Japanese Dialogue Corpus of Multi-Level Annotation

The Japanese Discourse Research Initiative
http://www.slp.cs.ritsumei.ac.jp/dtag/

Abstract

This paper describes a Japanese dialogue corpus annotated with multi-level information built by the Japanese Discourse Research Initiative, Japanese Society for Artificial Intelligence. The annotation information consists of speech, transcription delimited by slash units, prosodic, part of speech, dialogue acts and dialogue segmentation. In the project, we used the corpus for obtaining new findings by examining the relationship between linguistic information and dialogue acts, that between prosodic information and dialogue segment, and the characteristics of agreement/disagreement expressions and non-sentence elements.

1 Introduction

This paper describes a Japanese dialogue corpus annotated with multi-level information such as speech, linguistic and discourse information built by the Japanese Discourse Research Initiative, supported by Japanese Society for Artificial Intelligence.

Dialogue corpora are now indispensable to speech and language research communities. The corpora have been used not only for examining the relationship between speech and linguistic phenomena, but also for building speech and language understanding systems.

Sharing corpora among researchers is most desirable since creating the corpora needs considerable cost like writing and revising annotation manuals, annotating the data, and checking the consistency and reliability of the annotated data. Discourse Research Initiative was set up in March of 1996 by US, European, and Japanese researchers to develop standardized discourse annotation schemes (Carletta et al., 1997; Core et al., 1998).

The efforts of the initiative have been called 'standardization', but this naming is misleading at least. In typical standardizing efforts, as done in audio-visual and telecommunication technologies, commercial companies try to expand the market for their products or interfaces by the standard. The objective of standardizing efforts in discourse is to promote interactions among discourse researchers and thereby provide a solid foundation for corpus-based discourse research, dispensing with duplicating resource making efforts and increasing sharable resources.

In cooperation with this initiative, Japanese Discourse Research Initiative has started in Japan in May 1996, supported by Japanese Society for Artificial Intelligence (JDRI, 1996; Ichikawa et al., 1999). The activities of the initiative involve:

- creating and revising annotation schemes based on the survey of the existing schemes and annotation experiments,
- annotating corpora based on the proposed annotation schemes, and
- doing research using the corpora not only for examining the utility of the schemes and corpora but also for obtaining new findings.

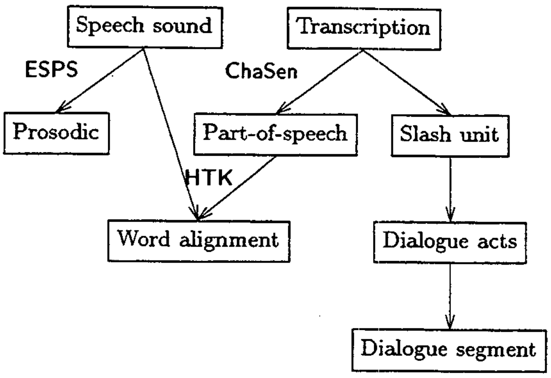

Figure 1: The relations among the annotation information

In the following, a Japanese dialogue corpus of multi-level annotation is demonstrated. The annotation schemes deal with the information for speech, transcription segmented by utterance units, called 'slash units,' prosody, part of speech, dialogue acts and dialogue segment. Figure 1 shows the relations among the annotation information.

2 Speech Sound and Transcription

The corpus consists of a collection of 14 task-oriented dialogues, each performed by two native speakers of Japanese. The total time of the 14 dialogues is 53 minutes. The tasks include scheduling, route guidance, telephone shopping, and so on. We set the roles of the two speakers and the goal of the task but no pre-defined scenarios. For example, in the scheduling task, the speakers were given the roles of a private secretary and a client, and asked to arrange a meeing appointment.

The speech sound of the two speakers participating in a dialogue was recorded on separate channels, which enables us to perform accurate acoustic/prosodic analysis even for overlapped talks. The transcription contains orthographic representations in Kanji and the starting and ending time of each utterance, where an utterance is defined as a continuous speech region delimited by pauses of 400 msec or longer.

3 Prosodic Information and Part-of-speech

The prosodic information and the part-of-speech tags were assigned (semi-)automatically using the speech sound and the transcription.

3.1 Prosodic information

Prosody has been widely recognized as one of the important factors which relate to discourse structures, dialogue acts, information status, and so on. Informative corpora should, in the first place, contain some form of prosodic information.

At this stage, our corpus merely includes, as prosodic information, raw values of fundamental frequency, voicing probability, and rms energy, which were obtained from the speech sound using speech analysis software ESPS/waves+ (Entropic, 1996) and simple post-processing for smoothing. The future corpus will contain more abstract descriptions of prosodic events such as accents and boundary tones.

3.2 Part-of-speech

The part-of-speech is another basic information for speech recognition, syntactic/semantic parsing, and dialogue processing as well as linguistic and psycholinguistic analysis of spoken discourse.

Part-of-speech tags were, first, obtained automatically from the transcription using the morphological analysis system ChaSen (Matsumoto et al., 1999), and, then, corrected manually. The tag set was extended to cover filled pauses and contracted forms peculiar to spontaneous speech, and some dialects. The tagged corpus will be used as a part of the training data for the statistical learning module of ChaSen to improve its performance for spontaneous speech, which can be used for future applications.

3.3 Word alignment

In some applications such as co-reference resolution utilizing prosodic correlates of given-new status of words, it is useful to know the prosodic information of particular words or phrases. In order to obtain such information, the correspondence between the word sequence and the speech sound must be given. Our corpus contains the information for the starting and the ending time of every word.

The time-stamp of each word in an utterance was obtained automatically from the speech sound and the part-of-speech using the forced alignment function of speech recognition software HTK (Odell et al., 1997) with the tri-phone model for Japanese speech developed by the IPA dictation software project (Shikano et al., 1998). Apparent errors were corrected manually with reference to sound waveforms and spectrograms obtained and displayed on a screen by ESPS/waves+ (Entropic, 1996).

4 Utterance Units

4.1 Slash units

In the transcription, an utterance is defined as a continuous speech region delimited by pauses of 400 msec or longer. However, this definition of the utterances does not correspond to the units for discourse annotation. For example, the utterances are sometimes interrupted by the partner. For reliable discourse annotation, analysis units must be constructed from the utterances defined above. Following Meteer and Taylor (1995), we call such a unit 'slash unit.'

4.2 Criteria for determining slash units

The criteria for determining slash units in Japanese were defined with reference to those for English (Meteer and Taylor, 1995). The slash units were annotated manually with reference to the speech sound and transcription of dialogues.

Single utterances as slash unit Single utterances which can be thought to represent sentences conceptually are qualified as a slash unit. Figure 2 shows examples of slash units by single utterances (slash units are delimited by the symbol '/').

```
A:  hai / ;{response}
 (yes.)

A:  kochira chiri annai sisutemu desu /
    ;{a single sentence}
 (This is the sightseeing guide
  system.)

A:  ryoukin niha fukumarete orimasen ga
  betto 1200 en de goyoui sasete
  itadakimasu /
    ;{a complex sentence}
(This is not included in the charge.
 We offer the service for
 the separate charge of 1200 yen.)
```

Figure 2: Examples of single utterances as slash unit

```
A:  shuppatsu chiten kara --
  (From the starting point)
A:  --  nishi gawa ni --
  (to the west)
A:  --  sukoshi dake ikimasu /
  (move a little)
```

Figure 3: An example of multiple utterances as slash unit

In the cases where the word order is inverted, the utterances are regarded as a slash unit only if the utterances with normalized word order are qualified as a slash unit.

A sequence of one speaker's speech that terminates with a hesitation, an interruption and a slip of the tongue, but does not continue in the speaker's next utterance is also qualified as a slash unit.

Multiple utterances as slash unit When collection of multiple utterances form a sentence, as in Figure 3, they are qualified as one slash unit. In slash units spanning multiple utterances, the symbol '--' is marked both at the end of the first utterance and at the start of the last utterance.

4.3 Non sentence elements

Non sentence elements consist of 'aiduti', conjunction markers, discourse markers, fillers

A: *sukoshi dake itte* /
(move a little)

B: *un* /
(ok)

A: {D *de*} *hidari naname shitani*
({D then} to your left and down)

Figure 4: An example of a slash unit defined by discourse markers

and non speech elements, which are enclosed by {B ...}, {C ...}, {D ...}, {F ...}, and {N ...}, respectively. These elements can be used to define a slash unit. For example, when 'aiduti' is expressed by the words such as "hai (yes, yeah, right)", "un (yes, yeah, right)" and "ee (mmm, yeah)" or by word repetition, it is regarded as an utterance. Otherwise, 'aiduti' is not qualified as an independent slash unit.

The main function of discourse markers is to show the relations between utterances, like starting a new topic, changing topics, and restarting an interrupted conversation. The words such as "mazu (first, firstly)", "dewa (then, ok)", "tsumari (I mean, that means that)" and "sorede (and so)" may become discourse markers when they appear at the head of the utterances. An utterance just before the one with discourse markers is qualified as a slash unit (Figure 4).

In the Switchboard project(Meteer and Taylor, 1995), our {B ...} (aiduti) category is not regarded as a separate category. However in Japanese dialogue, signals that indicate a hearer's attention to speaker's utterances, are expressed frequently. For this reason, we created 'aiduti' as a separate category. Otherwise {A ...}(aside), {E ...}(Explict editing term), the restart and the repair are not annotated in our scheme at the present stage.

5 Dialogue Acts

Identifying dialogue act of the slash unit is difficult task because the mapping between surface form and dialogue act is not obvious. In addition, some slash units have more than one function, e.g. answering question with stating additional information. Considering above problems, DAMSL architecture codes various functions at one utterance, such as forward looking function, backward looking function, etc.

However, it is difficult to determine the function of the isolated utterance. We had shown that assumptions of dialogue structure and exchange structure improved agreement score among coders (Ichikawa et al., 1999). Therefore, we define our dialogue act tagging scheme as hierarchical refinement from the exchange structure.

The annotation scheme for dialogue acts includes a set of rules to identify the function of each slash unit based on the theory of speech act (Searle, 1969) and discourse analysis (Coulthhard, 1992; Stenström, 1994). This scheme provides a basis for examining the local structure of dialogues.

- Task-oriented dialogue
  (Opening)
  Problem solving
  (Closing)
- Problem solving
  Exchange$^+$
- Exchange
  Initiation
  (Response)/Initiation*
  (Response)*
  (Follow-up)
  (Follow-up)

Figure 5: Model for task-oriented dialogues

In general, a dialogue [1] is modeled with problem solving subdialogues, sometimes preceded by opening subdialogue (e.g., greeting) and followed by closing subdialogue (e.g., expressing gratitude). A problem solving subdialogue consists of initiating and responding

[1] In this paper, we limit our attention to task-oriented dialogues, which are the main target of the study in computational linguistics and spoken dialogue research.

(Initiation)
41 A: *chikatetsu no ekimei ha?*
(What's the name of the subway station?)

(Response)
42 B: *chikatetsu no teramachi eki ni narimasu*
(The name of the subway station is Teramachi.)

(Follow-up)
43 A: *hai*
(Ok.)

Figure 6: An example problem solving subdialogue with the exchange structure

utterances, sometimes followed by following up utterances (Figure 5).

Figure 6 shows an example problem solving subdialogue with the exchange structure.

In this scheme, dialogue acts, the elements of the exchange structure, are classified into the tags shown in Figure 7.

- Dialogue management
  Open, *Close*
- Initiation
  Request, *Suggest*, *Persuade*, *Propose*, *Confirm*, *Yes-no question*, *Wh-question*, *Promise*, *Demand*, *Inform*, *Other assert*, *Other initiate.*
- Response
  Positive, *Negative*, *Answer*, *Other response.*
- Follow-up
  Understand
- Response with Initiation
  The element of this category is represented as *Response type / Initiation type.*

Figure 7: The classification of dialogue acts

6 Dialogue Structure and Constraints on Multiple Exchanges

6.1 Dialogue Segment

In the previous discourse model(Grosz and Sidner, 1986), a discourse segment has a beginning and an ending utterances and may have smaller discourse segments in it. It is not an easy task to identify such segments with the nesting structure for spoken dialogues, because the structure of a dialogue is often very complicated due to the interaction of two speakers. In a preliminary experiment of coding segments in spoken dialogues, there were a lot of disagreements on the granularity or the relation of the segments and on identifying ending utterances of the segment. An alternative scheme of coding the dialogue structure (DS) is necessary to build dialogue corpora annotated with the discourse level structure.

Our scheme annotates spoken dialogues with boundary marking of the DS, instead of identifying a beginning and an ending utterance of each DS. A building block of dialogue segments is identified based on the exchanges explained in Section 5. A dialogue segment (DS) tag is inserted before initiating utterances because the initiating utterances can be thought of as a start of new discourse segments.

The DS tag consists of a topic break index (TBI), a topic name and a segment relation. TBI signifies the degree of topic dissimilarity between the DSs. TBI takes the value of 1 or 2: the boundary with TBI 2 is less continuous than the one with TBI 1 with regard to the topic. The topic name is labeled by coders' subjective judgment. The segment relation indicates the one between the preceding and the following segments, which is classified into the following categories.

- *clarification*

  suspends the exchange and makes a clarification in order to obtain information necessary to answer the partner's utterance;

[2: room for a lecture:]
38 A: {F *e*} *heya wa dou simashou ka?*
(How about meeting room?)

[1: small-sized meeting room: clarification]
39 B: *heya wa shou-kaigishitsu wa aite masu ka?*
(Can I use the small-sized meeting room?)

40 A: {F *to*} *kayoubi no* {F *e*} *14 ji han kara wa* {F *e*} *shou-kaigisitsu wa aite imasen*
(The small meeting room is not available from 14:30 on Tuesday.)

[1:the large-sized meeting room:]
41 A: *dai-kaigishitsu ga tukae masu*
(You can use the large meeting room.)

[1: room for a lecture: return]
42 B: {D *soreja*} *dai-kaigishitsu de onegai shimasu*
(Ok. Please book the large meeting room.)

Figure 8: An example dialogue with the dialogue segment tags

- *interruption*

  starts a different topic from the previous one during or after the partner's explanatory utterances; and

- *return*

  goes back to the previous topic after the clarification or the interruption.

Figure 8 shows an example dialogue annotated with the DS tags.

6.2 Constraints on multiple exchanges

Annotation of dialogue segments mostly depends on the coders' intuitive judgment on topic dissimilarity between the segments. In order to lighten the burden of the coders' judgment, the structural constraints on multiple exchanges are experimentally introduced.

The constraints can be classified into two types: one concerns embedding exchanges (*relevance type 1*) and the other is neighboring exchanges (*relevance type 2*).

In *relevance type 1*, the relation of an initiating utterance and its responding utterance is shown by attaching the id number of the initiating utterance to the responding utterance. This id number can indicates non-adjacent initiation-response pairs including embedded exchanges inside.

In *relevance type 2*, the structures of neighboring exchanges such as chaining, coupling, elliptical coupling (Stenström, 1994) are introduced. Chaining takes the pattern of [A:I B:R] [A:I B:R] (in both exchanges, speaker A initiates an utterance and speaker B responds to A). Coupling is the pattern of [A:I B:R] [B:I A:R]. (speaker A initiates, speaker B both responds and initiates and speaker A responds to B). Elliptical coupling is the pattern of [A:I] [B:I A:R], equivalent to the one in which B's second response is omitted in coupling. *Relevance type 2* shows whether the above structures of neighboring exchanges can be observed or not. Figure 9 shows an example of annotation of relevance types 1 and 2.

7 Corpus Building Tools

In the experiments, various tools for transcription and annotation were used. For transcription, the automatics segmentizer (TIME) and the online transcriber (PV) were used (Horiuchi et al., 1999). The former lists up

[<Yes-no question> <relevance no>]
27 A: *hatsuka no jyuuji kara ha aite irun de syou ka?*
(Is the room available from 10am on the 20th?)

[<Yes-no question> <relevance yes>]
28 B: *kousyuu shitsu desu ka?*
(Are you mentioning the seminar room?)

[<Positive> <0028>]
29 A: *hai*
(Yes.)

[<Negative> <0027>]
30 B: *hatsuka ha aite orimasen*
(It is not available on the 20th.)

[<Understand>]
31 A: *soudesu ka*
(Ok.)

Figure 9: An example dialogue with relevance types 1 and 2

candidates for unit utterances according to the parameter for the length of silences. The latter displays energy measurement of each speaker's utterance on the two windows using a speech data file. Users can see any part of a dialogue using the scroll bar, and can hear speech for both speakers or each speaker by selecting any region of the windows using a mouse.

For prosodic and part of speech annotation, the speech analysis software ESPS/waves+ (Entropic, 1996), speech recognition software HTK (Odell et al., 1997) with the tri-phone model for Japanese speech developed by the IPA dictation software project (Shikano et al., 1998) and the morphological analysis system ChaSen (Matsumoto et al., 1999) were used.

For discourse annotation, Dialogue Annotation Tool (DAT) had been used in the previous experiments (Core and Allen, 1997). Although DAT had a consistency check between some functions in one sentence, we need more wide-ranging consistency check because our scheme has assumptions of dialogue structure and exchange structure. Therefore it is dissatisfying but the modification of the tool to our need is not easy. Thus, for the moment, we decided to use just a simple transcription viewer and sound player (TV) (Horiuchi et al., 1999), which enables us to hear the sound of utterances on the transcription.

Our project does not have any intention to create new tools. Rather we do want to use any existing tools if they suit or can be easily modified to satisfy our needs. The tools of MATE project(Carletta and Isard, 1999), which also directs multi-level annotation, can be a good candidate for our project. In the near future, we will examine if we can effectively use their tools in the project.

8 Conclusion

This paper described a Japanese dialogue corpus annotated with multi-level information built by the Japanese Discourse Research Initiative supported by Japanese Society for Artificial Intelligence. The annotation information includes speech, transcription delimited by slash units, prosodic, part of speech, dialogue acts and dialogue segmentation. In the project (JSAI, 2000), we used the corpus for obtaining new findings by examining:

- the relationship between linguistic information and dialogue acts
- the relationship between prosodic information and dialogue segment, and
- the characteristics of agreement/disagreement expressions and non-sentence elements.

This year we plan to quadruple the size of the corpus and make it publicly available as soon as we finish the annotation and its verification.

References

J. Carletta and A. Isard. 1999. The MATE Annotation Workbench: User Requirements. In *The Proceedings of the ACL'99 Workshop on Towards Standards and Tools for Discourse Tagging*, pages 11–17.

J. Carletta, N. Dahlback, N. Reithinger, and M. A. Walker. 1997. Standards for Dialogue Coding in Natural Language Processing. ftp://ftp.cs.uni-sb.de/pub/dagstuhl/reporte/97/9706.ps.gz.

M. Core and J. Allen. 1997. Coding Dialogues with the DAMSL Annotation Scheme. In *The Proceedings of AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35.

M. Core, M. Ishizaki, J. Moore, C. Nakatani, N. Reithinger, D. Traum, and S. Tutiya. 1998. The Report of the Third Workshop of the Discourse Research Initiative, Chiba Corpus Project. Technical Report 3, Chiba University.

M. Coulthhard, editor. 1992. *Advances in Spoken Discourse Analysis*. Routledge.

Entropic Research Laboratory, Inc. 1996. *ESPS/waves+ 5.1.1 Reference Guide.*

Grosz, B. J. and Sidner, C. L. 1986. Attention, Intentions, and the Structure of Discourse, *Computational Linguistics*, 12(3), pages 175–204.

Y. Horiuchi, Y. Nakano, H. Koiso, M. Ishizaki, H. Suzuki, M. Okada, M. Makiko, S. Tutiya, and A. Ichikawa. 1999. The Design and Statistical Characterization of the Japanese Map Task Dialogue Corpus. *Japanese Society of Artificial Intelligence*, 14(2).

A. Ichikawa, M. Araki, Y. Horiuchi., M. Ishizaki, S. Itabashi, T. Itoh, H. Kashioka, K. Kato, H. Kikuchi, H. Koiso, T. Kumagai, A. Kurematsu, K. Maekawa, S. Nakazato, M. Tamoto, S. Tutiya, Y. Yamashita, and T. Yoshimura. 1999. Evaluation of Annotation Schemes for Japanese Discourse. In *Proceedings of ACL'99 Workshop on Towards Standards and Tools for Discourse Tagging*, pages 26–34.

Japanese Discourse Research Initiative. http://www.slp.cs.ritsumei.ac.jp/dtag/.

Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. Japanese morphological analysis system ChaSen version 2.0 manual (2nd edition). 1999. Technical Report NAIST-IS-TR99012, Graduate School of Information Science, Nara Institute of Science and Technology. http://cl.aist-nara.ac.jp/lab/nlt/chasen/manual2/manual.pdf.

M. Meteer and A. Taylor. 1995. Dysfluency Annotation Stylebook for the Switchboard Corpus. ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps.gz.

Japanese Society for Artificial Intelligence. 2000. Technical Report of SIG on Spoken Language Understanding and Dialogue Processing. SIG-SLUD-9903.

J. Odell, D. Ollason, V. Valtchev, and P. Woodland. 1997. *The HTK Book (for HTK Version 2.1).* Cambridge University

J. R. Searle. 1969. *Speech Acts: An Essay in the Philosopy of Language.* Cambridge University Press.

K. Shikano, T. Kawahara, K. Ito, K. Takeda, A. Yamada, T. Utsuro, T. Kobayashi, N. Minematsu, and M. Yamamoto. 1998. *The Development of Basic Softwares for the Dictation of Japanese Speech: Research Report 1998.*

A. B. Stenström. 1994. *An Introduction to Spoken Interaction.* Longman.