Part-of-Speech Tagging Based on Hidden Markov Model Assuming Joint Independence

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 part-of-speech taggers based on hidden Markov models, which adopt a less strict Markov assumption to consider rich contexts. In models whose parameters are very specific like lexicalized ones, sparse-data problem is very serious and also conditional probabilities tend to be estimated unreliably. To overcome data-sparseness, a simplified version of the well-known back-off smoothing method is used. To mitigate unreliable estimation problem, our models assume joint independence instead of conditional independence because joint probabilities have the same degree of estimation reliability. In experiments for the Brown corpus, models with rich contexts achieve relatively high accuracy and some models assuming joint independence show better results than the corresponding HMMs.

1 Introduction

Part-of-speech (POS) tagging can be defined as a process in which a proper POS tag is assigned to each word in texts and so it can be viewed as a classification problem (Mitchell, 1997). Over a decade, many works for POS tagging have used a wide range of machine learning techniques such as a hidden Markov model (HMM) (Charniak et al., 1993), a maximum entropy model (Ratnaparkhi, 1996), transformation rules (Brill, 1994), a decision tree (Lee et al., 1999), relaxation labeling (Padró, 1996), Bayesian inference (Samuelsson, 1993), discriminative learning (Lin, 1992), a neural network (Schmid, 1994), and so on.

In this paper we propose hidden Markov models for part-of-speech tagging, which adopt a less strict Markov assumption(Cinlar, 1975) to consider rich contexts. Because such models have a large number of parameters, they must suffer from sparse-data problem unless they have an enough volume of training corpus. Moreover, because such models assume conditional independence, the probability estimates of their parameters may have statistically different reliability that depends on the number of samples of conditional terms. To overcome the first problem, a simplified version of the well-known back-off smoothing method is used. To mitigate unreliable estimation problem, our models assume joint independence between random variables instead of conditional independence because joint probabilities have the same degree of estimation reliability.

2 HMM-based POS tagging

Figure 1 shows a lattice structure of an English sentence, "Flies like a flower.", where each node has a word and its POS tag and where the sequence connected by bold lines indicates the most likely sequence.

2.1 Standard model

We basically follow the notation of (Charniak et al., 1993) to describe Bayesian models for POS tagging. In this paper, we assume that $\{w^1, w^2, \ldots, w^\omega\}$ is a set of

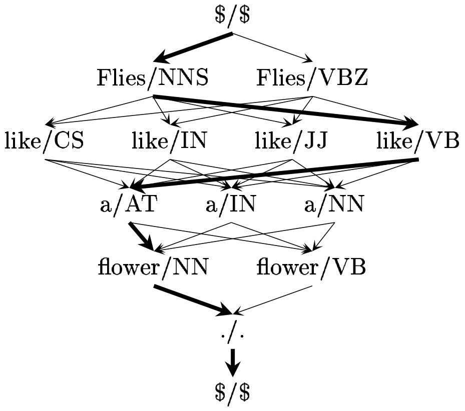

Figure 1: A lattice of "Flies like a flower ."

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, and a sequence of random variables $T_{1,n} = T_1 \; T_2 \; \ldots \; T_n$ is a sequence of n POS tags. 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:

$$\begin{aligned} T(w_{1,n}) & \\ &= \operatorname*{argmax}_{t_{1,n}} \Pr(T_{1,n} = t_{1,n} \mid W_{1,n} = w_{1,n}) \quad (1) \\ &= \operatorname*{argmax}_{t_{1,n}} \Pr(t_{1,n} \mid w_{1,n}) \quad (2) \\ &= \operatorname*{argmax}_{t_{1,n}} \frac{\Pr(t_{1,n}, w_{1,n})}{\Pr(w_{1,n})} \\ &= \operatorname*{argmax}_{t_{1,n}} \Pr(t_{1,n}, w_{1,n}) \quad (3) \end{aligned}$$

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[1]. In the standard model (K=1), for example, the probability of a node "a/AT" of the most likely sequence in Figure 1 is calculated as follows:

$$\begin{array}{l} \Pr(AT \mid NNS, VB) \\ \times \Pr(a \mid AT) \end{array}$$

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.

2.2 Extended models

An extended HMM, $\Lambda(T_{(K,J)}, W_{(L,I)})$, for POS tagging is defined by making the following two less strict Markov assumption, Eqn. 8 and Eqn. 9, as follows:

$$\Pr(t_i \mid t_{1,i-1}, w_{1,i-1}) \approx \Pr(t_i \mid t_{i-K,i-1}, w_{i-J,i-1}) \quad (8)$$

$$\Pr(w_i \mid t_{1,i}, w_{1,i-1}) \approx \Pr(w_i \mid t_{i-L,i}, w_{i-I,i-1}) \quad (9)$$

$$\begin{aligned} &\Lambda(T_{(K,J)}, W_{(L,I)}) \models \Pr(t_{1,n}, w_{1,n}) \\ &\approx \prod_{i=1}^{n} \left(\begin{array}{l} \Pr(t_i \mid t_{i-K,i-1}, w_{i-J,i-1}) \\ \times \Pr(w_i \mid t_{i-L,i}, w_{i-I,i-1}) \end{array} \right) \quad (10) \end{aligned}$$

In a model $\Lambda(T_{(K,J)}, W_{(L,I)})$, the probability of the current tag t_i conditionally depends on

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

both the previous K tags $t_{i-K,i-1}$ and the previous J words $w_{i-J,i-1}$ and the probability of the current word w_i conditionally depends on the current tag and the previous L tags $t_{i-L,i}$ and the previous I words $w_{i-I,i-1}$. 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.

In an extended model $\Lambda(T_{(2,2)}, W_{(2,2)})$, for example, the probability of a node "a/AT" of the most likely sequence in Figure 1 is calculated as follows:

$$\Pr(AT \mid NNS, VB, Flies, like)$$
$$\times \Pr(a \mid AT, NNS, VB, Flies, like)$$

3 Parameter estimation

Because the extended models have a large number of parameters, 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.

3.1 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 tag $(K+1)$-gram probability, $\Pr_{ML}(t_i \mid t_{i-K,i-1})$, is calculated as follows:

$$\Pr_{ML}(t_i \mid t_{i-K,i-1}) = \frac{\mathrm{Fq}(t_{i-K,i})}{\mathrm{Fq}(t_{i-K,i-1})} \qquad (11)$$

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.

(Chen, 1996), where various smoothing techniques was tested for a language model by using the perplexity measure, reported that a back-off smoothing(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 $\Pr_{SBO}(t_i \mid t_{i-K,i-1})$ is calculated as follows:

$$\Pr_{SBO}(t_i \mid t_{i-K,i-1}) =$$
$$\begin{cases} d_r \Pr_{ML}(t_i \mid t_{i-K,i-1}) & \text{if } r > 0 \\ \alpha(t_{i-K,i-1}) \Pr_{SBO}(t_i \mid t_{i-K+1,i-1}) & \text{if } r = 0 \end{cases} \qquad (12)$$

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

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

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)[2]. Eqn. 12 says that $\Pr_{SBO}(t_i \mid t_{i-K,i-1})$ is under-estimated by d_r than its maximum likelihood estimate, if $r > 0$, or is backed off by its smoothing term $\Pr_{SBO}(t_i \mid t_{i-K+1,i-1})$ in proportion to the value of the function $\alpha(t_{i-K,i-1})$ of its conditional term $t_{i-K,i-1}$, if $r = 0$.

However, because Eqn. 12 requires complicated computation in $\alpha(t_{i-K,i-1})$, we simplify it to get a function of the frequency of a conditional term, as follows:

$$\alpha(\mathrm{Fq}(t_{i-K,i-1}) = f) =$$
$$\Delta \times \frac{\mathrm{E}[\mathrm{Fq}(t_{i-K,i-1}) = f]}{\sum_{f=0}^{\infty} \mathrm{E}[\mathrm{Fq}(t_{i-K,i-1}) = f]} \qquad (13)$$

where

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

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

In Eqn. 13, the range of f is bucketed into 7 regions such as $f = 0, 1, 2, 3, 4, 5$ and $f \geq 6$ since it is also difficult to compute this equation for all possible values of f.

Using the formalism of our simplified back-off smoothing, each of probabilities whose ML estimate is zero is backed off by its corresponding smoothing term. In experiments, the smoothing terms of $\Pr_{SBO}(t_i \mid$

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

$t_{i-K,i-1}, w_{i-J,i-1})$ are determined as follows:

$$\begin{array}{ll}
\mathrm{Pr}_{SBO}(t_i \mid \begin{array}{c} t_{i-K+1,i-1}, \\ w_{i-J+1,i-1} \end{array}) & \text{if } K \geq 1, J > 1 \\
\mathrm{Pr}_{SBO}(t_i \mid t_{i-K,i-1}) & \text{if } K \geq 1, J = 1 \\
\mathrm{Pr}_{SBO}(t_i \mid t_{i-K+1,i-1}) & \text{if } K > 1, J = 0 \\
\mathrm{Pr}_{AD}(t_i) & \text{if } K = 1, J = 0
\end{array}$$

Also, the smoothing terms of $\mathrm{Pr}_{SBO}(w_i \mid t_{i-L,i}, w_{i-I,i-1})$ are determined as follows:

$$\begin{array}{ll}
\mathrm{Pr}_{SBO}(w_i \mid \begin{array}{c} t_{i-L+1,i}, \\ w_{i-I+1,i-1} \end{array}) & \text{if } L \geq 1, I > 1 \\
\mathrm{Pr}_{SBO}(w_i \mid t_{i-L,i}) & \text{if } L \geq 1, I = 1 \\
\mathrm{Pr}_{SBO}(w_i \mid t_{i-L+1,i}) & \text{if } L \geq 1, I = 0 \\
\mathrm{Pr}_{AD}(w_i) & \text{if } L = 0, I = 0
\end{array}$$

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

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

3.2 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.

$$\begin{array}{ll}
\Pr(a) = 0.01 & \Pr(d \mid a) = 0.1 \\
\Pr(b) = 0.001 & \Pr(d \mid b) = 0.1 \\
\Pr(c) = 0.0001 & \Pr(d \mid c) = 0.1
\end{array}$$

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× $\Pr(a)$) is bigger than others. Actually, this phenomenon is 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:

$$\begin{array}{l}
\Pr(t_i \mid t_{1,i-1}, w_{1,i-1}) \approx \\
\qquad \Pr(t_i \mid t_{i-K,i-1}, w_{i-J,i-1}) \\
\qquad \times \mathrm{W}(t_{i-K,i-1}, w_{i-J,i-1})
\end{array} \quad (14)$$

$$\begin{array}{l}
\Pr(w_i \mid t_{1,i}, w_{1,i-1}) \approx \\
\qquad \Pr(w_i \mid t_{i-L,i}, w_{i-I,i-1}) \\
\qquad \times \mathrm{W}(t_{i-L,i}, w_{i-I,i-1})
\end{array} \quad (15)$$

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:

$$\begin{array}{l}
\Pr(t_i \mid t_{1,i-1}, w_{1,i-1}) \approx \\
\qquad \Pr(t_i, t_{i-K,i-1}, w_{i-J,i-1})
\end{array} \quad (16)$$

$$\begin{array}{l}
\Pr(w_i \mid t_{1,i}, w_{1,i-1}) \approx \\
\qquad \Pr(w_i, t_{i-L,i}, w_{i-I,i-1})
\end{array} \quad (17)$$

The equations above assume that the probability of the current tag t_i jointly depends on both the previous K tags $t_{i-K,i-1}$ and the previous J words $w_{i-J,i-1}$ and that the probability of the current word w_i jointly depends on the current tag and the previous L tags $t_{i-L,i}$ and the previous I words $w_{i-I,i-1}$. 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 parameter is regarded as a score, JIMs have no problem.

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

$$\begin{array}{l}
\Phi(T_{(K,J)}, W_{(L,I)}) \models \Pr(t_{1,n}, w_{1,n}) \\
\approx \prod_{i=1}^{n} \left(\begin{array}{c} \Pr(t_i, t_{i-K,i-1}, w_{i-J,i-1}) \\ \times \Pr(w_i, t_{i-L,i}, w_{i-I,i-1}) \end{array} \right)
\end{array} \quad (18)$$

In an extended JIM, $\Phi(T_{(2,2)}, W_{(2,2)})$, for example, the probability of a node "a/AT" of the most likely sequence in Figure 1 is calculated as follows:

$$\Pr(AT, NNS, VB, Flies, like) \\ \times \Pr(a, AT, NNS, VB, Flies, like)$$

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

$$\Pr_{SBO}(t_i, t_{i-K,i-1}, w_{i-J,i-1}) = \\ \Pr_{SBO}(t_i \mid t_{i-K,i-1}, w_{i-J,i-1}) \\ \times \Pr_{AD}(t_{i-K,i-1}, w_{i-J,i-1})$$

$$\Pr_{SBO}(w_i, t_{i-L,i}, w_{i-I,i-1}) = \\ \Pr_{SBO}(w_i \mid t_{i-L,i}, w_{i-I,i-1}) \\ \times \Pr_{AD}(t_{i-L,i}, w_{i-I,i-1})$$

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

4 Experiments

For experiments, we used the Brown corpus which consists of 1,113,180 words and 53,885 sentences and is tagged with 82 POS tags[3]. 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 assign all possible tags to each word, we made two assumption: closed vocabulary assumption and open vocabulary assumption. For closed vocabulary assumption, we looked up a dictionary tailored to the Brown corpus. In this case, the average number of tags per word became 1.64. For open vocabulary assumption, we looked up a dictionary tailored only to a training set in order to assign possible tags to frequent words whose frequency is greater than 5. In case of rare words, tags in the dictionary were assigned and then 6 tags with highest score were assigned by using a naive Bayesian classifier(Mitchell, 1997) considering character features as follows:

$$\Pr(t_i, w_i) = \Pr(t_i) \times \Pr(w_i \mid t_i)$$

$$\approx \Pr(t_i) \times \prod_{j=1}^{F} \Pr(f_i^j \mid t_i)$$

where f_i^j indicates j-th character features of w_i and $F(=12)$ is the number of character feature types including prefixes (whose length is 1 through 4), suffixes (whose length is 1 through 4), if w_i contains numbers, if w_i contains an initial uppercase letter, if w_i contains any non-initial uppercase letter, if w_i contains hyphens. In this case, the average number of tags per word became 2.00 and the rate of words that have the correct tag among all assigned tags became 99.85%.

Figure 2 illustrates graphs showing the average accuracy rates of HMMs and JIMs under the closed vocabulary assumption. Here, labels in the x-axis specify models in the way that $\frac{K,J}{L,I}$ denotes $\Lambda(T_{(K,J)}, W_{(L,I)})$ or $\Phi(T_{(K,J)}, W_{(L,I)})$. 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 just below the figure.

In this figure, 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. It is statistically significant with confidence 99that the model, $\Phi(T_{(2,2)}, W_{(1,1)})$ (98.05%), is better than any other models including the standard bigram HMM, $\Lambda(T_{(1,0)}, W_{(0,0)})$ (97.27%) and the best HMM, $\Lambda(T_{(1,1)}, W_{(1,0)})$ (97.93%).

Figure 3 depicts graphs indicating the average accuracy rates of HMMs and JIMs under the open vocabulary assumption. Unlike Figure 2, the model, $\Lambda(T_{(2,0)}, W_{(1,0)})$, achieves the best accuracy rate (96.86%) with confidence 99%.

5 Conclusion

We have presented the extended HMMs for English POS tagging, which can consider rich

[3] Note that some sentences, which have composite tags(such as "HV+TO" in "*hafta*"), "ILLEGAL" tag, or "NIL" tag, were removed from the Brown corpus and tags with "*"(not) such as "BEZ*" were replaced by corresponding tags without "*" such as "BEZ".

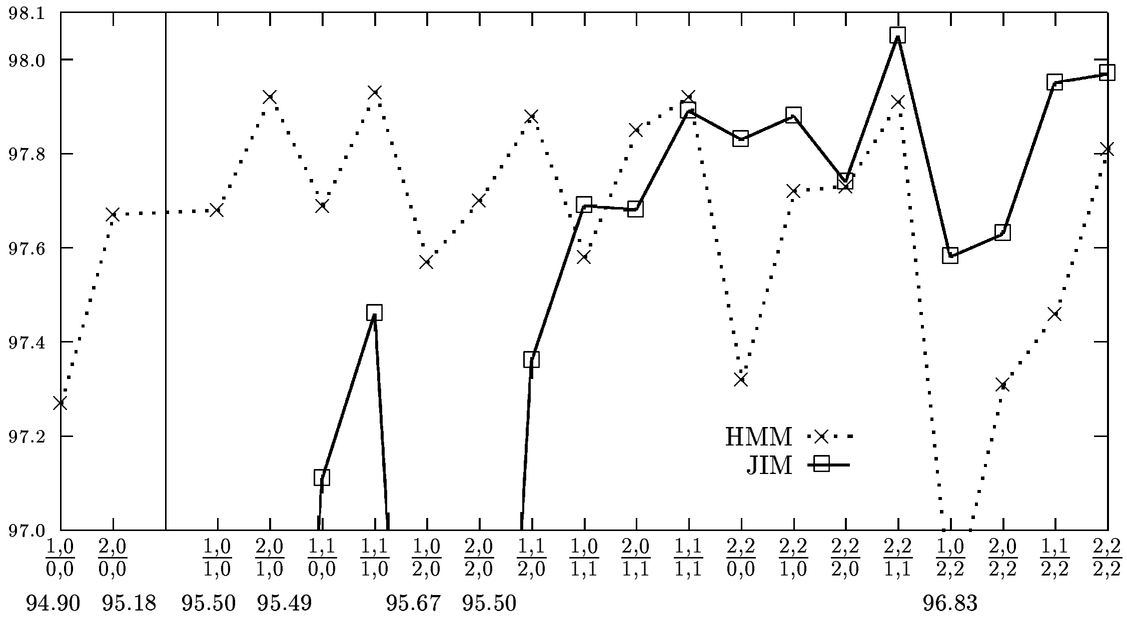

Figure 2: Results under the closed vocabulary assumption

information in contexts. In the models, a simplified version of back-off smoothing is used to mitigate data sparseness problem. The models assume joint independence between random variables in order to make the parameter estimation more reliable.

From the experiments, we have observed that extended models achieved even better results than the standard models in case of both HMMs and JIMs, that the simplified back-off smoothing technique mitigated data sparseness quite effectively, and that some extended JIMs outperformed the corresponding HMMs. Under the closed vocabulary assumption, the best JIM outperformed the best HMM. On the contrary, under the open vocabulary assumption, the best HMM outperformed the best JIM. Intuitively speaking, it is empirically proven that the joint independence assumption is more effective than the Markov assumption in some models that consult specific features such as lexicalized ones.

Generally, the uniform extension of models requires rapid increase of parameters, and hence suffers from large storage and sparse data. Recently in many areas where HMMs are used, many efforts to extend models non-uniformly have been made, sometimes resulting in noticeable improvement. For this reason, we are trying to transform our uniform models into non-uniform models, which may be more effective in terms of both space complexity and reliable estimation of paremeters, without loss of accuracy.

And also, we are trying to apply our models to different areas such as information extraction in the bio-molecular domain, noun phrase chuncking, and so on.

References

E. Brill. 1994. Some Advances in Transformation-Based Part of Speech Tagging. In *Proc. of the* 12^{th} *Nat'l Conf. on Artificial Intelligence(AAAI-94)*, 722–727.

E. Charniak, C. Hendrickson, N. Jacobson, and M. Perkowitz. 1993. Equations for Part-of-Speech Tagging. In *Proc. of the* 11^{th} *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.

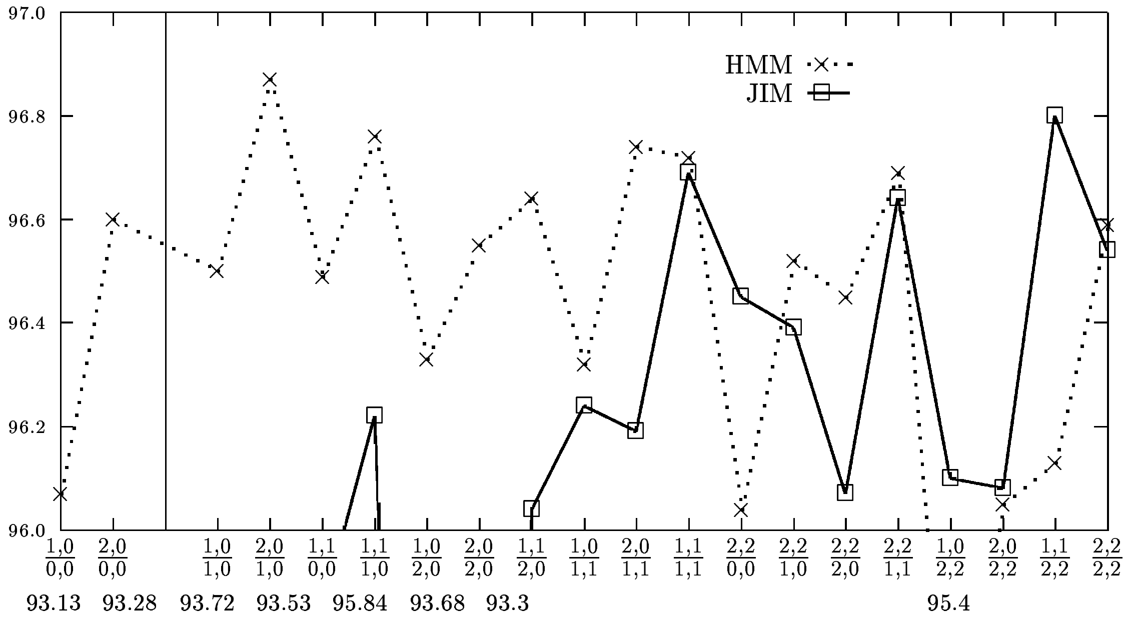

Figure 3: Results under the open vocabulary assumption

W. N. Francis and H. Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin Company, Boston, Massachusetts.

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.

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.

Y.-C. Lin, T.-H. Chiang, and K.-Y. Su. 1992. Discrimination Oriented Probabilistic Tagging. In *Proc. of the 5^{th} International Conference: Research on Computational Linguistics(ROCLING-V)*, 85–96.

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.

T. M. Mitchell. 1997. *Machine Learning.* New York: McGraw-Hill.

L. Padró. 1996. *POS Tagging Using Relaxation Labeling.* Research Report LSI-96-10-R. Department de Llenguatgatges i Sistemes Informàtics, Universitat Politècnica de Catalunya.

A. Ratnaparkhi. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proc. of the Empirical Methods in Natural Language Processing Conference(EMNLP-96)*, 133–142.

C. Samuelsson. 1993. Morphological Tagging Based Entirely on Bayesian Inference. In *Proc. of the 9^{th} Nordic Conference on Computational Linguistics*, 225–238.

H. Schmid. 1994. Part-of-Speech Tagging with Neural Networks. In *Proc. of the International Conference on Computational Linguistics(COLING-94)*, 172–176.