Morphological Annotation of a Large Spontaneous Speech Corpus in Japanese

Kiyotaka Uchimoto and **Hitoshi Isahara**
National Institute of Information and Communications Technology
3-5, Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{uchimoto,isahara}@nict.go.jp

Abstract

We propose an efficient framework for human-aided morphological annotation of a large spontaneous speech corpus such as the *Corpus of Spontaneous Japanese*. In this framework, even when word units have several definitions in a given corpus, and not all words are found in a dictionary or in a training corpus, we can morphologically analyze the given corpus with high accuracy and low labor costs by detecting words not found in the dictionary and putting them into it. We can further reduce labor costs by expanding training corpora based on active learning.

1 Introduction

The "Spontaneous Speech: Corpus and Processing Technology" project sponsored the construction of a large spontaneous Japanese speech corpus, the *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. The 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 words* and *long words*. The term *short word* approximates an item found in an ordinary Japanese dictionary, and *long word* represents various compounds. The length and morphological information of each are different, and every short word is included in a long word, which is shorter than a Japanese phrasal unit, a *bunsetsu*. For example, the short and long words in "*keitaisokaiseki ni tsuite ohanasi itashimasu*" (I would like to talk about morphological analysis) are represented as shown in Table 1. In this table, each line indicates a short word, and ten short words and four long words are shown.

Approximately 7.5 million short words were detected in the CSJ, which makes it the largest spontaneous speech corpus in the world. On the other hand, there were fewer long words because each long word consists of one or more short words. Approximately one-eighth of the words have been manually detected, and morphological information such as part-of-speech (POS) categories and conjugation types have been assigned to them. Human annotators tagged every morpheme in the one-eighth of the CSJ that had 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-eighth of the CSJ were both approximately 99.9%. The accuracies were evaluated by random sampling. Because it took over two years to tag one-eighth of the CSJ accurately, tagging the remainder with morphological information would take about twenty years. Therefore, the remaining seven-eighths of the CSJ were tagged 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 accurately tag a large spontaneous speech corpus. We propose an efficient framework for human-aided morphological annotation of a large spontaneous speech corpus.

We collectively call short and long words *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 morphological attributes such as a POS category.

2 Framework of Morphological Annotation

We call a series of processes for morphological analysis and maintenance of a corpus **morphological annotation**. Our framework of morphological annotation is illustrated in Figure 1. The purpose of this framework is, given an annotated corpus and a large raw corpus, to improve the quality of the annotated morphological information in both corpora with low labor costs. In this framework, a corpus-based morphological analyzer is used because the definition of a set of part-of-speech categories and that of word units are often changed in the middle of constructing a corpus, and a morphological analyzer must be robust to such changes.

The framework of morphological annotation consists of three parts: maintenance of an annotated corpus, analysis of a large raw corpus, and enhancement of linguistic resources. They are described in more detail in the following sections.

Table 1: Example of morphological analysis results.

Short word									Label	Long word						
OT	DicForm	Lemma		PT	POS	ConjType	ConjForm	Other		OT	DicForm	Lemma	POS	ConjType	ConjForm	Other
keitai	*keitai*	*keitai*	(form)	*kêtai*	Noun				Ba	*keitaiso-*	*keitaiso-*	*keitaisokaiseki*	Noun			
so	*so*	*so*	(element)	*so*	Suffix				I	*kaiseki*	*kaiseki*	(morphological				
kaiseki	*kaiseki*	*kaiseki*	(analysis)	*kaiseki*	Noun				Ia			analysis)				
ni	*ni*	*ni*		*ni*	PPP			case marker	B	*nitsuite*	*nitsuite*	*nitsuite*	PPP			case marker
tsui	*tsuku*	*tsuku*	(relate)	*tsui*	Verb	KA-GYO	ADF	euphonic change	I			(about)				& compound
te	*te*	*te*		*te*	PPP			conjunctive	I							word
o	*o*	*o*		*o*	Prefix				B	*ohanashi-*	*ohanashi-*	*ohanashiitasu*	Verb	SA-GYO	ADF	
hanashi	*hanasu*	*hanasu*	(talk)	*hanashi*	Verb	SA-GYO	ADF		Ia	*itashi*	*itasu*	(talk)				
itashi	*itasu*	*itasu*		*itashi*	Verb	SA-GYO	ADF		Ia							
masu	*masu*	*masu*		*masu*	AUX		ending form		Ba	*masu*	*masu*	*masu*	AUX		ending form	

OT: orthographic transcription, DicForm: dictionary form written in *kanji* and *hiragana* characters, Lemma: lemma written in *kanji* and *hiragana* characters, POS: part-of-speech, PT: phonetic transcription written in *katakana* characters, ConjType: conjugation type, ConjForm: conjugation form. PPP: post-positional particle, AUX: auxiliary verb, ADF: adverbial form.

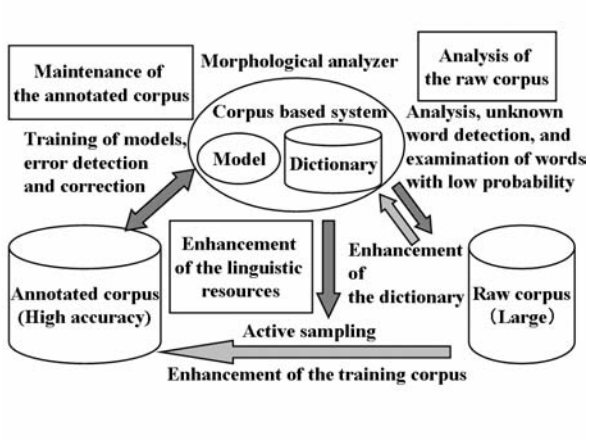

Figure 1: Framework of morphological annotation.

2.1 Maintenance of an Annotated Corpus

In general, when an annotated corpus or a training corpus has many errors, a corpus-based analyzer easily overfits the errors, and the analysis accuracy often decreases. Therefore, the errors should be detected and corrected to avoid overfitting them. The simplest way to detect and correct the errors in a training corpus is to examine the difference between the original annotated corpus and the corpus automatically annotated by a corpus-based analyzer trained by using the original annotated corpus. In previous studies, a boosting method and an anomaly detection method were applied to detect errors in a corpus [Abney *et al.*, 1999; Eskin, 2000], and a method for detecting and correcting errors in a corpus was proposed [Murata *et al.*, 2002]. In the method for detecting and correcting errors, the difference between the annotated labels of the original corpus and those of the automatically annotated corpus is first detected. For each difference, the probabilities of the labels in the original corpus and the automatically annotated corpus are calculated by a model trained using the original annotated corpus. Then, the labels in the original corpus whose probabilities are lower than those in the automatically annotated corpus are replaced with the corresponding labels in the automatically annotated corpus. This method was initially applied to the CSJ, and the morphological information replaced by the method was manually examined.

2.2 Analysis of a Large Raw Corpus and Enhancement of Linguistic Resources

Enhancement of the Dictionary

When a given raw corpus including unknown words that are in neither the dictionary nor the training corpus, detection errors tend to occur on the unknown words and words to the left and right of the unknown words. In many cases, not a single unknown word but a series of unknown words are found, or an unknown word may consist of known words and unknown characters. In this case, the accuracy of morphological information assigned to a raw corpus can be increased by detecting and putting the unknown words into the dictionary and manually examining words whose probability is estimated as low [Uchimoto *et al.*, 2003].

The cost of manual examination is high when word segments and their POS categories have several definitions. Because the CSJ has two types of word definition, the cost would double. However, when one type of word segment consists of compounds of other types of words, the manual examination of the shortest word segments improves the morphological analysis accuracy for other types of words if we use a method based on a chunking model [Uchimoto *et al.*, 2003].

Besides the above methods, in the morphological analysis of the CSJ, the number of unknown words was further reduced by expanding conjugational words in a dictionary based on a conjugation chart developed by the members of the National Institute for Japanese Language.

Active Sampling

In general, a corpus-based morphological analysis system requires a large training corpus. However, the accuracy does not improve in proportion to the simple increase of the training corpus. This is because a model for morphological analysis usually considers the relationship between adjacent words, and the simply supplemented data rarely includes relationships that were not found in the initial training corpus. Therefore, when expanding a training corpus, data must be selected that include as many sequences of words as possible

that are difficult for the morphological analysis model to analyze. The expansion should be done so that a big improvement can be achieved with as small a supplement as possible. Argamon-Engelson et al. reported that data that can usefully be added to training data can be selected by extracting sentences whose analysis results obtained by using randomly selected models include a great amount of consistency [Argamon-Engelson and Dagan, 1999]. However, they assumed that word boundaries were given and that the data had no unknown words. Their method cannot be simply applied to Japanese sentences when word boundaries are inconsistent because no blank spaces are used between words in Japanese sentences. In our research, we assumed that word boundaries were not given and that the data had unknown words. This section describes an active sampling method using a single statistical model for expanding a training corpus.

The active sampling is conducted as follows, under three assumptions. Under these assumptions, a set of sentences that minimize $\prod p$ is extracted, and words in the extracted sentences whose probabilities are below a threshold are examined. Here, $\prod p$ is the product of the probabilities estimated by the morpheme model for all words in a set of sentences. The three assumptions are as follows.

1. Similar errors tend to appear in the same speech.

   This is because specific words or sequences of words may only appear in a certain speech. Therefore, the data should be compiled from as varied speeches as possible to avoid examining the same erroneous words. That is why $x\%$ of words in each speech were examined. The maximum number of examined words was chosen to be $\frac{n \times x}{100}$ when the number of words in a speech was n.
2. The sentence will have more errors if the product of probabilities estimated for its words are low.

   Therefore, we preferred sentences with low probabilities.
3. Any word whose probability is over a certain threshold can be considered correct.

   Therefore, words whose probabilities are over a threshold are ignored when $\prod p$ mentioned above is calculated. The threshold should be set rather high to reduce the chance that errors will be ignored. In our preliminary experiments, the threshold was set at 95% because the accuracy obtained using this threshold was 99% or higher.

2.3 Models and Algorithms for Morphological Analysis

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 the potential to overcome the unknown word problem. Therefore, we used their method to morphologically analyze the CSJ.

In this paper, we assume that two types of word segments, short and long words, are defined and every long word consists of one or more short words, as in the CSJ.

Method Based on a Morpheme Model

Given a tokenized test corpus, namely, a set of strings, the problem of Japanese morphological analysis can be reduced to that 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 of 1 is thus assigned one of a number, n, of grammatical attributes assigned to morphemes, and the problem becomes assigning an attribute (from 0 to n) to every string in a given sentence.

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

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

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

where a (called a "future") is one of the categories for classification, and it can be one of $(n+1)$ tags from 0 to n; b (called a "history") is the contextual or conditioning information that enables us to make a decision among the space of futures; 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 Section 3.

Given a sentence, the 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 divisions of morphemes in the sentence, an optimal one is found by using the Viterbi algorithm or a branch and bound method. 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.

In the CSJ, transcriptions consist of basic forms and pronunciations. The pronunciation is transcribed separately from the basic form written in *kanji* and *hiragana* characters.

Speech sounds are faithfully transcribed in *katakana* characters as the pronunciation and represented as basic forms in *kanji* and *hiragana* characters. The text we targeted for morphological analysis is the transcription in the CSJ. When all possible divisions of morphemes in a sentence are obtained, they are matched with the pronunciation part in each line representing a *bunsetsu* by using a dynamic programming method. Then, morphemes whose phonetic transcription candidates do not match the aligned pronunciation part are eliminated before searching for the optimal division of the morphemes.

Method Based on the Chunking Model and Transformation Rules

Long word segments and their POS information are detected by using a method described below after detecting short word segments and their POS information by using a morpheme model.

Given the two types of word segments, the longer of which consists of compounds of the shorter, the problem of detecting long word segments and their POS information can be reduced to the problem of assigning one of four labels, as explained below, to each short word. We call the model that estimates the likelihood of the four labels a **chunking model**. We implemented this model within ME or support vector machine (SVM) based modeling frameworks. The four labels are as follows.

Ba: The beginning of a long word, and the POS information of the long word agrees with that of the short word.

Ia: The middle or end of a long word, and the POS information of the long word agrees with that of the short word.

B: The beginning of a long word, and the POS information of the long word does not agree with that of the short word.

I: The middle or end of a long word, and the POS information of the long word does not agree with that of the short word.

The label assigned to the leftmost constituent of a long word is "Ba" or "B". The labels assigned to the other constituents of long words are "Ia" and "I". A short word that "Ba" or "Ia" is assigned to has the same POS information as its corresponding long word. Here, the POS information represents a set of a POS category, conjugation type, conjugation form, and other detailed information on POS, as shown in Table 1. For example, in the table, the labels (in the "label" column) are assigned to the short words. If these labels are correctly detected, the POS information of the long words can be obtained from the short words to which "Ba" or "Ia" is assigned. In the example, the POS information can be detected for all the long words except "*nitsuite*". On the other hand, only the long word segment information can be assigned to "*nitsuite*" even if the assigned labels for its constituents are correct because it has additional POS information as a compound word that is different from the POS information of its constituents, "*ni*", "*tsui*", and "*te*". In this case, the POS information can be obtained by using the transformation rules mentioned later.

A given sentence is labeled from its beginning/end of to its end/beginning. When using an ME-based model, the optimal set of labels is obtained by finding a division that maximizes the product of the probabilities estimated for each label assigned to each short word. The model is represented by Eq. (1). In the equation, a can be one of the four labels. The optimal set of labels is found by using the Viterbi algorithm or a branch and bound method. On the other hand, SVM is a binary classifier. Therefore, we expanded it to a multi-class classifier by using multi-class methods such as a pairwise method and a one-versus-rest method. The optimal label determined using an SVM model is deterministically assigned to each short word. The features used in our experiments are described in Section 3.

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 Figure 2 was acquired from the example shown in Table 1. This rule indicates that when the labels "B", "I", and "I" are assigned to "*ni*" (post-positional particle), "*tsui*" (verb), and "*te*" (post-positional particle), respectively, the combination "*nitsuite*" is transformed into a long word having the morphological information, a post-positional particle, case marker, and compound word. 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 the posterior context
2. Delete the anterior and posterior contexts
3. Delete the anterior and posterior contexts and lexical information such as orthographic transcriptions, dictionary forms, lemmas, and phonetic transcriptions

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

The dictionary form and lemma of a long word are usually generated by concatenating those of short words. As for a spoonerism, for example, in case of a long word "*ipponbari*" consisting of three short words, "*ichi*", "*hon*", and "*hari*", information on the phonetic transcriptions of the short words is used to generate the dictionary form and the lemma of the long word when concatenating the short words.

3 Experiments and Discussion

3.1 Experimental Conditions

In our experiments, we used 868,243 short words and 721,978 long words for training and 63,039 short words and 51,699 long words for testing. Those words were extracted from the one-eighth of the CSJ that already had been manually tagged. The training corpus consisted of 377 speeches and the test corpus consisted of 19 speeches.

In our experiments, we used the basic forms and pronunciations of transcriptions as the input for morphological analysis.

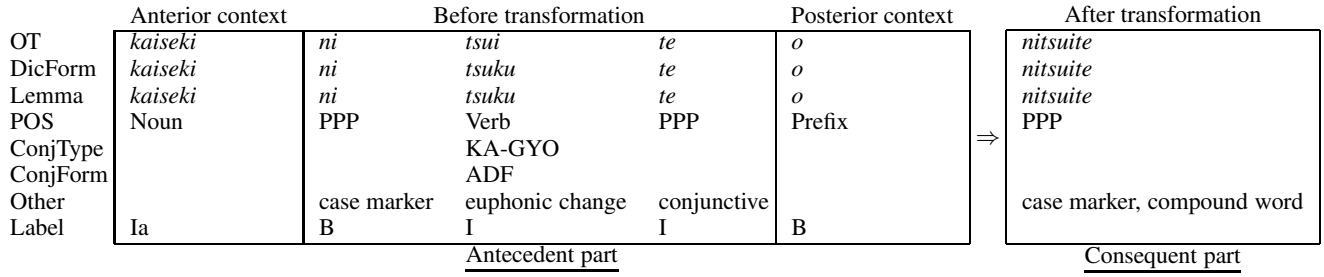

Figure 2: Example of transformation rule.

Because the sentences in the corpus do not have boundaries between them, we selected the places in the CSJ that were automatically detected as pauses of 500 ms or longer and designated them as sentence boundaries. In addition to these, we used utterance boundaries as sentence boundaries. These were automatically detected at places where short pauses (between 50 and 200 ms) follow the typical sentence-ending forms of predicates such as verbs, adjectives, and copulas.

In the CSJ, *bunsetsu* boundaries, which are the boundaries of Japanese phrasal units, were manually detected. Fillers and disfluencies were marked with the labels (F) and (D). In the experiments, we eliminated the 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. Candidate morphemes that crossed *bunsetsu* boundaries were ignored because morphemes do not cross them. The output was a sequence of morphemes with grammatical attributes, as shown in Table 1. The candidate morphemes whose POS information corresponded to that of the correct morphemes were used as positive examples, and the others were used as negative examples. We used the POS categories in the CSJ as grammatical attributes. We obtained 14 major POS categories for short and long words. Therefore, a in Eq. (1) can be one of 15 tags from 0 to 14 for short words.

Next, we describe the features used with the morpheme models in our experiments. Each feature consists of a type and a value, and it corresponds to j in the function $g_{i,j}(a, b)$ in Eq. (1). As the feature functions, we used the combinations of features and futures that appeared three times or more in the training corpus. The features used in our experiments are basically the same as those that Uchimoto et al. used[Uchimoto *et al.*, 2003]. The main differences are as follows:

Strings following to the right: One- and two-character strings directly to the right of the target word, their character types, and the combinations of them and the dictionary information on the target word were used as features.

Fine-grained categories of conjugation types and forms: Conjugation types and forms were divided into fine-grained categories according to the context.

In our experiments, SVM-based models were used to analyze long words because better results were obtained using SVM-based models than ME-based models in our preliminary experiments. A SVM-based chunker, YamCha[Kudo and Matsumoto, 2001], was used to assign the four chunking labels.

We selected the following parameters for YamCha based on our preliminary experiments.

- Degree of polynomial kernel: 3rd
- Analysis direction: Right to left
- Dynamic features: Two preceding chunk labels
- Multi-class method: One-versus-rest

We used the following information as features of target words:

- Morphological information: orthographic transcription, dictionary form, lemma, phonetic transcription, POS category, conjugation type, conjugation form, and other detailed POS information
- Boundary information: *bunsetsu* and various labels such as "filler".
- The same information as for the target word for the four closest words, the two on the left and the two on the right of the target word.

Morphological information for approximately 10% of the long words was generated by applying transformation rules.

3.2 Results and Discussion

The results of the morphological analysis obtained by using morpheme models are shown in Table 2. In the table, OOV indicates the out-of-vocabulary rates, which were calculated as the proportion of word and morphological information pairs that were not found in a dictionary to all pairs in the test corpus. The morphological information included orthographic transcriptions, dictionary forms, lemmas, POS categories, conjugation types, conjugation forms, and other detailed POS information. *Recall* is the percentage of morphemes in the test corpus for which the segmentation and all morphological information were identified correctly. *Precision* is the percentage of all morphemes identified by the system that were identified correctly. The *F-measure* is defined as the harmonic mean of recall and precision.

Table 2, except the bottom line, shows the results obtained when the output of a short word analysis was used as the input

Table 2: Accuracies of morphological analysis.

Word	Recall	Precision	F	OOV
Short	96.14% ($\frac{60,603}{63,039}$)	94.76% ($\frac{60,603}{63,957}$)	95.44	2.34%
	98.23% ($\frac{61,922}{63,039}$)	98.19% ($\frac{61,922}{63,101}$)	98.18	0%
Long	95.05% ($\frac{49,140}{51,699}$)	94.73% ($\frac{49,140}{51,876}$)	94.89	6.54%
	96.51% ($\frac{49,895}{51,699}$)	96.40% ($\frac{49,895}{51,756}$)	96.46	0%
Long	99.10% ($\frac{51,233}{51,699}$)	98.96% ($\frac{51,233}{51,773}$)	99.03	0%

of a long word analysis. The bottom line shows the results obtained when the long words were analyzed after correcting the errors of the output of the short word analysis. This indicates that all morphemes of the CSJ could be analyzed accurately if no unknown words were in the data. The accuracies obtained after detecting unknown words, examining them, and putting them into a dictionary is shown in Table 3. In the initial state, 1475 words were unknown, and these words were of 837 types. The unknown words were detected, examined, and registered in the dictionary through the following steps.

1. Examine the short words that are not found in the dictionary, put them into the dictionary, and reanalyze the corpus using the amended dictionary.

   After conducting this step twice, all the words obtained in the short word analysis were found in the dictionary, and 458 words were registered in the dictionary. The first and second lines in Table 3 show the accuracies of a short word analysis done after the first registration in this step, and the third and fourth lines show the accuracies after the second registration.

2. Examine the short words whose probabilities are within a window size, put them into the dictionary, and reanalyze the corpus using the amended dictionary.

   This step was conducted twice. The window size was set from 0 to 0.001 in the first examination and from 0.001 to 0.01 in the second examination. A total of 638 words were registered in the dictionary, which is approximately 76% of the initial number of unknown words. The fifth and sixth lines in Table 3 show the accuracies of short word analysis done after the first registrations in this step, and seventh and eighth lines show the accuracies after the second registration.

In the above steps, a total of 1478 short words were examined, which is close to the number of unknown words in the initial state. These results show that the small number of examinations in this experiment can dramatically reduce the number of unknown words and improve the accuracy to close to that obtained without unknown words.

Table 4 shows the results obtained by applying the active sampling mentioned in Section 2.2. Here, we assumed all words were known [1]. In this table, the examined word rate is the x described in Section 2.2, and the extracted word rate

[1]The accuracies obtained using a dictionary having approximately 76% of the initial unknown words had a similar tendency to those shown in Table 4, although they were approximately 0.3 less than those in F-measure.

Table 3: Accuracies obtained after unknown words were registered in dictionary.

Word	Recall	Precision	F	NOE (Ratio)
Short	97.15% ($\frac{61,242}{63,039}$)	96.41% ($\frac{61,242}{63,525}$)	96.78	772 (1.2%)
Long	95.92% ($\frac{49,592}{51,699}$)	95.75% ($\frac{49,592}{51,792}$)	95.84	
Short	97.62% ($\frac{61,540}{63,039}$)	96.83% ($\frac{61,540}{63,554}$)	97.22	942 (1.5%)
Long	96.26% ($\frac{49,768}{51,699}$)	96.04% ($\frac{49,768}{51,818}$)	96.15	
Short	97.93% ($\frac{61,735}{63,039}$)	97.47% ($\frac{61,735}{63,336}$)	97.70	1205 (1.9%)
Long	96.46% ($\frac{49,867}{51,699}$)	96.32% ($\frac{49,867}{51,773}$)	96.39	
Short	98.07% ($\frac{61,821}{63,039}$)	97.73% ($\frac{61,821}{63,255}$)	97.90	1478 (2.3%)
Long	96.54% ($\frac{49,909}{51,699}$)	96.41% ($\frac{49,909}{51,765}$)	96.48	

NOE: # of examination.

Table 4: Accuracies obtained after active sampling.

Word	Recall	Precision	F	Ext (Exm)
Short	98.27% ($\frac{61,950}{63,039}$)	98.19% ($\frac{61,950}{63,093}$)	98.23	2.3% (1%)
Long	96.54% ($\frac{49,912}{51,699}$)	96.44% ($\frac{49,912}{51,755}$)	96.49	
Short	98.58% ($\frac{62,146}{63,039}$)	98.53% ($\frac{62,146}{63,073}$)	98.56	13.6% (5%)
Long	96.78% ($\frac{50,035}{51,699}$)	96.69% ($\frac{50,035}{51,750}$)	96.73	
Short	98.85% ($\frac{62,311}{63,039}$)	98.81% ($\frac{62,311}{63,059}$)	98.83	30.1% (10%)
Long	96.90% ($\frac{50,098}{51,699}$)	96.80% ($\frac{50,098}{51,752}$)	96.85	

Ext: extracted word rate, Exm: examined word rate.

(Ext) is the percentage of words in the sentences extracted so that the percentage of examined words was $x\%$. For example, this table shows that the accuracies of the short and long words were approximately 99 and 97 in F-measure when about 10% were examined. Here, we assumed that all errors in the extracted short words were corrected. Actually, the percentage of words that needed correction was close to that of the examined words. In the CSJ, the examined word rate (Exm) was about 1%. Therefore, the final accuracies of the short and long words are expected to be approximately 98 and 96 in F-measure.

4 Conclusion

We proposed an efficient framework for human-aided morphological annotation of a large spontaneous speech corpus such as the *Corpus of Spontaneous Japanese* (CSJ). We demonstrated that within this framework, even when word units in a given corpus have several definitions and not all words are found in a dictionary or in a training corpus, the corpus can be morphologically analyzed with high accuracy and low labor costs by detecting words not found in the dictionary and putting them into it. We also demonstrated that labor costs can be further reduced by expanding the training corpus based on active learning. We applied this framework to the CSJ. The final accuracies of short and long words are expected to be approximately 98 and 96 in F-measure.

Acknowledgments

The authors would like to thank Prof. Sadaoki Furui, Mr. Kazuma Takaoka, Dr. Chikashi Nobata, Dr. Atsushi Yamada, Prof. Satoshi Sekine, Dr. Masaki Murata, and the members of the National Institute for Japanese Language, especially

Dr. Masaya Yamaguchi, Dr. Hideki Ogura, Mr. Ken'ya Nishikawa, Dr. Hanae Koiso, and Dr. Kikuo Maekawa, for their beneficial comments during the progress of this work. The authors also would like to thank anonymous reviewers for their helpful comments.

References

[Abney *et al.*, 1999] Steven Abney, Robert E. Schapire, and Yoram Singer. Boosting applied to tagging and PP attachment. In *In Proceedings of the EMNLP/VLC*, 1999.

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

[Berger *et al.*, 1996] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71, 1996.

[Eskin, 2000] Eleazar Eskin. Detecting errors within a corpus using anomaly detection. In *Proceedings of the 1st Meeting of the NAACL*, 2000.

[Jaynes, 1957] Edwin Thompson Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106:620–630, 1957.

[Jaynes, 1979] Edwin Thompson Jaynes. Where do we Stand on Maximum Entropy? In R. D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*. M. I. T. Press, 1979.

[Kashioka *et al.*, 1997] Hideki Kashioka, Stephen G. Eubank, and Ezra W. Black. Decision-Tree Morphological Analysis Without a Dictionary for Japanese. In *Proceedings of the NLPRS*, pages 541–544, 1997.

[Kudo and Matsumoto, 2001] Taku Kudo and Yuji Matsumoto. Chunking with Support Vector Machines. In *Proceedings of the 2nd Meeting of the NAACL*, pages 192–199, 2001.

[Maekawa *et al.*, 2000] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. Spontaneous Speech Corpus of Japanese. In *Proceedings of LREC2000*, pages 947–952, 2000.

[Mori and Nagao, 1996] Shinsuke Mori and Makoto Nagao. Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis. In *Proceedings of the 16th COLING96*, pages 1119–1122, 1996.

[Murata *et al.*, 2002] Masaki Murata, Masao Utiyama, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. Correction of Errors in a Modality Corpus Used for Machine Translation by Using Machine-learning Method. In *Proceedings of the 9th TMI*, pages 125–134, 2002.

[Nagata, 1999] Masaaki Nagata. A Part of Speech Estimation Method for Japanese Unknown Words Using a Statistical Model of Morphology and Context. In *Proceedings of the 37th ACL*, pages 277–284, 1999.

[Uchimoto *et al.*, 2001] Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. The Unknown Word Problem: a Morphological Analysis of Japanese Using Maximum Entropy Aided by a Dictionary. In *Proceedings of the 2001 Conference on EMNLP*, pages 91–99, 2001.

[Uchimoto *et al.*, 2003] Kiyotaka Uchimoto, Chikashi Nobata, Atsushi Yamada, Satoshi Sekine, and Hitoshi Isahara. Morphological Analysis of a Large Spontaneous Speech Corpus in Japanese. In *Proceedings of the 41st ACL*, pages 479–488, 2003.