Tagging Structure and Relationships in a Japanese Natural Dialogue Corpus

Shimpei Makimoto[1], *Hideki Kashioka*[123], *Nick Campbell*[123]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan
[2]National Institute of Information and Communications Technology,
[3]ATR Spoken Language Communication Laboratory, Keihanna, Japan
{shimpei-m,kashioka,nick}@is.naist.jp

Abstract

The purpose of our study is to develop a method for tagging structures and relationships between segments in spoken dialogues, in particular domain-free chatting and casual conversation, to deal with information from recorded dialogue resources for speech and natural language processing applications.

In this paper, we present the specification of our tagging set and discuss some insights gained from the study. We propose a novel structure and suggest tags for annotating discourse fragment relations in Japanese dialogues. We test this system with a large corpus of telephone dialogues, part of the JST-CREST/ATR Expressive Speech Processing corpus. We first detect 'utterance fragments', i.e., smaller discourse units than phrases or sentences, and then link these fragments using a set of relationship attributes. We present samples of tags and relationships tagged in this way for parts of the ESP_C corpus dialogues.

Index Terms: discourse, dialogue, language resources, tagging

1. Introduction

Dialogues are one of the most popular methods of human communication, and considerable amounts of information are exchanged every day through dialogue. While acoustic and storage technologies are improving dramatically, huge amounts of dialogue resources are now being stored and are becoming avaialable. Similarly, high quality dictation engines will soon be able to provide transcriptions of these resources. Such resources can be used for many applications of speech and natural language processing (dialogue summarization, speech synthesis, information extraction and so on).

However, natural conversational dialogues, in particular non-task-oriented and domain-free dialogues like chats, are less clearly organised, more complicated, and don't have a clearly-defined structure. We cannot easily analyze such resources using current natural-language tools in a normal way, and if we are to produce a technology which is able to react to such dialogues in future applications, considerable investment of research into their structural organisation will be needed.

Several previous studies have proposed dialogue and discourse structure annotation schemata (e.g., [1, 2]). However, most of them assume that the conversations are oriented to some business purpose and expect clearly task-oriented dialogues. They are not easily adaptable to general friendly conversations where phatic communion is as common as pragmatic function. Furthermore, their annotations require deep human understanding and world knowledge, so they are not suitable for automatic tagging with high accuracy (see e.g. [3]).

In the present study, we define a new tag set for annotating natural dialogues. First, we define the basic unit of processing: 'utterance fragments'. We then separate natural dialogue utterances into streams of linguistic and non-linguistic events as made up by these fragments.

By first chunking utterance units into their component fragments, the smaller unit facilitates representation of human utterance functions and relationships by heuristic semi-automatic methods. Subsequently, the fragments are linked together again by a set of relationship attributes. We propose 5 simple basic relationships which can be distinguished by surface information using utterance sounds and transcriptions. We give samples of these annotations using a natural telephone dialogue corpus ESP_C, a subset of JST/ATR ESP Corpus [?].

In this paper, we present our tagging specification and results and describe insights gained through the tagging of actual dialogues in the natural conversational speech corpus.

2. The Japanese Dialogue Corpus ESP_C

Our dialogue corpus ESP_C[7] is a subset of the JST/ATR ESP Corpus constituted by the 2000 to 2005 JST/CREST Expressive Speech Processing Project [?]. This section of the corpus comprises a set of high-quality recordings of 2-person telephone dialogues in Japanese. Participants are 10 male and female adults (4 of whom are not native Japanese speakers). Each session lasted for 30 minutes and usually 10 sessions were recorded for each pair (5 sessions each if the partner is a foreign speaker of Japanese). Participants were not given any instructions as to the content of the speech, other than the 30-minute time requirement for speaking, so the conversations are content-free and develop naturally as the partners become more familiar with each other over the period of the recordings. It is a corpus of very free dialogues and includes many non-linguistic interpersonal speech events, including fillers, laughs and disfluencies. All the conversations have been manually transcripted by human experts. An example ESP_C transcription for the first conversation of the series is shown in Fig. 1.

Several large scale dialogue corpora have already been made widely available, such as the CALLHOME[4] and Switchboard Corpora[5]. However, Switchboard is not free-content dialogue, so is less suitable for our purpose of understanding the human interactional element in natural conversational speech. In general, free-content (no-purpose) dialogues tend to include many more incomplete utterances; more so than written texts, monologues and dialogues produced with some specific purpose or goal. We selected the ESP_C natural dialogue corpus of conversational speech for this study because it allows us to research the different ways that people interact with each other through speech. It includes much more casual chat than CALLHOME and Switchboard.

```
                    :
JMA_JMB_J01 1.132 0.543 どうも (hello)
JMB_JMA_J01 2.094 0.425 ほ (ho)
JMA_JMB_J01 2.150 1.913 ハハハ (hahaha)
JMB_JMA_J01 2.528 1.809 今から始めるみたい
です (so it begins from now)
JMA_JMB_J01 4.101 0.937 あーほんまです
か (ah really?)
JMB_JMA_J01 4.936 0.421 はい (yes)
JMA_JMB_J01 5.510 1.840 もう普通にしゃべっ
とったらいいんですかねこれね
(and can we speak normally then?)
JMB_JMA_J01 7.458 1.079 あーたぶんなんか
(ah maybe, well ...)
JMB_JMA_J01 8.834 2.471 それ内容はそんな重
視しないってさっきも言ってたんで
(the contents are not so
important they said just now)
JMA_JMB_J01 10.526 0.648 あっ (ah)
JMA_JMB_J01 11.218 0.764 そうなんです
か (really?)
JMB_JMA_J01 11.776 0.890 はい (yes)
JMB_JMA_J01 13.054 2.656 普通に話ししとった
らいいみたいですけど
(we can just speak normally then)
JMA_JMB_J01 13.606 0.100 ん (n)
JMA_JMB_J01 14.845 0.837 あーー (ahh)
                    :
```

Figure 1: An example of ESP_C transcription, showing from left to right file id (speaker, partner, and conversation number), start time of the utterance,its duration, the Japanese transcription and a gloss in English.

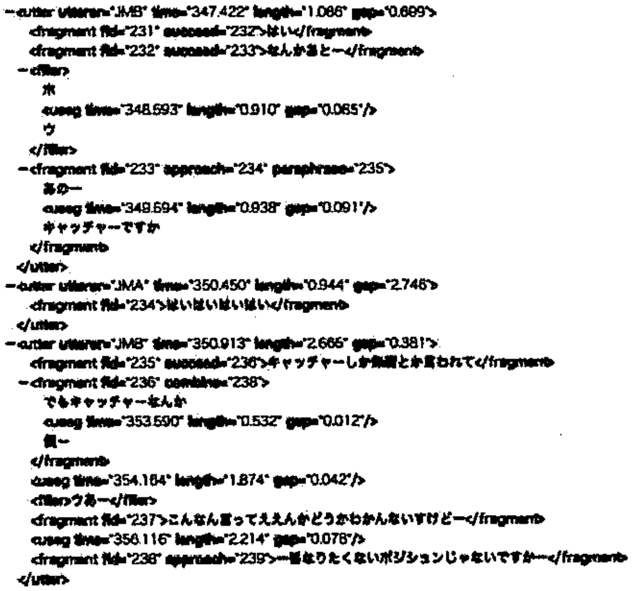
3. Tags and Tagging Specification

In this section, we describe our proposed tag set. It consists of definitions for ddetermining discourse fragments and presents 5 simple relationship attributes.

3.1. Tagging Policy

Human dialogues consist of multiple streams of events generated from factors which are not always fully represented on the surface, such as world knowledge, social and community factors, discourse management, interpersonal affect display, speaker emotion, etc. However, it is hard for a computational process to understand (i.e., recognise or model) such events. Thus, we need to annotate tags to the corpus by representing the surface information of the speech and its sounds through transcriptions without any deep understanding of the contents. We can't depend on morphological and syntactic analysis produced by current natural-language processing algorithms, because their performance is not at all good when confronted with transcriptions of natural unprompted speech.

In this paper, the utterance fragments are annotated automatically according to heuristics determined by trial and error, and are checked subsequently for human modification. Relationship attributes are each determined by human annotators. The final tagged corpus is represented as an XML document (see Fig. 2).



```
-<utter utterer="JMB" time="347.422" length="1.086" gap="0.699">
   <fragment fid="231" succeed="232">はい</fragment>
   <fragment fid="232" succeed="233">なんかあと－</fragment>
  -<filler>
     ホ
     <useg time="348.693" length="0.910" gap="0.065"/>
     ウ
   </filler>
  -<fragment fid="233" approach="234" paraphrase="235">
     あの－
     <useg time="349.694" length="0.938" gap="0.091"/>
     キャッチャーですか
   </fragment>
  </utter>
-<utter utterer="JMA" time="350.450" length="0.944" gap="2.746">
   <fragment fid="234">はいはいはいはい</fragment>
  </utter>
-<utter utterer="JMB" time="350.913" length="2.665" gap="0.381">
   <fragment fid="235" succeed="236">キャッチャーしか無理とか言われて</fragment>
  -<fragment fid="236" combine="238">
     でもキャッチャーなんか
     <useg time="353.590" length="0.532" gap="0.012"/>
     嫌－
   </fragment>
   <useg time="354.184" length="1.874" gap="0.042"/>
   <filler>うあ－</filler>
   <fragment fid="237">こんなん言ってええんかどうかわかんないすけど－</fragment>
   <useg time="356.116" length="2.214" gap="0.078"/>
   <fragment fid="238" approach="239">一番なりたくないポジションじゃないですか－</fragment>
  </utter>
```

Figure 2: The tagged corpus (XML style)

3.2. Tagging Specification

3.2.1. utterance units by pause

First, we roughly define utterance units by pauses in the flow of speech. This is the most objective way to determine the units. The tag name is *utter*. If a pause longer than 300 ms occurs in the acoustic signal from each speaker, we assume a prosodic word in the conversation. This threshold duration was determined by examining results of several candidate thresholds. The segment of speech transcribed between each such pause is considered as one minimal unit of the utterance.

3.2.2. filler, laugh and disfluency

In a dialogue, in particular free-domain chat, there are many utterance events which do not convey linguistic information such as fillers, laughs and disfluencies. In our tagging philosophy, we maintain that the content of a dialogue is changed by such events, and that the meaning of an utterance after such events doesn't identify with a meaning of that before the events. In other wowrds, we think that such events are important clues for structuring dialogues, thus, we define the tags *filler, laugh* and *disfluency*.

Studies of auto extraction of fillers in speech has been performed[8]. However, for this paper, we use hand-crafted pattern matching and human modification. Automatic detection of fillers and disfluencies in the transcriptions of conversatioal speech remains as a key theme of ongoing research.

3.2.3. fragment

The *fragment* is the smallest unit of our study. In our utterances, one meaning unit is constructed from one or more than one fragments. Tag *fragment* is tagged according to the following rules:

- A fragment does not include a pause longer than 300ms and contains at most one *utterance unit by pause*.
- non-linguistic information.

In an utterance unit, if there is *filler*, *laugh* or *disfluency*, this point is considered a bounday of a fragment unit.

- grammatical rules.
  if there is a auxiliary verb "ですdesu", 終助詞or conjunctive auxiliary, this point is considered a fragment boundary. However, inversions, which are common in dialogues, are added to the front unit. For example, in the utterance "負けずぎらいっていうのはよくいわれーましたね中学んときとか*makezugirai tte iu no ha yoku iwaree mashita ne chuugaku n toki toka (I was often said to hate to lose, in juniour high school days*", there is 終助詞"ね*ne*" in front of "中学んときとか*chuugaku n toki toka (in junior high school days)*". However, this utterance is considered one fragment as "中学のときとか*chuugaku n toki toka (in junior high school days)*" is an inversion.

3.2.4. *Relationship Attributes*

We assign each fragment attributes toward other related fragments. They take a role as a pointer from one fragment to the others.

For utterances in real conversational dialogues it is perhaps too complex to include all the attributes of relationships. So, in this study, we limit our definition to the main 5 relationship attributes. They can be tagged by surface information without deep understanding of the contents. Now, our relationship attributes are not defined for all attributes, and we retain the right to redefine them as necessary and to define new ones as the need arises in this corpus research. This is a continuing and future work.

The relationship attributes are as follows:

combine This relationship combines 2 fragments into one meaning unit, separated by another speaker's utterance, fillers or other reasons.

approach This relationship signals that the speaker is approaching another speaker for reaction (eg. asking questions or confidence), this tag ties those 2 fragments between speakers.

refer a fragment refers back in the past to other fragments with no approach especially.

succeed A speaker refer to his or her own utterance and speaks continuously, embellishing it.

paraphrase A speaker paraphrases or repeats a previous utterance.

In this paper, we limit our research to the above 5 attributes. However, there are of course need for more attributes to cover all relationships between fragments in dialogues; there are pairs of fragments which are not represented by our 5 relationship attributes. We discuss this problem in future works.

3.3. A Dialogue Graph

By tagging in this way, we can represent dialogues as directed graphs, fragments become nodes, and relationships edges. An example is presented in Fig. 3.

3.4. Layered Structure and Semantemes of Dialogues

In tagging dialogue structure, we must consider 2 categories of fragment relationship; one is *intra-relationship*, and other *inter-relationship*. Intra-relationship concerns relations between the speech fragments of any given speaker, while inter-relationships concern the relations between a fragment from one speaker and another or others from his or her conversational partner.

In our 5 defined relationships, *combine*, *refer*, *succeed* and *paraphrase* can be intra-relationship categories. While, *approach* and *refer* can be inter-relationship; *refer* can be both.

Furthermore, we work on the assumption that there are *communication layers* and *content layers* in a dialogues. In a communication layer, utterances are connected to those of the other person, and the facilitation of communication is made. In a content layer, afective information and discourse control management takes place, and the supply information is provided by each speaker. Thus, we can understand the flows of a dialogue by only processing content layers. And, we can produce semantemes, semantic units of a dialogues, by relating each chunk of fragments. In Fig. 3, we show an example of the two layers.

Fragments #1, #4, #5 and #9 are included in the content layer. While, *fragments #2, #3, #6, #7 and #8* are in the communication layer. In #2 and #6-7, Speaker A approaches towards Speaker B, and, in #3 and #8, Speaker B reacts to these approaches; *#2 "キャッチャーですかkyachaa desu ka (Is it a catcher)" / #3 "はいはいはいはい, hai hai hai hai (yeah yeah)". #6-7 "僕ー/ こんなん言ってええんかわかんないですけどboku / kon'nan itte een ka wakan'nai desu kedo (I / don't know I can say I like it)" / #8 "はいhai (yes)"*. While, in #1, #4, #5 and #9, Speaker A produces fragments of significant content. We can see the summarization of this part of the dialogue; *"なんかあとー/キャッチャーしか無理といわれて/でもキャッチャーなんか/ー番なりたくないポジションじゃないですかーnanka ato / kyacchaa shika muri to iwarete/ demo kyacchaa nannkaa/ ichiban naritaku nai pojishon ja naidesukaa (And/ I was told I couldn't be anything but a catcher/ but a catcher is / the most unwanted position, isn't it)"* .

4. Tagging and Discussion

In this section, we report the result of tagging the corpus dialogues using our tags as defined in the previous section. We tagged 2 male Japanese native speakers' 30-minute dialogues in ESP_C. We show the occurrence counts of each tag in Tab. 1 and Tab. 2. In 1, the counts of *utter,fragment* and non-linguistic events are shown. In 2, the counts of each relationship attributes are shown.

In Tab. 1, the figures for *utter* and *fragment* between the speakers are nearly equal. However, those of *filler* and *laugh* are very different. This perhaps reflects differences in each speaker's personal factor and mental state. Future work will build on this point for processing affective information in spoken dialogues.

From Tab. 2, we can see equality and differences in the number of attributes. Each speaker has attributes *combine* and *refer* in equal proportions. However, other attributes, in particular *approach* and *succeed*, are different. Speaker B has twice as many *approach* attributes than Speaker A, however, only half as many *succeed* attributes. From this, we can see that in this dialogue, Speaker B takes the dominant role and tends to ask more questions, while Speaker A mainly just responds to these approaches. As described, from the numbers of relationship attributes, we can make an estimate of the tendencies of the dialogues as follows roughly; *How active the dialogues are, Which speaker has the initiative, What the state of mind each speaker is in*, and so on.

By tagging structures and relationships in this way, we can quantify the states and overall characteristics of the dialogues and the affective states and discourse roles of the speakers. In

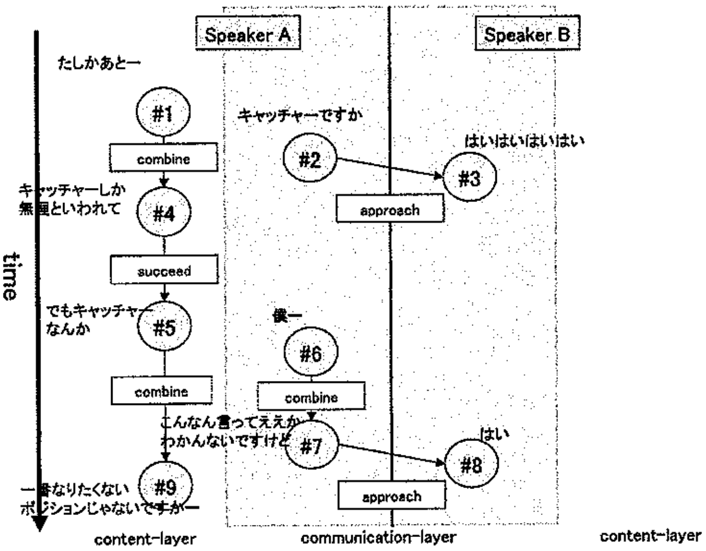


Figure 3: A part of dialogue graph and 2 layers..

this paper, we described only one session. We will obtain new aspects from further investigations, both more finely (quantifying local events) and more widely (quantifying events between sessions or speakers). This remains as ongoing and future work.

Table 1: Counts of utterance units, fragments and non-linguistic events in one session.

Spkr	utter	fragment	filler	laugh	disfluency
A	414	606	80	114	37
B	439	564	227	168	44
Sum.	853	1169	357	283	81

Table 2: Counts of relationship attributes in one session.

Spkr	combine	approach	refer	succeed	paraphrase
A	52	47	88	202	42
B	67	100	105	85	26
Sum.	119	147	193	287	68

5. Conclusion and Future Works

The purpose of this study was to propose a method for tagging fragments in dialogues, to determine their structures and relationships to the various dialogue elements. We defined utterance units, fragments of utterances and 5 relationship attributes. We then tagged a section of the ESP_C telephone-speech conversational corpus. In the present work, we have described 2 parallel and perhaps independent layers of communication in the dialogues, *communication-layer* and *content-layer*.

There are three major tasks to be carried out in the future. At present our annotation is manual and expensive so we cannot easily deal with large amounts of data. We are therefore building and training a semi-automatic annotation tool for this task. As described in Section 3.2.4, our relation attributes are still not complete. There is a need to define further new attributes and for the definitions of the existing attributes to be fixed. Thirdly, we need to investigated the validity of this tagging by means of multi-annotators.

6. References

[1] Araki, M., Ueda, K., Nishimoto, T. and Niimi, Y. "A semantic tagging tool for spoken dialogue corpus", Proc. 6th Int'l Conf. on Spoken Language Processing, Vol. 4, pp. 720–723, 2000.

[2] Allen, J.F. and Core, M. Draft of DAMSL: Dialogue act markup in several layers. http://www.cs.rochester.edu/research/trains/annotation.

[3] Samuel, K., Carberry, S. and Vijay-Shanker, K. "Dialogue act tagging with transformation-based learning". Proc. COLING-ACL, pp. 1150–1156, 1998.

[4] Wheatley, B., Kaneko, M., and Kobayashi, M. CALLHOME Japanese Transcripts. Linguistic Data Consortium, Philadelphia, 1996.

[5] Godfrey, J.J., Holliman, E.C., and McDaniel, J., "SWITCHBOARD: Telephone speech corpus for research and development" In Proc. IEEE-ICASSP, Vol. 1, pp. 517–520, 1992.

[7] Campbell. N., Selecting speech fragments for affect display in concatenative expressive speech synthesis" 2007 Spring Meeting of ASJ, 2007.

[8] Asahara, M. and Matsumoto, Y. "Filler and disfluency identification based on morphological analysis and chunking" In Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp. 163–166, 2003.