INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Recognition of Intentions of Users' Short Responses for Conversational News Delivery System

Hiroaki Takatsu[1], *Katsuya Yokoyama*[1], *Yoichi Matsuyama*[1],
Hiroshi Honda[2], *Shinya Fujie*[1,3], *Tetsunori Kobayashi*[1]

[1]Waseda University, Japan
[2]Honda R&D Co.,Ltd, Japan
[3]Chiba Institute of Technology, Japan

{takatsu,katsuya,matsuyama}@pcl.cs.waseda.ac.jp, Hiroshi_01_Honda@n.t.rd.honda.co.jp,
shinya.fujie@p.chibakoudai.jp, koba@waseda.jp

Abstract

In human-human conversations, listeners often convey intentions to their speakers through feedbacks comprising reflexive short responses. The speakers then recognize these intentions and dynamically change the conversational plans to transmit information more efficiently. For the design of spoken dialogue systems that deliver a massive amount of information, such as news, it is essential to accurately capture users' intentions from reflexive short responses to efficiently select or eliminate the information to be transmitted depending on the user's needs. However, such short responses from users are normally too short to recognize their actual intentions only from the prosodic and linguistic features of their short responses. In this paper, we propose a user's short-response intention-recognition model that accounts for the previous system's utterances as the context of the conversation in addition to prosodic and linguistic features of user's utterances. To achieve this, we define types of short response intentions in terms of effective information transmission and created new dataset by annotating over the interaction data collected using our spoken dialogue system. Our experimental results demonstrate that the classification accuracy can be improved using the linguistic features of the system's previous utterances encoded by Bidirectional Encoder Representations from Transformers (BERT) as the conversational context.

Index Terms: spoken dialogue system; intention recognition; neural networks

1. Introduction

In human–human conversations, listeners often convey their understanding level or degree of interest in the current topic to speakers through feedbacks comprising reflexive short responses. Based on the feedback from the listener, the speaker dynamically modifies his/her conversational plan to adjust the conversational situation to the state of the listener. A speaker may add supplemental information or change the topic when a listener's understanding level seems insufficient or when they seem to have no interest in the current topic, respectively. By applying such a series of processes to a human–system conversation, Fujie and Kobayashi developed a smooth and convenient conversation system [1, 2]. By providing or eliminating information depending on the user's needs [3], we develop a spoken dialogue system that efficiently delivers a massive amount of information. In the proposed system, any given written documents, such as news articles, can be translated into an utterance plan consisting of a primary plan for delivering main content and the associated subsidiary plans for supplementing the main content. By applying text summarization and style conversion techniques, a primary plan is automatically generated. The subsidiary plans are compiled by considering potential user/system interactions. To efficiently select or eliminate subsidiary plans depending on the user's needs, it is essential to accurately capture users' intentions from reflexive short responses while delivering the primary plan. Here, "intention" refers to certain types of user requests while consuming information, which are transmitted directly/indirectly and explicitly/implicitly by the users. An assumed example of the conversation between a user and our system is shown in Table 1. In the table, the utterances of U1 and U5 can be regarded as reflexive short responses that indicate the user's understanding. Note that U1 and U5 were articulated with the same Japanese pronunciation "*e?*" with different intents. U1 intends that the user did not know about Google Glass. U5 intends a surprise for wearing Google Glass for 18 hours a day. By accurately perceiving the difference between the two utterances, our system was able to define Google Glass and present more detailed information regarding Google Glass to the user. In this way, to transmit information efficiently, it is important to understand users' implicit information requests from reflexive short utterances. Conventionally, in spoken dialogue systems that recommend or transmit information, intentions are estimated from para-language [4] or explicit sentences [5] and are used for dialogue control. However, such users' reflexive short responses, such as U1 and U5 in Table 1, are normally too short to distinguish their actual intentions only by prosodic and linguistic features.

Table 1: *An assumed conversation between a user (U) and our system (S). Japanese pronunciations are written in Italic form.*

S1:	グーグル・グラスを Google Glass
U1:	え？ (*e?*) [**reflexive short response**] What's it?
S2:	グーグル・グラスはグーグルのメガネ型インターネット端末だよ Google Glass is Google's glasses-type Internet terminal.
U2:	へー I see.
S3:	それで、そのグーグル・グラスを着用した男性が依存症の治療を受けていたことが分かったんだって So, it seems it was found that the man who wore the Google Glass was being treated for addiction.
U4:	どれくらい着用してたの How long did he wear it?
S4:	1 日に最長 18 時間着用していたそうだよ He seems to have been wearing it for up to 18 hours a day.
U5:	え？ (*e?*) [**reflexive short response**] Too long!
S6:	男性は夢までもグーグル・グラス越しに見るようになったって言ってる The man says that he started viewing dreams as if they were projected through Google Glass's.
U6:	ふーん Hmmm.

In this paper, we propose a user's short response intention recognition model that takes the previous system's utterances into account as the context of the conversation in addition to prosodic and linguistic features of user's utterances. In the pro-

http://dx.doi.org/10.21437/Interspeech.2019-2121

posed model, users' intentions are identified using the prosodic features of the user's utterance extracted from the spectrogram and the linguistic features of the user's utterance and the previous system's utterances encoded by Bidirectional Encoder Representations from Transformers (BERT) [6]. The rest of the paper is organized as follows. Section 2 discusses related works done by other researchers. We introduce the dialogue data collected using our system in Section 3. Next, we describe classified intentions in the view of efficient information transmission and annotated labels on the utterance data in Section 4. Furthermore, we evaluate the effectiveness of the proposed model using this dataset in Section 5. Finally, we present a conclusion of this paper in Section 6.

2. Related Work

Spoken dialogue systems for information seeking tasks, such as item recommendation, typically control their dialogue flows based on users' explicit intentions, such as questions and recommendation queries, which are estimated from user's explicit utterances. There are various methods used for recognizing intentions from prosodic and linguistic features of the users' utterances. Yoshino et al. developed a news navigation system that transmits information adapted to the user's interests [5]. Using linguistic features, such as part-of-speech tags and semantic role labels, they proposed a method to estimate these intentions through logistic regression assuming that users' utterances are given in the form of explicit sentences. Conventionally, supplemental prosodic features such as F0 have been used for intention recognition [7, 8, 9, 10]. Fujie et al. proposed a method for estimating whether the user's attitude to the system is positive or negative via Bayesian discrimination using para-linguistic information such as fundamental frequency (F0) [4]. In recent years, numerous methods used for recognizing intentions or emotions by using spectrograms with rich information close to raw speech signals have been increasing [12, 13, 15, 16, 17]. Using the features obtained by inputting a spectrogram, phase information, and MGDCC [11] into a convolutional neural network (CNN), Guo et al. proposed a model for identifying emotions using bidirectional long short-term memory (LSTM)[12]. Luo et al. proposed a model that identifies emotions by combining the features obtained by inputting a spectrogram into a convolutional recurrent neural network (CRNN) and manually designed features such as F0 and MFCC [13]. Yenigalla et al. proposed a model to identify emotions by combining the features obtained by inputting a spectrogram into CNN and embedded representations of the phonemes obtained by word2vec [14] [15]. By combining linguistic features of user's utterance and the prosodic features extracted from a spectrogram [18, 19], we proposed a model to identify intentions.

In contrast, conversational media that assume that users primarily listen to a certain amount of information, e.g., listening to music, and sometimes spontaneously ask for clarification and details of the contents should accept users' explicit and implicit "pull" and "push" requests to retrieve a certain amount of information via natural methods. The users' intents may include backchannels showing that a user is listening to the news or showing his/her interests and questions with ambiguous short phrases (recall U1 and U5 in Table 1). However, such users' responses are normally too short to distinguish their actual intentions only from the para-linguistic and linguistic features of the users' responses. In this paper, we propose an intention recognition model that considers not only prosodic and linguistic features of user utterances but also previous system utterances.

Table 2: *User intention types, their effects, and corresponding system actions*

Effects	User intention	System action
Increasing information	*Question*	Answer
	Request Supplement	Provide explanations
	Request Repeat	Repeat
Decreasing information	*Disinterest*	Change topic
	Already Known	Omit details
Avoiding overlaps	*Wait Request*	Listen

Table 3: *Statistics of the dataset*

User intention	Majority vote	At least one vote
Question	2373	2913
Request Supplement	1218	6946
Request Repeat	23	380
Disinterest	394	2850
Already Known	134	732
Wait Request	296	1468

3. Dataset

We classified user intentions that demand the system to increase or decrease the amount of information to transmit [19]. Table 2 shows the user intention types, their effects, and corresponding system actions. We defined *Question*, *Request Supplement*, and *Request Repeat* as intention to demand increase of information to transmit. We defined *Disinterest* and *Already Known* as intention to demand decrease of information to transmit. In addition, we defined *Wait Request* to make the system wait to speak so that simultaneous utterances by the user and system would not occur. We employed seven annotators to annotate these intention labels to users' utterance data collected by our spoken dialogue system [3]. Among all the collected user utterances, we extracted short user utterances of less than 1.5 seconds using a voice activity detection (VAD) program. We allocated at least three annotators to annotate each short utterance. Table 3 shows the distribution of each label. The "Majority vote" column refers to the number of annotated utterance data that got a majority vote among more than three annotators. Additionally, the "At least one vote" column refers to the number of annotated utterance data that got at least one vote among more than three annotators.

4. Intention Recognition Model

An overview of the intention recognition model we propose is shown in Figure 1. First, a spectrogram is generated from a short time width fragment of speech. Next, the obtained spectrogram is input to an AutoEncoder including CNN (CNN-AutoEncder) and the prosodic features compressed in the intermediate layer is input to LSTM along time series. LSTM sequentially outputs the probability of the intention. When a speech recognition result is obtained, the prosodic features contained in LSTM, the linguistic features of the user's utterance, and the linguistic features of the previous system's utterances are integrated; then, the final probability of the intention is calculated using these features.

4.1. Design of Feature Extraction Part

In general, fundamental frequency (F0) is used as a feature to recognize users' intentions. However, due to the quasi-periodicity of the speech waveform, ambient noise, and variation in the F0 in the voice spanning over a wide range, it is difficult to accurately extract the F0. We developed a model

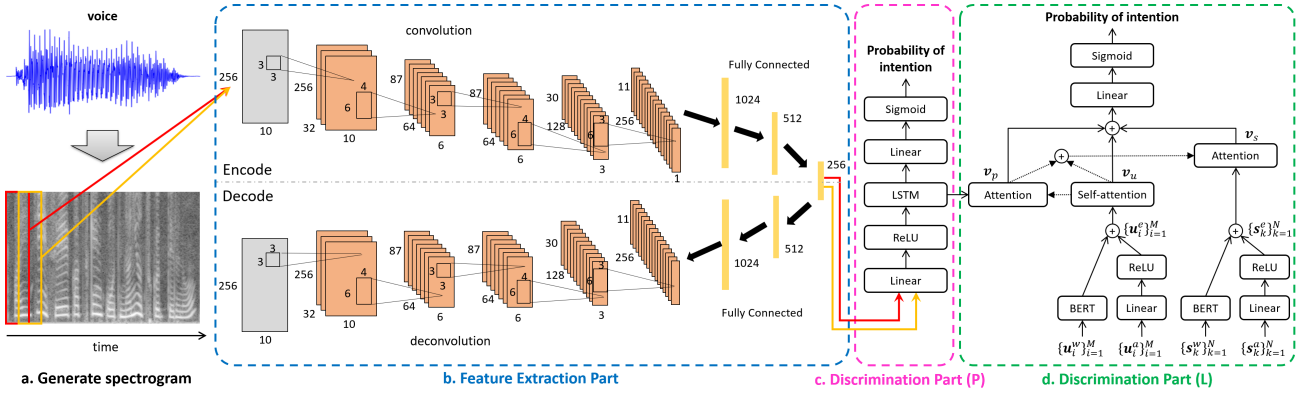

Figure 1: *Intention recognition model. (a) Generate a spectrogram from short time width fragment of speech. (b) Input the obtained spectrogram into AutoEncoder including CNN and obtain an intermediate layer vector as compressed prosodic features. (c) Input the obtained prosodic features into LSTM along time series and estimate intention sequentially. (d) Integrate the prosodic features contained in LSTM, the linguistic features of the user's utterance and those of the system's immediately preceding utterances to estimate the final intention.*

of extracting features directly from the time-frequency spectrum of speech without involving F0 estimation [18, 19]. The characteristics related to phoneme and voice height appear as a spectrogram pattern. Therefore, we consider this pattern as an image and train the CNN-AutoEncoder (Figure 1.b). We use the prosodic features compressed in the intermediate layer to recognize the users' intentions.

4.2. Design of Discrimination Part

The discrimination part comprises a discrimination part (P) and a discrimination part (L).

4.2.1. Discrimination Part (P)

In the discrimination part (P), the prosodic features extracted from the CNN-AutoEncoder are sequentially input to the LSTM to identify the users' intentions (Figure 1.c). Since speech has a variable length in the time direction, a model that can obtain the length in the time direction is necessary. Even for the sound of the same speech sentences, voice duration varies depending on said the speaker and their condition or state. Furthermore, the changes in the time direction of prosody are also useful in intention recognition, such as being easily recognized as *Question* when the pitch of the end of utterance rises. Therefore, we use the LSTM, which can deal with long time sequences for feature extraction.

4.2.2. Discrimination Part (L)

In the discrimination part (L), intentions are identified using the prosodic features included in the LSTM of discrimination part (P), the linguistic features of user's utterance, and the system's immediately preceding utterances. Here, the system's immediately preceding utterances represent the system's current and previous utterances when the user's utterance is obtained. For example, in the conversation of Table 1, when identifying the intention of the user's utterance U5, the range of the system's utterances used as the context is the current system's utterance S4 and the previous system's utterance S3. We used BERT [6] to encode linguistic features. We describe the flow of processing in the discrimination part (L). First, we analyze the speech recognition result of the user's utterance and obtain a sequence of words $\{\boldsymbol{u}_i^w\}_{i=1}^M$ and a sequence of auxiliary features (e.g., part-of-speech) $\{\boldsymbol{u}_i^a\}_{i=1}^M$. Similarly, for the system's utterances, we also obtain a sequence of words $\{\boldsymbol{s}_k^w\}_{k=1}^N$ and a sequence of auxiliary features $\{\boldsymbol{s}_k^a\}_{k=1}^N$. Next, a vector sequence $\{\boldsymbol{u}_i^e\}_{i=1}^M$ is calculated by combining a representation obtained by inputting the word sequence of the user's utterance into the BERT and a representation obtained by linear transformation of the auxiliary features. Similarly, for the system's utterances, a vector sequence $\{\boldsymbol{s}_k^e\}_{k=1}^N$ is calculated. Then, to obtain a vector $\boldsymbol{v}_u$, self-attention [20] is applied to the encoded result of the user's utterance. To obtain a context vector $\boldsymbol{v}_p$, we compute attention [21] for each state of the LSTM holding prosodic features using the vector $\boldsymbol{v}_u$ as a query. To obtain a context vector $\boldsymbol{v}_s$, we compute attention [21] for the encoded result of the system's utterances using a vector combined $\boldsymbol{v}_u$ and $\boldsymbol{v}_p$ as a query. Finally, using a vector combining $\boldsymbol{v}_p$, $\boldsymbol{v}_u$ and $\boldsymbol{v}_s$, the probability of the intention is calculated in the output layer.

5. Experiments

5.1. Learning of Feature Extraction Part

We use a large speech corpus (Corpus of Spontaneous Japanese: CSJ) [22] to train the CNN-AutoEncoder. The input to the CNN-AutoEncoder is a spectrogram generated by frame size of 800 (50ms), frameshift of 160 (10ms), and chunk size of 1024 and these sizes are 10×256.

5.2. Pre-training of BERT

In this section, we describe the experimental setup about pre-training of the BERT. We extracted adjacent sentence pairs from about 1 million articles of Wikipedia without duplication so that the paragraph was not exceeded. Among the sentence pairs, 5 million sentence pairs were used as a training set and 10 thousand sentence pairs were used as a development set, and the BERT was pre-trained with 2 tasks: masked language model and next sentence prediction [6]. The model parameters were set such that the number of blocks in the transformer was 2, the dimension of the hidden layer was 256, and the number of heads in the self-attention was 2.

5.3. Auxiliary Features

In this section, we describe the details of the auxiliary features. In our system that speaks based on scenarios, system's utter-

Table 4: *Experimental results. P: The result of estimation obtained by using only prosodic features of the user's utterance. P+U: The result of estimation obtained by using prosodic features and linguistic features of the user's utterance. P+U+S: The result of estimation obtained by using prosodic features, linguistic features of user's utterance, and linguistic features of the system's utterances.*

			w/ pre-training BERT w/ auxiliary features		w/ pre-training BERT w/o auxiliary features		w/o pre-training BERT w/o auxiliary features	
		P	P+U	P+U+S	P+U	P+U+S	P+U	P+U+S
Question	Majority vote	0.920	0.958	**0.962**	0.957	0.961	0.940	0.951
	At least one vote	0.902	0.957	0.961	0.953	0.956	0.923	0.938
Request Supplement	Majority vote	0.836	0.911	**0.913**	0.901	0.906	0.895	0.896
	At least one vote	0.827	0.882	0.887	0.881	0.887	0.847	0.865
Request Repeat	Majority vote	-	-	-	-	-	-	-
	At least one vote	0.700	0.826	**0.832**	0.795	0.805	0.784	0.805
Disinterest	Majority vote	0.893	**0.913**	0.908	0.903	0.898	0.898	0.897
	At least one vote	0.797	0.811	0.805	0.810	0.804	0.804	0.803
Already Known	Majority vote	0.818	**0.939**	**0.939**	**0.939**	**0.939**	**0.939**	**0.939**
	At least one vote	0.702	0.825	0.831	0.817	0.828	0.806	0.817
Wait Request	Majority vote	0.716	**0.885**	0.865	0.885	0.858	0.845	0.838
	At least one vote	0.614	0.688	0.692	0.688	0.689	0.634	0.672

ances can be analyzed in advance. Therefore, we can use rich features. We used JUMAN++ [23, 24] (Ver.1.02) as a morphological analyzer and KNP [25] (Ver.4.19) as a parser. We used part-of-speech of the word, conjugated form, category, domain, TF, IDF, TF-IDF, named entity class, whether the word is included in corner bracket or not, dependency type of the clause, depth of the clause in the dependency tree, and position of the clause from beginning of the sentence as features. On the other hand, the user's utterances need to be analyzed in real time. Therefore, to analyze user's utterances and used part-of-speech of the word, conjugated form, category, domain as features, we applied only JUMAN [26] (Ver.7.01).

5.4. Experimental Setup

The training set, development set, and test set used in model learning and evaluation were made by dividing the dataset in Table 3 at a ratio of 2:1:1 for each intention. The same number of negative examples as positive examples were randomly extracted from data of other intentions. The evaluation metric is accuracy. The discrimination part (P) is trained using a 256-dimensional vector, which is the intermediate layer vector of the CNN-AutoEncoder. To the CNN-AutoEncoder, a 100 ms spectrogram was shifted in intervals of 50 ms. In other words, there was an overlap of 50 ms with the input of the current time and the input of the next time. This was to let the LSTM learn information about the time change. We represent the result of calculating the accuracy for the test set based on the output of the final state of the LSTM as "P". We set the dimensions of the linear and hidden layers of the LSTM to be 64. Using the hidden states of the LSTM of the discrimination part (P) and linguistic features of the user's utterance, the discrimination part (L) that excluded the structure of processing the system's utterances from the network was trained. We represent the result of calculating the accuracy for the test set as "P+U". All of the networks of the discrimination part (L) were trained using the hidden states of the LSTM of the discrimination part (P), the linguistic features of the user's utterance, and the linguistic features of the system's utterances. We represent the result of calculating the accuracy for the test set as "P+U+S". In addition, we compared the model that the BERT was pre-trained and the model that the BERT was not pre-trained. We also compared the model that input auxiliary features and the model that did not input them. Furthermore, we compared the model that trained using "At least one vote" dataset in Table 3 and the model that trained using "Majority vote" dataset.

5.5. Experimental Results

The experimental results are shown in Table 4. For all intentions, we found that adding linguistic features of the user's utterance is better than only inputting prosodic features. In addition, it was found that accuracy can be improved by adding auxiliary features by pre-training the BERT. As for the quality of the dataset, the performance using the dataset that adopted a label annotated by majority annotators was higher than the performance using the dataset that adopted a label annotated by at least one annotator. With respect to the effects of taking into consideration the linguistic features of the system's utterances, the performance of *Question*, *Request Supplement*, *Request Repeat*, *Already Known*, and *Wait Request* (At least one vote) were improved. However, the performance of *Disinterest* and *Wait Request* (Majority vote) were slightly degraded. *Wait Request* tends to extend the duration of the end of the utterance and to be grammatically incomplete. Therefore, it is considered that the reason for this is that the feature of the user's utterance contributes more to the identification than the context of the conversation. Users' interest in topics varies. Therefore, as for *Disinterest*, it is thought that the result was worse because the variation in context is large.

6. Conclusion

To communicate information efficiently via spoken dialogues, we examined a method to identify the user's intention from reflexive short responses. Since the interpretation of reflexive short responses depends on the context of the conversation, we proposed an intention recognition model that considers not only prosodic features and linguistic features of user's utterance but also context information of system's utterances. We classified the intentions in terms of efficiently transmitting information and annotated them on the dialogue data collected by the actual conversation system. As a result of the experiment using this dataset, we confirmed that the discrimination performance of the intentions is improved by using the linguistic features of the immediately preceding system's utterances. Furthermore, we confirmed that the model performance is improved by adding auxiliary features by pre-training the BERT by improving the quality of the dataset. In the future, we would like to consider a multimodal intention recognition method that also considers visual cues such as facial expressions.

7. References

[1] S. Fujie, R. Miyake, and T. Kobayashi, "Spoken dialogue system using recognition of user's feedback for rhythmic dialogue," in *Proceedings of the Speech Prosody*, pp. 1–4, 2006.

[2] T. Kobayashi and S. Fujie, "Conversational robots: An approach to conversation protocol issues that utilizes the paralinguistic information available in a robot-human setting," in *Acoustical Science and Technology*, vol. 34, no. 2, pp. 64-72, 2013.

[3] H. Takatsu, I. Fukuoka, S. Fujie, Y. Hayashi, and T. Kobayashi, "A spoken dialogue system for enabling information behavior of various intention levels," in *Journal of the Japanese Society fot Artificial Intelligence*, vol. 33, no. 1, pp. 1–24, 2018. (in Japanese)

[4] S. Fujie, Y. Ejiri, H. Kikuchi, and T. Kobayashi, "Dialogue robot with an ability to understand para - linguistic information," in *Research Report of the Information Processing Society of Japan, Spoken Language Processing*, vol. 2003, no. 104, pp. 13–20, 2003. (in Japanese)

[5] K. Yoshino and T. Kawahara, "Conversational system for information navigation based on POMDP with user focus tracking," in *Computer Speech & Language*, vol. 34, no. 1, pp. 275–291, 2015.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, pp. 1–14, 2018.

[7] R. Nisimura, S. Omae, H. Kawahawa, and T. Irino, "Analyzing dialogue data for real-world emotional speech classification," in *Proceedings of the 7th Annual Conference of the International Speech Communication Association*, pp. 1822–1825, 2006.

[8] Y. Hayashi, S. Osaragi, and Y. Nakano, "Estimating utterance tags using prosodic features in collaborative learning,", in *Proceedings of the 39 Annual Conference of Japanese Society of Information and Systems in Education*, pp. 441–442, 2014.

[9] A. Ando, T. Asami, M. Okamoto, H. Masataki, and S. Sakauchi, "Agreement and disagreement utterance detection in conversational speech by extracting and integrating local features," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, pp. 2494–2498, 2015.

[10] M. Tamoto and T. Kawabata, "A schema for illocutionary act identification with prosodic feature," in *Research Report of the Information Processing Society of Japan, Spoken Language Processing*, vol. 1998, no. 68, pp. 55–60, 1998. (in Japanese)

[11] R. M. Hegde, , H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," in *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 1, pp. 190–202, 2006.

[12] L. Guo, L. Wang, J. Dang, L. Zhang, H. Guan, and X. Li, "Speech emotion recognition by combining amplitude and phase information using convolutional neural network," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pp. 1611–1615, 2018.

[13] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pp. 152–156, 2018.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of the 1st International Conference on Learning Representations*, pp. 1–12, 2013.

[15] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pp. 3688–3692, 2018.

[16] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, pp. 1089–1093, 2017.

[17] D. Tang, J. Zeng, and M. Li, "An end-to-end deep learning framework for speech emotion recognition of atypical individuals," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pp. 162–166, 2018.

[18] H. Takatsu, K. Yokoyama, H. Honda, S. Fujie, Y. Hayashi, and T. Kobayashi, "Utterance intention understanding for news articles transfer by conversation," in *Proceedings of the 32nd Annual Conference of the Japanese Society for Artificial Intelligence*, pp. 1–4, 2018. (in Japanese)

[19] K. Yokoyama, H. Takatsu, H. Honda, S. Fujie, and T. Kobayashi, "Investigation of users' short responses in actual conversation system and automatic recognition of their intentions", in *Proceedings of the the 2018 IEEE Workshop on Spoken Language Technology*, pp. 934–940, 2018.

[20] Z. Lin, M. Feng, C. N. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *Proceedings of the 5th International Conference on Learning Representations*, pp. 1–15, 2017.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3th International Conference on Learning Representations*, pp. 1–15, 2015.

[22] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12, 2003.

[23] H. Morita, D. Kawahara, and S. Kurohashi, "Morphological analysis for unsegmented languagesusing recurrent neural network language model," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2297, 2015.

[24] A. Tolmachev, D. Kawahara, and S. Kurohashi, "Juman++: A morphological analysis toolkit for scriptio continua," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations)*, pp. 54–59, 2018.

[25] D. Kawahara and S. Kurohashi, "A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 176–183, 2006.

[26] S. Kurohashi and D. Kawahara, "JUMAN (a user-extensible morphological analyzer for Japanese)," `http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN`.