Evaluation of Annotation Schemes for Japanese Discourse

Japanese Discourse Tagging Working Group
Ichikawa, A. (Chiba U.), Araki, M. (KIT), Horiuchi, Y. (Chiba U.),
Ishizaki, M. (JAIST), Itabashi, S. (Tsukuba U.), Itoh, T. (Shizuoka U.),
Kashioka, H. (ATR-ITL), Kato, K. (Tsukuba U.), Kikuchi, H. (Waseda U.),
Koiso, H. (NLRI), Kumagai, T. (NLRI), Kurematsu, A. (UEC),
Maekawa, K. (NLRI), Nakazato, S. (Meio U.), Tamoto, M. (NTT BRL),
Tutiya,S. (Chiba U.), Yamashita,Y. (Ritsumeikan U.) and Yoshimura,T. (ETL)

Abstract

This paper describes standardizing discourse annotation schemes for Japanese and evaluates the reliability of these schemes. We propose three schemes, that is, utterance unit, discourse segment and discourse markers. These schemes have shown to be incrementally improved based on the experimental results, and the reliability of these schemes are estimated as "good" range.

1 Introduction

Linguistic corpora are now indispensable of speech and language research communities. They are used not only for examining their characteristics, but also for (semi-)automatically learning rules for speech recognition, parsing and anaphora resolution, and evaluating the performance of speech and natural language processing systems.

Linguistic corpora can be used as they are, however, they are usually annotated with information such as part of speech and syntactic structures. Currently there are many large linguistic annotated corpora worldwide, but the types of annotation information are limited to morphological and syntactic information. While there are some corpora annotated with discourse information like speech act types and discourse structures, they are much smaller than that of the corpora with morphological and syntactic information. One of the major reasons for this difference in the size is due to the lack of computer tools such as morphological analyzers and syntactic parsers to semi-automatically annotate information. Of course we will be able to develop such tools for discourse information, but before that, we must create a base corpora by setting standards [1] for resource sharing, which can contribute to creating large resources for discourse.

To this end, the Discourse Research Initiative (DRI) was set up in March of 1996 by US, European, and Japanese researchers to develop standard discourse annotation schemes (Walker et al., 1996). In line with the effort of this initiative, a discourse tagging working group has started in Japan in May 1996, with the support of the Japanese Society of Artificial Intelligence. The working group consists of representatives from eight universities and four research institutes in Japan. In the first year, (1) we collected and analyzed existing annotation schemes for Japanese discourse from the viewpoints of annotation units and information types, (2) developed new annotation schemes and experimentally annotated actual data, and (3) analyzed the experimental results to im-

[1] The efforts have been called 'standardization', but we must admit this naming is misleading at least. In typical standardizing efforts, as done in audio-visual and telecommunication technologies, companies try to expand the market for their products by making their products or interfaces standards, and this profit directedness leaves room for negotiation. Even if the negotiation fails, they can appeal their products or interfaces for the market to judge. The objective of standardizing efforts in discourse is to promote interactions among different discourse researcher groups and thereby provide a solid foundation for corpus-based discourse research, which makes the researchers dispense with duplicate resource making efforts and increases the resources to be shared.

prove the coding schemes. In the second year, based on the examination results obtained in the first year's experiments, we have revised new annotation schemes and conducted the second round of coding experiments to verify them.

This paper describes our project of standardizing annotation schemes for Japanese discourse. In the following, annotation schemes for utterance units, discourse structure, and discourse markers are discussed based on our coding experiments.

2 Utterance Unit

2.1 First annotation scheme

Based on the survey of existing annotation schemes such as the schemes of several research groups in Japan (Kyoto Univ., Tsukuba Univ., Waseda Univ., ATR (Nagata, 1992)) and DRI (Allen and Core, 1996; Carletta et al., 1997a) for utterances (we call this utterance unit tags), we created the first annotation manual for illocutionary force type, mood information and exchange structures. Illocutionary force types come from speech act theory (Searle, 1969), and are one of the most popular set of describing communicative aspects of utterances. Mood information corresponds to the meaning of auxiliary verbs in Japanese, which has been hinted that there might be close relations with illocutionary act types. Exchange structures define minimal interactional units consisting of initiative, response and follow-up (Coulthhard, 1992; Stenstrom, 1994).

We carried out a first annotation experiments using the above three manuals, and obtained the following lessons for improving the schemes.

- The frequencies of the classifications:

  There exist exceedingly high and low frequency classifications in the illocutionary force types and mood information. The most frequent classification is *inform* in the illocutionary force types (54.9 %).

- The disagreement among coders:

  The disagreement among coders occurred due to three factors. The first is consistent decision errors caused by different interpretations of the category names (some coders classify utterances based on their interpretations of the category names, not on the functional definitions of the categories). The second is by the ambiguity of certain words and/or expressions. The last involves incomplete utterances like omission of the end part of utterances observed in Japanese spontaneous speech.

- The correlation between the information types:

  Most of the classifications for illocutionary force types and mood information show high correlation. This holds for exchange structure and speech act / mood except for *inform* category in the illocutionary force types.

2.2 Second annotation scheme

Based on the analysis of the experimental results, we revised the first annotation scheme by (1) unifying mood information into illocutionary force types, and (2) re-classifying some categories, i.e., further classifying high frequency categories by other information type and collapsing low frequency categories. The resultant scheme is composed of the illocutionary force types and the role of the utterances in the interaction unit.

To improve the disagreement among coders, we impose the constraint on the patterns of exchange structure (Figure 1).

In this new scheme, the tags (Figure 2) need to be an element of exchange structure except for those of dialogue management.

As in (Carletta et al., 1997a; Carletta et al., 1997b), we also created a decision tree to improve the coding reliability of this scheme. This decision tree consists of a set of questions concerning the functional character of target utterances.

2.3 Analysis of annotation results

In order to examine the reliability of this new scheme, we have carried out another

Basic pattern

⟨exchange structure⟩ →
⟨initiate⟩ ⟨response⟩ (⟨follow up⟩) (⟨follow up⟩)

Embedded pattern

⟨exchange structure⟩ →
⟨initiate⟩ ⟨embedded structure⟩* ⟨response⟩
(⟨ follow up⟩) (⟨ follow up⟩)

⟨embedded structure⟩ →
⟨response/initiate⟩ (⟨response⟩)

Figure 1: Patterns of exchange structure

- Dialogue management

  Open, *Close*

- Initiate

  Request, *Suggest*, *Persuade*, *Propose*, *Confirm*, *Yes-No question*, *Wh-question*, *Promise*, *Demand*, *Inform*, *Other assert*, *Other initiate.*

- Response

  Positive, *Negative*, *Answer*, *Hold*, *Other response.*

- Follow up

  Understand

- Response with Initiate

  The element of this category is represented as *Response Type / Initiate Type*.

Figure 2: Tag set of the second annotation scheme

tagging experiment for comparing the reliability of the first and the second scheme. We used five different types of task-oriented dialogues (Japanese Map Task, group scheduling, route direction, telephone shopping and appointment scheduling). An annotation unit is pre-defined based on (Meteer and Taylor, 1995), which roughly corresponds to one verb and related case phrases.

The experimental results show major improvements on the frequency of the categories (by avoiding the categories of high and low frequencies), and the reliability of the scheme.

- Frequency:

  The average quantity of information (entropy) are 1.65 in the first scheme, and 3.50 in the new scheme. The most frequent category in the new scheme is *Understand* (15.5 %), and other categories are evenly distributed.

- Reliability:

  The agreement among the coders is quantitatively evaluated with reliability in terms of the kappa coefficient K (Siegel and Castellan, 1988; Carletta et al., 1997b). In raw data, we cannot observe improvement, however, we found out a number of disagreements caused by consistent mistakes about the word "hai", which can be interpreted as either a positive response or a follow-up. Some coders neglected the constraints on follow-up introduced by the new manual: the constraint says that follow-ups must come only after response class utterances. This mistake can be alleviated by making a computer tagging tool display a warning message to the coders if they do not observe the constraint. To correctly evaluate the reliability of the schemes, the above simple problem should be discounted. Table 1 shows the agreement rate after substituting the mistaken follow-ups with the responses, in which we can clearly observe improvement on the reliability of the new scheme over the that of the first.

The reliability score of the new scheme is $K = 0.64$. This score is in "good" range according to (Siegel and Castellan, 1988), but does not seem to be the best. One reason for this is that our experiments were done with untrained subjects, which means that there can be more room for improvements on the reliability.

Table 1: Evaluation of utterance unit tagging

Data	scheme					
	first version			second version		
	agree 3	agree 2	disagree	agree 3	agree 2	disagree
Map task	60	51	1	41	54	18
group scheduling	38	8	0	30	12	4
route direction	35	8	3	31	6	9
telephone shopping	86	24	1	87	20	4
appointment scheduling	26	28	6	29	21	11
Total	245	119	11	218	113	46
Number of utterance	375			377		
P(A)	0.76			0.68		
P(E)	0.44			0.12		
κ	0.57			0.64		

3 Discourse Structure

3.1 First annotation scheme

Grosz and Sidner proposed a model of discourse structure, in which discourse structure is composed of the linguistic structure, the intentional structure, and the attentional state (Grosz and Sidner, 1986). We built the first annotation scheme of discourse structure in dialogue based on this model. The written instruction of the scheme describes as follows.

- Utterances both at the start and the end of segments are marked.
- Discourse segments may be nested. That is, a discourse segment can contain some smaller segments.
- Coders are allowed to decide the size of discourse segments.

In the first coding experiments, disagreements among the coders are incurred by three types of difficulties in segmenting dialogue.

- Identification of the end of discourse segments:

  This case often occurs due to the utterances which can be interpreted as responding to the preceding utterance while can be interpreted as initiating a new (but often related) topic, and the utterances followed by long series of responses, which are difficult to judge to be as initiating or responding.

- Disagreements of the nesting level of discourse segments:

  There are cases where coders can judge the relationship between adjacent discourse segments differently such as coordination and subordination. This results in different discourse structures, although the coders identically recognized the start and the end of the segment at the top level.

- Annotation units:

  Coders are allowed to change annotation units if necessary. Hence, for example, if some coder combine utterances in the given transcription, she might delete boundaries for segmenting discourse.

3.2 Second annotation scheme

We renewed the annotation scheme based on the analysis of disagreements in the first coding experiments.

In the second annotation scheme, the coders identify topic breaks between utterances based on the exchange structure,

32 A: '*Chikatetsu wa doko made noreba iidesu ka?*' [I]
(What station should I take the subway to?)

33 B: '*Hommachi eki kara Chuou eki made nori masu.*' [R]
(From Hommachi station to Chuou station.)

34 A: '*Hai.*' [F]
(Yes.)

35 B: '*Ikura kakari masu ka?*' [I]
(How much does it cost?)

36 A: '*Kuukou kara desu ka?*' [R&I]
(From the airport?)

37 B: '*Chikatetsu no hou dake wa?*' [R&I]
(How much only concerning the subway?)

38 A: '*Hommachi eki kara Chuou eki made 210 en desu.*' [R]
(210 yen from Hommachi station to Chuou station.)

38 B: '*Hai.*' [F]
(Yes.)

Figure 3: Exchanges in a Japanese dialogue.

which is explained in section 2. The topic break always starts a discourse segment. This modification can avoid the problem of identifying the segment ends. This scheme uses an exchange as a building block of discourse segments. Topic boundaries are marked before the Initiate and the Response-with-Initiate utterances, which start a new discourse segment. The Response and Follow-up utterances do not start a discourse segment. Figure 3 shows exchange structures with the utterance unit tags in a Japanese dialogue. In this Figure, [I], [R], [R&I], [F] denotes Initiate, Response, Response-with-Initiate, and Follow-up utterance, respectively. The topic boundaries are inserted before the utterances 32, 35, 36, and 37, in this example.

The second scheme is not concerned with the nesting structure of the discourse segments. This identification of topic breaks results in a flat structure of the discourse segments. Instead, each topic break is annotated in terms of two level *topic-break-index*(TBI), which indicates dissimilarity of the topics. The boundaries of the discourse segment with TBI=1 and =2 indicate a weak and a strong break of topic, respectively.

The tagging procedure of the second scheme is

1. recognizing exchange structures,

2. making tags immediately before all initiating utterances, and

3. assigning the strength of topic break for the tags.

3.3 Analysis of annotation results

We carried out tagging experiments for dialogues of two tasks, scheduling and route direction, based on two versions of annotation schemes. The agreement of tags between the coders is quantitatively evaluated with K.

Table 2 summarizes the average scores of reliability for paired comparisons among all coders. The number of coders is 4 and 5 for the route direction and the scheduling of the first experiments, respectively, and 10 for the second experiments. Table 2(a) shows reliability of existence of boundaries between all discourse segments ignoring the nesting structure and the strength of topic break. Table 2(b) shows reliability of structure of the discourse structure. The latter comparison considers the nesting level for the first annotation scheme and the TBI for the second annotation scheme. The second annotation scheme are confirmed to improve the reliability, especially for the segment structure. It successfully restricts the coder to mark start of the discourse segments using an exchange as a building block of the discourse segments. In the first experiment, reliability of segment structure was incurred by the difference of nesting structure the depth of which the coder determined. Replacing

Table 2: Reliability of Annotation of Discourse Structure

(a) for existence of boundaries

task	annotation scheme	
	1st	2nd
route direction	0.508	0.732
scheduling	0.756	0.570
average	0.632	0.653

(b) for segment structure

task	annotation scheme	
	1st	2nd
route direction	0.412	0.600
scheduling	0.478	0.529
average	0.445	0.564

the nesting by the TBI's for describing structure of the segments also improved coding reliability.

4 Discourse Markers

In English, some discourse markers have shown to be a cue to predict the boundary of discourse segments. In Japanese, discourse markers are expressed with the same vocabulary with aiduti (acknowledgment) and fillers.

Unlike English discourse markers, Japanese discourse markers are not lexical. Japanese words as "etto", "ano" and "ja" have no meaning themselves. However, there are abundant in Japanese discourse. Kawamori compared English discouse markders with Japanese. In Japanese coupus, half of the turns are started with these words, while English corpus shows that about 25 % of the turns start with corresponding expression(Kawamori et al., 1998).

The correlation between Japanese discourse markers and the boundary of discourse segments has not shown, which can be used to improve the identification of the discourse boundaries. In this section, the expressions which can be used for discourse markers, aiduti and fillers are enumerated based on the data survey, and the correlation between discourse markers and the discourse boundaries in Japanese is shown.

Table 3: Aiduti expressions selected by the coders

	4 coders	3 coders	2 coders	1 coder
hai	16	26	38	49
soudesuka	0	0	2	0
asoudesuka	0	0	2	0
e	0	1	1	0
nai	0	1	0	0
ha	1	0	1	0
Total	17	30	56	73

Table 4: Discourse marker expressions selected by the coders

	4 coders	3 coders	2 coders	1 coder
e	10	26	17	4
ano	2	9	6	4
de	1	0	5	7
dewa	1	1	2	2
a	0	19	7	15
eto	0	19	9	10
ja	0	6	3	0
aja	0	1	1	0
iya	0	1	0	0
Total	14	87	60	108

4.1 Surface expressions of discourse markers

Discourse markers and speech related phenomena are defined as utterances that function as a lubricant rather than contributing to achieving some task-related goals in conversations. In the first coding experiments, coders are instructed to annotate 'aiduti' (acknowledgments) and discourse markers based on their functional descriptions. Here filler was tentatively included in discourse markers.

Table 3 and Table 4 show words which were selected by 4 coders and their agreements of the selection.

The results show that surface forms can be used to distinguish between discourse markers and aiduti (and fillers), and the variety of the forms is rather limited. Based on the analysis of the results, we defined the functions and surface forms of aiduti, discourse markers and fillers as follows.

4.1.1 Aiduti

- Definition:
  Items which signify hearing of the other's speaking or prompting the next utterance (their function is not a definite answer rather a lubricant for conversations).

- Surface forms:
  "hai (yes, yeah, right)", "eto (well, aah, um)", "e (mmm, yeah)"

English corresponding expressions are shown in bracket for reference.

The above three expressions covered most of the cases for aiduti in the test-tagging experiment (for example, "hai" covered 81 % of all aiduti expressions), although we found out that there are a few expression different from the above. Candidate words sometimes have other functions than aiduti.

If "hai" functions as a definite answer, coders are instructed not to annotate it as aiduti.

4.1.2 Discourse markers

- Definition:
  Items which mainly contribute to clarifying discourse structure but not to problem solving

- Surface forms:
  "ja (ok)", "dewa (then, ok)", "soredewa (then, ok)", "soushitara (then, in that case)", "deshitara (then, in that case)", "souieba (I've just remembered, aah", "de (you see, so)", "sorede (and so)", "sousuruto (and so, in that case)", "soushimasuto (so you mean, in that case)", "tsumari (I mean, that means that)", "yousuruni (so you mean,)", "mazu (first, firstly)", "saishoni (first, firstly)", "kondo (then, next)", "tsugini (then, next)", "saigoni (last, lastly)", "maa (well)"

The phrases such as "hanashi wa kawarimasuga (by the way)" and "tsugi ni ikimasu (go ahead)" are also included in discourse markers, which are not identified by surface forms, but by their functions.

Table 5: Correlation between discourse markers and discourse boundaries

	Before	After	Else	Total
No Segment	50 (36 %)	121 (88 %)	633 (73 %)	804 (70 %)
Segment level 1	56 (41 %)	7 (5 %)	140 (16 %)	203 (18 %)
Segment level 2	32 (23 %)	10 (7 %)	94 (11 %)	136 (12 %)

4.1.3 Filler

- Definition:
  Items that fill up the gap between utterances and indicate the speaker's state like under consideration, hesitation and continuation.

- Candidate words:
  "eto (well, aah, um)", "e (mmm, yeah)", "ano (well, aah, um)", "a (oh)", "n (mmm)", "to (well)"

To limit candidate words, we suppose differences between corders decrease. We can annotate these words almost automatically.

4.2 Correlation between discourse markers and discourse boundaries

We examined the correlation between the discourse markers and the discourse boundary defined in section 3. In this experiment, 5 subjects were instructed to annotate the discourse boundaries, and 46 discourse markers were automatically selected by their surface forms in 5 dialogue data.

Table 5 shows that 64 % (41 % for segment level 1 and 23 % for segment level 2) of discourse markers are located directly after the discourse boundaries. The chance level is 30 %, and therefore, surface forms of discourse markers were found to be effective cue for recognizing discourse boundaries.

5 Conclusion

This paper summarized our efforts on standardizing discourse annotation schemes for Japanese and evaluated the reliability of the schemes. To improve the base reliability of the schemes, (1) interactional units are useful for constraining tag candidates and linking the utterance to the discourse structure level, and (2) discourse markers identified by their surface form can be used as a cue for indicating discourse boundaries.

The reliability issues involve various factors. For example, in the projects which attain high agreement rate of tagging such as the MapTask and Switchboard, they used syntactic cues in the coding manuals. This apparently contribute to the high agreement rate of tagging, although there leave some possibilities for confusing syntactic information with the meaning of the tags. In addition, in the MapTask, they include domain specific knowledge in the tags. The Switchboard project took the approach that the coders are allowed to tag utterances freely and then create the abstract classification relating to DAMSL coding schemes based on the first tagging experiment. Interestingly, the coders in the above two projects are all students, not researchers as in DRI and our project. The student coders are well-trained, while researchers of DRI and our project sometime have some biases to the coding schemes and often take little time for tagging experiments. The MapTask used the decision tree approach and was successful for attaining the high agreement rate. Since then, the decision tree approach has been believed to be a key to the high agreement rate. DRI and our project also adopted this approach, but the resultant agreement rate is not so high, comparing to the MapTask project. Considering various factoring involving the reliability, we should realise the decision tree approach cannot be a only key to the successful coding schemes. In this respect, our experiments are interesting. That is, we showed there is some room for improving coding schemes by introducing different dimensions to the original coding schemes. This kind of continuous efforts to improving coding schemes should not be looked over.

The computer tagging tools are necessary at least for creating consistent underlying representation of the tagging results. Moreover, for multi-level tagging, as in MATE and our project, the tools should provide easy access to different level. In both respects, the MATE tagging tool currently developed will be a very valuable resource for discourse (tagging) research community. However, if we want to create a large discursive annotated corpora, we must consider to build semi-automatically tagging tools used in morphological and syntactic tagging, which should include some kind of machine learning techniques.

References

J. Allen and M. Core. 1996. Draft of damsl: Dialog act markup in several layers. (ftp://ftp.cs.rochester.edu/pub/packages/dialog-annotation/manual.ps.gz).

J. Carletta, N. Dahlback, N. Reithinger, and M. A. Walker. 1997a. Standards for dialogue coding in natural language processing. Dagstuhl-Seminar-Report:167 (ftp://ftp.cs.uni-sb.de/pub/dagstuhl/ reporte/97/9706.ps.gz).

J. Carletta, A. Isard, S. Isard, J.C. Kowkto, G. Doherty-Sneddon, and A.H. Anderson. 1997b. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23:13–31.

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

B. J. Grosz and C. L. Sidner. 1986. Attention, intention and the structure of discourse. *Computational Linguistics*, 12:175–204.

M. Kawamori, T. Kawabata, and A. Shimazu. 1998. Discourse markers in spontaneous dialogue: A corpus based study of japanese and english. In *Proc. of ACL98 Workshop on Discourse Relations and Discourse Markers*, pages 93–99.

M. Meteer and A. Taylor. 1995. Dysfluency annotation stylebook for the switchboard corpus. Linguistic Data Consortium (ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps.gz).

M. Nagata. 1992. Using pragmatics to rule out recognition errors in cooperative task-oriented dialogues. In *Proc. of ICSLP.*

J. R. Searle. 1969. *Speech Acts.* Cambridge University Press.

S. Siegel and Jr. Castellan, N. J. 1988. *Non-parametric Statistics for the Behavioral Sciences.* McGraw-Hill, second edition.

A. B. Stenstrom. 1994. *An Introduction to Spoken Interaction.* Addison-Wesley.

M. Walker, L. Hirshman, J. Moore, and A. Joshi. 1996. IRCS workshops on discourse tagging. http://www.georgetown.edu/luperfoy/Discouse-Treebank/dri-kickoff.html.