HMM Parameter Learning for Japanese Morphological Analyzer

Koichi Takeuchi **Yuji Matsumoto**
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-01 Japan
{kouit-t, matsu}@is.aist-nara.ac.jp

Abstract

This paper presents a method to apply Hidden Markov Model (HMM) to parameter learning for Japanese morphological analyzer. We especially emphasize how the following two information sources affect the results of the parameter learning: 1) The initial value of parameters, i.e., the initial probabilities and 2) some grammatical constraints that hold in Japanese sentences independently of any domain. First and foremost, a simple application of HMM to Japanese corpus does not give a satisfactory results since word boundaries are not clear in Japanese texts because of lack of word separators. The first results of the experiments show that initial probabilities learned from correct tagged corpus affects greatly to the results and that a small tagged corpus is enough for the initial probabilities. The second result is that the incorporation of simple grammatical constraints works well in the improvements of the results. The final result gives that the total performance of the HMM-based parameter learning achieves almost the same level as the human developed rule-based Japanese morphological analyzer.

1 Introduction

Morphological analysis and part-of-speech tagging is an important preprocessing especially for analyses of unrestricted texts. We have been developing a rule-based Japanese morphological analyzer called JUMAN[8]. The rules are represented as costs to lexical entry and cost to pairs of adjacent parts-of-speech (connectivity cost), which are manually assigned. The cost for a lexical entry reflects the probability of the occurrence of the word, and a connectivity cost of a pair of parts-of-speech reflects the probability of an adjacent occurrence of the pair. Greater cost means less probability.

Since those costs vary according to the domain of texts, it requires much effort to estimate them for texts of a new domain. Some statistical methods have been proposed for part-of-speech tagging of English and other Indo-European languages. Church[4] proposed a method to use trigram probabilities obtained from tagged Brown corpus and achieved over 95% precision in English part-of-speech tagging. Cutting[5] used Hidden Markov Model to estimate probability parameters for the tagger and achieved 96% precision. This experiment was done on

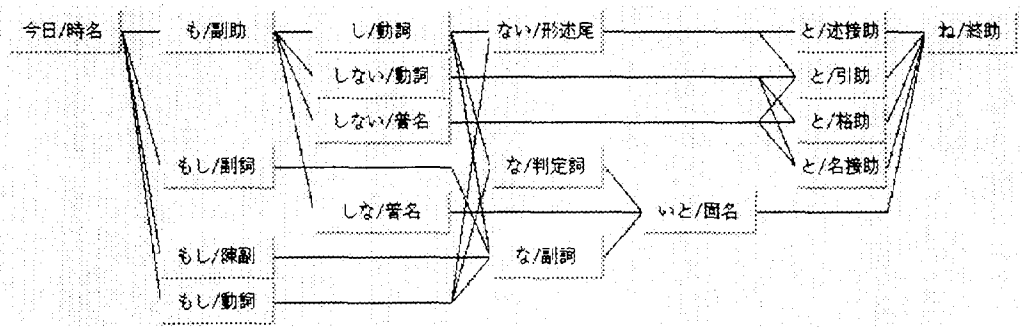

Figure 1: Sample result of Japanese morphological analysis

a large scale untagged text. Statistics works well for part-of-speech tagging of a language like English since words are separated by spaces and word order is comparatively more restricted than free word order languages like Japanese and Korean. We have pursued a similar approach based on HMM for Japanese part-of-speech tagging, resulting in a poor performance. The reason is that Japanese sentences do not have word separators, thus, word boundaries are not clear, causing spurious ambiguity in word segmentation. Chang and Chen[1] applied HMM to part-of-speech tagging of Chinese. However, they assumed a word-segmented corpus for the training data. We do not assume a large scale tagged corpora. The reasons are the following:

1. It is not easy to get a large scale tagged corpus, especially because there is no standard set of parts-of-speech for Japanese language. There is even no consensus on the definition of morphemes.
2. The probabilities of word occurrences and connectivities may vary according to the domain of texts. This necessitates to provide a tagged corpus virtually for each domain.

This paper describes how the difficulties in Japanese morphological analysis are overcome by the use of the HMM parameter learning. We put a special emphasis on the effect of the initial probabilities and some domain-independent grammatical constraints. By grammatical constraints we mean pairs of parts-of-speech or morphemes which never occurs in real texts.

Our Japanese morphological analyzer JUMAN and its relationship to HMM are introduced in the next section. Then, the effects of the initial probabilities and grammatical constraints are described by giving some experimental results.

2 JUMAN-HMM System

2.1 JUMAN morphological analyzer

JUMAN[8] is a cost based Japanese morphological analyzer developed at NAIST and Kyoto University. The morphological analysis is controlled by two types of cost functions, one for lexical entries and the other for connectivity of adjacent parts-of-speech. The result of an analysis is a lattice-like structure of

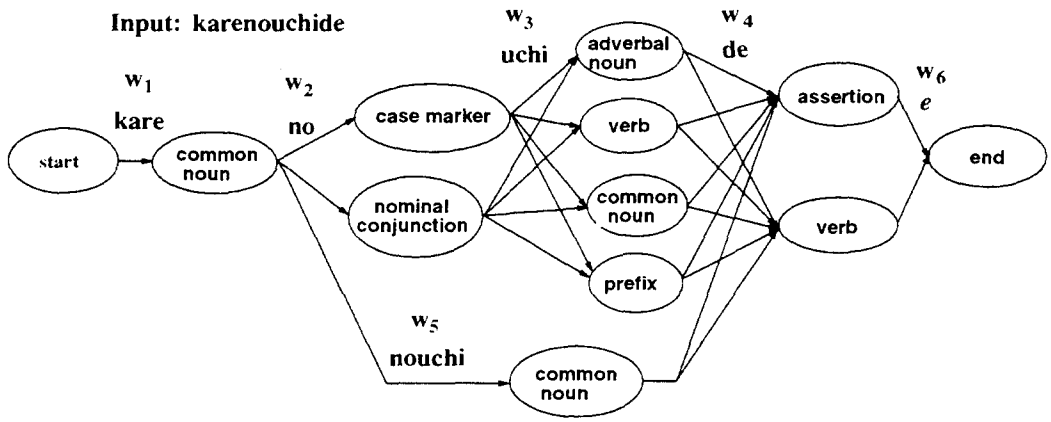

Figure 2: HMM state transition of Japanese input

words, of which the path with the least total cost is selected as the most plausible answer (see Figure 1 for an ambiguous result with the most plausible path selected on the top). The performance of the current JUMAN system is 93%~95% accuracy in word segmentation and part-of-speech tagging when tested on newspaper editorial articles.

The edges in the lattice-like structure produced by the system has the one-to-one mapping with the state transition of Hidden Markov Model if the cost is regarded as the inverse probability. Actually, when the absolute logarithmic value of the probability is regarded as a cost and multiplication of the probabilities is replaced with addition of the costs, the two models coincide with a little modification.

2.2 Hidden Markov Model of Japanese morphological analysis

When applying HMM parameter learning procedure to Japanese morphological analysis some modification is necessary since state transitions take place with an arbitrary portion of the input that makes up a lexical entry. A sample of state transition is shown in Figure 2, where two dummy states are assumed, one for initial state ('start' in the figure) and the other for the final state ('end' in the figure). Since the probability of the input sequence should be summed up for all possible paths from the initial state, the probability $P(L)$ of the input sequence L will be expressed as follows (state transition and word occurrence probabilities are assumed to depend on a single preceding state, i.e., are based on bigram model):

$$P(L) = \sum_{w_{1,n+1} \in L} \sum_{s_{0,n+1}} \prod_{i=1}^{n+1} P(s_i|s_{i-1}) P(w_i|s_i)$$

A little modification is necessary in the forward and backward probabilities since some transition with a symbol may come from distinct states with the same name. An example is the transitions by w_4='de,' where two paths come from distinct states of 'common noun.' In the following formulae, $\{s^1, \cdots, s^i, \cdots, s^\sigma\}$ is the set of states, w_k means the k-th word (k-th does not mean the position from

the initial state but the k-th portion of the input that makes up a possible word), w_k^- indicates the set of the numbers of the words that precede w_k and w_k^+ the set of the numbers of the word that follow w_k (e.g., $w_4^- = \{3,5\}$ and $w_1^+ = \{2,5\}$). $\alpha_j(k)$ is the forward probability of producing the sequence up to w_k ending up in state s^j. $\beta_i(k)$ is the backward probability of producing the sequence from w_k to the end of the input starting at state s^i.

$$\alpha_j(k) = \sum_{i=1}^{\sigma} \sum_{h \in w_k^-} \alpha_i(h) P(s^i \xrightarrow{w_k} s^j) = \sum_{i=1}^{\sigma} \sum_{h \in w_k^-} \alpha_i(h) P(s^j|s^i) P(w_k|s^j)$$

$$\beta_i(k) = \sum_{j=1}^{\sigma} \sum_{h \in w_k^+} P(s^i \xrightarrow{w_h} s^j) \beta_j(h) = \sum_{j=1}^{\sigma} \sum_{h \in w_k^+} P(s^j|s^i) P(w_h|s^j) \beta_j(h)$$

Then the probabilistic count of state transition is defined as the following. Here the modification is caused by the same fact that w_l may cause more than one transition from the state s^i to the states with the same name (i.e., s^j).

$$C(s^i \xrightarrow{w^l} s^j) = \frac{1}{P(L)} \sum_{k=1}^{e} \alpha_i(k) P(s^i \xrightarrow{w^l} s^j) \sum_{h \in w_k^+} \beta_j(h)$$

Then, the parameters are estimated based on the probabilistic counts in the same way as the normal HMM parameter estimation.

$$P_e(s^j|s^i) = \frac{\sum_{w^l \in L} C(s^i \rightarrow^{w^l} s^j)}{\sum_{w^l \in L} \sum_j C(s^i \rightarrow^{w^l} s^j)}$$

$$P_e(w_l|s^j) = \frac{\sum_i C(s^i \rightarrow^{w^l} s^j)}{\sum_{w^l \in L} \sum_i C(s^i \rightarrow^{w^l} s^j)}$$

HMM parameter learning starts with arbitral initial probabilities and the parameters (the probabilities) are estimated based on the above formulae with the transition counts obtained from the morphological analysis of a large training corpus. For a concise and comprehensive introduction to HMM parameter learning, see [2] or [7].

2.3 JUMAN-HMM system

The lattice-like structure produced by the JUMAN system (e.g., Figure 1) and the transition graph of HMM (e.g., Figure 2) have the one-to-one correspondence if the cost is regarded as the inverse logarithmic value of probability. Figure 3 shows the configuration of the integrated system of JUMAN and HMM parameter estimation system.

The HMM learning module is an independent system that learns the cost values for the JUMAN system using the HMM parameter estimation technique. The module assumes a large scale untagged Japanese corpus for its input. The

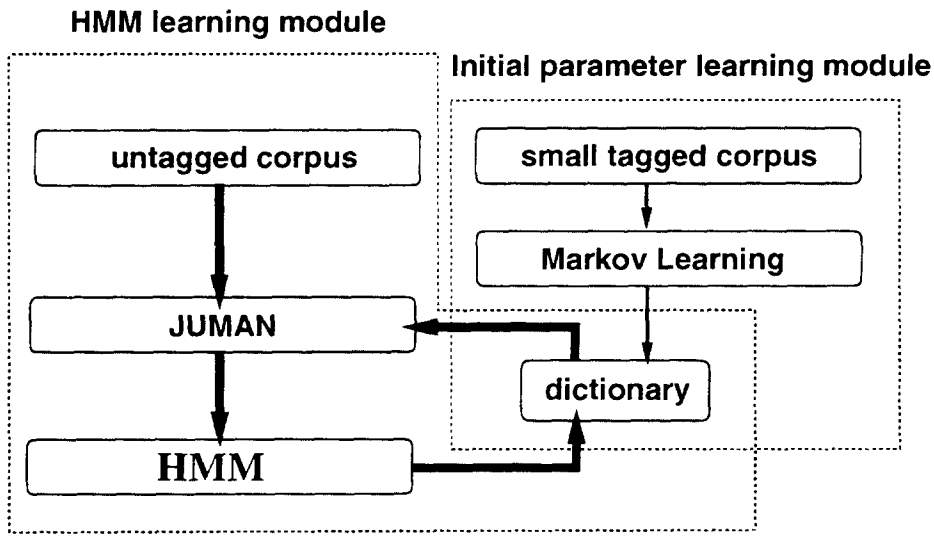

Figure 3: JUMAN-HMM System

probabilities of state transition and word occurrence are transformed into the cost values of the JUMAN system. (Alternatively the system may start with a set of cost values of the JUMAN system.) The input corpus is analyzed by the JUMAN system, producing graph structures. The HMM module uses the graph structures to estimate new probabilities. The process is repeated until it ends up at some stable state.

The initial parameter learning module counts the numbers of transitions and word occurrences and calculates the initial probabilities according to the Markov model. The initial probabilities are used as the initial parameters of the HMM-JUMAN system.

3 HMM Parameter Learning

When we undertook HMM parameter learning with Japanese newspaper editorial articles, the parameters fell into a local optimum with a poor performance. The resulting parameters give the accuracy of lower than 20%. Since Japanese texts do not specify word boundaries, a simple application of HMM parameter learning does not give good results compared with some similar works for the languages like English[3][5].

To overcome this defect and to improve the learning performance we incorporate two kinds of techniques to the HMM learning and try to figure out their effectiveness on the final performance. We found that the initial probabilities play an important role to achieve better results, and that some grammatical constraints, such as unacceptable adjacent occurrences of pairs of parts-of-speech or words, work well in preventing implausible word segmentation. In the following, we will show by some experiments how effective the initial probability and grammatical knowledge are on the final performance of Japanese morphological

initial corpus	tagged corpus 1 (300 sentences)	tagged corpus 2 (300 sentences)
1. EDR corpus	16.9(16.0)	14.8(13.6)
2. JUMAN corpus	9.9(8.7)	7.6(6.4)
3. tagged corpus 1	1.9(inside)(1.7)	6.4(5.5)
current JUMAN	7.6(6.1)	5.5(4.7)

traning corpus: editorial articles (200,000 sentences)

Table 1: Error rates based on initial probabilities

analysis.

3.1 Effect of initial probability

Initial probabilities of transitions and word occurrences are easily obtainable if there is a large scale tagged corpus, simply by counting the occurrences of each word and adjacent part-of-speech pairs and calculating the probabilities by the proportional values over the total events. Things are not so easy because we cannot expect large scale tagged corpora for a number of different domains. There is another difficulty especially in Japanese, where there is no standard set of parts-of-speech and even no standard treatment of inflections and classification of functional words such as auxiliary verbs and particles. It is not an easy task to transform a tagged corpus in a grammatical system into a one in another grammatical system. Though we now have a large scale tagged Japanese corpus distributed by EDR[6], we had a great difficulty in transforming it into another tagged corpus in the grammar system we are using at present.

We do not need a large tagged corpus but a 'good' initial probabilities so as to get better results after HMM parameter learning. To see the effect of initial probability we use our HMM parameter learning scheme with the probabilities calculated from the following (not necessarily correct) tagged corpora.

1. EDR tagged corpus: Since the tag set in our system is quite different from that of EDR corpus, only the word segmentation is used in the initial HMM parameter learning process.
2. Asahi Newspaper editorial articles tagged by JUMAN system (65,000 sentences): The corpus is tagged by the JUMAN system and the counts are used for the calculation of the initial probability (the tagging includes 5%~7% errors).
3. Manually tagged editorial articles (300 sentences): A very small corpus with very few errors.

For the training corpus we used Asahi Newspaper editorial articles (approx. 200,000 sentences). In the above initial corpora, 1. and 2. are relatively large scale but include some errors. On the other hand, 3. is very small but includes few errors. Our first evaluation is the direct evaluation of the initial probability setting. The initial probabilities are transformed directly to the cost values of the JUMAN system and some test data are analyzed under each setting. The results

initial corpus	tagged corpus 1 (300 sentences)	tagged corpus 2 (300 sentences)
2. JUMAN corpus	16.2(15.4)	14.0(13.3)
3. tagged corpus 1	3.8(inside)(3.6)	6.0(5.2)

traning corpus: editorial articles (200,000 sentences)

Table 2: Error rates of HMM trained results

initial corpus	tagged corpus 1 (300 sentences)	tagged corpus 2 (300 sentences)
tagged corpus 1	3.5(inside)(3.3)	5.4(4.7)
tagged corpus 2	6.9(6.4)	3.3(inside)(3.1)
current JUMAN	7.6(6.1)	5.5(4.7)

traning corpus: editorial articles (200,000 sentences)

Table 3: Error rates of HMM training with grammatical knowledge

are shown in Table 1. The figures are error rates, i.e., the ratio of the number of wrongly tagged or segmented morphemes over the total number of morphemes in the tagged corpus. Figures in the parentheses indicates the error rates when the categorization of Japanese postpositional particles are neglected. This is because fine categorization of postpositional particles is not easy only by referring to local contexts. Tagged corpus 1 is used both as the data for calculating the initial probabilities and a test corpus. Tagged corpus 2 is a manually tagged distinct corpus used only for the evaluation of the results. Naturally, the inside data gives the best result. The last row shows the error rates of the current rule-based JUMAN system. It is shown for the purpose of reference.

The results show that an erroneous corpus is far less useful than a small but correct corpus for obtaining the parameters. Since the EDR corpus does not give good initial probabilities, we decided not to use the result in later experiments and undertook the HMM training using the latter two initial probabilities using a training (untagged) corpus of 200,000 sentences (taken from newspaper editorial articles). Table 2 shows the results.

From this we can see that the HMM parameter learning improve the precision of the system a little for outside data but impoverish the learned results starting from the JUMAN corpus. This results also show that a small but correct initial corpus is much better than a large and erroneous corpus. Moreover, a small initial tagged corpus and HMM parameter learning could stands on a par with manually tuned rules.

3.2 Incorporating grammatical knowledge

The next experiment is to investigate how grammatical knowledge works well for the improvement of the results. We found that the HMM learned probabilities allow some grammatically unacceptable connections, such as, a prefix precedes

	tagged corpus A	tagged corpus B
initial corpus	(300 sentences)	(300 sentences)
tagged corpus A	3.0(inside)(2.8)	5.8(5.2)
tagged corpus B	5.5(4.9)	3.1(inside)(2.9)
current JUMAN	7.2(5.9)	6.3(5.0)

traning corpus: editorial articles (200,000 sentences)

Table 4: Error rates of HMM training with grammatical knowledge (2)

a postfix, a stem of a verb precedes a non-inflectional suffix, and so on. We therefore invalidated such unacceptable connections (about 15 rules) by fixing the probabilities of those adjacent occurrences to zero probability throughout the training process. Those rules are selected on the basis that they are never acceptable in Japanese sentences in any domain. The experimental results are shown in Table 3. Now the trained parameters outperform the current rule-based system.

Table 4 shows the results of experiments with the same setting except that the two tagged corpora are created by mixing up the sentences in the tagged corpora 1 and 2 and dividing them into two sets. They are named tagged corpora A and B. This shows almost the same results as above.

3.3 Effect of domain dependency

Since the rules of the current system have been tested and improved using the editorial articles as the test data, we made another experiment using a Japanese corpus of financial newspaper articles (Nikkei Newspaper). This experiment is to see the effect of the difference of the initial and test corpora. We used two manually tagged test corpora (100 sentences each) and an untagged corpus (200,000 sentences) for the training data, both of which are taken from Nikkei newspaper articles. The results are shown in Table 5. First two lines are the results where both initial and training corpora are in the same domain. The performance is almost same as the previous results (Table 3 and Table 4). The third row shows the error rates of the HMM trained system with the initial probabilities taken from a tagged editorial articles, and forth row shows the error rates of the current JUMAN system, both of which are tested on the tagged corpora of Nikkei articles.

These results show that the initial probabilities should be taken from the same domain as the training corpus even if the size of the initial tagged corpus is small. This can be read from the difference between the third row and the first two rows. This is noticeable since the size of the tagged corpus 1 taken from the editorial articles is three times larger than that of the initial corpora from Nikkei articles, still giving a worse result. Moreover, the domains of the above two initial data are not so different. Although Nikkei newspaper articles incline to economical and financial matters, both the articles are taken from newspapers, so the domains are not very different compared with novels, technical papers and spoken language. This means that even in newspapers, difference of topics potentially affects the

initial corpus Nikkei articles	tagged corpus 3 (100 sentences)	tagged corpus 4 (100 sentences)
tagged corpus 3	3.4(inside)(3.1)	6.0(4.9)
tagged corpus 4	6.5(5.4)	3.3(inside)(3.0)
tagged corpus 1	8.5(7.7)	7.4(6.9)
current JUMAN	7.8(6.7)	7.6(6.3)

traning corpus: Nikkei articles (200,000 sentences)

Table 5: HMM learning with initial corpora of different domains

performance of the morphological analysis. There need inevitably some technique to learn form real texts.

4 Conclusions

We proposed a method of applying HMM parameter learning to Japanese morphological analyzer and showed how the initial probabilities and grammatical knowledge perform well in improving the results of HMM parameter learning for Japanese morphological analysis. The results show that a small but correct tagged corpus and a large untagged training corpus could outperform manually tuned parameters of rule-based morphological analyzer.

From the series of experiments we found that even the parameters learned from inside test data fail to provide error rates less than 3%. It seems to show a limit of bigram-based HMM. We are now investigating a way to decide an appropriate set of HMM states.

Acknowledgements

We used the EDR Japanese corpus, the editorial articles of Asahi Newspaper and the newspaper articles of Nikkei Newspaper CD-ROM (1994 version) for the training and test corpora. We express sincere thanks to the permission of research use of the corpora.

References

[1] Chang, C.-H. and Chen, C.-D., HMM-based part-of-speech tagging for Chinese Corpora. *Proc. Workshop on Very Large Corpora*, pp.40-47, 1993.

[2] Charniak, E., Statistical Language Learning. *MIT Press*, 1993.

[3] Charniak, E., Hendrickson, C. Jacobson, N. and Perkowitz, M., Equations for Part-of-Speech Tagging. *Proc. the Eleventh National Conference on Artificial Intelligence (AAAI-93)*, pp.784-789, 1993.

[4] Church, K., A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Proc. ACL 2nd Conference on Applied Natural Language Processing*, pp.136–143, 1988.

[5] Cutting, D., Kupiec, J. Pedersen, J. and Sibun, P., A Practical Part-of-speech Tagger. *Proc. 3rd Conference on Applied Natural Language Processing*, pp.133–143, 1992.

[6] EDR Japanese Corpus, version 1. Japan Electronic Dictionary Research Institute. 1995.

[7] Huang, X.D., Ariki, Y.M. and Jack, A., Hidden Markov Models for Speech Recognition. Edinburgh University Press. 1990.

[8] Matsumoto, Y., et al., Japanese Morphological Analyzer JUMAN Manual (in Japanese). *Nara Institute of Science and Technology*, Technical Report NAIST-IS-TR94025, 1994.