Guessing Parts-of-Speech of Unknown Words Using Global Information

Tetsuji Nakagawa
Corporate R&D Center
Oki Electric Industry Co., Ltd.
2−5−7 Honmachi, Chuo-ku
Osaka 541−0053, Japan
nakagawa378@oki.com

Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
8916−5 Takayama, Ikoma
Nara 630−0101, Japan
matsu@is.naist.jp

Abstract

In this paper, we present a method for guessing POS tags of unknown words using local and global information. Although many existing methods use only local information (i.e. limited window size or intra-sentential features), global information (extra-sentential features) provides valuable clues for predicting POS tags of unknown words. We propose a probabilistic model for POS guessing of unknown words using global information as well as local information, and estimate its parameters using Gibbs sampling. We also attempt to apply the model to semi-supervised learning, and conduct experiments on multiple corpora.

1 Introduction

Part-of-speech (POS) tagging is a fundamental language analysis task. In POS tagging, we frequently encounter words that do not exist in training data. Such words are called unknown words. They are usually handled by an exceptional process in POS tagging, because the tagging system does not have information about the words. Guessing the POS tags of such unknown words is a difficult task. But it is an important issue both for conducting POS tagging accurately and for creating word dictionaries automatically or semi-automatically. There have been many studies on POS guessing of unknown words (Mori and Nagao, 1996; Mikheev, 1997; Chen et al., 1997; Nagata, 1999; Orphanos and Christodoulakis, 1999). In most of these previous works, POS tags of unknown words were predicted using only local information, such as lexical forms and POS tags of surrounding words or word-internal features (e.g. suffixes and character types) of the unknown words. However, this approach has limitations in available information. For example, common nouns and proper nouns are sometimes difficult to distinguish with only the information of a single occurrence because their syntactic functions are almost identical. In English, proper nouns are capitalized and there is generally little ambiguity between common nouns and proper nouns. In Chinese and Japanese, no such convention exists and the problem of the ambiguity is serious. However, if an unknown word with the same lexical form appears in another part with informative local features (e.g. titles of persons), this will give useful clues for guessing the part-of-speech of the ambiguous one, because unknown words with the same lexical form usually have the same part-of-speech. For another example, there is a part-of-speech named *sahen-noun* (verbal noun) in Japanese. Verbal nouns behave as common nouns, except that they are used as verbs when they are followed by a verb "*suru*"; e.g., a verbal noun "*dokusho*" means "reading" and "*dokusho-suru*" is a verb meaning to "read books". It is difficult to distinguish a verbal noun from a common noun if it is used as a noun. However, it will be easy if we know that the word is followed by "*suru*" in another part in the document. This issue was mentioned by Asahara (2003) as a problem of *possibility-based POS tags*. A possibility-based POS tag is a POS tag that represents all the possible properties of the word (e.g., a verbal noun is used as a noun or a verb), rather than a property of each instance of the word. For example, a sahen-noun is actually a noun that can be used as a verb when it is followed by "*suru*". This property cannot be confirmed without observing real usage of the word appearing with "*suru*". Such POS tags may not be identified with only local information of one instance, because the property that each instance has is only one among all the possible properties.

To cope with these issues, we propose a method that uses global information as well as local information for guessing the parts-of-speech of unknown words. With this method, all the occurrences of the unknown words in a document[1] are taken into consideration at once, rather than that each occurrence of the words is processed separately. Thus, the method models the whole document and finds a set of parts-of-speech by maximizing its conditional joint probability given the document, rather than independently maximizing the probability of each part-of-speech given each sentence. Global information is known to be useful in other NLP tasks, especially in the named entity recognition task, and several studies successfully used global features (Chieu and Ng, 2002; Finkel et al., 2005).

One potential advantage of our method is its

[1] In this paper, we use the word *document* to denote the whole data consisting of multiple sentences (training corpus or test corpus).

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 705–712,
Sydney, July 2006.

ability to incorporate unlabeled data. Global features can be increased by simply adding unlabeled data into the test data.

Models in which the whole document is taken into consideration need a lot of computation compared to models with only local features. They also cannot process input data one-by-one. Instead, the entire document has to be read before processing. We adopt Gibbs sampling in order to compute the models efficiently, and these models are suitable for offline use such as creating dictionaries from raw text where real-time processing is not necessary but high-accuracy is needed to reduce human labor required for revising automatically analyzed data.

The rest of this paper is organized as follows: Section 2 describes a method for POS guessing of unknown words which utilizes global information. Section 3 shows experimental results on multiple corpora. Section 4 discusses related work, and Section 5 gives conclusions.

2 POS Guessing of Unknown Words with Global Information

We handle POS guessing of unknown words as a sub-task of POS tagging, in this paper. We assume that POS tags of known words are already determined beforehand, and positions in the document where unknown words appear are also identified. Thus, we focus only on prediction of the POS tags of unknown words.

In the rest of this section, we first present a model for POS guessing of unknown words with global information. Next, we show how the test data is analyzed and how the parameters of the model are estimated. A method for incorporating unlabeled data with the model is also discussed.

2.1 Probabilistic Model Using Global Information

We attempt to model the probability distribution of the parts-of-speech of all occurrences of the unknown words in a document which have the same lexical form. We suppose that such parts-of-speech have correlation, and the part-of-speech of each occurrence is also affected by its local context. Similar situations to this are handled in physics. For example, let us consider a case where a number of electrons with spins exist in a system. The spins interact with each other, and each spin is also affected by the external magnetic field. In the physical model, if the state of the system is $\mathbf{s}$ and the energy of the system is $E(\mathbf{s})$, the probability distribution of $\mathbf{s}$ is known to be represented by the following Boltzmann distribution:

$$P(\mathbf{s})=\frac{1}{Z}\exp\{-\beta E(\mathbf{s})\}, \quad (1)$$

where β is inverse temperature and Z is a normalizing constant defined as follows:

$$Z=\sum_{\mathbf{s}}\exp\{-\beta E(\mathbf{s})\}. \quad (2)$$

Takamura et al. (2005) applied this model to an NLP task, semantic orientation extraction, and we apply it to POS guessing of unknown words here.

Suppose that unknown words with the same lexical form appear K times in a document. Assume that the number of possible POS tags for unknown words is N, and they are represented by integers from 1 to N. Let t_k denote the POS tag of the kth occurrence of the unknown words, let w_k denote the local context (e.g. the lexical forms and the POS tags of the surrounding words) of the kth occurrence of the unknown words, and let $\mathbf{w}$ and $\mathbf{t}$ denote the sets of w_k and t_k respectively:

$$\mathbf{w}=\{w_1,\cdots,w_K\},\ \mathbf{t}=\{t_1,\cdots,t_K\},\ t_k\in\{1,\cdots,N\}.$$

$\lambda_{i,j}$ is a weight which denotes strength of the interaction between parts-of-speech i and j, and is symmetric ($\lambda_{i,j}=\lambda_{j,i}$). We define the energy where POS tags of unknown words given $\mathbf{w}$ are $\mathbf{t}$ as follows:

$$E(\mathbf{t}|\mathbf{w})=-\left\{\frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\lambda_{t_k,t_{k'}}+\sum_{k=1}^{K}\log p_0(t_k|w_k)\right\}, \quad (3)$$

where $p_0(t|w)$ is an initial distribution (local model) of the part-of-speech t which is calculated with only the local context w, using arbitrary statistical models such as maximum entropy models. The right hand side of the above equation consists of two components; one represents global interactions between each pair of parts-of-speech, and the other represents the effects of local information.

In this study, we fix the inverse temperature $\beta=1$. The distribution of $\mathbf{t}$ is then obtained from Equation (1), (2) and (3) as follows:

$$P(\mathbf{t}|\mathbf{w})=\frac{1}{Z(\mathbf{w})}p_0(\mathbf{t}|\mathbf{w})\exp\left\{\frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\lambda_{t_k,t_{k'}}\right\}, \quad (4)$$

$$Z(\mathbf{w})=\sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w})}p_0(\mathbf{t}|\mathbf{w})\exp\left\{\frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\lambda_{t_k,t_{k'}}\right\}, \quad (5)$$

$$p_0(\mathbf{t}|\mathbf{w})\equiv\prod_{k=1}^{K}p_0(t_k|w_k), \quad (6)$$

where $\mathcal{T}(\mathbf{w})$ is the set of possible configurations of POS tags given $\mathbf{w}$. The size of $\mathcal{T}(\mathbf{w})$ is N^K, because there are K occurrences of the unknown words and each unknown word can have one of N POS tags. The above equations can be rewritten as follows by defining a function $f_{i,j}(\mathbf{t})$:

$$f_{i,j}(\mathbf{t})\equiv\frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\delta(t_k,i)\delta(t_{k'},j), \quad (7)$$

$$P(\mathbf{t}|\mathbf{w})=\frac{1}{Z(\mathbf{w})}p_0(\mathbf{t}|\mathbf{w})\exp\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{i,j}f_{i,j}(\mathbf{t})\right\}, \quad (8)$$

$$Z(\mathbf{w})=\sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w})}p_0(\mathbf{t}|\mathbf{w})\exp\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{i,j}f_{i,j}(\mathbf{t})\right\}, \quad (9)$$

where $\delta(i,j)$ is the Kronecker delta:

$$\delta(i,j)=\begin{cases}1 & (i=j),\\ 0 & (i\neq j).\end{cases} \quad (10)$$

$f_{i,j}(\mathbf{t})$ represents the number of occurrences of the POS tag pair i and j in the whole document (divided by 2), and the model in Equation (8) is essentially a maximum entropy model with the document level features.

As shown above, we consider the conditional joint probability of all the occurrences of the unknown words with the same lexical form in the document given their local contexts, $P(\mathbf{t}|\mathbf{w})$, in contrast to conventional approaches which assume independence of the sentences in the document and use the probabilities of all the words only in a sentence. Note that we assume independence between the unknown words with different lexical forms, and each set of the unknown words with the same lexical form is processed separately from the sets of other unknown words.

2.2 Decoding

Let us consider how to find the optimal POS tags $\mathbf{t}$ basing on the model, given K local contexts of the unknown words with the same lexical form (test data) $\mathbf{w}$, an initial distribution $p_0(t|w)$ and a set of model parameters $\Lambda=\{\lambda_{1,1},\cdots,\lambda_{N,N}\}$. One way to do this is to find a set of POS tags which maximizes $P(\mathbf{t}|\mathbf{w})$ among all possible candidates of $\mathbf{t}$. However, the number of all possible candidates of the POS tags is N^K and the calculation is generally intractable. Although HMMs, MEMMs, and CRFs use dynamic programming and some studies with probabilistic models which have specific structures use efficient algorithms (Wang et al., 2005), such methods cannot be applied here because we are considering interactions (dependencies) between all POS tags, and their joint distribution cannot be decomposed. Therefore, we use a sampling technique and approximate the solution using samples obtained from the probability distribution.

We can obtain a solution $\hat{\mathbf{t}}=\{\hat{t}_1,\cdots,\hat{t}_K\}$ as follows:

$$\hat{t}_k=\operatorname*{argmax}_t P_k(t|\mathbf{w}), \quad (11)$$

where $P_k(t|\mathbf{w})$ is the marginal distribution of the part-of-speech of the kth occurrence of the unknown words given a set of local contexts $\mathbf{w}$, and is calculated as an expected value over the distribution of the unknown words as follows:

$$\begin{aligned}P_k(t|\mathbf{w})&=\sum_{\substack{t_1,\cdots,t_{k-1},t_{k+1},\cdots,t_K\\ t_k=t}} P(\mathbf{t}|\mathbf{w}),\\ &=\sum_{\mathbf{t}\in\mathcal{T}(\mathbf{w})}\delta(t_k,t)P(\mathbf{t}|\mathbf{w}). \quad (12)\end{aligned}$$

Expected values can be approximately calculated using enough number of samples generated from the distribution (MacKay, 2003). Suppose that $A(\mathbf{x})$ is a function of a random variable $\mathbf{x}$, $P(\mathbf{x})$ is a distribution of $\mathbf{x}$, and $\{\mathbf{x}^{(1)},\cdots,\mathbf{x}^{(M)}\}$ are M samples generated from $P(\mathbf{x})$. Then, the expectation of $A(\mathbf{x})$ over $P(\mathbf{x})$ is approximated by the samples:

$$\sum_{\mathbf{x}}A(\mathbf{x})P(\mathbf{x})\simeq\frac{1}{M}\sum_{m=1}^{M}A(\mathbf{x}^{(m)}). \quad (13)$$

Thus, if we have M samples $\{\mathbf{t}^{(1)},\cdots,\mathbf{t}^{(M)}\}$ generated from the conditional joint distribution $P(\mathbf{t}|\mathbf{w})$, the marginal distribution of each POS tag is approximated as follows:

$$P_k(t|\mathbf{w})\simeq\frac{1}{M}\sum_{m=1}^{M}\delta(t_k^{(m)},t). \quad (14)$$

Next, we describe how to generate samples from the distribution. We use Gibbs sampling for this purpose. Gibbs sampling is one of the Markov chain Monte Carlo (MCMC) methods, which can generate samples efficiently from high-dimensional probability distributions (Andrieu et al., 2003). The algorithm is shown in Figure 1. The algorithm firstly set the initial state $\mathbf{t}^{(1)}$, then one new random variable is sampled at a time from the conditional distribution in which all other variables are fixed, and new samples are created by repeating the process. Gibbs sampling is easy to implement and is guaranteed to converge to the true distribution. The conditional distribution $P(t_k|\mathbf{w},t_1,\cdots,t_{k-1},t_{k+1},\cdots,t_K)$ in Figure 1 can be calculated simply as follows:

initialize $\mathbf{t}^{(1)}$
for m := 2 **to** M
  for k := 1 **to** K
    $t_k^{(m)} \sim P(t_k|\mathbf{w},t_1^{(m)},\cdots,t_{k-1}^{(m)},t_{k+1}^{(m-1)},\cdots,t_K^{(m-1)})$

Figure 1: Gibbs Sampling

$$\begin{aligned}&\frac{P(t_k|\mathbf{w},t_1,\cdots,t_{k-1},t_{k+1},\cdots,t_K)}{P(\mathbf{t}|\mathbf{w})}\\ =&\frac{P(\mathbf{t}|\mathbf{w})}{P(t_1,\cdots,t_{k-1},t_{k+1},\cdots,t_K|\mathbf{w})},\\ =&\frac{\frac{1}{Z(\mathbf{w})}p_0(\mathbf{t}|\mathbf{w})\exp\{\frac{1}{2}\sum_{k'=1}^{K}\sum_{\substack{k''=1\\k''\neq k'}}^{K}\lambda_{t_{k'},t_{k''}}\}}{\sum_{t_k^*=1}^{N}P(t_1,\cdots,t_{k-1},t_k^*,t_{k+1},\cdots,t_K|\mathbf{w})},\\ =&\frac{p_0(t_k|w_k)\exp\{\sum_{\substack{k'=1\\k'\neq k}}^{K}\lambda_{t_{k'},t_k}\}}{\sum_{t_k^*=1}^{N}p_0(t_k^*|w_k)\exp\{\sum_{\substack{k'=1\\k'\neq k}}^{K}\lambda_{t_{k'},t_k^*}\}}, \quad (15)\end{aligned}$$

where the last equation is obtained using the following relation:

$$\frac{1}{2}\sum_{k'=1}^{K}\sum_{\substack{k''=1\\k''\neq k'}}^{K}\lambda_{t_{k'},t_{k''}}=\frac{1}{2}\sum_{\substack{k'=1\\k'\neq k}}^{K}\sum_{\substack{k''=1\\k''\neq k,k''\neq k'}}^{K}\lambda_{t_{k'},t_{k''}}+\sum_{\substack{k'=1\\k'\neq k}}^{K}\lambda_{t_{k'},t_k}.$$

In later experiments, the number of samples M is set to 100, and the initial state $\mathbf{t}^{(1)}$ is set to the POS tags which maximize $p_0(\mathbf{t}|\mathbf{w})$.

The optimal solution obtained by Equation (11) maximizes the probability of each POS tag given $\mathbf{w}$, and this kind of approach is known as the maximum posterior marginal (MPM) estimate (Marroquin, 1985). Finkel et al. (2005) used simulated annealing with Gibbs sampling to find a solution in a similar situation. Unlike simulated annealing, this approach does not need to define a cooling

schedule. Furthermore, this approach can obtain not only the best solution but also the second best or the other solutions according to $P_k(t|\mathbf{w})$, which are useful when this method is applied to semi-automatic construction of dictionaries because human annotators can check the ranked lists of candidates.

2.3 Parameter Estimation

Let us consider how to estimate the parameter $\Lambda = \{\lambda_{1,1}, \cdots, \lambda_{N,N}\}$ in Equation (8) from training data consisting of L examples; $\{\langle \mathbf{w}^1, \mathbf{t}^1 \rangle, \cdots, \langle \mathbf{w}^L, \mathbf{t}^L \rangle\}$ (i.e., the training data contains L different lexical forms of unknown words). We define the following objective function $\mathcal{L}_\Lambda$, and find Λ which maximizes $\mathcal{L}_\Lambda$ (the subscript Λ denotes being parameterized by Λ):

$$\begin{aligned}
\mathcal{L}_\Lambda &= \log \prod_{l=1}^{L} P_\Lambda(\mathbf{t}^l|\mathbf{w}^l) + \log P(\Lambda), \\
&= \log \prod_{l=1}^{L} \frac{1}{Z_\Lambda(\mathbf{w}^l)} p_0(\mathbf{t}^l|\mathbf{w}^l) \exp\left\{ \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{i,j} f_{i,j}(\mathbf{t}^l) \right\} \\
&\quad + \log P(\Lambda), \\
&= \sum_{l=1}^{L} \left[-\log Z_\Lambda(\mathbf{w}^l) + \log p_0(\mathbf{t}^l|\mathbf{w}^l) + \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{i,j} f_{i,j}(\mathbf{t}^l) \right] \\
&\quad + \log P(\Lambda). \qquad (16)
\end{aligned}$$

The partial derivatives of the objective function are:

$$\begin{aligned}
\frac{\partial \mathcal{L}_\Lambda}{\partial \lambda_{i,j}} &= \sum_{l=1}^{L} \left[f_{i,j}(\mathbf{t}^l) - \frac{\partial}{\partial \lambda_{i,j}} \log Z_\Lambda(\mathbf{w}^l) \right] + \frac{\partial}{\partial \lambda_{i,j}} \log P(\Lambda), \\
&= \sum_{l=1}^{L} \left[f_{i,j}(\mathbf{t}^l) - \sum_{\mathbf{t} \in \mathcal{T}(\mathbf{w}^l)} f_{i,j}(\mathbf{t}) P_\Lambda(\mathbf{t}|\mathbf{w}^l) \right] + \frac{\partial}{\partial \lambda_{i,j}} \log P(\Lambda). \qquad (17)
\end{aligned}$$

We use Gaussian priors (Chen and Rosenfeld, 1999) for $P(\Lambda)$:

$$\log P(\Lambda) = -\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\lambda_{i,j}^2}{2\sigma^2} + C, \quad \frac{\partial}{\partial \lambda_{i,j}} \log P(\Lambda) = -\frac{\lambda_{i,j}}{\sigma^2}.$$

where C is a constant and σ is set to 1 in later experiments. The optimal Λ can be obtained by quasi-Newton methods using the above $\mathcal{L}_\Lambda$ and $\frac{\partial \mathcal{L}_\Lambda}{\partial \lambda_{i,j}}$, and we use L-BFGS (Liu and Nocedal, 1989) for this purpose[2]. However, the calculation is intractable because $Z_\Lambda(\mathbf{w}^l)$ (see Equation (9)) in Equation (16) and a term in Equation (17) contain summations over all the possible POS tags. To cope with the problem, we use the sampling technique again for the calculation, as suggested by Rosenfeld et al. (2001). $Z_\Lambda(\mathbf{w}^l)$ can be approximated using M samples $\{\mathbf{t}^{(1)}, \cdots, \mathbf{t}^{(M)}\}$ generated from $p_0(\mathbf{t}|\mathbf{w}^l)$:

$$\begin{aligned}
Z_\Lambda(\mathbf{w}^l) &= \sum_{\mathbf{t} \in \mathcal{T}(\mathbf{w}^l)} p_0(\mathbf{t}|\mathbf{w}^l) \exp\left\{ \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{i,j} f_{i,j}(\mathbf{t}) \right\}, \\
&\simeq \frac{1}{M} \sum_{m=1}^{M} \exp\left\{ \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{i,j} f_{i,j}(\mathbf{t}^{(m)}) \right\}. \qquad (18)
\end{aligned}$$

The term in Equation (17) can also be approximated using M samples $\{\mathbf{t}^{(1)}, \cdots, \mathbf{t}^{(M)}\}$ generated from $P_\Lambda(\mathbf{t}|\mathbf{w}^l)$ with Gibbs sampling:

$$\sum_{\mathbf{t} \in \mathcal{T}(\mathbf{w}^l)} f_{i,j}(\mathbf{t}) P_\Lambda(\mathbf{t}|\mathbf{w}^l) \simeq \frac{1}{M} \sum_{m=1}^{M} f_{i,j}(\mathbf{t}^{(m)}). \qquad (19)$$

In later experiments, the initial state $\mathbf{t}^{(1)}$ in Gibbs sampling is set to the gold standard tags in the training data.

2.4 Use of Unlabeled Data

In our model, unlabeled data can be easily used by simply concatenating the test data and the unlabeled data, and decoding them in the testing phase. Intuitively, if we increase the amount of the test data, test examples with informative local features may increase. The POS tags of such examples can be easily predicted, and they are used as global features in prediction of other examples. Thus, this method uses unlabeled data in only the testing phase, and the training phase is the same as the case with no unlabeled data.

3 Experiments

3.1 Data and Procedure

We use eight corpora for our experiments; the Penn Chinese Treebank corpus 2.0 (CTB), a part of the PFR corpus (PFR), the EDR corpus (EDR), the Kyoto University corpus version 2 (KUC), the RWCP corpus (RWC), the GENIA corpus 3.02p (GEN), the SUSANNE corpus (SUS) and the Penn Treebank WSJ corpus (WSJ), (cf. Table 1). All the corpora are POS tagged corpora in Chinese(C), English(E) or Japanese(J), and they are split into three portions; training data, test data and unlabeled data. The unlabeled data is used in experiments of semi-supervised learning, and POS tags of unknown words in the unlabeled data are eliminated. Table 1 summarizes detailed information about the corpora we used: the language, the number of POS tags, the number of open class tags (POS tags that unknown words can have, described later), the sizes of training, test and unlabeled data, and the splitting method of them. For the test data and the unlabeled data, unknown words are defined as words that do not appear in the training data. The number of unknown words in the test data of each corpus is shown in Table 1, parentheses. Accuracy of POS guessing of unknown words is calculated based on how many words among them are correctly POS-guessed.

Figure 2 shows the procedure of the experiments. We split the training data into two parts; the first half as sub-training data 1 and the latter half as sub-training data 2 (Figure 2, *1). Then, we check the words that appear in the sub-training

[2] In later experiments, L-BFGS often did not converge completely because we used approximation with Gibbs sampling, and we stopped iteration of L-BFGS in such cases.

Corpus (Lang.)	# of POS (Open Class)	# of Tokens (# of Unknown Words) [partition in the corpus]		
		Training	Test	Unlabeled
CTB (C)	34 (28)	84,937 [sec. 1–270]	7,980 (749) [sec. 271–300]	6,801 [sec. 301–325]
PFR (C)	42 (39)	304,125 [Jan. 1–Jan. 9]	370,627 (27,774) [Jan. 10–Jan. 19]	445,969 [Jan. 20–Jan. 31]
EDR (J)	15 (15)	2,550,532 [$id = 4n + 0, id = 4n + 1$]	1,280,057 (24,178) [$id = 4n + 2$]	1,274,458 [$id = 4n + 3$]
KUC (J)	40 (36)	198,514 [Jan. 1–Jan. 8]	31,302 (2,477) [Jan. 9]	41,227 [Jan. 10]
RWC (J)	66 (55)	487,333 [1–10,000th sentences]	190,571 (11,177) [10,001–14,000th sentences]	210,096 [14,001–18,672th sentences]
GEN (E)	47 (36)	243,180 [1–10,000th sentences]	123,386 (7,775) [10,001–15,000th sentences]	134,380 [15,001–20,546th sentences]
SUS (E)	125 (90)	74,902 [sec. A01–08, G01–08, J01–08, N01–08]	37,931 (5,760) [sec. A09–12, G09–12, J09–17, N09–12]	37,593 [sec. A13–20, G13–22, J21–24, N13–18]
WSJ (E)	45 (33)	912,344 [sec. 0–18]	129,654 (4,253) [sec. 22–24]	131,768 [sec. 19–21]

Table 1: Statistical Information of Corpora

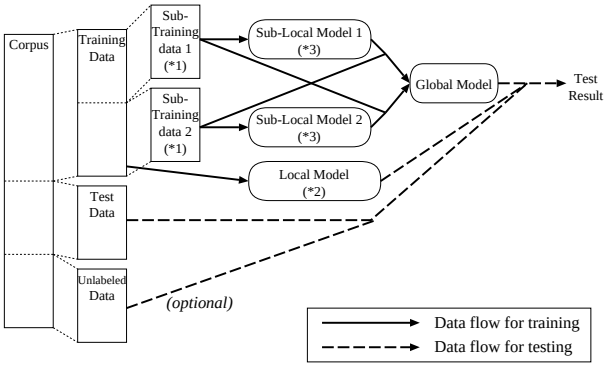

Figure 2: Experimental Procedure

data 1 but not in the sub-training data 2, or vice versa. We handle these words as (*pseudo*) unknown words in the training data. Such (two-fold) cross-validation is necessary to make training examples that contain unknown words[3]. POS tags that these pseudo unknown words have are defined as open class tags, and only the open class tags are considered as candidate POS tags for unknown words in the test data (i.e., N is equal to the number of the open class tags). In the training phase, we need to estimate two types of parameters; local model (parameters), which is necessary to calculate $p_0(t|w)$, and global model (parameters), i.e., $\lambda_{i,j}$. The local model parameters are estimated using all the training data (Figure 2, *2). Local model parameters and training data are necessary to estimate the global model parameters, but the global model parameters cannot be estimated from the same training data from which the local model parameters are estimated. In order to estimate the global model parameters, we firstly train sub-local models 1 and 2 from the sub-training data 1 and 2 respectively (Figure 2, *3). The sub-local models 1 and 2 are used for calculating $p_0(t|w)$ of unknown words in the sub-training data 2 and 1 respectively, when the global model parameters are estimated from the entire training data. In the testing phase, $p_0(t|w)$ of unknown words in the test data are calculated using the local model parameters which are estimated from the entire training data, and test results are obtained using the global model with the local model.

Global information cannot be used for unknown words whose lexical forms appear only once in the training or test data, so we process only non-unique unknown words (unknown words whose lexical forms appear more than once) using the proposed model. In the testing phase, POS tags of unique unknown words are determined using only the local information, by choosing POS tags which maximize $p_0(t|w)$.

Unlabeled data can be optionally used for semi-supervised learning. In that case, the test data and the unlabeled data are concatenated, and the best POS tags which maximize the probability of the mixed data are searched.

3.2 Initial Distribution

In our method, the initial distribution $p_0(t|w)$ is used for calculating the probability of t given local context w (Equation (8)). We use maximum entropy (ME) models for the initial distribution. $p_0(t|w)$ is calculated by ME models as follows (Berger et al., 1996):

$$p_0(t|w) = \frac{1}{Y(w)} \exp\left\{\sum_{h=1}^{H} \alpha_h g_h(w,t)\right\}, \qquad (20)$$

[3] A major method for generating such pseudo unknown words is to collect the words that appear only once in a corpus (Nagata, 1999). These words are called *hapax legomena* and known to have similar characteristics to real unknown words (Baayen and Sproat, 1996). These words are interpreted as being collected by the leave-one-out technique (which is a special case of cross-validation) as follows: One word is picked from the corpus and the rest of the corpus is considered as training data. The picked word is regarded as an unknown word if it does not exist in the training data. This procedure is iterated for all the words in the corpus. However, this approach is not applicable to our experiments because those words that appear only once in the corpus do not have global information and are useless for learning the global model, so we use the two-fold cross validation method.

Language	Features
English	Prefixes of ω_0 up to four characters, suffixes of ω_0 up to four characters, ω_0 contains Arabic numerals, ω_0 contains uppercase characters, ω_0 contains hyphens.
Chinese Japanese	Prefixes of ω_0 up to two characters, suffixes of ω_0 up to two characters, ψ_1, $\psi_{\|\omega_0\|}$, ψ_1 & $\psi_{\|\omega_0\|}$, $\bigcup_{i=1}^{\|\omega_0\|}\{\psi_i\}$ (set of character types).
(common)	$\|\omega_0\|$ (length of ω_0), τ_{-1}, τ_{+1}, τ_{-2} & τ_{-1}, τ_{+1} & τ_{+2}, τ_{-1} & τ_{+1}, ω_{-1} & τ_{-1}, ω_{+1} & τ_{+1}, ω_{-2} & τ_{-2} & ω_{-1} & τ_{-1}, ω_{+1} & τ_{+1} & ω_{+2} & τ_{+2}, ω_{-1} & τ_{-1} & ω_{+1} & τ_{+1}.

Table 2: Features Used for Initial Distribution

$$Y(w)=\sum_{t=1}^{N}\exp\left\{\sum_{h=1}^{H}\alpha_h g_h(w,t)\right\}, \quad (21)$$

where $g_h(w,t)$ is a binary feature function. We assume that each local context w contains the following information about the unknown word:

- The POS tags of the two words on each side of the unknown word: $\tau_{-2}, \tau_{-1}, \tau_{+1}, \tau_{+2}$.[4]
- The lexical forms of the unknown word itself and the two words on each side of the unknown word: $\omega_{-2}, \omega_{-1}, \omega_0, \omega_{+1}, \omega_{+2}$.
- The character types of all the characters composing the unknown word: $\psi_1, \cdots, \psi_{|\omega_0|}$. We use six character types: alphabet, numeral (Arabic and Chinese numerals), symbol, Kanji (Chinese character), Hiragana (Japanese script) and Katakana (Japanese script).

A feature function $g_h(w,t)$ returns 1 if w and t satisfy certain conditions, and otherwise 0; for example:

$$g_{123}(w,t)=\begin{cases}1 & (\omega_{-1}=\text{“President” and } \tau_{-1}=\text{“NNP” and } t=5),\\ 0 & (\text{otherwise}).\end{cases}$$

The features we use are shown in Table 2, which are based on the features used by Ratnaparkhi (1996) and Uchimoto et al. (2001).

The parameters α_h in Equation (20) are estimated using all the words in the training data whose POS tags are the open class tags.

3.3 Experimental Results

The results are shown in Table 3. In the table, *local*, *local+global* and *local+global w/ unlabeled* indicate that the results were obtained using only local information, local and global information, and local and global information with the extra unlabeled data, respectively. The results using only local information were obtained by choosing POS tags $\hat{\mathbf{t}}=\{\hat{t}_1,\cdots,\hat{t}_K\}$ which maximize the probabilities of the local model:

$$\hat{t}_k=\operatorname*{argmax}_t p_0(t|w_k). \quad (22)$$

[4] In both the training and the testing phases, POS tags of known words are given from the corpora. When these surrounding words contain unknown words, their POS tags are represented by a special tag *Unk*.

PFR (Chinese)	
+162	vn (verbal noun)
+150	ns (place name)
+86	nz (other proper noun)
+85	j (abbreviation)
+61	nr (personal name)
···	···
−26	m (numeral)
−100	v (verb)
RWC (Japanese)	
+33	noun-proper noun-person name-family name
+32	noun-proper noun-place name
+28	noun-proper noun-organization name
+17	noun-proper noun-person name-first name
+6	noun-proper noun
+4	noun-sahen noun
···	···
−2	noun-proper noun-place name-country name
−29	noun
SUS (English)	
+13	NP (proper noun)
+6	JJ (adjective)
+2	VVD (past tense form of lexical verb)
+2	NNL (locative noun)
+2	NNJ (organization noun)
···	···
−3	NN (common noun)
−6	NNU (unit-of-measurement noun)

Table 4: Ordered List of Increased/Decreased Number of Correctly Tagged Words

The table shows the accuracies, the numbers of errors, the p-values of McNemar's test against the results using only local information, and the numbers of non-unique unknown words in the test data. On an Opteron 250 processor with 8GB of RAM, model parameter estimation and decoding without unlabeled data for the eight corpora took 117 minutes and 39 seconds in total, respectively.

In the CTB, PFR, KUC, RWC and WSJ corpora, the accuracies were improved using global information (statistically significant at $p < 0.05$), compared to the accuracies obtained using only local information. The increases of the accuracies on the English corpora (the GEN and SUS corpora) were small. Table 4 shows the increased/decreased number of correctly tagged words using global information in the PFR, RWC and SUS corpora. In the PFR (Chinese) and RWC (Japanese) corpora, many proper nouns were correctly tagged using global information. In Chinese and Japanese, proper nouns are not capitalized, therefore proper nouns are difficult to distinguish from common nouns with only local information. One reason that only the small increases were obtained with global information in the English corpora seems to be the low ambiguities of proper nouns. Many verbal nouns in PFR and a few sahen-nouns (Japanese verbal nouns) in RWC, which suffer from the problem of possibility-based POS tags, were also correctly tagged using global information. When the unlabeled data was used, the number of non-unique words in the test data increased. Compared with the case without the unlabeled data, the accu-

Corpus (Lang.)	Accuracy for Unknown Words (# of Errors) [p-value] ⟨# of Non-unique Unknown Words⟩		
	local	local+global	local+global w/ unlabeled
CTB (C)	0.7423 (193)	0.7717 (171) [0.0000] ⟨344⟩	0.7704 (172) [0.0001] ⟨361⟩
PFR (C)	0.6499 (9723)	0.6690 (9193) [0.0000] ⟨16019⟩	0.6785 (8930) [0.0000] ⟨18861⟩
EDR (J)	0.9639 (874)	0.9643 (863) [0.1775] ⟨4903⟩	0.9651 (844) [0.0034] ⟨7770⟩
KUC (J)	0.7501 (619)	0.7634 (586) [0.0000] ⟨788⟩	0.7562 (604) [0.0872] ⟨936⟩
RWC (J)	0.7699 (2572)	0.7785 (2476) [0.0000] ⟨5044⟩	0.7787 (2474) [0.0000] ⟨5878⟩
GEN (E)	0.8836 (905)	0.8837 (904) [1.0000] ⟨4094⟩	0.8863 (884) [0.0244] ⟨4515⟩
SUS (E)	0.7934 (1190)	0.7957 (1177) [0.1878] ⟨3210⟩	0.7979 (1164) [0.0116] ⟨3583⟩
WSJ (E)	0.8345 (704)	0.8368 (694) [0.0162] ⟨1412⟩	0.8352 (701) [0.7103] ⟨1627⟩

Table 3: Results of POS Guessing of Unknown Words

Corpus (Lang.)	Mean±Standard Deviation	
	Marginal	S.A.
CTB (C)	0.7696±0.0021	0.7682±0.0028
PFR (C)	0.6707±0.0010	0.6712±0.0014
EDR (J)	0.9644±0.0001	0.9645±0.0001
KUC (J)	0.7595±0.0031	0.7612±0.0018
RWC (J)	0.7777±0.0017	0.7772±0.0020
GEN (E)	0.8841±0.0009	0.8840±0.0007
SUS (E)	0.7997±0.0038	0.7995±0.0034
WSJ (E)	0.8366±0.0013	0.8360±0.0021

Table 5: Results of Multiple Trials and Comparison to Simulated Annealing

racies increased in several corpora but decreased in the CTB, KUC and WSJ corpora.

Since our method uses Gibbs sampling in the training and the testing phases, the results are affected by the sequences of random numbers used in the sampling. In order to investigate the influence, we conduct 10 trials with different sequences of pseudo random numbers. We also conduct experiments using simulated annealing in decoding, as conducted by Finkel et al. (2005) for information extraction. We increase inverse temperature β in Equation (1) from $\beta = 1$ to $\beta \approx \infty$ with the linear cooling schedule. The results are shown in Table 5. The table shows the mean values and the standard deviations of the accuracies for the 10 trials, and *Marginal* and *S.A.* mean that decoding is conducted using Equation (11) and simulated annealing respectively. The variances caused by random numbers and the differences of the accuracies between *Marginal* and *S.A.* are relatively small.

4 Related Work

Several studies concerning the use of global information have been conducted, especially in named entity recognition, which is a similar task to POS guessing of unknown words. Chieu and Ng (2002) conducted named entity recognition using global features as well as local features. In their ME model-based method, some global features were used such as "when the word appeared first in a position other than the beginning of sentences, the word was capitalized or not". These global features are static and can be handled in the same manner as local features, therefore Viterbi decoding was used. The method is efficient but does not handle interactions between labels.

Finkel et al. (2005) proposed a method incorporating non-local structure for information extraction. They attempted to use *label consistency* of named entities, which is the property that named entities with the same lexical form tend to have the same label. They defined two probabilistic models; a local model based on conditional random fields and a global model based on log-linear models. Then the final model was constructed by multiplying these two models, which can be seen as unnormalized log-linear interpolation (Klakow, 1998) of the two models which are weighted equally. In their method, interactions between labels in the whole document were considered, and they used Gibbs sampling and simulated annealing for decoding. Our model is largely similar to their model. However, in their method, parameters of the global model were estimated using relative frequencies of labels or were selected by hand, while in our method, global model parameters are estimated from training data so as to fit to the data according to the objective function.

One approach for incorporating global information in natural language processing is to utilize consistency of labels, and such an approach have been used in other tasks. Takamura et al. (2005) proposed a method based on the spin models in physics for extracting semantic orientations of words. In the spin models, each electron has one of two states, *up* or *down*, and the models give probability distribution of the states. The states of electrons interact with each other and neighboring electrons tend to have the same spin. In their

method, semantic orientations (*positive* or *negative*) of words are regarded as states of spins, in order to model the property that the semantic orientation of a word tends to have the same orientation as words in its gloss. The mean field approximation was used for inference in their method.

Yarowsky (1995) studied a method for word sense disambiguation using unlabeled data. Although no probabilistic models were considered explicitly in the method, they used the property of label consistency named "one sense per discourse" for unsupervised learning together with local information named "one sense per collocation".

There exist other approaches using global information which do not necessarily aim to use label consistency. Rosenfeld et al. (2001) proposed whole-sentence exponential language models. The method calculates the probability of a sentence s as follows:

$$P(s)=\frac{1}{Z}p_0(s)\exp\left\{\sum_i \lambda_i f_i(s)\right\},$$

where $p_0(s)$ is an initial distribution of s and any language models such as trigram models can be used for this. $f_i(s)$ is a feature function and can handle sentence-wide features. Note that if we regard $f_{i,j}(\mathbf{t})$ in our model (Equation (7)) as a feature function, Equation (8) is essentially the same form as the above model. Their models can incorporate any sentence-wide features including syntactic features obtained by shallow parsers. They attempted to use Gibbs sampling and other sampling methods for inference, and model parameters were estimated from training data using the generalized iterative scaling algorithm with the sampling methods. Although they addressed modeling of whole sentences, the method can be directly applied to modeling of whole documents which allows us to incorporate unlabeled data easily as we have discussed. This approach, modeling whole wide-scope contexts with log-linear models and using sampling methods for inference, gives us an expressive framework and will be applied to other tasks.

5 Conclusion

In this paper, we presented a method for guessing parts-of-speech of unknown words using global information as well as local information. The method models a whole document by considering interactions between POS tags of unknown words with the same lexical form. Parameters of the model are estimated from training data using Gibbs sampling. Experimental results showed that the method improves accuracies of POS guessing of unknown words especially for Chinese and Japanese. We also applied the method to semi-supervised learning, but the results were not consistent and there is some room for improvement.

Acknowledgements

This work was supported by a grant from the National Institute of Information and Communications Technology of Japan.

References

Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. 2003. An introduction to MCMC for Machine Learning. *Machine Learning*, 50:5–43.

Masayuki Asahara. 2003. *Corpus-based Japanese morphological analysis*. Nara Institute of Science and Technology, Doctor's Thesis.

Harald Baayen and Richard Sproat. 1996. Estimating Lexical Priors for Low-Frequency Morphologically Ambiguous Forms. *Computational Linguistics*, 22(2):155–166.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

Stanley Chen and Ronald Rosenfeld. 1999. A Gaussian Prior for Smoothing Maximum Entropy Models. Technical Report CMUCS-99-108, Carnegie Mellon University.

Chao-jan Chen, Ming-hong Bai, and Keh-Jiann Chen. 1997. Category Guessing for Chinese Unknown Words. In *Proceedings of NLPRS '97*, pages 35–40.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. In *Proceedings of COLING 2002*, pages 190–196.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of ACL 2005*, pages 363–370.

D. Klakow. 1998. Log-linear interpolation of language models. In *Proceedings of ICSLP '98*, pages 1695–1699.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(3):503–528.

David J. C. MacKay. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Jose. L. Marroquin. 1985. Optimal Bayesian Estimators for Image Segmentation and Surface Reconstruction. A.I. Memo 839, MIT.

Andrei Mikheev. 1997. Automatic Rule Induction for Unknown-Word Guessing. *Computational Linguistics*, 23(3):405–423.

Shinsuke Mori and Makoto Nagao. 1996. Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis. In *Proceedings of COLING '96*, pages 1119–1122.

Masaki Nagata. 1999. A Part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context. In *Proceedings of ACL '99*, pages 277–284.

Giorgos S. Orphanos and Dimitris N. Christodoulakis. 1999. POS Disambiguation and Unknown Word Guessing with Decision Trees. In *Proceedings of EACL '99*, pages 134–141.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of EMNLP '96*, pages 133–142.

Ronald Rosenfeld, Stanley F. Chen, and Xiaojin Zhu. 2001. Whole-Sentence Exponential Language Models: A Vehicle For Linguistic-Statistical Integration. *Computers Speech and Language*, 15(1):55–73.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting Semantic Orientations of Words using Spin Model. In *Proceedings of ACL 2005*, pages 133–140.

Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. 2001. The Unknown Word Problem: a Morphological Analysis of Japanese Using Maximum Entropy Aided by a Dictionary. In *Proceedings of EMNLP 2001*, pages 91–99.

Shaojun Wang, Shaomin Wang, Russel Greiner, Dale Schuurmans, and Li Cheng. 2005. Exploiting Syntactic, Semantic and Lexical Regularities in Language Modeling via Directed Markov Random Fields. In *Proceedings of ICML 2005*, pages 948–955.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of ACL '95*, pages 189–196.