A New Discourse Structure Model for Spontaneous Spoken Dialogue

Tetsuro CHINO *Hiroyuki TSUBOI*

Toshiba Kansai Research Laboratories

8-6-26 Motoyama-Minami-Cho Higashinada-Ward, Kobe Japan.

{chino,tsuboi}@krl.toshiba.co.jp

ABSTRACT

In this paper, a new discourse structure model is proposed, and based on the model, we report the results of an analysis on Japanese-language dialogues over the telephone. As a result, a method for describing and analyzing the structure of spontaneous spoken dialogue is provided, and some characteristics of spontaneous spoken dialogue over the telephone were clarified.

1. INTRODUCTION

A consequence of the progress of natural language processing (NLP) and automatic speech recognition (ASR) is an increasing requirement for processing spontaneous spoken dialogues. Attempts have been made to integrate NLP and ASR; but the nature of spontaneous spoken dialogue prevents simple conventions of NLP and ASR from being effective and successful. In addition, the unique characteristics have not yet been clearly described. A further complication is that each utterance of spontaneous spoken dialogue cannot be handled independently of the context. Thus, a discourse structure model of spontaneous spoken dialogue for analysis and processing is required. In this paper we propose a new discourse structure model for spontaneous spoken dialogue which makes it possible to deal with some of the characteristics of spontaneous dialogues. We also report the results of an analysis of Japanese-language dialogues over the telephone, based on the model proposed here.

2. DIALOGUE DATABASE

2.1. Development of Database

Detailed investigations of naturally occurring data are indispensable to develop a discourse structure model which treats spontaneous spoken dialogues. Therefore, first of all, we had developed a dialogue database and analyzed its contents in order to clarify the necessary conditions for the model. The database has transcriptions of each utterance, timing information, markers for linguistic/prosodic clues, temporal relation between utterances, and so on[7].

Material. For our purpose, we collected 45 minutes of audio data of dialogues over the phone from Japanese-language radio program[4]. The audio data comprised 10 dialogues, more than 1,500 utterances, and 21 participants in total. This material was chosen for the following reasons. First, because the participants communicate by voice only, we are able to concentrate solely on linguistic/prosodic phenomena, since facial expressions, gestures, and so on are absent. Second, the topics of and the participants in each dialogue are not predetermined. Third, since the participants were unaware that their utterances would be analyzed, the dialogues are spontaneous.

Speech Fragments. It is difficult to identify a unit of utterance. In this research, we divided the dialogue sound data into segments of audio data by physical threshold (*e.g.* pause of over 100 msec) , and combine them with other factors (*e.g.* utterer of each utterance) to determine the border and the unit of utterances. After this, we refer them as speech-fragment(s) (*SF*s).

2.2. Analysis of Database

Length. Figure 1 shows the histogram of length of speech-fragments, and indicates the following characteristics of spontaneous spoken dialogues: (1) almost all utterances are short and fragmentary (average length is 1.75 sec), (2) a remarkable peak exists at the interval of 0.3-0.4 sec, corresponding to one interjection typically, (3) a gentle peak exists in the range of 0.7-1.2 sec, corresponding to syntactic form of one interjection plus one clause typically.

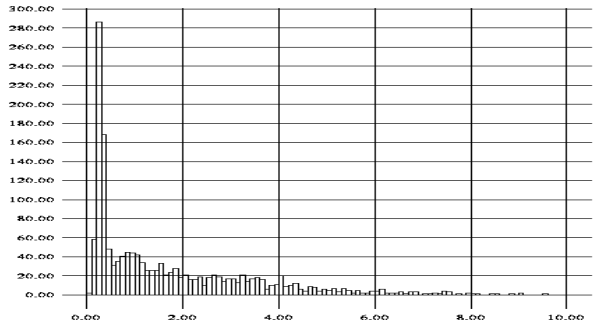

Figure 1: The histogram of length of speech-fragments. (x-axis is intervals of length of speech-fragments in second, y-axis is their frequency in each interval.)

Overlapping. Figure 2 shows the histogram of length of pause/overlap between adjoined speech-fragments. The right

half of the graph represents the distribution of length of pause, and the left half represents that of overlapping time.

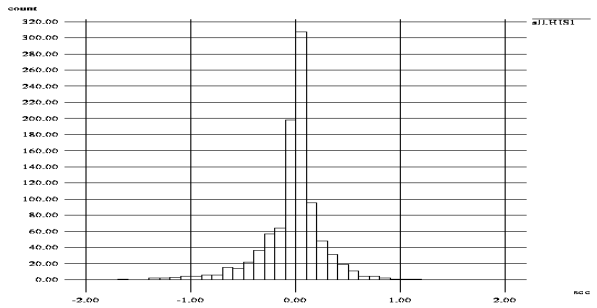

Figure 2: The histogram of length of pause/overlap between adjoined speech-fragments in dialogues. (x-axis is intervals of length of pause/overlap in second, y-axis is their frequency in each interval)

Figure 2 indicates many facts[7], but the most important and interesting result is that (4) exceeded our expectations: overlap of utterances occurs very frequently (about 40%) in spontaneous spoken dialogues.

Clues. An analysis of the database from the viewpoint of linguistic/prosodic clues was performed and revealed following: (5) almost all speech-fragments (over 90%) include more than one clue (linguistic clues: *e.g. interjections, connectives* : over 78%, prosodic clues: e.g. marked *pause, rise* or *rise and fall* of intonation contour : over 38%). Thus, it seems reasonable to suppose that speech-fragments can be a element of a discourse structure model for spontaneous spoken dialogues.

2.3. Requisites for the model

Judging from the above, the discourse structure model for spontaneous spoken dialogue has to be able to deal with (A) fragmentary spoken utterances, and (B) overlapped utterances. Since almost all previous researches concentrated on only written language or transcripts of spoken language, and paid little attention to the timing of utterances, they are not sufficient for our purpose. This is our motivation for developing the new discourse structure model proposed in section 3.

3. DISCOURSE STRUCTURE MODEL

3.1. Exchange Structure

Various discourse structure models have been proposed by other works. We examined many of them and chose Exchange Structure[3,5] as the basis of our model. In the theory of Exchange Structure, each utterance in discourse is classified by a set of utterer intentions (e.g. initiate, response, feedback), and analyzed in terms of the pattern of a series of such types.

The following is an example of a transcription of dialogue that forms an exchange structure (from [5]):

(I) Can you tell me why you eat all that food - yes?
(R) to keep me strong.
(F) to keep you strong, yes - to keep you strong.

Although the local structure of dialogues can be captured successfully by the exchange structure, well-formed utterances at each constituent of the exchange structure are supposed, and therefore it cannot be adopted for spontaneous spoken dialogues.

3.2. Overview

Figure 3 shows an example of out proposed discourse structure model. It is a tree structure whose leaves are speech fragments and represents the flow of arguments in dialogues.

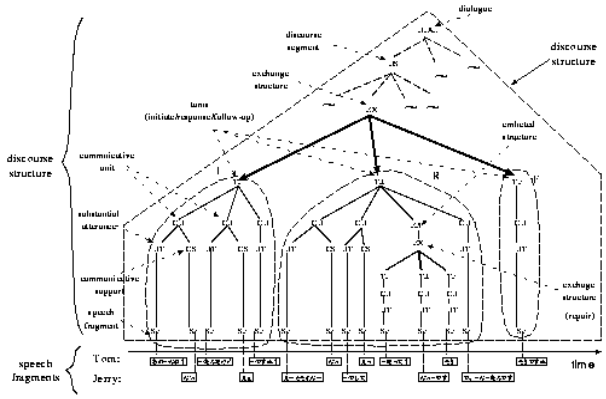

Figure. 3: An example of discourse structure.

In this model, each node represents some sub-structure of the discourse and is classified into thirteen categories that are layered in seven levels (*SF ~ UT*/*CS ~ CU*/*EB ~ TI*/*TR*/*TF* /*CR* /*CF ~ EX*/*DX ~ DS ~ DIAL*). In the following sections, we examine these categories and levels.

3.3. Fundamental Elements

Speech-Fragment (*SF*) ~ Substantial-Utterance (*UT*) / Communicative-Support (*CS*) ~ Communicative-Unit (*CU*). C*ommunicative-units* (*CU*s) are introduced to this model as the minimum units of communication among the conversational participants. They consist of more than one *substantial-utterance* (*UT*s) of one participant who keeps the *initiative* of the utterance at that point, and arbitrary number of (optional) *communicative-supports* (*CS*s) uttered by the conversational partner (hearer). Both a *UT* and a *CS* are one *speech-fragment* (*SF*). While the former can contain an *SF* with any types of information, the later can contain only an *SF* which is uttered with some intentions to control or maintain

the communication (In other words, *SF*s corresponding to the latter are uttered by some discourse purpose.). An agreeable response made to the utterance of a conversational partner is one of the typical instance of *CS*. The discrimination between *UT* and *CS* is performed by examining the surface characteristics (*e.g.* occurrence of linguistic/prosodic clues) of each SF. It should be noted that since the temporal order of each SF that shapes some *CU* is not restricted, this model can deal with overlapped and fragmentary utterances and fit for spontaneous spoken dialogues.

Turns (*TI*/*TR*/*TF*) ~ Exchange-Structure (*EX*). The *exchange-structure* (*EX*) is the basis of our model, and an *EX* sub-categorizes an obligatory *turn-of-initiate* (*TI*) of a participant, and an obligatory *turn-of-response* (*TR*) of the partner, and an optional *turn-of-feedback* (*TF*) by the participant who initiates the *EX*.

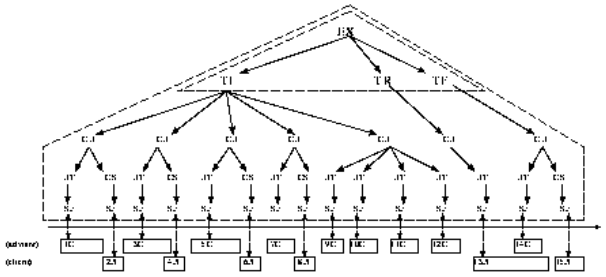

Figure 4: An Example sub-structure of a part of dialogue derived from the database, that shapes an *EX*. A series of *CU*s take the shape of *TI*, *TR*, and *TF*, and these are constituents of the *EX*.

As we have seen in the previous section, each utterance is fragmentarily spoken in real dialogue, and it is very hard to convey complex information or realize some kind of intention by only one utterance. Based on this observation, we define a *turn* as a parent node of a series of *CU*s, and one conversational participant keeps the initiative of utterance in each *CU* in a turn. And each *CU* in a turn serves to realize one intention: *initiate*, *response* or *feedback* in all. (On the contrary, in the conventional *exchange-structure*, a complete sentence (utterance) is supposed to consist of an exchange.)

Discourse-Segment (*DS*) ~ Dialogue (*DIAL*). Dialogues are created as a result of some acts of more than one participant. Even though in a cooperative dialogue, since each participant has his/her own goals or plans, utterances that occur in real or spontaneous spoken dialogues sometime fail to achieve the utterers' intended goal. In addition, there are many cases in which plural exchanges are spent to realize one goal.

Based on these observations, we introduced *discourse-segments* (*DS*s), and defined a *DS* as a parent node of a series of *EX*s that are initiated by one participant to realize his/her goal(s). While in previous works, the words "discourse segment" are used in many different senses and there is no consensus among the researchers[6], in our model it has been clearly defined as above and represents the range of plan or initiative of each participant in the discourse structure model. *Dialogue* (*DIAL*) is defined as a parent node of a series of *DS*s ,and is the root node of the discourse structure.

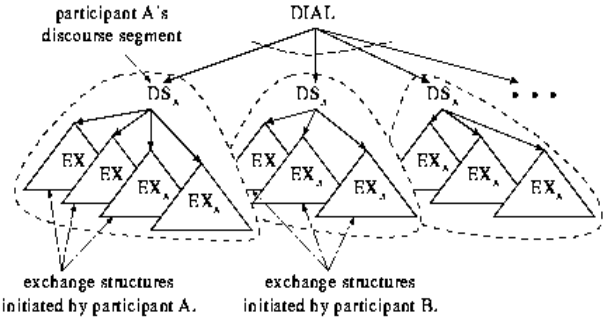

Figure 5: Relation among *EX*s, *DS*s, and the *DIAL*.

3.4. Additional Elements

As there are many phenomena that cannot be captured by conventional discourse models in spontaneous spoken dialogues, some additional elements are introduced to out model. Here, a typical example is presented.

Discarded Utterances and Canceled-Turn-of-Feedback (*CF*). Figure 6 shows an example of a part of a dialogue which contains what we call a "*discarded utterance*", and Figure 7 shows the corresponding sub-structure with a *canceled-turn-of-feedback* (*CF*).

SF_{A1}: *When are you going?*		
SF_{B2}: *Let me see, You see this meeting will be held in Tokyo*		
SF_{A3}:	*Uh-huh*	
SF_{B4}:	*Tomorrow,*	*I've got another meeting in Osaka.*
SF_{A5}*:		*I see, you're leaving tomorrow ...*
(SF_{B4}): *So, I'm leaving Friday*		
SF_{A6}:		*Oh.*
SF_{A7}: *You mean you're leaving Friday.*		

Figure 6: An example dialogue with a "*discarded utterance*" (SF_{A5}). (Notes: Vertical Bars in this figure represent timing information. This transcription is an artificial example for the purpose of explanation)

In this example, first, participant-***A*** asks a question and *initiates* this *EX* by SF_{A1} *(speech-fragment*-A1). Then, participant-***B*** starts to reply by SF_{B2}, and ***A*** makes an overlapped agreeable response SF_{A3}. At this point, the initiative of utterance moves to ***B***. Then ***B*** starts $SF_{B4.}$ While SF_{B4} is presented, ***A*** utters SF_{A5} to try to take back the initiative

of utterance, and to realize a *feedback* intention to the *response* from ***B***. But, ***B*** continues speaking and doesn't hand over the initiative of utterance, and ***A*** stops SF_{A5} and it becomes what we call a *discarded utterance*. Then, ***B*** finishes SF_{B4} and realizes the intention of *response*. Finally, ***A*** utters SF_{A7} as a rephrased *feedback.*

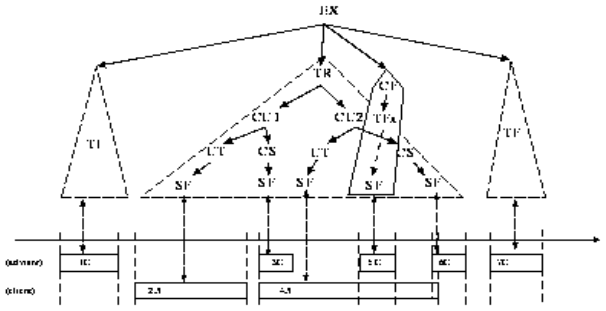

Figure 7: An example of discourse structure with a *canceled-turn-of-feedback*(*CF*)

We introduced into our model the notion of "*discarded utterances"* and additional categories (*CF, CR, and DX* : see below) for them. As a result, our model can capture phenomena of this kind that sometimes occur in real or spontaneous spoken dialogues, but are ignored in all of previous models. Thus, our model has a conspicuous advantage for spontaneous spoken dialogues

3.5. Formal Description

Based on segment grammar [1,2], which was designed as a cognitive model of incremental sentence generation, we have developed a formal description of our model. Therefore, our model has the potential to serve as a basis for incremental processing of spontaneous spoken dialogues.

4. CHARACTERISTICS OF JAPANESE DIALOGUE OVER THE TELEPHONE

In this section, we present the results of an analysis of the data in the database. An amount of data (includes 6 dialogues, 755 utterances, 1,200 sec) is examined based on the proposed model by hand, and the results are summarized as follows.

1. 130 *EXs* are extracted, and one sixth of them are embedded, and about 10% are classified as *discarded - exchange* (*DX*). (In a *DX*, some utterances in the *TI* intended to initiate an *EX* are discarded by the conversational partner.).

2. The size of each *TI* (measured by the number of *CUs*) is about twice that of a *TR* and almost all *TF*s consist of only one *CU*.

3. No correlation between the participant who has the initiative of utterance and the initiator of embedded exchange in the *turn* was observed.

These results lead to the following considerations. First, as shown in (1), discarded phenomena often occur in spontaneous spoken dialogues. (A) This fact corroborates the validity and advantages of our model. Additionally, (2) and (3) can be seen as evidence that supports the following assumptions. (B) In spontaneous spoken dialogues over the telephone, many *CU*s tend to be devoted to the initiation part of exchanges to clarify the intention intended to be conveyed. (C) In spontaneous spoken dialogues, when a problem in communication (*e.g.* misconception, mishearing) arises, an embedded sub-dialogue for the repair is initiated immediately regardless of who has the initiative of utterance at that point in time.

5. CONCLUSIONS

In this paper, we proposed a new discourse structure model that can deal with some phenomena that occur in real or spontaneous spoken dialogues. Additionally, an analysis based on the model is also performed and some characteristics of spontaneous spoken Japanese-language dialogues over the telephone are clarified.

Integration of a plan-goal structure into our model, and extensions of our model to handle multimodal communication are subjects for future work.

6. REFERENCES

1. Koenraad DE SMEDT, IPF: An incremental parallel formulator, In R. Dare(Eds.), Current Research in Natural Language Generation, Kluwer Academic Pub., pp.167-192, 1990.
2. Koenraad DE SMEDT, Segment Grammar: a formalism for incremental sentence generation. In C. L. Paris(Eds.), Natural Language Generation in Artificial Intelligence and Computational Linguistics, Kluwer Academic Pub., 1990.
3. K. MAYNARD, Senko, *Kaiwa Bunseki* (Conversation Analysis), Kuroshio Shuppan Pub., 1993.
4. NHK, Ku*rashi no Denwa Sudan,* NHK Radio Program, 1993.
5. Michael STUBBS, Discourse Analysis - The Socio-linguistic Analysis of Natural Language -, Basil Balckell Ltd., 1983.
6. Bonie WEBBER, Lynn, Discourse deixis: Reference to discourse segment, In ACL proceedings of the 13th International Conference on Computational Linguistics(COLING-88), Vol. 2, pp. 113-122, 1988.
7. Tetsuro CHINO, Development of Human-Human Spontaneous Spoken Dialogue Database, and Development of Discourse Model, IPSJ-SIG-SLP-2-9,(Japanese), pp.59-66,1994.
8. Tetsuro CHINO, An Analysis of Consultation Dialogue on Telephone Based on a Discourse Structure Model of Human-Human Spontaneous Spoken Dialogues, IPSJ-SIG-SLP/EIC-SIG-SP-94-67/EIC-SIG-NLC-94-36, (Japanese), pp.33-40, 1994.