Creation and Application of Dialogue Corpora - Characteristics of High Occurrence Frequency N-Grams of Spoken Japanese Corpora -

Shuichi ITAHASHI, Mikio YAMAMOTO, Kazuyuki TAKAGI and Naoko UEDA
Institute of Information Sciences & Electronics, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305 JAPAN
E-mail: itahashi@milab.is.tsukuba.ac.jp

1. Introduction

For years, many researchers have pointed out that spoken dialogue corpus is indispensable for promoting spoken dialogue research. However, much less discussion has been made on how to design it or what to choose when several corpora are available.

It is very difficult to have general discussions on the contents of spoken dialogue corpus, so we are obliged to deal with them individually. In terms of formal properties, however, we will be able to have a sort of design principles or guidelines for selecting the most suitable corpus.

For instance, when there are transcribed texts of simulated dialogues, we may obtain reference data to decide which material is more suitable for given objective by analyzing statistical characteristics of the texts.

As a first step, we investigated temporal characteristics of speech corpora. Next, we recorded telephone shopping dialogue as a typical example of simulated dialogues. Then we made statistical investigations of various speech corpora to extract useful information reflecting the contents of the corpus to some extent.

2. Temporal Characteristics of Read Speech and Spoken Dialogues

Temporal characteristics of utterances of read speech and spoken dialogues were analysed in order to see some acoustic differences between the read speech and the spontaneous speech. Speech data were segmented into "utterance units",each of which is a speech interval bounded by pauses, or an utterance such as interjection, filled pause, response word, false start. Then each utterance unit was labelled with its transcription and utterance category. Firstly, duration of pauses and utterance units were examined. Distribution of pause and utterance unit duration in spoken dialogues was wider than read speech, which reflected various disfluencies of spontaneous speech. Secondly, temporal characteristics of utterance units dynamically change as the dialogue develops. Duration of interjection words and filled pauses in topic boundaries was 1.7 times longer than the other place, which indicates duration of interjections and fillers is a sign of topic shift of the dialogue. Response time between the client and the agent speaker becomes shorter as the dialogue proceeds, indicating that interactions between the participants become active and smooth. Average speech rate was relatively low in the main part of the dialogue, where specific knowledge about the task is exchanged.

3. Recording of Telephone Shopping Dialogues

We took up telephone shopping dialogues as an example of spoken dialogue recording. Reasons are that, recently, many people are much familiar with telephone shopping, and that it is rather easy to avoid the case where only one speaker speaks more because a customer orders merchandise through a catalogue. We used high-quality microphones instead of telephone with each speaker separated by a shield; the dialogue was recorded by DAT.

Speakers playing the role of a customer were asked to read the merchandise list of 158 items as a practice of pronunciation as well as checking the recording level. Customer speakers placed an order using a data sheet which includes customer name, customer number, address, telephone number and credit card number as well as name/company name, address, telephone number for delivery.

The customer's name is the speaker's real one while name/company name and address for delivery are imaginary. The address is a combination of real names of prefectures, cities, towns and streets. Telephone and credit card numbers are, of course, imaginary with the same number of digits as real ones. Area codes of telephone numbers are chosen so as to conform to the corresponding address. The number of merchandise is 224 items chosen from a catalogue. A customer speaker recorded at least two dialogues without customer number at first and with it second time and after. A trained speaker played a sales clerk accepting orders according to a manual. An assistant helped her calculate the total amount. Forty seven telephone shopping dialogues were recorded with 13 male and six female speakers.

The speakers were university students in their twenties. Out of 47 dialogues, 35 were considered favourable with the shortest utterance time of 3 minutes 2 seconds, the longest of 8 minutes 36 seconds and 4 minutes 30 seconds on average. So far, we have transcribed 10 dialogues into text. The smallest number of utterances was 89, the largest 216 and the average 123. The smallest number of characters per utterance was two, the largesst 71 and average 11. Most of the dialogues were converted into CD-ROMs.

4. Statistical Characteristics of Various Speech Corpora

We investigated the occurrence frequencies of a phoneme, a two-phoneme sequence, a mora and an n-mora sequence (n-gram), using 55 phonemic units, which include palatalized phonemes such as /ky, sy, ny, .../, and 215 moraic units, which include CwV for transcribing loan words where C stands for a consonant and V a vowel.

A mora is a unit almost equivalent to a kana-character in the Japanese language. In this research, the total number of moras is set to be 215, including palatalized moras, those for loan words, syllabic nasal /N/, double consonant /Q/ and elongated vowel /-/.

4.1 Speech Corpora Investigated

Speech materials used for statistical analysis were taken from ASJ (Acoustical Society of Japan) continuous speech corpus volumes one to six and ATR Dialogue Corpus and simulated dialogue corpora recorded in this project.

1. ATR 503 sentences (503 sentences, Vols. 1-3)

   Phonetically balanced sentences, typical written language, which are selected from newspapers, journals, novels, letters, textbooks, among others, in a way that several phonetic environments occur at the same rate as possible as they can.

2. ASJ Simulated Dialogues (3129 sentences, Vol.7)

   Transcribed sentences of the speeches after simulated dialogues were recorded as a task of various guides.

3. ASJ Guide Task Sentences (1027 sentences, Vols. 4-6)

   Selected from the above mentioned simulated dialogue sentences, 12 dialogues were transcribed and corrected by eliminating interjections and mispronunciations, considering legibility, syntactic standards and politeness.

4. ATR Dialogues (10609 sentences)

   Dialogues on the phone regarding an application for international conferences or travel information were recorded and transcribed.

5. Telephone Shopping Dialogues (2060 sentences)

   Simulated dialogues our laboratory recorded the last fiscal year were transcribed. 25 dialogues are available with a task of telephone shopping through a catalogue.

6. Electronic Secretary Dialogues (480 sentences), Schedule Management Dialogues (755 sentences), and Japanese MAPTASK Corpus (482 sentences)

   Simulated dialogues recorded by Kyoto University, University of Electro-communications, and Chiba University, respectively.

4.2 N-gram counting

The number of moras is 215. When investigating the joint occurrence frequency, as the combination number increases, the number of units grows tremendously. The method to count the occurrence frequency by pattern matching with all the units entered in advance is extremely inefficient, requiring a lot of memory size to enter only the units. To avoid this, a more efficient statistical investigation method of N combined units (N¿1) is proposed by Nagao et al.

4.3 N-gram Occurrence Frequency

Our studies have revealed so far that written language can be distinguished from spoken language by investigating 2 or 3-mora combinations and that the occurrence frequency of redundant words can be found approximately.

This time, we conducted statistical investigation of 1 to 10-mora combinations for each speech corpus of 1 to 6 as stated above. Out of them, the results of 5-mora combinations are indicated in Table 1. This table lead to the following findings.

For ASJ Simulated Dialogues, ASJ Guide Task Sentences and ATR Dialogues, expressions particular to dialogues, such as "soudesune"(Is that right ?), "soudesuka"(Is that so ?) or "arimasuka"(Is there ...), occur in 5-mora combinations, suggesting that they are dialogue sentences. "desukeredo" (you are right, but ...) has a high occurrence frequency in ASJ Guide Task Sentences while "N'desukedo"(you're right, but ...) does in ASJ Simulated Dialogues, showing that ASJ Guide Task Sentences have more polite expressions. Actually, ASJ Guide Task Sentences are ASJ Simulated Dialogues corrected.

Moreover, we can presume the topic of dialogues to some extent from such combinations as "ninottekudasai"(take the train of ...) in 8-mora or "hibiyaseN'nokamiyacho(u)"(Kamiyacho station on Hibiya line), "kikaisiN'koukaika(N')"(Machinery Promotion Building) in 10-mora of ASJ Guide Task Sentences. Similarly, we can predict a task of reception or reservation from such combinations as "a-soudesukaha(i)"(well, is that so ? Yes) in 8-mora, "kuonegai-itasimasu"(Will you please do it ?), "arigatougozaimasu" (thank you very much), "haikasiko-marimasita"(Yes, I understand) or "yorosiidesyokahai"(Is it OK ? Yes) in 10-mora of ATR

Dialogues and Telephone Shopping Dialogues. Moreover, nouns inherent in the task occur, such as "gocyuumoN'" (your order) in 5-mora or "odeN'wabaN'gou" (your telephone number) in 8-mora of Telephone Shopping Dialogues.

5. Conclusion and future Issues

This paper reports the results of our statistical investigations in an attempt to evaluate the characteristics and contents of speech corpora quantitatively.

By investigating the frequencies of 1 to 10-mora combinations in existing eight corpora, we were able to find the distinction between written and spoken languages, keywords of dialogues and the rough task of dialogues.

The simple statistical investigation makes it possible to represent the contents of the corpus to some extent without conducting a complicated job such as morphological analysis.

The issue in the future is to review the efficiency of these methods to propose specific applicabilities in designing a speech corpus by investigating statistically a larger amount of speech corpora.

Publications

[1] K. Takagi, N. Houra and S. Itahashi,"Characteristics of Utterance Units and Temporal Structure of Spoken Dialogue", Proc. ISSD'93, pp. 287-290 (1993).

[2] K. Tanaka, K. Kinebuchi, N. Houra, K. Takagi, S. Itahashi, K. Itou, and S. Hayamizu, "Annotating illocutionary force types and phonological features into a spontaneous dialogue corpus: An experimental study", Proc. ICSLP94, Paper 30.11, pp. 1831-1834, Yokohama (1994).

[3] S. Itahashi, N. Houra and N. Ueda,"Designing, Recording and Statistical Analysis of Spoken Dialogue Corpora", Report of the Priority Area Project "Spoken Dialogue", pp. 1-7 (Dec. 1994).

[4] N. Houra and S. Itahashi,"Comparison of Statistical Characteristics among Speech Corpora", Preprints of the Spring Meeting of ASJ, Paper 2-Q-22, pp. 339-340 (1995).

[5] N. Ueda and S. Itahashi,"Recording and Analysis of Telephone Shopping Dialogue", Preprints of the Spring Meeting of ASJ, Paper 2-Q-21, pp. 337- 338 (1995).

[6] S. Itahashi,"Speech Research and Speech Database", Proc. Phonetic Society of Japan (PSJ) 1995 Annual Convention, Invited Lecture, pp. 5-10 (1995).

[7] N. Ueda, K. Takagi, M, Yamamoto and S. Itahashi,"Characteristics of High Occurrence Frequency N-Grams of Various Speech Corpora', Report of the Priority Area Project "Spoken Dialogue", pp. 1-4 (Dec. 1995).

[8] N. Ueda and S. Itahashi,"Characteristics of High Occurrence Frequency N-Grams of Spoken Japanese Corpora", to be published in Preprints of the ASJ Spring Meeting (1996).

Keywords: speech corpus, database, statistics, utterance unit, temporal characteristics, telephone shopping, spoken dialogue, n-gram, occurrence frequency