JDCFC: A Japanese Dialogue Corpus with Feature Changes

Tetsuaki Nakamura[†], **Daisuke Kawahara**[†‡]
†Kyoto University, ‡JST PRESTO
Yoshida-Honmachi, Sakyo-ku, Kyoto, Japan
tnakamura@nlp.ist.i.kyoto-u.ac.jp, dk@i.kyoto-u.ac.jp

Abstract

In recent years, the importance of dialogue understanding systems has been increasing. However, it is difficult for computers to deeply understand our daily conversations because we frequently use emotional expressions in conversations. This is partially because there are no large-scale corpora focusing on the detailed relationships between emotions and utterances. In this paper, we propose a dialogue corpus constructed based on our knowledge base, called the Japanese Feature Change Knowledge Base (JFCKB). In JFCKB and the proposed corpus, the feature changes (mainly emotions) of arguments in event sentences (or utterances) and those of the event sentence recognizers (or utterance recognizers) are associated with the event sentences (or utterances). The feature change information of arguments in utterances and those of the utterance recognizers, replies to the utterances, and the reasonableness of the replies were gathered through crowdsourcing tasks. We conducted an experiment to investigate whether a machine learning method can recognize the reasonableness of a given conversation. Experimental result suggested the usefulness of our proposed corpus.

Keywords: emotion, commonsense knowledge, dialogue corpus, crowdsourcing, neural network, LSTM

1. Introduction

In recent years, the importance of dialogue understanding systems has been increasing because interactive interfaces handling a natural language such as smart speakers have become popular. However, it is difficult for computer programs to understand our daily conversations because we frequently use emotional expressions in conversations. This is partially because there are no large-scale corpora focusing on the detailed relationships between emotions and utterances.

Many dialogue corpora have been developed because they are essential language resources needed to train and evaluate machine learning methods. For instance, the Dialog State Tracking Challenge (DSTC) dataset is used to estimate a user's goal in a spoken dialog system (Kim et al., 2016). While the DSTC corpus is made from manually transcribed Skype dialogues, there are corpora that consist of conversations extracted from SNS websites (Ritter et al., 2010; Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015). There is also a corpus based on a collection of logs extracted from Ubuntu-related chat rooms that is mainly composed of technical support conversations (Lowe et al., 2015; Lowe et al., 2016). The Dialogue Breakdown Detection Challenge database is a corpus used to detect incorrect replies generated by dialogue systems (Higashinaka et al., 2016). Although these corpora are very useful resources for understanding actual human-human dialogue or human-machine dialogue, it is difficult to understand a speaker's/replier's motivations because such corpora do not record a speaker's/replier's inner state (in particular, his/her emotions). Even if such corpora include some keywords as clues for inferring a speaker's/replier's inner state, it is necessary to develop a method to extract inner state information from the corpora, which are composed of raw text.

Dialogue corpora that include various feature changes of arguments in utterances and the reactions to speakers can be used to understand a speaker's motivations. In the dialogue corpus used in Hasegawa et al. (2013), each utterance is annotated with the addressee's emotions. Although this corpus is useful for understanding the relationships between utterances and emotions in a conversation, the understandable relationships are limited to the addressee's direct emotional expressions because the corpus is annotated based on an explicit keyword list. In the keyword list, explicit keywords such as "afraid" and "happy" are manually associated with emotions "fear" and "joy" respectively. There are other relationships between utterances and emotions in conversations, such as relationships that concern the speaker's emotion, addressee's emotion, and emotions of any arguments in the utterances. We think that these relationships are also important for understanding speakers' motivations in conversations (especially in emotional conversations) in addition to the relationships used in Hasegawa et al. (2013). It is necessary to construct corpora designed to treat both of explicit and implicit emotional expressions because explicit emotional expressions are not always used in daily conversations. For example, when someone says "my wife hit my child," he probably wants to convey some kinds of information about his "surprise," "anger," and "disgust."

In this paper, we propose a dialogue corpus constructed based on our knowledge base, called the Japanese Feature Change Knowledge Base (*JFCKB*) (Nakamura and Kawahara, 2018). In the proposed corpus, feature changes (mainly emotions) of arguments in utterances and those of the utterance recognizers (i.e., utterers and addressees) are associated with the utterances. Because of the lack of large-scale corpora focusing on detailed relationships between emotions and utterances, the dialogue corpora constructed based on JFCKB will be useful for developing robots and software that can handle natural language. To validate the usefulness of our dialogue corpus, we conducted an experiment to investigate whether a machine learning method can recognize the reasonableness of

Sentence	Case (word)	Probability	Trigger utterance	Reply
Tsuma ga kodomo wo hippataku. (My wife hits my child.)	*ga* (nominative) (wife)	*joy* (+, −, UNC) = (0, 1, 0)	*Tsuma ga kodomo wo hippataita yo.* (My wife hit my child.)	*Hidoi ne.* (How terrible.) (○, ×, UNK) = (1, 0, 0)
		:		
	wo (accusative) (child)	*anger* (+, −, UNC) = (1, 0, 0)		*Nande*? (Why?) (○, ×, UNK) = (1, 0, 0)
		:		
	ni (dative) (NULL)	N/A		*Bouryoku ha ikenai yo.* (Violence is bad.) (○, ×, UNK) = (1, 0, 0)
	reader (NULL)	*disgust* (+, −, UNC) = (0.99, 0, 0.01)		
		:		:

Table 1: Example of the proposed dialogue corpus (JDCFC). The actual corpus is in Japanese. Each sentence has various feature changes for *readers* and three cases (*ga*, *wo*, and *ni*), which are Japanese language syntactic cases that roughly correspond to the nominative, accusative, and dative, respectively. *Readers* are not arguments in the sentence. The left three columns (sentence, case, and probabilities) in JDCFC are the same information of JFCKB. The trigger utterance corresponds to the event sentence. Replies are given probabilities for their reasonableness. In the "Probability" column, symbols +, −, and UNC denote *increased*, *decreased*, and *unchanged*, respectively. In the "Reply" column, symbols ○, ×, and UNK denote *reasonable*, *unreasonable*, and *unknown*, respectively.

a given conversation (i.e., a dialogue). This corpus is for Japanese.

2. Proposed Dialogue Corpus Based on a Feature Change Knowledge Base

Since the publication of our previous work (Nakamura and Kawahara, 2016), we have been constructing a knowledge base of argument feature changes in event sentences with controlled granularity. We call this knowledge base JFCKB (Nakamura and Kawahara, 2018). In JFCKB, arguments in event sentences are associated with various feature changes caused by the events. The feature changes of sentence readers (i.e., sentence recognizers) are also associated with the sentences in the current version of JFCKB. For example, in the case of "my wife hits my child," "my child" is associated with some feature changes, such as *increase in pain*, *increase in anger*, *increase in disgust*, *decrease in joy*, and *decrease in trust*. The sentence is also associated with feature changes such as *increase in a reader's anger* and *increase in a reader's disgust*. We gathered such information through crowdsourcing.

In this paper, we propose a dialogue corpus constructed using JFCKB to address the lack of large-scale corpora that focus on detailed relationships between emotions and utterances. We first briefly explain JFCKB, then we explain the proposed dialogue corpus.

2.1. JFCKB

JFCKB is composed of three types of information for event sentences, as shown the left three columns (sentence, case, and probabilities) in Table 1. As shown in the table, for each sentence, arguments in the sentence are associated with various features. Each feature in each argument has a triple (*increased*, *decreased*, and *unchanged*) whose values are probabilities.

We controlled the granularity of knowledge (i.e., features), and designed the 47 features shown in Table 2. These features were designed to correspond with *basic level categories* in cognitive linguistics (Rosch et al., 1976; Taylor, 1995) as much as possible. This design was based on a traditional emotion study (Plutchik, 1980), Japanese thesauri (Ikehara, 1997; NINJAL, 2004), sentiment analysis studies (Tokuhisa et al., 2008; Tokuhisa et al., 2009), and features used in the VerbCorner project (Hartshorne et al., 2014). Although our final version of JFCKB will have all the features listed in Table 2, emotional and sensory features are mainly investigated in the current study.

Event sentences for JFCKB were created as follows. Step 1: the 200 most frequent verbs, 1,000 most frequent verbs, and all verbs were respectively extracted from the *Kyoto University Web Document Leads Corpus* (*KWDLC*) (Hangyo et al., 2012)[1], *Kyoto University Case Frames* (*KUCF*) (Kawahara and Kurohashi, 2006)[2], and the Japanese version of the Winograd Schema Challenge dataset (*JWSC*) (Levesque, 2011; Shibata et al., 2015). KWDLC is a Japanese text corpus that comprises 5,000 documents (15,000 sentences) with annotations of morphology, named entities, dependencies, predicate-argument structures including zero anaphora and coreferences. KUCF is a database of case frames automatically con-

[1] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KWDLC
[2] http://www.gsk.or.jp/en/catalog/gsk2008-b/

Category	Sub category	Feature
physical	form	**length**, **size**, width, thickness (around) thickness (depth)
	color	redness, orangeness yellowness, greenness blueness, purpleness brownness, whiteness blackness, brightness
	touch	**temperature**, rigidness roughness, stickiness
	smell	goodness, badness
	sound	silence
	taste	sweetness, sourness bitterness, astringency hotness (not temperature)
	density	denseness
	amount	quantity
mental	emotion	**joy**, **trust**, **surprise** **disgust**, **fear**, **sadness** **anger**, **anticipation**
	evaluation	polarity
sensory	sensory	**pain**, **sleepiness** **tiredness**
relation	relation	interaction, possession physical contact physical force existence social relationship
	position	closeness

Table 2: Features assumed in this study. These features were decided by considering various studies such as traditional psychological studies, studies on cognitive development of infants, Japanese thesauri, sentiment analysis studies, and VerbCorner project. Features with bold fonts have been investigated so far.

structed from a corpus of 10 billion Japanese sentences taken from the Web. Case frames describe what kinds of nouns are related to each verb. Many Japanese verbs have some meanings. Examples are shown in Table 3. In KUCF, each case frame is composed of case frame ID, verb, cases, nouns filled in the cases, frequencies of the nouns in the Web corpus. KUCF has about 110,000 predicates with 5.4 case frames on average for each predicate. WSC (JWSC) dataset is basically composed of two sentences including one anaphor, two antecedent candidates, and a correct antecedent. Step 2: For each case frame of the extracted verbs in KUCF, representative event sentences were created. These sentences were composed of representative words for three cases in Japanese grammar (*ga*, *wo*, and *ni*: these cases roughly correspond to nominative, accusative, and dative, respectively). Each representative word was one of the most frequent words for each case in KUCF. Each case frame has one or some representative sentences because each case in the case frame has one or a few representative words. Step 3: we conducted a crowdsourcing task to discard nonsense sentences from the created sentences. In total, 1,559 crowdsourcing workers participated in this task and were asked to answer whether the presented sentences were comprehensible or not.

After event sentence creation, we conducted a crowdsourcing task to gather feature changes of arguments in the event sentences. In this task, in addition to feature changes of the arguments in sentences, we attempted to gather those of sentence readers. Crowdsourcing workers were asked to answer the feature changes of the arguments presented in the sentences (e.g., *anger* of "my child" in "my wife hits my child") or those of the workers themselves (e.g., *anger* of each worker himself/herself when he/she reads the presented sentence "my wife hits my child"). In total, 33,683 workers participated in this task. As a result, feature changes of 9,073 event sentences (types) were acquired (including 975 verbs (types) and 19,052 arguments (tokens) (4,882 types), 5,647 case frames (types)).

For every crowdsourcing task described above, we calculated probabilities that each answer would be selected by crowdsourcing workers based on the aggregation method proposed by Whitehill et al. (2009). Unlike majority voting, this method calculates the probabilities based on worker agreements.

2.2. JDCFC: JFCKB Based Dialogue Corpus

The proposed dialogue corpus, *JDCFC*, is based on JFCKB. JDCFC is composed of records for event sentences, as shown in Table 1. Each record is composed of feature change information for three Japanese syntactic cases (*ga*, *wo*, and *ni*: roughly corresponding to nominative, accusative, and dative, respectively) and sentence readers (sentence recognizers), the trigger utterance corresponding to the event sentence, and reply candidates to the trigger utterance with their probabilities of reasonableness. In common with JFCKB, the features in Table 2 are used in JDCFC.

Trigger utterances in JDCFC were created based on event sentences in JFCKB. 2,428 sentences were extracted from JFCKB as trigger utterances because these sentences have one or more feature changes caused by the events in total. For each utterance, we regarded a feature whose probability of *increased* or *decreased* is 0.75 or more as a feature changed by the event. The difference between the trigger utterances and event sentences in JFCKB is the expression of the predicates. In the trigger utterances, predicates are in past tense and attached with the postposition *yo*[3], while those in JFCKB are in present tense. This arrangement of predicates in trigger sentences is based on our speculation that (in Japanese, at least) such sentences are more natural as utterances than those in present tense in conversations. After trigger utterance creation, JDCFC was constructed using two crowdsourcing tasks.

The first task was to acquire replies to trigger utterances. In this task, crowdsourcing workers were given a trigger sentence and asked to answer appropriate replies, as shown in Figure 1. A total of 8,370 workers participated in the

[3] *Yo* is a postposition in the Japanese grammar, which represents familiarity. For example, many Japanese people have more friendly feeling towards "*Ii tenki da yo* (It is nice weather)" than "*Ii tenki da*."

Verb: case frame ID	Case	Word
yaku: yaku1 (bake)	*ga* (nominative)	*watashi* (I): 114, *haha* (mom): 75, *musume* (daughter): 74, ...
	wo (accusative)	*pan* (bread): 54076, *ke-ki* (cake): 31693, *niku* (meat): 14059, ...
	de (tools/ingredients)	*koubo* (yeast): 888, *be-kari-* (bakery): 768, *o-bun* (oven): 515, ⋯
yaku: yaku2 (have difficulty)	*ga* (nominative)	*mina* (all persons): 23, *sensei* (teacher): 11, *hito* (person): 8, ...
	wo (accusative)	*te* (hand): 26449
	ni (dative)	*kodomo* (child): 168, *musuko* (son):108, ...
yaku: yaku3 (burn)	*ga* (nominative)	*daitouryou* (president): 1, śhidousya (mentor): 1, ...
	ni (dative)	CD:13812, DVD:12200, ...

Table 3: Case frame examples. Each row denotes one case frame. In the "Word" column, each number denotes the frequency of the noun in the Web corpus.

Please answer your reply to the speaker.	
speaker's utterance	**My wife hit my child.**
your reply	

Figure 1: Question used for a crowdsourcing task to gather dialogue replies. Although this example is written in English, it was written in Japanese in the actual task.

Is the pair below a reasonable conversation? (speaker B's utterance is the reply to speaker A) (select *yes*, *no*, or *neither yes nor no*)	
speaker A	**My wife hit my child.**
speaker B	**Why?**

Figure 2: Question used for a crowdsourcing task to determine the reasonableness of replies to trigger utterances. Although this example is written in English, the actual task was written in Japanese.

task. After the task, for each trigger utterance, overlapping replies and extremely low-quality replies (such as copied and pasted trigger utterances and empty replies) were discarded. As a result, 23,196 replies (2,428 types of trigger utterances) were acquired. Hence, the average number of replies for each trigger utterance is approximately 9.6.

The second task was to determine the reasonableness of replies acquired in the first task. Crowdsourcing workers different from those in the first task were given a dialogue (a trigger utterance and one of its acquired replies) and asked to judge whether the given pair is reasonable (Figure 2). In total, 5,605 workers participated in this task, and ten workers were assigned to each dialogue. As a result, judgements on the reasonableness of the 23,196 dialogues were acquired. The probabilities of the reasonableness were calculated using the aggregation method proposed by Whitehill et al. (2009), just as used in JFCKB construction. The example shown in Table 1 presents typical acquired data. We regarded dialogues whose probability of the reasonableness was more than or equal to 0.8 as reasonable dialogues. The number of such dialogues was 22,357. We used the 22,357 dialogues in the evaluation experiment described in the next section.

2.3. Natural Dialogues versus Semi-artificial Dialogues

Our dialogue corpus is constructed semi-artificially because trigger utterances are collected artificially and replies are made through crowdsourcing tasks. It is desirable to construct a corpus through gathering naturally occurring dialogues from somewhere and annotating them with feature change information. However, to do this ideal construction procedure, there are problems described below at least: (1) it is difficult to get large scale dialogues, (2) there is no guarantee that utterances gathered from natural dialogues always have simple structures. Especially, in Japanese conversations, sentences often have no predicates or no essential arguments such as subjects and objects because ellipses are frequently used.

We are planning to start with applying feature change information to understanding of simple texts. Therefore, we made artificial trigger utterances based on event sentences in JFCKB that are based on highly frequent words in various Web documents.

3. Evaluation

In a dialogue, it is often necessary to use emotional knowledge to make a response to an utterance. For example, it is difficult to make a response "It must be hard for you" to an utterance "I dispatched my students to the battlefront" when we do not know the utterer's emotions (emotions of "I") such as *increase in disgust*, *decrease in joy* and *increase in fear*. Moreover, it is difficult to understand the reasonableness of the reply when we do not know such emotional information. Considering this, to validate the usefulness of JDCFC, we conducted an experiment to investigate whether a machine learning method could appropriately estimate the reasonableness of given dialogues.

3.1. Data

We made 22,357 sets composed of a positive example and a negative one. As described in section 2.2., all the trigger utterances in JDCFC had more than or equal to one feature change caused by the events. As described in the previous section, we used the 22,357 reasonable dialogues. That is, each of the 22,357 replies was regarded as a positive example S_r^+ of the corresponding trigger utterance S_t. For each S_r^+, we randomly selected one reply that do not overlap with S_r^+ from replies of the other trigger utterances as a negative example S_r^- of S_t.

3.2. Model

In this evaluation experiment, we used a bidirectional long short-term memory (LSTM) model as shown in Figure 3. This model is based on the model used in Lowe et al. (2015). In our model, the reasonableness of a given dialogue is calculated as follows: (1) when a trigger utterance and reply candidate pair is given, the word embeddings of words in the sentences are used as input for the LSTM. The trigger utterance and reply are sequences of words, denoted as $w_{t1}, w_{t2}, ..., w_{tn}$ and $w_{r1}, w_{r2}, ..., w_{rm}$, respectively. The dimension of word embeddings is 128. (2) The hidden layer vectors of the trigger utterance and the reply are calculated by the LSTM. The dimension of the hidden layer vectors is 100. (3) The sentence vector $\boldsymbol{v_t}$ is calculated by concatenating $\boldsymbol{h_t^f}$, $\boldsymbol{h_t^b}$, and $\boldsymbol{v_f}$, where $\boldsymbol{v_t}$, $\boldsymbol{h_t^f}$, $\boldsymbol{h_t^b}$, and $\boldsymbol{v_f}$ denote the sentence vector of the trigger utterance, the final forward hidden layer vector (i.e., the hidden vector corresponding to w_{tn}), the final backward hidden layer vector (i.e., the hidden vector corresponding to w_{t1}), and the feature change vector of the predicate in the trigger utterance. The sentence vector $\boldsymbol{v_r}$ is calculated by concatenating $\boldsymbol{h_r^f}$ and $\boldsymbol{h_r^b}$, where $\boldsymbol{v_r}$, $\boldsymbol{h_r^f}$, and $\boldsymbol{h_r^b}$ denote the sentence vector of the reply, the final forward hidden layer vector (i.e., the hidden vector corresponding to w_{rm}), and the final backward hidden layer vector (i.e., the hidden vector corresponding to w_{r1}). We used eight emotional feature changes in Table 2. The feature change vector of each predicate is composed of four feature change vectors of cases (*ga*, *wo*, *ni*, and *reader*). Each feature change vector of each case is composed of a triple (*increased*, *decreased*, and *unchanged*) whose values are probabilities. Therefore, in this study, the dimension of predicate feature change vectors is 96. (4) Output o is the reasonableness of the given dialogue and is calculated by Equation (1), where f, $\boldsymbol{W}$, and b denote the activation function, weighting matrix, and bias, respectively.

$$o = f(\boldsymbol{v_t^T W v_r} + b) \quad (1)$$

In this study, we used a sigmoid function as the activation function. We used Adam (Kingma and Ba, 2017) for optimizing the parameters.

3.3. Evaluation Settings

We compared the following two types of models. (1) Baseline model: this model is the same as that of Figure 3, except that the model does not use the feature change information in the figure. (2) Proposed model: this model is a bidirectional LSTM model shown in Figure 3. Note that the number of layers of LSTM is two in both of the models although that in Figure 3 is one.

In the training phase, each model was trained to output 1 and 0 for the given positive and negative examples, respectively. In the test phase, for each test set, when the output value for the positive example was greater than that for the negative example, the output was regarded as a correct estimation.

All of the 22,357 sets were used in this evaluation, where 80%, 10%, and 10% of the sets were used as training data, development data, and test data, respectively. For this evaluation, we conducted 10 training epochs.

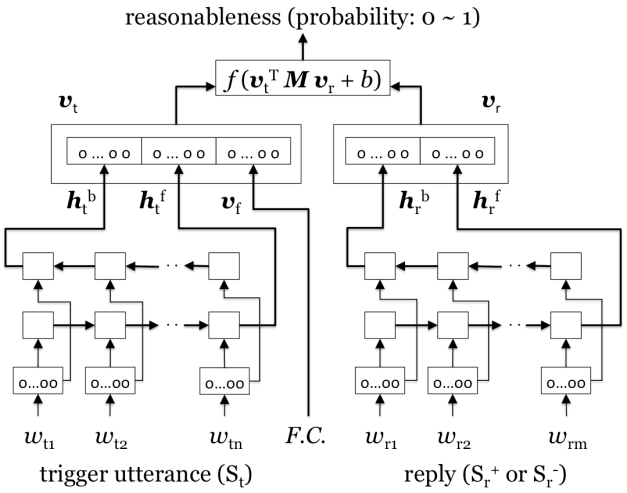

Figure 3: Bidirectional LSTM model to evaluate the usefulness of our dialogue corpus. Symbol *F.C.* denotes the feature change vector of the predicate in the trigger utterance.

	Baseline	**Proposal**
Accuracy	64.2%	71.0%

Table 4: Evaluation result.

3.4. Results and Discussions

The results of the evaluation are shown in Table 4. As shown in the table, the proposed model outperformed the baseline model. This result suggests the usefulness of JDCFC; that is, the feature change information benefits the estimation of the reasonableness of a given dialogue.

One example of the cases for which feature change information worked well is the pair (Trigger: "I aimed to become a surgeon." Positive reply: "Keep trying!" Negative reply: "I will regret it"). As for this trigger utterance, feature changes of "I" and the utterance recognizers are shown in Figure 4. As shown in the figure, the subject ("I") is associated with *increase in anticipation*, *increase in fear*, *increase in joy*, and *increase in trust*. The utterance recognizers (the utterer and/or the addressee) are associated with *increase in anticipation*. Considering this association, it seems that the reply reflects feelings of the arguments in the trigger utterance and the utterance recognizers. We speculate that the feature change information influences the estimation of the reasonableness of a given dialogue when the replies are associated with the feature changes of arguments in the trigger utterances or those of the utterance recognizers.

4. Conclusion

In this study, we constructed a dialogue corpus focusing on detailed relationships between emotions and utterances. In the corpus, both of the emotional changes of the arguments in trigger utterances and those of the utterance recognizers are associated with the trigger utterances. The corpus is based on our feature change knowledge base in which arguments in various event sentences are associated with various feature changes caused by the events. In the knowledge

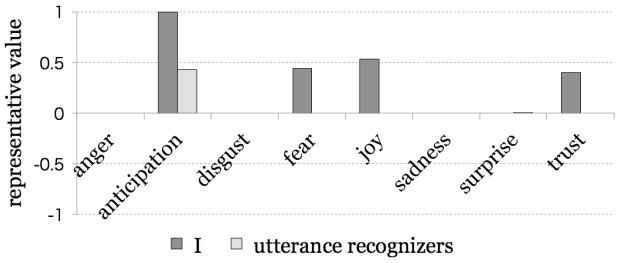

Figure 4: Feature changes of "I" in the utterance "I aimed to become a surgeon" and the utterance recognizers. The representative value of each feature is the weighted average of the feature, where weights of *increased*, *decreased*, and *unchanged* are $+1$, -1, and 0, respectively.

base, emotional change information of the event sentence readers is also associated with the sentences. The feature change information in the knowledge base was gathered through crowdsourcing tasks.

To construct the dialogue corpus, we created trigger utterances based on event sentences in our knowledge base. After creating the trigger utterances, we conducted crowdsourcing tasks to gather replies to the trigger utterances and to determine the reasonableness of the replies.

To validate the usefulness of our dialogue corpus, we conducted an experiment to investigate whether a machine learning method could appropriately estimate the reasonableness of a given dialogue based on our dialogue corpus. In this experiment, we compared two types of bidirectional LSTM models. The difference between these models was that whether the emotional change information was used. As a result of the experiment, the model using the emotional change information outperformed the other. This result suggests the usefulness of our proposed corpus.

5. Acknowledgements

This work was supported by JST PRESTO Grant Number JPMJPR1402, Japan.

6. Bibliographical References

Hangyo, M., Kawahara, D., and Kurohashi, S. (2012). Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of the 26th Pacific Asia Conference on Language Information and Computing*, pages 535–544.

Hartshorne, J. K., Bonial, C., and Palmer, M. (2014). The verbcorner project: Findings from phase 1 of crowdsourcing a semantic decomposition of verbs. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 397–402.

Hasegawa, T., Kaji, N., Yoshinaga, N., and Toyoda, M. (2013). Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 964–972.

Higashinaka, R., Funakoshi, K., Kobayashi, Y., and Inaba, M. (2016). The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3146–3150.

Ikehara, S. (1997). *Nihongo Goi Taikei*. Iwanami Syoten, Tokyo.

Kawahara, D. and Kurohashi, S. (2006). Case frame compilation from the web using high-performance computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, pages 1344–1347.

Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., Henderson, M., and Yoshino, K. (2016). The fifth dialog state tracking challenge. In *Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT 2016)*, pages 511–517.

Kingma, D. and Ba, J. (2017). Adam: A method for stochastic optimization. In *arXiv:1412.6980v9*.

Levesque, H. J. (2011). The Winograd Schema Challenge. In *Proceedings of AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Lowe, R., Pow, N., Serban, I. V., and Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference*, pages 285–294.

Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). On the evaluation of dialogue systems with next utterance classification. In *Proceedings of the SIGDIAL 2016 Conference*, pages 264–269.

Nakamura, T. and Kawahara, D. (2016). Constructing a dictionary describing feature changes of arguments in event sentences. In *Proceedings of the 4th Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 46–50.

Nakamura, T. and Kawahara, D. (2018). JFCKB: Japanese feature change knowledge base. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

NINJAL. (2004). *Bunrui Goihyo*. Dainippon Tosho, Tokyo.

Plutchik, R., (1980). *A General Psychoevolutionary Theory of Emotion, 1*, pages 3–33. Academic Press.

Ritter, A., Cherry, C., and Dolan, B. (2010). Unsupervised modeling of twitter conversations. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL*, pages 172–180.

Ritter, A., Cherry, C., and Dolan, B. (2011). Data-driven response generation in social media. In *Proceedings of EMNLP 2011*, pages 583–593.

Rosch, E., Mervisa, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

Shang, L., Lu, Z., and Li, H. (2015). Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1577–1586.

Shibata, T., Kohama, S., and Kurohashi, S. (2015). Nihongo Winograd Schema Challenge no kouchiku to bun-

seki (in Japanese). In *Proceedings of NLP2015*, pages 493–496.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL*, pages 196–205.

Taylor, J. R. (1995). *Linguistic Categorization: Prototypes in Linguistic Theory*. Clarendon Press.

Tokuhisa, R., Inui, K., and Matsumoto, Y. (2008). Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 881–888.

Tokuhisa, R., Inui, K., and Matsumoto, Y. (2009). Emotion classification using massive examples extracted from the web. *Journal of Information Processing (in Japanese)*, 50(4):1365–1374.

Whitehill, J., Ruvolo, P. L., Ting-fan Wu, J. B., and Movellan, J. R. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, et al., editors, *Advances in Neural Information Processing Systems 22*, pages 2035–2043. Curran Associates, Inc.