Dialogue Act Annotation and Identification in a Japanese Multi-party Conversation Corpus

Takashi Yamamura, Masato Hino, Kazutaka Shimada
Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka Japan
{t_yamamura, m_hino, shimada}@pluto.ai.kyutech.ac.jp

Abstract

Identifying dialogue acts of an utterance has an important role in various tasks of natural language processing, such as dialogue systems and summarization. In this paper, we describe an annotation task of dialogue acts based on an existing tag set for a Japanese multi-party conversation corpus and evaluate agreement scores between annotators. Then, we propose a method that identifies dialogue acts for an input utterance by using the classifier's outputs for each dialogue act tag. We apply utterance features and audio-visual features for the dialogue act classification. We also evaluate the effectiveness of our dialogue act annotation through a summarization task.

Keywords: dialogue acts, multi-party

1. Introduction

Dialogue acts denote dialogue functions and discourse structure of utterances in a conversation. Dialogue acts provide key information that indicates a role of an utterance and a relationship between utterances. Therefore, many researchers have utilized dialogue acts for understanding multi-party conversations in various tasks of natural language processing. In particular, for dialogue systems, identifying dialogue acts of an utterance contributes to the selection of appropriate responses (Higashinaka et al., 2014). For another task, Murray et al. (2010) have proposed a meeting summarization system that creates abstract summaries of specific aspects of a meeting by using dialogue acts.

Many researchers have constructed conversation corpora with dialogue act tags (Carletta, 2007; Shriberg et al., 2004). However, most of the existing tag sets and annotation manuals of dialogue acts have been mainly designed for the target corpora. Therefore, we often cannot utilize them directly in the annotation process of other corpora. In recent years, Bunt et al. (2012) have proposed an application-independent dialogue act annotation scheme, ISO 24617-2, which can adequately deal with typed, spoken, and multimodal dialogues. There have been several studies of construction of corpora using the ISO 24617-2 annotation schema, such as Hiraoka et al. (2014).

In this paper, we focus on a dialogue act annotation task based on ISO standard 24617-2 and a dialogue act classification task for a Japanese multi-party conversation corpus. We use the Kyutech corpus that is freely available on the web (Yamamura et al., 2016). The Kyutech corpus is a Japanese conversation corpus annotated for summarization tasks and consists of conversations about a decision-making task with four participants. Although the Kyutech corpus contains several annotations, dialogue acts have not been annotated. Therefore, we annotate dialogue acts to the Kyutech corpus. We also propose a method for identifying dialogue acts in the corpus.

In addition, we evaluate the effectiveness of the dialogue act tags through another task: conversation summarization. We compare the performance of a summarization method using dialogue acts and a method without dialogue acts. This paper shows that the method with dialogue acts outperforms the baseline without dialogue acts in the summarization task for the Kyutech corpus.

The contributions of this paper are as follows:

- We annotate dialogue acts based on ISO standard 24617-2 to the Kyutech corpus.
- We propose a supervised learning method that classifies dialogue acts based on ISO standard 24617-2. We use utterance features and audio-visual features.
- We show that our dialogue act annotation is effective for conversation summarization.

2. Related Work

There have been some previous studies on dialogue act annotation. DAMSL (Allen and Core, 1997) is a famous dialogue act annotation scheme developed especially for task-oriented dialogues. In addition, Jurafsky et al. (1997) and Shriberg et al. (2004) have proposed annotation schemes based on DAMSL. Also, Carletta et al. (1997) have been proposed a famous dialogue structure coding scheme that was devised for use with the HCRC Map Task corpus (Anderson et al., 1997). The HCRC Map Task corpus is a cooperative task that two participants share information about each other's map and a route marked on the one map. The scheme has been mainly designed for the task.

For Japanese conversation, Araki et al. (1999) have proposed a standard utterance-unit tagging scheme. They designed twenty types of dialogue act tags and its annotation scheme. Generally, a suitable dialogue act tag set and an annotation scheme for a corpus depend on target conversation types and purposes of use. Therefore, many existing annotation schemes depend on target corpora. In contrast, Bunt et al. (2012) have unified various annotation schemes and proposed a standard annotation scheme for dialogue acts, ISO 24617-2. The ISO 24617-2 contains various and useful tags for various corpus type. Hiraoka et al. (2014)

have annotated dialogue acts on the basis of ISO standard 24617-2 for a Japanese conversation corpus. The ISO 24617-2 also supports task-oriented conversations. Therefore, we apply the ISO 24617-2 to our tag sets for dialogue act annotation.

The features for dialogue act classification are mainly divided into two categories: utterance features and audio-visual features. Bag-of-words (BOW) is a popular vector representation that describes the occurrence of words within a document. In particular, bag-of-ngrams representations contribute to dialogue act classification (Stolcke et al., 2000; Verbree et al., 2006; Tavafi et al., 2013). Many researchers have proposed various utterances features, such as words at the end of utterances (Moldovan et al., 2011), part-of-speech tags (Bangalore et al., 2006), function word n-grams (Omuya et al., 2013), and utterance length (Ferschke et al., 2012; Tavafi et al., 2013). Verbree et al. (2006) and Kim et al. (2010) have reported that dialogue act tags in previous utterances provided an important clue to the dialogue act classification for a current utterance. In real conversations, gesture and speech information have an important role in the communication between speakers. Therefore, audio-visual features are also valuable for dialogue act classification as well as utterance features. Several researchers have proposed some audio-visual features, such as acoustic information of utterances and speech rate (Surendran and Levow, 2006), pose and gesture (Ezen-Can et al., 2015), and facial expression (Boyer et al., 2011). Therefore, we apply utterance features and audio-visual features to dialogue act classification.

3. The Kyutech Corpus

We use the Kyutech corpus (Yamamura et al., 2016). The Kyutech corpus contains multi-party conversations with four participants randomly selected from sixteen male students and four female students. The participants pretended managers of a virtual shopping mall in a virtual city and then determined a new restaurant, as an alternative to a closed restaurant, from three candidates. Before the discussion, the participants read a 10-page document including information about the three candidates, the closed restaurant and the existing restaurants in the mall, the city information, statistical information about the shopping mall, and so on. They read the document for 10 minutes, then discussed the candidates for 20 minutes and finally determined one restaurant as a newly-opened restaurant. The Kyutech corpus consists of nine conversations based on four scenarios of which task settings differ from each other.

The transcription rules were based on the construction manual of the Corpus of Spontaneous Japanese (CSJ) by (Maekawa et al., 2000). All utterances in the corpus were separated by 0.2-second interval by the guideline and annotated some tags such as filler, question, and so on. Each utterance was not always sentence-level because it depended on the 0.2-second interval rule. Therefore, other tags were appended to the end of each utterance for sentence-level identification. The corpus consists of 4,509 utterances in nine conversations, with a total of 2,810 sentences.

The Kyutech corpus is developed for summarization tasks and contains the annotations for summarization tasks: topic tags and reference summaries. However, communicative functions are also an important role in multi-party conversation understanding. Therefore, we annotate dialogue acts based on ISO standard 24617-2, which is introduced in the latter section.

4. Dialogue Act Annotation

In this section, we explain the dialogue act annotation based on ISO standard 24617-2 for the Kyutech corpus and then report the results.

4.1. Annotation unit

We annotate dialogue act tags for each utterance of the nine conversations contained in the Kyutech corpus. Although the original Kyutech corpus divided each utterance by 0.2-second interval, we divide the sentences of the Kyutech corpus into *long utterance-unit*[1] (Den et al., 2010) that is a scheme for annotating utterance-level units in Japanese dialogs. Long utterance-units correspond to basic units of interaction in sharing information between a speaker and a listener. We apply this scheme to the Kyutech corpus. As a result, we obtained 3,302 utterances in long utterance-units.

4.2. Tag set

As dialogue act tags, we use the general-purpose functions (GPF) and the dimension-specific communicative functions (DSCF) defined by ISO standard 24617-2. GPF tags are dialogue act tags for utterances with communicative functions for advancing discussion and accomplishing a task. In contrast, DSCF tags focus on interactive features between speakers, such as correcting own utterance or other speaker's utterance and indicating understanding or not. While GPF tags can be annotated with any DSCF tags, two or more GPF tags cannot be assigned together for one utterance.

Although there are various communicative functions in the ISO 24617-2, some communicative functions might not be necessary for the Kyutech corpus. In addition, too many functions for annotation often cause a decrease in the agreement between annotators. Therefore, we performed the preliminary annotation to decide tags for the Kyutech corpus. As a result, we excluded some redundant communicative functions. We also merged some functions with the low agreement in the preliminary annotation. After all, we selected ten GPF tags and nine DSCF tags from the communicative functions of ISO standard 24617-2. In the selection of GPF tags, we adopted the superordinate tags in the GPF tags of ISO standard 24617-2. Moreover, we added two tags to GPF tags and one tag to DSCF tags because we judged them to be necessary for the Kyutech corpus.

Table 1 lists the dialogue act tags in the Kyutech corpus. "Inform" has the three subclass of "Agreement", "Disagreement", and "Answer" in a parent-child relationship. We explain the added tags and the merged tags. "Monologue" is a function for muttering to oneself. "Vague" is used to

[1] http://www.jdri.org/resources/manuals/uu-doc-2.1.pdf

GPF tags	Question, Inform, Agreement†, Disagreement†, Answer†, Offer, Suggest, Request, Address Suggest, Address Request, Monologue*, Vague*
DSCF tags	Positive Feedback**, Negative Feedback**, Feedback Elicitation, Stalling, Pausing, Self Correction, Self Completion*, Retraction, Completion, Correct Misspeaking

Table 1: GPF tags and DSCF tags in the Kyutech corpus. The dagger (†) denotes the subclass for "Inform". The asterisk ($*$) denotes the added tags. The double asterisk ($**$) denotes the merged tags.

annotate utterances when an annotator cannot judge a tag because the annotator does not understand what a speaker is saying. These tags are independent of other GPF tags.
The ISO 24617-2 contains "Self Correction" tag that is a function of the correction to the previous utterance. We used "Self Correction" tag in the preliminary annotation. However, we found that speakers often corrected own utterances by adding some explanation in the Kyutech corpus. Therefore, we created "Self Completion" tag that is a function for adding some explanation to the previous utterance. We also merged "autoPositive" and "alloPositive" in the DFCS of ISO standard 24617-2, and then defined "Positive Feedback". In the same way, we also defined "Negative Feedback".

4.3. Annotation process

Nine annotators[2] annotate the dialogue act tags for each utterance. We applied three annotators into one conversation. In this process, each annotator selects at most one GPF tag and two DSCF tags; annotators have to annotate at least one tag for each utterance. During the annotation, annotators can check the audio-visual data of the conversations.

4.4. Results and analysis

To evaluate the reliability of the annotation, we computed the inter-annotator agreement about tags. For each conversation, we severally computed the score between two annotators in three annotators. We used two measures: the Dice coefficient and Cohen's Kappa (Carletta, 1996). We used the Dice coefficient because an utterance can be annotated more than one tag in this annotation. We computed the Dice score for each utterance between tags of two annotators and then calculated an overall score for the conversation by averaging them. Table 2 shows the Dice scores between two annotators and the averaged Dice scores for each conversation. The Dice score between two annotators was 0.587 on average. This score shows that two of the three tags are almost the same tags for two annotators. We also computed the Cohen's Kappa that is used in assessing agreement between annotators (Carletta, 1996). When an utterance contains more than one tag, we selected one tag with high agreement between annotators. Specifically, if an annotator annotates "Inform" and "Self Completion" tag and another annotator annotates "Inform" tag, we regard the first annotator's tag as "Inform". Table 3 shows the Kappa scores for each conversation. The Kappa score between two annotators was 0.478 on average. This score is not high. Hence, we need to improve the reliability of the annotation.

[2]Nine Annotators consist of seven male students and two female students.

Conv. ID	Dice Score			
	A_1-A_2	A_1-A_3	A_2-A_3	Avg.
0313_C1	0.597	0.668	0.557	0.607
0320_C1	0.674	0.554	0.532	0.587
0320_C4	0.529	0.532	0.613	0.558
0323_C3	0.554	0.779	0.564	0.632
0326_C1	0.588	0.563	0.518	0.556
0326_C2	0.515	0.539	0.600	0.551
0326_C4	0.670	0.655	0.634	0.653
0327_C2	0.562	0.533	0.597	0.564
0327_C3	0.537	0.602	0.580	0.573
Avg.	0.581	0.603	0.577	0.587

Table 2: The Dice scores for each conversation.

Conv. ID	Kappa Score			
	A_1-A_2	A_1-A_3	A_2-A_3	Avg.
0313_C1	0.460	0.450	0.594	0.501
0320_C1	0.615	0.456	0.479	0.517
0320_C4	0.555	0.425	0.432	0.471
0323_C3	0.701	0.449	0.431	0.527
0326_C1	0.426	0.489	0.463	0.459
0326_C2	0.364	0.451	0.405	0.407
0326_C4	0.593	0.479	0.559	0.544
0327_C2	0.474	0.417	0.434	0.442
0327_C3	0.406	0.432	0.474	0.438
Avg.	0.510	0.450	0.475	0.478

Table 3: The Kappa scores for each conversation.

We analyze tags with the low agreement. As a result, we found the pairs that are frequently annotated different tags between two annotators: ("Positive Feedback", "Agreement"), ("Monologue", "Inform"), and ("Monologue", "Stalling"). These pairs showed a different tendency for annotators to give priority to the tags. To solve this problem, we need to modify our annotation manual.

The pairs of ("Inform", "Answer") and ("Inform", "Agreement") also caused the decrease in agreement. In this situation, the utterance length was generally long. In addition, the utterances contained a reference to other speakers and expressions of an opinion. We instructed annotators to give priority to annotate "Inform" for the situation on the annotation manual because "Inform" is the superclass. However, annotators could not select "Inform" tag. Annotators might not be able to understand the instruction for this situation well. Therefore, we should simplify our annotation manual.

Tag	Ratio (%)	Number of tags
Inform	43.82	1,446
Positive Feedback	18.76	619
Mismatch	6.81	225
Monologue	6.00	198
Vague	5.97	197
Question	5.06	167
Agreement	4.15	137
Answer	3.91	129
Self Completion	3.76	124
Stalling	1.76	58

Table 4: Distribution of tags (the top 10 tags).

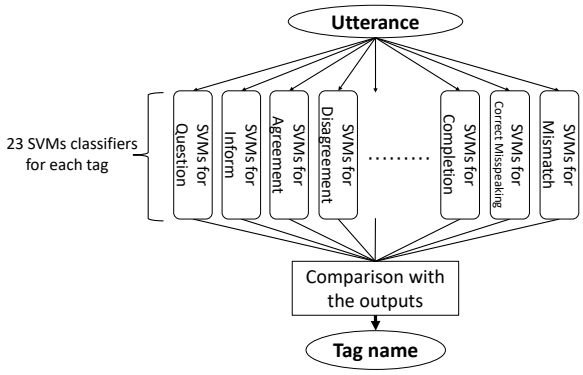

Figure 1: The overview of the classification method.

4.5. Annotation unification

We unify the annotation results by the three annotators for each conversation on the basis of following steps:

Step:1 If the three annotators annotate the same tag to an utterance, we select the tag as the final tag of the utterance.

Step:2 If we cannot select a tag in **Step1**, we select the majority tag (annotated by two annotators) as the final tag of the utterance.

Step:3 If we cannot select a tag in **Step2** and the annotators annotate "Agreement", "Disagreement", or "Answer" tags, we replace them with the superclass "Inform". Then we repeat **Step2**.

Step:4 If we cannot select any tag after **Step3**, we select the new tag, "Mismatch".

Table 4 shows the distribution of the tags in descending order of the number of tags annotated for the Kyutech corpus. As mentioned in Table 4, the sum of "Inform" and "Positive Feedback" accounts for 60% of the total number of tags. The previous studies reported that some tags are unevenly distributed in corpora (Godfrey et al., 1992; Shriberg et al., 2004). For the Kyutech corpus, we obtained the similar result.

The number of "Mismatch" tags was approximately seven percent. The length of the utterances with "Mismatch" was almost short and the utterances contained little information to select a suitable tag. In contrast, some long utterances were also annotated "Mismatch". The annotators could not annotate suitable tags for the utterances because the utterances contained a large amount of information and several tags were considered as candidate tags. Future work should address this issue.

5. Dialogue Act Classification

In this section, we explain a dialogue act classification task and a method that estimates a suitable tag of each utterance. In the current annotation, an utterance contains several tags by the unification between the three annotators (in Section 4.5.). In this experiment, we select one tag from unified tags as the correct class of a dialogue act for the utterance. Specifically, we give priority to GPF tag if an utterance contains GPF tag and DSCF tags[3]. In the Kyutech corpus, the transcripts contain "<laugh>" that indicates laughing. As a preprocessing, we excluded utterances with only "<laugh>". Finally, the number of utterances after the preprocessing is 3,120.

ID	Description
f_1	The word n-grams
f_2	The part-of-speech (POS) n-grams
f_3	The detailed POS n-grams
f_4	The function word n-grams
f_5	The POS n-grams of function word
f_6	The detailed POS n-grams of function word
f_7	The three head-word n-grams
f_8	The three last-word n-grams
f_9	The number of morphemes
f_{10}	The agreement between the current speaker and the last speaker
f_{11}	The agreement between the current speaker and the next speaker
f_{12}	The position in topic segments
f_{13}	The previous dialogue act tags

Table 5: The utterance features for dialogue act classification.

In this paper, we propose a method that estimates the 23 classes consisting of 22 dialogue act tags as mentioned in Section 4.2. and "Mismatch", using Support Vector Machines (SVMs) (Vapnik, 1999). As shown in Figure 1, we apply the SVM classifiers for each class to our method. Our proposed method utilizes each result of the SVMs and outputs the class with the highest reliability as the correct class for an utterance. We use utterance features and audio-visual features.

5.1. Utterance feature

In this section, we explain the utterance features. We use MeCab[4] as a Japanese morphological analyzer and define

[3] Our annotation has a possibility of annotating DSCF tags only for an utterance. However, we did not find such a case.

[4] http://taku910.github.io/mecab/

the stopwords as the words that appear only once in training data. Table 5 lists our features and the descriptions. We adopt the bag-of-ngrams representation[5] described in Section 2. We introduce other n-grams representations, such as part-of-speech (POS) and function words. We use the two POS features: the simple POS information such as "noun" and "verb" and the more detailed POS information such as "proper noun" and "quantity". O'Shea et al. (2010) have been reported that function words are more useful to characterize dialogue acts than content words. Therefore, we utilize the n-grams focused on function words. We replaced all content words with the token "$*$", and then obtained the n-grams of function words.

Next, we describe the other features. We assume that the speaker switching information is valuable because some dialogue acts relate to information exchange between speakers, such as "Inform" and "Suggest". Therefore, we use the speaker switching information for the previous speaker or the following speaker. We also focus on the topic information. The Kyutech corpus contains the three topic tags for each utterance: one main tag and two additional tags. In the Kyutech corpus, the same tags often continue across several utterances, and we call it the topic segment in this paper. We believe that the utterance position in topic segments provides a clue to dialogue act classification. This is because it is possible to capture the relation with the previous utterance. For example, it is considered that "Agreement" and "Positive Feedback" tend to occur in the same topic sequence. In contrast, "Question" and "Suggest" tend to change the current topic and cause a new topic. Therefore, we apply the utterance position in topic segments to our method. We also utilize the dialogue act tags of the previous utterances as the feature. According to the results of the preliminary experiment, we set the number of previous utterances to four.

5.2. Audio-visual feature

We propose the audio-visual features obtained by the audio-visual data for the conversations. In this paper, we extract the acoustic information and the body pose information from the audio-visual data and apply them to our method. Table 6 lists the audio-visual features and the descriptions.

First, we explain the acoustic features. Pitch is a fundamental frequency that is a basic acoustic feature of speech. Power denotes a sound power level of an utterance. We consider that the audio information is useful because it captures some information not obtained in the features. For example, the meaning of the word "Okay" depends on the pitch at the end of the word. If the pitch at the end of the word is raised, the word means the confirmation of understanding for other speakers. In the case of lowering of the pitch, the word generally means a favorable answer for other speakers. For these reasons, we introduce the pitch features and the power features as the acoustic information. We compute the pitch and power values of the 60-msec interval for an utterance by using wavesufer[6]. We use the average, standard deviation, maximum, and minimum values of pitch and power values. To capture intonations at the end of utterances, we also use the distribution of pitch and power values and the ratio of pitch and power values in the last two 200-msec intervals. Dialogue acts are considered to be related to time information. This is because speakers tend to need more time for speaking with complicated dialogue acts like "Question" and "Inform" than speaking utterances with simple dialogue acts like "Agreement". As the other acoustic features, we introduce the speech rate and the time interval between a current utterance and its previous utterance.

Next, we explain the body pose features. Head nods and head shakes are gestures that indicate agreement and disagreement respectively. Therefore, we use the moving distances of a nose between frames as the moving distances of the speaker's head. We obtain the body pose information of speakers by using OpenPose (Cao et al., 2017; Simon et al., 2017; Wei et al., 2016). In addition, we focus on the shoulder motions of speakers. Ezen-Can et al. (2015) have utilized the one-hand-to-face feature for dialogue act classification. We consider that the shoulder angles and widths are captured from speaker's positions for other speakers. Therefore, we apply the shoulder angles and widths to our method.

ID	Description
f_{14}	The pitch
f_{15}	The variation in pitch changes between intervals
f_{16}	The distribution of pitch
f_{17}	The ratio of two pitch values in the last two intervals
f_{18}	The power
f_{19}	The variation in power changes between intervals
f_{20}	The distribution of power
f_{21}	The ratio of two power values in the last two intervals
f_{22}	The speech rate
f_{23}	The time interval between current and previous utterances
f_{24}	The moving distance of the speaker's head
f_{25}	The shoulder angle
f_{26}	The ratio of two average speaker's shoulder angles during utterance and whole conversation.
f_{27}	The shoulder widths (distances between one and another shoulder points)
f_{28}	The ratio of the two average shoulder widths in the utterance and in the whole conversation.

Table 6: The audio-visual features for dialogue act classification.

5.3. Experiment for dialogue act classification

We evaluated our method on the Kyutech corpus with nine-fold cross-validation for the nine conversations. In other words, we evaluated one test conversation with the model that was generated from the other conversations and repeated this process for all conversations. In this experiment, we first investigated the best combination of the utterance features and defined the model with the features as the base-

[5] We use the three types of n-grams for each n-gram feature: uni-gram, bi-gram, and tri-gram.

[6] http://www.speech.kth.se/wavesurfer

Feature	Macro Avg.	Micro Avg.
Base	63.99	63.72
+ Pitch (f_{14-17})	64.03	63.69
+ Power (f_{18-21})	63.93	63.62
+ Speech rate (f_{22})	64.07	63.81
+ Time interval (f_{23})	63.89	63.72
+ Head (f_{24})	63.80	63.72
+ Shoulder (f_{25-28})	63.72	63.49
+ All	63.91	63.59
+ Best	**64.50**	**64.17**

Table 7: Micro and macro-averaged accuracy of the nine conversations.

Conv. ID	# of utterances	# of important utterances
0313_C1	759	240
0320_C1	505	124
0326_C1	502	76
0326_C2	566	160
0327_C2	284	52
0323_C3	324	102
0327_C3	445	118
0320_C4	637	69
0326_C4	487	98

Table 8: The number of utterances and important utterances of each conversation in the Kyutech corpus.

line. Then, we added each audio-visual feature to the baseline and evaluated them. Table 7 shows micro-averaged accuracy and macro-averaged accuracy of the nine conversations.

In Table 7, "Base" is the best combination of the utterance features (f_1, f_3, f_7, f_8, f_{10}, and f_{13}). "All" denotes all audio-visual features with "Base". "Best" denotes the best combination of the audio-visual features (f_{14}, f_{15}, f_{19}, f_{22}, and f_{24}) with "Base". Although each audio-visual feature was not always effective, we found that some combinations of the audio features contributed to the improvement of the accuracy. In contrast, the models with the body pose features did not work well on the whole. Since the speakers of the Kyutech corpus were talking on the conversation, the speaker's poses were almost unchanged. In this work, we focused on the body pose information during speaking only. However, head nods and head shakes could occur before and after speaking. In our future work, we analyze the speaker's motions in the overall conversation more deeply.

6. Contribution of Dialogue Act for Other Task

One of the main purposes of this study is dialogue act annotation for a multi-party conversation corpus because dialogue acts are useful for conversation understanding. We expect that our dialogue act annotation also contributes to other tasks. Since the Kyutech corpus is annotated for a summarization task, we evaluate the effectiveness of the dialogue act tags through the extractive summarization task.

Yamamura and Shimada (2018) have annotated the extractive summaries for the Kyutech corpus. They annotated whether each utterance is important or not for the conversation and defined the important utterances as the extractive summaries for each conversation. Table 8 shows the number of utterances and important utterances of each conversation in the Kyutech corpus. They proposed an extractive summarization method using Conditional Random Fields (Lafferty et al., 2001). We consider that dialogue acts are valuable for extractive summarization. This is because the utterances with "Question" and "Inform" tags are considered to be essential utterance for understanding the conversation. Therefore, we apply our dialogue acts to the summarization method.

We used the summarization method in (Yamamura and Shimada, 2018). They applied the several features to the summarization method. In addition to these features, we applied the dialogue act feature using dialogue act tags of utterances to the method. We evaluate the original method and the method with our dialogue act feature. The following are the feature names and the descriptions used in this experiment:

- Features in (Yamamura and Shimada, 2018)
  - Normalized utterance position: The utterance numbers normalized by the total number of utterances in the conversation.
  - Speaker information: The ranking of the number of utterances for each speaker.
  - Topic Tag: The three topic tags for each utterance.
  - Length: The number of morphemes.
  - Utterance timing: The time difference between the start times of the current utterance and the previous utterance.
- Our additional feature
  - Dialogue act: The unified dialogue act tag as mentioned in Section 5.

Since the extractive summaries are annotated to the original utterance-units and the dialogue act tags are annotated to the long utterance-units, we annotate the dialogue act tags to the original utterance-units. We divide the long utterance-units into the original utterance-units while we keep the dialogue act tags of the long utterance-units.

We evaluated the original method and our method with nine-fold cross-validation for the nine conversations. We also investigated some combination patterns of the features on both methods. We computed the precision, recall rates, and F-measure for each conversation and took an average of the overall scores (macro-averaging).

Table 9 shows the experimental result. We found that our method was best among the models with the dialogue act feature when we use all the features except for the feature of speaker information. The scores of our method were slightly higher than the scores of the original method on

	Precision	Recall	F-measure
Yamamura and Shimada (2018)	0.411	0.294	0.326
Our method	**0.465**	**0.303**	**0.354**

Table 9: Macro-averaged precision, recall, and F-measure scores.

all criteria. In particular, our method contributed to the improvement of the precision and the F-measure score. Although we tested for statistical significance in all scores for each criteria using a paired t-test, there was no significant difference. We focus on analyzing the effectiveness of dialogue acts more deeply in future work.

7. Conclusions

In this paper, we explained the annotation task of dialogue acts for a Japanese multi-party conversation corpus. We defined the 23 dialogue act tags based on ISO standard 24617-2. The Dice score between the annotators was 0.587 and not sufficient. Therefore, we need to improve an agreement in the dialogue act annotation and address this issue by revising our annotation manual in future work.

In this work, we examined the models with the audio-visual features for dialogue act classification. As a result, we obtained 64.5 percent of the accuracy on the best model. Future work will mainly focus on the relationship between the other audio-visual information and dialogue acts. We also need to apply other approaches, such as neural network-based approaches (Ortga and Vu, 2017) and unsupervised approaches (Jo et al., 2017).

We also evaluated the effectiveness of the dialogue act tags through a conversation summarization task. As a result, the performance of our method with dialogue acts was slightly higher than that of the method without dialogue acts in the summarization task for the Kyutech corpus. Future work analyzes the effectiveness of dialogue acts more deeply and evaluate the effectiveness of dialogue acts for other tasks.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 17H01840.

References

Allen, J. and Core, M. (1997). Draft of DAMSL: Dialog act markup in several layers.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1997). The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

Araki, M., Itoh, T., Kumagai, T., and Ishizaki, M. (1999). Proposal of a standard utterance-unit tagging scheme. *Journal of Japanese Society for Artificial Intelligence (In Japanese)*, 14(2):251–260.

Bangalore, S., Fabbrizio, G. D., and Stent, A. (2006). Learning the structure of task-driven human-human dialogues. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 201–208.

Boyer, K. E., Grafsgaard, J. F., Ha, E. Y., Phillips, R., and Lester, J. C. (2011). An affect-enriched dialogue act classification model for task-oriented dialogue. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hosida, K., Petukhova, V., Porescu-Belis, A., and Trauim, D. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 430–437.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J. C., and Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–32.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. In *Computational Linguistics*, volume 22, pages 249–254.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Den, Y., Koiso, H., Maruyama, T., Maekawa, K., Takanashi, K., Enomoto, M., and Yoshida, N. (2010). Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1483–1486.

Ezen-Can, A., Grafsgaard, J. F., Lester, J. C., and Boyer, K. E. (2015). Classifying student dialogue acts with multimodal learning analytics. In *Proceedings of the 5th International Conference on Learning Analytics and Knowledge*.

Ferschke, O., Gurevych, I., and Chebotar, Y. (2012). Behind the article: Recognition dialogue acts in Wikipedia talk page. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 777–786.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE-ICASSP 1992)*, volume 1, pages 517–520.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (2014). Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*,

pages 928–939.

Hiraoka, T., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2014). Construction and analysis of a persuasive dialogue corpus. In *Proceedings of the 5th International Workshop on Spoken Dialog Systems (IWSDS 2014)*, pages 213–223.

Jo, Y., Yoder, M. M., Jang, H., and Rosé, C. P. (2017). Modeling dialogue acts with content word filtering and speaker preference. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. Technical report, Institue of Cognitive Science, University of Colorado.

Kim, S. N., Cavedon, L., and Baldwin, T. (2010). Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 282–289.

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of Japanese. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, volume 2, pages 947–952.

Moldovan, C., Rus, V., and Graesser, A. C. (2011). Automated speech act classification for online chat. In *Proceedings of the 22nd Midwest Artificial Intelligence and Cognitive Science Conference (MAICS 2011)*.

Murray, G., Carenini, G., and Ng, R. T. (2010). Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference (INLG 2010)*, pages 105–113.

Omuya, A., Prabhakaran, V., and Rambow, O. (2013). Improving the quality of minority class identification in dialog act tagging. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 802–807.

Ortga, D. and Vu, N. T. (2017). Neural-based context representation learning for dialog act classification. In *Proceedings of the 18th SIGdial Workshop on Discourse and Dialogue (SIGDIAL 2017)*.

O'Shea, J., Bander, Z., and Crockett, K. (2010). A machine language approach to speech act classification using function words. In *Proceedings of the 4th KES International Conference on Agent and Multi-agent Systems: Technologies and Applications, Part II*, volume 6071 of *KES-AMSTA'10*, pages 82–91.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvev, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue (SIGDIAL 2004)*, pages 97–100.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Ess-Dykema, C. V., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. In *Computational Linguistics*, volume 26, pages 339–373.

Surendran, D. and Levow, G.-A. (2006). Dialog act tagging with support vector machines and hidden markov model. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 1950–1953.

Tavafi, M., Mehdad, Y., Joty, S., Carenini, G., and Ng, R. (2013). Dialogue act recognition in synchronous and asynchronous conversations. In *Proceedings of the 14th SIGdial Workshop on Discourse and Dialogue (SIGDIAL 2013)*, pages 117–121.

Vapnik, V. N. (1999). *Statistical Learning Theory*. Wiley.

Verbree, D., Rienks, R., and Heylen, D. (2006). Dialogue-act tagging using smart feature selection: Result on multiple corpora. In *Spoken Language Technology Workshop*, pages 70–73. IEEE.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *CVPR*.

Yamamura, T. and Shimada, K. (2018). Annotation and analysis of extractive summaries for the Kyutech corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3216–3220.

Yamamura, T., Shimada, K., and Kawahara, S. (2016). The Kyutech corpus and the topic segmentation using a combined method. In *Proceedings of the 12th Workshop on Asian Language Resources*, pages 95–104.