Morphological Analysis of a Large Spontaneous Speech Corpus in Japanese

Kiyotaka Uchimoto[†] **Chikashi Nobata**[†] **Atsushi Yamada**[†]

Satoshi Sekine[‡] **Hitoshi Isahara**[†]

[†]Communications Research Laboratory
3-5, Hikari-dai, Seika-cho, Soraku-gun,
Kyoto, 619-0289, Japan
{uchimoto,nova,ark,isahara}@crl.go.jp

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

Abstract

This paper describes two methods for detecting word segments and their morphological information in a Japanese spontaneous speech corpus, and describes how to tag a large spontaneous speech corpus accurately by using the two methods. The first method is used to detect any type of word segments. The second method is used when there are several definitions for word segments and their POS categories, and when one type of word segments includes another type of word segments. In this paper, we show that by using semi-automatic analysis we achieve a precision of better than 99% for detecting and tagging short words and 97% for long words; the two types of words that comprise the corpus. We also show that better accuracy is achieved by using both methods than by using only the first.

1 Introduction

The "Spontaneous Speech: Corpus and Processing Technology" project is sponsoring the construction of a large spontaneous Japanese speech corpus, *Corpus of Spontaneous Japanese (CSJ)* (Maekawa et al., 2000). The CSJ is a collection of monologues and dialogues, the majority being monologues such as academic presentations and simulated public speeches. Simulated public speeches are short speeches presented specifically for the corpus by paid non-professional speakers. The CSJ includes transcriptions of the speeches as well as audio recordings of them. One of the goals of the project is to detect two types of word segments and corresponding morphological information in the transcriptions. The two types of word segments were defined by the members of The National Institute for Japanese Language and are called *short word* and *long word*. The term *short word* approximates a dictionary item found in an ordinary Japanese dictionary, and *long word* represents various compounds. The length and part-of-speech (POS) of each are different, and every short word is included in a long word, which is shorter than a Japanese phrasal unit, a *bunsetsu*. If all of the short words in the CSJ were detected, the number of the words would be approximately seven million. That would be the largest spontaneous speech corpus in the world. So far, approximately one tenth of the words have been manually detected, and morphological information such as POS category and inflection type have been assigned to them. Human annotators tagged every morpheme in the one tenth of the CSJ that has been tagged, and other annotators checked them. The human annotators discussed their disagreements and resolved them. The accuracies of the manual tagging of short and long words in the one tenth of the CSJ were greater than 99.8% and 97%, respectively. The accuracies were evaluated by random sampling. As it took over two years to tag one tenth of the CSJ accurately, tagging the remainder with morphological information would take about twenty years. Therefore, the remaining nine tenths of the CSJ must be tagged automatically or semi-automatically.

In this paper, we describe methods for detecting

the two types of word segments and corresponding morphological information. We also describe how to tag a large spontaneous speech corpus accurately. Henceforth, we call the two types of word segments *short word* and *long word* respectively, or merely *morphemes*. We use the term *morphological analysis* for the process of segmenting a given sentence into a row of morphemes and assigning to each morpheme grammatical attributes such as a POS category.

2 Problems and Their Solutions

As we mentioned in Section 1, tagging the whole of the CSJ manually would be difficult. Therefore, we are taking a semi-automatic approach. This section describes major problems in tagging a large spontaneous speech corpus with high precision in a semi-automatic way, and our solutions to those problems.

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. Two statistical approaches have been applied to this problem. One is to find 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)). Uchimoto et al. used both approaches. They proposed a morphological analysis method based on a maximum entropy (ME) model (Uchimoto et al., 2001). Their method uses a model that estimates how likely a string is to be a morpheme as its probability, and thus it has a potential to overcome the unknown word problem. Therefore, we use their method for morphological analysis of the CSJ. However, Uchimoto et al. reported that the accuracy of automatic word segmentation and POS tagging was 94 points in F-measure (Uchimoto et al., 2002). That is much lower than the accuracy obtained by manual tagging. Several problems led to this inaccuracy. In the following, we describe these problems and our solutions to them.

- Fillers and disfluencies

  Fillers and disfluencies are characteristic expressions often used in spoken language, but they are randomly inserted into text, so detecting their segmentation is difficult. In the CSJ, they are tagged manually. Therefore, we first delete fillers and disfluencies and then put them back in their original place after analyzing a text.

- Accuracy for unknown words

  The *morpheme model* that will be described in Section 3.1 can detect word segments and their POS categories even for unknown words. However, the accuracy for unknown words is lower than that for known words. One of the solutions is to use dictionaries developed for a corpus on another domain to reduce the number of unknown words, but the improvement achieved is slight (Uchimoto et al., 2002). We believe that the reason for this is that definitions of a word segment and its POS category depend on a particular corpus, and the definitions from corpus to corpus differ word by word. Therefore, we need to put only words extracted from the same corpus into a dictionary. We are manually examining words that are detected by the morpheme model but that are not found in a dictionary. We are also manually examining those words that the morpheme model estimated as having low probability. During the process of manual examination, if we find words that are not found in a dictionary, those words are then put into a dictionary. Section 4.2.1 will describe the accuracy of detecting unknown words and show how much those words contribute to improving the morphological analysis accuracy when they are detected and put into a dictionary.

- Insufficiency of features

  The model currently used for morphological analysis considers the information of a target morpheme and that of an adjacent morpheme on the left. To improve the model, we need to consider the information of two or more morphemes on the left of the target morpheme. However, too much information often leads to overtraining the model. Using all the information makes training the model difficult when there is too much of it. Therefore, the best way to improve the accuracy of the morphological information in the CSJ within the limited

time available to us is to examine and revise the errors of automatic morphological analysis and to improve the model. We assume that the smaller the probability estimated by a model for an output morpheme is, then the greater the likelihood is that the output morpheme is wrong. Therefore, we examine output morphemes in ascending order of their probabilities. The expected improvement of the accuracy of the morphological information in the whole of the CSJ will be described in Section 4.2.1

Another problem concerning unknown words is that the cost of manual examination is high when there are several definitions for word segments and their POS categories. Since there are two types of word definitions in the CSJ, the cost would double. Therefore, to reduce the cost, we propose another method for detecting word segments and their POS categories. The method will be described in Section 3.2, and the advantages of the method will be described in Section 4.2.2

The next problem described here is one that we have to solve to make a language model for automatic speech recognition.

- Pronunciation

  Pronunciation of each word is indispensable for making a language model for automatic speech recognition. In the CSJ, pronunciation is transcribed separately from the basic form written by using *kanji* and *hiragana* characters as shown in Fig. 1. Text targeted for morphological analysis is the basic form of the CSJ and it does not have information on actual pronunciation. The result of morphological analysis, therefore, is a row of morphemes that do not have information on actual pronunciation. To estimate actual pronunciation by using only the basic form and a dictionary is impossible. Therefore, actual pronunciation is assigned to results of morphological analysis by aligning the basic form and pronunciation in the CSJ. First, the results of morphological analysis, namely, the morphemes, are transliterated into *katakana* characters by using a dictionary, and then they are aligned with pronunciation in the CSJ by using a dynamic programming method.

Basic form	Pronunciation
0017 00051.425-00052.869 L:	
(Fえー)	(Fエー)
形態素解析	ケータイソカイセキ
0018 00053.073-00054.503 L:	
について	ニツイテ
0019 00054.707-00056.341 L:	
お話しいたします	オハナシイタシマス

"Well, I'm going to talk about morphological analysis."

Figure 1: Example of transcription.

In this paper, we will mainly discuss methods for detecting word segments and their POS categories in the whole of the CSJ.

3 Models and Algorithms

This section describes two methods for detecting word segments and their POS categories. The first method uses morpheme models and is used to detect any type of word segment. The second method uses a chunking model and is only used to detect long word segments.

3.1 Morpheme Model

Given a tokenized test corpus, namely a set of strings, 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 it is a morpheme. When a string is a morpheme, a grammatical attribute is assigned to it. A tag designated as a 1 is thus assigned one of a number, n, of grammatical attributes assigned to morphemes, and the problem becomes to assign an attribute (from 0 to n) to every string in a given sentence.

We define a model that estimates the likelihood that a given string is a morpheme and has a grammatical attribute $i(1 \leq i \leq n)$ as a *morpheme model*. We implemented this model within an ME modeling framework (Jaynes, 1957; Jaynes, 1979; Berger et al., 1996). The model is represented by Eq. (1):

$$p_\lambda(a|b) = \frac{\exp\left(\sum_{i,j} \lambda_{i,j} g_{i,j}(a,b)\right)}{Z_\lambda(b)} \quad (1)$$

Short word						Long word					
Word		Pronunciation		POS	Others	Word		Pronunciation		POS	Others
形態	(form)	ケータイ	(*keitai*)	Noun		形態素解析	(morphological analysis)	ケータイソカイセキ	(*keitaisokaiseki*)	Noun	
素	(element)	ソ	(*so*)	Suffix							
解析	(analysis)	カイセキ	(*kaiseki*)	Noun							
に		ニ	(*ni*)	PPP	case marker	について	(about)	ニツイテ	(*nitsuite*)	PPP	case marker, compound word
つい	(relate)	ツイ	(*tsui*)	Verb	KA-GYO, ADF, euphonic change						
て		テ	(*te*)	PPP	conjunctive						
お		オ	(*o*)	Prefix		お話しいたし	(talk)	オハナシシタシ	(*ohanashiitasi*)	Verb	SA-GYO, ADF
話し	(talk)	ハナシ	(*hanashi*)	Verb	SA-GYO, ADF						
いたし	(do)	イタシ	(*itashi*)	Verb	SA-GYO, ADF						
ます		マス	(*masu*)	AUX	ending form	ます		マス	(*masu*)	AUX	ending form

PPP : post-positional particle , AUX : auxiliary verb , ADF : adverbial form

Figure 2: Example of morphological analysis results.

$$Z_\lambda(b) = \sum_a \exp\left(\sum_{i,j} \lambda_{i,j} g_{i,j}(a,b)\right), \quad (2)$$

where a is one of the categories for classification, and it can be one of $(n+1)$ tags from 0 to n (This is called a "future."), b is the contextual or conditioning information that enables us to make a decision among the space of futures (This is called a "history."), and $Z_\lambda(b)$ is a normalizing constant determined by the requirement that $\sum_a p_\lambda(a|b) = 1$ for all b. The computation of $p_\lambda(a|b)$ in any ME model is dependent on a set of "features" which are binary functions of the history and future. For instance, one of our features is

$$g_{i,j}(a,b) = \begin{cases} 1 : \text{if } has(b, f_j) = 1 \ \& \ a = a_i \\ \qquad f_j = \text{"POS}(-1)\text{(Major)} : \text{verb,"} \\ 0 : \text{otherwise.} \end{cases} \quad (3)$$

Here "$has(b, f_j)$" is a binary function that returns 1 if the history b has feature f_j. The features used in our experiments are described in detail in Section 4.1.1.

Given a sentence, probabilities of n tags from 1 to n are estimated for each length of string in that sentence by using the morpheme model. From all possible division of morphemes in the sentence, an optimal one is found by using the Viterbi algorithm. Each division is represented as a particular division of morphemes with grammatical attributes in a sentence, and the optimal division is defined as a division that maximizes the product of the probabilities estimated for each morpheme in the division. For example, the sentence "形態素解析についてお話いたします" in basic form as shown in Fig. 1 is analyzed as shown in Fig. 2. "形態素解析" is analyzed as three morphemes, "形態 (noun)", "素 (suffix)", and "解析 (noun)", for short words, and as one morpheme, "形態素解析 (noun)" for long words.

In conventional models (e.g., (Mori and Nagao, 1996; Nagata, 1999)), probabilities were estimated for candidate morphemes that were found in a dictionary or a corpus and for the remaining strings obtained by eliminating the candidate morphemes from a given sentence. Therefore, unknown words were apt to be either concatenated as one word or divided into both a combination of known words and a single word that consisted of more than one character. However, this model has the potential to correctly detect any length of unknown words.

3.2 Chunking Model

The model described in this section can be applied when several types of words are defined in a corpus and one type of words consists of compounds of other types of words. In the CSJ, every long word consists of one or more short words.

Our method uses two models, a morpheme model for short words and a chunking model for long words. After detecting short word segments and their POS categories by using the former model, long word segments and their POS categories are detected by using the latter model. We define four labels, as explained below, and extract long word segments by estimating the appropriate labels for each short word according to an ME model. The four labels are listed below:

Ba: Beginning of a long word, and the POS category of the long word agrees with the short word.

Ia: Middle or end of a long word, and the POS category of the long word agrees with the short word.

B: Beginning of a long word, and the POS category of the long word does not agree with the short word.

I: Middle or end of a long word, and the POS category of the long word does not agree with the short word.

A label assigned to the leftmost constituent of a long word is "Ba" or "B". Labels assigned to other constituents of a long word are "Ia", or "I". For example, the short words shown in Fig. 2 are labeled as shown in Fig. 3. The labeling is done deterministically from the beginning of a given sentence to its end. The label that has the highest probability as estimated by an ME model is assigned to each short word. The model is represented by Eq. (1). In Eq. (1), a can be one of four labels. The features used in our experiments are described in Section 4.1.2.

Short word			Long word	
Word	POS	Label	Word	POS
形態	Noun	Ba	形態素解析	Noun
素	Suffix	I		
解析	Noun	Ia		
に	PPP	Ba	について	PPP
つい	Verb	I		
て	PPP	Ia		
お	Prefix	B	お話しいたし	Verb
話し	Verb	Ia		
いたし	Verb	Ia		
ます	AUX	Ba	ます	AUX

PPP : post-positional particle , AUX : auxiliary verb

Figure 3: Example of labeling.

When a long word that does not include a short word that has been assigned the label "Ba" or "Ia", this indicates that the word's POS category differs from all of the short words that constitute the long word. Such a word must be estimated individually. In this case, we estimate the POS category by using transformation rules. The transformation rules are automatically acquired from the training corpus by extracting long words with constituents, namely short words, that are labeled only "B" or "I". A rule is constructed by using the extracted long word and the adjacent short words on its left and right. For example, the rule shown in Fig. 4 was acquired in our experiments. The middle division of the consequent part represents a long word "てみ" (auxiliary verb), and it consists of two short words "て" (post-positional particle) and "み" (verb). If several different rules have the same antecedent part, only the rule with the highest frequency is chosen. If no rules can be applied to a long word segment, rules are generalized in the following steps.

1. Delete posterior context
2. Delete anterior and posterior contexts
3. Delete anterior and posterior contexts and lexical entries.

If no rules can be applied to a long word segment in any step, the POS category noun is assigned to the long word.

4 Experiments and Discussion

4.1 Experimental Conditions

In our experiments, we used 744,204 short words and 618,538 long words for training, and 63,037 short words and 51,796 long words for testing. Those words were extracted from one tenth of the CSJ that already had been manually tagged. The training corpus consisted of 319 speeches and the test corpus consisted of 19 speeches.

Transcription consisted of basic form and pronunciation, as shown in Fig. 1. Speech sounds were faithfully transcribed as pronunciation, and also represented as basic forms by using *kanji* and *hiragana* characters. Lines beginning with numerical digits are time stamps and represent the time it took to produce the lines between that time stamp and the next time stamp. Each line other than time stamps represents a *bunsetsu*. In our experiments, we used only the basic forms. Basic forms were tagged with several types of labels such as fillers, as shown in Table 1. Strings tagged with those labels were handled according to rules as shown in the rightmost columns in Table 1.

Since there are no boundaries between sentences in the corpus, we selected the places in the CSJ that

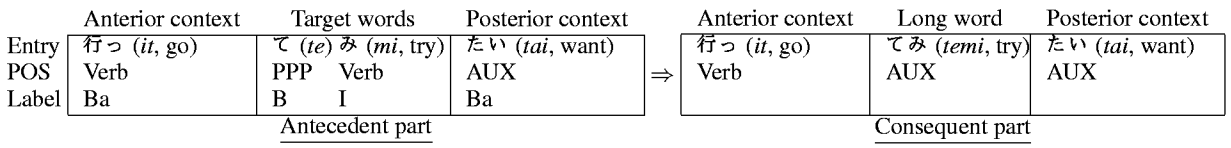

Figure 4: Example of transformation rules.

Table 1: Type of labels and their handling.

Type of Labels	Example	Rules
Fillers	(F あの)	delete all
Disfluencies	(D こ) これ、 これ (D2 は) が	delete all
No confidence in transcription	(? タオンダー)	leave a candidate
Entirely	(?)	delete all
Several candidates exist	(? あのー, あんのー)	leave the former candidate
Citation on sound or words	(M わ) は (M は) と表記	leave a candidate
Foreign, archaic, or dialect words	(O ザッツファイン)	leave a candidate
Personal name, discriminating words, and slander	○○研の (R △△) さんが	leave a candidate
Letters and their pronunciation in *katakana* strings	(A イーユー;EU)	leave the former candidate
Strings that cannot be written in *kanji* characters	(K い (F んー) ずみ; 炭)	leave the latter candidate

are automatically detected as pauses of 500 ms or longer and then designated them as sentence boundaries. In addition to these, we also used utterance boundaries as sentence boundaries. These are automatically detected at places where short pauses (shorter than 200 ms but longer than 50 ms) follow the typical sentence-ending forms of predicates such as verbs, adjectives, and copula.

4.1.1 Features Used by Morpheme Models

In the CSJ, *bunsetsu* boundaries, which are phrase boundaries in Japanese, were manually detected. Fillers and disfluencies were marked with the labels (F) and (D). In the experiments, we eliminated fillers and disfluencies but we did use their positional information as features. We also used as features, *bunsetsu* boundaries and the labels (M), (O), (R), and (A), which were assigned to particular morphemes such as personal names and foreign words. Thus, the input sentences for training and testing were character strings without fillers and disfluencies, and both boundary information and various labels were attached to them. Given a sentence, for every string within a *bunsetsu* and every string appearing in a dictionary, the probabilities of a in Eq. (1) were estimated by using the morpheme model. The output was a sequence of morphemes with grammatical attributes, as shown in Fig. 2. We used the POS categories in the CSJ as grammatical attributes. We obtained 14 major POS categories for short words and 15 major POS categories for long words. Therefore, a in Eq. (1) can be one of 15 tags from 0 to 14 for short words, and it can be one of 16 tags from 0 to 15 for long words.

Table 2: Features.

Number	Feature Type	Feature value (Number of value) (Short:Long)
1	String(0)	(113,474:117,002)
2	String(-1)	(17,064:32,037)
3	Substring(0)(Left1)	(2,351:2,375)
4	Substring(0)(Right1)	(2,148:2,171)
5	Substring(0)(Left2)	(30,684:31,456)
6	Substring(0)(Right2)	(25,442:25,541)
7	Substring(-1)(Left1)	(2,160:2,088)
8	Substring(-1)(Right1)	(1,820:1,675)
9	Substring(-1)(Left2)	(11,025:12,875)
10	Substring(-1)(Right2)	(10,439:13,364)
11	Dic(0)(Major)	Noun, Verb, Adjective, . . . Undefined (15:16)
12	Dic(0)(Minor)	Common_noun, Topic_marker, Basic_form. . . (75:71)
13	Dic(0)(Major&Minor)	Noun&Common_noun, Verb&Basic_form, . . . (246:227)
14	Dic(-1)(Minor)	Common_noun, Topic_marker, Basic_form. . . (16:16)
15	POS(-1)	Noun, Verb, Adjective, . . . (14:15)
16	Length(0)	1, 2, 3, 4, 5, 6_or_more (6:6)
17	Length(-1)	1, 2, 3, 4, 5, 6_or_more (6:6)
18	TOC(0)(Beginning)	Kanji, Hiragana, Number, Katakana, Alphabet (5:5)
19	TOC(0)(End)	Kanji, Hiragana, Number, Katakana, Alphabet (5:5)
20	TOC(0)(Transition)	Kanji→Hiragana, Number→Kanji, Katakana→Kanji, . . . (25:25)
21	TOC(-1)(End)	Kanji, Hiragana, Number, Katakana, Alphabet (5:5)
22	TOC(-1)(Transition)	Kanji→Hiragana, Number→Kanji, Katakana→Kanji, . . . (16:15)
23	Boundary	Bunsetsu(Beginning), Bunsetsu(End), Label(Beginning), Label(End), (4:4)
24	Comb(1,15)	(74,602:59,140)
25	Comb(1,2,15)	(141,976:136,334)
26	Comb(1,13,15)	(78,821:61,813)
27	Comb(1,2,13,15)	(156,187:141,442)
28	Comb(11,15)	(209:230)
29	Comb(12,15)	(733:682)
30	Comb(13,15)	(1,549:1,397)
31	Comb(12,14)	(730:675)

The features we used with morpheme models in

our experiments are listed in Table 2. Each feature consists of a type and a value, which are given in the rows of the table, and it corresponds to j in the function $g_{i,j}(a,b)$ in Eq. (1). The notations "(0)" and "(-1)" used in the feature-type column in Table 2 respectively indicate a target string and the morpheme to the left of it. The terms used in the table are basically as same as those that Uchimoto et al. used (Uchimoto et al., 2002). The main difference is the following one:

Boundary: Bunsetsu boundaries and positional information of labels such as fillers. "(Beginning)" and "(End)" in Table 2 respectively indicate whether the left and right side of the target strings are boundaries.

We used only those features that were found three or more times in the training corpus.

4.1.2 Features Used by a Chunking Model

We used the following information as features on the target word: a word and its POS category, and the same information for the four closest words, the two on the left and the two on the right of the target word. Bigram and trigram words that included a target word plus bigram and trigram POS categories that included the target word's POS category were used as features. In addition, bunsetsu boundaries as described in Section 4.1.1 were used. For example, when a target word was "に" in Fig. 3, "素", "解析", "に", "つい", "て", "Suffix", "Noun", "PPP", "Verb", "PPP", "解析 & に", "に & つい", "素 & 解析 & に", "に & つい & て", "Noun&PPP", "PPP&Verb", "Suffix&Noun&PPP", "PPP&Verb&PPP", and "Bunsetsu(Beginning)" were used as features.

4.2 Results and Discussion

4.2.1 Experiments Using Morpheme Models

Results of the morphological analysis obtained by using morpheme models are shown in Table 3 and 4. In these tables, OOV indicates Out-of-Vocabulary rates. Shown in Table 3, OOV was calculated as the proportion of words not found in a dictionary to all words in the test corpus. In Table 4, OOV was calculated as the proportion of word and POS category pairs that were not found in a dictionary to all pairs in the test corpus. *Recall* is the percentage of morphemes in the test corpus for which the segmentation and major POS category were identified correctly. *Precision* is the percentage of all morphemes identified by the system that were identified correctly. The *F-measure* is defined by the following equation.

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

Table 3: Accuracies of word segmentation.

Word	Recall	Precision	F	OOV
Short	97.47% ($\frac{61,444}{63,037}$)	97.62% ($\frac{61,444}{62,945}$)	97.54	1.66%
	99.23% ($\frac{62,553}{63,037}$)	99.11% ($\frac{62,553}{63,114}$)	99.17	0%
Long	96.72% ($\frac{50,095}{51,796}$)	95.70% ($\frac{50,095}{52,346}$)	96.21	5.81%
	99.05% ($\frac{51,306}{51,796}$)	98.58% ($\frac{51,306}{52,047}$)	98.81	0%

Table 4: Accuracies of word segmentation and POS tagging.

Word	Recall	Precision	F	OOV
Short	95.72% ($\frac{60,341}{63,037}$)	95.86% ($\frac{60,341}{62,945}$)	95.79	2.64%
	97.57% ($\frac{61,505}{63,037}$)	97.45% ($\frac{61,505}{63,114}$)	97.51	0%
Long	94.71% ($\frac{49,058}{51,796}$)	93.72% ($\frac{49,058}{52,346}$)	94.21	6.93%
	97.30% ($\frac{50,396}{51,796}$)	96.83% ($\frac{50,396}{52,047}$)	97.06	0%

Tables 3 and 4 show that accuracies would improve significantly if no words were unknown. This indicates that all morphemes of the CSJ could be analyzed accurately if there were no unknown words. The improvements that we can expect by detecting unknown words and putting them into dictionaries are about 1.5 in F-measure for detecting word segments of short words and 2.5 for long words. For detecting the word segments and their POS categories, for short words we expect an improvement of about 2 in F-measure and for long words 3.

Next, we discuss accuracies obtained when unknown words existed. The OOV for long words was 4% higher than that for short words. In general, the higher the OOV is, the more difficult detecting word segments and their POS categories is. However, the difference between accuracies for short and long words was about 1% in recall and 2% in precision, which is not significant when we consider that the difference between OOVs for short and long words was 4%. This result indicates that our morpheme models could detect both known and unknown words accurately, especially

long words. Therefore, we investigated the recall of unknown words in the test corpus, and found that 55.7% (928/1,667) of short word segments and 74.1% (2,660/3,590) of long word segments were detected correctly. In addition, regarding unknown words, we also found that 47.5% (791/1,667) of short word segments plus their POS categories and 67.3% (2,415/3,590) of long word segments plus their POS categories were detected correctly. The recall of unknown words was about 20% higher for long words than for short words. We believe that this result mainly depended on the difference between short words and long words in terms of the definitions of compound words. A compound word is defined as one word when it is based on the definition of long words; however it is defined as two or more words when it is based on the definition of short words. Furthermore, based on the definition of short words, a division of compound words depends on its context. More information is needed to precisely detect short words than is required for long words. Next, we extracted words that were detected by the morpheme model but were not found in a dictionary, and investigated the percentage of unknown words that were completely or partially matched to the extracted words by their context. This percentage was 77.6% (1,293/1,667) for short words, and 80.6% (2,892/3,590) for long words. Most of the remaining unknown words that could not be detected by this method are compound words. We expect that these compounds can be detected during the manual examination of those words for which the morpheme model estimated a low probability, as will be shown later.

The recall of unknown words was lower than that of known words, and the accuracy of automatic morphological analysis was lower than that of manual morphological analysis. As previously stated, to improve the accuracy of the whole corpus we take a semi-automatic approach. We assume that the smaller the probability is for an output morpheme estimated by a model, the more likely the output morpheme is wrong, and we examine output morphemes in ascending order of their probabilities. We investigated how much the accuracy of the whole corpus would increase. Fig. 5 shows the relationship between the percentage of output morphemes whose probabilities exceed a threshold and their

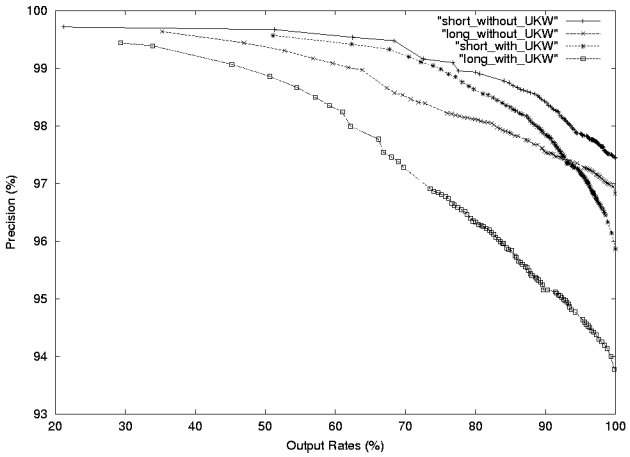

Figure 5: Partial analysis.

precision. In this figure, "short_without_UKW", "long_without_UKW」", "short_with_UKW", and "long_with_UKW" represent the precision for short words detected assuming there were no unknown words, precision for long words detected assuming there were no unknown words, precision of short words including unknown words, and precision of long words including unknown words, respectively. When the output rate in the horizontal axis increases, the number of low-probability morphemes increases. In all graphs, precisions monotonously decrease as output rates increase. This means that tagging errors can be revised effectively when morphemes are examined in ascending order of their probabilities.

Next, we investigated the relationship between the percentage of morphemes examined manually and the precision obtained after detected errors were revised. The result is shown in Fig. 6. Precision represents the precision of word segmentation and POS tagging. If unknown words were detected and put into a dictionary by the method described in the fourth paragraph of this section, the graph line for short words would be drawn between the graph lines "short_without_UKW" and "short_with_UKW", and the graph line for long words would be drawn between the graph lines "long_without_UKW" and "long_with_UKW". Based on test results, we can expect better than 99% precision for short words and better than 97% precision for long words in the whole corpus when we examine 10% of output mor-

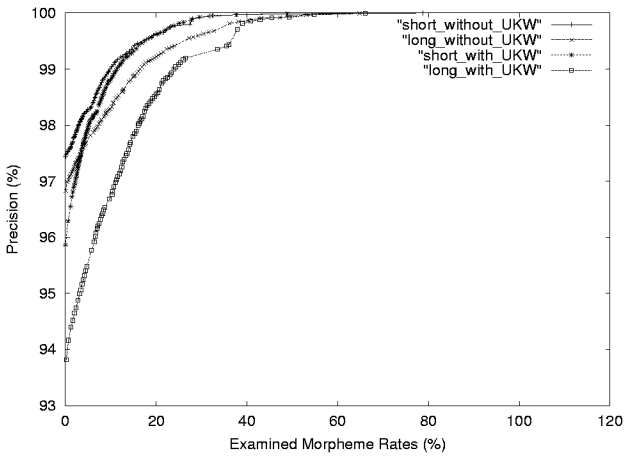

Figure 6: Relationship between the percentage of morphemes examined manually and precision obtained after revising detected errors (when morphemes with probabilities under threshold and their adjacent morphemes are examined).

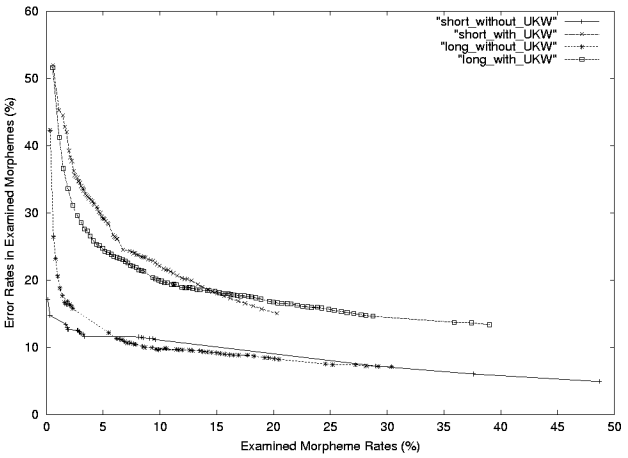

Figure 7: Relationship between percentage of morphemes examined manually and error rate of examined morphemes.

phemes in ascending order of their probabilities.

Finally, we investigated the relationship between percentage of morphemes examined manually and the error rate for all of the examined morphemes. The result is shown in Fig. 7. We found that about 50% of examined morphemes would be found as errors at the beginning of the examination and about 20% of examined morphemes would be found as errors when examination of 10% of the whole corpus was completed. When unknown words were detected and put into a dictionary, the error rate decreased; even so, over 10% of examined morphemes would be found as errors.

4.2.2 Experiments Using Chunking Models

Results of the morphological analysis of long words obtained by using a chunking model are shown in Table 5 and 6. The first and second lines

Table 5: Accuracies of long word segmentation.

Model	Recall	Precision	F
Morph	96.72% ($\frac{50,095}{51,796}$)	95.70% ($\frac{50,095}{52,346}$)	96.21
Chunk	97.65% ($\frac{50,580}{51,796}$)	97.41% ($\frac{50,580}{51,911}$)	97.54
Chunk	98.84% ($\frac{51,193}{51,796}$)	98.66% ($\frac{51,193}{51,888}$)	98.75

Table 6: Accuracies of long word segmentation and POS tagging.

Model	Recall	Precision	F
Morph	94.71% ($\frac{49,058}{51,796}$)	93.72% ($\frac{49,058}{52,346}$)	94.21
Chunk	95.59% ($\frac{49,513}{51,796}$)	95.38% ($\frac{49,513}{51,911}$)	95.49
Chunk	98.56% ($\frac{51,051}{51,796}$)	98.39% ($\frac{51,051}{51,888}$)	98.47
Chunk w/o TR	92.61% ($\frac{47,968}{51,796}$)	92.40% ($\frac{47,968}{51,911}$)	92.51

TR : transformation rules

show the respective accuracies obtained when OOVs were 5.81% and 6.93%. The third lines show the accuracies obtained when we assumed that the OOV for short words was 0% and there were no errors in detecting short word segments and their POS categories. The fourth line in Table 6 shows the accuracy obtained when a chunking model without transformation rules was used.

The accuracy obtained by using the chunking model was one point higher in F-measure than that obtained by using the morpheme model, and it was very close to the accuracy achieved for short words. This result indicates that errors newly produced by applying a chunking model to the results obtained for short words were slight, or errors in the results

obtained for short words were amended by applying the chunking model. This result also shows that we can achieve good accuracy for long words by applying a chunking model even if we do not detect unknown long words and do not put them into a dictionary. If we could improve the accuracy for short words, the accuracy for long words would be improved also. The third lines in Tables 5 and 6 show that the accuracy would improve to over 98 points in F-measure. The fourth line in Tables 6 shows that transformation rules significantly contributed to improving the accuracy.

Considering the results obtained in this section and in Section 4.2.1, we are now detecting short and long word segments and their POS categories in the whole corpus by using the following steps:

1. Automatically detect and manually examine unknown words for short words.
2. Improve the accuracy for short words in the whole corpus by manually examining short words in ascending order of their probabilities estimated by a morpheme model.
3. Apply a chunking model to the short words to detect long word segments and their POS categories.

As future work, we are planning to use an active learning method such as that proposed by Argamon-Engelson and Dagan (Argamon-Engelson and Dagan, 1999) to more effectively improve the accuracy of the whole corpus.

5 Conclusion

This paper described two methods for detecting word segments and their POS categories in a Japanese spontaneous speech corpus, and describes how to tag a large spontaneous speech corpus accurately by using the two methods. The first method is used to detect any type of word segments. We found that about 80% of unknown words could be semi-automatically detected by using this method. The second method is used when there are several definitions for word segments and their POS categories, and when one type of word segments includes another type of word segments. We found that better accuracy could be achieved by using both methods than by using only the first method alone.

Two types of word segments, short words and long words, are found in a large spontaneous speech corpus, CSJ. We found that the accuracy of automatic morphological analysis for the short words was 95.79 in F-measure and for long words, 95.49. Although the OOV for long words was much higher than that for short words, almost the same accuracy was achieved for both types of words by using our proposed methods. We also found that we can expect more than 99% of precision for short words, and 97% for long words found in the whole corpus when we examined 10% of output morphemes in ascending order of their probabilities as estimated by the proposed models.

In our experiments, only the information contained in the corpus was used; however, more appropriate linguistic knowledge than that could be used, such as morphemic and syntactic rules. We would like to investigate whether such linguistic knowledge contributes to improved accuracy.

References

S. Argamon-Engelson and I. Dagan. 1999. Committee-Based Sample Selection For Probabilistic Classifiers. *Artificial Intelligence Research*, 11:335–360.

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

E. T. Jaynes. 1957. Information Theory and Statistical Mechanics. *Physical Review*, 106:620–630.

E. T. Jaynes. 1979. Where do we Stand on Maximum Entropy? In R. D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, page 15. M. I. T. Press.

H. Kashioka, S. G. Eubank, and E. W. Black. 1997. Decision-Tree Morphological Analysis Without a Dictionary for Japanese. In *Proceedings of NLPRS*, pages 541–544.

K. Maekawa, H. Koiso, S. Furui, and H. Isahara. 2000. Spontaneous Speech Corpus of Japanese. In *Proceedings of LREC*, pages 947–952.

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

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

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

K. Uchimoto, C. Nobata, A. Yamada, S. Sekine, and H. Isahara. 2002. Morphological Analysis of The Spontaneous Speech Corpus. In *Proceedings of COLING*, pages 1298–1302.