Morphological Analysis of The Spontaneous Speech Corpus

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 a project tagging a spontaneous speech corpus with morphological information such as word segmentation and parts-of-speech. We use a morphological analysis system based on a maximum entropy model, which is independent of the domain of corpora. In this paper we show the tagging accuracy achieved by using the model and discuss problems in tagging the spontaneous speech corpus. We also show that a dictionary developed for a corpus on a certain domain is helpful for improving accuracy in analyzing a corpus on another domain.

1 Introduction

In recent years, systems developed for analyzing written-language texts have become considerably accurate. This accuracy is largely due to the large amounts of tagged corpora and the rapid progress in the study of corpus-based natural language processing. However, the accuracy of the systems developed for written language is not always high when these same systems are used to analyze spoken-language texts. The reason for this remaining inaccuracy is due to several differences between the two types of languages. For example, the expressions used in written language are often quite different from those in spoken language, and sentence boundaries are frequently ambiguous in spoken language. The "Spontaneous Speech: Corpus and Processing Technology" project was implemented in 1999 to overcome this problem. Spoken language includes both monologue and dialogue texts; the former (e.g. the text of a talk) was selected as a target of the project because it was considered to be appropriate to the current level of study on spoken language.

Tagging the spontaneous speech corpus with morphological information such as word segmentation and parts-of-speech is one of the goals of the project. The tagged corpus is helpful for us in making a language model in speech recognition as well as for linguists investigating distribution of morphemes in spontaneous speech. For tagging the corpus with morphological information, a morphological analysis system is needed. 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 of segmenting a given sentence into a row of morphemes and assigning to each morpheme grammatical attributes such as part-of-speech (POS) and 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. 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 (M.E.) model (Uchimoto et al., 2001). We used their method to tag a spontaneous speech corpus. Their 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. This model is independent of the domain of corpora; in this paper we demonstrate that this is true by applying our model to the spontaneous speech corpus, *Corpus of Spontaneous Japanese (CSJ)* (Maekawa et al., 2000). We also show that a dictionary developed for a corpus on a certain domain is helpful for improving accuracy in analyzing a corpus on another domain.

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 "futures" in the M.E. formulation of our problem of morphological analysis. The M.E. model 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. The computation of $P(f|h)$ in any M.E. model 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

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

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 what part-of-speech the adjacent morpheme is.

Given a set of features and some training data, the M.E. estimation process produces a model, which is represented as follows (Berger et al., 1996; Ristad, 1997; Ristad, 1998):

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

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

We define a model which 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*. This model is represented by Eq. (2), in which f can be one of $(n+1)$ tags from 0 to n.

Given a sentence, it 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.

3 Experiments and Discussion

3.1 Experimental Conditions

We used the spontaneous speech corpus, CSJ, which is a tagged corpus of transcriptions of academic presentations and simulated public speech. Simulated public speech is short speech spoken specifically for the corpus by paid non-professional speakers. For training, we used 805,954 morphemes from the corpus, and for testing, we used 68,315 morphemes from the corpus. Since there are no boundaries between sentences in the corpus, we used two types of boundaries, utterance boundaries, which are automatically detected at the place where a pause of 200 ms or longer emerges in the CSJ, and sentence boundaries assigned by the sentence boundary identification system, which is based on hand-crafted rules which use the pauses as a clue. In the CSJ, fillers and disfluencies are marked with tags (F) and (D). In the experiments, we did not use those tags. Thus the input sentences for testing are character strings without any tags. The output is a sequence of morphemes with grammatical attributes. As the grammatical attributes, we define the part-of-speech categories in the CSJ. There are 12 major categories. Therefore, the number of grammatical attributes is 12, and f in Eq. (2) can be one of 13 tags from 0 to 12.

Given a sentence, for every string consisting of five or fewer characters and every string appearing in a dictionary, 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. We collected all morphemes from the training corpus except disfluencies and used them as dictionary entries. We denote the entries with a *Corpus* dictionary. The maximum length for a morpheme was set at five because morphemes consisting of six or more characters are mostly compound words or words consisting of *katakana* characters. We assumed that compound words that do not appear in the dictionary can be divided into strings consisting of five or fewer characters because compound words tend not to appear in dictionaries. *Katakana* strings that are not found in the 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. The assigned part-of-speech in the optimal set is selected from all the categories of the M.E. model except the one in which the string is not a morpheme.

The features used in our experiments are listed in Table 1. Each feature consists of a type and a value, which are given in the rows of the table. The features are basically some attributes of the morpheme itself or attributes of the morpheme to the left of it. We used the features 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 to the left of it.

The terms used in the table are as follows:

String: Strings appearing as a morpheme three or more times in the training corpus

Substring: Characters used in a string. "(Left1)" and "(Right1)" respectively represent the leftmost and rightmost characters of a string. "(Left2)" and "(Right2)" respectively represent the leftmost and rightmost character bigrams of a string.

Dic: Entries in the Corpus dictionary. As minor categories we used inflection types such as a basic form as well as minor part-of-speech categories. "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.

Length: Length of a string

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 character in a string. "TOC(-1)(Transition)" represents the transition from the rightmost character in the adjacent morpheme on the left to the leftmost character 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.

POS: Part-of-speech.

3.2 Results and Discussion

Results of the morphological analysis obtained by our method are shown 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. The *F-measure* is defined by the following equation.

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

This result shows that there is no significant difference between accuracies obtained by using two types of sentence boundaries. However, we found that the errors that occurred around utterance boundaries were reduced in the result obtained with sentence boundaries assigned by the sentence boundary identification system. This shows that there is a high possibility that we can achieve better accuracy if we use boundaries assigned by the sentence boundary identification system as sentence boundaries and if we use utterance boundaries as features.

In these experiments, we used only the entries with a Corpus dictionary. Next we show the experimental results with dictionaries developed for a corpus on a certain domain. We added to the Corpus dictionary all the approximately 200,000 entries of the JUMAN dictionary (Kurohashi and Nagao, 1999). We also added the entries of a dictionary developed by ATR. We call it the ATR dictionary.

Results obtained with each dictionary or each combination of dictionaries are shown in Table 3. In this table, OOV indicates Out-of-Vocabulary rates. The accuracy obtained with the JUMAN dictionary or the ATR dictionary was worse than the accuracy obtained without those dictionaries. This is because the segmen-

Table 1: Features.

Feature number	Feature type	Feature value (Number of value)
1	String(0)	(223,457)
2	String(-1)	(20,769)
3	Substring(0)(Left1)	(2,492)
4	Substring(0)(Right1)	(2,489)
5	Substring(0)(Left2)	(74,046)
6	Substring(0)(Right2)	(73,616)
7	Substring(-1)(Left1)	(2,237)
8	Substring(-1)(Right1)	(2,489)
9	Substring(-1)(Left2)	(12,726)
10	Substring(-1)(Right2)	(12,241)
11	Dic(0)(Major)	Noun, Verb, Adj, ... Undefined (13)
12	Dic(0)(Minor)	Common_noun, Topic_marker, Basic_form... (223)
13	Dic(0)(Major&Minor)	Noun&Common_noun, Verb&Basic_form, ... (239)
14	Length(0)	1, 2, 3, 4, 5, 6_or_more (6)
15	Length(-1)	1, 2, 3, 4, 5, 6_or_more (6)
16	TOC(0)(Beginning)	Kanji, Hiragana, Number, Katakana, Alphabet (5)
17	TOC(0)(End)	Kanji, Hiragana, Number, Katakana, Alphabet (5)
18	TOC(0)(Transition)	Kanji→Hiragana, Number→Kanji, Katakana→Kanji, ... (25)
19	TOC(-1)(End)	Kanji, Hiragana, Number, Katakana, Alphabet (5)
20	TOC(-1)(Transition)	Kanji→Hiragana, Number→Kanji, Katakana→Kanji, ... (18)
21	POS(-1)	Verb, Adj, Noun, ... (12)
22	Comb(1,21)	Combinations Feature 1 and 21 (142,546)
23	Comb(1,2,21)	Combinations Feature 1, 2 and 21 (216,431)
24	Comb(1,13,21)	Combinations Feature 1, 13 and 21 (29,876)
25	Comb(1,2,13,21)	Combinations Feature 1, 2, 13 and 21 (158,211)
26	Comb(11,21)	Combinations Feature 11 and 21 (156)
27	Comb(12,21)	Combinations Feature 12 and 21 (1,366)
28	Comb(13,21)	Combinations Feature 13 and 21 (1,518)

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

Boundary	Recall	Precision	F-measure
utterance	93.97% (64,198/68,315)	93.25% (64,198/68,847)	93.61
sentence	93.97% (64,195/68,315)	93.18% (64,195/68,895)	93.57

tation of morphemes and the definition of part-of-speech categories in the JUMAN and ATR dictionaries are different from those in the CSJ.

Given a sentence, for every string consisting of five or fewer characters as well as every string appearing in a dictionary, whether or not the string is a morpheme was determined by our morpheme model. However, we speculate that we can ignore strings consisting of two or more characters when they are not found in the dictionary when OOV is low. Therefore, we carried out the additional experiments ignoring those strings. In the experiments, given a sentence, for every string consisting of one character and every string appearing in a dictionary, whether or not the string is a morpheme is determined by our morpheme model. Results obtained under this condition are shown in Table 4. We compared the accuracies obtained with dictionaries including the Corpus dictionary, whose OOVs are relatively low. The accuracies obtained with the additional dictionaries increased while those obtained only with the Corpus dictionary decreased. These results show that a dictionary whose OOV in the test corpus is low contributes to increasing the accuracy when ignoring the possibility that strings that consist of two or more characters and are not found in the dictionary become a morpheme.

These results show that a dictionary developed for a corpus on a certain domain can be used to improve accuracy in analyzing a corpus on another domain.

The accuracy in segmentation and major POS tagging obtained for spontaneous speech was worse than the approximately 95% obtained for newspaper articles. We think the main reason for this is the errors and the inconsistency of the corpus, and the difficulty in recognizing characteristic expressions often used in spoken language such as fillers, mispronounced words, and disfluencies. The inconsistency of the corpus is due to the way the corpus was made, i.e., completely by human beings, and it is also due

Table 3: Results of Experiments (Segmentation and major POS tagging).

Dictionary	Boundary	Recall	Precision	F	OOV
Corpus	utterance	92.64% (63,288/68,315)	91.83% (63,288/68,917)	**92.24**	1.84%
Corpus	sentence	92.61% (63,265/68,315)	91.79% (63,265/68,923)	92.20	1.84%
JUMAN	utterance	90.28% (61,676/68,315)	90.07% (61,676/68,478)	90.17	6.13%
JUMAN	sentence	90.33% (61,710/68,315)	90.22% (61,710/68,403)	90.27	6.13%
ATR	utterance	89.80% (61,348/68,315)	90.12% (61,348/68,073)	89.96	8.14%
ATR	sentence	89.96% (61,453/68,315)	90.30% (61,453/68,057)	90.13	8.14%
Corpus+JUMAN	utterance	92.03% (62,872/68,315)	91.77% (62,872/68,507)	91.90	0.52%
Corpus+JUMAN	sentence	92.09% (62,913/68,315)	91.80% (62,913/68,534)	91.95	0.52%
Corpus+ATR	utterance	92.35% (63,086/68,315)	92.03% (63,086/68,547)	92.19	0.64%
Corpus+ATR	sentence	92.30% (63,057/68,315)	91.94% (63,057/68,585)	92.12	0.64%
JUMAN+ATR	utterance	91.60% (62,579/68,315)	91.57% (62,579/68,339)	91.59	4.61%
JUMAN+ATR	sentence	91.66% (62,618/68,315)	91.67% (62,618/68,311)	91.66	4.61%
Corpus+JUMAN+ATR	utterance	91.72% (62,658/68,315)	91.66% (62,658/68,357)	91.69	0.47%
Corpus+JUMAN+ATR	sentence	91.72% (62,657/68,315)	91.62% (62,657/68,391)	91.67	0.47%

∗ For training 1/5 of all the training corpus (163,796 morphemes) was used.

Table 4: Results of Experiments (Segmentation and major POS tagging).

Dictionary	Boundary	Recall	Precision	F	OOV
Corpus	utterance	92.80% (63,395/68,315)	90.47% (63,395/70,075)	91.62	1.84%
Corpus	sentence	92.71% (63,333/68,315)	90.48% (63,333/70,000)	91.58	1.84%
Corpus+JUMAN	utterance	92.45% (63,154/68,315)	91.60% (63,154/68,942)	92.02	0.52%
Corpus+JUMAN	sentence	92.48% (63,179/68,315)	91.71% (63,179/68,893)	92.09	0.52%
Corpus+ATR	utterance	92.91% (63,474/68,315)	91.81% (63,474/69,137)	**92.36**	0.64%
Corpus+ATR	sentence	92.75% (63,361/68,315)	91.76% (63,361/69,053)	92.25	0.64%
Corpus+JUMAN+ATR	utterance	92.30% (63,055/68,315)	91.57% (63,055/68,858)	91.94	0.47%
Corpus+JUMAN+ATR	sentence	92.28% (63,039/68,315)	91.55% (63,039/68,860)	91.91	0.47%

∗ For training 1/5 of all the training corpus (163,796 morphemes) was used.

to the definition of morphemes. Several inconsistencies in the test corpus existed, such as: “ 東京 (*tokyo*, Noun)(Tokyo), 都 (*to*, Other)(the Metropolis), 立 (*ritsu*, Other)(founded), 大学 (*daigaku*, Noun)(university),” and “都立 (*toritsu*, Noun)(metropolitan), 大学 (*daigaku*, Noun)(university).” Both of these are the names representing the same university. The “都立” is partitioned into two in the first one while it is not partitioned into two in the second one according to the definition of morphemes. When such inconsistencies in the corpus exist, it is difficult for our model to discriminate among these inconsistencies because we used only bigram information as features. To achieve better accuracy, therefore, we need to use trigram or longer information. To correctly recognize characteristic expressions often used in spoken language, we plan to extract typical patterns used in the expressions, to generalize the patterns manually, and to generate possible expressions using the generalized patterns, and finally, to add such patterns to the dictionary. We also plan to expand our model to skip fillers, mispronounced words, and disfluencies because those expressions are randomly inserted into text and it is impossible to learn the connectivity between those randomly inserted expressions and others.

References

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.

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

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

K. Maekawa, H. Koiso, S. Furui, and H. Isahara. 2000. Spontaneous Speech Corpus of Japanese. In *Proceedings of the 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 the 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 the ACL*, pages 277–284.

E. S. Ristad. 1997. Maximum Entropy Modeling for Natural Language. ACL/EACL Tutorial Program, Madrid.

E. S. Ristad. 1998. Maximum Entropy Modeling Toolkit, Release 1.6 beta. http://www.mnemonic.com/software/memt.

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 the EMNLP*, pages 91–99.