Dialogue acts annotation for NICT Kyoto tour dialogue corpus to construct statistical dialogue systems

Kiyonori Ohtake, Teruhisa Misu, Chiori Hori, Hideki Kashioka, Satoshi Nakamura

MASTAR Project, NICT, 3-5 Hikaridai Keihannna Science City, Japan
Kiyonori.ohtake (at) nict.go.jp

Abstract

This paper introduces a new corpus of consulting dialogues designed for training a dialogue manager that can handle consulting dialogues through spontaneous interactions from the tagged dialogue corpus. We have collected more than 150 hours of consulting dialogues in the tourist guidance domain. This paper outlines our taxonomy of dialogue act (DA) annotation that can describe two aspects of an utterance: the communicative function (speech act (SA)), and the semantic content of the utterance. We provide an overview of the Kyoto tour guide dialogue corpus and a preliminary analysis using the DA tags. We also show a result of a preliminary experiment for SA tagging via Support Vector Machines (SVMs). In addition, we mention the usage of our corpus for the spoken dialogue system that is being developed.

1. Introduction

This paper introduces a new dialogue corpus for consulting in the tourist guidance domain. The corpus consists of speech, transcripts, speech act tags, morphological analysis results, dependency analysis results, and semantic content tags. In this paper, we describe the current status of a dialogue corpus that is being developed by our research group, focusing on two types of tags: speech act tags and semantic content tags. These speech act and semantic content tags were designed to express the dialogue act of each utterance.

Many studies have focused on developing spoken dialogue systems. Their typical task domains included the retrieval of information from databases or making reservations, such as airline information e.g. Defense Advanced Research Projects Agency (DARPA) Communicator (Walker et al., 2001) and train information e.g. Automatic Railway Information Systems for Europe (ARISE) (Bouwman et al., 1999) and Multimodal-Multimedia Automated Service Kiosk (MASK) (Lamel et al., 2002). Most studies assumed a definite and consistent user objective, and the dialogue strategy was usually designed to minimize the cost of information access. Other target tasks include tutoring and trouble-shooting dialogues (Boye, 2007). In such tasks, dialogue scenarios or agendas are usually described using a (dynamic) tree structure, and the objective is to satisfy all requirements.

In this paper, we introduce our corpus, which is being developed as part of a project to construct consulting dialogue systems, that helps the user in making a decision. Thus far, several projects have been organized to construct speech corpora such as the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000). The size of CSJ is very big, and a large part of the corpus consists of monologues. Although, CSJ includes some dialogues, the size of the dialogues is not enough to construct a dialogue system via recent statistical techniques. In addition, as compared to consulting dialogues, the existing large dialogue corpora covered very clear tasks in limited domains.

However, consulting is a frequently used and very natural form of human interaction. We often consult with a sales clerk while shopping or with staff at a concierge desk in a hotel. Such dialogues usually form part of a series of information retrieval dialogues that have been investigated in many previous studies. They also contain various exchanges, such as clarifications and explanations. The user may explain his/her preferences vaguely by listing examples. The server would then sense the user's preferences from his/her utterances, provide some information, and then request a decision.

It is almost impossible to handcraft a scenario that can handle such spontaneous consulting dialogues; thus, the dialogue strategy should be bootstrapped from a dialogue corpus. If an extensive dialogue corpus is available, we can model the dialogue using machine learning techniques such as partially observable Markov decision processes (POMDPs) (Thomson et al., 2008). Hori et al. (Hori et al., 2008) have also proposed an efficient approach to organize a dialogue system using weighted finite-state transducers (WFSTs); the system obtains the structure of the transducers and the weight for each state transition from an annotated corpus. Thus, the corpus must be sufficiently rich in information to describe the consulting dialogue to construct the statistical dialogue manager via such techniques.

In addition, a detailed description would be preferable when developing modules that focus on spoken language understanding and generation modules. In this study, we adopt DAs (Bunt, 2000; Shriberg et al., 2004; Bangalore et al., 2006; Rodriguez et al., 2007; Levin et al., 2002) for this information and annotate DAs in the corpus.

In this paper, we describe the design of the Kyoto tour guide dialogue corpus in Section 2. Our design of the DA annotation is described in Section 3. Sections 4 and 5 respectively describe two types of tag sets, namely, the SA tag and the semantic content tag. Section 6 describe the usage of the Kyoto tour guide dialogue corpus to construct our spoken dialogue system.

2. NICT Kyoto Tour Guide Dialogue Corpus

We are currently developing a dialogue corpus based on tourist guidance for Kyoto City as the target domain. Thus far, we have collected itinerary planning dialogues in Japanese, in which users plan a one-day visit to Kyoto City.

There are three types of dialogues in the corpus: face-to-face (F2F), Wizard of OZ (WOZ), and telephonic (TEL) dialogues. The corpus consists of 114 face-to-face dialogues, 80 dialogues using the WOZ system, and 62 dialogues obtained from telephone conversations with the interface of the WOZ system. Moreover, we also collected 48 English F2F dialogues in the beginning of 2009, and the dialogues have not been transcribed.

The overview of these three types of dialogues is shown in Table 2. Each dialogue lasts for almost 30 min. All of the dialogues have been manually transcribed. Table 2 also shows the average number of utterances per a dialogue.

Each face-to-face dialogue involved a professional tour guide and a tourist. Three guides, one male and two females, were employed to collect the dialogues. All three guides were involved in almost the same number of dialogues. The guides used maps, guidebooks, and a PC connected to the internet.

In the WOZ dialogues, two female guides were employed. Each of them participated in 40 dialogues. The WOZ system consists of two Internet browsers, a speech synthesis program, and an integration program for the collaborative work. Collaboration was required because in addition to the guide, operators were employed to operate the WOZ system and support the guide. The guide and the operators had their own individual computers that were connected to each other; further, they collaboratively operated the WOZ system to serve the user (tourist).

In the telephone dialogues, the same two female guides as for the WOZ dialogues were employed. In these dialogues, we used the WOZ system, but we did not need the speech synthesis program. The guide and a tourist shared the same interface in different rooms, and they could talk to each other through the hands-free headset.

Dialogues to plan a one-day visit consist of several conversations for choosing the places to visit. The conversations usually included sequences of requests from the users and provision of information by the guides as well as consultation in the form of explanation and evaluation. It should be noted that in this study, unlike information kiosk systems such as those developed in (Lamel et al., 2002) or (Thomson et al., 2008), enabling the user to access information is not an objective in itself. The objective is similar to the problem-solving dialogue of the study by (Ferguson and Allen, 1998); in other words, accessing information is just an aspect of consulting dialogues.

An example of dialogue via face-to-face communication is shown in Table 1. This dialogue is part of a consultation to decide on a sightseeing spot to visit. The user asks the location of a spot, and the guide answers it. Then, the user provides a follow-up by evaluating the answer. The task is challenging because there are many utterances that affect the flow of the dialogue during a consultation. The utterances are listed in the order of their start times with the utterance ids (UID). From the column 'Time' in the table, it is easy to see that there are many overlaps.

3. Annotation of Dialogue Acts

We annotate DAs in the corpus to describe a user's intention and a system's (or the tour guide's) action. Recently, several studies have addressed multilevel annotation of dialogues (Bangalore et al., 2006; Rodriguez et al., 2007; Levin et al., 2002); in our study, we focus on the two aspects of a DA indicated by Bunt (Bunt, 2000). One is the communicative function that corresponds to how the content should be used to update the context, and the other is a semantic content that corresponds to what the act is about. We consider both as important information to handle the consulting dialogue.

We designed two different tag sets to annotate DAs in the corpus. The SA tag is used to capture the communicative functions of an utterance using domain-independent multiple function layers. The semantic content tag is used to describe the semantic content of an utterance using domain-specific hierarchical semantic classes.

UID	SA tag
56	WH–Question_Where
57	State_Answer→56
58	State_Inversion
59	State_Evaluation→57
60	Pause_Grabber
61	Y/N–Question
62	State_Acknowledgment→59
63	State_AffirmativeAnswer→61
64	State_Opinion
65	State_Acknowledgment→64_Evaluation→64

Tags are concatenated by a delimiter '_' and omitting the null values.
The number following the '→' denotes the target utterance of the function.

Table 3: Example of SA annotation for the data shown in Table 1

3.1. Speech Act Tags

We introduce the SA tag set that describes communicative functions of utterances. Table 3 shows the example of speech act tags.

3.1.1. Annotation unit

There have been numerous discussions on the base unit of an SA annotation. As the simplest base unit, we can use a sentence or an utterance. However, sentence boundaries are not necessarily obvious in human-human dialogue. In addition, a long sentence tends to contain multiple dialogue functions. Thus, it is desirable to define a short unit so that the tags can elaborate the utterance. In addition, if the SA tag is used as an input of a dialogue system, the unit should be detected automatically (not manually). Therefore, we apply the clause boundary annotation program (CBAP) (Kashioka and Maruyama, 2004) to the transcript of the dialogue session, and adopt a clause as the base unit of tag annotation. Thus, in the following discussions, 'utterance' denotes a clause. We have already tagged more than 55 dialogues with SA tags. Roughly speaking, one dialogue consists of one thousand utterances.

UID	Time (ms)	Speaker	Transcript
56	76669–78819	User	*Ato* (and) *Ohara ga* (Ohara) *dono heN ni* (whereabouts) *narimasuka* (be?) (*Where is Ohara?*)
57	80788–81358	Guide	*kono* (here) *heN desune* (around be) (*Around here.*)
58	81358–81841	Guide	*Ohara wa* (Ohara)
59	81386–82736	User	*Chotto* (a bit) *hanaresugite masune* (be too far) (*Ohara seems to be too far from Kyoto st.*)
60	83116–83316	Guide	*A* (ah)
61	83136–85023	User	*kore demo* (it) *ichinichi dewa* (one day) *doudeshou* (how about?) (*Can I do Ohara in a day?*)
62	83386–84396	Guide	*Soudesune* (let me see)
63	85206–87076	Guide	*Ichinichi* (one day) *areba* (if be) *jubuN* (enough) *ikemasu* (can go) (*One day is enough to visit Ohara.*)
64	88392–90072	Guide	*Oharamo* (Ohara) *sugoku* (very) *kireidesuyo* (be a beautiful) (*Ohara is a very beautiful place.*)
65	89889–90759	User	*Iidesune* (sounds nice)

Table 1: Example dialogue from the Kyoto tour guide dialogue corpus

dialogue type	F2F (ja)	WOZ (ja)	TEL (ja)	F2F (en)
Number of dialogues	114	80	102	48
Number of guides	3	2	2	1
Average number of utterances per dialogue (guide)	365.4	165.2	–	–
Average number of utterances per dialogue (tourists)	301.7	112.9	–	–

Table 2: Overview of Kyoto tour guide dialogue corpus

3.1.2. Tag Specifications

There are two major policies in SA annotation. One is to select exactly one label from the tag set (e.g., the Augmented Multi-party Interaction (AMI) corpus[1]). The other is to annotate with as many labels as required. Meeting Recorder Dialog Act (MRDA) (Shriberg et al., 2004) and Dynamic Interpretation Theory (DIT) and DIT++ (Bunt, 2000) are defined on the basis of the second policy. We believe that the utterances are generally multifunctional and this multifunctionality is an important aspect for managing consulting dialogues through spontaneous interactions. Therefore, we have adopted the latter policy.

By extending the MRDA (Shriberg et al., 2004) tag set and DIT++ (Bunt, 2000), we defined our speech act tag set that consists of six layers to describe six groups of function: *General*, *Response*, *Check*, *Constrain*, *ActionDiscussion*, and *Others*.

The descriptions of the layers are as follows:

General layer Each tag of this layer represents the basic form of the unit. Most of the tags in this layer are used to describe forward-looking functions. The tags are classified into three large groups: 'Question,' 'Fragment,' and 'Statement.' The tag 'Statement==' denotes the continuation of the utterance. The following are the tags of the General layer.

Statement, **Pause**, **Backchannel**, **Y/N-Question**, **WH-Question**, **OR-Question**, **OR-segment-after-Y/N**, **Open-Question**

In the *General* layer, there are two sublayers for the labels: *Pause* and *WH-Question*. The *Pause* sublayer

[1] `http://corpus.amiproject.org`

Tag	Percentage(%) User	Guide	Tag	Percentage(%) User	Guide	Tag	Percentage(%) User	Guide
(General)			**(Response)**			**(ActionDiscussion)**		
Statement	45.25	44.53	Acknowledgment	19.13	5.45	Opinion	0.52	2.12
Pause	12.99	15.05	Accept	4.68	6.25	Wish	1.23	0.05
Backchannel	26.05	9.09	PartialAccept	0.02	0.10	Request	0.22	0.19
Y/N-Question	3.61	2.19	AffirmativeAnswer	0.08	0.20	Suggestion	0.16	1.12
WH-Question	1.13	0.40	Reject	0.25	0.11	Commitment	1.15	0.29
Open-Question	0.32	0.32	PartialReject	0.04	0.03			
OR–after-Y/N	0.05	0.02	NegativeAnswer	0.10	0.10			
OR-Question	0.05	0.03	Answer	1.16	2.57			
Statement==	9.91	27.79						
(Constrain)			**(Check)**					
Reason	0.64	2.52	RepetitionRequest	0.07	0.03			
Condition	0.61	3.09	UnderstandingCheck	0.19	0.20			
Elaboration	0.28	4.00	DoubleCheck	0.36	0.15			
Evaluation	1.35	2.01	ApprovalRequest	2.01	1.07			

Table 4: List of speech act tags and their occurrence in the experiment

consists of Hold, Grabber, Holder, and Releaser. The *WH* sublayer labels the WH-Question type.

Response layer The tags of this layer denote the responses directed to a specific previous utterance made by the addressee. The following are the tags of the Response layer.

Answer, **Acknowledgment**, **Accept**, **PartialAccept**,**AffirmativeAnswer**, **Reject**, **PartialReject**, **NegativeAnswer**

Check layer The tags of this layer denote the confirmation of a certain expected response. The following are the tags of the Check layer.

RepetitionRequest, **DoubleCheck**, **UnderstandingCheck**, **ApprovalRequest**

Constrain layer The tags of this layer denote the functions to restrict or complement the target of the utterance. The following are the tags of the Constrain layer.

Reason, **Condition**, **Elaboration**, **Evaluation**

ActionDiscussion layer The tags of this layer mark the functions of the utterances that pertain to a future action. The following are the tags of the Action Discussion layer.

Wish, **Opinion**, **Suggestion**, **Request**, **Commitment**

Others layer The tags of this layer describe various functions of the utterance, e.g. Greeting, SelfTalk, Welcome, Apology, etc. The following are the tags of the Others layer.

Greeting, **Introduction**, **Thank**, **Apology**, **Welcome**, **SelfRepair**, **Correct**, **CollaborativeComplementation**, **SelfTalk**, **Repeat**, **Mimic**, **Maybe**, **Inversion**

	General layer	All layers
Agreement ratio	86.7%	74.2%
Kappa statistic	0.74	0.68

Table 5: Agreement among labellers

3.1.3. Evaluation of the annotation

We performed a preliminary annotation of the SA tags in the F2F corpus. Thirty dialogues (900 minutes; 23,169 utterances) were annotated by three labellers. When annotating the dialogues, we took into account textual information, audio information, and contextual information. The result was cross-checked by another labeller.

The frequencies of the tags, expressed in percentages, are shown in Table 4. In the General layer, nearly half of the utterances were *Statement*. This bias is acceptable because 66% of the utterances that are tagged as *Statement* had tag(s) of other layers.

The percentages of the tags in the *Constrain* layer are relatively higher than those of the tags in the *ActionDiscussion* and *Check* layers. They are also higher than the corresponding percentage figures for MRDA (Shriberg et al., 2004) and SWBD-DAMSL (Jurafsky et al., 1997).

These statistics characterize the consulting dialogue of sightseeing planning, where elaborations and evaluations play an important role during the decision process.

We investigated the inter-annotator agreement for SA tags. Three labellers were employed to make six annotated dialogues from two dialogues (2,087 utterances). Each dialogue was annotated by the three labellers and the agreement among them was examined. These results are listed in Table 5. The agreement ratio is the average of all the combinations of the three individual agreements. In the same way, we also computed the average Kappa statistic, which is often used to measure the agreement by considering the chance rate. At present, 55 dialogues in the F2F and TEL corpora have been already annotated with SA tags.

A high concordance rate was obtained for the *General* layer. When the specific layers and sublayers are taken into account, The Kappa statistic was 0.68, which is considered a good result for this type of task. (e.g. (Shriberg et al., 2004), etc.)

We then investigated the tendencies of tag occurrence through a dialogue to clarify how consulting is conducted in the corpus. We annotated the boundaries of the episodes that determined the spots to visit to carefully investigate the structure of the decision-making processes. In our corpus, users were asked to write down their itinerary for a practical one-day tour. Thus, the beginning and ending of an episode can be determined on the basis of this itinerary.

As a result, we found 192 episodes. We selected 122 episodes that had more than 50 utterances, and analyzed the tendency of tag occurrence. The episodes were divided into five segments so that each segment had an equal number of utterances. An example of the tendencies of tag occurrence is shown in Figure 1. The relative occurrence rate is obtained by dividing the number of times the tags appeared in each segment by the total number of occurrences throughout the dialogues.

We found three patterns in the tendencies of occurrence. The tags corresponding to the first pattern frequently appear in the early part of an episode; this typically applies to Open-Question, WH-Question, and Wish. The tags of the second pattern frequently appear in the later part; this typically applies to Evaluation, Commitment, and Opinion. The tags of the third pattern appear uniformly over an episode; this typically applies to Y/N-Question, Accept, and Elaboration.

These statistics characterize the dialogue flow of sightseeing planning, where the guide and the user first clarify the latter's interests (Open, WH-Questions) and then list and evaluate candidates (Evaluation), following which the user takes a decision (Commitment).

This progression indicates that the management of a session or a dialogue phase requires wide contextual information within an episode to manage the consulting dialogue, even though the test-set perplexity[2], which was calculated by a 3-gram language model trained with the SA tags, was not high (4.25 using the general layer and 14.75 using all layers).

We also carried out a preliminary experiment to estimate the SA tags by using SVMs (Support Vector Machines). The SA tagged corpus is being developed and the corpus may not be clean. However, we tried to construct a SA tagger via SVMs.

We can see a SA tagging as a sequential labeling problem. We prepared 36 dialogues of F2F corpus with SA tags, in which we used 34 dialogues as learning data and two dialogues were used as a test data. We construct a classifier using only the labels of General layer. There are 16 labels of General layer in the learning data and the test data includes 13 labels.

The features used to construct a classifier are as follows: the role of the speaker, length of the utterance (second), barge-in flag, last three morphemes of the utterance, etc. The feature vector for the label of utterance u_i is extracted from $u_{i-4}, u_{i-3}, ..., u_i, u_{i+1}, u_{i+2}$. The kernel function of SVM is a 2nd-degree polynomial function. To achieve a multi-class classifier via SVM, we constructed the SVMs by the pairwise method.

The accuracy of our first trial was 73.02%. We have to consider the feature extraction to improve the accuracy. We will try to use features of all sorts.

The SA tagged corpus should be brushed up, because the agreement ratio between human labellers as shown in Table 5 does not reach 90% for the general layer. In other words, the maximum accuracy is estimated at around 86%. From these numbers, the accuracy 73% of our first try seems very promising.

3.2. Semantic Content Tags

The semantic content tag set was designed to capture the contents of an utterance. Some might consider semantic representations by HPSG (Pollard and Sag, 1994) or LFG (Dalrymple et al., 1994) for an utterance. Such frameworks require knowledge of grammar and experiences to describe the meaning of an utterance. In addition, the utterances in a dialogue are often fragmentary, which makes the description more difficult.

We focused on the predicate-argument structure that is based on dependency relations. Annotating dependency relations is more intuitive and is easier than annotating the syntax structure; moreover, a dependency parser is more robust for fragmentary expressions than syntax parsers.

Table 3 shows the example of the semantic content tags (semantic class labels.)

We introduced semantic classes to represent the semantic content of an utterance. Semantic class labels are applied to each unit of the predicate-argument structure. The task that identifies the semantic classes is very similar to named entity recognition, because the classes of the named entities can be equated to the semantic classes that are used to express semantic content. However, both nouns and predicates are very important for capturing the semantic content of an utterance. For example, '10 a.m.' might denote the current time in the context of planning, or it might signify the opening time of a sightseeing spot. Thus, we represent the semantic content on the basis of the predicate-argument structure. Each predicate and argument is assigned a semantic category.

For example, the sentence "I would like to see Kinkakuji temple." is annotated as shown in Figure 2. In this figure, the semantic content tag *(preference).action* indicates that the predicate portion expresses the speaker's *preference* for the speaker's action, while the semantic content tag *(preference).(spot).name* indicates the *name* of the *spot* as the object of the speaker's *preference*.

Although we do not define the semantic role (e.g. object (*Kinkakuji temple*) and subject (*I*)) of each argument item in this case, we can use conventional semantic role labelling techniques (Gildea and Jurafsky, 2002) to estimate them.

[2]The perplexity was calculated by a 10-fold cross validation of the 30 dialogues.

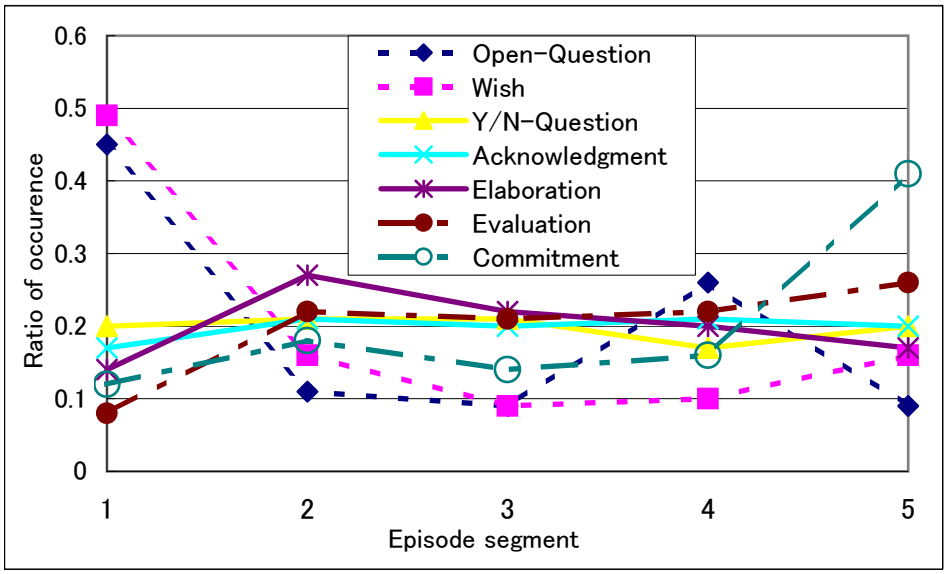

Figure 1: Progress of episodes vs. the occurrence of SA tags

UID	Transcript	Semantic content tag
56	*Ato* (and)	null
	Ohara ga (Ohara)	(activity),location
	dono heN ni (whereabouts)	(activity),(demonst),interr
	narimasuka (be?)	(activity),predicate
57	*kono* (here)	(demonst),kosoa
	heN desune (around be)	(demonst),noun
58	*Ohara wa* (Ohara)	location
59	*Chotto* (a bit)	(trsp),(cost),(distance),adverb-phrase
	hanaresugite masune (be too far)	(trsp),(cost),(distance),predicate
60	*A* (ah)	null
61	*kore demo* (it)	null
	ichinichi dewa (one day)	(activity),(planning),duration
	doudeshou (how about?)	(activity),(planning),(demonst),interr
62	*Soudesune* (let me see)	null
63	*Ichinichi* (one day)	(activity),(planning),(entity),day-window
	areba (if be)	(activity),(planning),predicate
	jubuN (enough)	(consulting),(activity),adverb-phrase
	ikemasu (can go)	(consulting),(activity),action
64	*Oharamo* (Ohara is)	(recommend),(activity),location
	sugoku (very)	(recommend),(activity),adverb-phrase
	kireidesuyo (beautiful)	(recommend),(activity),predicate
65	*Iidesune* (sounds nice)	(consulting),(activity),predicate

Table 6: Example of semantic content tag annotation for the data shown in Table 1

3.2.1. Tag Specifications

We defined the hierarchical semantic classes to annotate the semantic content tags. There are 33 labels (classes) at the top hierarchical level. The labels include **activity**, **event**, **meal**, **spot**, **transportation**, **cost**, **consulting**, and **location**, and are shown in Figure 3. There are two kinds of labels, nodes, and leaves. A node must have at least one child, a node, or a leaf. A leaf has no children. The number of types for nodes is 47, and the number of types for leaves is 47. The labels of the leaves are very similar to the labels for named entity recognition. For example, there are 'year', 'date', 'time', 'organizer', 'name', etc. in the labels of the leaves.

One of the characteristics of the hierarchical structure of the semantic classes is that the lower level structures are shared by many upper nodes. Thus, the lower level structure can be used in any other domain or target task.

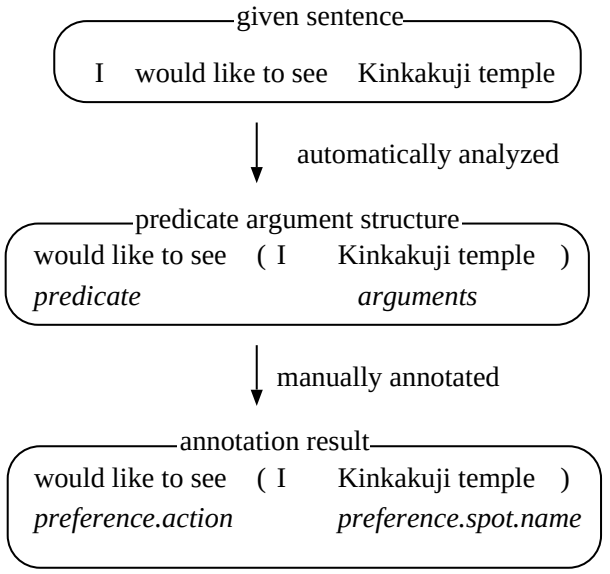

Figure 2: Example of annotation with semantic content tags

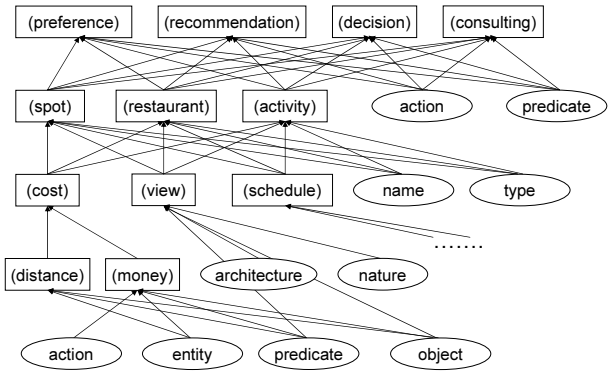

Figure 3: A part of the semantic category hierarchy

3.2.2. Annotation of semantic contents tags

The annotation of semantic contents tags is performed by the following four steps. First, an utterance is analyzed by a morphological analyzer, ChaSen[3]. Second, the morphemes are chunked into dependency unit (*bunsetsu*). Third, dependency analysis is performed using a Japanese dependency parser, CaboCha[4]. Finally, we annotate the semantic content tags for each *bunsetsu* unit by using our annotation tool. An example of an annotation is shown in Table 6. Each row in column "Transcript" denotes the divided *bunsetsu* units.

At present, the annotations of semantic content tags are being carried out for 40 dialogues. Approximately 26,800 paths, including paths that will not be used, exist if the layered structure is fully expanded. In the 40 dialogues, 1,980 tags (or paths) are used.

In addition, not only the annotation of semantic content tags but also the correction of the morphological analysis and dependency analysis results is being carried out. If we complete the annotation, we will also obtain the correctly tagged data of the Kyoto tour guide corpus. These corpora can be used to develop analyzers such as morphological analyzers and dependency analyzers via machine learning techniques or to adapt the analyzers for this domain.

[3] http://sourceforge.jp/projects/chasen-legacy/

[4] http://chasen.org/~taku/software/cabocha/

4. Usage of the Kyoto Tour Guide Corpus

In this section, we discuss the usage of the Kyoto tour guide corpus. We can see that a dialogue system consists of a speech recognition module, a dialogue management module, a speech synthesis module, and a database for target domain. Recently, most of those modules have been based on statistical methods that require corpora.

4.1. Speech Recognition

We constructed the language model that is used in the speech recognition module of our dialogue system. To construct the language model, the morphological analysis results of the dialogue corpus were used. It is required that the domain specific n-gram entries are included in the language model to achieve high performance for speech recognition. Only maintaining the recognition dictionaries does not lead us to the satisfactory recognition results.

4.2. Dialogue Management

One of the most significant roles of the dialogue model in a spoken language dialogue system seems to appropriately represent a contextual interpretation of the user utterances. This allows the system to generate the most adequate system response without limiting the dialogue to a succession of questions and answers. This role should also enable the system to anticipate/predict, raise ambiguities, correct errors, explain system decisions, and trigger the corresponding actions throughout the dialogue to suitably manage other processing modules.

We are now adopting the corpus annotated with the DA tags to construct a dialogue system using WFSTs as dialogue management modules. To achieve dialogue management via WFSTs, we have to prepare not only the DA tags but also the tags for the system's action. As such, we are now preparing such action tags to construct a dialogue management module using WFSTs.

In addition, the corpus consists of real conversions between the guide and the travellers. Important and valuable information is buried in the corpus. If we apply data-mining techniques to the corpus, we will obtain much valuable information for travelling in Kyoto city and we can store this information in the database of the spoken dialogue system.

4.3. Speech Synthesis

Recent speech synthesis techniques such as concatenative synthesis or statistical parametric synthesis require large speech corpora. We can use conventional speech synthesis modules for a spoken dialogue system and the performance of the module as a text-to-speech module seems very high. However, we want to construct a more natural speech synthesis module that is suitable for a spoken dialogue system. Most of the conventional speech synthesis modules make only one speech from the text. In other words, it is hard to synthesize different speeches from the same text.

We have corpora with speech act tags, and we want to use this information to synthesize different speeches from the same text. In Japanese, "*hai* (yes)" is used in many ways, such as acknowledgment, back-channel, etc. We are now constructing speech synthesis modules using our dialogue corpus using two approaches. One is by constructing a speech synthesis system that directly uses the recorded speech data of the guide. The other one is by constructing a speech synthesis system that uses a new speech corpus recorded with voice actors/actresses. For these recordings, we prepared the scripts from the transcripts of the corpus.

5. Conclusion

In this paper, we have introduced our spoken dialogue corpus for developing consulting dialogue systems. We designed a dialogue act annotation scheme that describes two aspects of a DA: speech act (SA) and semantic content. The SA tag set was designed by extending the MRDA tag set. The design of the semantic content tag set is almost complete. If we complete the annotation, we will obtain SA tags and semantic content tags, as well as manual transcripts, morphological analysis results, and dependency analysis results. As a preliminary analysis, we have evaluated the SA tag set in terms of the agreement between labellers and investigated the patterns of tag occurrences. In addition, we tried to construct a SA tagger via SVMs as a first step to use the tagged corpus and the result was promising. We also mentioned the corpus usage in the development of our spoken dialogue system.

Next, we will investigate the features for automatic SA and semantic content tagging. We will construct taggers for both SA and semantic content tags using the annotated corpora and machine learning techniques.

6. References

Srinivas Bangalore, Giuseppe Di Fabbrizio, and Amanda Stent. 2006. Learning the structure of task-driven human-human dialogs. In *Proceedings of COLING/ACL*, pages 201–208.

Gies Bouwman, Janienke Sturm, and Louis Boves. 1999. Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the ARISE Project. In *Proc. ICASSP*.

Johan Boye. 2007. Dialogue Management for Automatic Troubleshooting and Other Problem-solving Applications. In *Proc. of 8th SIGdial Workshop on Discourse and Dialogue*, pages 247–255.

Harry Bunt. 2000. Dialogue pragmatics and context specification. In Harry Bunt and William Black, editors, *Abduction, Belief and Context in Dialogue*, pages 81–150. John Benjamins.

Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, and Anni e Zaenen, editors. 1994. *Formal Issues in Lexical-Functional Grammar*. CSLI Publications.

George Ferguson and James F. Allen. 1998. TRIPS: An intelligent integrated problem-solving assistant. In *Proc. Fifteenth National Conference on Artificial Intelligence*, pages 567–573.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Chiori Hori, Kiyonori Ohtake, Teruhisa Misu, Hideki Kashioka, and Satoshi Nakamura. 2008. Dialog Management using Weighted Finite-state Transducers. In *Proc. Interspeech*, pages 211–214.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical report, University of Colorado at Boulder & SRI International.

Hideki Kashioka and Takehiko Maruyama. 2004. Segmentation of Semantic Unit in Japanese Monologue. In *Proc. ICSLT-O-COCOSDA*.

Lori F. Lamel, Samir Bennacef, Jean-Luc Gauvain, H. Dartigues, and J. N. Temem. 2002. User evaluation of the MASK kiosk. *Speech Communication*, 38(1):131–139.

Lori Levin, Donna Gates, Dorcas Wallace, Kay Peterson, Along Lavie, Fabio Pianesi, Emanuele Pianta, Roldano Cattoni, and Nadia Mana. 2002. Balancing expressiveness and simplicity in an interlingua for task based dialogue. In *Proceedings of ACL 2002 workshop on Speech-to-speech Translation: Algorithms and Systems*.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of Japanese. In *Proceedings of the Second International Conference of Language Resources and Evaluation (LREC2000)*, pages 947–952.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press.

Kepa Joseba Rodriguez, Stefanie Dipper, Michael Götze, Massimo Poesio, Giuseppe Riccardi, Christian Raymond, and Joanna Rabiega-Wisniewska. 2007. Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus. In *Proc. Linguistic Annotation Workshop*, pages 148–155.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proc. 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100.

Blaise Thomson, Jost Schatzmann, and Steve Young. 2008. Bayesian update of dialogue state for robust dialogue systems. In *Proceedings of ICASSP '08*.

Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and Qualitative Evaluation of DARPA Communicator Spoken Dialogue Systems. In *Proc. of 39th Annual Meeting of the ACL*, pages 515–522.