Committee-based Decision Making in Probabilistic Partial Parsing

INUI Takashi* and **INUI Kentaro***†

$*$ Department of Artificial Intelligence, Kyushu Institute of Technology
† PRESTO, Japan Science and Technology Corporation
{t_inui,inui}@pluto.ai.kyutech.ac.jp

Abstract

This paper [1] explores two directions for the next step beyond the state of the art of statistical parsing: probabilistic partial parsing and committee-based decision making. Probabilistic partial parsing is a probabilistic extension of the existing notion of partial parsing, which enables fine-grained arbitrary choice on the trade-off between accuracy and coverage. Committee-based decision making is to combine the outputs from different systems to make a better decision. While various committee-based techniques for NLP have recently been investigated, they would need to be further extended so as to be applicable to probabilistic partial parsing. Aiming at this coupling, this paper gives a general framework to committee-based decision making, which consists of a set of weighting functions and a combination function, and discusses how it can be coupled with probabilistic partial parsing. Our experiments have so far been producing promising results.

1 Introduction

There have been a number of attempts to use statistical techniques to improve parsing performance. While this goal has been achieved to a certain degree given the increasing availability of large tree banks, the remaining room for the improvement appears to be getting saturated as long as only statistical techniques are taken into account. This paper explores two directions for the next step beyond the state of the art of statistical parsing: probabilistic partial parsing and committee-based decision making.

Probabilistic partial parsing is a probabilistic extension of the existing notion of partial parsing (e.g. (Jensen et al., 1993)) where a parser selects as its output only a part of the parse tree that are probabilistically highly reliable. This decision-making scheme enables a fine-grained arbitrary choice on the trade-off between accuracy and coverage. Such trade-off is important since there are various applications that require reasonably high accuracy even sacrificing coverage. A typical example is the paraphrasing task embedded in summarization, sentence simplification (e.g. (Carroll et al., 1998)), etc. Enabling such trade-off choice will make state-of-the-art parsers of wider application. Partial parsing has also been proven useful for bootstrapping learning.

One may suspect that the realization of partial parsing is a trivial matter in probabilistic parsing just because a probabilistic parser inherently has the notion of "reliability" and thus has the trade-off between accuracy and coverage. However, there has so far been surprisingly little research focusing on this matter and almost no work that evaluates statistical parsers according to their coverage-accuracy (or recall-precision) curves. Taking the significance of partial parsing into account, therefore in this paper, we evaluate parsing performance according to coverage-accuracy curves.

Committee-based decision making is to combine the outputs from several different systems (e.g. parsers) to make a better decision. Recently, there have been various attempts to ap-

[1] [note for revision] The results of the experiments reported in the original version of this paper have been found partly invalid: While the (total) accuracy achieved by KNP (Kurohashi et al., 1994) for our test data is wrongly reported to be 87.25%, the actual accuracy it achieves is 91.25%. We apologize deeply to the creators of the parser for assessing the performance of KNP improperly. This paper is a revised version of the paper, in which every invalid description is annotated with a footnote. **Readers are NOT expected to refer to any result related to KNP shown in the body of this paper.**

ply committee-based techniques to NLP tasks such as POS tagging (Halteren et al., 1998; Brill et al., 1998), parsing (Henderson and Brill, 1999), word sense disambiguation (Pedersen, 2000), machine translation (Frederking and Nirenburg, 1994), and speech recognition (Fiscus, 1997). Those works empirically demonstrated that combining different systems often achieved significant improvements over the previous best system.

In order to couple those committee-based schemes with probabilistic partial parsing, however, one would still need to make a further extension. Aiming at this coupling, in this paper, we consider a general framework of committee-based decision making that consists of a set of weighting functions and a combination function, and discuss how that framework enables the coupling with probabilistic partial parsing. To demonstrate how it works, we report the results of our parsing experiments on a Japanese tree bank.

2 Probabilistic partial parsing

2.1 Dependency probability

In this paper, we consider the task of deciding the dependency structure of a Japanese input sentence. Note that, while we restrict our discussion to analysis of Japanese sentences in this paper, what we present below should also be straightforwardly applicable to more wide-ranged tasks such as English dependency analysis just like the problem setting considered by Collins (1996).

Given an input sentence s as a sequence of *Bunsetsu*-phrases (BPs)[2], $b_1\ b_2\ \ldots\ b_n$, our task is to identify their inter-BP dependency structure $R = \{r(b_i, b_j) | i = 1, \ldots, n\}$, where $r(b_i, b_j)$ denotes that b_i depends on (or modifies) b_j. Let us consider a *dependency probability* (DP), $P(r(b_i, b_j)|s)$, a probability that $r(b_i, b_j)$ holds in a given sentence s: $\forall i.\ \sum_j P(r(b_i, b_j)|s) = 1$.

2.2 Estimation of DPs

Some of the state-of-the-art probabilistic language models such as the bottomup models $P(R|s)$ proposed by Collins (1996) and Fujio et al. (1998) directly estimate DPs for a given input, whereas other models such as PCFG-based topdown generation models $P(R, s)$ do *not* (Charniak, 1997; Collins, 1997; Shirai et al., 1998). If the latter type of models were totally excluded from any committee, our committee-based framework would not work well in practice. Fortunately, however, even for such a model, one can still estimate DPs in the following way if the model provides the n-best dependency structure candidates coupled with probabilistic scores.

Let R_i be the i-th best dependency structure $(i = 1, \ldots, n)$ of a given input s according to a given model, and let $\mathcal{R}_H$ be a set of R_i. Then, $P(r(b_i, b_j)|s)$ can be estimated by the following approximation equation:

$$P(r(b_i, b_j)|s) \approx \frac{P^r_{\mathcal{R}_H}}{P_{\mathcal{R}_H}} \tag{1}$$

where $P_{\mathcal{R}_H}$ is the probability mass of $R \in \mathcal{R}_H$, and $P^r_{\mathcal{R}_H}$ is the probability mass of $R \in \mathcal{R}_H$ that supports $r(b_i, b_j)$. The approximation error ϵ is given by $\epsilon \leq \frac{P_{\mathcal{R}} - P_{\mathcal{R}_H}}{P_{\mathcal{R}}}$, where $P_{\mathcal{R}}$ is the probability mass of all the dependency structure candidates for s (see (Poole, 1993) for the proof). This means that the approximation error is negligible if $P_{\mathcal{R}_H}$ is sufficiently close to $P_{\mathcal{R}}$, which holds for a reasonably small number n in most cases in practical statistical parsing.

2.3 Coverage-accuracy curves

We then consider the task of selecting dependency relations whose estimated probability is higher than a certain threshold σ $(0 < \sigma \leq 1)$. When σ is set to be higher (closer to 1.0), the accuracy is expected to become higher, while the coverage is expected to become lower, and vice versa. Here, coverage C and accuracy A are defined as follows:

$$C = \frac{\#\text{ of the decided relations}}{\#\text{ of all the relations in the test set}} \tag{2}$$

$$A = \frac{\#\text{ of the correctly decided relations}}{\#\text{ of the decided relations}} \tag{3}$$

Moving the threshold σ from 1.0 down toward 0.0, one can obtain a coverage-accuracy curve (C-A curve). In probabilistic partial parsing, we evaluate the performance of a model according to its C-A curve. A few examples are shown in Figure 1, which were obtained

[2] A *bunsetsu* phrase (BP) is a chunk of words consisting of a content word (noun, verb, adjective, etc.) accompanied by some functional word(s) (particle, auxiliary, etc.). A Japanese sentence can be analyzed as a sequence of BPs, which constitutes an inter-BP dependency structure

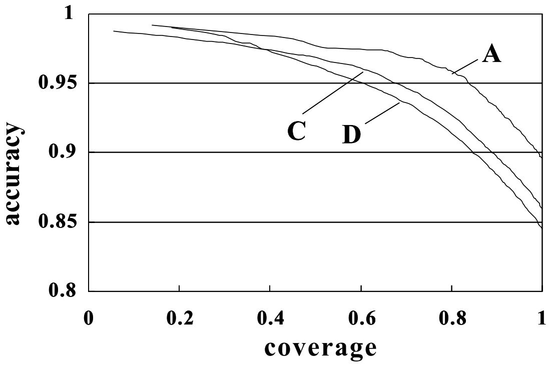

Figure 1: C-A curves

in our experiment (see Section 4). Obviously, Figure 1 shows that model A outperformed the other two. To summarize a C-A curve, we use the 11-point average of accuracy (11-point accuracy, hereafter), where the eleven points are $C = 0.5,\ 0.55, \ldots,\ 1.0$. The accuracy of total parsing corresponds to the accuracy of the point in a C-A curve where $C = 1.0$. We call it total accuracy to distinguish it from 11-point accuracy. Note that two models with equal achievements in total accuracy may be different in 11-point accuracy. In fact, we found such cases in our experiments reported below. Plotting C-A curves enable us to make a more fine-grained performance evaluation of a model.

3 Committee-based probabilistic partial parsing

We consider a general scheme of committee-based probabilistic partial parsing as illustrated in Figure 2. Here we assume that each committee member M_k $(k = 1, \ldots, m)$ provides a DP matrix $P_{M_k}(r(b_i, b_j)|s)$ $(b_i,\ b_j \in s)$ for each input s. Those matrices are called input matrices, and are given to the committee as its input.

A committee consists of a set of weighting functions and a combination function. The role assigned to weighting functions is to standardize input matrices. The weighting function associated with model M_k transforms an input matrix given by M_k to a weight matrix W_{M_k}. The majority function then combines all the given weight matrices to produce an output matrix O, which represents the final decision of the committee. One can consider various options for both functions.

3.1 Weighting functions

We have so far considered the following three options.

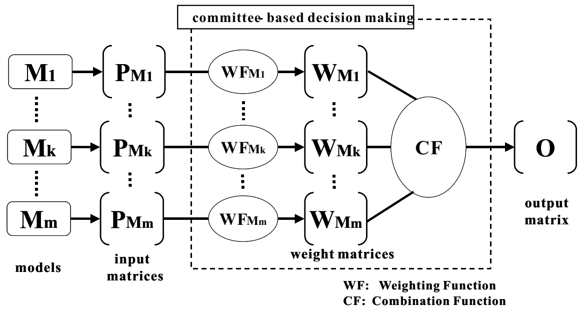

Figure 2: Committee-based probabilistic partial parsing

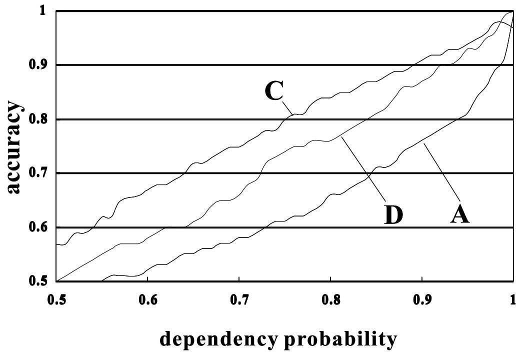

Figure 3: P-A curves

Simple The simplest option is to do nothing:

$$w_{ij}^{M_k} = P_{M_k}(r(b_i, b_j)|s) \tag{4}$$

where $w_{ij}^{M_k}$ is the (i, j) element of W_{M_k}.

Normal A bare DP may not be a precise estimation of the actual accuracy. One can see this by plotting probability-accuracy curves (P-A curves) as shown in Figure 3. Figure 3 shows that model A tends to overestimate DPs, while model C tends to underestimate DPs. This means that if A and C give different answers with the same DP, C's answer is more likely to be correct. Thus, it is not necessarily a good strategy to simply use given bare DPs in weighted majority. To avoid this problem, we consider the following weighting function:

$$w_{ij}^{M_k} = \alpha_i^{M_k} A_{M_k}(P_{M_k}(r(b_i, b_j)|s)) \tag{5}$$

where $A_{M_k}(p)$ is the function that returns the expected accuracy of M_k's vote with its dependency probability p, and $\alpha_i^{M_k}$ is a normalization factor. Such a function can be trained by plotting a P-A curve for training data. Note that training data should be shared by all the committee members. In practice, for training a P-A curve, some smoothing technique should be applied to avoid overfitting.

Class The standardization process in the above option **Normal** can also be seen as an effort for reducing the averaged cross entropy of the model on test data. Since P-A curves tend to defer not only between different models but also between different problem classes, if one incorporates some problem classification into (5), the averaged cross entropy is expected to be reduced further:

$$w_{ij}^{M_k} = \beta_i^{M_k} A_{M_k C_{b_i}} (P_{M_k}(r(b_i, b_j)|s)) \qquad (6)$$

where $A_{M_k C_{b_i}}(p)$ is the P-A curve of model M_k only for the problems of class C_{b_i} in training data, and $\beta_i^{M_k}$ is a normalization factor. For problem classification, syntactic/lexical features of b_i may be useful.

3.2 Combining functions

For combination functions, we have so far considered only simple weighted voting, which averages the given weight matrices:

$$o_{ij} = \frac{1}{m} \sum_{k=1}^{m} w_{ij}^{M_k} \qquad (7)$$

where o_{ij} is the (i, j) element of O.

Note that the committee-based partial parsing framework presented here can be seen as a generalization of the previously proposed voting-based techniques in the following respects:

(a) A committee accepts probabilistically parameterized votes as its input.

(d) A committee accepts multiple voting (i.e. it allow a committee member to vote not only to the best-scored candidate but also to all other potential candidates).

(c) A committee provides a means for standardizing original votes.

(b) A committee outputs a probabilistic distribution representing a final decision, which constitutes a C-A curve.

For example, none of simple voting techniques for word class tagging proposed by van Halteren et al. (1998) does not accepts multiple voting. Henderson and Brill (1999) examined constituent voting and naive Bayes classification for parsing, obtaining positive results for each. Simple constituent voting, however, does not accept parametric votes. While Naive Bayes seems to partly accept parametric multiple voting, it does not consider either standardization or coverage/accuracy trade-off.

Table 1: The total / 11-point accuracy achieved by each individual model

	total	11-point
A	0.8974	0.9607
B	0.8551	0.9281
C	0.8586	0.9291
D	0.8470	0.9266
E	0.7885	0.8567

4 Experiments

4.1 Settings

We conducted experiments using the following five statistical parsers:

- KANA (Ehara, 1998): a bottom-up model based on maximum entropy estimation. Since dependency score matrices given by KANA have no probabilistic semantics, we normalized them for each row using a certain function manually tuned for this parser.
- CHAGAKE (Fujio et al., 1998): an extension of the bottom-up model proposed by Collins (Collins, 1996).
- Kanayama's parser (Kanayama et al., 1999): a bottom-up model coupled with an HPSG.
- Shirai's parser (Shirai et al., 1998): a top-down model incorporating lexical collocation statistics. Equation (1) was used for estimating DPs.
- Peach Pie Parser (Uchimoto et al., 1999): a bottom-up model based on maximum entropy estimation.

Note that these models were developed fully independently of each other, and have significantly different characters (for a comparison of their performance, see Table 1). In what follows, these models are referred to anonymously.

For the source of the training/test set, we used the Kyoto corpus (ver.2.0) (Kurohashi et al., 1997), which is a collection of Japanese newspaper articles annotated in terms of word boundaries, POS tags, BP boundaries, and inter-BP dependency relations. The corpus originally contained 19,956 sentences. To make the training/test sets, we first removed all the sentences that were rejected by any of the above five parsers (3,146 sentences). For the remaining 16,810 sentences, we next checked the consistency of the BP boundaries given by the parsers since they had slightly different criteria for BP segmentation from each other. In this process, we tried to recover as many inconsistent boundaries as possible. For example,

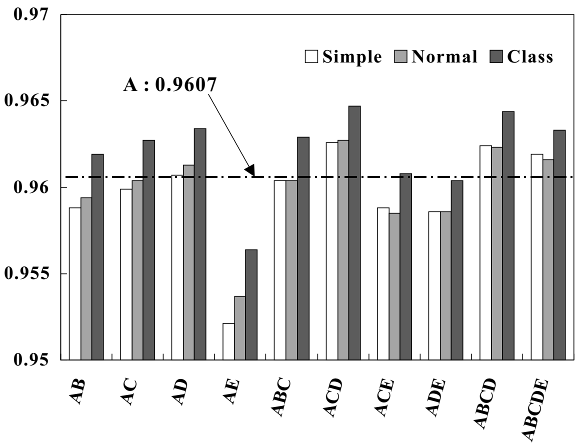

Figure 4: 11-point accuracy: A included

we found there were quite a few cases where a parser recognized a certain word sequence as a single BP, whereas some other parser recognized the same sequence as two BPs. In such a case, we regarded that sequence as a single BP under a certain condition. As a result, we obtained 13,990 sentences that can be accepted by all the parsers with all the BP boundaries consistent [3]. We used this set for training and evaluation.

For closed tests, we used 11,192 sentences (66,536 BPs[4]) for both training and tests. For open tests, we conducted five-fold cross-validation on the whole sentence set.

For the classification of problems, we manually established the following twelve classes, each of which is defined in terms of a certain morphological pattern of depending BPs:

1. nominal BP with a case marker "*wa* (topic)"
2. nominal BP with a case marker "*no* (POS)"
3. nominal BP with a case marker "*ga* (NOM)"
4. nominal BP with a case marker "*o* (ACC)"
5. nominal BP with a case marker "*ni* (DAT)"
6. nominal BP with a case marker "*de* (LOC/...)"
7. nominal BP (residue)

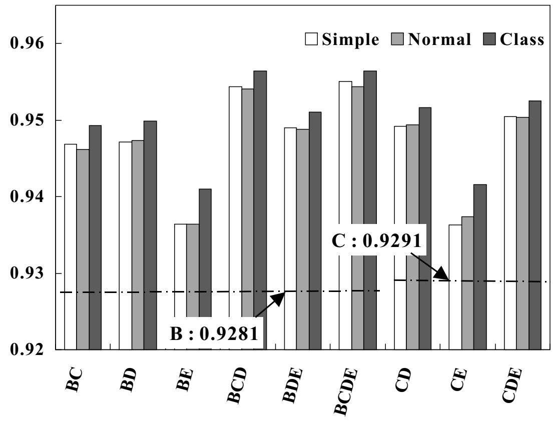

Figure 5: 11-point accuracy: B/C included

8. adnominal verbal BP
9. verbal BP (residue)
10. adverb
11. adjective
12. residue

4.2 Results and discussion

Table 1 shows the total/11-point accuracy of each individual model. The performance of each model widely ranged from 0.96 down to 0.86 in 11-point accuracy. Remember that A is the optimal model, and there are two second-best models, B and C, which are closely comparable. In what follows, we use these achievements as the baseline for evaluating the error reduction achieved by organizing a committee.

The performance of various committees is shown in Figure 4 and 5. Our primary interest here is whether the weighting functions presented above effectively contribute to error reduction. According to those two figures, although the contribution of the function **Normal** were nor very visible, the function **Class** consistently improved the accuracy. These results can be a good evidence for the important role of weighting functions in combining parsers. While we manually built the problem classification in our experiment, automatic classification techniques will also be obviously worth considering.

We then conducted another experiment to examine the effects of multiple voting. One can straightforwardly simulate a single-voting committee by replacing w_{ij} in equation (7) with w'_{ij} given by:

$$w'_{ij} = \begin{cases} w_{ij} & (\text{if } j = \arg\max_k w_{ik}) \\ 0 & (\text{otherwise}) \end{cases} \tag{8}$$

[3] In the BP concatenation process described here, quite a few trivial dependency relations between neighboring BPs were removed from the test set. This made our test set slightly more difficult than what it should have been.

[4] This is the total number of BPs excluding the right-most two BPs for each sentence. Since, in Japanese, a BP always depends on a BP following it, the right-most BP of a sentence does not depend on any other BP, and the second right-most BP always depends on the right-most BP. Therefore, they were not seen as subjects of evaluation.

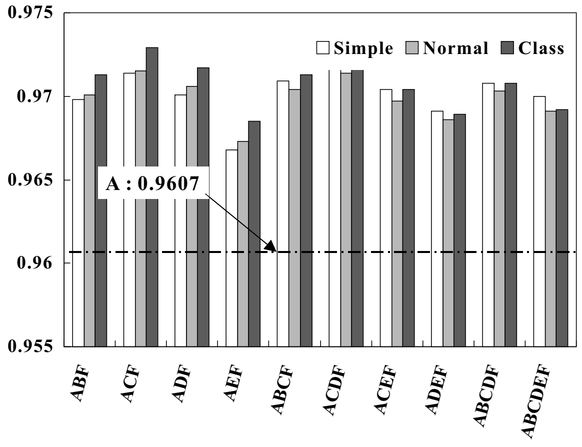

Figure 6: 11-point accuracy: +KNP

The results are shown in Figure 7, which compares the original multi-voting committees and the simulated single-voting committees. Clearly, in our settings, multiple voting significantly outperformed single voting particularly when the size of a committee is small.

The next issues are whether a committee always outperform its individual members, and if not, what should be considered in organizing a committee. Figure 4 and 5 show that committees not including the optimal model A achieved extensive improvements, whereas the merit of organizing committees including A is not very visible. This can be partly attributed to the fact that the competence of the individual members widely diversed, and A significantly outperforms the other models.

Given the good error reduction achieved by committees containing comparable members such as BC, BD and BCD, however, it should be reasonable to expect that a committee including A would achieve a significant improvement if another nearly optimal model was also incorporated. To empirically prove this assumption, we conducted another experiment, where we add another parser KNP (Kurohashi et al., 1994) to each committee that appears in Figure 4. KNP is much closer to model A in total accuracy than the other models (0.8725 in total accuracy) [5]. However, it does not provide DP matrices since it is designed in a rule-based fashion — the current version of KNP provides only the best-preferred parse tree for each input sentence without any scoring annotation. We thus let KNP to simply vote its total accuracy. The results are shown in Figure 6. This time all the committees achieved significant improvements, with the maximum error reduction rate up to 31% [6].

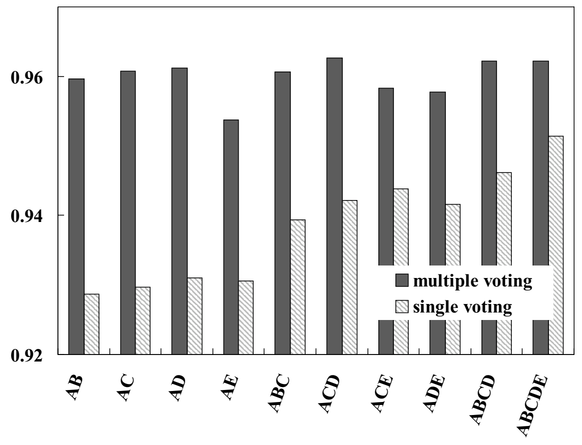

Figure 7: Single voting vs. Multiple voting

As suggested by the results of this experiment with KNP, our scheme allows a rule-based non-parametric parser to play in a committee preserving its ability to output parametric DP matrices. To push the argument further, suppose a plausible situation where we have an optimal but non-parametric rule-based parser and several suboptimal statistical parsers. In such a case, our committee-based scheme may be able to organize a committee that can provide DP

[5] [note for revision] 0.8725 is the total accuracy achieved by KNP, but with an unnecessary experimental option setting wrongly specified in execution. With the proper setting, KNP achieves a much higher accuracy, 0.9125, for the same test data. KNP thus outperforms model A, in actual fact.

[6] [note for revision] Figure 6 shows the gains in 11-point accuracy achieved by combining the statistical parsers in hand and KNP with the improper option setting as mentioned above. With the proper setting, the best achieved accuracy turned out to be up to 0.9753 (11-point accuracy). We should note, however, that KNP with the proper setting outpermed the optimal statistical parser A, and therefore the baseline of the committees should be not at A's accuracy but at KNP's accuracy. Since we cannot evaluate the 11-point accuracy of KNP, the best we can do is to make comparisons in *total* accuracy. According to our experiment, the best achieved *total* accuracy of the committee is 0.9200, and the gain (0.0075 points from the baseline) is much less significant than those shown in Figure 6. Given this result, we need to totally reconsider and change the discussion in this paragraph.

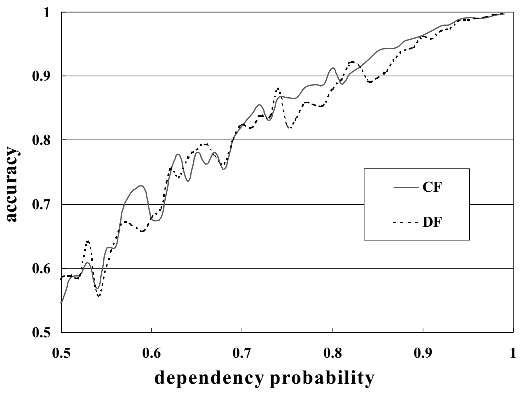

Figure 8: P-A curves: +KNP

matrices while preserving the original total accuracy of the rule-based parser. To see this, we conducted another small experiment, where we combined KNP with each of C and D, both of which are less competent than KNP. The resulting committees successfully provided reasonable P-A curves as shown in Figure 8, while even further improving the original total accuracy of KNP (0.8725 to 0.8868 for CF and 0.8860 for DF)[7]. Furthermore, the committees also gained the 11-point accuracy over C and D (0.9291 to 0.9600 for CF and 0.9266 to 0.9561 for DF). These results suggest that our committee-based scheme does work even if the most competent member of a committee is rule-based and thus non-parametric.

5 Conclusion

This paper presented a general committee-based framework that can be coupled with probabilistic partial parsing. In this framework, a committee accepts parametric multiple votes, and then standardizes them, and finally provides a probabilistic distribution. We presented a general method for producing probabilistic multiple votes (i.e. DP matrices), which allows most of the existing probabilistic models for parsing to join a committee. Our experiments revealed that (a) if more than two comparably competent models are available, it is likely to be worthwhile to combine them, (b) both multiple voting and vote standardization effectively work in committee-based partial parsing, (c) our scheme also allows a non-parametric rule-based parser to make a good contribution. While our experiments have so far been producing promising results, there seems to be much room left for investigation and improvement.

Acknowledgments

We would like to express our special thanks to all the creators of the parsers used here for enabling all of this research by providing us their systems. We would also like to thank the reviewers for their suggestive comments.

References

Brill, E. and J. Wu. Classifier Combination for Improved Lexical Disambiguation. In *Proc. of the 17th COLING*, pp.191–195, 1998.

Carroll, J. ,G. Minnen, Y. Canning, S. Devlin and J. Tait. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*,1998.

Charniak, E. Statistical parsing with a context-free grammar and word statistics. In *Proc. of the AAAI*, pp.598–603, 1997.

Collins, M. J. A new statistical parser based on bigram lexical dependencies. In *Proc. of the 34th ACL*, pp.184–191, 1996.

Collins, M. J. Three generative, lexicalised models for statistical parsing. In *Proc. of the 35th ACL*, pp.16–23, 1997.

Ehara, T. Estimating the consistency of Japanese dependency relations based on the maximam entropy modeling. In *Proc. of the 4th Annual Meeting of The Association of Natural Language Processing*, pp.382–385, 1998. (In Japanese)

Fiscus, J. G. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *EuroSpeech*, 1997.

Frederking, R. and S. Nirenburg. Three heads are better than one. In *Proc. of the 4th ANLP*, 1994.

Fujio, M. and Y. Matsumoto. Japanese dependency structure analysis based on lexicalized statistics. In *Proc. of the 3rd EMNLP*, pp.87–96, 1998.

Henderson, J. C. and E. Brill. Exploiting Diversity in Natural Language Processing: Combining Parsers. In *Proc. of the 1999 Joint SIGDAT Conference on EMNLP and VLC*, pp.187–194.

Jensen, K., G. E. Heidorn, and S. D. Richardson, editors. *natural language processing: The PLNLP Approach.* Kluwer Academic Publishers, 1993.

Kanayama, H., K. Torisawa, Y. Mitsuisi, and J. Tsujii. Statistical Dependency Analysis with an HPSG-based Japanese Grammar. In *Proc. of the NLPRS*, pp.138–143, 1999.

[7] [note for revision] Again, 0.8725 is the total accuracy achieved by KNP with an unnecessary experimental option setting wrongly specified in execution.

Kurohashi, S. and M. Nagao. Building a Japanese parsed corpus while improving the parsing system. In *Proc. of NLPRS*, pp.151–156, 1997.

Kurohashi, S. and M. Nagao. KN Parser : Japanese Dependency/Case Structure Analyzer. In *Proc. of The International Workshop on Sharable Natural Language Resources*, pp.48-55, 1994.

Poole, D. Average-case analysis of a search algorithm for estimating prior and posterior probabilities in Bayesian networks with extreme probabilities. *the 13th IJCAI*, pp.606–612, 1993.

Pedersen, T. A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation In *Proc. of the NAACL*, pp.63-69, 2000.

Shirai, K., K. Inui, T. Tokunaga and H. Tanaka An empirical evaluation on statistical parsing of Japanese sentences using a lexical association statistics. *the 3rd EMNLP*, pp.80-87, 1998.

Uchimoto, K., S. Sekine, and H. Isahara. Japanese dependency structure analysis based on maximum entopy models. In *Proc. of the 13th EACL*, pp.196-203, 1999.

van Halteren, H., J. Zavrel, and W. Daelemans. Improving data driven wordclass tagging by system combination. In *Proc. of the 17th COLING*, 1998.