Hidden Markov Model-Based Korean Part-of-Speech Tagging Considering High Agglutinativity, Word-Spacing, and Lexical Correlativity

Sang-Zoo Lee and **Jun-ichi Tsujii**
Department of Information Science
University of Tokyo
Hongo 7-3-1, Bunkyo-ku
Tokyo 113-0033, Japan
{lee,tsujii}@is.s.u-tokyo.ac.jp

Hae-Chang Rim
Department of Computer Science
Korea University
1 5-Ga Anam-Dong, Seongbuk-Gu
Seoul 136-701, Korea
rim@nlp.korea.ac.kr

Abstract

In this paper we present hidden Markov models for Korean part-of-speech tagging, which consider Korean characteristics such as high agglutinativity, word-spacing, and high lexical correlativity. In order ot consider rich information in contexts, the models adopt a less strict Markov assumption. In the models, sparse-data problem is very serious and their parameters tend to be estimated unreliably because they have a large number of parameters. To overcome sparse-data problem, our model uses a simplified version of the well-known back-off smoothing method. To mitigate unreliable estimation problem, our models assume joint independence instead of conditional independence because joint probabilities have the same degree of estimation reliability. Experimental results show that models with rich contexts perform even better than standard HMMs and that joint independent assumption is effective in some models.

1 Introduction

Korean is an highly agglutinative language which has word-spacing orthography. It makes Korean part-of-speech (POS) tagging different from English POS tagging. Generally English POS tagging can be regarded as a process in which a proper POS tag is assigned to each word in texts. However, in Korean POS tagging, each word is tagged with a proper combination of categories and lexical forms of morphemes(Lee et al., 1999) because Korean words can be freely formed by agglutinating morphemes and so the number of categories of Korean words can be (theoretically) infinite.

Over a decade, many works for Korean POS tagging have used a wide range of machine learning techniques such as a hidden Markov model (HMM) (Lee, 1995)(Kim et al., 1998), a maximum entropy model (Kang, 1998), transformation rules (Lim, 1997), a decision tree (Lee et al., 1999), discriminative learning (Kim et al., 1995), a fuzzy net (Kim et al., 1993), a neural network (Lee, 1994), and so on.

In this paper we propose hidden Markov models for Korean POS tagging, which adopt a less strict Markov assumption(Cinlar, 1975) to consider rich contexts and which consider Korean characteristics such as high agglutinativity, word-spacing, and high lexical correlativity. In the models, sparse-data problem is very serious because they have a large number of parameters. To overcome sparse-data problem, our model uses a simplified version of the well-known back-off smoothing method. If the parameters are very specific like lexicalized ones, they tend to have very different estimation reliability, making the Markov assumption implausible. To mitigate this problem, our models assume joint independence between random variables instead of conditional independence because joint probabilities have the same degree of estimation reliability. Experimental results for the KUNLP corpus (Lee et al., 1999) show that models

with rich contexts perform even better than standard HMMs and that joint independent assumption is effective in some models.

2 Lexical correlativity of Korean

In Korean, the same word form can be made from different morpheme sequences with the same tag sequence. For instance, a word form *Na-Neun* can correspond to two different morpheme sequences with the same tag sequence, *Na*/V(=to sprout)+*Neun*/E(=case marker) and *Nal*/V(=to fly)+*Neun*/E[1]. We call this ambiguity "homo-categorial" ambiguity.

Usually homo-categorial ambiguity is not easy to resolve without consulting lexical information in contexts. For example, *Na-Neun* is tagged with *Na*/V+*Neun*/E in "*SSag-i Na-Neun Jung-i-Da (= A bud is sprouting)*" and with *Nal*/V+*Neun*/E in "*Sae-Ga Na-Neun Jung-i-Da (= A bird is flying)*". Because these sentences have the same tag context "N+P V+E N+I+E"[2], they cannot be discriminated by considering only POS tag information in contexts. Moreover, although two lexical probabilities, $Pr(Na \mid V)$ and $Pr(Nal \mid V)$, are considered, the word can not be correctly tagged since the tag with larger probability is always selected in both sentences.

However, such ambiguity can be resolved by referring lexical relations in contexts. For example, *Na-Neun* can be correctly tagged if we consider lexical relations between *SSag-Gi* and *Na-Neun* and between *Sae-Ga* and *Na-Neun*.

3 HMM-based Korean POS tagging

Figure 1 shows a morpheme-unit lattice structure of a Korean sentence, "*Neo-Neun Hal Su iss-Da.*" , where each node has a morpheme and its POS tag and where the sequence connected by bold lines indicates the most likely sequence. Because Korean has word-spacing orthography, transitions between nodes can be distinguished by a word boundary. Transitions across a word boundary, which are depicted by a solid line, are distinguished from transitions within a word, which are depicted by a dotted line.

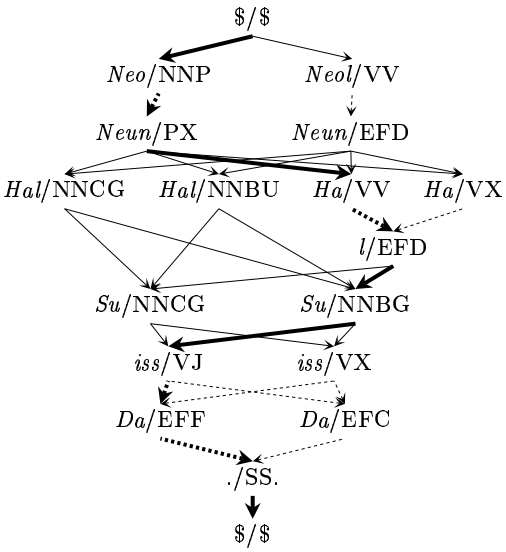

Figure 1: A morpheme-unit lattice of "*NeoNeun Hal Su issDa.*"(= You can do it.)

3.1 Standard word-unit model

We basically follow the notation of (Charniak et al., 1993) to describe Bayesian models. In this paper, we assume that $\{w^1, w^2, \ldots, w^\omega\}$ is a set of words, $\{t^1, t^2, \ldots, t^\tau\}$ is a set of POS tags, a sequence of random variables $W_{1,n} = W_1\ W_2\ \ldots\ W_n$ is a sentence of n words, a sequence of random variables $T_{1,n} = T_1\ T_2\ \ldots\ T_n$ is a sequence of n word categories. Because each of random variables W can take as its value any of the words in the vocabulary, we denote the value of W_i by w_i and a particular sequence of values for $W_{i,j}$ $(i \leq j)$ by $w_{i,j}$. In a similar way, we denote the value of T_i by t_i and a particular sequence of values for $T_{i,j}$ $(i \leq j)$ by $t_{i,j}$. For generality, terms $w_{i,j}$ and $t_{i,j}$ $(i > j)$ are defined as being empty.

The purpose of Bayesian models for POS tagging is to find the most likely sequence of POS tags for a given sequence of words, as follows:

$T(w_{1,n})$

[1] V denotes a verbal stem, and E a verbal ending.

[2] N denotes a noun, P a postposition, and I a copula.

$$= \underset{t_{1,n}}{\operatorname{argmax}} \Pr(T_{1,n} = t_{1,n} \mid W_{1,n} = w_{1,n}) \quad (1)$$

$$= \underset{t_{1,n}}{\operatorname{argmax}} \Pr(t_{1,n} \mid w_{1,n}) \quad (2)$$

$$= \underset{t_{1,n}}{\operatorname{argmax}} \frac{\Pr(t_{1,n}, w_{1,n})}{\Pr(w_{1,n})}$$

$$= \underset{t_{1,n}}{\operatorname{argmax}} \Pr(t_{1,n}, w_{1,n}) \quad (3)$$

Eqn. 1 becomes Eqn. 2 because reference to the random variables themselves can be omitted. Eqn. 2 is then transformed into Eqn. 3 since $\Pr(w_{1,n})$ is constant for all $t_{1,n}$.

Then, the probability $\Pr(t_{1,n}, w_{1,n})$ is broken down into Eqn. 4 by using the chain rule.

$$\Pr(t_{1,n}, w_{1,n}) = \prod_{i=1}^{n} \left(\begin{array}{l} \Pr(t_i \mid t_{1,i-1}, w_{1,i-1}) \\ \times \Pr(w_i \mid t_{1,i}, w_{1,i-1}) \end{array} \right) \quad (4)$$

However, it is either implausible or impossible to compute $\Pr(t_i \mid t_{1,i-1}, w_{1,i-1})$ and $\Pr(w_i \mid t_{1,i}, w_{1,i-1})$ in Eqn. 4.

The standard HMM simplifies them by making the following two strict Markov assumption (conditional independence), Eqn. 5 and Eqn. 6, to get a more tractable form, Eqn. 7.

$$\Pr(t_i \mid t_{1,i-1}, w_{1,i-1}) \approx \Pr(t_i \mid t_{i-K,i-1}) \quad (5)$$

$$\Pr(w_i \mid t_{1,i}, w_{1,i-1}) \approx \Pr(w_i \mid t_i) \quad (6)$$

$$\Pr(t_{1,n}, w_{1,n}) \approx \prod_{i=1}^{n} \left(\begin{array}{l} \Pr(t_i \mid t_{i-K,i-1}) \\ \times \Pr(w_i \mid t_i) \end{array} \right) \quad (7)$$

The standard HMM assumes that the probability of a current tag t_i conditionally depends on only the previous K tags $t_{i-K,i-1}$ and that the probability of a current word w_i conditionally depends on only the current tag[3].

Generally, the standard HMM has a limitation that it can not solve complicated ambiguities because it does not consider rich contexts. To overcome this limitation, the standard HMM should be extended so that it can consult rich information in contexts.

Moreover, the standard word-unit model can not be used effectively for tagging highly agglutinative languages like Korean. Therefore, the word-unit model should be transformed into a morpheme-unit model.

[3]Usually, K is determined as 1 (bigram as in (Charniak et al., 1993)) or 2 (trigram as in (Merialdo, 1991)).

3.2 Extended morpheme-unit model

Bayesian models for morpheme-unit tagging find the most likely sequence of morphemes and corresponding tags for a given sequence of words, as follows:

$$T(w_{1,n}) = \underset{c_{1,u}, m_{1,u}}{\operatorname{argmax}} \Pr(c_{1,u}, m_{1,u} \mid w_{1,n}) \quad (8)$$

$$\approx \underset{c_{1,u}, m_{1,u}}{\operatorname{argmax}} \Pr(c_{1,u}, p_{2,u}, m_{1,u}) \quad (9)$$

In the above equations, $u(\geq n)$ denotes the number of morphemes in a sequence corresponding the given word sequence, c denotes a morpheme-unit tag, m denotes a morpheme, and p denotes a type of transition from the previous tag to the current tag. p can have one of two values, "#" denoting a transition across a word boundary and "+" denoting a transition within a word. Because it is difficult to calculate Eqn. 8, the word sequence term $w_{1,n}$ is usually ignored as in Eqn. 9. Instead, we introduce p in Eqn. 9 to consider word-spacing[4].

The probability $\Pr(c_{1,u}, p_{2,u}, m_{1,u})$ is also broken down into Eqn. 10 by using the chain rule.

$$\Pr(c_{1,u}, p_{2,u}, m_{1,u}) = \prod_{i=1}^{u} \left(\begin{array}{l} \Pr(c_i, p_i \mid c_{1,i-1}, p_{2,i-1}, m_{1,i-1}) \\ \times \Pr(m_i \mid c_{1,i}, p_{2,i}, m_{1,i-1}) \end{array} \right) \quad (10)$$

Because Eqn. 10 is not easy to compute, it is simplified by making a Markov assumption to get a more tractable form.

An extended HMM for morpheme-unit tagging can be defined by making a less strict Markov assumption, as follows:

$$\Lambda(C_{[s](K,J)}, M_{[s](L,I)}) \models \Pr(c_{1,u}, p_{2,u}, m_{1,u})$$

$$\approx \prod_{i=1}^{u} \begin{array}{l} \Pr(c_i[, p_i] \mid \begin{array}{l} c_{i-K,i-1}[, p_{i-K+1,i-1}], \\ m_{i-J,i-1} \end{array}) \\ \times \Pr(m_i \mid c_{i-L,i}[, p_{i-L+1,i}], m_{i-I,i-1}) \end{array} \quad (11)$$

In a model $\Lambda(C_{[s](K,J)}, M_{[s](L,I)})$, the probability of the current morpheme tag c_i conditionally depends on both the previous K tags

[4]Most previous HMM-based Korean taggers except (Kim et al., 1998) did not consider word-spacing.

$c_{i-K,i-1}$ (optionally, the types of their transition $p_{i-K+1,i-1}$) and the previous J morphemes $m_{i-J,i-1}$ and the probability of the current morpheme m_i conditionally depends on the current tag and the previous L tags $c_{i-L,i}$ (optionally, the types of their transition $p_{i-L+1,i}$) and the previous I morphemes $m_{i-I,i-1}$. In experiments, we set K as 1 or 2, J as 0 or K, L as 1 or 2, and I as 0 or L. If J and I are zero, the above models are non-lexicalized models. Otherwise, they are lexicalized models.

For example, the extended model $\Lambda(C_{s(2,2)}, M_{(2,2)})$, where word-spacing is considered only in the tag probabilities, calculate the probability of a node "*Su*/NNBG" of the most likely sequence in Figure 1 as follows:

$$\Pr(NNBG, \# \mid VV, EFD, +, Ha, l) \times \Pr(Su \mid VV, EFD, NNBG, Ha, l)$$

4 Parameter estimation

The extended models have a large number of parameters, as compared to the standard models. Therefore, they must suffer from both sparse-data problem and unreliable estimation problem. The models adopt a simplified back-off smoothing technique as a solution to the first problem, and joint independence assumption as a solution to the second.

4.1 Simplified back-off smoothing

In supervised learning, the simpliest parameter estimation is the maximum likelihood(ML) estimation(Duda et al., 1973) which maximizes the probability of a training set. The ML estimate of morpheme tag $(K+1)$-gram probability, $\Pr_{ML}(c_i \mid c_{i-K,i-1})$, is calculated as follows:

$$\Pr_{ML}(c_i \mid c_{i-K,i-1}) = \frac{\mathrm{Fq}(c_{i-K,i})}{\mathrm{Fq}(c_{i-K,i-1})} \qquad (12)$$

where the function $\mathrm{Fq}(x)$ returns the frequency of x in the training set. When using the ML estimation, data sparseness is even more serious in the extended models than in the standard models because the former has even more parameters than the latter.

In (Chen, 1996), where various smoothing techniques was tested for a language model by using the perplexity measure, it was reported that the back-off smoothing method(Katz, 1987) performs better on a small traning set than other methods. In the back-off smoothing, the smoothed probability of tag $(K+1)$-gram $\Pr_{SBO}(c_i \mid c_{i-K,i-1})$ is calculated as follows:

$$\Pr_{SBO}(c_i \mid c_{i-K,i-1}) = \begin{cases} d_r \Pr_{ML}(c_i \mid c_{i-K,i-1}) & \text{if } r > 0 \\ \alpha(c_{i-K,i-1}) \Pr_{SBO}(c_i \mid c_{i-K+1,i-1}) & \text{if } r = 0 \end{cases} \qquad (13)$$

where $r = \mathrm{Fq}(c_{i-K,i})$, $r^* = (r+1)\frac{n_{r+1}}{n_r}$

$$d_r = \frac{\frac{r^*}{r} - \frac{(r+1)\times n_{r+1}}{n_1}}{1 - \frac{(r+1)\times n_{r+1}}{n_1}}$$

In the equation above, n_r denotes the number of $(K+1)$-gram whose frequency is r, and the coefficient d_r is called the discount ratio, which reflects the Good-Turing estimate(Good, 1953)[5]. Eqn. 13 says that $\Pr_{SBO}(c_i \mid c_{i-K,i-1})$ is under-estimated by d_r than its maximum likelihood estimate, if $r > 0$, or is backed off by its smoothing term $\Pr_{SBO}(c_i \mid c_{i-K+1,i-1})$ in proportion to the value of the function $\alpha(c_{i-K,i-1})$ of its conditional term $c_{i-K,i-1}$, if $r = 0$.

However, because Eqn. 13 requires complicated computation in $\alpha(c_{i-K,i-1})$, we simplify it to get a function of the frequency of a conditional term, as follows:

$$\alpha(\mathrm{Fq}(c_{i-K,i-1}) = f) = \Delta \times \frac{\mathrm{E}[\mathrm{Fq}(c_{i-K,i-1}) = f]}{\sum_{f=0}^{\infty} \mathrm{E}[\mathrm{Fq}(c_{i-K,i-1}) = f]} \qquad (14)$$

where

$$\Delta = 1 - \frac{\sum_{c_{i-K,i}, r>0} \Pr_{SBO}(c_i | c_{i-K,i-1})}{\sum_{c_{i-K,i}, r>0} \Pr_{ML}(c_i | c_{i-K,i-1})},$$

$$\mathrm{E}[\mathrm{Fq}(c_{i-K,i-1}) = f] = \sum_{c_{i-K+1,i}, r=0, Fq(c_{i-K,i-1})=f} \Pr_{SBO}(c_i | c_{i-K+1,i-1})$$

In Eqn. 14, the range of f is bucketed into 7 regions such as $f = 0, 1, 2, 3, 4, 5$ and $f \geq 6$

[5] In (Katz, 1987) $d_r = 1$ if $r > 5$.

since it is also difficult to compute this equation for all possible values of f.

In the formalism of the simplified back-off smoothing, each probability whose ML estimate is zero is backed off by its corresponding smoothing term. In experiments, the smoothing term of $\Pr_{SBO}(c_i[,p_i] \mid c_{i-K,i-1}[,p_{i-K+1,i-1}], m_{i-J,i-1})$ is determined as follows:

$$\Pr_{SBO}(c_i[,p_i] \mid \begin{matrix} c_{i-K+1,i-1} \\ [,p_{i-K+2,i-1}], \\ m_{i-J+1,i-1} \end{matrix}) \quad \text{if } \begin{matrix} K \geq 1, \\ J > 1 \end{matrix}$$

$$\Pr_{SBO}(c_i[,p_i] \mid \begin{matrix} c_{i-K,i-1} \\ [,p_{i-K+1,i-1}] \end{matrix}) \quad \text{if } \begin{matrix} K \geq 1, \\ J = 1 \end{matrix}$$

$$\Pr_{SBO}(c_i[,p_i] \mid \begin{matrix} c_{i-K+1,i-1} \\ [,p_{i-K+2,i-1}] \end{matrix}) \quad \text{if } \begin{matrix} K > 1, \\ J = 0 \end{matrix}$$

$$\Pr_{AD}(c_i) \quad \text{if } \begin{matrix} K = 1, \\ J = 0 \end{matrix}$$

The smoothing term of $\Pr_{SBO}(m_i \mid c_{i-L,i}[,p_{i-L+1,i}], m_{i-I,i-1})$ is determined as follows:

$$\Pr_{SBO}(m_i \mid \begin{matrix} c_{i-L+1,i} \\ [,p_{i-L+2,i}], \\ m_{i-I+1,i-1} \end{matrix}) \quad \text{if } L \geq 1, I > 1$$

$$\Pr_{SBO}(m_i \mid \begin{matrix} c_{i-L,i} \\ [,p_{i-L+1,i}] \end{matrix}) \quad \text{if } L \geq 1, I = 1$$

$$\Pr_{SBO}(m_i \mid \begin{matrix} c_{i-L+1,i} \\ [,p_{i-L+2,i}] \end{matrix}) \quad \text{if } L \geq 1, I = 0$$

$$\Pr_{AD}(m_i) \quad \text{if } L = 0, I = 0$$

In the equations above, the unigram probabilities are calculated by using the additive smoothing with $\delta = 10^{-2}$, which is chosen through experiments. The equation for the additive smoothing(Chen, 1996) is as follows:

$$\Pr_{AD}(c_i \mid c_{i-K,i-1}) = \frac{\mathrm{Fq}(c_{i-K,i}) + \delta}{\sum_{c_i}(\mathrm{Fq}(c_{i-K,i}) + \delta)}$$

4.2 Joint independence

The parameters of an HMM may have different degree of statistical reliability because parameter reliability depends on the frequency of conditional term. For example, let a corpus consist of 1 million words and let the following parameters be extracted from the corpus by using the maximum likelihood estimation.

$$\Pr(a) = 0.01 \quad \Pr(d \mid a) = 0.1$$
$$\Pr(b) = 0.001 \quad \Pr(d \mid b) = 0.1$$
$$\Pr(c) = 0.0001 \quad \Pr(d \mid c) = 0.1$$

In this case, three conditional probabilities, $\Pr(d \mid a)$, $\Pr(d \mid b)$, and $\Pr(d \mid c)$ are all 0.1 but $\Pr(d \mid a)$ is statistically more reliable than others because its sample size (10,000 words = 1 million$\times \Pr(a)$) is bigger than others. Actually, this problem becomes very serious in extended models, even though parameters of the models are seen in the training corpus.

To consider such statistical reliability of a probability estimate, we introduce the concept of weighting Markov assumption, as follows:

$$\Pr(c_i \mid c_{1,i-1}, m_{1,i-1}) \approx \Pr(c_i \mid c_{i-K,i-1}, m_{i-J,i-1}) \times \mathrm{W}(c_{i-K,i-1}, m_{i-J,i-1}) \quad (15)$$

$$\Pr(m_i \mid c_{1,i}, m_{1,i-1}) \approx \Pr(m_i \mid c_{i-L,i}, m_{i-I,i-1}) \times \mathrm{W}(c_{i-L,i}, m_{i-I,i-1}) \quad (16)$$

If the probability function, Pr, is used as the weight function, W, the equations above become equations assuming joint independence between random variables as follows:

$$\Pr(c_i \mid c_{1,i-1}, m_{1,i-1}) \approx \Pr(c_i, c_{i-K,i-1}, m_{i-J,i-1}) \quad (17)$$

$$\Pr(m_i \mid c_{1,i}, m_{1,i-1}) \approx \Pr(m_i, c_{i-L,i}, m_{i-I,i-1}) \quad (18)$$

The equations above assume that the probability of the current morpheme tag c_i jointly depends on both the previous K tags $c_{i-K,i-1}$ and the previous J words $m_{i-J,i-1}$ and that the probability of the current word m_i jointly depends on the current tag and the previous L tags $c_{i-L,i}$ and the previous I words $m_{i-I,i-1}$. If a Bayesian model assumes joint independence, we call it a joint independence model (JIM).

Actually, using the probability function as the weight function is mathematically incorrect and implausible. For example, while the sum of probabilities of all sentences with the same length becomes 1.0 in an HMM, it becomes naturally less than 1.0 in a JIM. Therefore, JIMs should not be used in calculating the probability of a sentence. However, if we want to find the most likely sequence for each sentence and the joint probability of each pa-

rameter is regarded as a score, JIMs work well.

By replacing corresponding parameters, an extended morpheme-unit HMM can be transformed into the corresponding JIM, which is defined as follows:

$$\Phi(C_{[s](K,J)}, M_{[s](L,I)}) \models \Pr(c_{1,u}, p_{2,u}, m_{1,u}) \approx \prod_{i=1}^{u} \Pr(c_i[, p_i], \begin{matrix} c_{i-K,i-1}[, p_{i-K+1,i-1}], \\ m_{i-J,i-1} \end{matrix}) \times \Pr(m_i, c_{i-L,i}[, p_{i-L+1,i}], m_{i-I,i-1}) \quad (19)$$

In the extended JIM, $\Phi(C_{s(2,2)}, M_{(2,2)})$, the probability of a node "*Su*/NNBG" of the most likely sequence in Figure 1 is calculated as follows:

$$\Pr(NNBG, \#, VV, EFD, +, Ha, l) \times \Pr(Su, VV, EFD, NNBG, Ha, l)$$

The parameters of a JIM are estimated by using the parameters of the corresponding HMM as follows:

$$\Pr_{SBO}(c_i[, p_i], \begin{matrix} c_{i-K,i-1}[, p_{i-K+1,i-1}], \\ m_{i-J,i-1} \end{matrix}) = \Pr_{SBO}(c_i[, p_i] \mid \begin{matrix} c_{i-K,i-1}[, p_{i-K+1,i-1}], \\ m_{i-J,i-1} \end{matrix}) \times \Pr_{AD}(\begin{matrix} c_{i-K,i-1}[, p_{i-K+1,i-1}], \\ m_{i-J,i-1} \end{matrix})$$

$$\Pr(m_i, c_{i-L,i}[, p_{i-L+1,i}], m_{i-I,i-1}) = \Pr(m_i \mid c_{i-L,i}[, p_{i-L+1,i}], m_{i-I,i-1}) \times \Pr(c_{i-L,i}[, p_{i-L+1,i}], m_{i-I,i-1})$$

$$\Pr_{AD}(c_{i-K,i}) = \frac{\mathrm{Fq}(c_{i-K,i}) + \delta}{\sum_{c_{i-K,i}} (\mathrm{Fq}(c_{i-K,i}) + \delta)}$$

5 Experiments

In experiments, we used the KUNLP corpus which consists of 167,115 words and 15,211 sentences and is tagged with 65 POS tags. It was segmented into two parts, the training set of 90% and the test set of 10%, in the way that each sentence in the test set was extracted from every 10 sentence. In the same way, we made 10-fold data set for 10-fold cross validation.

In order to morphologically analyze each word, we used the Korean morphological analyzer (Lee, 1999) which is consistent with the KUNLP corpus. By using the morphological analyzer, the average number of possible analyses per word becomes 3.41.

Figure 2-5 illustrate graphs showing the average accuracy rates of HMMs and JIMs, without considering word-spacing, with considering word-spacing only in the lexical probabilities, with considering word-spacing only in the tag probabilities, and with considering word-spacing in both the tag and lexical probabilities, respectively. Here, labels in x-axis specify models in the way that $\frac{K,J}{L,I}$ denotes $\Lambda(C_{[s](K,J)}, M_{[s](L,I)})$ or $\Phi(C_{[s](K,J)}, M_{[s](L,I)})$. The models are arranged by the ascending order of theoretical number of parameters. The first two models are standard models and the others are extended models. The average accuracy rates beyond the range of each graph are intentionally omitted.

In these figures, we can observe that the simplified back-off smoothing technique mitigates sparse-data problems in both HMMs and JIMs. As expected, JIMs achieves higher accuracy than the corresponding HMMs in some extended models consulting rich contexts. Consulting word-spacing makes slight improvement in some of both HMMs and JIMs. It is statistically significant with confidence 99that the best model, $\Lambda(C_{s(2,2)}, M_{s(2,2)})$ (96.97%), is better than any other models including the previous standard model $\Lambda(C_{(1,0)}, M_{(0,0)})$ (94.95%)(Lee, 1995), the previous model $\Lambda(C_{s(1,0)}, M_{(0,0)})$ (94.96%) (Kim et al., 1998), and the best JIM, $\Lambda(C_{(1,1)}, M_{s(1,1)})$ (96.95%).

6 Conclusion

We have presented the extended HMMs for Korean POS tagging, which assume joint independence between random variables, which are based on the morpheme-unit lattice structure, and which consider word-spacing and rich information in contexts. In the models, a simplified version of back-off smoothing is used to mitigate data sparseness problem.

From the experimental results, we have observed that extended models achieved even better results than the standard models in case of both HMMs and JIMs, that the simpli-

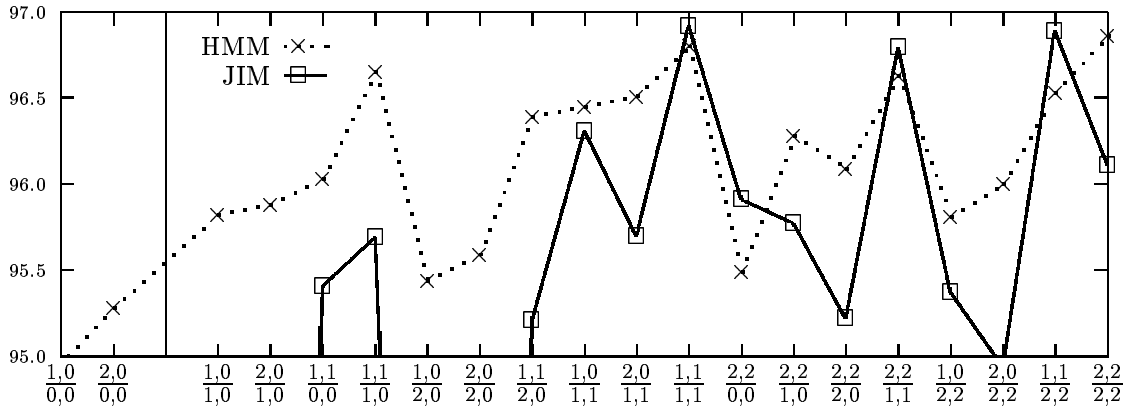

Figure 2: Without considering word-spacing

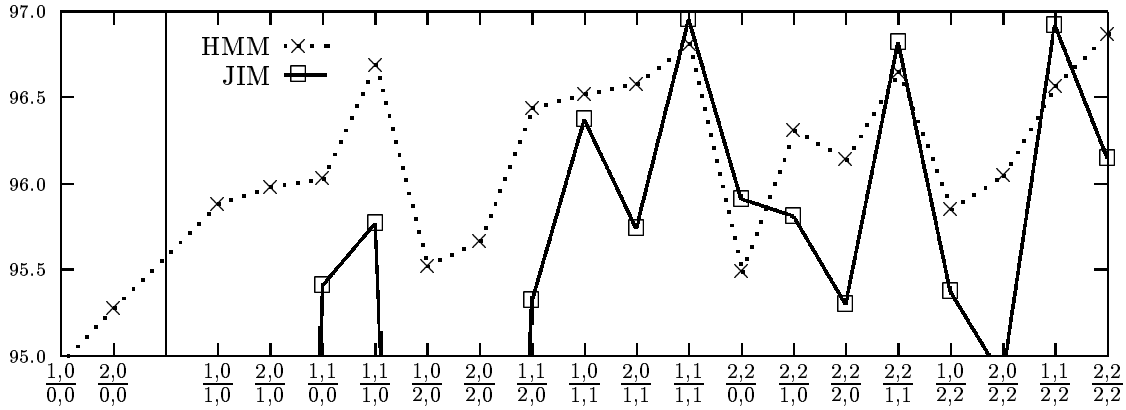

Figure 3: With considering word-spacing only in the lexical probabilities

fied back-off smoothing technique mitigated data sparseness quite effectively, that consulting word-spacing made slight improvement of accuracy, and that some extended JIMs outperformed the corresponding HMMs.

Now, we are implementing and evaluating various smoothing techniques in order to find more effective smoothing technique for HMM/JIM-based Korean POS tagging. And also, we are trying to apply JIMs to different areas such as information extraction in the bio-molecular domain, noun phrase chuncking, and so on.

References

E. Charniak, C. Hendrickson, N. Jacobson, and M. Perkowitz. 1993. Equations for Part-of-Speech Tagging. In *Proc. of the 11th Nat'l Conf. on Artificial Intelligence(AAAI-93)*, 784–789.

S. F. Chen. 1996. *Building Probabilistic Models for Natural Language.* Doctoral Dissertation, Harvard University, USA.

E. Cinlar. 1975. *Introduction to Stochastic Processes.* Prentice-Hall, New Jersey.

R. O. Duda and R. E. Hart. 1973. *Pattern Classification and Scene Analysis.* John Wiley.

I. J. Good. 1953. "The Population Frequencies of Species and the Estimation of Population Parameters," In *Biometrika*, 40(3-4):237–264.

S. M. Katz. 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. In *IEEE Transactions on Acoustics, Speech and Signal Processing(ASSP)*, 35(3):400–401.

I.-H. Kang, J.-H. Kim, and G.-C. Kim. 1998. Korean Part-of-Speech Tagging Using Maximum Entropy Model. In *Proc. of the 10th National Conference on Korean Information Processing*, 9–14.

J.-H. Kim, et al. 1993. Korean Part-of-Speech Tagging by Using a Fuzzy net. In *Proc. of the*

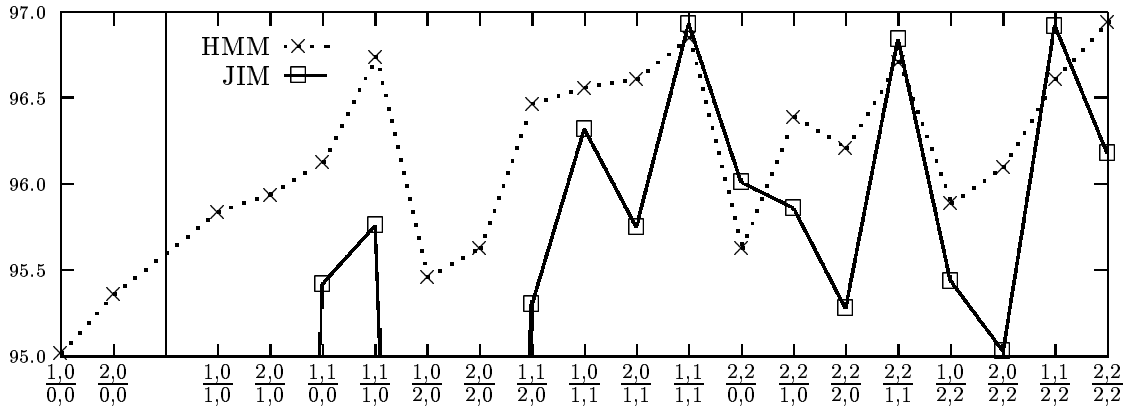

Figure 4: With considering word-spacing only in the tag probabilities

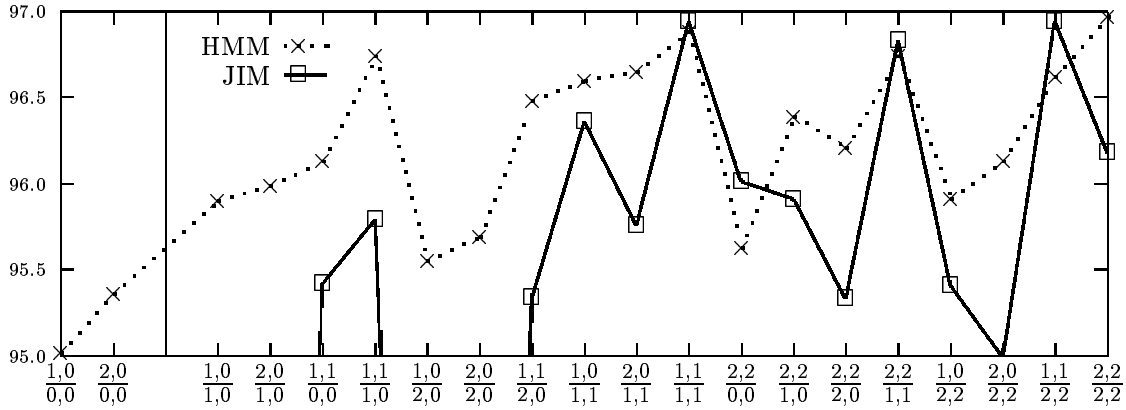

Figure 5: With considering word-spacing in both the tag and lexical probabilities

5th National Conference on Korean Information Processing, 593–603.

J.-H. Kim and G.-C. Kim. 1995. Discriminative Learning in Part-of-Speech Tagging. In *Proc. of the National Conference on Korean Information Science Society*, Spring, 627–630.

J.-D. Kim, S.-Z. Lee, and H.-C. Rim. 1998. A Morpheme-Unit POS Tagging Model Considering Word-Spacing. In *Proc. of the 10th National Conference on Korean Information Processing*, 3–8.

S.-J. Lee. 1994. *Prediction and Disambiguation of Korean Word Category by Using a Neural Network.* Doctoral Dissertation, Seoul National University, Korea.

S.-H. Lee. 1995. *Korean POS Tagging System Considering Unknown Words.* Master Thesis, Korea Advanced Institute of Science and Technology(KAIST), Korea.

S.-Z. Lee, J.-D. Kim, W.-H. Ryu, and H.-C. Rim. 1999. A Part-of-Speech Tagging Model Using Lexical Rules Based on Corpus Statistics. In *Proc. of the International Conference on Computer Processing of Oriental Languages(ICCPOL-99)*, 385–390.

S.-Z. Lee. 1999. *New Statistical Models for Automatic POS Tagging.* Doctoral Dissertation, Korea University, Korea.

H.-S. Lim. 1997. *Korean Part-of-Speech Tagging by Using Lingusitic Knowledge and Statistical Information.* Doctoral Dissertation, Korea University, Korea.

B. Merialdo. 1991. Tagging Text with a Probabilistic Model. In *Proc. of the International Conference on Acoustic, Speech and Signal Processing(ICASSP-91)*, 809–812.