JAIST Annotated Corpus of Free Conversation

Kiyoaki Shirai, Tomotaka Fukuoka
Japan Advanced Institute of Science and Technology, Nextremer Co., Ltd
1-1 Asahidai, Nomi, Ishikawa, Japan, 1-30-13 Narimasu, Itabashi, Tokyo, Japan
kshirai@jaist.ac.jp, tomotaka.fukuoka@nextremer.com

Abstract

This paper introduces an annotated corpus of free conversations in Japanese. It is manually annotated with two kinds of linguistic information: dialog act and sympathy. First, each utterance in the free conversation is annotated with its dialog act, which is chosen from a coarse-grained set consisting of nine dialog act labels. Cohen's kappa of the dialog act annotation between two annotators was 0.636. Second, each utterance is judged whether the speaker expresses his/her sympathy or antipathy toward the other participant or the current topic in the conversation. Cohen's kappa of sympathy tagging was 0.27, indicating the difficulty of the sympathy identification task. As a result, the corpus consists of 92,031 utterances in 97 dialogs. Our corpus is the first annotated corpus of Japanese free conversations that is publicly available.

Keywords: Annotated corpus, Dialog Act, Sympathy, Free Conversation

1. Introduction

In recent years, study of an open domain conversation system or free conversation system, which can freely talk with users, has attracted much research interest (Libin and Libin, 2004; Higashinaka et al., 2014a; Higashinaka et al., 2014b). Unlike a task oriented dialog system, an open domain conversation system can chat with users about various topics. Such systems can be used as robotic pets or nursing care robots that can enrich our daily life.

Obviously, language resources are indispensable for the study of free conversation systems. Especially, corpora of free conversations annotated with some linguistic information are valuable. However, for the Japanese language, there is no annotated corpus in the domain of free conversations that is publicly available.

This paper introduces an annotated corpus of free conversations in Japanese, called "JAIST Annotated Corpus of Free Conversations"[1]. It consists of dialogs of two participants, where they freely talk about various topics. Each utterance in the dialogs is annotated with two kinds of tags. One is a dialog act (or speech act), which is the type of utterance that represents the speaker's intention. The other is sympathy. In this paper, sympathy means that the speaker shows interest in the current topic in the conversation. We will report in detail how to construct the corpus as well its statistics. Furthermore, two usage cases of this corpus will be reported: the classification of the dialog acts and the identification of sympathy.

The rest of the paper is organized as follows. Section 2 will discuss related work. Section 3 will report the details of our corpus, including annotation guidelines, the size of the corpus, the distribution of the tags, and inter-annotator agreement. Section 4 will describe two case studies using our corpus. Finally, the paper is concluded in Section 5 .

2. Related Work

One of the well-known dialog corpora is the Switchboard Dialog Act Corpus (University of Colorado at Boulder, 2000). Stolcke et al. (2000) reported that it consisted of a substantial portion of the Switchboard corpus (Godfrey et al., 1992), which was a collection of human–human conversational telephone speech. A total of 1,155 conversations were labeled, comprising 205,000 utterances and 1.4 million words. The SWBD-DAMSL tag set, which was based on the Dialogue Act Markup in Several Layers (DAMSL) tag set (Core and Allen, 1997), was used for annotation. It consisted of 42 dialog act labels. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus (ICSI, 2004) is a collection of 72 hours of speech from 75 naturally occurring meetings (Shriberg et al., 2004). Eleven general tags and thirty-nine specific tags were used for the dialog act annotation. The corpus contained 180,000 utterances with dialog acts in total.

As for Japanese, several studies have been devoted to supervised learning of dialog act classification in free conversations. Isomura et al. (2009) applied Conditional Random Field (CRF) with the features of word unigrams and bigrams that occurred twice or more in the training data as well as the dialog act of the previous utterance. They reported that the accuracy of their method was 75.77%. Meguro et al. (2013a) identified dialog acts in conversations on a microblog, i.e., Twitter. It was a challenging task, since a wide variety of topics and words were used and the sentences were often ungrammatical. The features for machine learning were the n-gram of the semantic classes derived from a thesaurus as well as the n-gram of the characters. Higashinaka et al. (2014a) proposed an open-domain conversational system and developed a dialogue-act estimation module in it. To identify the dialog act of a user's utterance, a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) was trained with features such as the character n-grams, word n-grams, and semantic categories. They used a dialog tag set consisting of 33 dialog acts, which was proposed by Meguro et al. (2013b). The accuracy of the dialog act classification was 45%, while the inter-annotator agreement was 59%.

In these studies, corpora annotated with dialog acts were constructed and used for training classifiers as well as the evaluation of the proposed methods. However, these cor-

[1] "JAIST" is the acronym for the affiliation of the first author.

pora were not released. There is no annotated public corpus of free conversations in Japanese, which can be used by every researcher.
One of the important characteristics in a free conversation is the sympathy of a speaker for topics in a conversation (Anderson and Keltner, 2002; Higashinaka et al., 2008). In the past studies of dialog systems, sympathy was usually considered as one of the dialog act classes (Minami et al., 2010; Sekino and Inoue, 2010; Meguro et al., 2013b; Shriberg et al., 2004). On the other hand, sympathy is independently tagged to the utterance in our corpus, since we think that sympathy plays an important role in a free conversation system. Topics in free conversations are not fixed but could be changed by the speakers at any time. To make the conversation natural and smooth, however, the free conversation system can not arbitrarily change the topics. It is uncomfortable for the user if the system were to suddenly change the topic when the user wants to continue talking on the current topic, or if the system were to keep the same topic when the user is bored and does not want to talk on that topic any more. The sympathy of the user is one of the useful clues to guess what would be a good time for changing the topic. If the user shows sympathy for the current topic, the system should continue the conversation with the same topic. On the other hand, if the user does not display sympathy, the system should provide another topic. The peculiarity of our corpus is that the dialog act and sympathy are tagged separately.

3. The Construction of the Corpus

3.1. The Raw Corpus

The Nagoya University conversation corpus (Fujimura et al., 2012b) was chosen as the texts for the annotation. It consists of transcriptions of 120 free conversations between two or more participants. The total duration of the dialog is about 100 hours. Each utterance was transcribed by hand. The corpus was developed by Fujimura et al. (2012) at Nagoya University, but has now been released by the National Institute for Japanese Language and Linguistics (NINJAL). It is freely available at the web site of NINJAL. Not all, but 97 dialogs, where only two people participate in the conversation, were chosen for the annotation. The statistics of the sub-corpus are shown in Table 1. It indicates that each dialog is rather long.

Number of dialogs	97
Number of utterances	92,020
Average number of utterances per dialog	949

Table 1: Statistics of corpus

3.2. Overview of Annotation

Each utterance in our corpus has the following information.

- Speaker ID

  An identification number of the speaker. It has been already annotated in the Nagoya University conversation corpus.

- Turn taking

  A flag indicating whether the speaker has changed or not. This was automatically annotated.

- Dialog act

  A dialog act of an utterance.

- Sympathy tag

  A tag that represents whether the speaker shows sympathy or antipathy.

We have manually annotated the utterances with the dialog act and sympathy tags.

3.3. Annotation with Dialog Act

Nine dialog acts were formulated for the annotation. Table 2 shows these dialog acts, their definitions, and examples of utterances.
The annotation guidelines of the dialog act are as follows:

- Decide on the dialog act for the utterance, considering its context. To consider the context, the utterances should be annotated in the same order as they occur in the dialog.

- Choose one dialog act for each utterance. When two or more dialog acts are possible, choose the primary one. However, it is allowed to assign multiple dialog acts, but only when the annotator cannot decidedly choose one dialog act.

- Guidelines for Response(Yes/No)
  "Response(Yes/No)" should be the tag for an utterance that consists of a short phrase such as "yes" or "no." The annotation of "Response(Yes/No)" is restricted to the context just after an utterance of "Question(Yes/No)," "Question(What)," "Confirmation," or "Request." It is not necessary to always give the tag "Response(Yes/No)" to the response to an utterance of a "Question(Yes/No)". If the speaker replies to a yes/no question by a declarative sentence, not "Response(Yes/No)," but "Response(Declaration)" should be chosen.

- Guideline for Response(Declaration)
  The tag "Response(Declaration)" should be applied to a speaker's response presented by declarative sentences. The annotation of "Response(Declaration)" is restricted to the context just after the utterance of "Question(Yes/No)," "Question(What)," "Confirmation," or "Request." It is not necessary to always apply the tag "Response(Declaration)" to the response to an utterance of "Question(What)." If the speaker replies to a wh*-question by "yes" or "no," not "Response(Declaration)," but "Response(Yes/No)" should be chosen.

Table 3 presents the numbers of dialog acts and their proportions in the constructed corpus. The most frequent dialog act is "Self-disclosure," followed by "Backchannel," "Response(Declaration)," and "Question(Yes/No)." On the

ID	dialog act	definition	example
d_1	Self-disclosure	Speaker expresses his/her opinion or fact.	*In short, he has a meager vocabulary.*
d_2	Question(Yes/No)	Speaker asks a yes/no question.	*Can I turn on the light for a moment?*
d_3	Question(What)	Speaker asks a question (what, who, when, how, etc.).	*Which country did he come from?*
d_4	Response(Yes/No)	Speaker replies to a question with a short phrase.	*Yes, please.*
d_5	Response(Declaration)	Speaker replies to a question with a declarative sentence.	*Yeah, so, he came from Brazil.*
d_6	Backchannel	Speaker gives a short response	*Uh-huh.*
d_7	Filler	Speaker utters a short phrase to just fill in the time.	*Wow.*
d_8	Confirmation	Speaker confirms the hearer's understanding.	*Really?*
d_9	Request	Speaker requests something of a hearer.	*Please introduce that person to me.*

Table 2: Definition of dialog acts

ID	dialog act	frequency	proportion
d_1	Self-disclosure	53,701	58.35%
d_2	Question(Yes/No)	6,430	6.99%
d_3	Question(What)	3,950	4.29%
d_4	Response(Yes/No)	2,130	2.31%
d_5	Response(Declaration)	7,508	8.16%
d_6	Backchannel	9,216	10.01%
d_7	Filler	4,405	4.79%
d_8	Confirmation	3,940	4.28%
d_9	Request	751	0.82%

Table 3: Distribution of dialog acts

other hand, the utterance of a "Request" seldom appears in free conversations. Its proportion is only 0.8%.

Although it is allowed to assign two or more dialog acts to one utterance, the annotators are required to assign one dialog act as much as possible. As a result, only 11 utterances were annotated with two dialog acts.

A dialog act is assigned by one annotator for each utterance, although two annotators work for the construction of the whole annotated corpus. To check the inter-annotator agreement, only three dialogs were annotated by two annotators. The agreement ratio is 0.773 and Cohen's kappa is 0.636.

3.4. Annotation with Sympathy Tag

Three sympathy tags are defined for the annotation.

Sympathy

This is assigned if a speaker expresses sympathy with the other participant's previous utterance or a current topic in a conversation.

(example) *That is great!*

Antipathy

This is assigned if a speaker expresses antipathy towards the other participant's previous utterance or a current topic in a conversation.

(example) *I can't agree with you.*

Neutral

This is chosen if a speaker expresses neither sympathy nor antipathy.

The annotation guidelines for the sympathy tags are as follows:

- The utterance is likely to be sympathetic or antipathetic when the previous utterance is subjective. For example, if one participant expresses his/her opinion, sentiment, or impression, another participant may express sympathy or antipathy.
- The "Sympathy" tag can be assigned when an utterance shows sympathy or approval. However, if a speaker just shows agreement with the other participant, the "Sympathy" tag is not assigned.

  (example)
  P1: *We will arrive at around 11 tonight.*
  P2: *Yes.*

  Speaker P2 agrees with P1, but does not show any sympathy with the fact said by P1. Therefore, P2 should be tagged as "Neutral."
- The "Antipathy" tag can be assigned when an utterance shows antipathy or bad feeling. As with the sympathy tag, if a speaker just disagrees with the other participant, the "Antipathy" tag is not assigned.

  (example)
  P1: *He is a cunning fellow, isn't he?*
  P2: *I don't think so.*

  The speaker disagrees with P1's comment, but does not express any ill feeling. Therefore, "Neutral" should be chosen for the utterance of P2.

Table 4 shows the number of occurrences of each sympathy tag and its proportion in the constructed corpus. It is found that the number of sympathy and antipathy tags is quite small. Most of utterances are tagged as "Neutral."

sympathy tag	frequency	proportion
Sympathy	1,067	1.16%
Antipathy	222	0.24%
Neutral	90,731	98.60%

Table 4: Distribution of sympathy tags

The same annotators who worked for the dialog act tagging also annotated the corpus with the sympathy tags. Two annotators gave the sympathy tags to only three dialogs, this was done to measure the inter-annotator agreement; the rest

of the dialogs were annotated by one annotator. Cohen's kappa of the sympathy annotation is 0.27. This indicates the difficulty of the sympathy identification task. In particular, the judgment of implicit sympathy tends to be inconsistent. Here implicit sympathy means a sympathetic utterance without any linguistic features that clearly indicate sympathy. Sympathy in such an utterance can be identified by its context and/or prosody. Intuitively, prosody is an important clue to judging the sympathy of a speaker. However, only the transcriptions of the utterances were used: no speech information was used for the annotation. In the future, the definition of "sympathetic utterance" should be clarified in order to have better guidelines for consistent annotation.

4. The Case Studies

Two case studies of the JAIST Annotated Corpus of Free Conversations are reported. In the first case, the corpus was used as labeled data to train a model for the classification of the dialog acts. In the second case, the corpus was used for training a classifier of sympathy identification.

4.1. Classification by Dialog Acts

The task considered in this subsection is to classify a given utterance by its dialog act. Fukuoka and Shirai proposed a method for dialog act classification and evaluated their method on the JAIST Annotated Corpus of Free Conversations (Fukuoka and Shirai, 2017). The present paper briefly introduces their method and the results of the experiment.

4.1.1. Method

Figure 1 shows an overview of the classification model. For each dialog act, a binary classifier that judges whether a given utterance has a dialog act is trained. An optimized set of the features is empirically determined for each classifier of the dialog act. The binary classifiers also calculate the reliability of the judgment. After nine classifiers are applied, one dialog act is chosen by considering the judgment and the reliability of the nine classifiers. L2-regularized logistic regression is used for training the binary classifiers. The probability given by LIBLINEAR (Fan et al., 2008) is used as the reliability of the classification.

Table 5 shows a list of the types of the proposed features. These features were designed by manually analyzing free conversations. The linguistic characteristics of the dialog acts were carefully considered for the feature engineering.

The set of the feature types is optimized for each dialog act by removing ineffective feature types. Figure 2 shows the algorithm of the feature type optimization. E stands for the initial feature set consisting of all proposed features, while E' stands for the optimized one. The function $f(X)$ denotes the F-measure of binary classification of the dialog act for the development data, where the binary classifier was trained with the feature set X. For each feature type f_i in E, if $f(E \setminus \{f_i\})$ is less than or equal to $f(E)$, f_i is regarded as effective and added to E'. After checking all the feature types in E, the optimization is terminated if no more feature type is removed (line 6). Otherwise, we update E by E' (line 11) and repeat the same procedure.

$f(E')$ sometimes becomes lower than $f(E)$ at line 7. This means that removal of each ineffective feature increases the

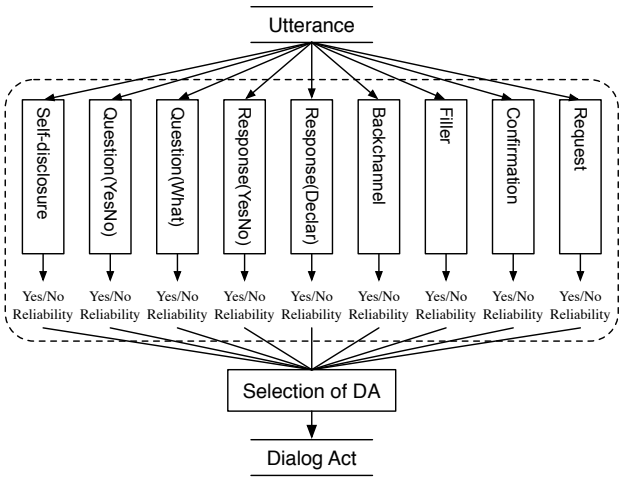

Figure 1: Overview of dialog act classification

F-measure, but the simultaneous removal of two or more ineffective features causes a decrease of the F-measure. In such a case, we choose the most ineffective feature f_x, where the difference between $f(E \setminus \{f_i\})$ and $f(E)$ is the maximum, and remove only f_x from E (line 9).

Input: $E = \{f_1, f_2, \cdots, f_n\}$
Output: E'

```
1:  while true do
2:      E' ← ∅
3:      for all f_i ∈ E do
4:          if f(E) ≥ f(E\{f_i}) then E' ← E'∪{f_i}
5:      end for
6:      if E = E' then return E'
7:      if f(E) > f(E') then
8:          f_x = arg max_{f_i} f(E \ {f_i}) − f(E)
9:          E ← E \ {f_x}
10:     else
11:         E ← E'
12:     end if
13: end while
```

Figure 2: Algorithm of feature set optimization

After the judgment and its reliability for nine dialog acts have been obtained, one dialog act is chosen by Equation (1). We compare the reliability of each classifier ($r(d_i)$) and choose the dialog act with the maximum reliability.

$$\hat{d} = \arg\max_{d_i} r(d_i) \tag{1}$$

In addition, two additional procedures are introduced, as in Equation (2).

$$\hat{d} = \begin{cases} \arg\max_{d_i} w_i \cdot r(d_i) & \text{if rank(1)=Self-disclosure} \\ \text{classify}(\text{rank}(1), \text{rank}(2)) & \\ \qquad \text{if } \{\text{rank(1),rank(2)}\} = \{d_6, d_7\} \text{ or } \{d_2, d_8\} & \\ \arg\max_{d_i} r(d_i) & \text{if otherwise} \end{cases} \tag{2}$$

f_1: word n-gram	f_{14}: keyword of "Backchannel"
f_2: word n-gram in previous utterance	f_{15}: keyword of "Filler"
f_3: content word	f_{16}: key phrase of "Request" at the end of utterance
f_4: content word in previous utterance	f_{17}: key phrase of "Backchannel" at the end of utterance
f_5: word n-gram at the end of utterance	f_{18}: previous dialog acts of the hearer
f_6: word n-gram at the end of previous utterance	f_{19}: previous dialog acts of the speaker
f_7: sequence of function words at the end of utterance	f_{20}: length of utterance
f_8: sequence of function words at the end of previous utterance	f_{21}: turn taking
f_9: pair of word n-gram at the end of current and previous utterances	f_{22}: existence of content word
f_{10}: pair of sequence of function words at the end of current and previous utterances	f_{23}: repetition of content words (1)
f_{11}: keyword of question	f_{24}: repetition of content words (2)
f_{12}: keyword of "Question(Yes/No)"	f_{25}: repetition of content words (3)
f_{13}: keyword of "Response(Yes/No)"	f_{26}: utterance formed by one content word
	f_{27}: utterance formed by one function word
	f_{28}: duplication of words in current utterance

Table 5: Features for dialog act classification

In the preliminary experiment, many utterances were wrongly classified as "Self-disclosure," because the reliability of "Self-disclosure" is usually much higher than the others. This is because the proportion of "Self-disclosure" utterances is the greatest in our corpus, as shown in Table 3. To alleviate the imbalance of the reliability of the dialog acts, the weighted reliability of the dialog acts are compared when the highest ranked dialog act is "Self-disclosure" as in the first case in Equation (2). The weights of the dialog acts, w_i, are optimized on the development data. Furthermore, it was found that two specific pairs of dialog acts are very difficult to distinguish: d_6(Backchannel) & d_7(Filler) and d_2(Question(Yes/No)) & d_8(Confirmation). Therefore, if the first and second ranked dialog acts are (d_6,d_7) or (d_2,d_8), other classifiers that select one of these dialog acts are used to make the final determination of the dialog act, as in the second case of Equation (2). The classifiers are separately trained with the union of the optimized feature sets of two dialog acts.

4.1.2. Experiment

In the experiment, the JAIST Annotated Corpus of Free Conversations was randomly divided into three sets, as shown in Table 6.

	# of dialog	# of utterance
Training set	77	74,228
Development set	10	8,984
Test set	10	8,694

Table 6: Data sets

Table 7 presents the precision (P), recall (R) and F-measure (F) of the classification of each dialog act, as well as their macro- and micro-averages. BL_s stands for the baseline, where a unique feature set was used for training the classifier. The feature set was chosen by the algorithm of Figure 2 so that the F-measure of the classification of all dialog acts is maximized. Pro_p is the proposed method that simply chooses the dialog act with the highest reliability, as in Equation (1). Pro_b is the proposed method that chooses the dialog act by Equation (2).

Pro_b achieved satisfactory results, i.e., 82.5% of the micro-average of the F-measure. Furthermore, the proposed methods outperformed the baseline. It was confirmed, by the McNemar's test, that the difference between BL_s and Pro_b was statistically significant at the 5% level. Although the feature set was optimized for the individual dialog acts, the F-measures of d_8(Confirmation) and d_9(Request) are still low. This may be because the numbers of occurrences of d_8 and d_9 are too small, as shown in Table 1. That is, there are much fewer positive samples than negative samples. One way to resolve this is to apply a technique to learn a classifier from an imbalanced training dataset, such as SMOTE(Chawla et al., 2002).

4.2. Identification of Sympathy

The task to be discussed in this subsection is to judge whether a given utterance is sympathetic or not. Fukuoka and Shirai used the JAIST Annotated Corpus of Free Conversations to develop a method to identify sympathetic utterances in free conversations. The present paper briefly introduces their case study. For more details, see (Fukuoka and Shirai, 2015).

SVM[2] was applied to train a binary classifier to judge whether the given utterance was sympathetic. The features used for training are summarized in Table 8.

F_{rw1} and F_{rw2} are introduced since speakers often show their sympathy by repeating a word in the previous utterance of the other. F_{rw1} simply checks whether the same content word appears in both the current and the previous utterance. On the other hand, F_{rw2} more strictly checks the presence of a repetition of content words: F_{rw2} is activated if either of the conditions below is fulfilled.

- The last predicative word in the previous utterance is also found in the current utterance.
- There is only one content word in the current utterance and it also appears in the previous utterance.

[2] LIBSVM (Chang and Lin, 2011) is used for training SVM.

	BL_s			Pro_p			Pro_b		
	P	R	F	P	R	F	P	R	F
d_1 Self-disclosure	0.851	0.951	0.898	0.852	0.953	0.900	0.859	0.949	0.901
d_2 Question(Yes/No)	0.762	0.745	0.753	0.754	0.751	0.752	0.760	0.753	0.756
d_3 Question(What)	0.787	0.672	0.725	0.807	0.689	0.743	0.797	0.706	0.749
d_4 Response(Yes/No)	0.874	0.900	0.887	0.876	0.880	0.878	0.876	0.880	0.878
d_5 Response(Declaration)	0.819	0.772	0.795	0.818	0.812	0.815	0.811	0.839	0.824
d_6 Backchannel	0.768	0.730	0.748	0.758	0.724	0.741	0.790	0.699	0.741
d_7 Filler	0.608	0.412	0.491	0.607	0.399	0.482	0.627	0.553	0.588
d_8 Confirmation	0.634	0.318	0.424	0.678	0.265	0.381	0.687	0.276	0.394
d_9 Request	0.724	0.214	0.331	0.773	0.173	0.283	0.643	0.184	0.286
Macro-average	0.759	0.635	0.672	0.769	0.628	0.664	0.761	0.649	0.680
Micro-average	0.819	0.819	0.819	0.821	0.821	0.821	0.825	0.825	0.825

Table 7: Result of dialog act classification

F_{ng}	Word unigrams, bigrams, and trigrams of the current and previous utterances
F_{len}	Length of utterance
F_{tu}	Turn taking
F_{rw1}	Repetition of word (1)
F_{rw2}	Repetition of word (2)
F_{rc1}	Repetition of semantic class (1)
F_{rc2}	Repetition of semantic class (2)
F_{da}	Dialog act
F_{end}	Sequence of function words at the end of utterance

Table 8: Features for identification of sympathy

F_{rc1} and F_{rc2} are similar to F_{rw1} and F_{rw2}, but not the repetition of the word but of the semantic class or concept derived from the Japanese thesaurus *Bunruigoihyo* (NINJAL, 2004) is considered.

Combination features, i.e., arbitrary pairs of the features in Table 8, are also used for training.

Since the numbers of the word n-gram feature (F_{ng}) and of combination features are extremely high, a simple feature selection procedure is introduced. The correlation between a sympathy class and a feature f_i is measured by its χ^2 value. Th word n-gram feature and combination feature are discarded when χ^2 is less than certain thresholds T_{ng} and T_{comb}, respectively. These thresholds are optimized on the development data.

In the experiment, the utterances with the "Sympathy" tag are regarded as sympathetic, while the utterances with the "Antipathy" and "Neutral" tags are regarded as non-sympathetic. The JAIST Annotated Corpus of Free Conversations is divided into training, development, and test data, as shown in Table 6. In addition, since the number of sympathetic utterances is small, as shown in Table 4, a balanced dataset including the same number of positive and negative samples was also used for evaluation. It was made by keeping all positive samples and randomly choosing an equal number of negative samples for the training, development, and test datasets.

Tables 9 and 10 show the precision (P), recall (R) and F-measure (F) of the imbalanced (original) and balanced datasets. The baseline system is the classifier trained with only the word n-gram feature (F_{ng}). In both the balanced and imbalanced datasets, Fukuoka's method outperformed the baseline. However, the performance on the imbalanced dataset was not good. Note that the identification of sympathetic utterances in our corpus is very difficult since the positive samples are much fewer than the negative samples. Through an error analysis, we found a few major causes of errors. First, errors often crop up when a previous utterance is long and consists of several sentences. Even when a speaker talks for a long time, the hearer may show sympathy with only one sentence of the many sentences spoken by the speaker. The current system extracts the features from the previous (long) utterance, but most of them are irrelevant to sympathy identification. As a result, too many irrelevant features cause classification errors. Second, many false negative errors are caused by the feature F_{end}, which is a typical sentence-end expression that indicates the sympathy of a speaker. However, such sentence-end expressions do not always appear in sympathetic utterances. Especially, an utterance including this feature is not sympathetic when it and its previous utterance are short. In such a case, the number of extracted features is small and the feature F_{end} causes the misclassification of a non-sympathetic utterance as sympathetic.

	P	R	F
Baseline (F_{ng})	0.23	0.11	0.15
(Fukuoka and Shirai, 2015)	0.28	0.13	0.18

Table 9: Results of sympathy identification on imbalanced data

	P	R	F
Baseline (F_{ng})	0.80	0.73	0.76
(Fukuoka and Shirai, 2015)	0.81	0.76	0.80

Table 10: Results of sympathy identification on balanced data

5. Conclusion

In this paper, we introduced the JAIST Annotated Corpus of Free Conversations. It was the corpus of the free conversa-

tions between two participants. It was manually annotated with two kinds of linguistic information: a dialog act and a sympathy tag. Our corpus is the first annotated corpus of free conversations in Japanese that is publicly available. It is distributed by the non-profit organization *Gengo Shigen Kyokai* (literally "Language Resources Association")[3]. Furthermore, two case studies of this corpus (a classification of the dialog acts and an identification of sympathy) were also presented.

In the future, we plan to enrich the annotation of this corpus with two additional tags: topic shift and subjectivity. A topic shift tag for a conversation may be useful for a study of a dialog control manager in a free conversation system. Another additional piece of information is the subjectivity of the content of the utterance. In general, a subjective utterance reflects more the feeling or emotion of the speaker than an objective utterance. Subjectivity may be an important clue to make a conversation between a user and a system smooth and natural.

6. Bibliographical References

Anderson, C. and Keltner, D. (2002). The role of empathy in the formation and maintenance of social bonds. *Behavioral and Brain Sciences*, 25(1):21–22.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):Article 27.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Core, M. G. and Allen, J. F. (1997). Coding dialogs with the DAMSL annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Fujimura, I., Chiba, S., and Ohso, M. (2012). Lexical and grammatical features of spoken and written Japanese in contrast: Exploring a lexical profiling approach to comparing spoken and written corpora. In *Proceedings of the VIIth GSCP International Conference: Speech and Corpora*, pages 393–398.

Fukuoka, T. and Shirai, K. (2015). Identification of sympathy in free conversation. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 1–9.

Fukuoka, T. and Shirai, K. (2017). Dialog act classification using features intrinsic to dialog acts in an open-domain conversation. *Journal of Natural Language Processing*, 24(4):523–546. (in Japanese).

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

Higashinaka, R., Dohsaka, K., and Isozaki, H. (2008). Effects of self-disclosure and empathy in human–computer dialogue. In *Proceedings of the Spoken Language Technology Workshop*, pages 109–112.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (2014a). Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 928–939.

Higashinaka, R., Kobayashi, N., Hirano, T., Miyazaki, C., Meguro, T., Makino, T., and Matsuo, Y. (2014b). Syntactic filtering and content-based retrieval of Twitter sentences for the generation of system utterances in dialogue systems. In *Proceedings of the 5th International Workshop on Spoken Dialog Systems*, pages 113–123.

Isomura, N., Toriumi, F., and Ishii, K. (2009). Auto-tagging method for evaluating non-task-oriented dialogue agent. *The Transactions of the Institute of Electronics, Information and Communication Engineers A*, 92(11):795–805.

Libin, A. V. and Libin, E. V. (2004). Person–robot interactions from the robopsychologists' point ofview: The robotic psychology and robotherapy approach. In *Proceedings of the IEEE 92*, volume 92, issue 11, pages 1789–1803.

Meguro, T., Higashinaka, R., Sugiyama, H., and Minami, Y. (2013a). Dialogue act tagging for microblog utterances using semantic category patterns. *IPSJ SIG Technical Report*, 2013-SLP-98(1):1–6. (in Japanese).

Meguro, T., Minami, Y., Higashinaka, R., and Dohsaka, K. (2013b). Learning to control listening-oriented dialogue using partially observable Markov decision processes. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(4):Article 15.

Minami, Y., Higashinaka, R., Dohsaka, K., Meguro, T., and Maeda, E. (2010). Trigram dialogue control using POMDPs. In *Proceedings of the Spoken Language Technology Workshop, IEEE*, pages 336–341.

Sekino, T. and Inoue, M. (2010). Fine-grained dialog tagging of utterances. In *Tohoku-Section Convention of Information Processing Society of Japan, 10-6-B3-2*. (in Japanese).

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

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

7. Language Resource References

National Institute for Japanese Language and Linguistics (NINJAL). (2004). *Bunrui Goi Hyo*. Dainippon Tosho.

[3] http://www.gsk.or.jp/en/catalog/gsk2017-b/

Fujimura, Itsuko and Chiba, Shoju and Ohso, Mieko. (2012b). *Nagoya University conversation corpus*. http://mmsrv.ninjal.ac.jp/nucc/.

International Computer Science Institute (ICSI). (2004). *ICSI Meeting Corpus*. http://www1.icsi.berkeley.edu/Speech/mr/.

University of Colorado at Boulder. (2000). *Switchboard Dialog Act Corpus*. http://compprag.christopherpotts.net/swda.html.