ATIS spoken language corpus Hemphill, C.T., Godfrey, J.J., Doddington, G.R. 'The ATIS spoken language systems pilot corpus", *Proceedings of 1990 DARPA Speech and Natural Language Workshop*, 1990. The well-known standard method to infer a probabilistic context-free grammar from a bracketed/unbracketed corpus without nonterminal labels is so-called inside-outside algorithm which was originally proposed by Baker Baker, J.K., "Trainable grammars for speech recognition", *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America (Klatt, D.H. and Wolf, J.J. eds.)*, 1979, pp. 547-550 and was implemented as applications for speech and language in Lari, K. and Young, S.J. "The estimation of stochastic context-free grammars using the inside-outside algorithm", *Computer speech and recognition*, Vol. 4, 1990, pp. 35-56, Pereira, F. and Schabes, Y., "Inside-outside reestimation from partially bracketed corpora", *Proceedings of 30th Annual Meeting of the ACL* , 1992, pp. 128-135 and Schabes, Y., Roth M. and Osborne, R., "Parsing the Wall Street Journal with the inside-outside algorithm", *Proceedings of 6th European Chapter of ACL* , 1993, pp. 341-347. Although encouraging results were shown in these works, the derived grammars were restricted to Chomsky normal-form CFGs and there were problems of the small size of acceptable training corpora and the relatively high computation time required for training the grammars.

Towards the problems, this paper proposes a new method which can learn a standard CFG with less computational cost by adopting techniques of clustering analysis to construct a context-sensitive probabilistic grammar from a bracketed corpus where nonterminal labels are not annotated. Another claim of this paper is that statistics from a large bracketed corpus without nonterminal labels combined with clustering techniques can help us construct a probabilistic grammar which produces an accurate natural language statistical parser. In this method,

nonterminal labels for brackets in a bracketed corpus can be automatically assigned by making use of local contextual information which is defined as a set of category pairs of left and right words of a constituent in the phrase structure of a sentence. In this research, based on the assumption that not all contexts are useful in every case, effectiveness of contexts is also investigated. By using only effective contexts, it is possible for us to improve training speed and memory space without a sacrifice of accuracy. Finally, a statistical parsing model based on the acquired grammar is provided and the performance is shown through some experiments using the WSJ corpus.

2. Grammar Acquisition as Clustering Process

In the past, Theeramunkong Theeramunkong, T. and Okumura, M., "Towards automatic grammar acquisition from a bracketed corpus", *Proceedings of 4th International Workshop on Very Large Corpora*, 1996, pp. 168-177 proposed a method of grouping brackets in a bracketed corpus (with lexical tags but no nonterminal labels), according to their local contextual information, as a first step towards the automatic acquisition of a context-free grammar. The basic idea is to apply clustering analysis to find out a number of groups of similar brackets in the corpus and then to assign each group with a same nonterminal label. Clustering analysis is a generic name of a variety of mathematical methods that can be used to find out which objects in a set are similar. Its applications on natural language processing are varied such as in areas of word classification, text categorization and so on Iwayama, M. and Tokunaga, T., "Hierarchical bayesian clustering for automatic text classification", *IJCAI*, 1995, pp. 1322-1327Pereira, F., Tishby, N. and Lee, L. "Distributional clustering of English words", *Proceedings of 31st Annual Meeting of the ACL* 1993, pp. 183-190. However, there is still few researches which apply clustering analysis for grammar inference and parsing Mori, S. and Nagao M., "Parsing without grammar", *Proceedings of the 4th International Workshop on Parsing Technologies*, 1995, pp. 174-185. This section gives an explanation of grammar acquisition based on clustering analysis. In the first place, let us consider the following example of the parse structures of two sentences in the corpus in Figure 1.

Sentence (1) : A big man slipped on the ice.
Parse Tree (1) : (((DT,"a")((JJ,"big")(NN,"man")))
((VB,"slipped")((IN,"on")
((DT,"the")(NN,"ice")))))
Sentence (2) : The boy dropped his wallet somewhere.
Parse Tree (2) : (((DT,"the")(NN,"boy"))
(((VB,"dropped")((PRP\$,"his")
(NN,"wallet")))
(RB,"somewhere")))

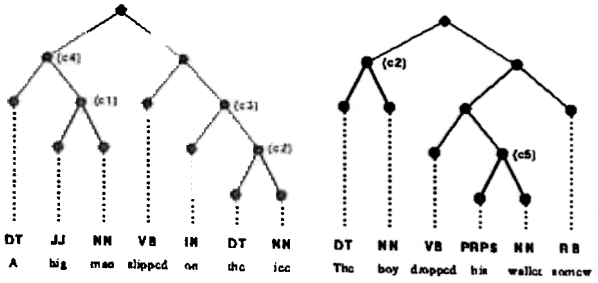

Figure 1. The graphical representation of the parse structures of "A big man slipped on the ice" and "the boy dropped his wallet somewhere"

In the parse structures, leaf nodes are given tags while there is no label for intermediate nodes. Note that each node corresponds to a bracket in the corpus. With this corpus, the grammar learning task corresponds to a process to determine the label for each intermediate node. In other words, this task is concerned with the way to cluster the brackets into some certain groups based on their similarity and give each group a label. For instance, in figure 1, it is reasonable to classify the brackets (c2),(c4) and (c5) into a same group and give them a same label (e.g., NP(noun phrase)). As the result, we obtain three grammar rules: *NP → (DT)(NN), NP → (PRP$)(NN) and NP → (DT)(c1)*. To do this, the grammar acquisition algorithm operates in five steps as follows.

Assign a unique label to each node of which lower nodes are assigned labels. At the initial step, such node is one whose lower nodes are lexical categories. For example, in figure 1, there are three unique labels derived: $c1 \rightarrow$ *(JJ)(NN)*, $c2 \rightarrow$ *(DT)(NN)* and $c5 \rightarrow$ *(PRP$)(NN)*. This process is performed throughout all parse trees in the corpus.
Calculate the similarity of every pair of the derived labels.
Merge the most similar pair to a single new label(i.e., a label group) and recalculate the similarity of this new label with other labels.

4. Repeat (3) until a termination condition is detected. Finally, a certain set of label groups is derived.
5. Replace labels in each label group with a new label in the corpus. For example, if *(DT)(NN)* and *(PRP$)(NN)* are in the same label group, we replace them with a new label (such as *NP*) in the whole corpus.
6. Repeat (1)-(5) until all nodes in the corpus are assigned labels.

To compute the similarity of labels, the concept of local contextual information is applied. In this work, the local contextual information is defined as categories of the words immediately before and after a label. This information is shown to be powerful for acquiring phrase structures in a sentence in Brill, E., "Automatically acquiring phrase structure using distributional analysis", *Proceedings of Speech and Natural Language Workshop*, 1992, pp. 155-159. In our preliminary experiments, we also found out that the information are potential for characterizing constituents in a sentence.

2.1 Distributional Similarity

While there are a number of measures which can be used for representing the similarity of labels in the step 2, measures which make use of relative entropy (Kullback-Leibler distance) are practical interest and scientific. One of these measures is divergence which has a symmetrical property. Its application on natural language processing was firstly proposed by Harris Harris, Z., "Structural Linguistics", Chicago, University of Chicago Press, 1951 and was shown successfully for detecting phrase structures in Brill, E., "Automatically acquiring phrase structure using distributional analysis", *Proceedings of Speech and Natural Language Workshop*, 1992, pp. 155-159Pereira, F., Tishby, N. and Lee, L., "Distributional clustering of English words", *Proceedings of 31st Annual Meeting of the ACL*, 1993, pp. 183-190. Basically, divergence, as well as relative entropy, is not exactly similarity measure instead it indicates distributional dissimilarity. That means the large value it gets, the less similarity it means. The detail of divergence is illustrated below.

Let P_{c1} and P_{c2} be two probability distributions of labels *c1* and *c2* over contexts, *CT*. The relative entropy between P_{c1} and P_{c2} is:

$$D(P_{c1}||P_{c2}) = \sum_{e \in CT} p(e|c1) \times \log (p(e|c1)/p(e|c2))$$

Relative entropy $D(P_{c1}||P_{c2})$ is a measure of the amount of extra information beyond P_{c1} needed to describe P_{c2}. The divergence between P_{c1} and P_{c2} is defined as $D(P_{c1}||P_{c2})+D(P_{c2}||P_{c1})$, and is a measure of how difficult it is to distinguish between the two distributions. The context is defined as a pair of words immediately before and after a label(bracket). Any two labels are considered to be identical when they are distributionally similar, i.e., the divergence is low. From the practical point view, this measure addresses a problem of sparseness in limited data. Particularly, when $p(e|c2)$ is zero, we cannot calculate the divergence of two probability distributions because the denominator becomes zero. To cope with this problem, the original probability can be modified by a popular technique into the following formula.

$$p(e|ci) = \lambda(N(ci,e)/N(c\_i)) + (\ \lambda)(1/|CT|)$$

where, $N(ci)$ and $N(ci,e)$ are the occurrence frequency of *ci* and *(ci,e)*, respectively. $|CT|$ is the number of possible contexts and λ is an interpolation coefficient. As defining contexts by the left and right lexical categories, $|CT|$ is the square of the number of existing lexical categories. In the formula, the first term means the original estimated probability and the second term expresses a uniform distribution, where the probability of all events is estimated to a fixed uniform number. λ is applied as a balancing weight between the observed distribution and the uniform distribution. In our experimental results, λ is assigned with a value of 0.6 which seems to make a good estimate.

2.2 Termination Condition

During iteratively merging the most similar labels, all labels will finally be gathered to a single group. Due to this, a criterion is needed for determining whether this merging process should be continued or terminated. In this section, we describe a criterion named *differential entropy* which is a measure of entropy (perplexity) fluctuation before and after merging a pair of labels. Let *c1* and *c2* be the most similar pair of labels. Also let *c3* be the result label. $p(e|c1)$, $p(e|c2)$ and $p(e|c3)$ are probability distributions over contexts *e* of *c1*, *c2* and *c3*, respectively. $p(c1)$, $p(c2)$ and $p(c3)$ are estimated probabilities of *c1*, *c2* and *c3*, respectively. The differential entropy (*DE*) is defined as follows.

$$DE = \textit{Consequence Entropy} - \textit{Previous Entropy}$$
$$= -p(c3) \times \Sigma_e\, p(e|c3) \log p(e|c3)$$
$$+p(c1) \times \Sigma_e\, p(e|c1) \log p(e|c1)$$
$$+p(c2) \times \Sigma_e\, p(e|c2) \log p(e|c2)$$

where $\Sigma_e\, p(e|ci) \log p(e|ci)$ is the total entropy over various contexts of label *ci*. The larger *DE* is, the larger the information fluctuation before and after merging becomes. In general, a small fluctuation is preferred to a larger one because when *DE* is large, the current merging process introduces a large amount of information fluctuation and its reliability becomes low.

3. Local Context Effective

As the similarity of any two labels is estimated based on local contextual information which is defined by a set of category pairs of left and right words, there is an interesting question of which contexts are useful for calculation of similarity. In the past, effectiveness of contexts is indicated in some previous researches Bartell, B.T., Cottrell, G.W. and Belew, R.K., "Representing documents using an explicit model of their similarities", *Journal of the American Society for Information Science*, Vol. 46, No. 4, 1995, pp. 254-271. One of suitable measures for representing effectiveness of a context is dispersion of the context on labels. This measure expresses that the number of useful contexts should be diverse for different labels. From this, the effectiveness (*E*) of a context (*c*) can be defined using variance as follow:

$$E(c) \quad \Sigma_{a \in A} ((N(a,c)-N^*(c))^2 / |A|$$
$$N^*(c) \quad \Sigma_{a \in A} N(a,c) / |A|$$

where *A* is a set of all labels and *a* is one of its individual member. $N(a,c)$ is the number of times a label *a* and a context *c* are cooccurred. $N^*(c)$ is an averaged value of $N(a,c)$ on a label *a*. In order to take large advantage of context in clustering, it is preferable to choose a context c with a high value of E(c) because this context trends to have a high discrimination for characterizing labels. Ranking the contexts by the effectiveness value *E*, some rank higher contexts are selected for clustering the labels instead of all contexts. This enables us to decrease computation time and space without sacrificing the accuracy of the clustering results and sometimes also helps us to remove some noises due to useless contexts. Some experiments were done to support this assumption and their results are shown in the next section.

4. Statistical Parsing Model

This section describes a statistical parsing model which takes a sentence as input and produce a phrase-structure tree as output. 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. While a model of a simple probabilistic 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 follow.

$$P(T|S) = \Pi_{(ri,ci) \in T}\; P(ri,ci)$$

where *ri* is an application rule in the tree and *ci* is the left and right contexts at the place the rule is applied. Similar to most probabilistic models and our clustering process, 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, D.M. and Marcus, M.P., "Pearl: a probabilistic chart parser", *Proceedings of the European ACL Conference*, 1991. The modified model is:

$$P(T|S)=(\Pi_{(ri,ci)\in T}(\alpha*P(ri,ci)+(1-\alpha)*(1/(NrNc)))^{1/|T|}$$

Here, α is a balancing weight between the observed distribution and the uniform distribution and it is assigned with 0.95 in our experiments. The applied parsing algorithm is a simple bottom-up chart parser whose scoring function is based on this model. The grammar used is one trained by the algorithm described in section **Grammar Acquisition as Clustering Process**. A dynamic programming algorithm is used: 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.

5. Experimental Evaluation

To give some support to our suggested grammar acquisition method and statistical parsing model, three following evaluation experiments are made. The experiments use texts from the Wall Street Journal (WSJ) Corpus and its bracketed version provided by the Penn Treebank. Out of nearly 49,000 sentences (1,222,065 words), we extracted 48,000 sentences (1,172,710 words) as possible material source for training a grammar and 2000 sentences (49,355 words) as source for testing.

The first experiment involves an evaluation of performance of our proposed grammar learning method shown in the section **Grammar Acquisition as Clustering Process**. In this preliminary experiment, only rules which have lexical categories as their right hand side are considered and the acquired nonterminal labels are compared with those assigned in the WSJ corpus. The second experiment stands for investigating effectiveness of contexts described in section **Local Context Effective**. The purpose is to find out useful contexts and use them instead of all contexts based on the assumption that not all contexts are useful for clustering brackets in grammar acquisition. Reducing the number of contexts will help us to improve the computation time and space. The last experiment is carried out for evaluating the whole grammar which is learned based on local contextual information and indicating the performance of our statistical parsing model using the acquired grammar. The measures used for this evaluation are bracketing recall, precision and crossing.

5.1 Evaluation of Clustering in Grammar Acquisition

This subsection shows some results of our preliminary experiments to confirm effectiveness of the proposed grammar acquisition techniques. The grammar is learned from the WSJ bracketed corpus where all nonterminals are omitted. In this experiment, we focus on only the rules with lexical categories as their right hand side. For instance, *c1 →(JJ)(NN)*, *c2→(DT)(NN)* and *c5→(PRP$)(NN)* in figure 1. Due to the reason of computation time and space, we use the rule tokens which appear more than 500 times in the corpus. The number of initial rules is 51. From these rules, the most similar pair is calculated and merged to a new label. The merging process is carried out in iterative way. In each iterative step of the merging process, differential entropies are calculated. During the merging process, there are some sharp peaks indicating the rapid fluctuation of entropy. These sharp peaks can be used as a step to terminate the merging process. In the experiments, a peak with $DE > 0.12$ is applied. As the result, the process is halted up at the 45th step and 6 groups are obtained.

This result is evaluated by comparing the system's result with nonterminal symbols given in the WSJ corpus. The evaluation method utilizes a contingency table model which is introduced in Swets, J.A., "Effectiveness of information retrieval methods", *American Documentation*, Vol. 20, 1969, pp. 72-89 and widely used in Information Retrieval and
Psychology Agarwal, Rajeev, "Evaluation of semantic clusters", *Proceedings of 33rd Annual Meeting of the ACL* , 1995, pp. 284-286Iwayama, M. and Tokunaga, T., "Hierarchical bayesian clustering for automatic text classification", *IJCAI*, 1995, pp. 1322-1327. The following measures are considered.

Positive Recall (PR) $a/(a$
Positive Precision (PP) $a/(a+b)$
Negative Recall (NR) $d/(b+d)$
Negative Precision (NP) $d/(c+d)$
F-measure(FM) : $((\beta^2+ \times PP \times PR)/(\beta^2. PP+PR)$

where a is the number of the label pairs which the WSJ corpus assigns in the same group and so does the system, b is the number of the pairs which the WSJ corpus does not assign in the same group but the system does, c is the number of the pairs which the WSJ assigned but the system does not, and d is the number of the pairs which both the WSJ and the system does not assign in the same group. The F-measure is used as a combined measure of recall and precision, where β is the weight of recall relative to precision. Here, we use $\beta = 1.0$, equal weight.

The result shows 0.93 % PR, 0.93 % PP, 0.92 % NR, 0.92 % NP and 0.93 % FM, which are all relatively good values. Especially, PP shows that almost all same labels in the WSJ are assigned in same groups. In order to investigate whether the application of differential entropy to cut off the merging process is appropriate, we plot values of these measures at all merging steps as shown in figure 2. From the graphs, we found out that the best solution is located at around 44th-45th merging step. This is consistent with the grouping result of our approach. Moreover, the precision equals 100 % from 1st-38nd steps, indicating that the merging process is suitable.

5.2 Checking Context Effectiveness

As another experiment, we examine effectiveness of contexts in the clustering process in order to reduce the computation time and space. Variance is used for expressing effective of a context. The assumption is that a context with has the highest variance is the most effective. The experiment is done by selecting the top N of contexts and use it instead of all contexts in the clustering process.

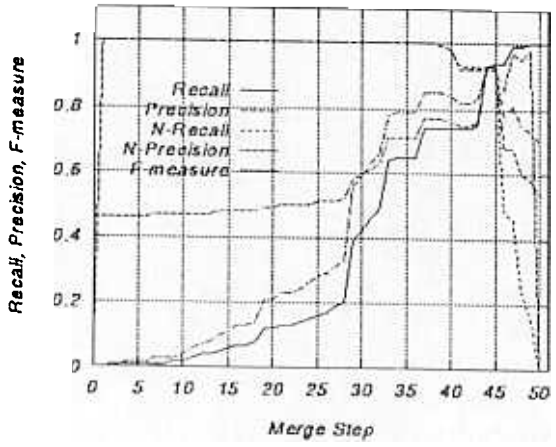

Figure 2. The transition of PR, PP, NR, NP and FM during the merging process

Besides cases of N = 10, 50, 200, 400 and all (2401), a case that 200 contexts are randomly chosen from all contexts, is taken into account in order to examine the assumption that variance is efficient. In this case, 3 trials are made and the average value is employed. Due to the limit of paper space, we show only F-measure in Figure 3. The graphs tell us that the case of top 200 seems superior to the case of 200 random contexts in all merging step. This means that variance seems to be a good measure for selecting a set of effective contexts in the clustering process.

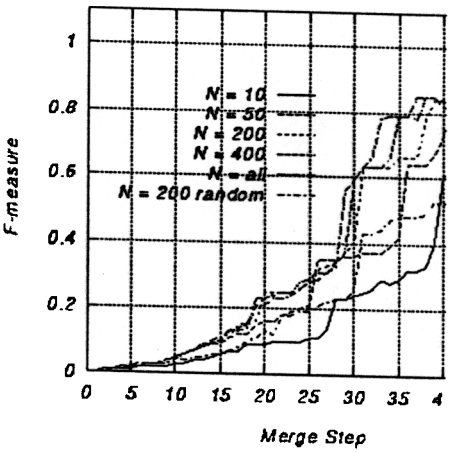

Figure 3. The transition of PR, PP, NR, NP and FM during the merging process

Furthermore, we can observe that a high accuracy can be achieved even if not all contexts are taken into account. From this result, the best F-measures are all 0.93 and the number of groups are 2, 5, 5 and 6 for each case, i.e., 10, 50, 200 and 400. Excepts the case of 10, all cases shows a good result compared with all contexts (0.93, 6 groups). This result tells us that it is reasonable to select contexts with large values of variance to ones with small variance and a relatively large number of contexts are enough for the clustering process. By preliminary experiments, we found out that the following criterion is sufficient for determining the number of contexts. Contexts are selected in the order of their variance and a context will be accepted when its variance is more than 10 % of the average variance of the previous contexts.

5.3 Performance of Statistical Parsing Model

Utilizing top N contexts, we learn the whole grammar based on the algorithm given in section **Grammar Acquisition as Clustering Process**. Brackets (rules) which are occurred more than 40 times in the corpus are considered and the number of contexts used is determined by the criterion described in the previous subsection. As the result of the grammar acquisition process, 1396 rules are acquired. These rules are attached with the conditional probability based on contexts (the left and right categories of the rules). The chart parser tries to find the best parse of the sentence. 48,000 sentences are used for training a grammar and 2000 sentences are for a test set. To evaluate the performance, the PARSEVAL measures as defined in Black, E. and al., et, "A procedure for quantitatively comparing the syntactic coverage of English grammars", *Proceedings of 1991 DARPA Speech and Natural Language Workshop*, 1991, pp. 306-311 are used:

$$\text{Precision} = \frac{\text{number of correct brackets in proposed parses}}{\text{number of brackets in proposed parses}}$$

$$\text{Recall} = \frac{\text{number of correct brackets in proposed parses}}{\text{number of brackets in treebank parses}}$$

The parser generates the most likely parse based on context-sensitive condition probability of the grammar. Among 2000 test sentences, only 1874 sentences can be parsed owing to two following reasons: (1) our algorithm considers rules which occur more than 40 times in the corpus, (2) test sentences have different characteristics from training sentences. Table 1 displays the detail results of our statistical parser evaluated against the WSJ corpus.

93 % of sentences can be parsed with 71 % recall, 52 % precision and 4.5 crossings per sentence. For short sentences (3-9 words), the parser achieves up to 88 % recall and 71 % precision with only 0.71 crossings. For moderately long sentences (10-19 and 20-30 words), it works with 60-71 % recall and 41-51 % precision. From this result, the proposed parsing model is shown to succeed with high bracketing recalls to some degree. Although our parser cannot achieve good precision, it is not so a serious problem because our parser tries to give more detail bracketing for a sentence than that given in the WSJ corpus. In the next section, the comparison with other researches will be discussed.

Table 1. Parsing accuracy using the WSJ Corpus

Sent. Length	3-9	3-15	10-19	20-30	3-40
Comparisons	393	988	875	484	1862
Avg. Sent. Len.	7.0	10.3	14.0	24.0	16.33
TBank Parses	4.81	6.90	9.37	15.93	10.85
System's Parse	10.86	16.58	23.14	40.73	27.18
Crossings/Sent.	0.72	1.89	3.36	7.92	4.52
Sent. Cross=0 (%)	56.7	33.1	13.6	2.5	19.0
Sent. Cross≤1 (%)	79.4	50.4	25.4	6.0	30.3
Sent. Cross≤2 (%)	93.4	67.0	41.5	9.5	41.8
Recall	88.2	79.3	71.2	59.7	70.8
Precision	71.9	60.6	51.3	41.2	52.1

6. Related Works and Discussion

In this section, our approach is compared with some previous interesting methods. These methods can be classified into non-grammar-based and grammar-based approaches. For non-grammar-based approaches, the most successful probabilistic parser named SPATTER is proposed by Magerman Magerman, D.M., "Statistical decision-tree models for parsing", *Proceedings of 33rd Annual Meeting of the ACL*, 1995, pp. 276-283. The parser is constructed by using decision-tree learning techniques and can succeed up to 86-90 % of bracketing accuracy(both recall and precision) when training with the WSJ corpus, a fully-parsed corpus with nonterminal labels. Later Collins Collins, M.J., "A new statistical parser based on bigram lexical dependencies", *Proceedings of 34th Annual Meeting of the ACL*, 1996, pp. 184-191 introduced a statistical parser which is based on probabilities of bigram dependencies between head-words in a parse tree. At least the same accuracy as SPATTER was acquired for this parser. These two methods utilized a corpus which includes both lexical categories and nonterminal categories. However, it seems a hard task to assign nonterminal labels for a corpus and the way to assign a nonterminal label to each constituent in the parsed sentence is arduous and arbitrary. It follows that it is worth trying to infer a grammar from corpora without nonterminal labels.

One of the most promising results of grammar inference based on grammar-based approaches is the inside-outside algorithm proposed by Lari, K. and Young, S.J., "The estimation of stochastic context-free grammars

using the inside-outside algorithm", *Computer speech and recognition*, Vol. 4, 1990, pp. 35-56 to construct the grammar from unbracketed corpus. This algorithm is an extension of forward-backward algorithm which infers the parameters of a stochastic context-free grammar. In this research the acquired grammar is evaluated based on its entropy or perplexity where the accuracy of parsing is not taken into account. As another research, Pereira and Schabes Pereira, F. and Schabes, Y., "Inside-outside reestimation from partially bracketed corpora", *Proceedings of 30th Annual Meeting of the ACL*, 1992, pp. 128-135Schabes, Y., Roth M. and Osborne, R., "Parsing the Wall Street Journal with the inside-outside algorithm", *Proceedings of 6th European Chapter of ACL*, 1993, pp. 341-347 proposed a modified method to infer a stochastic grammar from a partially parsed corpus and evaluated the results with a bracketed corpus. This approach gained up to around 90 % bracketing recall for short sentences(0-15 words) but it suffered with a large amount ambiguity for long ones(20-30) where 70 % recall is gained. The acquired grammar is normally in Chomsky normal-form which is a special case of grammar although he claimed that all of CFGs can be in this form. This type of the grammar makes all output parses of this method be in the form of binary-branching trees and then the bracketing precision cannot be taken into account because correct parses in the corpus need not be in this form. On the other hand, our proposed approach can learn a standard CFG with 88 % recall for short sentences and 60 % recall for long ones. This result shows that our method gets the same level of accuracy as the inside-outside algorithm does. However, our approach can learn a grammar which is not restricted to Chomsky normal-form and performs with less computational cost compared with the approaches applying the inside-outside algorithm.

7. Conclusion

In this paper, we proposed a method of applying clustering analysis to learn a context-sensitive probabilistic grammar from an unlabeled bracketed corpus. Supported by some experiments, local contextual information which is left and right categories of a constituent was shown to be useful for acquiring a context-sensitive conditional probability context-free grammar from a corpus. A probabilistic parsing model using the acquired grammar was described and its potential was examined. Through experiments, our parser can achieve high parsing accuracy to some extent compared with other previous approaches with less computational cost. As our further work, there are still many possibilities for improvement which are encouraging. For instance, it is possible to use lexical information and head information in clustering and constructing a probabilistic grammar.

Reference

[1] Agarwal, Rajeev, "Evaluation of semantic clusters", *Proceedings of 33rd Annual Meeting of the ACL*, 1995, pp. 284-286

[2] Baker, J.K., "Trainable grammars for speech recognition", *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America (Klatt, D.H. and Wolf, J.J. eds.)*, 1979, pp. 547-550

[3] Bartell, B.T., Cottrell, G.W. and Belew, R.K., "Representing documents using an explicit model of their similarities", *Journal of the American Society for Information Science*, Vol. 46, No. 4, 1995, pp. 254-271

[4] Black, E. and al., et, "A procedure for quantitatively comparing the syntactic coverage of English grammars", *Proceedings of 1991 DARPA Speech and Natural Language Workshop*, 1991, pp. 306-311

[5] Black, E., Jelinek, F., Lafferty, J., Magerman, D.M., Mercer, R. and Roukos, S., "Towards history-based grammars: Using richer models for probabilistic parsing", *Proceedings of 1992 DARPA Speech and Natural Language Workshop*, 1992, pp. 134-139

[6] Brill, E., "Automatically acquiring phrase structure using distributional analysis", *Proceedings of Speech and Natural Language Workshop*, 1992, pp. 155-159

[7] Collins, M.J., "A new statistical parser based on bigram lexical dependencies", *Proceedings of 34th Annual Meeting of the ACL*, 1996, pp. 184-191

[8] EDR, "EDR Electronic Dictionary User's Manual" (in Japanese), *Japan Electronic Dictionary Research Institute*, 1994, Ed. 2.1

[9] Harris, Z., "Structural Linguistics", Chicago, University of Chicago Press, 1951

[10] Hemphill, C.T., Godfrey, J.J., Doddington, G.R., "The ATIS spoken language systems pilot corpus", *Proceedings of 1990 DARPA Speech and Natural Language Workshop*, 1990

[11] Iwayama, M. and Tokunaga, T., "Hierarchical bayesian clustering for automatic text classification", *IJCAI*, 1995, pp. 1322-1327

[12] Lari, K. and Young, S.J., "The estimation of stochastic context-free grammars using the inside-outside algorithm", *Computer speech and recognition*, Vol. 4, 1990, pp. 35-56

[13] Magerman, D.M. and Marcus, M.P., "Pearl: a probabilistic chart parser", *Proceedings of the European ACL Conference*, 1991

[14] Magerman, D.M., "Statistical decision-tree models for parsing", *Proceedings of 33rd Annual Meeting of the ACL* , 1995, pp. 276-283

[15] Mori, S. and Nagao M., "Parsing without grammar", *Proceedings of the 4th International Workshop on Parsing Technologies*, 1995, pp. 174-185

[16] Pereira, F. and Schabes, Y., "Inside-outside reestimation from partially bracketed corpora", *Proceedings of 30th Annual Meeting of the ACL* , 1992, pp. 128-135

[17] Pereira, F., Tishby, N. and Lee, L., "Distributional clustering of English words", *Proceedings of 31st Annual Meeting of the ACL* , 1993, pp. 183-190

[18] Schabes, Y., Roth M. and Osborne, R., "Parsing the Wall Street Journal with the inside-outside algorithm", *Proceedings of 6th European Chapter of ACL* , 1993, pp. 341-347

[19] Swets, J.A., "Effectiveness of information retrieval methods", *American Documentation*, Vol. 20, 1969, pp. 72-89

[20] Theeramunkong, T. and Okumura, M., "Towards automatic grammar acquisition from a bracketed corpus", *Proceedings of 4th International Workshop on Very Large Corpora*, 1996, pp. 168-177

Thanaruk THEERAMUNKONG

Born in 1967. He received a bachelor degree in Electric and Electronics and master and doctor degrees in Computer Science from Tokyo Institute of Technology in 1990, 1992, and 1995 respectively. From 1995 to 1997, he worked as research associates at Graduate School of Information Science, Japan Advanced Institute of Science and Technology. At present, he is working as assistant professor at Sirindhorn International Institute of Technology. His current research interests include natural language processing, machine learning, corpus-based processing, information retrieval and database technology.