Data Collection and Characteristics of Spontaneous Speech on the Scheduling Task

Akira KUREMATSU, Tomomi Nakasuji, and Hideaki YAMAMOTO
Department of Electronics Engineering
Univ. of Electro-Communications
Chofu-shi, Tokyo 182, JAPAN
e-mail: kure@electra.ee.uec.ac.jp

1. Introduction

The importance of database of spontaneous speech database has been highly emphasized as a basis of fundamental researches on spoken language to realize a robust man-machine spoken dialogue system . We have collected human-to-human dialogues data of scheduling task in which two speakers arrange a meeting by negotiation. The design of spontaneous database and the procedure of data collection and transcription are issues to be studied in detail. After collecting spontaneous dialogue speech data, it is important to transcribe to be used for analysis and design spoken dialogue components. Analysis on linguistic characteristics of spontaneous spoken dialogues in the database have been made mainly in terms of the statistics of the part of speech. We have investigated the method to subcategorize word cluster of the part of speech. It was shown that the perplexity of part-of-speech bigram grammar decreased by subcategorize word classes.

2. Data Collection of Spoken Dialogue of Scheduling Task

We have selected the scheduling task as the appropriate task of the spontaneous spoken dialogue. In the scheduling task, two speakers talk each other looking at each individual calendar. The speakers are instructed to assume that they will make an appointment to meet for some specific purpose. They exchange conversations to settle the convenient date and time of their appointment. The calendar is for the next two weeks or one month. The topic of the negotiation is free and the speakers take their own suitable topic such as a meeting or an event plan. When the common agreement is settled, the conversation will be concluded. Each speaker is given a calendar and asked to use the calendar. There are 25 kinds of calendars which is filled in either two or four weeks period during one month. Due to this fact the database contains dialogues over the same constraints of scenario as performed in various speakers. The calendar is that of a college professor.

Both speakers' native language is Japanese. Speakers are allowed to speak standard Japanese as spoken, but they are not required to have had such special training in speaking, for example, as an announcers or an actor/actress. Each speakers is permitted to speak in the dialect he or she generally uses in family or private life. Most of the speakers are students and staffs in the university for ease of collecting persons. Background information is taken for each speaker; this includes age, sex, education and places the speaker has lived. Dialogues are spoken under the action partners of professor vs. professor and some of the dialogues are taken under the actions of professor vs. secretary and secretary vs. secretary. The scenario involves third party scheduling and the calendars include such events as weekday's business appointments schedule, weekend private activities and extended vacations.

The dialogues are recorded in a regular office environment or meeting room. The speakers sit in the same room but do not face each other in order to avoid non-verbal communication.

As for the condition of the speaking, we set up two types. One is the free speaking style which allows the overlapping of utterance. Another is the style that the speaking turn is controlled and the overlapping of utterance is reduced. In this case, speaker's turn is controlled by a lamp switch system. The speaker pushes a button at the end of his or her turn and the lamp light is changed. The speaker is allowed to speak while the speaker's lamp is light up. This procedure is sufficient to prevent the speakers from interrupting each other during a dialogue.

The number of collected and transcribed data is as follows. The free utterance allowing overlapped conversation (type A) has 13 dialogues. The controlled utterance not allowing overlapped conversation (type B) has 6 dialogues. Another 200 dialogues of type B is also analyzed for reference.

The recording system uses SennHeiser HMD410 microphones and records on a DAT deck. The DAT's sampling rate is 48KHz and the signal data is resampled at 16KHz by using workstations. All recordings of two speakers are made using the DAT recorder in two different channels.

3 Transcriptions

It is worthwhile to transcribe the linguistic description and annotate the tag for undertaking statistical analysis of phonetical and linguistically usage. It requires a lot of man power to transcribe and check the quality. The procedure of automatic or semi-automatic labelling is required to cope with the efficient labelling of a large amount of data. The collected dialogues have been transcribed in kanji-kana sentences and Roman script sentences. Both speech and non-speech events have been transcribed. The speech has been transcribed including words, human noises, non-human noises, silence, false starts, and transcriber comments. Hesitation i.e. human noises and horticultural human noises i.e. human non-phonemic noises are enclosed by angle brackets. Hiragana characters are used in kanji-kana sentences for the hesitations. Hesitation is sometimes confused with interjection in Japanese. In hearing the utterance, hesitation is decided when the pronunciation is hesitated. Part-of-speech is useful for the language model of spoken dialogue. Part-of-speech is tagged by use of morphological analyzer. The 27 kinds of part-of-speech are followed by "Daijirin" dictionary.

4. Analysis of Spoken Dialogue

Issues of linguistically characteristics of spoken dialogue of scheduling task are analyzed in order to develop a language model. Analyzed data are 13 dialogues of a type A, 6 dialogues of type B and 200 dialogues of type B. It requires a large number of data for the statistical analysis, however, the current data can be analyzed to induce a useful information. The data of the scheduling task has been analyzed statistically. The total number of utterance of 13 dialogues of type A is 1019. The average number of words per one dialogue is about 7. Since there is an overlap of two utterance, the utterance is short and the average number of words is small. There are 750 different words out of total words of 8624. Analysis of frequency of each part-of-speech is also calculated.

Peculiar characteristics are observed in the spontaneous dialogues such as false start, hesitation, insertion, and stoppage in the middle of sentence. The peculiar expressions of spoken dialogues of type A are extracted. Sometimes, the ending of the sentence is not completed from the grammatical point of view. That is, the endings are nouns, or in the middle of sentence. Young persons use their own compounded words or local expressions of the current fashion. These words become undefined words. In the usage of interjection and adverb, there is a individual peculiarities.

Interjection is a word which occurs at the beginning of a sentence and expresses motto, calling, or response. The ratio of occurrence of interjection per one dialogue is about 12.5% in the more frequent case and 5.02% in the less frequent case, and in average 8.89%. The style of conversation differs according to the social position of a speaker in Japanese. For example, in the case of conversation between intimate friends, pronunciation becomes vague and sentences become ill-formed from the grammatical standpoint. On the other hand, when a person speaks to a higher ranked person or meet each other for the first time, pronunciation and grammar becomes accurate comparatively.

5 Classification of Word Category

Word bigram has been conventionally used to predict the preceding word In the language model of the continuous speech recognition. It is usually not worthwhile to use it in the open speech sentence which is different from the ones used in the training of the bigram statistics, especially when the volume of training data is small. It is considered to use a bigram of part-of-speech instead of that of words . In the case of bigram of part-of-speech, the constraint becomes more loose than the case of word bigram. When many words are contained in the same part-of-speech like noun, the prediction is not accurate. The subclassification of part-of-speech is introduced for noun, verb, interjection. The clustering based on the mutual information is conducted. In this case, the value of mutual information becomes large when pair of words appears often simultaneously.

The procedure of clustering is as follows: First, one word is assigned to one class, and then a pair of two classes are merged into one class and the average mutual information is calculated. By examining all pairs, we can find the pair of clusters for which the loss in average mutual information is least. This pair of classes is merged into one class. This process is continued until we get the desired number of classes. Here, the information is calculated as the mutual information is obtained for all pairs of words across clusters and they are averaged. When a word is contained in the cluster, the estimation of the mutual information is conducted for all pairs of words in the class and then the average is calculated.

20 numbers of clusters are obtained by applying the above mentioned procedure to 206 nouns and 117 verbs respectively. Interjections are classified to 5 subclasses. The classification by use of the Japanese thesaurus dictionary is also conducted for normal nouns. In order to evaluate the effectiveness of the classification, the perplexity in the average branching factor in the sense of an information theory. Here, the probability of appearance of word bigram is considered. We can see that the perplexity in the case of subcategorization of the part-of-speech has become smaller than the case of original part-of-speech classes. The value of perplexity in the case of classification based on the mutual information was about the same as that of the case of using thesaurus dictionary. The value of perplexity is large in the case of spontaneous speech.

Keywords: spontaneous speech, speech database, mutual information, perplexity