A Part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context

Masaaki NAGATA
NTT Cyber Space Laboratories
1-1 Hikari-no-oka Yokosuka-Shi Kanagawa, 239-0847 Japan
nagata@nttnly.isl.ntt.co.jp

Abstract

We present a statistical model of Japanese unknown words consisting of a set of length and spelling models classified by the character types that constitute a word. The point is quite simple: different character sets should be treated differently and the changes between character types are very important because Japanese script has both ideograms like Chinese (*kanji*) and phonograms like English (*katakana*). Both word segmentation accuracy and part of speech tagging accuracy are improved by the proposed model. The model can achieve 96.6% tagging accuracy if unknown words are correctly segmented.

1 Introduction

In Japanese, around 95% word segmentation accuracy is reported by using a word-based language model and the Viterbi-like dynamic programming procedures (Nagata, 1994; Yamamoto, 1996; Takeuchi and Matsumoto, 1997; Haruno and Matsumoto, 1997). About the same accuracy is reported in Chinese by statistical methods (Sproat et al., 1996). But there has been relatively little improvement in recent years because most of the remaining errors are due to unknown words.

There are two approaches to solve this problem: to increase the coverage of the dictionary (Fung and Wu, 1994; Chang et al., 1995; Mori and Nagao, 1996) and to design a better model for unknown words (Nagata, 1996; Sproat et al., 1996). We take the latter approach. To improve word segmentation accuracy, (Nagata, 1996) used a single general purpose unknown word model, while (Sproat et al., 1996) used a set of specific word models such as for plurals, personal names, and transliterated foreign words.

The goal of our research is to assign a correct part of speech to unknown word as well as identifying it correctly. In this paper, we present a novel statistical model for Japanese unknown words. It consists of a set of word models for each part of speech and word type. We classified Japanese words into nine orthographic types based on the character types that constitute a word. We find that by making different models for each word type, we can better model the length and spelling of unknown words.

In the following sections, we first describe the language model used for Japanese word segmentation. We then describe a series of unknown word models, from the baseline model to the one we propose. Finally, we prove the effectiveness of the proposed model by experiment.

2 Word Segmentation Model

2.1 Baseline Language Model and Search Algorithm

Let the input Japanese character sequence be $C = c_1 \ldots c_m$, and segment it into word sequence $W = w_1 \ldots w_n$ [1]. The word segmentation task can be defined as finding the word segmentation $\hat{W}$ that maximize the joint probability of word sequence given character sequence $P(W|C)$. Since the maximization is carried out with fixed character sequence C, the word segmenter only has to maximize the joint probability of word sequence $P(W)$.

$$\hat{W} = \arg\max_W P(W|C) = \arg\max_W P(W) \qquad (1)$$

We call $P(W)$ the segmentation model. We can use any type of word-based language model for $P(W)$, such as word ngram and class-based ngram. We used the word bigram model in this paper. So, $P(W)$ is approximated by the product of word bigram probabilities $P(w_i|w_{i-1})$.

$$P(W) \approx P(w_1|\text{<bos>}) \prod_{i=2}^{n} P(w_i|w_{i-1}) P(\text{<eos>}|w_n) \qquad (2)$$

Here, the special symbols <bos> and <eos> indicate the beginning and the end of a sentence, respectively.

Basically, word bigram probabilities of the word segmentation model is estimated by computing the

[1] In this paper, we define a word as a combination of its surface form and part of speech. Two words are considered to be equal only if they have the same surface form and part of speech.

Table 1: Examples of word bigrams including unknown word tags

word bigram		frequency
の/no/particle	<U-noun>	6783
<U-verb>	し/shi/inflection	1052
<U-number>	円/yen/suffix	407
<U-adjectival-verb>	な/na/inflection	405
<U-adjective>	い/i/inflection	182
<U-adverb>	と/to/particle	139

relative frequencies of the corresponding events in the word segmented training corpus, with appropriate smoothing techniques. The maximization search can be efficiently implemented by using the Viterbi-like dynamic programming procedure described in (Nagata, 1994).

2.2 Modification to Handle Unknown Words

To handle unknown words, we made a slight modification in the above word segmentation model. We have introduced unknown word tags <U-t> for each part of speech t. For example, <U-noun> and <U-verb> represents an unknown noun and an unknown verb, respectively.

If w_i is an unknown word whose part of speech is t, the word bigram probability $P(w_i|w_{i-1})$ is approximated as the product of word bigram probability $P(\text{<U-t>}|w_{i-1})$ and the probability of w_i given it is an unknown word whose part of speech is t, $P(w_i|\text{<U-t>})$.

$$\begin{aligned} P(w_i|w_{i-1}) &= P(\text{<U-t>}|w_{i-1})P(w_i|\text{<U-t>}, w_{i-1}) \\ &\approx P(\text{<U-t>}|w_{i-1})P(w_i|\text{<U-t>}) \end{aligned} \quad (3)$$

Here, we made an assumption that the spelling of an unknown word solely depends on its part of speech and is independent of the previous word. This is the same assumption made in the hidden Markov model, which is called output independence.

The probabilities $P(\text{<U-t>}|w_{i-1})$ can be estimated from the relative frequencies in the training corpus whose infrequent words are replaced with their corresponding unknown word tags based on their part of speeches [2] .

Table 1 shows examples of word bigrams including unknown word tags. Here, a word is represented by a list of surface form, pronunciation, and part of speech, which are delimited by a slash '/'. The first example "の/no/particle <U-noun>" will appear in the most frequent form of Japanese noun phrases "A の B", which corresponds to "B of A" in English.

As Table 1 shows, word bigrams whose infrequent words are replaced with their corresponding part of speech-based unknown word tags are very important information source of the contexts where unknown words appears.

3 Unknown Word Model

3.1 Baseline Model

The simplest unknown word model depends only on the spelling. We think of an unknown word as a word having a special part of speech <UNK>. Then, the unknown word model is formally defined as the joint probability of the character sequence $w_i = c_1 \ldots c_k$ if it is an unknown word. Without loss of generality, we decompose it into the product of word length probability and word spelling probability given its length,

$$\begin{aligned} P(w_i|\text{<UNK>}) = P(c_1 \ldots c_k|\text{<UNK>}) = \\ P(k|\text{<UNK>})P(c_1 \ldots c_k|k, \text{<UNK>}) \end{aligned} \quad (4)$$

where k is the length of the character sequence. We call $P(k|\text{<UNK>})$ the word length model, and $P(c_1 \ldots c_k|k, \text{<UNK>})$ the word spelling model.

In order to estimate the entropy of English, (Brown et al., 1992) approximated $P(k|\text{<UNK>})$ by a Poisson distribution whose parameter is the average word length λ in the training corpus, and $P(c_1 \ldots c_k|k, \text{<UNK>})$ by the product of character zerogram probabilities. This means all characters in the character set are considered to be selected independently and uniformly.

$$P(c_1 \ldots c_k|\text{<UNK>}) \approx \frac{\lambda^k}{k!} e^{-\lambda} p^k \quad (5)$$

where p is the inverse of the number of characters in the character set. If we assume JIS-X-0208 is used as the Japanese character set, $p = 1/6879$.

Since the Poisson distribution is a single parameter distribution with lower bound, it is appropriate to use it as a first order approximation to the word length distribution. But the Brown model has two problems. It assigns a certain amount of probability mass to zero-length words, and it is too simple to express morphology.

For Japanese word segmentation and OCR error correction, (Nagata, 1996) proposed a modified version of the Brown model. Nagata also assumed the word length probability obeys the Poisson distribution. But he moved the lower bound from zero to one.

$$P(k|\text{<UNK>}) \approx \frac{(\lambda - 1)^{k-1}}{(k-1)!} e^{-(\lambda-1)} \quad (6)$$

[2] Throughout in this paper, we use the term "infrequent words" to represent words that appeared only once in the corpus. They are also called "hapax legomena" or "hapax words". It is well known that the characteristics of hapax legomena are similar to those of unknown words (Baayen and Sproat, 1996).

Instead of zerogram, He approximated the word spelling probability $P(c_1 \ldots c_k|k, <\text{UNK}>)$ by the product of word-based character bigram probabilities, regardless of word length.

$$P(c_1 \ldots c_k|k, <\text{UNK}>) \approx$$
$$P(c_1|<\text{bow}>) \prod_{i=2}^{k} P(c_i|c_{i-1}) P(<\text{eow}>|c_k) \quad (7)$$

where <bow> and <eow> are special symbols that indicate the beginning and the end of a word.

3.2 Correction of Word Spelling Probabilities

We find that Equation (7) assigns too little probabilities to long words (5 or more characters). This is because the lefthand side of Equation (7) represents the probability of the string $c_1 \ldots c_k$ in the set of all strings whose length are k, while the righthand side represents the probability of the string in the set of all possible strings (from length zero to infinity).

Let $P_b(c_1 \ldots c_k|<\text{UNK}>)$ be the probability of character string $c_1 \ldots c_k$ estimated from the character bigram model.

$$P_b(c_1 \ldots c_k|<\text{UNK}>) =$$
$$P(c_1|<\text{bow}>) \prod_{i=2}^{k} P(c_i|c_{i-1}) P(<\text{eow}>|c_k) \quad (8)$$

Let $P_b(k|<\text{UNK}>)$ be the sum of the probabilities of all strings which are generated by the character bigram model and whose length are k. More appropriate estimate for $P(c_1 \ldots c_k|k, <\text{UNK}>)$ is,

$$P(c_1 \ldots c_k|k, <\text{UNK}>) \approx \frac{P_b(c_1 \ldots c_k|<\text{UNK}>)}{P_b(k|<\text{UNK}>)} \quad (9)$$

But how can we estimate $P_b(k|<\text{UNK}>)$? It is difficult to compute it directly, but we can get a reasonable estimate by considering the unigram case.

If strings are generated by the character unigram model, the sum of the probabilities of all length k strings equals to the probability of the event that the end of word symbol <eow> is selected after a character other than <eow> is selected $k-1$ times.

$$P_b(k|<\text{UNK}>) \approx (1 - P(<\text{eow}>))^{k-1} P(<\text{eow}>) \quad (10)$$

Throughout in this paper, we used Equation (9) to compute the word spelling probabilities.

3.3 Japanese Orthography and Word Length Distribution

In word segmentation, one of the major problems of the word length model of Equation (6) is the decomposition of unknown words. When a substring of an unknown word coincides with other word in the dictionary, it is very likely to be decomposed into the dictionary word and the remaining substring. We find the reason of the decomposition is that the word length model does not reflect the variation of the word length distribution resulting from the Japanese orthography.

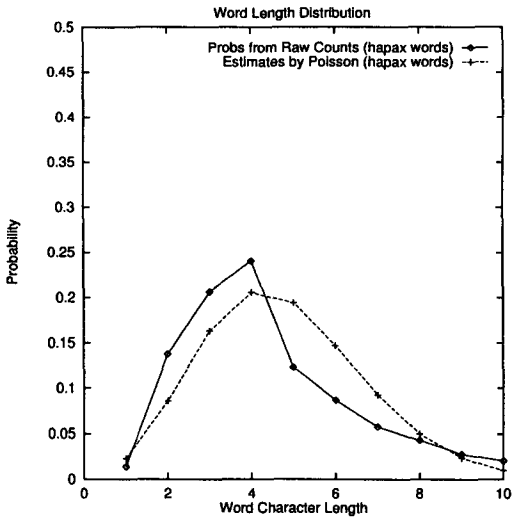

Figure 1: Word length distribution of unknown words and its estimate by Poisson distribution

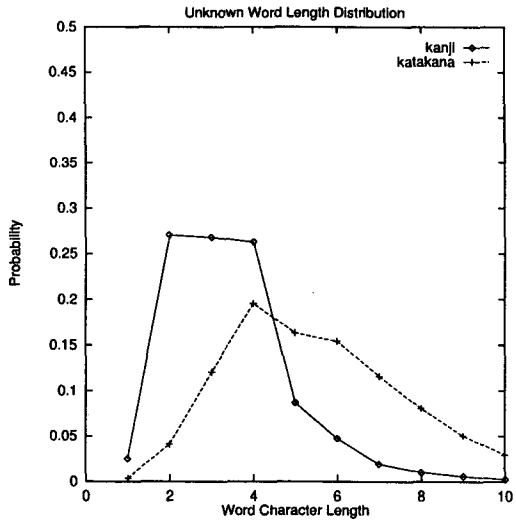

Figure 2: Word length distribution of *kanji* words and *katakana* words

Figure 1 shows the word length distribution of infrequent words in the EDR corpus, and the estimate of word length distribution by Equation (6) whose parameter ($\lambda = 4.8$) is the average word length of infrequent words. The empirical and the estimated distributions agree fairly well. But the estimates by Poisson are smaller than empirical probabilities for shorter words (<= 4 characters), and larger for longer words (> characters). This is because we rep-

Table 2: Character type configuration of infrequent words in the EDR corpus

character type sequence	percent	examples
kanji	45.1%	温泉街
katakana	11.4%	エスプレッソ
katakana-kanji	6.5%	ベル研究所
kanji-hiragana	5.6%	玉ねぎ, 極ま
hiragana	3.7%	なれそめ
kanji-katakana	3.4%	交通ルール
katakana-symbol-katakana	3.0%	ビル・ゲーツ
number	2.6%	００７
kanji-hiragana-kanji	2.4%	飲み会, 競り合
alphabet	2.0%	ＶＳＯＰ
kanji-hiragana-kanji-hiragana	1.7%	思い違い
hiragana-kanji	1.3%	えい児, おけ屋

resented all unknown words by one length model.

Figure 2 shows the word length distribution of words consists of only *kanji* characters and words consists of only *katakana* characters. It shows that the length of *kanji* words distributes around 3 characters, while that of *katakana* words distributes around 5 characters. The empirical word length distribution of Figure 1 is, in fact, a weighted sum of these two distributions.

In the Japanese writing system, there are at least five different types of characters other than punctuation marks: *kanji*, *hiragana*, *katakana*, Roman alphabet, and Arabic numeral. *Kanji* which means 'Chinese character' is used for both Chinese origin words and Japanese words semantically equivalent to Chinese characters. *Hiragana* and *katakana* are syllabaries: The former is used primarily for grammatical function words, such as particles and inflectional endings, while the latter is used primarily to transliterate Western origin words. Roman alphabet is also used for Western origin words and acronyms. Arabic numeral is used for numbers.

Most Japanese words are written in *kanji*, while more recent loan words are written in *katakana*. *Katakana* words are likely to be used for technical terms, especially in relatively new fields like computer science. *Kanji* words are shorter than *katakana* words because *kanji* is based on a large ($> 6{,}000$) alphabet of ideograms while *katakana* is based on a small (< 100) alphabet of phonograms.

Table 2 shows the distribution of character type sequences that constitute the infrequent words in the EDR corpus. It shows approximately 65% of words are constituted by a single character type. Among the words that are constituted by more than two character types, only the kanji-hiragana and hiragana-kanji sequences are morphemes and others are compound words in a strict sense although they

Table 3: Examples of common character bigrams for each part of speech in the infrequent words

part of speech	character bigram		frequency
noun	ー	<eow>	1343
number	<bow>	1	484
adjectival-verb	的	<eow>	327
verb	け	<eow>	213
adjective	し	<eow>	69
adverb	り	<eow>	63

are identified as words in the EDR corpus [3] .

Therefore, we classified Japanese words into 9 word types based on the character types that constitute a word: <sym>, <num>, <alpha>, <hira>, <kata>, and <kan> represent a sequence of symbols, numbers, alphabets, *hiraganas*, *katakanas*, and *kanjis*, respectively. <kan-hira> and <hira-kan> represent a sequence of *kanjis* followed by *hiraganas* and that of *hiraganas* followed by *kanjis*, respectively. The rest are classified as <misc>.

The resulting unknown word model is as follows. We first select the word type, then we select the length and spelling.

$$P(c_1 \ldots c_k|\text{<UNK>}) = P(\text{<WT>}|\text{<UNK>})P(k|\text{<WT>},\text{<UNK>}) P(c_1 \ldots c_k|k,\text{<WT>},\text{<UNK>}) \quad (11)$$

3.4 Part of Speech and Word Morphology

It is obvious that the beginnings and endings of words play an important role in tagging part of speech. Table 3 shows examples of common character bigrams for each part of speech in the infrequent words of the EDR corpus. The first example in Table 3 shows that words ending in '—' are likely to be nouns. This symbol typically appears at the end of transliterated Western origin words written in *katakana*.

It is natural to make a model for each part of speech. The resulting unknown word model is as follows.

$$P(c_1 \ldots c_k|\text{<U-t>}) = P(k|\text{<U-t>})P(c_1 \ldots c_k|k,\text{<U-t>}) \quad (12)$$

By introducing the distinction of word type to the model of Equation (12), we can derive a more sophisticated unknown word model that reflects both word

[3] When a Chinese character is used to represent a semantically equivalent Japanese verb, its root is written in the Chinese character and its inflectional suffix is written in *hiragana*. This results in kanji-hiragana sequence. When a Chinese character is too difficult to read, it is transliterated in *hiragana*. This results in either hiragana-kanji or kanji-hiragana sequence.

type and part of speech information. This is the unknown word model we propose in this paper. It first selects the word type given the part of speech, then the word length and spelling.

$$P(c_1 \ldots c_k|\text{<U-t>}) = P(\text{<WT>}|\text{<U-t>})P(k|\text{<WT>},\text{<U-t>}) P(c_1 \ldots c_k|k,\text{<WT>},\text{<U-t>}) \quad (13)$$

The first factor in the righthand side of Equation (13) is estimated from the relative frequency of the corresponding events in the training corpus.

$$P(\text{<WT>}|\text{<U-t>}) = \frac{C(\text{<WT>},\text{<U-t>})}{C(\text{<U-t>})} \quad (14)$$

Here, $C(\cdot)$ represents the counts in the corpus. To estimate the probabilities of the combinations of word type and part of speech that did not appeared in the training corpus, we used the Witten-Bell method (Witten and Bell, 1991) to obtain an estimate for the sum of the probabilities of unobserved events. We then redistributed this evenly among all unobserved events [4] .

The second factor of Equation (13) is estimated from the Poisson distribution whose parameter $\lambda_{\text{<WT>,<U-t>}}$ is the average length of words whose word type is <WT> and part of speech is <U-t>.

$$P(k|\text{<WT>},\text{<U-t>}) = \frac{(\lambda_{\text{<WT>,<U-t>}}-1)^{k-1}}{(k-1)!} e^{-(\lambda_{\text{<WT>,<U-t>}}-1)} \quad (15)$$

If the combinations of word type and part of speech that did not appeared in the training corpus, we used the average word length of all words.

To compute the third factor of Equation (13), we have to estimate the character bigram probabilities that are classified by word type and part of speech. Basically, they are estimated from the relative frequency of the character bigrams for each word type and part of speech.

$$f(c_i|c_{i-1},\text{<WT>},\text{<U-t>}) = \frac{C(\text{<WT>},\text{<U-t>},c_{i-1},c_i)}{C(\text{<WT>},\text{<U-t>},c_{i-1})} \quad (16)$$

However, if we divide the corpus by the combination of word type and part of speech, the amount of each training data becomes very small. Therefore, we linearly interpolated the following five probabilities (Jelinek and Mercer, 1980).

$$P(c_i|c_{i-1},\text{<WT>},\text{<U-t>}) = \alpha_1 f(c_i,\text{<WT>},\text{<U-t>}) + \alpha_2 f(c_i|c_{i-1},\text{<WT>},\text{<U-t>}) + \alpha_3 f(c_i) + \alpha_4 f(c_i|c_{i-1}) + \alpha_5(1/V) \quad (17)$$

Where $\alpha_1+\alpha_2+\alpha_3+\alpha_4+\alpha_5 = 1$. $f(c_i,\text{<WT>},\text{<U-t>})$ and $f(c_i|c_{i-1},\text{<WT>},\text{<U-t>})$ are the relative frequencies of the character unigram and bigram for each word type and part of speech. $f(c_i)$ and $f(c_i|c_{i-1})$ are the relative frequencies of the character unigram and bigram. V is the number of characters (not *tokens* but *types*) appeared in the corpus.

[4] The Witten-Bell method estimates the probability of observing novel events to be $r/(n+r)$, where n is the total number of events seen previously, and r is the number of symbols that are distinct. The probability of the event observed c times is $c/(n+r)$.

Table 4: The amount of training and test sets

	training set	test set-1	test set-2
sentences	100,000	100,000	5,000
word tokens	2,460,188	2,465,441	122,064
char tokens	3,897,718	3,906,260	192,818

4 Experiments

4.1 Training and Test Data for the Language Model

We used the EDR Japanese Corpus Version 1.0 (EDR, 1991) to train the language model. It is a manually word segmented and tagged corpus of approximately 5.1 million words (208 thousand sentences). It contains a variety of Japanese sentences taken from newspapers, magazines, dictionaries, encyclopedias, textbooks, etc..

In this experiment, we randomly selected two sets of 100 thousand sentences. The first 100 thousand sentences are used for training the language model. The second 100 thousand sentences are used for testing. The remaining 8 thousand sentences are used as a heldout set for smoothing the parameters.

For the evaluation of the word segmentation accuracy, we randomly selected 5 thousand sentences from the test set of 100 thousand sentences. We call the first test set (100 thousand sentences) "test set-1" and the second test set (5 thousand sentences) "test set-2". Table 4 shows the number of sentences, words, and characters of the training and test sets.

There were 94,680 distinct words in the training test. We discarded the words whose frequency was one, and made a dictionary of 45,027 words. After replacing the words whose frequency was one with the corresponding unknown word tags, there were 474,155 distinct word bigrams. We discarded the bigrams with frequency one, and the remaining 175,527 bigrams were used in the word segmentation model.

As for the unknown word model, word-based character bigrams are computed from the words with

Table 5: Cross entropy (CE) per word and character perplexity (PP) of each unknown word model

unknown word model	CE per word	char PP
Poisson+zerogram	59.4	2032
Poisson+bigram	37.8	128
WT+Poisson+bigram	33.3	71

frequency one (49,653 words). There were 3,120 distinct character unigrams and 55,486 distinct character bigrams. We discarded the bigram with frequency one and remaining 20,775 bigrams were used. There were 12,633 distinct character unigrams and 80,058 distinct character bigrams when we classified them for each word type and part of speech. We discarded the bigrams with frequency one and remaining 26,633 bigrams were used in the unknown word model.

Average word lengths for each word type and part of speech were also computed from the words with frequency one in the training set.

4.2 Cross Entropy and Perplexity

Table 5 shows the cross entropy per word and character perplexity of three unknown word model. The first model is Equation (5), which is the combination of Poisson distribution and character zerogram (Poisson + zerogram). The second model is the combination of Poisson distribution (Equation (6)) and character bigram (Equation (7)) (Poisson + bigram). The third model is Equation (11), which is a set of word models trained for each word type (WT + Poisson + bigram). Cross entropy was computed over the words in test set-1 that were not found in the dictionary of the word segmentation model (56,121 words). Character perplexity is more intuitive than cross entropy because it shows the average number of equally probable characters out of 6,879 characters in JIS-X-0208.

Table 5 shows that by changing the word spelling model from zerogram to bigram, character perplexity is greatly reduced. It also shows that by making a separate model for each word type, character perplexity is reduced by an additional 45% (128 → 71). This shows that the word type information is useful for modeling the morphology of Japanese words.

4.3 Part of Speech Prediction Accuracy without Context

Figure 3 shows the part of speech prediction accuracy of two unknown word model without context. It shows the accuracies up to the top 10 candidates. The first model is Equation (12), which is a set of word models trained for each part of speech (POS + Poisson + bigram). The second model is Equation (13), which is a set of word models trained for each part of speech and word type (POS + WT + Poisson + bigram). The test words are the same 56,121 words used to compute the cross entropy.

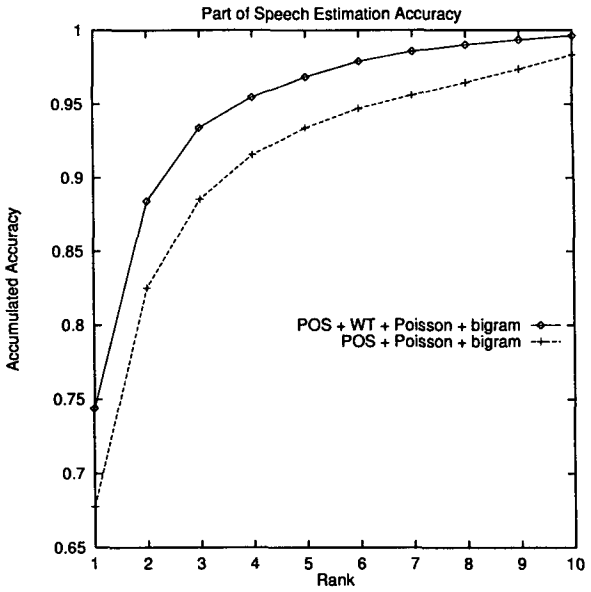

Figure 3: Accuracy of part of speech estimation

Since these unknown word models give the probability of spelling for each part of speech $P(w|t)$, we used the empirical part of speech probability $P(t)$ to compute the joint probability $P(w,t)$. The part of speech t that gives the highest joint probability is selected.

$$\hat{t} = \arg\max_t P(w,t) = P(t)P(w|t) \qquad (18)$$

The part of speech prediction accuracy of the first and the second model was 67.5% and 74.4%, respectively. As Figure 3 shows, word type information improves the prediction accuracy significantly.

4.4 Word Segmentation Accuracy

Word segmentation accuracy is expressed in terms of recall and precision as is done in the previous research (Sproat et al., 1996). Let the number of words in the manually segmented corpus be Std, the number of words in the output of the word segmenter be Sys, and the number of matched words be M. *Recall* is defined as M/Std, and *precision* is defined as M/Sys. Since it is inconvenient to use both recall and precision all the time, we also use the F-measure to indicate the overall performance. It is calculated by

$$F = \frac{(\beta^2 + 1.0) \times P \times R}{\beta^2 \times P + R} \qquad (19)$$

where P is precision, R is recall, and β is the relative importance given to recall over precision. We set

Table 6: Word segmentation accuracy of all words

	rec	prec	F
Poisson+bigram	94.5	93.1	93.8
WT+Poisson+bigram	94.4	93.8	94.1
POS+Poisson+bigram	94.4	93.6	94.0
POS+WT+Poisson+bigram	94.6	93.7	94.1

Table 7: Word segmentation accuracy of unknown words

	rec	prec	F
Poisson + bigram	31.8	65.0	42.7
WT+Poisson+bigram	45.5	62.0	52.5
POS+Poisson+bigram	39.7	61.5	48.3
POS+WT+Poisson+bigram	42.0	66.4	51.4

$\beta = 1.0$ throughout this experiment. That is, we put equal importance on recall and precision.

Table 6 shows the word segmentation accuracy of four unknown word models over test set-2. Compared to the baseline model (Poisson + bigram), by using word type and part of speech information, the precision of the proposed model (POS + WT + Poisson + bigram) is improved by a modest 0.6%. The impact of the proposed model is small because the out-of-vocabulary rate of test set-2 is only 3.1%.

To closely investigate the effect of the proposed unknown word model, we computed the word segmentation accuracy of unknown words. Table 7 shows the results. The accuracy of the proposed model (POS + WT + Poisson + bigram) is significantly higher than the baseline model (Poisson + bigram). Recall is improved from 31.8% to 42.0% and precision is improved from 65.0% to 66.4%.

Here, recall is the percentage of correctly segmented unknown words in the system output to the all unknown words in the test sentences. Precision is the percentage of correctly segmented unknown words in the system's output to the all words that system identified as unknown words.

Table 8 shows the tagging accuracy of unknown words. Notice that the baseline model (Poisson + bigram) cannot predict part of speech. To roughly estimate the amount of improvement brought by the proposed model, we applied a simple tagging strategy to the output of the baseline model. That is, words that include numbers are tagged as numbers, and others are tagged as nouns.

Table 8 shows that by using word type and part of speech information, recall is improved from 28.1% to 40.6% and precision is improved from 57.3% to 64.1%.

Other than the usual recall/precision measures, we defined another precision (prec2 in Table 8), which roughly correspond to the tagging accuracy in English where word segmentation is trivial. Prec2 is defined as the percentage of correctly tagged unknown words to the correctly segmented unknown words. Table 8 shows that tagging precision is improved from 88.2% to 96.6%. The tagging accuracy in context (96.6%) is significantly higher than that without context (74.4%). This shows that the word bigrams using unknown word tags for each part of speech are useful to predict the part of speech.

5 Related Work

Since English uses spaces between words, unknown words can be identified by simple dictionary lookup. So the topic of interest is part of speech estimation. Some statistical model to estimate the part of speech of unknown words from the case of the first letter and the prefix and suffix is proposed (Weischedel et al., 1993; Brill, 1995; Ratnaparkhi, 1996; Mikheev, 1997). On the contrary, since Asian languages like Japanese and Chinese do not put spaces between words, previous work on unknown word problem is focused on word segmentation; there are few studies estimating part of speech of unknown words in Asian languages.

The cues used for estimating the part of speech of unknown words for Japanese in this paper are basically the same for English, namely, the prefix and suffix of the unknown word as well as the previous and following part of speech. The contribution of this paper is in showing the fact that different character sets behave differently in Japanese and a better word model can be made by using this fact.

By introducing different length models based on character sets, the number of decomposition errors of unknown words are significantly reduced. In other words, the tendency of over-segmentation is corrected. However, the spelling model, especially the character bigrams in Equation (17) are hard to estimate because of the data sparseness. This is the main reason of the remaining under-segmented and over-segmented errors.

To improve the unknown word model, feature-based approach such as the maximum entropy method (Ratnaparkhi, 1996) might be useful, because we don't have to divide the training data into several disjoint sets (like we did by part of speech and word type) and we can incorporate more linguistic and morphological knowledge into the same probabilistic framework. We are thinking of re-implementing our unknown word model using the maximum entropy method as the next step of our research.

Table 8: Part of speech tagging accuracy of unknown words (the last column represents the percentage of correctly tagged unknown words in the correctly segmented unknown words)

	rec	prec	F	prec2
Poisson+bigram	28.1	57.3	37.7	88.2
WT+Poisson+bigram	37.7	51.5	43.5	87.9
POS+Poisson+bigram	37.5	58.1	45.6	94.3
POS+WT+Poisson+bigram	40.6	64.1	49.7	96.6

6 Conclusion

We present a statistical model of Japanese unknown words using word morphology and word context. We find that Japanese words are better modeled by classifying words based on the character sets (*kanji, hiragana, katakana*, etc.) and its changes. This is because the different character sets behave differently in many ways (historical etymology, ideogram vs. phonogram, etc.). Both word segmentation accuracy and part of speech tagging accuracy are improved by treating them differently.

References

Harald Baayen and Richard Sproat. 1996. Estimating lexical priors for low-frequency morphologically ambiguous forms. *Computational Linguistics*, 22(2):155–166.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jennifer C. Lai, and Robert L. Mercer. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

Jing-Shin Chang, Yi-Chung Lin, and Keh-Yih Su. 1995. Automatic construction of a Chinese electronic dictionary. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 107–120.

EDR. 1991. EDR electronic dictionary version 1 technical guide. Technical Report TR2-003, Japan Electronic Dictionary Research Institute.

Pascale Fung and Dekai Wu. 1994. Statistical augmentation of a Chinese machine-readable dictionary. In *Proceedings of the Second Workshop on Very Large Corpora*, pages 69–85.

Masahiko Haruno and Yuji Matsumoto. 1997. Mistake-driven mixture of hierachical tag context trees. In *Proceedings of the 35th ACL and 8th EACL*, pages 230–237.

F. Jelinek and R. L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397.

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 the 16th International Conference on Computational Linguistics*, pages 1119–1122.

Masaaki Nagata. 1994. A stochastic Japanese morphological analyzer using a forward-dp backward-A^* n-best search algorithm. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 201–207.

Masaaki Nagata. 1996. Context-based spelling correction for Japanese OCR. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 806–811.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 22(3):377–404.

Koichi Takeuchi and Yuji Matsumoto. 1997. HMM parameter learning for Japanese morphological analyzer. *Transaction of Information Processing of Japan*, 38(3):500–509. (in Japanese).

Ralph Weischedel, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmucci. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, 19(2):359–382.

Ian H. Witten and Timothy C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transaction on Information Theory*, 37(4):1085–1094.

Mikio Yamamoto. 1996. A re-estimation method for stochastic language modeling from ambiguous observations. In *Proceedings of the Fourth Workshop on Very Large Corpora*, pages 155–167.