Layered Speech-Act Annotation for Spoken Dialogue Corpus

Yuki Irie[*⋆] **, Shigeki Matsubara**[†]**, Nobuo Kawaguchi**[†]**,**
Yukiko Yamaguchi[†], **Yasuyoshi Inagaki**[‡]

[*]Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
irie@el.itc.nagoya-u.ac.jp
[†] Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
{matubara,kawaguti,yamaguchi}@itc.nagoya-u.ac.jp
[‡]Faculty of Information Science and Technology, Aichi Prefectural University
1522-3, Ibaragabasama, Kumabari, Ngakute-cho, Aichi-gun, 480-1198, Japan
inagaki@ist.aichi-pu.ac.jp

Abstract

This paper describes the design of speech act tags for spoken dialogue corpora and its evaluation. Compared with the tags used for conventional corpus annotation, the proposed speech intention tag is specialized enough to determine system operations. However, detailed information description increases tag types. This causes an ambiguous tag selection. Therefore, we have designed an organization of tags, with focusing attention on layered tagging and context-dependent tagging. Over 35,000 utterance units in the CIAIR corpus have been tagged by hand. To evaluate the reliability of the intention tag, a tagging experiment was conducted. The reliability of tagging is evaluated by comparing the tagging among some annotators using kappa value. As a result, we confirmed that reliable data could be built. This corpus with speech intention tag could be widely used from basic research to applications of spoken dialogue. In particular, this would play an important role from the viewpoint of practical use of spoken dialogue corpora.

1. Introduction

In recent years, large-scale speech corpora can be used for diverse research purposes and play important roles of basic resource for developing spoken dialogue systems. In order to utilize the collected dialogue data for upgrading a system, we need not only simple recording and transcription of speech but also various advanced information. Especially, understanding user's speech intention exactly is an essential to behave appropriately. It is preferable that speech intention tag is given to each utterance in the data.

This paper describes a design of speech intention tag and its evaluation using the CIAIR in-car spoken dialogue corpus. Compared with the tags used for conventional corpus annotation, this speech intention tag is specialized in development of spoken dialogue systems. For building a dialogue corpus with speech intention tag, we have used the CIAIR transcribed corpus. For each utterance unit about restaurant search on the corpus, we provided the intention tags by hand. At this time, we have tagged over 35,000 utterance units. To evaluate the reliability of intention tags, a tagging experiment was conducted. As a result, we confirmed that reliable data could be built.

2. Speech intention tag

Various tags expressing illocutionary force have been proposed as speech act tags (Alexandersson et al., 1997; Allen & Core, 1996; Walker & Passonneau, 2001). Speech act theory, proposed by Austin (Austin, 1962) and Searle (Searle, 1969), had no small effect on most of these tags. These uses between several to 20 kinds of intentions, such as "yn- question", "wh- question", "request". By giving these tags to each utterance, they have built the corpus with dialogue acts.

Understanding user's speech intentions exactly enables a dialogue system to act more adequately. However, an illocutionary force level of speech intention understanding isn't necessarily enough to determine system responses and operations. If a system determines a user's speech intention of "What time is it open until?" as "wh- question", it is not clear what user does request concretely. So, the system needs an additional processing such as reasoning.

In our study, we have designed speech intention tags specialized enough to determine a system operation and those are tagged on a corpus. In the previous example, we have given a tag expressing "the user requests the shop information regarding a business hour". However, a detailed information description like this increases tag types. This results in an ambiguous tag selection and thus it causes the following issues:

- If a speech intention tag gives a detailed description of the speaker's intention to a task-dependent level as referred to above, it includes from abstract information such as dialogue act to detailed information such as an object of the act. For example, a tag expressing "the user requests the shop information regarding a business hour" describes an intention given a shape to "request", and it includes more detailed information. But some dialogue systems would use the level of tag information selectively. (**multipurpose issue**)

- A speech intention could appear in speaker's facial expression or gesture, so it isn't always decided uniquely

[⋆] Current affiliation is DENSO CORPORATION.

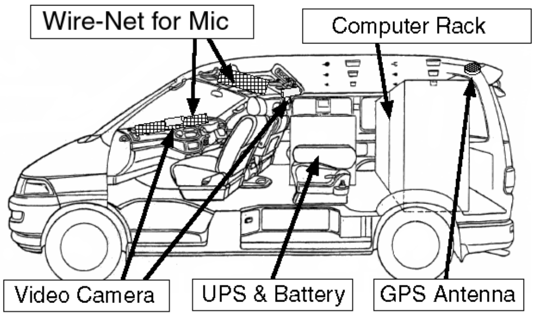


Figure 1: Recording environment for in-car speech

only from transcripts. For example, some people would consider “ima ai-te iru-kana. (Is it open now?)” as “the user asks whether the shop is open now or not”, other people would consider it as “the user asks whether there is any vacant seat now or not.” So, it is difficult for an annotator who isn’t a dialogue participant to understand utterances exactly. Also when the utterances are semantically ambiguous, given tags differ depending on the annotator’s interpretation. (**reliability isssue**)

On the other hand, we have designed an organization of tags, focused attention on layered tagging and context-dependent tagging:

- **Layered tagging:** In reference to a multipurpose problem, we have divided a tag into several layers according to degree of abstraction. More detailed speech intention can express by combining some levels of intention tag.
- **Context-dependent tagging:** In regard to a reliability problem, we respect participants’ judgments and assume that participants in the dialogue are cooperative enough. So we decide the intention tag based on how the listener understood. Specifically, we select a tag referring to the corresponding response utterance. According to this criterion, if a response to “ima ai-te iru-kana (Is it open now?)” is “The business hour is from 9 to 20”, then the speech intention of this utterance is regarded as “the user asks the business hour”, if a response is “That is full now” then the intention is “the user asks if a seat is available”.

3. Design of speech intention tag and annotation

For building a dialogue corpus with speech intention tag, we have used the CIAIR transcribed corpus (Kishida et al., 2003).

3.1. CIAIR in-car spoken dialogue corpus

At the Center for Integrated Acoustic Information Research (CIAIR), Nagoya University, we had collected an in-car spoken dialogue corpus aiming at realization of a robust spoken dialogue system in a real world environment (Kawaguchi et al., 2001; Kawaguchi et al., 2002; Kawaguchi et al., 2004). Figure 1 shows the recording environment for in-car speech. This corpus is a multi-modal corpus consisting of audio, videos, driving information and transcripts, and the world’s largest scale corpus recording the dialogues between a driver and a navigator with around 800 subjects, the volume of language data is about 1.03 million morphemes. Large-scale corpora can become the important resources for promoting various researches, and it is expected to be used by many researchers.

The transcription of dialogue speech was based on the transcription criteria for the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000). An example of a transcript is shown in Figure 2.

Table 1: Layered intention tag (part of)

Discouse act (1st)	Action (2nd)	Object (3rd)	Argument (4th)
Request	Confirm	Shop	ShopName
Propose	Exhibit	Parking	Genre
Express	Search	ShopInfo	Price
Suggest	Select	SearchResult	Place
Statement	Guide	ParkingInfo	Date

Table 2: Size of the corpus

Item	Numbers
Subject	1,256
Dialogue	3,641
Driver’s utterance	16,224
Operater’s utterance	19,187

3.2. Organization of intention tag

We have designed the organization of a speech intention tag according to the above concepts. Figure 3 shows a part of the organization of intention tags. And Table 1 shows a part of layered intention tags (LIT). LIT is composed of four layers, “Discourse act”, “Action”, “Object” and “Argument”. “Discourse act” layer denotes the role of the utterance unit in the dialogue. “Action” layer denotes the action of the utterance unit. “Object” layer denotes the object of the action such as “Shop”,”Parking”, etc. “Argument” layer denotes the other miscellaneous information about the utterance unit. Most of the argument layer tags can be decided directly from the specific keywords in the sentence.

All “Discourse act” layer tags is independent on tasks. Other layer tags express more detailed information, and include task-dependent tags. As Figure 3 shows, the upper-layered intention tag and the lower-layered one depends on each other. For example, “Object” layer tag of the utterance tagged “Express” on “Discourse act” layer is is either “Guide” or “Reserve”.

3.3. Annotation of spoken dialogue corpus

For building a dialogue corpus with LIT, we have used the CIAIR transcribed corpus (Kishida et al., 2003). For each utterance unit about restaurant search on the corpus, we provided the speech intention tag by hand. At this time, we

0003 – 00:04:955 – 00:06:560 M:D:N:O:		
じゃあ		&ジャー
マック	[McDonald’s]	&マック
教えてください<SB>	[Please tell me]	&オシエテクダサイ<SB>
0004 – 00:08:101 – 00:09:952 F:O:N:I:		
はい	[Yes]	&ハイ
マクドナルドですね<SB>	[McDonald’s]	&マクドナルドデスネ<SB>
0005 – 00:10:665 – 00:14:111 F:O:N:O:		
この先	[Around here]	&コノサキ
二百メートル先に	[200 meters away form here]	&ニヒャクメートルサキニ
マクドナルドが	[McDonald’s]	&マクドナルドガ
あります<SB>	[There is]	&アリマス<SB>

Figure 2: Transcript of in-car speech corpus

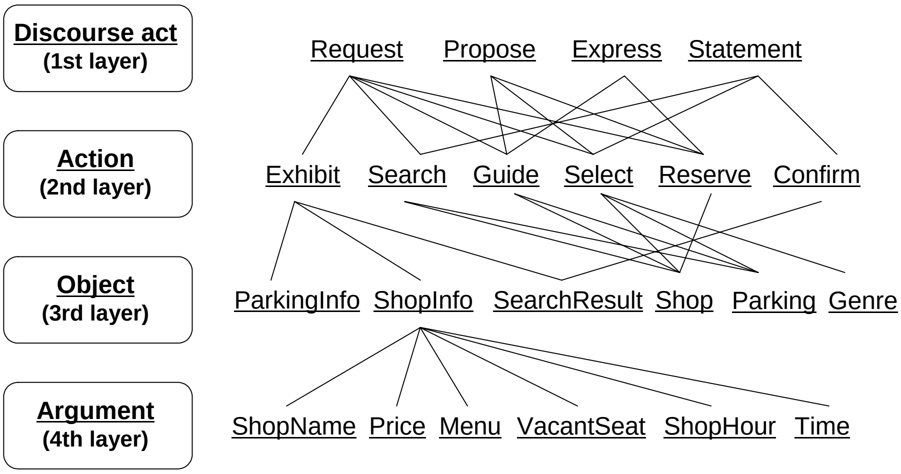


Figure 3: Organization of layered intention tag

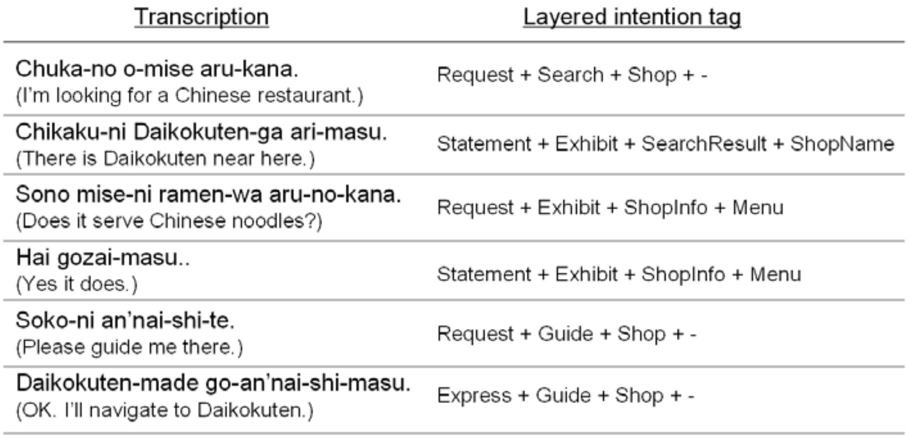
have tagged for over 35,000 utterance units. Figure 4 shows an example of a dialogue corpus with layered intention tags. We tagged two kinds of conversations, human-human conversation and human-WOZ conversation. This enables to analyze the effect of the difference in performance between dialogue parties (Kishida et al., 2003).

For tagging LIT, we have made an instruction manual. This manual gives a detailed explanation such as a procedure for annotation, detailed information of LIT, a connection restriction among layers, and an annotation unit. When we built the transcribed corpus, an utterance was divided into utterance units by a pause of 200 ms or more. In general, an utterance unit isn’t necessarily corresponding to an annotation unit such as sentence. In our restaurant search task, however, most utterance units correspond with a sentence. So, one speech intention tag is given to one utterance unit in principle. But the following exceptions are allowed:

- When the speech intention is over several utterance units, several utterance units are combined, and one intention tag is given to them. For example, two consecutive utterance units “ninki-no aru udonya desu-to (a popular noodle shop is)”, “Kanematsu ga kono-saki-ni ari-masu (Kanematsu down the road)” are combined. And we give one intention tag to “ninki-no aru udonya desu-to Kanematsu ga kono-saki-ni ari-masu (A popular noodle shop is Kanematsu down the road.)”

- When the utterance unit has several speech intentions, we divide the utterance unit into clauses, which is corresponding to the clause in English roughly, and one intention tag is given to one divided unit. For example, one utterance unit “Kanematsu-ni-wa chushajo-ga ari-masen-ga, yoroshi-desu-ka (Though Kanematsu doesn’t have the parking area, is it OK?)” is divided into two units “Kanematsu-ni-wa chushajo-ga ari-masen-ga (Though Kanematsu doesn’t have the parking area)” , “yoroshi-desu-ka (is it OK?)”, and then one intention tag is given to each unit.

Table 2 shows the size of the spoken dialogue corpus with LIT. LIT which appeared in the spoken dialogue corpus is 95 types. It counts a combination of “Discourse act”, “Action”, “Object” and “Argument” layer.



Transcription	Layered intention tag
Chuka-no o-mise aru-kana. (I'm looking for a Chinese restaurant.)	Request + Search + Shop + -
Chikaku-ni Daikokuten-ga ari-masu. (There is Daikokuten near here.)	Statement + Exhibit + SearchResult + ShopName
Sono mise-ni ramen-wa aru-no-kana. (Does it serve Chinese noodles?)	Request + Exhibit + ShopInfo + Menu
Hai gozai-masu.. (Yes it does.)	Statement + Exhibit + ShopInfo + Menu
Soko-ni an'nai-shi-te. (Please guide me there.)	Request + Guide + Shop + -
Daikokuten-made go-an'nai-shi-masu. (OK. I'll navigate to Daikokuten.)	Express + Guide + Shop + -

Figure 4: Example of layered intention tag annotation

Table 3: Experimental result

	Experiment I	Experiment II
$P(O)$	0.853	0.705
$P(E)$	0.071	0.052
κ	0.842	0.689

Table 4: Experimental result (Experiment I)

	Discourse act (1st)	Action (2nd)	Object (3rd)	Argument (4th)
$P(O)$	0.930	0.911	0.904	0.881
$P(E)$	0.341	0.252	0.184	0.302
κ	0.907	0.881	0.883	0.829

Table 5: Experimental result (Experiment II)

	Discourse act (1st)	Action (2nd)	Object (3rd)	Argument (4th)
$P(O)$	0.821	0.795	0.833	0.821
$P(E)$	0.356	0.230	0.168	0.302
κ	0.722	0.733	0.799	0.744

4. Evaluation of organization of intention tag

To evaluate the reliability of layered intention tag proposed in this paper, an evaluation experiment was conducted. If the selected tag varies among annotators, the conclusion derived from the tagged data could not be considered to be reliable.

Several researches discussed the reliability of a tag (Core & Allen, 1997; JDRI, 2000). In these researches, the reliability of a tag was evaluated by the comparison of tagging among some annotators. As an indicator of a quantitative evaluation, how many subjective judgments correspond among several annotators, Cohen's kappa value is frequently used (Carletta, 1996; Core & Allen, 1997; JDRI, 2000). So we have also used it as a measure of reliability. The kappa value measures an agreement among a set of tagging annotators, correcting for expected chance agreement.

$$\kappa = \frac{P(O) - P(E)}{1 - P(E)} \tag{1}$$

where $P(O)$ is the proportion of times that the annotators agree and $P(E)$ is the proportion of times that we would expect them to agree by chance (For complete instruction on how to calculate κ, see (Siegel & Castellan Jr., 1988)). When there is no agreement other than that which would be expected by chance, $\kappa = 0$. When there is total agreement, $\kappa = 1$.

A specialist who has expert knowledge and a general person who isn't familiar with this field are regarded as an annotator. In this study, we made the following experiments using a part of the spoken dialogue corpus with LIT.

- **Experiment I:** 2 persons, who are designers of LIT, give LIT to 28 dialogues (total 296 utterances).
- **Experiment II:** 4 persons, who aren't trained in tagging, give LIT to 51 dialogues (total 528 utterances). Each dialogue is tagged by 2 persons.

In both cases, annotators referred to the manual during experiments.

The results of a concordance rate are shown in Table 3, Table 4 and Table 5. Table 3 shows the value considered two tags are matched when all layer tags matched. And Table 4 and Table 5 show values calculated for each layer. In spoken dialogue research, there isn't an absolute criterion of acceptable level of agreement. Krippendorff has discussed what makes an acceptable level of agreement, while giving the caveat that it depends entirely on what one intends to do (Krippendorff, 1980). Carletta and Core & Allen say that $0.80 < \kappa$ is good reliability, $0.67 < \kappa < 0.80$ is usable quality for a concordance rate of 2 persons (Carletta, 1996; Core & Allen, 1997).

The conclusion derived from the tagged data could be reliable, because the kappa value of developers is within good reliability according to their literatures (Carletta, 1996; Core & Allen, 1997). Even though there are more types (95 types) than traditional tags, the high value has been got.

The use of a clear criterion for tag selection could be considered as one reason why we could get such a high value. The kappa values of all layers are within good reliability, so the reliable data could be built on any layers as shown in Table 4. This means that the conclusion from the data which use some layers selectively is reliable.

In the future, when building larger scale corpora with LIT, it is not absolutely necessary that tagging is performed by an expert. Even in the case of annotators who didn't have any prior training for tagging, the kappa values are within usable reliability (see Table 3 and 5) and consequently the proposed speech intention tag can be used for building reliable data even if annotator doesn't have specialized knowledge.

5. Conclusion

This paper describes the design of speech intention tags based on the CIAIR in-car speech dialogue corpus and its evalutation. Compared with the tags used for conventional corpus annotation, this proposed speech intention tag is specialized in spoken dialogue systems. For building a dialogue corpus with LIT, we have used the CIAIR transcribed corpus. For each utterance unit about restaurant search on the corpus, we provided the LIT by hand. At this time, we have tagged for over 35,000 utterance units. As a result of an evaluation experiment, we confirmed that reliable data could be built.

The spoken dialogue corpus with speech intention tag built in this way can be widely used from basic research to applications of spoken dialogue. In particular, this would play an important role from the viewpoint of practical use of spoken dialogue corpora. We have already obtained the results of discourse analysis (Irie et al., 2003; Kato et al., 2005; Kishida et al., 2003), speech intention understanding (Irie et al., 2004), and development of a spoken dialogue system (Hayashi et al., 2004).

In future work, this dialogue corpus will be effectively utilized for not only realization of robust spoken dialogue systems, but also analysis of the relation between spoken language grammar and speech intention, acquisition of dialogue grammar and knowledge acquisition, and so on.

6. Acknowledgments

The authors would like to thank all members of CIAIR, Nagoya University for their contribution to the construction of the in-car spoken dialogue corpus. They are grateful to Mr. Hiroya Murao of Sanyo Electric Co., Ltd. for valuable comments on tag design. This research was partially supported by the Grant-in-Aid for Scientific Research (No. 15300045) of the Ministry of Education, Science, Sports and Culture, Japan.

7. References

J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, E. Maier, N. Reithinger, B. Schmitz and M. Siegel. (1997). Dialogue acts in verbmobil-2. *Verbmobile Report* 204.

J. Allen and M. Core. (1996). draft of DAMSL: dialog act markup in several layers. [1]

J. L. Austin. (1962). *How to do things with words.* Harvard Univ. Press.

J. Carletta. (1996). Assessing agreement on classification tasks. *Computational Linguistics*, Vol.22, No.2, pp.249-254.

M. G. Core and J. F. Allen. (1997). Coding dialogs with the DAMSL annotation scheme. *Proceedings of the American Association for Artificial Intelligence Fall Symposium on Communicative Action in Humans and Machines*, pp. 28-35.

T. Fukada, D. Koll, A. Waibel and K. Tanigaki. (1998). Probabilistic Dialogue Act Extraction for Concept Based Multilingual Translation Systems. *Proceedings of the 5th International Conference on Spoken Language Processing*, Vol.6. pp.2771-2774.

K. Hayashi, Y. Irie, Y. Yamaguchi, S. Matsubara and N. Kawaguchi. (2004). Speech Understanding, Dialogue Management and Response Generation in Corpus-Based Spoken Dialogue System. *Proceedings of the 8th International Conference on Spoken Language Processing*.

Y. Irie, N. Kawaguchi, S. Matsubara, I. Kishida, Y. Yamaguchi, K. Takeda, F. Itakura and Y. Inagaki. (2003). An advanced Japanese speech corpus for in-car spoken dialogue research. *Proceedings of International Conference on Speech Databases and Assessment*, pp.209-216.

Y. Irie, S. Matsubara, N. Kawaguchi, Y. Yamaguchi and Y. Inagaki. (2004). Speech Intention Understanding based on Decision Tree Learning. *Proceedings of the 8th International Conference on Spoken Language Processing*.

JDRI: The Japanese Discourse Research Initiative JDRI. (2000). Japanese Dialogue Corpus of Multi-level Annotation. *Proceedings of 1st SIGdial Workshop on Discourse and Dialogue*.

S. Kato, S. Matsubara, Y. Yamaguchi and N. Kawaguchi. (2005). Dialogue Structure Annotation of In-car Speech Corpus based on Speech-Act Tag. *Proceedings of International Conference on Speech Databases and Assessment*, pp.159-163.

N. Kawaguchi, S. Matsubara, K. Takeda and F. Itakura. (2001). Construction of speech corpus in moving car environment. *Proceedings of the 7th European Conference on Speech Communication and Technology*, pp.2027-2030.

N. Kawaguchi, S. Matsubara, K. Takeda and F. Itakura. (2002). Multi-dimensional data acquisition for integrated acoustic information research. *Proceedings of the 3rd International Language Resources and Evaluation Conference*, pp.2043-2046.

N. Kawaguchi, S. Matsubara, Y. Yamaguchi, K. Takeda and F. Itakura. (2004). CIAIR in-car speech database. *Proceedings of the 8th International Conference on Spoken Language Processing*.

I. Kishida, Y. Irie, Y. Yamaguchi, S. Matsubara, N. Kawaguchi and Y. Inagaki. (2003). An advanced Japanese speech corpus for in-car spoken dialogue research. *Proceedings of the 8th European Conference on Speech Communication and Technology*, pp.1581-1584.

K. Krippendorff. (1980). *Content Analysis: An introduction to its methodology*. Sage Publication.

[1] http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html

K. Maekawa, H. Koiso, S. Furui and H. Isahara. (2000). Spontaneous speech corpus of Japanese. *Proceedings of the 2nd International Language Resources and Evaluation Conference*, pp.947-952.

J. R. Searle. (1969). *Speech acts: an essay in the philosophy of language*, Cambridge Univ. Press.

S. Siegel and N. J. Castellan Jr. (1988). *Nonparametric Statistics -for the Behavioral Science-*. McGraw-Hill, second Edition.

M. Walker and R. Passonneau. (2001). DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems. *Proceedings of the 1st International Conference on Human Language Technology Research*, pp.66-73.