Marking Up the Dialogs with <utterance> Tags: The Unit of Utterance in the Technical Sense

Syun TUTIYA
Department of Philosophy, Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba, 263 Japan
tutiya@cogsci.l.chiba-u.ac.jp

In the process of developing corpora of spoken dialogs is always involved the task of tagging the text as a sequence of utterances. Interestingly enough, an utterance is not necessarily a sentence in the grammatical sense, nor a sentence an utterance. Besides, the sentences might not be completed for various obvious reasons. Responses from the interlocuter might not be verbal but could be kinetic. Most typically, one interlocuter's utterance might interrupt the other's utterance, giving the observer the impression that the utterances do not follow each other but overlap. All these casual observations lead us to a serious consideration of the notion of "utterance" and the way to tag utterances. We still need to compromise about the notion of "utterance" in the technical sense.

By definition, a dialog is a sequence of utterances. In terms of SGML, the mark-up language we have decided to adopt as the basis of our tags of the transcription of dialogs, the content model of the element <text type=dialog> is one or more of the ordered sequence of the elements <u>. Technically put, the problem is where to put the start tag <u> and the end tag </u> for each utterance. Take for example the difficult case of semi-interruption. In the dialogs in Japanese, speakers, more often than observed in the dialogs in English, tend to utter interjective phrases or non- lexical human voices as signs of assent. Assuming we are equipped with an SGML mechanism of handling overlapping phenomena as in the line of TEI, we still have problems deciding on the status of such assenting sounds and the continuity of the utterance to which such assents are addressed.

The literature from the preceding attempts to mark up dialogs mainly in English show two distinctive policies in handling the unity of the utterance. In the tradition of conversation analysis and discourse understanding, where the notion of "turn taking" plays an important role, the utterance is more or less synonymous with the turn in their sense. An utterance would continue as long as the other interlocuter take up the right of utterance. In their analysis, the kinds of ascending voices would not mark the end of the interlocuter's utterance. On the other hand, in the tradition of cognitive science and discourse analysis, it is more customary to "chop" the dialogs in smaller pieces, searching for a unit which is just a little longer than a linguistic phrase.

We have experimentally tagged sample dialogs from the recording of map task dialogs in Japanese according to the two different policies and collected statistical data. After analysis, we are inclined to decide that the dialog transcriptions tagged with the notion of utterance in the second tradition would provide more reliable basis for further research in both speech recognition/generation and discourse understanding.

Advanced Research for Modeling of Spontaneous Spoken Dialogue Understanding System Using Prosody

ICHIKAWA Akira, IMIYA Atsushi and HORIUCHI Yasuo
Department of Information and Computer Sciences, CHIBA UNIVERSITY
1-33 Yayoi-cho, Inage-ku, Chiba-shi, Chiba 263, Japan
e-mail address: ichikawa@ics.tj.chiba-u.ac.jp
tel: +81-043-290-3256 fax: +81-043-290-3269

Modern technologies for computer-human interface are unbalance and excessively depend on some part of media, for examples the "mouse" and "windows". These technologies are rather difficult to use for advanced age people and people with disabilities. It is essential for all people to realize a spontaneous spoken language interface system under the natural dialogue conditions.

To understand natural dialogue utterances under real-time conditions, they should include the following information which may be realized by the prosody in addition to the segmental features.

a. Information which organizes the units (morpheme, word, phrase, sentence, paragraph, etc.)

b. Sentence structure information for clarifying the kind of sentence, as well as its meaning (syntax structure information)

c. Information which expresses mental state (repairs, restart, feeling, emphasis, stagnation, etc.)

d. Dialogue control information (the control right to speak, nod, etc.)

In this research we analyzed prosody of spoken dialogue utterances and Japanese Sign Language. Then we proposed a new model of a spontaneous spoken dialogue understanding system using prosodical information and multi agents system architecture.

(1) Roles of Prosody as Functions for Turn-taking and Cushion Words[1]

Before analyzing the supra segmental information of the spoken dialogue language, we survey two types of examination for those of the write style utterances. First one is the comparison between the results of the comprehension test of reading and those of hearing for Japanese garden pass sentences. Second is the prosodical analysis of the announced discourses of the radio news. These two experimental results show that the roles of the supra segmental information of speech is extremely important.

To analyze spoken dialogue sentences, 128 spontaneous conversations (total length is about 21 hours) were collected from 64 subjects through a map task[2]. The design as well as materials and procedures followed that of The HCRC Map Task Corpus[3], but the feature names were carefully scrutinized so they possibly represent interesting phonological modifications in Japanese and the recording system was enhanced. Giver and Follower's speech were recorded from independent microphones and separately stored on a DAT. Also, subject's maps and faces were video-recorded so that they show a Giver's hand movement on a map and a Follower's drawing a root and their eye- contact situations.

Also we developed a tool in the work station environment by which we can analyze the speech waveforms of dialogues and make them into texts. The speech sound can be output from any text of the dialogue.

Then using the corpus (named the CHIBA CORPUS), some characteristics of the prosody and cushion words were analyzed.

As the control function of turn-taking, three points are investigated;

(A) timing of nod-back or response

Nods were made just after speech or keywords in the partner's speech. It start less than 0.4 seconds after cues appear. This result means that to realize a system that understand spoken dialogue utterance which allows the users to speak in a natural manner, the system should be able to perform nods.

(B) discrimination cue between interruption and nod-back

If the user speaks when the system is speaking, the user's voice is the interruption or the nod. We examine whether or not we can obtain discrimination cues between interruption or nod through prosody. For maximum pitch there is the great difference in that interruption and nods similar to interruption have higher pitch frequencies than that of nod. From this, we note the next three points.

1. When a listener wants to obtain the right to speak by interruption, he expresses his request with a high pitch.
2. When he furthers his partner's speech with a nod, he lowers the maximum pitch frequency so as not to obstruct his partner's speech.
3. When he wants to express his agreement, he uses a nod similar to interruption and obtains the right to speak.

(C) discrimination cue between the end of utterances and stagnation of thinking

A pause in speech may indicate the end of speech or stagnation due to thinking. Since it is rather difficult to discriminate clearly between the end of speech and stagnation due to thinking by the use of end pitch, the rate of the end of speech as a function of the end pitch frequency is investigated. The frequency of speech termination is high in the areas of the lowest and the higher range of pitch frequency. The causes are presumed to be due to two types of speech termination; one type is the transfer of the right to speak, the other type is renunciation. In the case of renunciation, speech ends with a low pitch, and in the case of transfer, speech ends with high pitch to seek a reply from the partner.

The correlation between the rate of speech termination and the duration from termination to the next utterance of the partner is investigated as a function of sentence final-pitch frequencies. The duration until the partner's utterance is very short; therefore discrimination between the end of speech and stagnation due to thinking is very difficult. We thus assume that speech ends only in the case when the next utterance starts very quickly. Consequently, if there is no information in the end pitch frequencies which enables discrimination between the end of speech or stagnation due to thinking, the decision is made by the partner who may or not want to speak.

There are many kinds of cushion words in spontaneous speech. In this research, the typical cushion word "ne" was investigated, and we find three types of cushion word "ne". It is seen that each "ne" correlates strongly with its respective type of pitch pattern.;

(A) indication of interrogation

The pitch pattern indicates the rapidly rising type.

(B) request of affirmation

The pitch pattern is the slowly rising type.

(C) indication of self affirmation

The pitch pattern is the falling type.

Consequently we conclude that the speaker's psychological state can be inferred from the pitch pattern.

(2) On the Supra Segmental Features of Spoken Language and Sign Language [5]

Here it is claimed that spoken language and sign language are categorized as the natural dialogue languages. Thereupon, to clear the properties of natural dialogue languages and to find clues of actualization of the system, we made a plan to compare the spoken language with the sign language. (We also believe that the results of this research should make a great contribution to realize a translation system between the spoken language and sign language.)

They have to have functions which indicate the semantic structure of the utterance. The function of spoken language exists in the supra segmental features, ex. prosody, but that of sign language has not investigated yet. So we assume some supra segmental features in the sign language and confirm these hypotheses.

It is also shown that the prosody of the sign language has important information of its sentence structure through a sentence recognition experiment using motion speed of hands under some specific task conditions.

Firstly we try to classify the sign language words, phrases and sentences from the view point of the supra segmental structures. We got six types of the structures. (Those suggest that the structure of the sign language is similar to those of the KANJI characters because of direct visual understandability.)

Secondly time structures of the sign language sentences were analyzed. It becomes clear that the distribution of the gesture speed of the pauses between words in a sentence is highly correlative to the structure of sentence. And the gesture speed of sign indicates the importance of the word; the more slowly, the more important.

(3) Coding Method of Pitch Pattern to Analyze Conversational Speech

It was reported by Komatsu previously that the piece-wise-liner approximation of the pitch patterns on a log scale gave good results [6]. Pitch patterns are approximated as piece-wise linear lines using a newly proposed algorithm which is a modified method based on Hough Transform for line detection real-timely, and coded by the vector-quantization.

The conventional Hough Transform algorithm is the method to extract edge lines for the image transform.

Some functions of the prosody above mentioned in the conversational speech are related to certain sequences of the codes and will be identified by the HMM method.

(4) Real-Time Estimation of Sentence Structure from Prosody

A new method is proposed by which the sentence structures of utterances can be constructed under the real-time conditions from the prosodical information.

Our method consists of a modified Hough Transform algorithm. The conventional Hough Transform algorithm is modified as to be able to extract the approximated straight lines of pitch patterns under the real-time conditions. The sensitivity of the approximation can be established at will. A straight line under a lower sensitivity corresponds to a higher component of the sentence structure. So, parallel transform processes with different sensitivities can give multi level component of an utterance structure under the real-time conditions. The structure is not exactly the grammatical one but rather psychological one or one of the speaker's intentions.

(5) Similarity of Sentence Structure[4]

It was proposed that an on-line matching algorithm for spoken sentences which runs in real time. Our new algorithm is divided into two stages. At first, we transform a time-signal into a tree of the sentence by the modified Hough transform above mentioned. Secondly, we transform tree to regular trees by adding dummy nodes to express tree as lists.

Thus, editing distance of tree is computed recursively by using a lists matching technique.

(6) Modeling of a Spontaneous Spoken Dialogue Understanding System

These results will be useful to develop the spontaneous spoken dialogue human-machine interface system which should not obstruct the user's consideration for his task. So, we propose a new concept of a real-time spontaneous dialogue understanding system model using prosodical information.

They also suggest that a parallel processing is essential to comprehend dialogue utterances and the process should extremely differ from that for the written languages. It is also claimed that these surveys are important to realize the multi-modal human interfaces.

The system consists as a multi-agent system. Each agent processes its own job parallely; ex., one agent extracts sentence structure from the prosody under the real-time condition, one recognizes phonemes, one predicts next utterances, one analyzes an intention of the utterance, one estimates the combination of the outputs of the each agent, etc. And each agent is in capable of learning to take suitable steps and re-organized, and the system will be self-organized to adapt itself to a new situation.

Acknowledgement

The authors would like to thank Shinji Satoh, Yuhko Ishida, Tetsuya Sekiya, Fumihiko Nishida, Nagayuki Shimizu and Hajime Sasaki for their cooperation in spoken dialogue research, and Tadashisa Hashido, Hisashi Tanaka, Seiji Hayashi, Syuuko Unno, Chiyo Takayama and Atsuko Ichikawa for their cooperation in sign language research. CHIBA CORPUS was developed by the Chiba Corupus Project.

References

[1] Ichikawa, A. and Satoh, S,"Some Prosodical Characteristics in Spontaneous Spoken Dialogue," Proc. ICSLP 94, 5.11, pp.147-150 (1994)

[2] Aono, M. etal.," The Japanese map task corpus: an interim report," Japan Artificial Intelligence Association, report of SIG-SLUD-9402, pp.25-30 (Oct.1994) (in Japanese)

[3] Anderson, A.H. et al.,"The HCRC Map Task Corpus," Speech and Language, Vol. 34, 4, pp.351-166(1991)

[4] Imiya, A., Nishisda, F. and Ichikawa, A., "An Approximate Tree Matching Algorithm," Japan Artificial Intelligence Association ,report of SIG-SLUD-9402, pp.43-48 (Oct.1994) (in Japanese)

[5] Ichikawa, A. Yagi, K. and Hashido, T., "Spoken Language and Sign Language - Natural Dialogue Language for People with Disabilities - ," Proc. 9th Synposium on Human Interface , pp.299-306,(Oct.1993) (in Japanese)

[6] Komatsu, A., Oohira, E. and Ichikawa, A.,"Spontaneous Speech Understanding Based on Cooperative Problem Solving" IEICE Trans., Vol. E74, No.7, jul. 1991

Keywords: real-time dialogue understanding model, prosody, turn-taking, cushion word, corpus, nod, coding of pitch, estimation of sentence structure, similarity of sentence structure, multi-agent system, sign language