Type of Response Selection utilizing User Utterance Word Sequence, LSTM and Multi-task Learning for Chat-like Spoken Dialog Systems

Kengo Ohta*, Ryota Nishimura†, Norihide Kitaoka‡
* National Institute of Technology, Anan College, Tokushima, Japan
E-mail: kengo@anan-nct.ac.jp Tel: +81-88-423-7228
† Tokushima University, Tokushima, Japan
‡ Toyohashi University of Technology, Aichi, Japan

Abstract—**This paper describes a method of automatically selecting types of responses, such as back-channel responses, changing the topic or expanding the topic, in conversational spoken dialog systems by using an LSTM-RNN-based encoder-decoder framework and multi-task learning. In our dialog system architecture, response utterances are generated after the response type is explicitly determined in order to generate more appropriate and cooperative response than the conventional end-to-end approach which generate response utterances directly. As a response type selector, an encoder and two decoders share states of hidden layers and are trained with the interpolated loss function of the two decoders. One of the decoders is for selecting types of responses and the other is for estimating the word sequence of the response utterances. In an evaluation experiment using a corpus of dialogs between elderly people and an interviewer, our proposed method achieved better performance than the standard method using single-task learning, especially when the amount of training data was limited.**

I. Introduction

Task-oriented spoken dialog systems, such as personal assistants (e.g. Amazon's Alexa[1], Apple's Siri[2], Microsoft's Cortana[3], and Google's Now[4]), which are designed to fulfill a user's verbal requests, are now being widely used on a daily basis. Furthermore, non-task-oriented spoken dialog systems, such as conversation robots [1] (also known as chatbots), are expected to be widely used in future applications such as cognitive training or increasing the communication opportunities of elderly people. We believe that these chat-like interfaces will also be important for communication with humanoid robots [2] in the future. Based on a general recognition of these developments, a balanced corpus of daily conversation has been developed for the analysis of turn-taking during conversations [3]. The primary aim of such non-task-oriented conversation systems is for users to enjoy the conversation itself, thus it is more important for chatbots to be able to prolong a natural conversation as long as possible than to satisfy a user's specific demands. In human conversation, speakers select a response from various types of possible responses, such as back-channel responses (e.g., ”uh-huh”, ”hmm”, “really?”, “wow”, etc.), changing the topic, expanding the topic, etc., so chat-like spoken dialog systems should also have the ability to imitate this behavior. In order to imitate this behavior, in our spoken dialog system architecture, the type of response is explicitly determined by the response type selector, and then, the response utterance which matches the type of response is synthesized (as described in Section IV-A). We consider that this architecture enables us to generate more appropriate and cooperative response than the conventional end-to-end architecture which tends to generate highly generic responses such as ”I don't know” regardless of the user's intension.

In this paper, we propose a method of selecting the type of system response based on the word sequence of the user's utterances in a non-task-oriented conversational dialog system, in a manner which is likely to prolong a conversation. In a previous study, we proposed a support vector machine (SVM) [4] based framework [5] for this task. However, the performance of this method was limited for two reasons. First, point-wise classifiers such as SVM are not suitable for sequence classification problems in which previous samples affect succeeding samples, or for considering word order in each utterance sample. Second, the SVM used in our previous study was trained using only the input utterances from a limited training corpus. To address these limitations, in this paper we introduce an improved framework which employs an encoder-decoder model and uses multi-task learning and multiple decoders. We also use recurrent neural networks (RNNs) with long short-term memory units (LSTM-RNN), which are suitable for considering the word sequence in each utterance. Our framework is evaluated in an experiment using a self-developed conversation corpus containing exchanges between elderly people and an interviewer. One of our research goals is to develop a dialog system for reminiscence therapy for the elders, so we collected the utterances of elderly people for use in this study. The utterances of the elderly people during the interviews are used to represent a system user's input utterances to the system, and the interviewer's responses serve as a reference for the selection of an appropriate type of response.

[1] https://www.alexa.com/
[2] https://www.apple.com/ios/siri/
[3] https://www.microsoft.com/en-us/windows/cortana
[4] https://www.google.com/search/about/

This paper is organized as follows. We first discuss some related studies in Section II. In Section III we describe the development of our speech corpus. In Section IV we explain our response selection method. We then evaluate the proposed method in Section V and conclude the paper in Section VI..

II. Related Work

LSTM-RNNs have been applied in many areas of natural language and spoken language processing. Encoder-decoder frameworks based on RNNs have been especially successful when applied for machine translation [6][7]. In these frameworks, the encoder receives the word sequence of a user's utterance in chronological order and embeds it in a fixed-length feature vector. The decoder then converts this feature vector into an output sentence in the target language. A similar approach has also been accepted for use in dialog systems for the task of response generation [8] or response reranking [9] . In these applications, the RNN encoder is trained to embed necessary information into the vector for generating target sentences. In contrast, in this study we try to train our encoder to extract the necessary information for selecting an appropriate type of response for a spoken dialog system.

Another trend in the area of deep learning research is multi-task learning, which shares parameters or loss functions among multiple networks. Such a learning strategy has been successfully applied in the areas of natural language and spoken language processing [10][11]. We use three networks in our proposed method, namely, an encoder, a decoder for selecting the type of response and a decoder for estimating the word sequences of response utterances. These networks share the cell states of hidden layers and a loss function for more effective training.

TABLE I
Labels for Nine Types of Responses

Label	Response Type	Frequency
back	Back-channel response (neutral)	858
p-back	Back-channel response (positive)	276
n-back	Back-channel response (negative)	78
exp	Expand on the current topic	102
gin-up	Ginger/Liven up the conversation	79
change	Change the topic	32
smile	Smile	108
emp	Show empathy	49
non	Do nothing	405
Total		1,987

TABLE II
Number of Utterances of each type for each Speaker

Speaker	A	B	C	D	E	F	G	H
back	211	396	131	307	50	256	136	34
p-back	77	96	62	109	27	35	82	9
n-back	14	49	19	17	2	15	19	1
exp	46	33	27	13	11	9	19	5
gin-up	35	19	18	21	10	7	25	7
change	10	8	10	10	10	10	7	9
smile	41	38	35	19	9	27	24	3
emp	15	10	8	19	7	7	13	2
non	87	190	64	117	33	71	88	15
Total	536	839	374	632	159	437	413	85

III. Conversation Corpus

As mentioned above, one of the goals of this research is to build a reminiscence therapy dialog system for elderly people. In order to train and evaluate our classifier for response selection, we built a Japanese language conversation corpus of dialogs between elderly people and an interviewer, in cooperation with a nursing faculty. All of the dialogs were recorded in a low-noise environment. In each dialog, an elderly person speaks freely in response to ten questions asked by an interviewer (e.g., "Did you go somewhere recently?"). A total of 3,475 utterances from eight speakers were collected and manually classified. Here, each utterance is a unit of speech segmented by periods of silence of 200 milliseconds or longer. As the result of a preliminary investigation, these utterances were classified into nine categories, as shown in Table I, and all of the utterances were annotated with these labels for the supervised training of our classifier. The number of speech segments of each type is also shown in Table I. The word sequences of the utterances of both the elderly participants and the interviewer were also manually transcribed. The number of speech segments of each type for each speaker (A-H) is shown in Table II.

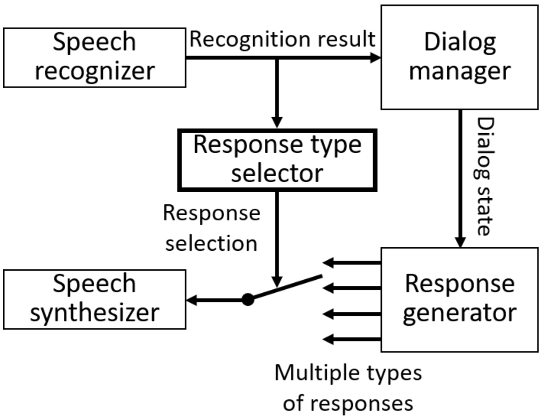

Fig. 1. Our System Architecture

IV. Proposed Method

A. Our System Architecture

The system architecture of our spoken dialog system is shown in Figure 1. A user's input utterance is first recognized by the speech recognizer. Based on the recognition result, the response type selector decides which type of response should be given, and the dialog manager tracks the state of the dialog. The response generator then generates multiple types of responses depending on the dialog state. Finally, a response which matches the response type determined by the response type selector is synthesized by the speech synthesizer as the system's next utterance.

In the next section, the model architecture of the response type selector (the bold block in Figure 1) is described in detail.

B. Response Selector

In our proposed method, the appropriate type of response to the user's input utterance is selected using an LSTM based

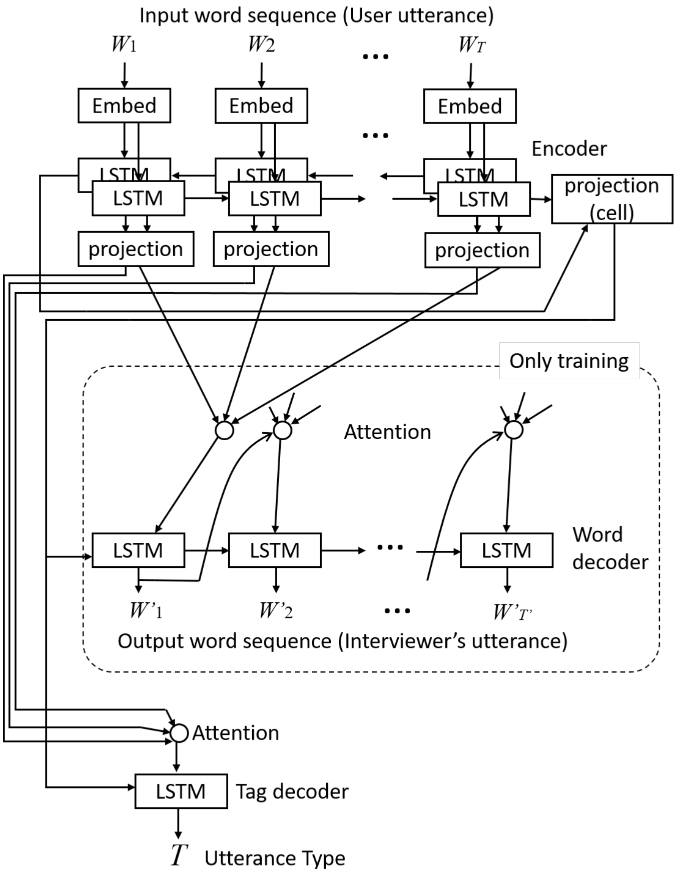

Fig. 2. The Model Architecture of Our Proposed Response Selector
Each input word is embedded in a 200-dimensional vector. The LSTMs in the encoder and the decoder have 200 hidden nodes. The “projection (cell)” has a fully connected feed-forward layer with 400 = 200 x 2 inputs and 200 outputs, which integrate the LSTM's forward and backward internal states. Each “projection” also has a fully connected feed-forward layer with 400 = 200 x 2 inputs and 200 outputs. The output word sequence decoder is only used in the training stage of multi-task learning.

encoder-decoder model.

An overview of our model is shown in Figure 2. An encoder is constructed using an attention-based bidirectional LSTM, and the two decoders are both constructed using a unidirectional LSTM. Here, one decoder (the “tag decoder”) is used for selection of the type of response, and the other decoder (the “word decoder”) is used for estimation of the word sequence for the response utterance. The size of the hidden layer of each LSTM is set to 200.

The training procedure for this model is as follows. First, word sequences in both the user's input utterance and the corresponding response utterance are converted into word sequences in distributed representations using word2vec[5], which is an implementation of Mikolov's method [12]. The distributed representation is trained using the articles of the Japanese edition of Wikipedia[6] from July 1st, 2017, which is tokenized using MeCab (ver. 0.996), a Japanese morphological analyzer with a custom dictionary （mecab-ipadic-NEologd ver. 0.0.5 [13]） which contains new words extracted from web documents. We adopted a skip- gram model for training, and the number of dimensions of the representation was set to 200. Then, the distributed word sequences of the input utterance and the response utterance, as well as the reference label of the type of response, are fed into the encoder, word decoder and tag decoder, respectively. Each network is trained with the training corpus, using the shared hidden layers and shared loss function described in next section.

TABLE III
Classification Result

Number of Training Samples for each Label	α	Accuracy
100	0	0.339
	0.9	0.350
500	0	0.356
	0.9	0.361

During the test step only the user's input utterance is fed into our model, and the utterance type is directly estimated using the encoder and the tag decoder.

C. Loss Function for Multi-task Learning

During the training of the previously described encoder-decoder model, back propagation is performed with the global loss function L, which is defined using a linear interpolation of L_{word} (the loss of the word decoder) and L_{tag} (the loss of tag decoder) as follows. Here, α represents an interpolation weight between 0 and 1:

$$L = \alpha L_{word} + (1 - \alpha) L_{tag} \tag{1}$$

L_{word} is defined as the sum of mean square errors. The tag decoder should output 0/1, so L_{tag} is a cross entropy loss.

V. Evaluation Experiment

A. Experimental Set-up

In order to evaluate our proposed method, we conducted evaluation experiments using the conversation corpus described in Section III. We compared a standard method using single-task learning, in which $\alpha = 0$ in Eq. (1), with our proposed method with multi-task learning, in which $\alpha = 0.9$. Note that, as shown in Table I, the frequencies of the various response labels are unbalanced. For example, there were far more “back” responses than “change” or “emp” responses. In order to avoid over-fitting to this bias, a balanced data set was used for training our model, which involved reducing the large samples and copying the small samples for each response label. Specifically, we compared balancing the number of training samples for each label at 100 and at 500. Similarly, the number of test samples for each response label was balanced at 20.

B. Experimental Results

Our classification results were evaluated on the basis of the classification accuracy for the nine types of response labels shown in Table III. As we can see, when the proposed method

[5] https://code.google.com/p/word2vec/
[6] https://ja.wikipedia.org/

TABLE IV
CONFUSION MATRIX（$\alpha = 0$, SINGLE-TASK LEARNING）

	Classified Class								
True Class	**back**	**p-back**	**n-back**	**exp**	**gin-up**	**change**	**smile**	**emp**	**non**
back	9	1	2	2	1	0	3	1	1
p-back	4	3	1	6	4	0	2	0	0
n-back	5	5	4	3	1	0	1	0	1
exp	2	3	2	6	1	0	5	0	1
gin-up	0	2	0	4	9	1	3	0	1
change	2	0	2	5	2	0	4	0	5
smile	5	2	1	3	2	0	4	0	3
emp	1	0	2	0	1	0	1	14	1
non	3	0	1	3	1	0	0	0	12

TABLE V
CONFUSION MATRIX（$\alpha = 0.9$, MULTI-TASK LEARNING）

	Classified Class								
True Class	**back**	**p-back**	**n-back**	**exp**	**gin-up**	**change**	**smile**	**emp**	**non**
back	7	2	2	4	2	0	2	1	0
p-back	2	4	1	4	2	1	5	0	1
n-back	5	1	7	3	1	1	2	0	0
exp	2	2	3	5	1	0	5	1	1
gin-up	1	3	1	5	6	1	3	0	0
change	2	1	2	3	1	1	6	0	4
smile	3	1	1	5	2	0	5	0	3
emp	1	1	2	1	0	0	0	15	0
non	2	1	1	2	1	0	0	0	13

with multi-task learning ($\alpha = 0.9$) was applied, classification performance was higher than the standard method with single-task learning ($\alpha = 0$). In particular, when the smaller amount of training data was used, classification performance was significantly improved by applying the proposed method, and it achieved comparable performance to the standard model with the larger amount of training data. These results suggest that multi-task learning is particularly effective when the amount of training data is limited.

The confusion matrix of our classification results when using standard, single-task learning is shown in Table IV, and the results when using the proposed, multi-task learning method is shown in Table V. For both methods, the number of training samples for each response label was set at 100. As shown in Tables IV and V, when multi-task learning was applied, classification accuracy for low frequency labels, such as ”n-back”, particularly improved. This suggests that the proposed encoder-decoder model was able to robustly learn the characteristics of these kinds of low frequency phenomena by utilizing not only the information from the user's input utterances, but also that from the response utterances.

VI. CONCLUSIONS

We proposed a method for selecting the type of response to be used by a spoken dialog system, using an LSTM-RNN based encoder-decoder model and multi-task learning. Both input utterances and response utterances were included in the training corpus, and the response utterances were also utilized to train the encoder-decoder model, which had multiple decoders which utilized a multi-task learning strategy. Since our target application was a reminiscence therapy system for the elders, we performed evaluation experiments using transcription data from conversations between elderly people and an interviewer. Our results showed the efficacy of the use of multi-task learning, especially when the amount of training data is limited. Moreover, the proposed method achieved the same level of classification performance as a standard method when the amount of training data is simply padded. Our results also suggested that our proposed method is particularly effective for the classification of low frequency labels, as it slightly outperformed the standard method.

In this study, we only utilized linguistic information from the word sequences of user utterances. In future studies, we plan to investigate the effects of using both linguistic and acoustic features for word sequencing, as we have confirmed that the use of acoustic features is effective in a previous study [5]. We would also like to investigate the effectiveness of using a longer dialog history. In addition, we intend to incorporate our response type selection technique into actual spoken dialog systems and evaluate the impressions of user's in subjective experiments.

VII. ACKNOWLEDGMENT

This work was supported in part by the Strategic Information and Communications R&D Promotion Program (SCOPE) of the Ministry of Internal Affairs and Communications of Japan, and also by the Grants-in-Aid for Scientific Research (19H01125, 19K04311) from the Japan Society for the Promotion of Science. We would also like to thank Taiki Yamamoto and Chen Jiahao of Tokushima University for their collaboration on this research.

REFERENCES

[1] N. Roy, G. Baltus, D. Fox, F. Gemperle, J. Goetz, T. Hirsch, D. Margaritis, M. Montemerlo, J. Pineau, J. Schulte *et al.*, "Towards personal service robots for the elderly," in *Workshop on Interactive Robots and Entertainment (WIRE)*, vol. 25, p. 184, 2000.

[2] K. Inoue, P. Milhorat, D. Lala, T. Zhao, and T. Kawahara, "Talking with erica, an autonomous android," in *the Annual Meeting of the Special Interest Group on Discourse and Dialogue(SIGDIAL)*, pp. 212–215, 2016.

[3] H. Koiso, T. Tsuchiya, R. Watanabe, D. Yokomori, M. Aizawa, and Y. Den, "Survey of conversational behavior: Towards the design of a balanced corpus of everyday japanese conversation," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pp. 4434–4439, 2017.

[4] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[5] K. Ohta, R. Marumotoa, R. Nishimuraa, and N. Kitaoka, "Selecting type of response for chat-like spoken dialogue systems based on acoustic features of user utterances," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1248–1252, 2017.

[6] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 3, no. 39, pp. 1700–1709, 2013.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[8] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[9] M. Inaba and K. Takahashi, "Neural utterance ranking model for conversational dialogue systems," in *The Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, vol. 3, no. 39, pp. 393–403, 2016.

[10] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," in *International Conference on Learning Representations (ICLR)*, 2016.

[11] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4835–4839, 2017.

[12] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 3111–3119, 2013.

[13] T. H. Toshinori Sato and M. Okumura, "Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese)," in *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*, 2017.