MORPHOLOGICAL ANALYSIS OF THE CORPUS OF SPONTANEOUS JAPANESE

Kiyotaka Uchimoto[†], *Chikashi Nobata*[†], *Atsushi Yamada*[†], *Satoshi Sekine*[‡], *and Hitoshi Isahara*[†]

[†]Communications Research Laboratory
2-2-2, 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 a method for accurately tagging a large spontaneous speech corpus. In this paper, we show that by using semi-automatic analysis we can expect a precision of over 99% for detecting and tagging short words and 97% for long words; the two types of words comprising the corpus.

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)* [1]. 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. 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. The accuracies of the manual detection and tagging of short and long words in one tenth of the CSJ are greater than 99.8% and 97%, respectively. 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 POS category.

2. MODELS AND ALGORITHMS

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

2.1. Morpheme Model

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 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 defined a model that estimates the likelihood that a given string is a morpheme and has the grammatical attribute $i(1 \leq i \leq n)$ as a *morpheme model* [2]. We implemented this model within an maximum entropy (ME) framework [3]. The features used in our experiments are described in Section 3.1.

Given a sentence, for each length of string in the sentence, probabilities of n tags from 1 to n are estimated by using the morpheme model. Among every possible division of morphemes in the sentence, an optimal one is found by using the

Viterbi algorithm. The optimal division is defined as a particular division of morphemes with grammatical attributes that maximize the product of the probabilities estimated for each morpheme in a division of morphemes in a sentence. 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.

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."

Fig. 1. Example of transcription.

Short word			Long word		
Word		POS	Word		POS
形態	(form)	Noun	形態素解析	(morphological analysis)	Noun
素	(element)	Suffix			
解析	(analysis)	Noun			
に		PPP	について	(about)	PPP
つい	(relate)	Verb			
て		PPP			
お		Prefix	お話しいたし	(talk)	Verb
話し	(talk)	Verb			
いたし	(do)	Verb			
ます		AUX	ます		AUX

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

Fig. 2. Example of morphological analysis results.

2.2. Chunking Model

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". The short words shown in Fig. 2, for example, 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 features used in our experiments are described in Section 3.1.

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

Fig. 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.

3. EXPERIMENTS AND DISCUSSION

3.1. 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 338 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*, which is a Japanese phrasal unit. In our experiments, we used only the basic forms.

	Anterior context	Target words	Posterior context		Anterior context	Long word	Posterior context
Entry	行っ(*it*, go)	て (*te*) み (*mi*, try)	たい (*tai*, want)		行っ(*it*, go)	てみ (*temi*, try)	たい (*tai*, want)
POS	Verb	PPP Verb	AUX	⇒	Verb	AUX	AUX
Label	Ba	B I	Ba				
		Antecedent part				Consequent part	

Fig. 4. Example of transformation rules.

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

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 various labels and boundary information were attached to them. Given a sentence, for every string within a *bunsetsu* and every string appearing in a dictionary, the probabilities 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.

The features we used with morpheme models in our experiments are basically as same as those that Uchimoto et al. used [4]. The main difference was in boundary information that indicated whether the left and right side of the target strings were boundaries. Bunsetsu boundaries and positional information of labels such as fillers were used as features. We used only those features that were found three or more times in the training corpus.

In our experiments using the chunking model, we used the following information as features on the target word: a word and the POS category to which it belonged, and the same information on 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 were used.

3.2. Results and Discussion

3.2.1. Experiments Using Morpheme Models

Results of the morphological analysis obtained by using morpheme models are shown in Table 1 and 2. In these tables, OOV indicates Out-of-Vocabulary rates. In Table 2, 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.

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

Table 1. 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 2. 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 1 and 2 show that accuracies would improve significantly if there were no unknown words. Especially, the accuracy for long words was close to that in the current corpus. This indicates that all morphemes of the CSJ could be analyzed accurately if there were no unknown 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 with their context. This was 77.6% (1,293/1,667) for short words, and 80.6% (2,892/3,590) for long words.

The accuracy of automatic morphological analysis was lower than that of manual morphological analysis. To improve the accuracy for 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 the relationship between the percentage of morphemes examined manually and the precision obtained after detected errors were revised. The result is shown in Fig. 5. 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, the precision for long

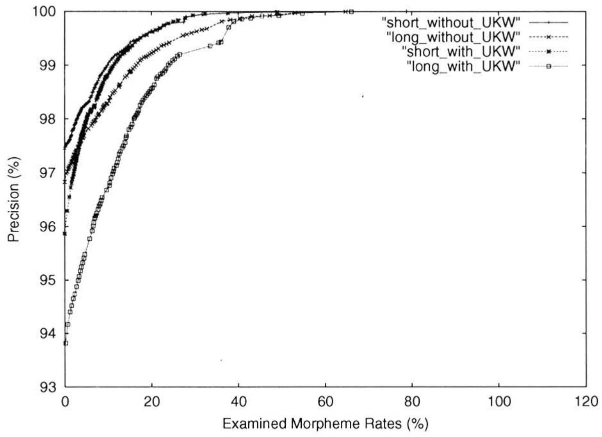

Fig. 5. Relationship between the percentage of morphemes examined manually and the precision obtained after detected errors were revised (when not only morphemes whose probabilities are under a threshold but also their adjacent morphemes are examined).

words detected assuming there were no unknown words, precision of short words including unknown words, and precision of long words including unknown words, respectively. 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 third paragraph in this section, the graph for short words would be drawn between the graphs "short_without_UKW" and "short_with_ UKW", and the graph for long words would be drawn between the graphs "long_without_UKW" and "long_with_UKW". Based on test results, we can expect over 99% precision for short words and over 97% precision for long words in the whole corpus when we examine 10% of output morphemes in ascending order of their probabilities.

3.2.2. Experiments Using Chunking Models

Results of the morphological analysis of long words obtained by using a chunking model are shown in Tables 3 and 4. The first and second lines show the accuracies obtained when OOVs are 5.81% and 6.93%, respectively. The third lines show the accuracies obtained when we assumed that the OOV for short words was 0% and there was no error for detecting short word segments and their POS categories.

Table 3. 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	97.55% ($\frac{50,527}{51,796}$)	97.41% ($\frac{50,527}{51,873}$)	97.48
Chunk	98.74% ($\frac{51,145}{51,796}$)	98.63% ($\frac{51,145}{51,855}$)	98.69

Table 4. 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	95.52% ($\frac{49,475}{51,796}$)	95.38% ($\frac{49,475}{52,346}$)	95.45
Chunk	98.50% ($\frac{51,019}{51,796}$)	98.39% ($\frac{51,019}{51,855}$)	98.44

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 words for long words and do not put them into a dictionary, though we must do so when we apply a morpheme model to long words. If we could improve the accuracy for short words, the accuracy for long words would improve to over 98 points in F-measure.

4. CONCLUSION

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

There are two types of word segments, short words and long words, in a large spontaneous speech corpus, CSJ. We found that the accuracies of automatic morphological analysis for the short and long words were 95.79 and 95.45, respectively, in F-measure. Although the OOV for long words was much higher than that for short words, almost the same accuracies was achieved for both words by using our proposed methods. We also found that we can expect over 99% and 97% of precision, respectively, for the two types of words in the whole corpus when we examine 10% of output morphemes in ascending order of their probabilities as estimated by the proposed model.

5. REFERENCES

[1] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous Speech Corpus of Japanese," in *Proceedings of LREC*, 2000, pp. 947–952.

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

[3] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[4] K. Uchimoto, C. Nobata, A. Yamada, S. Sekine, and H. Isahara, "Morphological Analysis of The Spontaneous Speech Corpus," in *Proceedings of COLING*, 2002, pp. 1298–1302.