Extended Models and Tools for High-performance Part-of-speech Tagger

Masayuki Asahara and **Yuji Matsumoto**
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0101, Japan
{masayu-a,matsu}@is.aist-nara.ac.jp

Abstract

Statistical part-of-speech(POS) taggers achieve high accuracy and robustness when based on large scale manually tagged corpora. However, enhancements of the learning models are necessary to achieve better performance. We are developing a learning tool for a Japanese morphological analyzer called *ChaSen*. Currently we use a fine-grained POS tag set with about 500 tags. To apply a normal tri-gram model on the tag set, we need unrealistic size of corpora. Even, for a bi-gram model, we cannot prepare a moderate size of an annotated corpus, when we take all the tags as distinct. A usual technique to cope with such fine-grained tags is to reduce the size of the tag set by grouping the set of tags into equivalence classes. We introduce the concept of *position-wise grouping* where the tag set is partitioned into different equivalence classes at each position in the conditional probabilities in the Markov Model. Moreover, to cope with the data sparseness problem caused by exceptional phenomena, we introduce several other techniques such as word-level statistics, smoothing of word-level and POS-level statistics and a *selective tri-gram model*. To help users determine probabilistic parameters, we introduce an error-driven method for the parameter selection. We then give results of experiments to see the effect of the tools applied to an existing Japanese morphological analyzer.

1 Introduction

Along with the increasing availability of annotated corpora, a number of statistic POS taggers have been developed which achieve high accuracy and robustness. On the other hand, there is still continuing demand for the improvement of learning models when sufficient quantity of annotated corpora are not available in the users domains or languages. Flexible tools for easy tuning of learning models are in demand. We present such tools in this paper. Our tools are originally intended for use with the Japanese morphological analyzer, *ChaSen* (Matsumoto et al., 1999), which at present is a statistical tagger based on the variable memory length Markov Model (Ron et al., 1994). We first give a brief overview of the features of the learning tools.

The part-of-speech tag set we use is a slightly modified version of the IPA POS tag set (RWCP, 2000) with about 500 distinct POS tags. The real tag set is even larger since some words are treated as distinct POS tags. The size of the tag set is unrealistic for building tri-gram rules and even bi-gram rules which take all the tags as distinct. The usual technique for coping with such fine-grained tags is to reduce the size of the tag set by grouping the set of tags into equivalence classes (Jelinek, 1998). We introduce the concept of *position-wise grouping* where the tag set is partitioned into different equivalence classes at each position in the conditional probabilities in the Markov Model. This feature is especially useful for Japanese language analysis since Japanese is a highly conjugated language, where conjugation forms have a great effect on the succeeding morphemes, but have little to do with the preceding morphemes. Moreover, in colloquial language, a number of contracted expressions are common, where two or more morphemes are contracted into a single word. The contracted word behaves as belonging to different parts-of-speech by connecting to the previous word or to the next word. Position-wise grouping enables users to group such words differently according to the positions in which they appear.

Data sparseness is always a serious problem when dealing with a large tag set. Since it is unrealistic to adopt a simple POS tri-gram model to our tag set, we base our model on a bi-gram model and augment it with *selective tri-grams*. By selective tri-gram, we mean that only special contexts are conditioned by tri-gram model and are mixed with the ordinary bi-gram model. We also incorporate some smoothing techniques for coping with the data sparseness problem.

By combining these methods, we constructed the learning tools for a high-performance statistical morphological analyzer that are able to learn the probability parameters with only a moderate size tagged corpus.

The rest of this paper is structured as follows.

Section 2 discusses the basic concepts of the statistical morphological analysis and some problems of the statistical approach. Section 3 presents the characteristics of the our learning tools. Section 4 reports the result of some experiments and the accuracy of the tagger in several settings. Section 5 discusses related works. Finally, section 6 gives conclusions and discusses future works.

Throughout this paper, we use *morphological analysis* instead of part-of-speech tagging since Japanese is an agglutinative language. This is the standard terminology in Japanese literatures.

2 Preliminaries

2.1 Statistical morphological analysis

The POS tagging problem or the Japanese morphological analysis problem must do tokenization and find the sequence of POS tags $T = t_1, \ldots, t_n$ for the word sequence $W = w_1, \ldots, w_n$ in the input string S. The target is to find T that maximizes the following probability:

$$arg\max_T P(T|W)$$

Using the Bayes' rule of probability theory, $P(W,T)$ can be decomposed as a sequence of the products of tag probabilities and word probabilities.

$$\begin{aligned} arg\max_T P(T|W) &= arg\max_T \frac{P(T,W)}{P(W)} \\ &= arg\max_T P(T,W) \\ &= arg\max_T P(W|T)P(T) \end{aligned}$$

We assumed that the word probability is constrained only by its tag, and that the tag probability is constrained only by its preceding tags, either with the bi-gram or the tri-gram model:

$$\begin{aligned} P(W|T) &= \prod_{i=1}^{n} P(w_i|t_i) \\ P(T) &= \prod_{i=1}^{n} P(t_i|t_{i-1}) \end{aligned}$$

$$\left(\quad P(T) = \prod_{i=1}^{n} P(t_i|t_{i-2,i-1}) \quad \right)$$

The values are estimated from the frequencies in tagged corpora using maximum likelihood estimation:

$$\begin{aligned} P(w_i|t_i) &= \frac{F(w_i,t_i)}{F(t_i)} \\ P(t_i|t_{i-1}) &= \frac{F(t_{i-1},t_i)}{F(t_{i-1})} \\ P(t_i|t_{i-2,i-1}) &= \frac{F(t_{i-2},t_{i-1},t_i)}{F(t_{i-2},t_{i-1})} \end{aligned}$$

Using these parameters, the most probable tag sequence is determined using the Viterbi algorithm.

2.2 Hierarchical Tag Set

We use the IPA POS tag set (RWCP, 2000). This tag set consist of three elements: the part-of-speech, the type of conjugation and the form of conjugation (the latter two elements are necessary only for words that conjugate).

The POS tag set has a hierarchical structure: *The top POS level* consists of 15 categories(e.g., Noun, Verb, ...). The second and lower levels are *the subdivision level*. For example, Noun is further subdivided into common nouns(general), proper nouns, numerals, and so on. Proper Noun is subdivided into General, Person, Organization and Place. Person and Place are subdivided again. The bottom level of the subdivision level is *the word level*, which is conceptually regarded as a part of the subdivision level.

In the Japanese language, verbs, adjectives and auxiliary verbs have conjugation. These are categorized into a fixed set of conjugation types(CTYPE), each of which has a fixed set of conjugation forms(CFORM). It is known that in Japanese that the CFORM varies according to the words appearing in the succeeding position. Thus, at the conditional position of the estimated tag probabilities, the CFORM plays an important role, while in the case of other positions, they need not be distinguished. Figure 1 illustrates the structure of the tag set.

2.3 Problems in statistical models

On the one hand, most of the problems in statistical natural language processing stem from the sparseness of training data. In our case, the number of the most fine-grained tags (disregarding the word level) is about 500. Even when we use the bi-gram model, we suffer from the data sparseness problem. The situation is much worse in the case of the tri-gram model. This may be remedied by reducing the tag set by grouping the tags into a smaller tag set.

On the other hand, there are various kinds of exceptions in language phenomena. Some words have different contextual features from others in the same tag. Such exceptions require a word or some group of words to be taken itself as a distinct part-of-speech or its statistics to be taken in distinct contexts. In our statistical learning tools, those exceptions are handled by position-wise grouping, word-level statistics, smoothing of word-level and POS-level, and selective tri-gram model, which are described in turn in the next section. These features enable users to

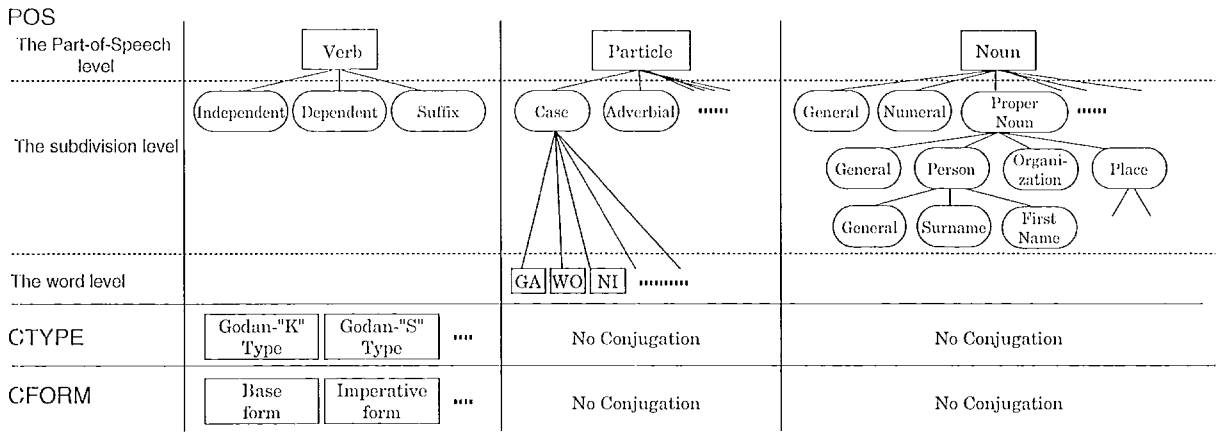

Figure 1: The examples of the hierarchical tag set

adjust the balance between fine and coarse grained model settings.

3 Features of the tools

This section overviews characteristic features of the learning tools for coping with the above mentioned problems.

3.1 Position-wise grouping of POS tags

Since we use a very fine-grained tag set, it is important to classify them into some equivalence classes to reduce the size of probabilistic parameters. Moreover, as is discussed in the previous section, some words or POS behaves differently according to the position they appear. In Japanese, for instance, the CFORM play an important role only to disambiguate the words at their succeeding position. In other words, the CFORM should be taken into account only when they appear at the position of t_{i-1} in either bi-gram or tri-gram model (t_{i-1} in $P(t_i|t_{i-1})$ and $P(t_i|t_{i-2}, t_{i-1})$). This means that when the statistics of verbs are taken, they should be grouped differently according to the positions. Note that, we named the positions; The current position means the position of t_i in the bi-gram statistics $P(t_i|t_{i-1})$ or the tri-gram statistics $P(t_i|t_{i-2}, t_{i-1})$. The preceding position means the position of t_{i-1}. The second preceding position means the position of t_{i-2}.

There are quite a few contracted forms in colloquial expressions. For example, auxiliary verb "chau" is a contracted forms consisting of two words "te(particle) + simau(auxiliary verb)" and behaves quite differently from other words. One way to learn its statistical behavior is to collect various usages of the word and add the data to the training data after correctly annotating them. In contrast, the idea of point-wise grouping provides a nice alternative solution to this problem. By simply group this word into the same equivalence class of "te" for the current position t_i and group it into the same equivalent class of "simau" for the preceding position t_{i-1} in $P(t_i|t_{i-1})$, it learns the statistical behavior from these classes.

We now describe the point-wise grouping in a more precise way. For simplicity, we assume bi-gram model. Let $\mathcal{T} = \{A, B, \cdots\}$ be the original tag set. We introduce two partitions of the tag set, one is for the current position $\mathcal{T}^c = \{A^c, B^c, \cdots\}$, and the other is for the preceding position $\mathcal{T}^p = \{A^p, B^p, \cdots\}$. We define the equivalence mapping of the current position: $I^c(\mathcal{T} \to \mathcal{T}^c)$, and another mapping of the preceding position: $I^p(\mathcal{T} \to \mathcal{T}^p)$.

Figure 2 shows an example of the partitions by those mappings, where the equivalence mappings are:

$$I^c = \{A \to A^c, B \to A^c, C \to A^c, D \to B^c, E \to B^c, \ldots\}$$

$$I^p = \{A \to A^p, B \to A^p, C \to B^p, D \to B^p, E \to C^p, \ldots\}$$

Suppose we express the equivalence class to which the tag t belongs as $[t]^c$ for the current position and $[t]^p$ for the preceding position, then:

$$P(w_i|t_i) = \frac{F(w_i, [t_i]^c)}{F([t_i]^c)} = \frac{F(w_i, t_i)}{F([t_i]^c)}$$

$$P(t_i|t_{i-1}) = \frac{F([t_{i-1}]^p, [t_i]^c)}{F([t_{i-1}]^p)}$$

3.2 Word-level statistics

Some words behave differently from other words even in the same POS. Especially Japanese particles, auxiliary verbs and some affixes are known to have different contextual behavior. The tools can define

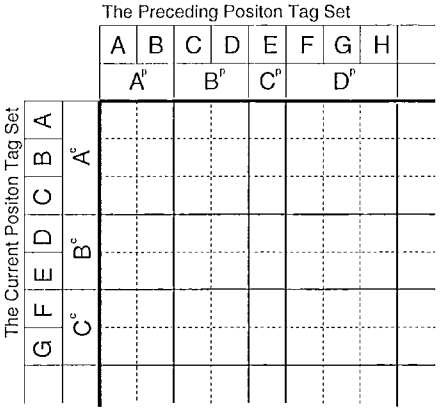

Figure 2: Position-wise grouping of tags

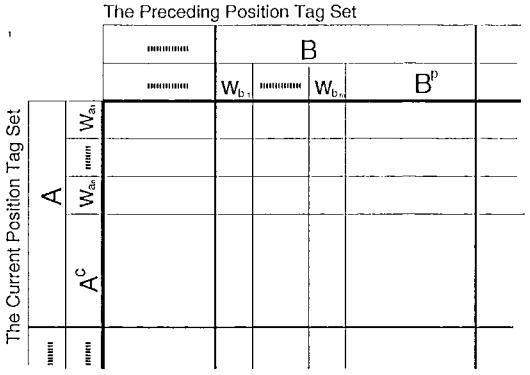

Figure 3: the word extended tag set

some words as distinct POS and their statistics are taken individually.

The tag set $\mathcal{T}$ extends to a new tag set $\mathcal{T}^{ext}$ that defines some words as individual POSs (*the word level*). Modification to the probability formulas for such word level tags is straightforward.

Note that the statistics for POS level should be modified when some words in the same group are individuated. Suppose that the tags A and B are defined in the $\mathcal{T}$ and some words $W_{a_1}, \ldots, W_{a_n} \in A$ and $W_{b_1}, \ldots, W_{b_m} \in B$ are individuated in $\mathcal{T}^{ext}$. We define tags $A_{ext}, B_{ext} \in \mathcal{T}^{ext}$ as follows:

$$\begin{aligned} A_{ext} &= A \setminus \{w_{a_1}, \ldots, w_{a_n}\} \\ B_{ext} &= B \setminus \{w_{b_1}, \ldots, w_{b_m}\} \end{aligned}$$

To estimate the probability for the connection A-B, the frequency $F(A_{ext}, B_{ext})$ is used rather than the total frequency $F(A, B)$. Figure 3 illustrate the tag set extension of this situation.

These tag set extension is actually a special case of position-wise grouping. The equivalence mappings are from all word level tags to $\mathcal{T}^{ext}$. The mapping I^c maps all the words in A_{ext} into A_{ext} and maps each of $\{W_{a_1}, \ldots, W_{a_n}\}$ into itself. In the same way, I^p maps all the words in B_{ext} into B_{ext} and maps each of $\{W_{b_1}, \ldots, W_{b_m}\}$ into itself.

3.3 Smoothing of word and POS level statistics

When a word is individuated while its occurrence frequency is not high, we have to accumulate instances to obtain enough statistics. Another solution is to smooth the word level statistics with POS level statistics. In order to back-off the sparseness of the words, we use the statistics of the POS to which the words belong.

We define two smoothing coefficients: λ_c is the smoothing ratio for the current position and λ_p is the smoothing ratio of the preceding position. Those values can be defined for each word.

Suppose the word w_i is individuated and its POS is t_i. If the current position is smoothed, then the tag probability is defined as follows (note that w_i itself is an individuated tag):

$$\tilde{P}(w_i|t_{i-1}) = ((1-\lambda_c)P(t_i|t_{i-1}) + \lambda_c P(w_i|t_{i-1}))$$

If the word at the preceding positions is smoothed (assume t_{i-1} is the POS of w_{i-1}):

$$\tilde{P}(t_i|w_{i-1}) = (1-\lambda_p)P(t_i|t_{i-1}) + \lambda_p P(t_i|w_{i-1})$$

If the both words of the positions is extend:

$$\begin{aligned} &\tilde{P}(w_i|w_{i-1}) \\ &\quad = \lambda_p((1-\lambda_c)P(t_i|w_{i-1}) + \lambda_c P(w_i|w_{i-1})) \\ &\qquad +(1-\lambda_p)((1-\lambda_c)P(t_i|t_{i-1}) + \lambda_c P(w_i|t_{i-1})) \end{aligned}$$

3.4 Selective tri-gram model

Simple tri-gram models are not feasible for a large tag set. As a matter of fact, only limited cases require as long contexts as tri-grams. We propose to take into account only limited tri-gram instances, which we call *selective tri-grams*. Our model is a mixture of such tri-gram statistics with bi-gram ones.

The idea of mixture of different context length is not new. Markov Models with variable memory length are proposed by Ron(Ron et al., 1994), in which a mixture model of n-grams with various value of n is presented as well as its learning algorithms. In such a model, the set of contexts (the set of states of the automata) should be mutually disjoint for the automata to be deterministic and well-defined.

We give a little different interpretation to tri-gram statistics. We consider a tri-gram as an exceptional

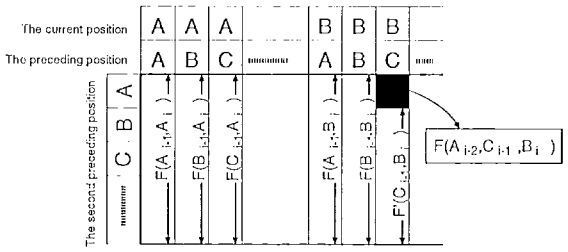

Figure 4: Selective tri-gram

context. When a bi-gram context and a tri-gram context have some intersection, the tri-gram context is regarded as an exception within the bi-gram context. In this sense, all the contexts are mutually disjoint as well in our model, and it is possible to convert our model into Ron's formulation. However, we think that our formulation is more straighforward if the longer contexts are interpreted as exceptions to the shorter contexts.

We assume that the grouping at the current position ($\mathcal{T}^c$) share the same grouping of the bi-gram case. But for the preceding position and the second preceding position, we can define different groupings of tag sets from those of the bi-gram case. We introduce the two new tag sets for the preceding positions:

The tag set of the preceding position:

$$\mathcal{T}^{p'} = \{A^{p'}, B^{p'}, \cdots\}$$

The tag set of the second preceding position:

$$\mathcal{T}^{pp'} = \{A^{pp'}, B^{pp'}, \cdots\}$$

We define the equivalence mapping for the preceding position: $I^{p'}(\mathcal{T} \to \mathcal{T}^{p'})$, and the mapping for the second preceding position: $I^{pp'}(\mathcal{T} \to \mathcal{T}^{pp'})$. Assuming that an equivalence classes for t defined by the mapping $I^{pp'}$ is expressed as $[t]^{pp'}$, the tri-gram probability is defined naturally as follows:

$$\begin{aligned} P(t_i|t_{i-2}, t_{i-1}) &= P([t_i]^c|[t_{i-2}]^{pp'}, [t_{i-1}]^{p'}) \\ &= \frac{F([t_{i-2}]^{pp'}, [t_{i-1}]^{p'}, [t_i]^c)}{F([t_{i-2}]^{pp'}, [t_{i-1}]^{p'})} \end{aligned}$$

Figure 4 shows an image of frequency counts for tri-gram model.

In case some bi-gram context overlaps with a tri-gram context, the bi-gram statistics are taken by excluding the tri-gram statistics.

For example, if we include the tri-gram context $A-C-B$ in our model, then the statistics of the bi-gram context $C-B$ is taken as follows (F stands for *true* frequency in training corpora while F' stands for estimated frequency to be used for probability calculation):

$$F'(C, B) = F(C, B) - F(A, C, B)$$

Since selection of tri-gram contexts is not easy task, the tools supports the selection based on an error-driven method. We omit the detail because of the space limitation.

3.5 Estimation for unseen words in corpus

Since not all the words in the dictionary appear in the training corpus, the occurrence probability of unseen words should be allocated in some way. There are a number of method for estimating unseen events. Our current tool adopts Lidstone's law of succession, which add a fixed count to each observation.

$$P(w|t) = \frac{F'(w, t) + \alpha}{\sum_{v \in t} F(v, t) + \alpha \cdot |t|}$$

At present, the default frequency count α is set to 0.5.

4 Experiments and Evaluation

For evaluating how the proposed extension improves a normal bi-gram model, we conducted several experiments. We group verbs according to the conjugation forms at the preceding position, take word level statistics for all particles, auxiliary verbs and symbols, each of which is smoothed with the immediately higher POS level. Selective tri-gram contexts are defined for discriminating a few notoriously ambiguous particle "no" and auxiliary verbs "nai" and "aru." This is a very simple extension but suffices for evaluating the effect of the learning tools.

We use 5-fold cross evaluation over the RWCP tagged corpus (RWCP, 2000). The corpus date size is 37490 sentences(958678 words). The errors of the corpus are manually modified. The annotated corpus is divided into the training data set(29992 sentences, 80%) and the test data set(7498 sentences, 20%). Experiments were repeated 5 times, and the results were averaged.

The evaluation is done at the following 3 levels:

- level1: only word segmentation (tokenization) is evaluated
- level2: word segmentation and the top level part-of-speech are evaluated
- level3: all information is taken into account for evaluation

Using the tools, we create the following six models:

D: normal bi-gram model

D_w: D + word level statistics for particles, etc.

Table 1: Results for test data (F-value %)

dataset	level1	level2	level3
D	98.69	98.12	96.91
D_w	98.75	98.24	97.22
D_{wg}	98.80	98.26	97.20
D_{ws}	98.76	98.27	97.23
D_{wgt}	98.78	98.35	97.27

Table 2: Results for learning data (F-value %)

dataset	level1	level2	level3
D	98.84	98.36	97.36
D_w	98.96	98.58	97.81
D_{wg}	98.92	98.46	97.61
D_{ws}	98.96	98.58	97.80
D_{wgt}	98.92	98.55	97.70

D_{wg}: D_w + grouping

D_{ws}: D_w + smoothing of word level with POS level

D_{wgt}: D_{wg} + selective tri-gram

The smoothing rate between the part-of-speech and the words is fixed to 0.9 for each word.

To evaluate the results, we use the F-value defined by the following formulae:

$$Recall = \frac{\texttt{number of correct words}}{\texttt{number of words in corpus}}$$

$$Precision = \frac{\texttt{number of correct words}}{\texttt{number of words by system output}}$$

$$F_\beta = \frac{(\beta^2 + 1) \cdot Recall \cdot Precision}{\beta^2 \cdot (Precision + Recall)}$$

For each model, we evaluate the F-value (with $\beta = 1$) for the learning data and test data at each level. The results are given in the Tables 1 and 2.

From the results the following observation is possible:

Smoothing improve on grouping dataset in test data slightly. But in the other environments the accuracy isn't improved. In this experiment, the smoothing rate for all words is fixed. We need to make the different rate for each word in the future work.

The grouping performs good result for the test dataset. It is natural that the grouping is not good for learning dataset since all the word level statistics are learned in the case of learning dataset.

Finally, the selective tri-gram (only 25 rules added) achieves non-negligible improvement at level2 and level3. Compared with the normal bi-gram model, it improves about 0.35% on level3 and about 0.2% on level2.

5 Related work

Cutting introduced grouping of words into equivalence classes based on the set of possible tags to reduce the number of the parameters (Cutting et al., 1992) . Schmid used the equivalence classes for smoothing. Their classes define not a partition of POS tags, but mixtures of some POS tags (Schmid, 1995) .

Brill proposed a transformation-based method. In the selection of tri-gram contexts we will use a similar technique (Brill, 1995) .

Haruno constructed variable length models based on the mistake-driven methods, and mixed these tag models. They do not have grouping or smoothing facilities (Haruno and Matsumoto, 1997).

Kitauchi presented a method to determine refinement of the tag set by a mistake-driven technique. Their method determines the tag set according to the hierarchical definition of tags. Word level discrimination and grouping beyond the hierarchical tag structure are out of scope of their method (Kitauchi et al., 1999).

6 Conclusion and Future works

We proposed several extensions to the statistical model for Japanese morphological analysis. We also gave preliminary experiments and showed the effects of the extensions.

Counting some words individually and smoothing them with POS level statistics alleviate the data sparseness problem. Position-wise grouping enables an effective refinement of the probability parameter settings. Using selective tri-gram provides an easy description of exceptional language phenomena.

In our future work, we will develop a method to refine the models automatically or semi-automatically. For example, error-driven methods will be applicable to the selection of the words to be individuated and the useful tri-gram contexts.

For the morphological analyzer *ChaSen*, we are using the mixture model: Position-wise grouping used for conjugation. Smoothing of the word level and the POS level used for particles.

The analyzer and the learning tools are available publicly[1].

References

E. Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4):543–565.

D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. 1992. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing.*

[1] `http://cl.aist-nara.ac.jp/lab/nlt/chasen/`

M. Haruno and Y. Matsumoto. 1997. Mistake-Driven Mixture of Hierarchical Tag Context Trees. In *35th Annual Meeting of the Association for Compuational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 230–237, July.

F. Jelinek. 1998. *Statistical Methods for Speech Recognition.* MIT Press.

A. Kitauchi, T. Utsuro, and Y. Matsumoto. 1999. Probabilistic Model Learning for Japanese Morphological Analysis by Error-driven Feature Selection (in Japanese). *Transaction of Information Processing Scicty of Japan*, 40(5):2325–2337, 5.

Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. 1999. Japanese Morphological Analyzer ChaSen Users Manual version 2.0. Technical Report NAIST-IS-TR99012, Nara Institute of Science and Technology Technical Report.

D. Ron, Y. Singer, and N. Tishby. 1994. Learning Probabilistic Automata with Variable Memory Length. In *COLT-94*, pages 35 46.

RWCP. 2000. RWC Text Database. `http://www.rwcp.or.jp/wswg/rwcdb/text/`.

H. Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *EACL SIGDAT workshop*, pages 47- 50.