The Unknown Word Problem: a Morphological Analysis of Japanese Using Maximum Entropy Aided by a Dictionary

Kiyotaka Uchimoto†, Satoshi Sekine‡ and **Hitoshi Isahara†**

†Communications Research Laboratory
2-2-2, Hikari-dai, Seika-cho, Soraku-gun,
Kyoto, 619-0289 Japan
[uchimoto, isahara]@crl.go.jp

‡New York University
715 Broadway, 7th floor
New York, NY 10003, USA
sekine@cs.nyu.edu

Abstract

In this paper we describe a morphological analysis method based on a maximum entropy model. This method uses a model that can not only consult a dictionary with a large amount of lexical information but can also identify unknown words by learning certain characteristics. The model has the potential to overcome the unknown word problem.

1 Introduction

Morphological analysis is one of the basic techniques used in Japanese sentence analysis. A morpheme is a minimal grammatical unit, such as a word or a suffix, and morphological analysis is the process segmenting a given sentence into a row of morphemes and assigning to each morpheme grammatical attributes such as a part-of-speech (POS) and an inflection type. One of the most important problems in morphological analysis is that posed by unknown words, which are words found in neither a dictionary nor a training corpus, and there have been two statistical approaches to this problem. One is to acquire unknown words from corpora and put them into a dictionary (e.g., (Mori and Nagao, 1996)), and the other is to estimate a model that can identify unknown words correctly (e.g., (Kashioka et al., 1997; Nagata, 1999)). We would like to be able to make good use of both approaches. If words acquired by the former method could be added to a dictionary and a model developed by the latter method could consult the amended dictionary, then the model could be the best statistical model which has the potential to overcome the unknown word problem. Mori and Nagao proposed a statistical model that can consult a dictionary (Mori and Nagao, 1998). In their model the probability that a string of letters or characters is a morpheme is augmented when the string is found in a dictionary. The improvement of the accuracy was slight, however, so we think that it is difficult to efficiently integrate the mechanism for consulting a dictionary into an n-gram model. In this paper we therefore describe a morphological analysis method based on a maximum entropy (M.E.) model. This method uses a model that can not only consult a dictionary but can also identify unknown words by learning certain characteristics. To learn these characteristics, we focused on such information as whether or not a string is found in a dictionary and what types of characters are used in a string. The model estimates how likely a string is to be a morpheme according to the information on hand. When our method was used to identify morpheme segments in sentences in the Kyoto University corpus and to identify the major parts-of-speech of these morphemes, the recall and precision were respectively 95.80% and 95.09%.

2 A Morpheme Model

This section describes a model which estimates how likely a string is to be a morpheme. We implemented this model within an M.E. framework.

Given a tokenized test corpus, the problem of Japanese morphological analysis can be reduced to the problem of assigning one of two tags to each string in a sentence. A string is tagged with a 1 or a 0 to indicate whether or not it is a morpheme. When a string is a morpheme, a grammatical attribute is assigned to it. The 1 tag is thus divided into the number, n, of grammatical attributes assigned to morphemes, and the problem is to assign an attribute (from 0 to n) to every string in a given sentence. The $(n+1)$ tags form the space of "fu-

tures" in the M.E. formulation of our problem of morphological analysis. The M.E. model, as well as other similar models, enables the computation of $P(f|h)$ for any future f from the space of possible futures, F, and for every history, h, from the space of possible histories, H. A "history" in M.E. is all of the conditioning data that enable us to make a decision in the space of futures. In the problem of morphological analysis, we can reformulate this in terms of finding the probability of f associated with the relationship at index t in the test corpus:

$$P(f|h_t) = P(f|\text{Information derivable from the test corpus related to relationship } t)$$

The computation of $P(f|h)$ in any M.E. models is dependent on a set of "features" which would be helpful in making a prediction about the future. Like most current M.E. models in computational linguistics, our model is restricted to those features which are binary functions of the history and future. For instance, one of our features is

$$g(h,f) = \begin{cases} 1: \text{ if has}(h,x) = \text{true}, \\ \quad x = \text{"POS}(-1)\text{(Major)} : \text{verb,"} \\ \quad \& \ f = 1 \\ 0: \text{ otherwise.} \end{cases} \quad (1)$$

Here "has(h,x)" is a binary function that returns true if the history h has feature x. In our experiments, we focused on such information as whether or not a string is found in a dictionary, the length of the string, what types of characters are used in the string, and the part-of-speech of the adjacent morpheme.

Given a set of features and some training data, the M.E. estimation process produces a model in which every feature g_i has an associated parameter α_i. This enables us to compute the conditional probability as follows (Berger et al., 1996):

$$P(f|h) = \frac{\prod_i \alpha_i^{g_i(h,f)}}{Z_\lambda(h)} \quad (2)$$

$$Z_\lambda(h) = \sum_f \prod_i \alpha_i^{g_i(h,f)}. \quad (3)$$

The M.E. estimation process guarantees that for every feature g_i, the expected value of g_i according to the M.E. model will equal the empirical expectation of g_i in the training corpus. In other words,

$$\sum_{h,f} \tilde{P}(h,f) \cdot g_i(h,f) = \sum_h \tilde{P}(h) \cdot \sum_f P_{M.E.}(f|h) \cdot g_i(h,f). \quad (4)$$

Here $\tilde{P}$ is an empirical probability and $P_{M.E.}$ is the probability assigned by the model.

We define part-of-speech and bunsetsu boundaries as grammatical attributes. Here a bunsetsu is a phrasal unit consisting of one or more morphemes. When there are m types of parts-of-speech, and the left-hand side of each morpheme may or may not be a bunsetsu boundary, the number, n, of grammatical attributes assigned to morphemes is $2 \times m$. [1] We propose a model which estimates the likelihood that a given string is a morpheme and has the grammatical attribute $i (1 \leq i \leq n)$. We call it a *morpheme model*. This model is represented by Eq. (2), in which f can be one of $(n+1)$ tags from 0 to n.

A given sentence is divided into morphemes, and a grammatical attribute is assigned to each morpheme so as to maximize the sentence probability estimated by our morpheme model. Sentence probability is defined as the product of the probabilities estimated for a particular division of morphemes in a sentence. We use the Viterbi algorithm to find the optimal set of morphemes in a sentence and we use the method proposed by Nagata (Nagata, 1994) to search for the N-best sets.

3 Experiments and Discussion

3.1 Experimental Conditions

The part-of-speech categories that we used follow those of JUMAN (Kurohashi and Nagao, 1999). There are 53 categories covering all possible combinations of major and minor categories as defined in JUMAN. The number of grammatical attributes is 106 if we include the detection of whether or not the left side of a morpheme is a bunsetsu boundary. We do not identify inflection types probabilistically since

[1] Not only morphemes but also bunsetsus can be identified by considering the information related to their bunsetsu boundaries.

they can be almost perfectly identified by checking the spelling of the current morpheme after a part-of-speech has been assigned to it. Therefore, f in Eq. (2) can be one of 107 tags from 0 to 106.

We used the Kyoto University text corpus (Version 2) (Kurohashi and Nagao, 1997), a tagged corpus of the Mainichi newspaper. For training, we used 7,958 sentences from newspaper articles appearing from January 1 to January 8, 1995, and for testing, we used 1,246 sentences from articles appearing on January 9, 1995.

Given a sentence, for every string consisting of five or less characters and every string appearing in the JUMAN dictionary (Kurohashi and Nagao, 1999), whether or not the string is a morpheme was determined and then the grammatical attribute of each string determined to be a morpheme was identified and assigned to that string. The maximum length was set at five because morphemes consisting of six or more characters are mostly compound words or words consisting of katakana characters. The stipulation that strings consisting of six or more characters appear in the JUMAN dictionary was set because long strings not present in the JUMAN dictionary were rarely found to be morphemes in our training corpus. Here we assume that compound words that do not appear in the JUMAN dictionary can be divided into strings consisting of five or less characters because compound words tend not to appear in dictionaries, and in fact, compound words which consist of six or more characters and do not appear in the dictionary were not found in our training corpus. Katakana strings that are not found in the JUMAN dictionary were assumed to be included in the dictionary as an entry having the part-of-speech "Unknown(Major), Katakana(Minor)." An optimal set of morphemes in a sentence is searched for by employing the Viterbi algorithm under the condition that connectivity rules defined between parts-of-speech in JUMAN must be met. The assigned part-of-speech in the optimal set is not always selected from the parts-of-speech attached to entries in the JUMAN dictionary, but may also be selected from the 53 categories of the M.E. model. It is difficult to select an appropriate category from the 53 when there is little training data, so we assume that every entry in the JUMAN dictionary has all possible parts-of-speech, and the part-of-speech assigned to each morpheme is selected from those attached to the entry corresponding to the morpheme string.

The features used in our experiments are listed in Table 1. Each row in Table 1 contains a feature type, feature values, and an experimental result that will be explained later. Each feature consists of a type and a value. The features are basically some attributes of the morpheme itself or those of the morpheme to the left of it. We used the 31,717 features that were found three or more times in the training corpus. The notations "(0)" and "(-1)" used in the feature type column in Table 1 respectively indicate a target string and the morpheme on the left of it.

The terms used in the table are the following:

String: Strings which appeared as a morpheme five or more times in the training corpus

Length: Length of a string

POS: Part-of-speech. "Major" and "Minor" respectively indicate major and minor part-of-speech categories as defined in JUMAN.

Inf: Inflection type as defined in JUMAN

Dic: We use the JUMAN dictionary, which has about 200,000 entries (Kurohashi and Nagao, 1999). "Major&Minor" indicates possible combinations between major and minor part-of-speech categories. When the target string is in the dictionary, the part-of-speech attached to the entry corresponding to the string is used as a feature value. If an entry has two or more parts-of-speech, the part-of-speech which leads to the highest probability in a sentence estimated from our model is selected as a feature value. JUMAN has another type of dictionary, which is called a *phrase dictionary*. Each entry in the phrase dictionary consists of one or more morphemes such as "と (*to*, case marker), は (*wa*, topic marker), いえ (*ie*, say)." JUMAN uses this dictionary to detect morphemes which need a longer context to be identified correctly. When the target string corresponds to the string of the left most morpheme in the phrase dictionary in JUMAN, the part-of-speech at-

Table 1: Features.

Feature number	Feature type	Feature value (Number of value)	Accuracy without each feature set		
			Recall	Precision	F-measure
1	String(0)	(4,331)	93.66%	93.81%	93.73
2	String(-1)	(4,331)	(−2.14%)	(−1.28%)	(−1.71)
3	Dic(0)(Major)	Verb, Verb&Phrase, Adj, Adj&Phrase, ... (28)	94.64%	92.87%	93.75
4	Dic(0)(Minor)	Common_noun, Common_noun&Phrase, Topic_marker, ... (90)	(−1.16%)	(−2.22%)	(−1.69)
5	Dic(0)(Major&Minor)	Noun&Common_noun, Noun&Common_noun&Phrase, ... (103)			
6	Length(0)	1, 2, 3, 4, 5, 6_or_more (6)	95.52%	94.11%	94.81
7	Length(-1)	1, 2, 3, 4, 5, 6_or_more (6)	(−0.28%)	(−0.98%)	(−0.63)
8	TOC(0)(Beginning)	Kanji, Hiragana, Symbol, Number, Katakana, Alphabet (6)	95.17%	93.89%	94.52
9	TOC(0)(End)	Kanji, Hiragana, Symbol, Number, Katakana, Alphabet (6)	(−0.63%)	(−1.20%)	(−0.92)
10	TOC(0)(Transition)	Kanji→Hiragana, Number→Kanji, Katakana→Kanji, ... (30)			
11	TOC(-1)(End)	Kanji, Hiragana, Symbol, Number, Katakana, Alphabet (6)			
12	TOC(-1)(Transition)	Kanji→Hiragana, Number→Kanji, Katakana→Kanji, ... (30)			
13	POS(-1)(Major)	Verb, Adj, Noun, Unknown, ... (15)	95.60%	95.31%	95.45
14	POS(-1)(Minor)	Common_noun, Sahen_noun, Numeral, ... (45)	(−0.20%)	(+0.22%)	(+0.01)
15	POS(-1)(Major&Minor)	[nil], Noun&Common_noun, Noun&Common_noun&Phrase, ... (54)			
16	Inf(-1)(Major)	Vowel verb, ... (33)	95.66%	95.00%	95.33
17	Inf(-1)(Minor)	Stem, Basic_form, Imperative_form, ... (60)	(−0.14%)	(−0.09%)	(−0.11)
18	BB(-1)	[nil], [exist] (2)	95.82%	95.25%	95.53
19	BB(-1) & POS(-1)(Major&Minor)	Noun&Common, noun&Bunsetsu_boundary, Noun&Common, noun&Within_a_bunsetsu, ... (106)	(+0.02%)	(+0.16%)	(+0.09)

tached to the entry plus the information that it is in the phrase dictionary (such as "Verb&Phrase") is used as a feature value.

TOC: Types of characters used in a string. "(Beginning)" and "(End)" respectively represent the leftmost and rightmost characters of a string. When a string consists of only one character, the "(Beginning)" and "(End)" are the same character. "TOC(0)(Transition)" represents the transition from the leftmost character to the rightmost one in a string. "TOC(-1)(Transition)" represents the transition from the rightmost character in the adjacent morpheme on the left to the leftmost one in the target string. For example, when the adjacent morpheme on the left is "先生 (*sensei*, teacher)" and the target string is "に (*ni*, case marker)," the feature value "Kanji→Hiragana" is selected.

BB: Indicates whether or not the left side of a morpheme is a bunsetsu boundary.

3.2 Results and Discussion

Some results of the morphological analysis are listed in Table 2. *Recall* is the percentage of morphemes in the test corpus whose segmentation and major POS tag are identified correctly. *Precision* is the percentage of all morphemes identified by the system that are identified correctly. F represents the *F-measure* and is defined by the following equation.

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

Table 2 shows results obtained by using our method, by using JUMAN, and by using JUMAN plus KNP (Kurohashi, 1998). We show the result obtained using JUMAN plus KNP because JUMAN alone assigns an "Unknown" tag to katakana strings when they are not in the dictionary. All katakana strings not found

Table 2: Results of Experiments (Segmentation and major POS tagging).

	Recall	Precision	F-measure
Our method	95.80% (29,986/31,302)	95.09% (29,986/31,467)	95.44
JUMAN	95.25% (29,814/31,302)	94.90% (29,814/31,417)	95.07
JUMAN+KNP	98.49% (30,830/31,302)	98.13% (30,830/31,417)	98.31

in the dictionary are therefore evaluated as errors. KNP improves on JUMAN by replacing the "Unknown" tag with a "Noun" tag and disambiguating part-of-speech ambiguities which arise during the process of parsing when there is more than one JUMAN analysis with the same score.

The accuracy in segmentation and major POS tagging obtained with our method and that obtained with JUMAN were about 3% worse than that obtained with JUMAN plus KNP. We think the main reason for this was an insufficient amount of training data and feature sets and the inconsistency of the corpus. The number of sentences in the training corpus was only about 8,000, and we did not use as many combined features as were proposed in Ref. (Uchimoto et al., 1999). We were unable to use more training data or more feature sets because every string consisting of five or less characters in our training corpus was used to train our model, so the amount of tokenized training data would have become too large and the training would not have been completed on the available machine if we had used more training data or more feature sets. The inconsistency of the corpus was due to the way the corpus was made. The Kyoto University corpus was made by manually correcting the output of JUMAN plus KNP, and it is difficult to manually correct all of the inconsistencies in the output. The use of JUMAN plus KNP thus has an advantage over the use of our method when we evaluate a system's accuracy by using the Kyoto University corpus. For example, the number of morphemes whose rightmost character is "者" was 153 in the test corpus, and they were all the same as those in the output of JUMAN plus KNP. There were three errors (about 2%) in the output of our system. There were several inconsistencies in the test corpus such as "生産 (*seisan*, Noun), 者 (*sha*, Suffix)(producer)," and "消費者 (*shouhi-sha*, Noun)(consumer)." They should have been corrected in the corpus-making process to "生産 (*seisan*, Noun), 者 (*sha*, Suffix)(producer)," and "消費 (*shouhi*, Noun), 者 (*sha*, Suffix)(consumer)." It is difficult for our model to discriminate among these without over-training when there are such inconsistencies in the corpus. Other similar inconsistencies were, for example, "芸術家 (*geijutsu-ka*, Noun)(artist)" and "工芸 (*kougei*, Noun), 家 (*ka*, Suffix)(craftsman)," "警視庁 (*keishi-cho*, Noun)(the Metropolitan Police Board)" and "検察 (*kensatsu*, Noun), 庁 (*cho*, Noun)(the Public Prosecutor's Office)," and "現実的 (*genjitsu-teki*, Adjective)(realistic)" and "理想 (*risou*, Noun), 的 (*teki*, Suffix)(ideal).". If these had been corrected consistently when making the corpus, the accuracy obtained by our method could have been better than that shown in Table 2. A study on corpus revision should be undertaken to resolve this issue. We believe it can be resolved by using our trained model. There is a high possibility that a morpheme lacks consistency in the training corpus when its probability, re-estimated by our model, is low. Thus a method which detects morphemes having a low probability can identify those lacking consistency in the training corpus. We intend to try this in the future.

3.3 Features and Accuracy

In our model, dictionary information and certain characteristics of unknown words are reflected as features, as shown in Table 1. "String" and "Dic" reflect the dictionary information, [2] and "Length" and "TOC"(types of characters) reflect the characteristics of unknown words. Therefore, our model can not only consult a dictionary but can also detect unknown words. Table 1 shows the results of an

[2] "String" indicates strings that make up a morpheme and were found five or more times in the training corpus. Using this information as features in our M.E. model corresponds to consulting a dictionary constructed from the training corpus.

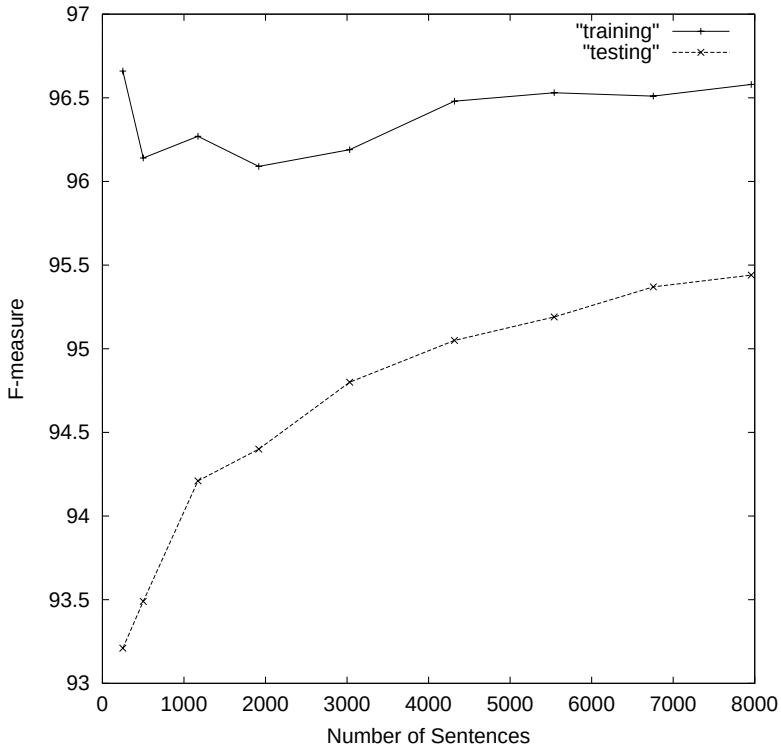

Figure 1: Relation between accuracy and the number of training sentences.

analysis without the complete feature set. Almost all of the feature sets improved accuracy. The contribution of the dictionary information was especially significant.

There were cases, however, in which the use of dictionary information led to a decrease in the accuracy. For example, we found these erroneous segmentations:
"／海 (*umi*, sea) ／に (*ni*, case marker) ／かけた (*kaketa*, bet) ／ロマンは (*romanha*, the Romantic school) ／" and "／荒波 (*aranami*, raging waves) ／に (*ni*, case marker) ／負け (*make*, lose) ／ない心 (*naishin*, one's inmost heart) ／と (*to*, case marker) ／" (Underlined strings were errors.) when the correct segmentations were:
"／海 (*umi*, sea) ／に (*ni*, case marker) ／かけた (*kaketa*, bet) ／ロマン (*roman*, romance) ／は (*wa*, topic marker) ／" and "／荒波 (*aranami*, raging waves) ／に (*ni*, case marker) ／負けない (*makenai*, not to lose) ／心 (*kokoro*, heart) ／と (*to*, case marker) ／" ("／" indicates a morphological boundary.).

These errors were caused by nonstandard entries in the JUMAN dictionary. The dictionary had not only the usual notation using kanji characters, "ロマン派" and "内心," but also the uncommon notation using hiragana strings, "ロマンは" and "ない心". To prevent this type of error, it is necessary to remove nonstandard entries from the dictionary or to investigate the frequency of such entries in large corpora and to use it as a feature.

3.4 Accuracy and the Amount of Training Data

The accuracies (F-measures) for the training corpus and the test corpus are shown in Figure 1 plotted against the number of sentences used for training. The learning curve shows that we can expect improvement if we use more training data.

3.5 Unknown Words and Accuracy

The strength of our method is that it can identify morphemes when they are unknown words and can assign appropriate parts-of-speech to them. For example, the nouns "漱石 (Souseki)" and "露伴 (Rohan)" are not found in the JU-

Table 3: Accuracy for unknown words (Recall).

	Segmentation and major POS tagging	Segmentation and minor POS tagging
	For words not found in the dictionary nor in our training corpus	
Our method	69.90% (432/618)	27.51% (170/618)
JUMAN+KNP	79.29% (490/618)	20.55% (127/618)
	For words not found in the dictionary nor in our features	
Our method	76.17% (719/944)	32.20% (304/944)
JUMAN+KNP	85.70% (809/944)	27.22% (257/944)
	For words not found in the dictionary	
Our method	82.40% (1,138/1,381)	49.24% (680/1,381)
JUMAN+KNP	89.79% (1,240/1,381)	38.60% (533/1,381)

MAN dictionary. JUMAN plus KNP analyzes them simply as "漱 (Noun) 石 (Noun)" and "露 (Adverb) 伴 (Noun)," whereas our system analyzes both of them correctly. Our system correctly identified them as names of people even though they were not in the dictionary and did not appear as features in our M.E. model. Since these names, or proper nouns, are newly coined and can be represented by a variety of expressions, no proper nouns can be included in a dictionary, nor can they appear in a training corpus; this means that proper nouns could easily be unknown words. We investigated the accuracy of our method in identifying morphemes when they are unknown words, and the results are listed in Table 3. The first row in each section shows the recall for the morphemes that were unknown words. The second row in each section shows the percentage of morphemes whose segmentation and "minor" POS tag were identified correctly. The difference between the first and second lines, the third and fourth lines, and fifth and sixth lines is the definition of unknown words. Unknown words were defined respectively as words not found in the dictionary nor in our training corpus, as words not found in the dictionary nor in our features, and as words not found in the dictionary. Our accuracy, shown as the second rows in Table 3 was more than 5% better than that of JUMAN plus KNP for each definition. These results show that our model can efficiently learn the characteristics of unknown words, especially those of proper nouns such as the names of people, organizations, and locations.

4 Related Work

Several methods based on statistical models have been proposed for the morphological analysis of Japanese sentences. An F-measure of about 96% was achieved by a method based on a hidden Markov model (HMM) (Takeuchi and Matsumoto, 1997) and by one based on a variable-memory Markov model (Haruno and Matsumoto, 1997; Kitauchi et al., 1999). Although the accuracy obtained with these methods was better than that obtained with ours, their accuracy cannot be compared directly with that of our method because their part-of-speech categories differ from ours. And an advantage of our model is that it can handle unknown words, whereas their models do not handle unknown words well. In their models, unknown words are divided into a combination of a word consisting of one character and known words. Haruno and Matsumoto (Haruno and Matsumoto, 1997) achieved a recall of about 96% when using trigram or greater information, but achieved a recall of only 94% when using bigram information. This leads us to believe that we could obtain better accuracy if we use trigram or greater information. We plan to do so in future work.

Two approaches have been used to deal with unknown words: acquiring unknown words from corpora and putting them into a dictionary (e.g., (Mori and Nagao, 1996)) and developing a model that can identify unknown words

correctly (e.g., (Kashioka et al., 1997; Nagata, 1999)). Nagata reported a recall of about 40% for unknown words (Nagata, 1999). As shown in Table 3, our method achieved a recall of 69.90% for unknown words. Our accuracy was about 30% better than his. It is difficult to compare his method with ours directly because he used a different corpus (the EDR corpus), but the part-of-speech categories and the definition of morphemes he used were similar to ours. Thus, this comparison is helpful in evaluating our method. There are no spaces between morphemes in Japanese. In general, therefore, detecting whether a given string is an unknown word or is not a morpheme is difficult when it is not found in the dictionary, nor in the training corpus. However, our model learns whether or not a given string is a morpheme and has a huge amount of data for learning what in a corpus is not a morpheme. Therefore, we believe that the characteristics of our model led to its good results for identifying unknown words.

Mori and Nagao proposed a model that can consult a dictionary (Mori and Nagao, 1998); they reported an F-measure of about 92 when using the EDR corpus and of about 95 when using the Kyoto University corpus. Their slight improvement in accuracy by using dictionary information resulted in an F-measure of about 0.2, while our improvement was about 1.7. Their accuracy of 95% when using the Kyoto University corpus is similar to ours, but they added to their dictionary all of the words appearing in the training corpus. Therefore, their experiment had to deal with fewer unknown words than ours did.

With regard to the morphological analysis of English sentences, methods for part-of-speech tagging based on an HMM (Cutting et al., 1992), a variable-memory Markov model (Schütze and Singer, 1994), a decision tree model (Daelemans et al., 1996), an M.E. model (Ratnaparkhi, 1996), a neural network model (Schmid, 1994), and a transformation-based error-driven learning model (Brill, 1995) have been proposed, as well as a combined method (Màrquez and Padró, 1997; van Halteren et al., 1998). On available machines, however, these models cannot handle a large amount of lexical information. We think that our model, which can not only consult a dictionary with a large amount of lexical information, but can also identify unknown words by learning certain characteristics, has the potential to achieve good accuracy for part-of-speech tagging in English. We plan to apply our model to English sentences.

5 Conclusion

This paper described a method for morphological analysis based on a maximum entropy (M.E.) model. This method uses a model that can not only consult a dictionary but can also identify unknown words by learning certain characteristics. To learn these characteristics, we focused on such information as whether or not a string is found in a dictionary and what types of characters are used in a string. The model estimates how likely a string is to be a morpheme according to the information on hand. When our method was used to identify morpheme segments in sentences in the Kyoto University corpus and to identify the major parts-of-speech of these morphemes, the recall and precision were respectively 95.80% and 95.09%. In our experiments without each feature set shown in Tables 1, we found that dictionary information significantly contributes to improving accuracy. We also found that our model can efficiently learn the characteristics of unknown words, especially proper nouns such as the names of people, organizations, and locations.

References

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.

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.

Doung Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. 1992. A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 133–140.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gills. 1996. MBT: A Memory-

Based Part-of-Speech Tagger-Generator. In *Proceedings of the 4th Workshop on Very Large Corpora*, pages 1–14.

Masahiko Haruno and Yuji Matsumoto. 1997. Mistake-Driven Mixture of Hierarchical-Tag Context Trees. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 230–237.

Hideki Kashioka, Stephen G. Eubank, and Ezra W. Black. 1997. Decision-Tree Morphological Analysis without a Dictionary for Japanese. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pages 541–544.

Akira Kitauchi, Takehito Utsuro, and Yuji Matsumoto. 1999. Probabilistic Model Learning for Japanese Morphological Analysis by Error-driven Feature Selection. *Transactions of Information Processing Society of Japan*, 40(5):2325–2337. (in Japanese).

Sadao Kurohashi and Makoto Nagao. 1997. Building a Japanese Parsed Corpus while Improving the Parsing System. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pages 451–456.

Sadao Kurohashi and Makoto Nagao, 1999. *Japanese Morphological Analysis System JUMAN Version 3.61.* Department of Informatics, Kyoto University.

Sadao Kurohashi, 1998. *Japanese Dependency/Case Structure Analyzer KNP Version 2.0b6.* Department of Informatics, Kyoto University.

Lluís Màrquez and Lluís Padró. 1997. A Flexible POS Tagger Using an Automatically Acquired Language Model. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 238–252.

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 (COLING96)*, pages 1119–1122.

Shinsuke Mori and Makoto Nagao. 1998. An Improvement of a Morphological Analysis by a Morpheme Clustering. *Journal of Natural Language Processing*, 5(2):75–103. (in Japanese).

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 (COLING94)*, pages 201–207.

Masaaki Nagata. 1999. A Part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 277–284.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Model for Part-Of-Speech Tagging. In *Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

Helmut Schmid. 1994. Part-Of-Speech Tagging with Neural Networks. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING94)*, pages 172–176.

Hinrich Schütze and Yoram Singer. 1994. Part-of-Speech Tagging Using a Variable Memory Markov Model. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 181–187.

Koichi Takeuchi and Yuji Matsumoto. 1997. HMM Parameter Learning for Japanese Morphological Analyzer. *Transactions of Information Processing Society of Japan*, 83(3):500–509. (in Japanese).

Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. 1999. Japanese Dependency Structure Analysis Based on Maximum Entropy Models. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 196–203.

Hans van Halteren, Jakub Zavrel, and Walter Daelemans. 1998. Improving Data Driven Wordclass Tagging by System Combination. In *Proceedings of the COLING-ACL '98*, pages 491–497.