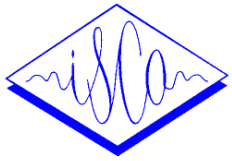

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

Corpus and Text Analysis of Spontaneous Japanese

Hitoshi Isahara

Communications Research Laboratory

ABSTRACT

There are three major parts of the "Spontaneous Speech: Corpus and Processing Technology" project; (1) compilation of large spontaneous speech corpus, (2) establishment of spoken language engineering based on the corpus, and (3) developing a prototype of a spoken language summarization system. This paper describes how we help to develop this large corpus, i.e., (1), using technology developed as a part of (2). Firstly, we discuss how to annotate whole corpus morphologically. Secondly, we explain how we annotate sentence boundaries. And thirdly we discuss discourse annotation for CSJ. This paper describes overviews of these works and details of the works described in this paper are explained in the other papers in this volume.

1. INTRODUCTION

The "Spontaneous Speech: Corpus and Processing Technology" project sponsors constructing a Japanese large spontaneous speech corpus, *Corpus of Spontaneous Japanese* (CSJ) [1]. The CSJ is a collection of monologues and dialogues, and the majority is monologue such as academic presentation speech and simulated public speech. Simulated public speech is short speech spoken specifically for the corpus by paid non-professional speakers. The total speech length will be 800 hours (roughly 7M words). One-tenth of the utterances ("Core") will be manually given orthographic and phonetic transcription.

Whole corpus will be tagged morphologically using our morphological analyzer described in section 2. We are now giving several kinds of tags to the core of the corpus, which include syntactic (dependency) tag, sentence boundary tag, summarization tag and discourse tag. Dependency tagging will be done manually. Sentence boundary tagging will be done using hand-coded rules and manual post-editing. As for summarization, we are developing three kinds of "summarization". First, we extract important sentences from talks in CSJ. 10% and 50% of the whole transcriptions will be extracted by three annotators. After the extraction, annotators are requested to paraphrase and delete some parts of the sentences so that extracted text is more natural than before. These tasks will be done within an extracted sentence, however, we also ask annotators to make 10% and 50% free summarizations. For 50% free summarization, annotators are allowed to delete some parts of the text and paraphrase the end of the sentences. For 10% free summarization, annotators can paraphrase any words or phrases in the text. Part of the core will be added discourse tags manually.

There are three major research targets of this project; (1) compilation of large spontaneous speech corpus, (2) establishment of spoken language engineering based on the corpus, and (3) developing a prototype of a spoken language summarization system. The following part of this paper describes how we help to develop this large corpus, i.e., (1), using technology developed as a part of (2). Firstly, we discuss how to annotate whole corpus morphologically. Secondly, we explain how we annotate sentence boundaries. And thirdly we discuss discourse annotation for CSJ. This paper describes overviews of these works and details of the works described in this paper are explained in the other papers in this volume [2, 3 and 4].

2. MORPHOLOGICAL ANALYSIS OF THE CORPUS OF SPONTANEOUS JAPANESE [2]

2.1 Overview

Here, two methods for detecting word segments and their morphological information in CSJ, and a method for tagging a large spontaneous speech corpus accurately will be discussed in this section.

The CSJ includes transcriptions of the speech as well as their speech sound. One of the goals of the project is to detect two types of word segments and their morphological information in the transcriptions. The two types of word segments were defined by the members of The National Institute for Japanese Language, and they are called *short word* and *long word*. Short word approximates a dictionary item of an ordinary Japanese dictionary, and long word represents various compounds. Their length and part-of-speech (POS) categories differ from each other, 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 become approximately seven million. 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 for short and long words in the one tenth of the CSJ are over 99.8% and 97%, respectively.

Our task here is to tag the remaining nine tenths of the CSJ automatically or semi-automatically, using the one tenth as a training set for the morphological tagger. So far, we can expect over 99% precision for short words and 97% precision for long word in the whole corpus by using semi-automatic analysis.

2.2 Morpheme Model and Chunking Model

We developed two methods for detecting word segments and their POS categories. The first method, which uses morpheme models, is applicable to detecting any word segments. The second method, which uses a chunking model, is applied to detecting long word segments.

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 becomes to assign an attribute (from 0 to n) to every string in a given sentence.

We define a model which estimates the likelihood that a given string is a morpheme and has the grammatical attribute i ($1 \leqq i \leqq n$) as a *morpheme model* [5]. We implemented this model within an Maximum Entropy (ME) framework.

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.

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, i.e., chunking model. We define four labels and extract long word segments by estimating the appropriate labels for each short word according to an ME model.

2.3 Semi-automatic Processing

The experiments for CSJ using these method show that accuracies would improve significantly if there were no unknown words. Especially, the accuracy for long words is 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. 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. It was 77.6% for short words, and 80.6% for long words. This means that about 80% of unknown words could be semi-automatically detected by using this method and could be stored in the dictionary.

The accuracy of automatic morphological analysis is lower than that of manual morphological analysis. To improve the accuracy of the whole corpus we take a semi-automatic approach. We assume that the smaller the probability of an output morpheme estimated by a model is, 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 are revised.

We found that we can expect respectively over 99% and 97% of precision for two types of words in the whole corpus when we examine 10% of output morphemes in ascending order of their probabilities estimated by the proposed model.

3. IDENTIFICATION OF "SENTENCE" IN SPONTANEOUS JAPANESE --- DETECTION AND MODIFICATION OF CLAUSE BOUNDARIES --- [3]

3.1 Introduction

In written language processing, a sentence is generally used as a basic unit for syntactic parsing, translation, text summarization, and so on. In spoken language processing, however, it is difficult to use a sentence as a basic unit because spoken language corpus often contains no punctuation. Therefore, it is needed to find some reasonable "sentence" in spoken language, instead of a sentence in written language, which will be useful for the automatic text summarization, parsing the dependencies between *bunsetsu*s (Japanese phrasal units), and analyzing discourse structure.

We are developing the method of semi-automatic detection of "sentence" boundaries from the CSJ. At first we extracted clause boundaries automatically as candidates,

then modified the result manually along the criteria we have defined in advance. The modification was applied to some characteristic phenomena in spoken language, like noun final clauses, shared topics, quotations, insertions, and inversions.

3.2 Automatic segmentation

Japanese is SOV language, and verb phrases are placed at the end of clauses. Clause boundaries are marked by conjugated forms of verb phrases or conjunctive particles. We can extract various types of boundaries quite precisely referring part-of-speech (POS) tags. We developed a program which segment the transcription of the CSJ into clauses automatically referring POS tags.

The program we developed to detect clause boundaries from the CSJ is in reality a set of conversion rules which find particular patterns of one to three morphemes concatenation and insert labels after the boundaries. When a particular concatenation of morphemes is accepted as an input, the program compares it with the boundary patterns prepared manually. Each morpheme is formed by four pieces of tags, including surface form, POS, conjugation form, and conjugation type. If the input matches some boundary pattern, boundary labels are inserted into the text.

3.2 Manual modification

As described above, automatic segmentation rule determines the boundary by referring only the local concatenation of morphemes. However, there are some characteristic phenomena which cannot be extracted nor treated appropriately by the local segmentation rule: noun final clause, shared topic, quotation, inserted clause, and so on. In these cases, it is required to modify the default boundaries manually referring to the recorded speech for constructing the processing unit which is syntactically well-formed as well as semantically adequate. Annotators are required to modify the result of automatic segmentation based on several prescribed criteria, which consist of definitions of phenomena and operations required.

Based on our criteria, 43 monologues, including fifteen academic presentations speech and 28 simulated public speech, were modified manually and "sentences" in the CSJ were extracted. The result of the automatic segmentation of each lecture was tagged by two or more annotators. The results of modification by each annotator were compared and put together.

Therefore the work should be supported by some annotation tools which can decrease the inevitable errors as well as the strain on annotators in the process of modification. We developed an annotation tool for the manual modification. This tool permits the annotators only the prescribed uses of symbols, and restricts the correspondences between symbols and obligatory comments showing the kinds of operation. This tool also enables easy comparison of the results of annotators and efficient data management.

4. COMMITTEE BASED DISCOURSE PURPOSE ASSIGNMENT: DISCOURSE STRUCTURE ANNOTATIONS OF SPONTANEOUS JAPANESE MONOLOGUE [4]

4.1 Introduction

In the projects of corpus construction, there are some works in which discourse structure are manually annotated. Nakatani and her colleagues [6] propose an instruction guides (Instruction for Annotating Discourse (IAD)) to annotate such corpora. Their instruction is based on the discourse structure model that proposed by the work of Grosz and Sidners [7] (hereafter GS-model). We analyzed the problems and made the extensions of applying their instruction to annotating our corpus.

4.2 Annotation Procedure

As a preprocessing of an annotation, annotating monologues are manually transcribed and divided into sentences. In transcribing process, not only words that the speaker uttered but also some inarticulate sounds (pauses, slight trip of the tongue, etc.) are transcribed into text. The sentences that we divide into in advance are minimal units of discourse [3].

In the first step of annotation, an annotator is required to listen to the whole discourse. Then, the annotator divides the discourse into segments and assigns a purpose to each segment. Finally, the annotator has to check over the annotation from the very beginning to the end.

We develop an on-line marking tool, which provides at least the following two advantages in annotations: One is that the annotator may examine any part of pair of the discourse transcription and its sound at any time. The other is that the annotator can easily know which part s/he is annotating in the overall structure because the tool displays the discourse in a tree structure.

4.3 committee based Purpose assignment

In order to identify stable boundaries of sub-stories and to assign their purpose, we call a 'committee', in which the members discuss how to segment a monologue before annotators start the deeper analysis of it. In the discussion

of the committee, some ‘guidelines’ helps the committee to assign stable boundaries of sub-stories and their purposes.

In actual annotation, the committee consists of the half of our annotators. They give the sub-stories based on the guidelines that we propose in the rest of this paper. The rest of them find cohesive patterns in each of the given sub-stories through the bottom-up.

The definition of the guidelines is based on the results of an experiment, in which three researchers of this area divided a monologue into ‘flat’ sub-stories and assigned their purposes with no hierarchical structure. It contrasts with hierarchical segmentation of IAD in GS-model and is an extended point in this work.

In the experiment, the number of sub-stories into which the monologue should be divided is assumed to be from 5 to 15 (each researchers divided each monologues into about 10 sub-stories in practice). As a result, we found that they stably assigned more than half of the boundaries of sub-stories and that the rest of them can be categorized into two types.

The above mentioned procedure will contribute to make the un-stable boundaries to be generalized. The further investigation of the collected purposes of such sub-stories will lead us to discover what the appropriate discourse purpose is really.

REFERENCES

[1] K. Maekawa, “Corpus of Spontaneous Japanese: Its Design and Evaluation”, in Proc. of SSPR-2003. 2003.

[2] K. Uchimoto et al., “Analysis of The Corpus of Spontaneous Japanese”, in Proc. of SSPR-2003. 2003.

[3] K. Takanashi et al., “Identification of “Sentence” in Spontaneous Japanese --- Detection and modification of clause boundaries ---”, in Proc. of SSPR-2003. 2003.

[4] K. Takeuchi et al., “Committee based Discourse Purpose Assignment: Discourse Structure Annotations of Spontaneous Japanese Monologue”, in Proc. of SSPR-2003. 2003.

[5] K. Uchimoto et al., “The Unknown Word Problem; a Morphological Analysis of Japanese Using Maximum Entropy Aided by a dictionary”, in Proc. of EMNLP 2001, pp.91-99.

[6] C.H Nakatani et al. *Instructions for annotating discourse*, Technical Report 21-95, Center for Research in Computing Technology, 1995.

[7] B.J. Grosz and C.L. Sidner, “Attention, Intention, and the Structure of Discourse”. *Computational Linguistics, 12(3),* 175-204, 1986.