A supplementary feature set for sentiment analysis in Japanese dialogues

PETER LAJOS IHASZ, Graduate School of Information Science and Engineering, Ritsumeikan University[1]
MATE KOVACS, Graduate School of Information Science and Engineering, Ritsumeikan University
IAN PIUMARTA, College of Information Science and Engineering, Ritsumeikan University
VICTOR V. KRYSSANOV, College of Information Science and Engineering, Ritsumeikan University

Recently real-time affect-awareness is being applied in several commercial systems, such as dialogue systems and computer games. Real-time recognition of affective states, however, requires the application of costly feature extraction methods and/or labor-intensive annotation of large datasets, especially in the case of Asian languages where large annotated datasets are seldom available. To improve recognition accuracy we propose the use of cognitive context in the form of 'emotion-sensitive' intentions. Intentions are often represented through dialogue acts and, as an emotion-sensitive model of dialogue acts, a tagset of interpersonal relations-directing *interpersonal acts* (the IA model) is proposed. The model's adequacy is assessed using a sentiment classification task in comparison with two well-known dialogue act models, the SWBD-DAMSL and the DIT++. For the assessment, five Japanese in-game dialogues were annotated with labels of sentiments and the tags of all three dialogue act models which were used to enhance a baseline sentiment classifier system. The adequacy of the IA tagset is demonstrated by a 9% improvement to the baseline sentiment classifier's recognition accuracy, outperforming the other two models by more than 5%.

CCS Concepts:

• **Information systems → Information retrieval → Retrieval tasks and goals** → Sentiment analysis; • **Computing methodologies → Artificial intelligence→ Natural language processing** → Information extraction; Discourse, dialogue and pragmatics; Language resources; ; • **Computing methodologies → Machine learning →Learning paradigms**; Supervised learning; *Supervised learning by classification*; **→Learning paradigms**; Machine learning approaches; *Neural networks*; **→ Learning paradigms**; Machine learning algorithms; *Ensemble methods*

KEYWORDS

Affect-awareness, Sentiment recognition, Dialogue acts, Gaming data, Japanese language

ACM Reference format:

Peter Lajos Ihasz, Mate Kovacs, Ian Piumarta and Victor V. Kryssanov. 2018. A SUPPLEMENTARY FEATURE SET FOR SENTIMENT ANALYSIS IN JAPANESE DIALOGUES. *:ACM Trans. Asian Low-Resour. Lang. Inf. Process.* :XXXX, :XXXX. :XXXX (:XXXX :2018), 21 pages.
DOI: :XXXX

1 INTRODUCTION

Recently there is a growing need for real-time recognition of the affective state of emotions or sentiments in the field of affective computing. In several applications, delayed inference of affective states, which would allow for feature engineering and extraction, is not available and real-time recognition is needed. Dialogue systems and affect-aware games [Szwoch and Szwoch 2014], where the system tries to adapt its content according to the perceived emotions/sentiments of the human interlocutor, are typical examples of such applications.

Real-time emotion/sentiment recognition has been realized mostly in non-commercial, academic projects through the recognition of physiological features [Yoon et al. 2013] or facial expressions [Obaid et al. 2008]. Although showing very promising results in a laboratory environment, these methods rely on carefully positioned, costly sensors, and cannot yet be applied efficiently in commercial applications. Dialogue

[1] Address of department: 1 Chome-1-1 Nojihigashi, Kusatsu, Shiga Prefecture 525-8577, Japan
E-mail address of the corresponding author: ihasz17@gmail.com

systems, for example, are often used as telephone-customer-service agents, thus they can rely only on audio features. Even in the case of computer games, where facial recognition is often feasible, a shadow on the user's face, an un-leveled camera or the presence of facial hair, could lead to incorrect classification [Duncan et al. 2016]. The use of headphones (that is common in multiplayer gaming sessions) poses an even bigger challenge due to the 'noisy' representation of the player's face. Extracting information from the textual and/or audio content of the users' utterances would provide a less technology-sensitive, and thus less easily-perturbed set of features that could serve as a reliable and inexpensive means for emotion recognition, suitable for use in commercial software development.

One example of text processing in commercial software is the game Facade [Mateas and Stern 2005] that uses a rule-based approach for real-time emotion recognition. Systems developed lately, however, such as the commercial tool of EmoVoice [Vogt et al. 2008], or the system proposed by Fayek et al. [2015], achieve significantly better real-time emotion recognition with machine learning-based audio data classifiers. Nevertheless, supervised learning requires classifiers to be pre-trained on labeled data before they can be deployed for real-time recognition. Owing to the diversity of vocabulary and audio features, emotion/sentiment recognition in spontaneous dialogues is a complex task that demands a large amount of labeled data to ensure satisfactory recognition accuracy [Tian et al. 2015].

The advancement of emotion/sentiment recognition is therefore necessary to allow for reasonably accurate classification results while working with small, and therefore relatively easily-prepared, labeled datasets. The latter is especially important in the case of Asian languages, for which there is a lack of large datasets labeled with emotion-related psychological constructs.

A possible approach to advancing dialogue-based recognition of affective states is to consider cognitive context (rather than physiological context) in the form of intentions. The 'intentional' context, hand-labeled or inferred from textual and prosodic cues, is conventionally represented as dialogue acts—pragmatic-level linguistic units. The use of dialogue acts for emotion/sentiment recognition was considered in several studies [Ang et al. 2002; Lee and Narayanan 2005; Bartliner et al. 2003; Liscombe et al. 2005]. The acts discussed, however, mostly relate to 'communication maintenance' or 'domain related' intentions, which do not correlate well with emotions. Consequently, the best improvement achieved through the application of dialogue acts in these studies was 4% [Ang et al. 2002] in a binary classification of emotion-types.

The presented study examines the use of 'emotion-sensitive' dialogue acts (instead of 'conventional' dialogue acts). Emotion-interdependent intentions, inferable form audio and textual data, would improve emotion/sentiment recognition even on small sets of labeled data and could be applied to the pre-training of commercial games, dialogue systems and other applications requiring real-time recognition of affective states. For 'emotion-sensitive' dialogue acts the authors propose a taxonomy of *interpersonal acts*, originally developed as a computational model of conversational Japanese [Ihasz et al. 2015]. Interpersonal acts represent 'interpersonal relations-controlling' intentions, modeling the self-esteem and identity-directing aspects of interactions. The authors chose to validate the applicability of the proposed model in the cooperative dialogic environment of in-game conversations, representative of software demanding real-time emotion recognition. Accordingly, the proposed model has been developed considering the dialogue-specific aspects of a cooperative gaming environment.

The IA model was evaluated by measuring its adequacy for augmenting a baseline sentiment classifier in comparison with two commonly used 'conventional' dialogue act taxonomies, the DIT++ [Bunt 2009] and the SWBD-DAMSL [Jurafsky et al. 1997] models. For the evaluation, a corpus was selected consisting of dialogues conducted in a cooperative gaming environment. The dialogues are in Japanese, for which no large datasets labeled with interpersonal relation-indicating tags exist.

The remainder of this paper is organized as follows. Section 2 provides the conceptual basis and technical background of using dialogue acts to improve sentiment/emotion recognition. Section 3 describes the IA model and the design of the proposed computational method for the assessment of the model's adequacy. Section 4 presents the data and annotation method used in the study. Section 5 discusses the sentiment classification experiments, conducted separately, using the three dialogue act models. Section 6 discusses the main results of the experiments. Section 7 summarizes contributions of the study and outlines future work.

2 BACKGROUND

2.1 Dialogue Acts

Emotions influence intention, and vice versa [Frijda 1987; Plutchik 2001]. Accordingly, the expressed intention, represented in the content of the given utterance, can serve both as a cue for the underlying emotion/sentiment of the speaker and as a cue for the emotion/sentiment it will elicit in the addressee. Content can be subdivided into semantic and functional content. Semantic content includes the objects, propositions, and events defined in an utterance. Functional content specifies the communicative function of the utterance, "the way an addressee should use the utterance's semantic content to update his information state" [Bunt 2011]. In other words, it specifies the intentions of the speakers behind their utterances. Artificial inference of the affective states from the semantic content, however, is impractical due to the variability in the vocabulary and the multiple, context-dependent meanings of the words. (Even human inference usually necessitates multiple channels of communication for correct interpretation.) On the other hand, automatic inference of emotional states from functional content appears more practical because the possible range of content is narrower and easier to group into specific 'dimensions'.

Communicative functions (functional contents) are usually represented as *dialogue acts*—intention-conveying, pragmatic-level dialogic units. Dialogue acts can be categorized in many ways, with a particular categorization covering either one communicative function dimension (with mutually-exclusive tags for each annotated segment) or several dimensions (with multiple tags for each annotated segment) [Popescu-Belis 2008].

2.2 Dialogue acts for the recognition of affective states

The use of dialogue acts for emotion recognition purposes was considered in several studies. [Ang et al. 2002] augmented lexical and prosodic features with dialogue acts (*repeat, repair, neither*) of the current turn to improve emotion recognition. The addition of the dialogue acts resulted in a 4% maximum improvement when classifying the emotional states of *annoyance-frustration* vs. *else* (the latter includes all the remaining emotion types), and *frustration* vs. *else*. In the study of [Lee and Narayanan 2005], the emotional salience word score (representing the context-wise appearance likelihood) and dialogue acts (*rejection, repeat, rephrase, ask-start over, other*) were used together with prosodic and lexical features, yielding a 3% improvement in the 'binary' classification of *negative* and *non-negative* sentiments. Likewise, [Bartliner et al. 2003] augmented lexical and prosodic data with discourse information of dialogue acts (*introduce, request, suggest*), obtaining a 1.2% improvement when differentiating between the cognitive states of *emotional* and *neutral*. [Liscombe et al. 2005] considered prosodic, lexical, and dialogue act features, in addition to contextual data (65 categories discriminated by call-types of the HMIHY 0300 corpus, e.g., *asking for customer representative, requesting information about account balance,* etc.). The application of the dialogue acts led to a 2.6% improvement in the classification of *non-negative* vs. *negative* sentiments. These and many other studies with similar goals and results used dialogic data with pre-annotated dialogue act labels to evaluate the applicability of the supplementary feature set.

Recently there has been a declining interest in the idea of enhancing emotion/sentiment recognition through the use of dialogue acts. This may be due, at least in part, to the fact that accurate extraction of dialogue act features also requires pre-training on large collections of labeled data. Annotation of dialogue acts then becomes excessively labor-intensive, especially when measured against the rather modest improvements it would result in emotion recognition.

3 PROPOSED SOLUTION AND METHOD

3.1 Modelling Dialogue Acts

3.1.1 Emotion-sensitive dialogue acts

Certain dimensions of communicative functions, represented through dialogue acts, show stronger correlation with emotions than others. One example is the 'turn management' dimension of DIT++ [Bunt 2009] consisting of acts such as 'turn accept', 'turn take', and so on. Each of these acts can affect (or be affected by) any possible emotion/sentiment.

It is assumed by the authors that 'interpersonal relations-directing intentions' such as 'criticizing' or 'indiscrete commenting' would provide an 'emotion-sensitive' communicative function dimension, likely to affect or be affected by only a limited range of emotions/sentiment. An act of 'criticizing', for example, often affects the addressee's self-esteem negatively and would probably elicit an emotional reaction of negative valence, such as 'fear' or 'anger'. Furthermore, 'criticizing' is likely to be expressed under the influence of a negative valence emotion such as 'anger' or 'disgust'. It then appears natural to expect that the emotion 'anger' should be associated with a dialogue act representing 'interpersonal relation control', such as 'criticizing' more consistently than with a dialogue act representing 'turn management' such as 'turn take', for example.

3.1.2 The proposed model

The authors developed a dialogue act model representing 'interpersonal relations managing' communicative functions called the Interpersonal Acts (IA) model (see Table 1). It is a modified version of a dialogue act model for Japanese conversation proposed in a previous study [Ihasz et al. 2015] and is based on the Politeness theory of Brown and Levinson [Brown and Levinson 1987] and its Japanese critics [Matsumoto 1988]. 'Interpersonal relations managing' intentions are complex, sometimes ambiguous, constructs. Japanese social practice, however, is more rigid than in most cultures and Japanese interactions are therefore assumed to be easier to annotate with such constructs. Accordingly, the proposed model is tailored to be applicable to Japanese conversations, containing acts such as "accepting as superior", accounting for the Japanese culture where deference is often displayed not only through conjugational forms but also through the use of specific phrases. Although the proposed model is culture-specific, it can also be applied to other languages after alteration considering the social practice of the target language. To the knowledge of the authors, there are no other culture-specific dialogue act models that incorporate interpersonal relations managing communicative functions. Some universal (not explicitly specified for any language) models such as the SWBD-DAMSL and DIT++ contain certain acts or dimensions of acts that deal with interpersonal relations, but only in a non-comprehensive manner. The model has also been tailored to fit verbal interactions in a co-operative environment (such as gaming) with the addition of 'partner-unrelated commenting' tags.

The IA categorization is intended to be used as a one-dimensional, independent model when needed, as well as an extension of other multi-dimensional models. As a one-dimensional tagset, it can be used to define labels for the recognition of 'interpersonal relations managing' intentions or as a supplementary feature set for emotion/sentiment recognition purposes. When incorporated into a multi-dimensional tagset, the IA model serves as a dimension of 'interpersonal relations managing' communicative functions. The IA model conforms to the ISO Standard for Dialogue Act Annotation 24617-2 [Bunt et al. 2012; Bunt et al. 2017] as detailed in Appendix II.

Interpersonal acts could be used to represent the previous turns' stimuli for emotions/sentiments, as well as results of the cognitive process influenced by the affective states of the current turn. Accordingly, the model is assumed to be applicable to improving the recognition of affective states of emotions/sentiments.

Table 1. Taxonomy of interpersonal acts

Categories, subcategories				Examples
Interpersonal acts	Non-face threatening acts	Partner-unrelated commenting	**P-u. positive commenting**	“よし見つけた” (“Finally! I found it!”)
			P-u. negative commenting	“やばい” (“That’s bad!”)
			P-u. neutral commenting	“どこだ？” (“Where is it?”)
		Trying to ground/maintain good relationship	**Paying attention**	“うん” (“Mhmm/Yep”)
			Empathizing	“マジで？” (“Seriously?”, in reaction to statement)
			Accepting as superior (showing deference)	“わかりました” (“Understood!”)
			Agreeing	”そうだよう一” (“Yes, it is!”)
			Self-image justification	”すぐ倒せるし” (“I can defeat them in an instant as well”)
	Face-threatening acts (FTA)	Positive FTA	**Criticizing**	“え、あれでいいの?｛笑｝” (“Are you sure you will be alright like that? [laughter]”)
			Inadequate commenting	“また死んだ？” (“You died again?”)
		Negative FTA	**Indebting partner**	“取っていってやる” (“I will take it for you.”)
			Commanding/ requesting	“じゃたまり場来て” (“Come to the gathering spot!”)

The categories are illustrated with actual examples selected from the data after annotation (see Section 4.1.2). For more detailed examples including conversational context, see Appendix I.

3.2 Validation

3.2.1 Validation method

The applicability of the proposed tagset will be tested in four sentiment analysis scenarios. For applications that use real-time recognition of affective states like affect-aware games or customer-service dialogue systems, sentiment recognition is preferred to the more fine-grained, but less reliable, emotion recognition. In the work of [Ang et al. 2002; Lee and Narayanan 2005; Bartliner et al. 2003; Liscombe et al. 2005] the proposed dialogue acts were tested in binary-classification scenarios.

In the validation experiments conducted in this study, the interpersonal act tags will be used to enhance a text- and audio-based sentiment classifier (detailed in Section 3.2.3) in comparison with

- two identical classifiers which use acts of other well-known dialogue act models, and
- a baseline classifier which does not use dialogue acts as a supplementary feature set for sentiment classification.

The four classifiers (one baseline classifier and three augmented classifiers, processing dialogue acts) were trained and tested on the same audio streams of Japanese dialogues and their transcriptions of gaming sessions to analyze interpersonal acts in an environment typical for real-time sentiment recognition. Each

utterance in the transcriptions was therefore annotated with four labels in total: one sentiment label and three dialogue act labels (one of the proposed tagset and two of the other dialogue act models used for comparison). Figure 1 shows the overall design of the four classification scenarios. Interpersonal acts are proposed as acts to be automatically inferred or to be hand-labeled before their application to emotion/sentiment recognition. Following the line of previous studies [Ang et al. 2002; Lee and Narayanan 2005; Bartliner et al. 2003; Liscombe et al. 2005] this study worked with hand-labeled dialogue act tags, concentrating only on measuring the proposed model's adequacy for augmenting sentiment recognition. The comparative performance augmentation of the dialogue act models could therefore be fully assessed, unhindered by their varying recognizability by automatic means.

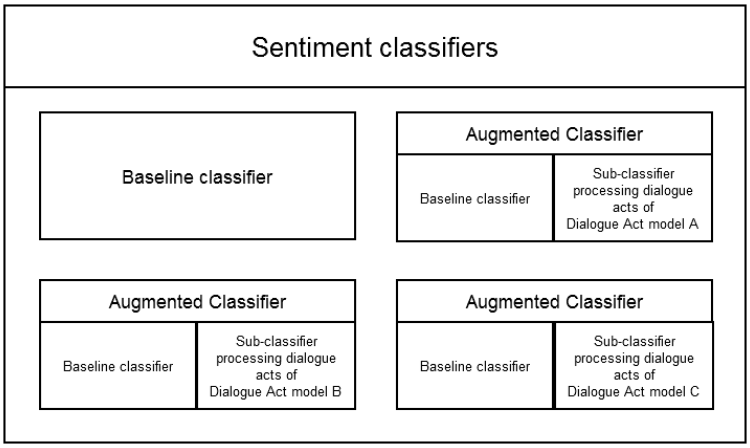

Figure 1. Sentiment classification scenarios

3.2.2 Dialogue Act Taxonomies Used for Comparison

Two other dialogue act taxonomies were used in this study. For the purpose of mutual comparison their tagsets, and the tagset of the IA model were used to augment the automatic recognition of sentiments with the intentional content of players' utterances. SWBD-DAMSL and DIT++ were selected because they are widely known and used, and represent communicative functions using one-dimensional and multi-dimensional approaches, respectively.

The SWBD-DAMSL tagset defines dialogue acts for 42 one-dimensional (mutually-exclusive) intentions. Although it contains a few dialogue acts for social obligations, it does not include acts accounting for social status or intentions that would influence self-esteem, and a wide range of subsequent emotions.

The multi-dimensional DIT++ consists of one set of 'general-purpose communicative functions' (intentions) and nine tagsets constituting 'dimension-specific communicative functions' such as 'auto-feedback', 'allo-feedback', and so on. (For the dimension of 'task/activity' no tagset of communicative functions is defined.) Well-formed tags on functional segments are pairings of <D,F> where D is one of the ten dimensions and F is a communicative function of the corresponding dimension (e.g., <autoFeedback, request>). DIT++ assumes that every functional segment of the dialogue is initially annotated with one tag from the dimension of 'general purpose communicative functions'. In addition, every functional segment can be optionally tagged with up to nine tags, one for each 'dimension-specific communicative function' dimension. As each dimension contains mutually-exclusive tags, one segment can be annotated with between one and ten tags in total. DIT++ has a function-specific dimension of 'social obligations', containing communicative functions such as 'greeting', but its acts do not account for non-obligatory interpersonal relation management. This study considered 22 'general purpose' acts, including all specifications described by Bunt [2009], and ten acts from the dimension of 'social obligations'. However, since the acts from the dimension of 'social obligations' (e.g., <social obligations, greeting>) defined in DIT++ can only co-occur

with one 'general purpose' act, a linearized tagset would contain 32 different acts in total, considering all possible combinations among the two dimensions.

3.2.3 Sentiment classification

Four classification scenarios were investigated to verify the applicability of the interpersonal acts for the improvement of sentiment analysis. In one scenario the sentiments are classified by a baseline classifier, while in the other three scenarios by an augmented classifier each processing a different set of dialogue act labels as an additional feature set (see Figure 1).

The baseline classifier processed only the audio streams of the dialogues and their textual transcriptions. Figure 2 depicts the architecture of the classifier, consisting of two sub-classifiers. Sub-classifier #1, a Gated Recurrent Unit (GRU) Neural Network, processed the word embeddings of textual transcriptions of the audio streams; it can recall its previous internal states to process sequences of inputs and find possible dependencies within long sequences of embedded utterances (functional segments) [Chung et al 2104]. The word embedding was created with the GloVe embedding algorithm [Pennington et al. 2104] which was trained on the Wikipedia dump data [Wikimedia Project Editors 2016].

Sub-classifier #2 processes the audio data. The audio files, containing each of the five dialogues, were partitioned into functional segments. Each functional segment was then saved as a 3-second audio file, lengthening the original segments with silence or shortening them if they were longer than 3 seconds. (Most segments were originally between 1 and 3 seconds long. Lengthening the segments to more than 3 seconds was assumed to introduce too much noise during vectorization for the training of the classifiers.) The segments were transformed with the OpenEar software [Eyben et al. 2009] into a set of low- level audio

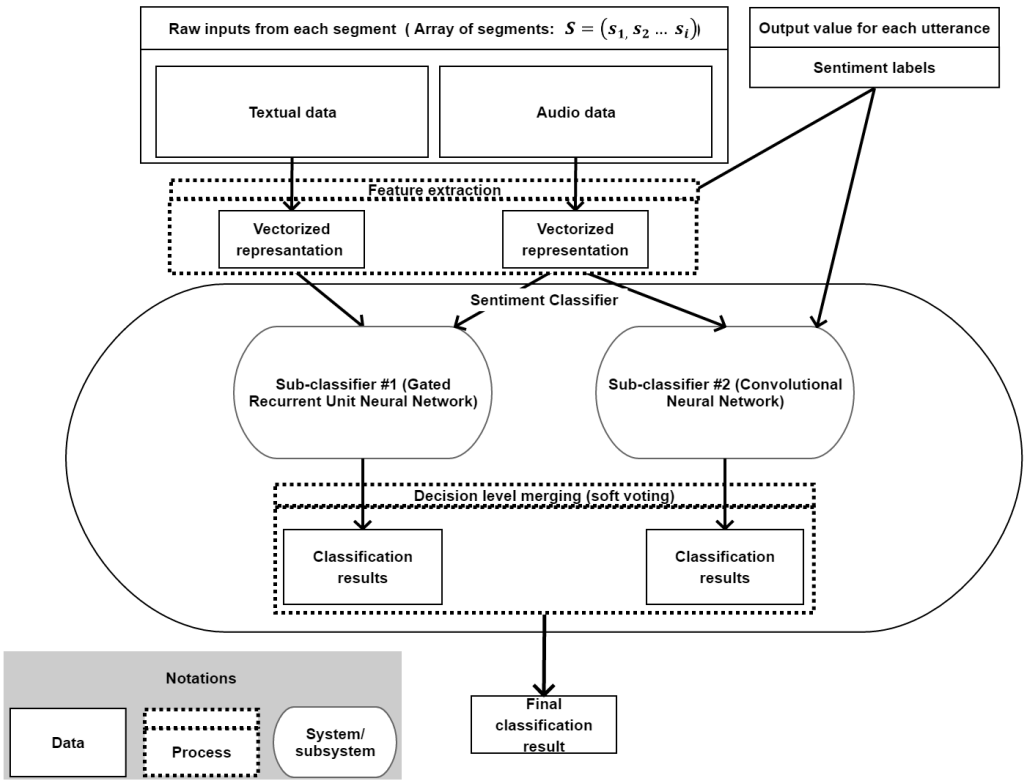

Figure 2. Baseline sentiment classifier

spectral bands. The audio feature vectors were processed by a one-dimensional Convolutional Neural Network (CNN) which can effectively extract the important vectors among a large number of others through its several convolutional and pooling layers [Abdel-Hamid et al. 2013]. (For further information about the number and order of layers of the sub-classifiers, see Appendix III.) Both sub-classifiers are trained and tested on the same set of sentiment labels. The audio and textual features extracted from the functional segments were processed in the order they occurred in the conversation, to help the GRU of sub-classifier #1 find meaningful dependencies between them. Using the ensemble learning method of soft voting [Opitz and Maclin 1999], the results of the two independently-trained sub-classifiers were merged at the decision level. Specifically, three fully-connected feed-forward network layers were trained on the classification results to acquire weights for them. The average of the sums of the weighted results was then computed.

In the other three scenarios, the baseline classifier was augmented with a third sub-classifier, another GRU, processing dialogue act labels as textual data (to find the possible dependencies in the sequence of labels). In each scenario the sub-classifier processed dialogue act labels from one of the dialogue act models IA, DIT++, or SWBD DAMSL. Similarly to the baseline method, the classification results of each sub-classifier were weighted and merged through soft-voting. Through weighting the results, the system could learn which sub-classifier contributed the most to the correct classification result. The dialogue act classifying sub-classifier #3, for example, was assumed to contribute less than sub-classifier #1 or #2. Hard voting would not allow for such learning; it would simply output the result produced by at least two sub-classifiers, or choose among the results arbitrarily if all outputs of the sub-classifiers were different. The authors elected to use decision-level merging instead of feature-level merging (where the audio feature vectors, word embeddings, and digitalized dialogue act labels would be merged into one tensor before being fed into the network) as preliminary experiments showed that decision- level merging allows for better classification results (the latter was also observed by Planet and Iriondo [2012]).

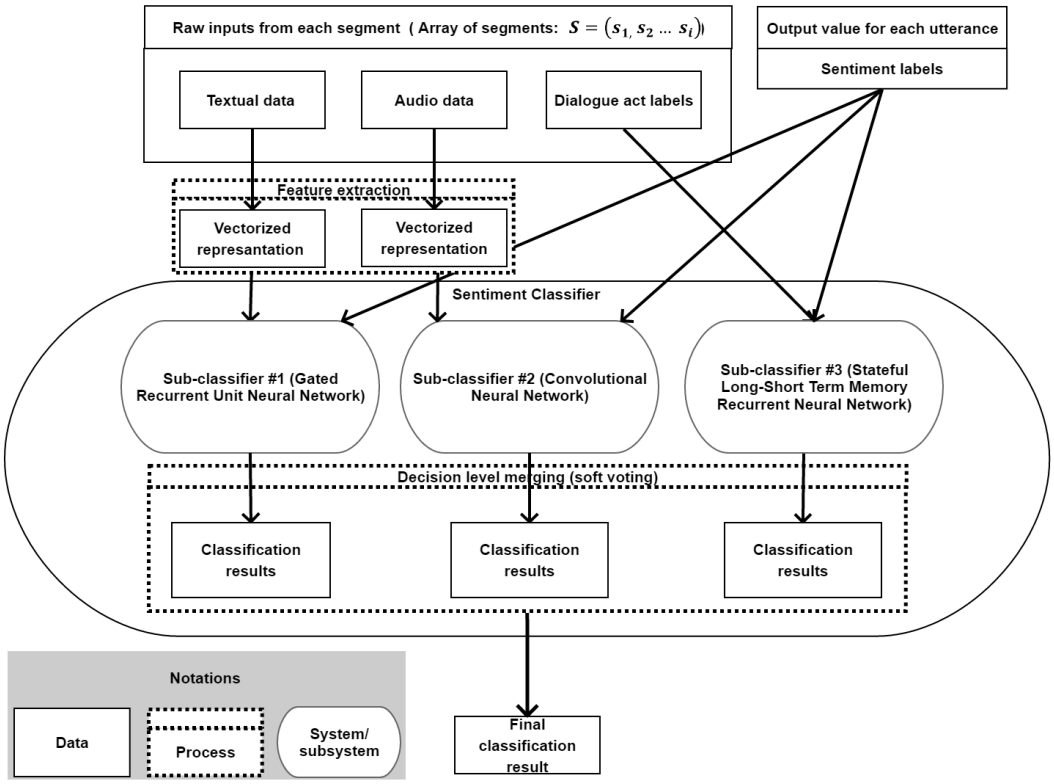

Figure 3. Sentiment classifier augmented with a dialogue act label processing sub-classifier

The target labels for the training and testing of sub-classifier #3 were also sentiment labels. Figure 3 depicts the architecture of the augmented baseline method used in the other three scenarios.

4 DATA

4.1 Annotated Corpus

Five natural language dialogues from the Online Gaming Voice Chat Corpus with Emotional Labels (OGVC) [Arimoto and Mori 2017] were selected for the experiments. The conversations were performed in Japanese during massively-multiplayer online role-playing game (MMORPG) sessions. The specific games involved were *Ragnarok Online*, *Monster Hunter Frontier*, and *Red Stone*. The in-game context of these dialogues demands co-operation, is rich in stimuli, and potentially provides for a wide range of emotions. The dialogs were performed in Japanese, for which no large datasets labeled with interpersonal relation-indicating (or other emotion-related) tags exist.

The five conversations consist of a total of 6,902 spontaneous utterances. Three dialogues were performed by three pairs of male players (4,397 utterances), and two dialogues by two pairs of female players (2,505 utterances). The corpus contains both transcriptions and audio recordings of each conversation. The conversations were segmented into individual utterances, with each interlocutor identified.

To provide a proper context for the annotators, the textual and audio data was reassembled into a dialogic form based on the timestamps in the sound files. (These files do not include network delay data, and for that reason the exact timing of the conversations could not be precisely recovered). The conversations were re-segmented into functional segments (not necessarily corresponding to single utterances) by the authors and by a native Japanese speaker, making a total of 6,934 functional segments.

For the purpose of evaluation, each segment had to be annotated with tags from DIT++, SWBD-DAMSL, and IA, and also with the emotions experienced by the interlocutors (later converted to sentiments). Annotators were employed to assign tags from the three models, and to complete the emotion tagging. (The corpus creators annotated only 80% of the original utterances with basic emotion tags.)

The sizes of the functional segments differ in each model. For example, the functional segment for the act <general-purpose, answer> from DIT++ was often expressed in a single utterance segment, while ‘empathizing’ from the IA model tended to be expressed through two or three utterances. The smallest possible segments were therefore chosen during functional segmentation, considering all the three models. In the cases where a functional segment of a given model (typically IA) covered several smaller segments, all those segments were annotated with the same tag. On average, a dialogue contained 1,386 segments corresponding to approximately 62 minutes of audio.

4.1.1 Emotion Annotation

Eight of the ten emotion tags employed by the compilers of the corpus are identical to the basic emotions defined by Plutchik [2001]: joy, sadness, anger, fear, acceptance, disgust, surprise, and anticipation. The remaining two tags, ‘neutral’ and ‘other’, are complementary; their purpose is to account for emotions that cannot be classified into any of the original eight categories. In the case of segments that had already been tagged with emotions by the corpus compilers, only tags assigned by at least two of the compilers were retained. When all three compilers assigned different tags, one tag was selected (based on the judgment of the authors) and retained. Since the goal of the study is to help improve real-time emotion/sentiment recognition, fine-grained emotion recognition was deemed unnecessary or even inapplicable. For that reason and similarly to previous work (see Section 1), the emotion labels were collapsed to negative (consisting of anger, fear, sadness, and disgust), positive (surprise, joy, acceptance, and anticipation), and neutral (consisting of neutral and other) sentiment labels. These labels correspond to the valance-categories of negative (angry, afraid, sad and annoyed), positive (astonished, happy, pleased/satisfied and excited) and neutral (neutral) from Russel’s circumplex of emotions [1980], with the addition of the emotion other to the neutral category. No original tags were provided for 1,355 of the functional segments. Tags were added to them by three native Japanese speakers, employed for this study, using the three sentiment labels. The

annotators were three male university students, between the ages of 20 and 23, each having over 100 hours of online-gaming experience. Transcripts and audio recordings of the dialogues were provided to the annotators who were asked to determine the underlying sentiment of the interlocutor for each segment. All segments therefore received one sentiment tag. Before the actual tagging procedure, each annotator participated in a brief training session, where 150 consecutive example segments (from the same corpus but not used in the study) were shown with suggested sentiment labels. The segments were annotated by the authors (having advanced-level Japanese language proficiency) to show how the labels should be attached in common and uncommon cases (e.g. a certain sentiment being expressed through not one but several segments). The inter-annotator agreement for sentiment tags assessed with Fleiss' Kappa was 78.6%.

4.1.2 Dialogue Act Annotation

The corpus was then annotated with the dialogue act tags of SWBD-DAMSL and DIT++, and with the interpersonal act tags of the IA model. Three native Japanese speakers (different from those who annotated the emotions) were employed. Each added tags from one of the three tagsets to the transcriptions of the five dialogues while listening to the corresponding audio recordings. Since three models were used, the annotation was conducted in three iterations, each iteration for a different dialogue act model. The annotators were university students, two male and one female, between the ages of 21 and 25, with more than 80 hours of online gaming experience each. They were instructed to determine the interlocutor's intention for each segment and to label it with the most appropriate dialogue act tag from each dialogue act model. The annotators first participated in a training session, similar to the one conducted before the emotion labeling. This training session involved the same 150 consecutive functional segments, repeated three times. Each time the 150 segments were labeled with one of the dialogue act models used in the study, showing how the dialogue acts of the given model could be expressed through one or several segments. The annotators were cautioned that certain acts of certain models (typically the acts of the IA model) tend to be expressed through several functional segments. The inter-annotator agreement (estimated again with Fleiss' Kappa) was 69.1% for the DAMSL SWBD tagset, 71.7% for DIT++, and 66.2% for the IA model. The IA model had the lowest score because the functional content dimension it covers permits more subjectivity than do the other models.

Similarly to the case of emotion tags, a single dialogue act tag from each taxonomy was assigned to each segment (hence three tags per segment). Any tag selected by at least two annotators was retained for the analysis; otherwise, one of the three different tags assigned by the annotators was retained (based on the judgment of the authors).

A preliminary analysis revealed that the DIT++ tagset is over-specified for the given experimental data. To compensate for the dataset's limited size, several optional class specifications were omitted. This was deemed reasonable as it improved the performance of DIT++ in the experiments. The number of acts in each taxonomy considered during analysis was further decreased by disregarding those not assigned to any functional segment. For the analysis, 28 acts of SWBD-DAMSL (3rd party, Acknowledge, Affirmative non-yes, Agree, Apology, Appreciation, Backchannel question, Conventional close, Declarative question, Directive, Hedging, Maybe, Negative answer, No-answer, Non-verbal, Non-understandable signal, Offer, Open-question, Or-clause, Other answers, Response acknowledgement, Statement, Statement-opinion, Summarize, reformulate, Tag-question, Thanking, Wh-question, Yes-no-question)
and 17 acts of DIT++ (Address request, Address suggestion, Agreement, Answer, Apology, Check-question, Confirm, Disagreement, Disconfirm, Inform, Instruct, Offer, Propositional question, Request, Set question, Suggestion, Thanking) were retained.
All 12 acts of the IA model occurred in at least one of the five dialogues.

5 EXPERIMENTS

Three experimental setups for sentiment recognition were developed, based on the architecture detailed in Section 3.2.2, and used separately for all the three augmented classifiers processing a given dialogue act model. In each setup the input labels of sub-classifier #3 were different, to account for the differences in the number of functional segments, in which the dialogue acts of each given model tend to be expressed. The authors tried to find the optimal setups in the case of each dialogue act model, resulting in the best sentiment classification accuracy of each sentiment classifiers' sub-classifier #3. Similarly to the previous work, the

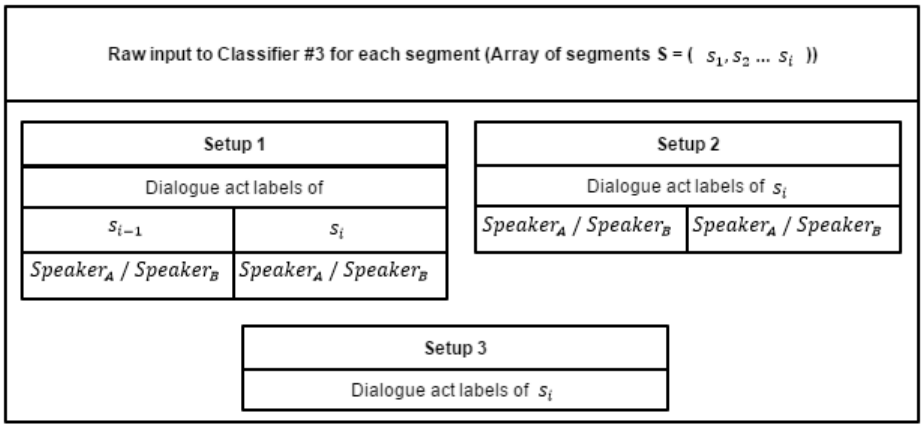

Figure 4. Dialogue act input labels for Sub-Classifier #3 in the various experimental setups

dialogue acts were not parsed by a sub-system but used 'as-is', to reveal the maximal extent of their applicability for sentiment recognition (see Section 3.2.1). Figure 4 specifies the inputs for the various setups.

Setup 1: The dialogue act of the preceding functional segment (s_{i-1}), performed by Speaker-A can represent the intention-level stimuli for the sentiment of the i^{th} segment (s_i) performed by Speaker-B (especially an interpersonal act). Subsequently, the dialogue act in Speaker-B's s_i is assumed to represent the outcome of a cognitive process influenced by their sentiment (see Section 2.1). To account for the causal connection between the consecutive utterances in Setup 1, each batch processed by sub-classifier #3 consists of the dialogue acts of s_{i-1} and s_i, labeled with the sentiment of s_i.

Since functional segment lengths are not consistent between the three different dialogue act taxonomies, intentions are sometimes expressed through several consecutive functional segments. In such cases, s_{i-1} and s_i are performed by the same speaker. The interpersonal act of s_{i-1} and s_i then represents an ongoing mental state (intentional context). This can still serve as a cue for the emotional state of the same speaker, expressed in the current segment s_i. To help sub-classifier #3 differentiate between these scenarios, each dialogue act label in Setup 1 indicates the performer as Speaker-A or Speaker-B. The number of possible dialogue acts is therefore doubled for each taxonomy. (In the case of the IA tagset, for example, 'criticizing' is subdivided into 'criticizing_A' and 'criticizing_B'.)

A shift between topics may negate causative or continuation relationships between consecutive utterances. However, because accounting for topic shifts would further increase the number of training labels, the authors elected not to consider them in the context of such a small dataset.

Setup 2: Dialogue acts of the SWDB-DAMSL and DIT++ models are presumed to have weaker sensitivity to emotions/sentiments, and may not serve as a stimulus for them. An experiment was therefore conducted with batches containing only the dialogue act labels and sentiment labels of s_i, where the considered two models may perform better. Speaker-A and Speaker-B were, however, still differentiated.

Setup 3: It is possible that, by increasing the number of dialogue acts through differentiating between dialogue acts performed by Speaker-A and Speaker-B, we add too much noise to the data. Setup 3 is intended to minimize the noise, where dialogue act tags were not differentiated by speaker and each batch contained only the dialogue act and sentiment label of s_i.

When comparing the final classification results of the four sentiment classifiers (the baseline classifier and the three augmented classifiers utilizing the different dialogue act models), each augmented classifier uses its own sub-classifier#3 with a setup optimized for the attributes of the dialogue act categorization

deployed. All sentiment classifiers used in the experiments were trained and tested on the same sets of functional segments, through 10-fold cross-validation. Specifically, the aggregated sets of the five dialogues were randomly partitioned into 10 equal-sized subsamples, from which a single subsample was retained for testing the model with the remaining 9 subsamples used as training data. (The order of the functional segments within each subsample and the order of the subsamples themselves were preserved to ensure the neural networks can learn from the structure of the conversations.) Thus the test set always consists of a subset of utterances of one or at most two dialogues, conducted by one or two pairs of speakers, while the training set consists of a subset of utterances from three or four dialogues conducted by three or four pairs of speakers. Accordingly, the test set is speaker-independent. To reduce variability, the testing was performed over 10 iterations, each time using a subsample as test set that had not been used previously, and producing 10 recognition accuracy results. The average of the 10 recognition results gives a less biased estimate of the overall recognition accuracy.

6 RESULTS AND DISCUSSION

6.1 Recognition rate of the sub-classifiers

Table 2 lists the experimental results obtained for sentiment recognition through the three sub-classifiers. In the case of sub-classifier #3, all cases of processing the three dialogue act models are analyzed separately according to the three proposed experimental setups.

The overall low recall on the negative sentiment indicates bias in the distribution of sentiments within the dataset. This may be due to the fact that online gameplay requires cooperation that would easily be ruined if

Table 2. Recognition accuracy of the separate sub-classifiers obtained in the experiments

Method	Precision			Recall			F1-score			Overall Acc.
	NEG	NEU	POS	NEG	NEU	POS	NEG	NEU	POS	
Sub-classifier #1	0.38	0.45	0.69	0.27	0.63	0.60	0.32	0.42	0.64	56.18%
Sub-classifier #2	**0.41**	**0.34**	**0.64**	**0.29**	**0.52**	**0.56**	**0.34**	**0.41**	**0.60**	**51.91%**
Sub-classifier #3 processing labels from the DIT ++, Setup1	0.14	0.17	0.23	0.07	0.20	0.24	0.09	0.18	0.23	19.60%
Sub-classifier #3 processing labels from the DIT ++, Setup2	**0.18**	**0.21**	**0.25**	**0.11**	**0.25**	**0.29**	**0.14**	**0.23**	**0.27**	**23.72%**
Sub-classifier #3 processing labels from the DIT ++, Setup3	0.13	0.16	0.21	0.06	0.20	0.23	0.08	0.18	0.22	18.84%
Sub-classifier #3 processing labels from the SWBD-DAMSL, Setup1	0.15	0.21	0.23	0.07	0.24	0.25	0.10	0.22	0.24	20.42%
Sub-classifier #3 processing labels from the SWBD-DAMSL, Setup2	**0.15**	**0.18**	**0.24**	**0.09**	**0.24**	**0.26**	**0.11**	**0.21**	**0.25**	**21.06%**
Sub-classifier #3 processing labels from the SWBD-DAMSL, Setup3	0.14	0.18	0.23	0.07	0.25	0.27	0.09	0.21	0.25	20.84%
Sub-classifier #3 processing labels from the IA model, Setup1	**0.25**	**0.24**	**0.36**	**0.13**	**0.30**	**0.33**	**0.17**	**0.27**	**0.34**	**30.93%**
Sub-classifier #3 processing labels from the IA model, Setup2	0.24	0.24	0.32	0.12	0.29	0.33	0.16	0.26	0.32	28.89%
Sub-classifier #3 processing labels from the IA model, Setup3	0.20	0.21	0.29	0.11	0.26	0.32	0.14	0.25	0.30	26.06%

negative sentiment were to be expressed excessively. Audio vectors seem to be slightly better indicators of negative sentiment while word embeddings are of neutral and positive sentiments.

Sub-classifier #1, processing the textual data, showed a 56.18% recognition accuracy while sub-classifier #2, processing the audio data, achieved 51.91%. This implies that the word vectors trained on GloVe served as a more consistent cue for sentiment recognition than the low-level audio feature vectors. Presumably, the convolutional neural network of sub-classifier #2 was not able to generalize well enough on such a small dataset.

As expected, sub-classifier #3, processing only the one-dimensional textual data of dialogue acts, performed significantly worse. The setups containing the best overall recognition accuracy for each dialogue act model (processed by sub-classifier #3) are highlighted in bold type. With the best-performing setups for processing the given dialogue act model, sub-classifier #3 achieved 21.06%, 23.72%, and 30.93% recognition accuracy when trained on the SWBD-DAMSL, DIT++ and IA tags, respectively. The use of the IA tagset (in the best-performing setup) yielded 9.87% and 7.21% better recognition accuracy compared to the best performance of the SWBD-DAMSL and DIT++ tagsets, respectively. Furthermore, in the case of the IA model, the best performance was achieved through Setup 1 (considering dialogue acts of preceding utterances and differentiating between speakers). In the cases of SWBD and DIT++, however, the best performances were achieved with Setup 2 (differentiating between speakers but not considering preceding dialogue acts), which implies that dialogue acts that are unrelated to affective states (and cannot serve as stimuli for them) are less suitable for harvesting contextual information during sentiment/emotion classification tasks.

In general, sub-classifier #3 shows similar performance when trained on the SWBD-DAMSL and DIT++ tags in terms of precision, recall and accuracy of the given sentiments. However, training the sub-classifier on the IA tagset resulted in a noticeably higher precision in negative sentiment. This fact suggests that the IA tagset, accounting for "face-threatening" interpersonal verbal-actions, has more acts consistently co-occurring with negative sentiments, thus, is more adequate to serve as a cue for them than the other two dialogue act models. All these results can be interpreted as strong evidence in favor of the definition and use of emotion-sensitive dialogue acts specifically for augmenting emotion/sentiment recognition systems.

Table 3. Recognition accuracy of the separate baseline and augmented-classifiers obtained in the experiments

Method	Precision			Recall			F1_scorre			Overall Acc.
	NEG	NEU	POS	NEG	NEU	POS	NEG	NEU	POS	
Baseline method (Sub-classifier #1 and #2)	0.42	0.50	0.77	0.30	0.70	0.67	0.35	0.58	0.72	62.33%
Baseline method + Sub-classifier #3 processing labels from the DIT ++ (with best performing setup)	0.45	0.58	0.76	0.23	0.70	0.81	0.30	0.63	0.78	66.10%
Baseline method + Sub-classifier #3 processing labels from SWBD-DAMSL (with best performing setup)	0.46	0.57	0.75	0.22	0.69	0.81	0.30	0.62	0.78	65.20%
Baseline method + Sub-classifier #3 processing labels from the IA model (with best performing setup)	0.58	0.60	0.79	0.30	0.74	0.84	0.40	0.66	0.80	71.42%

6.2 Recognition rate of the augmented classifiers

Table 3 shows the experimental results obtained for sentiment recognition of the baseline classifier, and the three augmented classifiers using different dialogue act models. The output of the baseline classifier, merged using a soft-voting process (Section 3.2.2), could predict the correct sentiments with an accuracy of 62.33%. This moderate accuracy reflects well the complexity of the task of recognizing sentiments when working with a small dataset.

Although dialogue act tags appear to be poor cues for sentiment classification, when used as single features, as a supplementary feature set (through decision-level merging) they improved the baseline model's recognition accuracy. (Appendix IV. details the averaged bias-weights, used in the soft-voting process for the various sub-classifiers (computed by a fully connected neural network.)) The use of SWBD-DAMSL improved overall recognition accuracy by a maximum of 2.87%, with DIT++ by 3.77%, and with IA by 9.09%. Only the setup in which sub-classifier #3 performed best (relative to a given dialogue act model) was selected for augmenting the baseline method.

6.3 Performance of the IA tagset as supplementary feature set

In previous studies the addition of dialogue act labels resulted in an improvement of 4% at most [Ang et al. 2002], using only two affective-types and with larger datasets. Thus, in the context of such a small dataset, these results are considered to be meaningful, demonstrating the usefulness of cognitive context (in the form of dialogue acts) for sentiment/emotion recognition.

A single factor Anova test, computed from the validation scores of each classifier's ten-fold cross-validation process, shows that the improvement achieved with the IA model is significant. Table 4 gives the results of the Anova tests computed through the results of each classifier's best-performing recognition setup, in comparison to the results of the best-performing setup of the classifier that applies to the IA tagset.

In the case of larger datasets, with more labeled utterances, it can be expected that the difference in performance of the various dialogue act models (for augmenting sentiment/emotion classification) would diminish. However, annotating large datasets with labels of emotion-related constructs is a highly labor-intensive task. Also, in the gaming domain, no such large datasets for the Japanese language currently exist.

To further scrutinize the applicability of the proposed tagset, an interpersonal act processing sub-classifier #3 (with the best performing Setup 1) was used to augment the text processing sub-classifier #1 and audio-processing sub-classifier #2 respectively, improving their recognition accuracy by 5.14% and 7.01 % (see Table 5.). Since the merging of the sub-classifiers was done by soft voting, (weighting and averaging mid-classification results), these results are not surprising.

Table 4. Significance of improvement yielded by the application of interpersonal acts

Classifier	p-value	F-score	F-critical
Baseline method vs. Augmented method processing IA - best performing setup (bfs.)	<0.005	151.80	4.41
Augmented method processing IA – bfs vs. Augmented method processing DIT++ - bfs.	<0.005	21.77	4.41
Augmented method processing IA – bfs vs. Augmented method processing SWBD-DAMSL – bfs.	<0.005	43.21	4.41

Table 5. Improving the separate sub-classifiers by the application of interpersonal acts

Method	Improvement	Overall Acc.
Sub-classifier #1 augmented with Sub-classifier #3 processing IA - best performing setup (bfs)	5.14%	61.32%
Sub-classifier #2 augmented with Sub-classifier #3 processing IA - best performing setup (bfs)	8.01 %	59.92%

In the case of merging the outputs of all three sub-classifiers, the weighting is more balanced, further improving the final classification result. Both sub-classifiers #1 and #2 perform better than #3, obtaining stronger weights in the soft voting process. Sub-classifier #1, however, processing the more reliable (at least in this dataset) word-embedding vectors, performs better than #2, thus obtaining even stronger weights during the soft- voting process, and not letting interpersonal acts heavily influence the final classification result. Sub-classifier #2, on the other hand, is a slightly weaker classifier, letting sub-classifier #3's results dominate. Thus the use of interpersonal acts improves sub-classifier #2 even more than it does sub-classifier #1, but results in a weaker overall classification accuracy.

6.4 Applicability of the IA tagset for automatic recognition

The experiments conducted utilized dialogic data with pre-annotated dialogue act labels to fully evaluate the applicability of the additional feature sets for sentiment recognition. To measure the applicability of the IA model more thoroughly, however, the amount of training data needed for a satisfactory level of automatic classification needs to be assessed. In particular, satisfactory-level recognition in this scenario would indicate a minimum level of classification accuracy that ensures that the automatically-annotated interpersonal act labels would improve sentiment recognition as a complementary feature set. Assessing the trade-off between annotation-cost and the improvement obtained is a complex task and is beyond the scope of this study.

Table 6. Results of interpersonal acts classification

IA tagset	**Precision**	**Recall**	**F1-score**
Inadequate commenting	0.11	0.10	0.10
Commanding/requesting	0.41	0.48	0.45
Criticizing	0.23	0.16	0.19
Indebting partner	0.50	0.11	0.18
Self-image justification	0.05	0.09	0.06
Partner-unrelated positive commenting	0.49	0.35	0.41
Partner-unrelated negative commenting	0.54	0.59	0.56
Partner-unrelated neutral commenting	0.49	0.66	0.57
Paying attention	0.62	0.63	0.63
Accepting as superior	0.45	0.16	0.23
Empathizing	0.76	0.63	0.69
Agreeing	0.32	0.29	0.30
Total	**0.53**	**0.55**	**0.54**
Overall accuracy		**55.09%**	

Further computational experiments are required to measure the learning rate while utilizing a complementary feature set of interpersonal act labels, as opposed to using only output (sentiment) labels on larger datasets.

Nevertheless, trained and tested on the same dataset with audio and textual input and interpersonal act label output, the baseline-classifier (consisting of the GRU-based sub-classifier #1 and CNN-based sub-classifier #2) showed promising results in their automatic classification. Table 6 summarizes the recognition results. An overall 55.09% recognition accuracy for 12 acts indicates that the acts of the IA model are easier to automatically classify from textual and audio features than sentiments (having classified with 62.33% overall recognition accuracy for three categories by the baseline-classifier).

7 CONCLUSIONS

This study presented the taxonomy of *interpersonal acts*, an emotion-sensitive dialogue act model aiming to advance text- and audio-based emotion recognition using a less annotation-intensive feature set. In an evaluation conducted for sentiment recognition, the IA model outperformed two well-known dialogue act taxonomies, the SWB-DAMSL and the DIT++ models. IA also appears to perform better than models used in previous studies known to the authors and surveyed in this research.

The IA model can therefore contribute to the advancement of emotion/sentiment recognition, as demonstrated by the present results. It could be used for the pre-training of commercial software demanding real-time emotion/sentiment recognition, especially in the context of the Japanese language for which large annotated datasets are rarely available. Since the IA model was developed in accordance with the ISO Standard 24617-2 for Dialogue Act Annotation, it can be used both as a stand-alone dialogue act model, or as one component in a multi-dimensional model.

In future work, the authors plan to

- further investigate the affective dynamics of human-human and human-computer dialogues, as revealed through topic-partitioned sequences of consecutive interpersonal acts.
- conduct further experiments to measure the amount of data needed for satisfactory-level recognition (yielding improvement in sentiment-classification) of the interpersonal acts
- extend/alter the IA model to be applicable to other languages

Appendix I. Interpersonal acts with contextual examples

Interpersonal act categories	Examples	Interpersonal act and sentiment labels retained for training/testing
P-u. positive commenting	B：”どこだ？” (“Where is it?”), A：”謎の骨だ” (“It’s a bone (game content) of mystery.”), **B：”よし見つけた” (“Finally! I found it!”),** B：”どちらから行こう？” (“From where should we approach?”) A：”早えよ。、｛笑｝” (“That was fast! {Laughter}”)	P-u. neutral commenting, ANGER Empathizing, JOY **P-u. positive commenting, JOY** P-u. neutral comm., NEUTRAL Empathizing ACCEPTANCE
P-u. negative commenting	**A：“やばい” (“That’s bad!”)** A：”間違えて剣出すの三角ボタンだとおもちゃっ、｛笑｝” (“I mistakenly thought that it’s the triangle button for equipping the sword...{Laughter}”), B：”｛笑｝” (“{Laughter}”),	**P-u. negative comm., DISGUST** Self-image justification, JOY Empathizing, ACCEPTANCE
P-u. neutral commenting	**B：”どこだ？” (“Where is it?”),** A：”謎の骨だ” (“It’s a bone (game content) of mystery.”), B：”よし見つけた” (“Finally! I found it!”), B：”どちらから行こう？” (“From where should we approach?”), A：”早えよ。、｛笑｝” (“That was fast! {Laughter}”)	**P-u. neutral comm., ANGER** Empathizing, JOY P-u. positive commenting, JOY P-u. neutral comm., NEUTRAL Empathizing, ACCEPTANCE
Paying attention	A：”ジュウバンか。” (“Is it the tenth?”), **B：“うん” (“Mhmm , I see.”),** B：“かな” (“Or is it?”), B：“ああジュウもいそうだね。” (“Oh yeah, it looks like there are ten of them”),	P-u. neutral comm., NEUTRAL **Paying attention, NEUTRAL** P-u. neutral comm.., NEUTRAL P-u. negative comm., SAD
Empathizing	A：”ロクバンいなければいいんだけどね。” (“It would be nice if there would be no number six”), B：“うん” (“Yep”), A：”と思ったら、ここにイッピキいた” (“But here is one…”), **B：“マジで？” (“Seriously?”),**	P-u. neutral comm., DISGUST Paying attention, NEUTRAL P-u. negative comm., DISGUST **Empathizing, FEAR**
Accepting as superior (showing deference)	A：“じゃたまり場来て” (“Come to the gathering spot!”), **B：“わかりました” (“Understood!”),**	Commanding/req. NEUTRAL **Accepting as sup., NEUTRAL**
Agreeing	B：”もうちょっと遅く出りゃいいのに” (“Couldn’t it appear a little bit later?”), A：”｛笑｝” (“{Laughter}”), A：”ほんとひどいよ。” (“It really is cruel.”), **A：”そうだようー” (“Yes, it is!”)**	P-u. negative comm., ANGER Empathizing, JOY Empathizing, ACCEPTANCE **Agreeing, ACCEPTANCE**
Self-image justification	A：“いっぱいいるねー。” (“There is a lot here!”) B：”ね。” (“Yep.”),	P-u. positive comm., NEUTRAL Paying attention, NEUTRAL

	A：”すぐ倒せるし” (“I can defeat them in an instant as well”),	**Self-image justification, JOY**
Criticizing	**A : “え、あれでいいの？｛笑｝”（“Are you sure you will be alright like that? [laughter]”）,** B：“ん?” (“What?”), B：“まあ自分でヒールできるしね” (“Well I can heal myself, so..”),	**Criticizing, JOY** Paying attention, NEUTRAL Self-image justification, NEUTRAL
Inadequate commenting	A : “また死んだ？” (“You died again?”), B : “死んだ。” (“I did”), B : “こいつ強すぎる。” (“It’s too strong”),	Inadequate commenting, JOY Self-image justification, SADNESS Self-image justification, SADNESS
Indebting partner	A：”やべ、俺が来たらなんかランポス復活してるんですけど。” (“This is bad! Now, that i have arrived, lanpos (game content) is somehow revived (and nowhere to be found). **B : “取っていってやる” (“I will take it for you.”),**	P-u. negative commenting., FEAR **Ind. partner, ACCEPTANCE**
Commanding / requesting	**A：“じゃたまり場来て” (“Come to the gathering spot!”),** B：“わかりました” (“Understood!”),	**Commanding/req., NEUTRAL** Accepting as sup., NEUTRAL

APPENDIX II.

The IA model conforms to the ISO Standard for Dialogue Act Annotation 24617-2 [Bunt et al. 2012; Bunt et al. 2017] in the following:

- The categorization differentiates between semantic and functional content.
- The dialogue acts defined in the model represent communicative functions.
- The represented communicative functions can be associated with a specific dimension. Therefore, the proposed tagset can be used as a function-specific dimension, and can be combined with all general and function-specific acts defined by the standard or by its representative tagset, the DIT++. For example, the utterance "うん" ("Mhmm/Yep") can be regarded as a 'paying attention' interpersonal act, and as either the DIT++ <'general-purpose, answer'> act or the function-specific <'auto feedback, auto-positive feedback'> act. (For further details on the DIT ++ dialogue acts, see Section 3.2.2.)
- The defined acts are intended to correspond to functional segments (minimal stretches of behavior having one or more communicative functions).

The IA model does not conform to the following aspects of the ISO 24617-2 standard:

- Functional dependency relations, feedback dependency relations, and rhetorical relations are not accounted for by the proposed model. The responsive interpersonal acts of 'paying attention' and 'empathizing' are typically used to represent functional- and feedback-dependency relations. However, their purpose in the developed taxonomy is solely to account for the dimension of interpersonal relations-managing actions, which includes these actions of responsive nature. All other acts can have functional- and feedback-dependency relations with each other, and all 12 acts can perform rhetorical functions, based on the dialogic situation.
- Although qualifiers (e.g., certainty, conditionality, partiality, or sentiment) can be attached to the proposed tags, the IA tagset used in the study does not assume the use of qualifiers, owing to the following:
    - Allowing for qualifiers would lead to a large number of possible tags that would make the annotation process unnecessarily confusing and prohibitively time-consuming.
    - Due to the large number of possible tags, each particular tag would be associated rarely, if ever, with a given emotion, even if a larger corpus were used.
    - Interpersonal acts have already been defined, with the intention to serve as indirect sentiment qualifiers themselves.

Appendix III. Hyperparameters of the Sub-classifiers' neural networks

Sub-classifier	Layers	Activation function	Loss function	Optimizer
Sub-classifier #1	GRU	-	Categorical cross entropy	AdaMax*
	GRU	-		
	Fully connected layer	Softmax		
Sub-classifier #2	Convolutional 1D	ReLU	Categorical cross entropy	Adam
	Pooling layer	-		
	Convolutional 1D	ReLU		
	Pooling layer	-		
	Convolutional 1D	ReLU		
	Pooling Layer	-		
	Fully connected layer	ReLU		
	Fully connected layer	Softmax		
Sub-classifier #3	GRU	-	Categorical cross entropy	Adam
	Fully connected layer	Softmax		
Ensemble method (soft voting)	Fully connected layer	ReLU	Categorical cross entropy	Adam
	Fully connected layer	ReLU		
	Fully connected layer	Softmax		

*For details on AdaMax and Adam optimization methods see [Kingma and Ba 2015]

Appendix IV. Averaged weights of sub-classification results used in the soft-voting process

Method	Sub-classifier #1 (GRU)	Sub-classifier #2 (CNN)	Sub-classifier #3 (GRU)
Baseline method (Sub-classifier #1 and #2)	0.69	0.31	-
Baseline method + Sub-classifier #3 processing labels from the DIT ++ (with best performing setup)	0.43	0.38	0.19
Baseline method + Sub-classifier #3 processing labels from SWBD-DAMSL (with best performing setup)	0.54	0.34	0.12
Baseline method + Sub-classifier #3 processing labels from the IA model (with best performing setup)	0.40	0.31	0.29

REFERENCES

ABDEL-HAMID, O. DENG, L. and YU, D. Exploring convolutional neural network structures and optimization techniques for speech recognition. Proceedings of 14th Annual Conference of the International Speech Communication Association, INTERSPEECH, 2013,

ANG, J. DHILLON, R. KRUPSKI, A. SHRIBERG, E. and STOLCKE, A. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. Proceedings of the 7th International Conference on Spoken Language Processing, INTERSPEECH, 2002.

ARIMOTO, Y. MORI, H. Emotion category mapping to emotional space by cross-corpus emotion labeling. Proceedings of The International Conference on Situated Interaction, INTERSPEECH, 2017.

BATLINER, A. FISCHER, K. HUBER, R. SPILKER, J. and NOTH, E. 2003. How to find trouble in communication, Speech communication, 40(1), 117-143.

BROWN, P. LEVINSON, S. C. Politeness: Some universals in language usage (Vol. 4), Cambridge University Press, Cambridge, 1987.

BUNT, H. The DIT++ taxonomy for functional dialogue markup. Proceedings of AAMAS 2009 Work., 2009, 13-24.

BUNT, H. 2011. Multifunctionality in dialogue. Comput. Speech Lang. (25), 222–245.

BUNT, H. ALEXANDERSSON, J. CHOE, J. FANG, A.C. HASIDA, K. PETUKHOVA, V. POPESCU-BELIS, A. and TRAUM, D. ISO 24617-2 : A semantically-based standard for dialogue annotation. Proceedings of LREC 2012, 2012, 430–437.

BUNT, H, PETUKHOVA, V. TRAUM, D. and ALEXANDERSSON, J. Dialogue Act Annotation with the ISO 24617-2 Standard. Deborah Dahl (ed.) Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything, Springer, Berlin, 2017, 109-135.

BURKHARDT, F. VAN Ballegooy, M. ENGELBRECHT, K.P. POLZEHL, T. and STEGMANN, J. 2009, September. Emotion detection in dialog systems: applications, strategies and challenges. Proceedings of 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, 2009,1-6.

CHUNG, J. GULCEHRE, C. CHO, K. and BENGIO Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv1–9. 2014.

DUNCAN, D. SHINE, G. and ENGLISH, C. Facial Emotion Recognition in Real Time, Report, Stanford, 2016.

EKMAN, P. Universals and cultural differences in facial expressions of emotion. Proceedings of the Nebraska Symposium on Motivation, 1971, 207–282.

ELLSWORTH, P. and SCHERER, K. 2003. Appraisal processes in emotion. Handbook of Affective Sciences, Oxford University Press, Oxford, 2003, 572–595.

EYBEN F. WOLMILLER, M. and SCHULLER, B. OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. Proceedings of the International Conference on Affective Computing and Intelligent Interaction, IEEE, 2009, 1-6.

FAYEK, H. M. LECH, M. and CAVEDON, L. Towards real-time speech emotion recognition using deep neural networks. Proceedings of the Conference on Signal Processing and Communication Systems (ICSPCS), IEEE, 2015, 1-5.

FRIJDA, N. H. 1987. Emotion, cognitive structure, and action tendency. Cogn. Emot. (1), 115–143.

IHASZ, P. L. VAN, T. H. KRYSSANOV, V. V. A Computational Model for Conversational Japanese, Proceedings of 2015 International Conference on Culture and Computing, 2015, 64–71.

JURAFSKY, D. SHRIBERG, E. and BIASCA, D. Switchboard SWBD-DAMSL shallow discourse-function annotation (coders manual, draft 13). Technical Report 97-02, University of Colorado, Institute of Cognitive Science, Colorado, 1997.

KINGMA, D. P. and BA, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

LAZARUS, R.S. Emotion and Adaptation, Oxford University Press, Oxford, 1991.
LEE, C. M. and NARAYANAN, S. S. 2005. Toward detecting emotions in spoken dialogs, IEEE Transactions on speech and audio processing; 13(2), 293-303.
LISCOMBE, J. RICCARDI, G. and HAKKANI-TUR, D. Using context to improve emotion detection in spoken dialog systems. Proceedings of the 9th European Conference on Speech Communication and Technology. INTERSPEECH, 2005.
MATEAS, M. and STERN, A. Structuring Content in the Façade Interactive Drama Architecture. Proceedings of the First Artificial Intelligence and Interactive Digital Entertainment Conference, 2005, 93-98.
MATSUMOTO, Y. 1988. Reexamination of the universality of face: Politeness phenomena in Japanese, Journal of pragmatics, 12(4), 403–426.
OBAID, M. HAN, C. and BILLINGHURST, M. Feed the Fish: an affect-aware game. Proceedings of the 5th Australasian Conference on Interactive Entertainment, ACM, 2008.
OPITZ, D. W. MACLIN, R. 1999. Popular ensemble methods: An empirical study, Journal of Artificial Intelligence Resources, 11, 169-198.
PANG, B. and LEE, L. 2008. Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, 2(1–2), 1-135.
PENNINGTON, J. SOCHER, R. MANNING, C.D. GloVe: Global Vectors for Word Representation. Proceedings of 2014 Conf. Empir. Methods Nat. Lang. Process., 2014, 532–543.
PLANET, S. and IRIONDO, I. Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition. Proceedings of the 7th Iberian Conference on Information Systems and Technologies (CISTI), IEEE, 2012, 1-6.
PLUTCHIK, R. 2001. The Nature of Emotions. American scientist, 89(4), 344-350.
POPESCU-BELIS, A. 2008. Dimensionality of dialogue act tagsets: An empirical analysis of large corpora. Language Res. Eval. (42), 99–107.
RUSSEL J. A. 1980. A circumplex model of affect. Journal of personality and social psychology. 39(6), 1161.
SZWOCH, M. and SZWOCH, W. 2014. Emotion Recognition for Affect Aware Video Games, Image Processing & Communications Challenges; 6 (313), 227.
TIAN, L. MOORE, J. D. and LAI, C. Emotion recognition in spontaneous and acted dialogues. Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction, IEEE, 2015, 698-704.
VOGT, T. ANDRE, E. and BEE, N. EmoVoice—A framework for online recognition of emotions from voice. Proceedings of the 4th Tutorial and Research Workshop on Perception in multimodal dialogue systems, IEEE, 2008, 188-199.
WALDEN, T. A. and SMITH, M. C. 1997. Emotion regulation, Motivation and emotion; 21(1), 7-25.
WIKIMEDIA PROJECT EDITORS. Wikimedia database dump of the Japanese Wikipedia on July 20, 2016, https://archive.org/details/jawiki-20160720, Last accessed: 2017/09/04.
YOON, H. PARK, S. LEE, Y. K. and JANG, J. H. Emotion recognition of serious game players using a simple brain computer interface. Proceedings of ICT Convergence (ICTC), IEEE, 2013, 783-786.