Speech Communication 133 (2021) 23–30

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Response type selection for chat-like spoken dialog systems based on LSTM and multi-task learning

Kengo Ohta [a,*], Ryota Nishimura [b], Norihide Kitaoka [c]

[a] National Institute of Technology, Anan College, Japan
[b] Tokushima University, Japan
[c] Toyohashi University of Technology, Japan

ARTICLE INFO

Keywords:
Spoken dialog system
Response type selection
Encoder–decoder model
Multi-task learning

ABSTRACT

We propose a method of automatically selecting appropriate responses in conversational spoken dialog systems by explicitly determining the correct response type that is needed first, based on a comparison of the user's input utterance with many other utterances. Response utterances are then generated based on this response type designation (back channel, changing the topic, expanding the topic, etc.). This allows the generation of more appropriate responses than conventional end-to-end approaches, which only use the user's input to directly generate response utterances. As a response type selector, we propose an LSTM-based encoder–decoder framework utilizing acoustic and linguistic features extracted from input utterances. In order to extract these features more accurately, we utilize not only input utterances but also response utterances in the training corpus. To do so, multi-task learning using multiple decoders is also investigated.

To evaluate our proposed method, we conducted experiments using a corpus of dialogs between elderly people and an interviewer. Our proposed method outperformed conventional methods using either a point-wise classifier based on Support Vector Machines, or a single-task learning LSTM. The best performance was achieved when our two response type selectors (one trained using acoustic features, and the other trained using linguistic features) were combined, and multi-task learning was also performed.

1. Introduction

Spoken dialog systems are being more widely used, in a variety of different applications. Task-oriented spoken dialog systems that attempt to fulfill a user's verbal requests, which include personal assistants such as Amazon's Alexa,[1] Apple's Siri,[2] Microsoft's Cortana[3] and Google's Now,[4] are already being widely used by the public. Non-task-oriented spoken dialog systems, such as conversation robots (Roy et al., 2000) (also known as 'chatbots'), are currently being used to answer frequently asked questions, for entertainment and in toys, and are expected to be widely used in the future in applications such as cognitive training and to increase communication opportunities for elderly people (Saczynski et al., 2006; Fratiglioni et al., 2000). It is likely that these interfaces will also be used for communication with humanoid robots (Inoue et al., 2016) in the future.

The primary aim of 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 natural conversations as long as possible than to satisfy a user's specific demands. To achieve this, a balanced corpus of everyday conversation is developed for the analysis of turn-taking during conversations (Koiso et al., 2017). During conversation, human speakers choose from a range of possible types of responses, such as back-channel responses (e.g., "uh-huh", "hmm", "really?", "wow", etc.), changing the topic, expanding the topic, etc. Chat-like spoken dialog systems also need to be able to imitate this behavior in order to maintain natural conversations. To achieve this, the architecture of our proposed spoken dialog system first determines the appropriate type of response using a response-type selector, then a response utterance which is consistent with that response type is synthesized. We believe this will enable our system to generate more appropriate and cooperative responses than conventional end-to-end architectures (Vinyals

* Corresponding author.
E-mail addresses: kengo@anan-nct.ac.jp (K. Ohta), nishimura@is.tokushima-u.ac.jp (R. Nishimura), kitaoka@tut.jp (N. Kitaoka).

[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/.

https://doi.org/10.1016/j.specom.2021.07.003
Received 22 April 2020; Received in revised form 2 May 2021; Accepted 3 July 2021
Available online 15 July 2021

and Le, 2015; Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015), which tend to generate highly generic responses (known as "dull responses"), such as "I don't know", regardless of the context (Serban et al., 2016; Li et al., 2015).

Therefore, in this paper we propose a method of selecting the correct type of system response in non-task-oriented, conversational dialog systems, using acoustic and linguistic features extracted from the user's utterances, in a manner which is likely to prolong a conversation. We introduce a novel framework which employs an encoder–decoder model based on recurrent neural networks (RNNs Robinson and Fallside, 1988) with long short-term memory units (LSTMs Hochreiter and Schmidhuber, 1997), which are suitable for evaluating the word sequence in each utterance. In order to enhance feature extraction, we utilize both input utterances and response utterances in the training corpus.

Multi-task learning using multiple decoders is also investigated as a further extension of our proposed framework. An encoder and two decoders share states of their hidden layers, and these components are trained using the interpolated loss function of the two decoders. One of the decoders selects the type of response, and the other estimates the word sequence of the response utterance.

Our proposed method has two advantages over conventional methods. First, our LSTM-based classifier is superior to point-wise classifiers used by methods such as support vector machine (SVM) (Vapnik, 2013) when solving sequence classification problems, such as those in which previous samples affect succeeding samples, or when the word order in each utterance sample is being analyzed. Second, our framework employs an encoder–decoder model which uses multi-task learning (Caruana, 1993, 1997) and multiple decoders, allowing it to utilize not only input utterances but also response utterances from the training corpus. This results in robust and efficient training of the framework, even when using a training corpus of limited size.

To experimentally evaluate our proposed framework, we used a self-developed conversation corpus consisting of dialogs between elderly study participants and an interviewer (Kitaoka et al., 2018). The utterances of the elderly participants are used to represent the user's input utterances to the dialog system, and the interviewer's responses serve as references for the selection of appropriate types of responses. We collected the utterances of elderly people for use in this study because one of our research goals is to develop a reminiscence therapy (Butler, 1963) dialog system for the elderly. However, it should be noted that our proposed method can be applied to any kind of conversational dialog system that is not task-oriented, i.e., not only to dialog systems for the elderly.

In previous studies (Ohta et al., 2017, 2019), we performed independent evaluations of an SVM-based classifier using acoustic features, and an LSTM-based classifier using linguistic features, respectively. In this paper, we:

- Systematically compare the performance of response selection models using acoustic features with that of models using linguistic features.
- Systematically compare the performance of response selection models based on SVM with that of models based on LSTM.
- Systematically compare the performance of response selection models based on LSTMs with and without attention.
- Evaluate selection performance when interpolating multiple response selection models.
- Discuss the results of these evaluations.

The contributions of this work are as follows. First, we propose a novel framework for dialog systems which explicitly classifies the types of responses which are required. This framework allows us to avoid, in a natural manner, the dull responses that occur in conventional end-to-end dialog systems. Second, we propose a sophisticated response-type selector based on an LSTM with multi-task learning, which outperforms conventional methods using an SVM-based point-wise classifier or a single-task-based LSTM. Third, we present the results of detailed experiments which include comparisons of the effectiveness of using acoustic features versus linguistic features, as well as using both types of features, with or without attention, using various metrics such as classification accuracy, precision, recall and F-measure, for each class as well as for the entire evaluation data.

We have organized the rest of this paper as follows. In Section 2, we discuss related studies, and in Section 3 we describe the development of our corpus of interview dialog speech. In Section 4, we explain in detail our proposed response selection method, as well as the architecture of our spoken dialog system. We describe our evaluation of the proposed method in Section 5, and then conclude the paper in Section 6.

2. Related work

In recent years, response generation methods for chat dialog systems have gradually evolved from simple, example-based approaches to sequence-to-sequence approaches. For example, several example-based approaches have been proposed (Nisimura et al., 2005; Seto et al., 2018) which retain pairs of input and response utterances, and retrieve the appropriate responses using input similarity metrics. As an enhancement of such example-based approaches, dialog systems which are composed entirely of modules based on natural language processing techniques (Higashinaka et al., 2014) have also been developed. Sequence-to-sequence generation approaches based on neural models, which maximize the probability of generating an appropriate response given the context, have also been proposed (Vinyals and Le, 2015; Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015). Although these end-to-end approaches enable us to incorporate rich context when generating a response, a problem arises, which is that these systems tend to generate dull responses, such as "I don't know", regardless of the context (Serban et al., 2016; Li et al., 2015). This is due to the high frequency of such generic responses in the training set, which can be associated with a diverse range of contexts. Therefore, we believe that it is more effective to determine the type of response needed before generating the response utterance (Ohta et al., 2019), in order to generate more appropriate responses. Using our approach, we can quickly generate responses which should be immediate, such as backchannel responses, while using more computational time when generating more fluent responses. Additionally, our approach enables us to avoid the previously mentioned dull responses in a natural manner.

The context information used for response generation are the acoustic and linguistic features of dialog speech, which have been explored in previous studies. As examples of methods which use the acoustic features of user speech to select the type of response in spoken dialogue systems, Ohsuga et al. (2005) proposed a method of determining whether or not speakers observe turn-taking in conversations, based on prosodic features of the participants' speech, such as fundamental frequency (F0), power and duration, which are extracted from their utterances. Kitaoka et al. (2005) used decision trees to determine timing for system-generated back-channel responses, as well as for turn-taking. Prosodic information, such as pitch and power gradients at the end of user utterances, and linguistic information, such as the part of speech of the last word spoken, as well as the identity of the last content word in the last utterance, are used as features of a decision tree. These studies demonstrate that acoustic information can be used for estimating the timing of turn-taking and back-channel responses. On the other hand, approaches which use distributed representations of words as linguistic features, such as word2vec (Mikolov et al., 2013), are also being used in various modules of dialog systems. In this study, we evaluate the effect of using both acoustic and linguistic features for response type selection.

LSTMs are now being used in many natural and spoken language processing applications, while encoder–decoder frameworks based on RNNs have demonstrated good performance in the field of machine translation (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2015).

When used in a machine translation framework, the encoder receives the word sequence of a speaker's utterance in chronological order and embeds this sequence in a fixed-length feature vector. The decoder then converts the feature vector into an output sentence in the target language. A similar approach is also being used in dialog systems for the task of response generation (Vinyals and Le, 2015) and response re-ranking (Inaba and Takahashi, 2016), where the RNN encoder is trained to embed the needed information into the vector used for generating the target sentences. In contrast, our proposed method applies an encoder to extract the information necessary for selecting the appropriate type of response in a spoken dialog system (Ohta et al., 2019).

Multi-task learning is another trend in the area of deep learning research, in which parameters or loss functions are shared among multiple networks. Such learning strategies have shown promising results in the areas of natural and spoken language processing (Luong et al., 2016; Kim et al., 2017). In our proposed method, we use three learning networks; an encoder to extract the needed information from the language corpus, a decoder for selecting the appropriate type of response, and a decoder for estimating the proper word sequence of the response utterance. For more effective training, these three networks share the cell states of hidden layers, as well as a loss function.

Regarding possible applications, our proposed method can be used in various spoken dialog systems, especially chat systems. The use of such systems by elderly people has attracted attention recently, for example as a method of dementia prevention. Reminiscence therapy, proposed by Butler (1963), involves prompting users to talk about their life histories, and has become widely accepted. Since speech has been found to be a simpler and more natural modality for human–computer interaction (Acartürk et al., 2015), spoken dialog systems which can conduct reminiscence therapy are expected to play an important role in delaying or mitigating dementia in the elderly. Although such a spoken dialog system has been developed using Japanese (Shitaoka et al., 2017), the types of responses which can be generated by this system are limited to just a few types, such as backchannel or empathy. A similar system was also proposed in Su et al. (2017), however the types of responses the system can generate are limited to several question-and-answer patterns. Our proposed method allows such a system to respond in a less structured, more diverse manner.

3. Conversation corpus

One of the goals of our research is to build a reminiscence therapy dialog system for elderly Japanese users, so we compiled a Japanese language conversation corpus containing dialogs between elderly people and an interviewer, in cooperation with a nursing facility, in order to train and evaluate the response selection performance of our classifier.

In each dialog, all of which were recorded in a low-noise environment, an elderly person speaks freely in response to ten questions asked by an interviewer, such as, "Have you gone anywhere recently?" After the study participant responds, the interviewer replies by expanding on the same topic, giving back-channel responses, expressing empathy, etc. A total of 3478 utterances were collected from eight speakers and manually classified. Here, each utterance is a unit of speech preceded and followed by periods of silence of 200 ms or longer. Based on the results of a preliminary investigation, we classified the interviewer responses into nine categories, as shown in Table 1. All of the utterances, of both the interviewer and elderly interviewees, were annotated with these labels to allow the supervised training of our classifier. The number of interviewer response segments of each type is also shown in Table 1. The word sequences of the utterances of both the interviewees and the interviewer were also manually transcribed, and the number of interviewer responses associated with each response type for each interviewee (A-H) is shown in Table 2. For example:

Table 1
Labels for the nine types of interviewer responses.

Label	Response type	Frequency
back	Back-channel response (neutral)	1,522
p-back	Back-channel response (positive)	497
n-back	Back-channel response (negative)	136
exp	Expand on the current topic	163
gin-up	Ginger/Liven up the conversation	142
change	Change the topic	74
smile	Smile or laugh	196
emp	Show empathy	83
non	Say nothing	665
Total		3,478

Table 2
Number of utterances linked to each type of response 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

Speaker: "My dog died a year ago".

Interviewer: "I'm sorry to hear that".(emp)

Speaker: "I left my bank card in the cash machine".

Interviewer: "Oh, no!" (emp)

These interactions would count as two empathy responses elicited from the interviewer by that particular elderly speaker.

4. Proposed method

4.1. System architecture

Fig. 1 shows the system architecture of our proposed spoken dialog system, which operates as follows. An input utterance is first recognized by the speech recognizer. The response-type selector then decides which type of response should be given, based on the recognition result, while the dialog manager tracks the state of the dialog. The response generator generates multiple responses of various types, based on the dialog state. A response which matches the response type determined by the response type selector is then synthesized by the speech synthesizer as the dialog system's next utterance.

The model architecture of the response-type selector (the bold, central block in Fig. 1) will be described in detail in the following section.

4.2. Response-type selection

Our proposed method selects the appropriate type of response to a user's input utterance using an LSTM-based encoder–decoder model, an overview of which is shown in Fig. 2. The encoder is constructed using an attention-based, bidirectional LSTM, while the decoder is constructed using a unidirectional LSTM. Each LSTM contains a hidden layer whose size is set to 200.

We then train our encoder–decoder model as follows. First, the word sequence in the user's input utterance is converted into a word

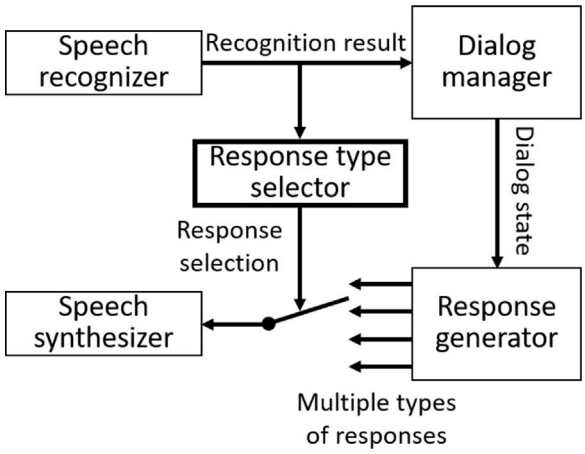

Fig. 1. Architecture of the proposed spoken dialog system.

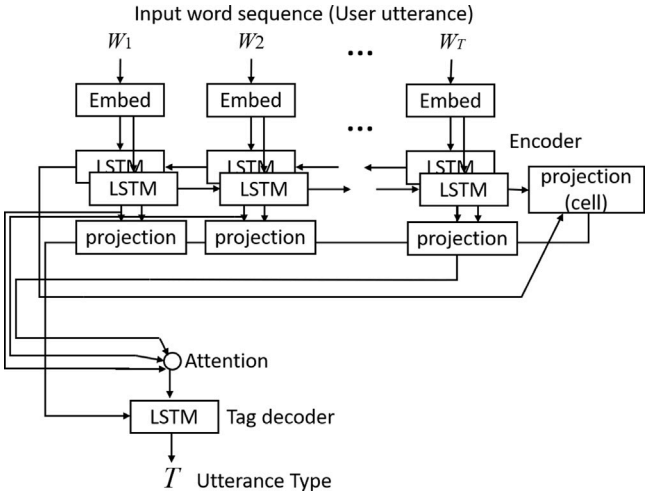

Fig. 2. Model architecture of proposed response type selector.

sequence in the form of distributed representations using word2vec,[5] which is an implementation of Mikolov's method (Mikolov et al., 2013). We trained the word2vec converter using all of the articles from the Japanese edition of Wikipedia[6] as it appeared on July 1st, 2017. The articles were tokenized using MeCab (ver. 0.996) (Kudo et al., 2004), a Japanese morphological analyzer with a custom dictionary (mecab-ipadic-NEologd ver. 0.0.5) (Sato et al., 2017) containing all of the new words extracted from the web documents. For training, we used a skip-gram model, and set the number of dimensions of the representation at 200. The distributed word sequence of an input utterance, along with the reference label for the appropriate response type, are then fed into the encoder and the decoder.

4.3. Multi-task learning

As a further extension of our proposed framework, we introduce multi-task learning to our model by including an additional decoder in our encoder–decoder model. The details of the architecture of the extended model and its learning strategy are described in the following sections.

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

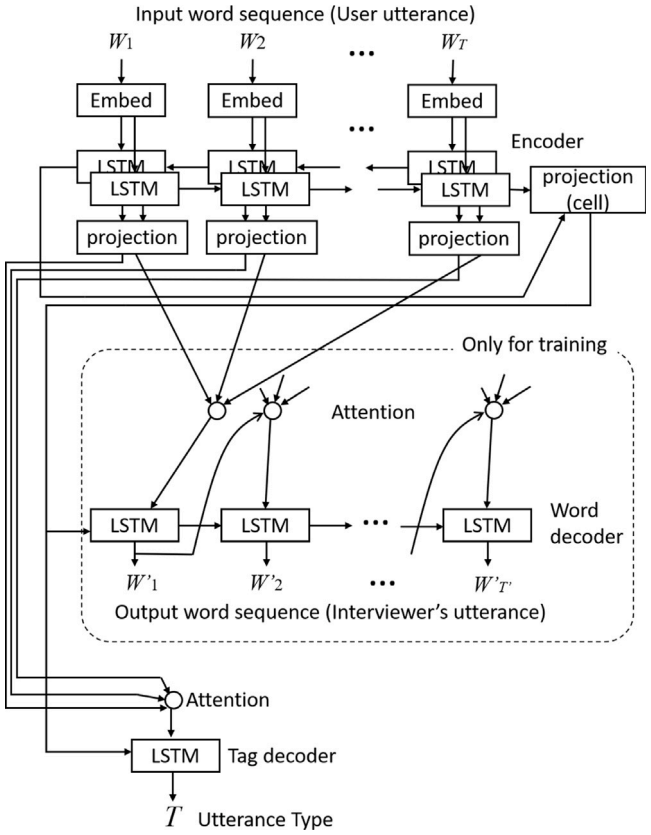

Fig. 3. Overview of the architecture of our proposed response-type selector model using multi-task learning. Each of the input words is embedded in a 200-dimensional vector, and each of the LSTMs in the encoder and decoder have 200 hidden nodes. The "projection (cell)" has a fully connected feed-forward layer which contains 400 = 200 x 2 inputs and 200 outputs, which are used to integrate the LSTM's forward and backward internal states. Likewise, each "projection" also has a fully connected feed-forward layer with 400 = 200 x 2 inputs and 200 outputs. Note that the output word sequence decoder is only used in the training stage of multi-task learning.

4.3.1. Encoder–decoder model for multi-task learning

An overview of the architecture of our encoder–decoder model for multi-task learning is shown in Fig. 3. The encoder contains an attention-based, bidirectional LSTM, while the two decoders each contain unidirectional LSTMs. One decoder (the "tag decoder") is used to select the type of response, and the other decoder (the "word decoder") is used to estimate the word sequence of the response utterance. The hidden layer of each LSTM is set to a size of 200 nodes.

During the training of this model, word sequences in the user's input utterances, as well as those in the corresponding response utterances, are converted into word sequences in the form of distributed representations, in the same manner as previously described for single-task learning in Section 4.2. The distributed word sequences of the input and response utterances, as well as the reference label indicating the type of response which is needed, are then fed into the encoder, word decoder and tag decoder, respectively. The three networks, which each contain shared hidden layers and a shared loss function, are trained using the training corpus, the details of which will be described in the next section.

During testing, only the user's input utterance is fed into our model. The type of response utterance needed is then directly estimated using the encoder and the tag decoder.

4.3.2. Loss function for multi-task learning

During the training of the encoder–decoder model, as described in the previous section, back propagation is performed using the global

loss function L, which is defined using a linear interpolation of L_{word} (loss of the word decoder) and L_{tag} (loss of the tag decoder) as shown in the following equation:

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

Here, α represents an interpolation weight between 0 and 1, and L_{word} is defined as the sum of mean square errors used to output word embeddings. The tag decoder should output one-hot vectors, so L_{tag} is a cross entropy loss.

5. Evaluation experiment

5.1. Experimental set-up

In order to evaluate our proposed method, we conducted evaluation experiments using the conversation corpus described in Section 3. Our classification results were evaluated on the basis of classification accuracy, precision, recall, and F-measure, using the nine types of response labels shown in Table 1.

We conducted three experiments as follows. In the first experiment, we compared classification performance when using only acoustic information and when using only linguistic information. As baselines, we used an SVM-based classifier with acoustic features, based on the Interspeech 2010 paralinguistic challenge feature set (Schuller et al., 2010), and the same SVM-based classifier with linguistic features based on distributed representation, and compared their performance with that of the proposed method using only acoustic features, or only linguistic features, respectively. We also evaluated naive classifiers such as a blind classifier which selects the response type randomly (based on uniform distribution or frequency distribution in the training data), and a majority guess classifier which always selects the response type with the highest frequency in the training data. Additionally, we investigated the effect of using attention in our proposed, LSTM-based model.

In the second experiment, we compared a standard classification method using single-task learning, in which $\alpha = 0$ in Eq. (1), with our proposed methods when using multi-task learning, in each of three configurations; where $\alpha = 0.7$, $\alpha = 0.8$ or $\alpha = 0.9$, respectively.

In the third experiment, we evaluated classification accuracy when the proposed method using acoustic information and the proposed method using linguistic information were combined, using the linear interpolation of the output probabilities for each response-type label.

We used 8-fold cross validation in each experiment, where the data of seven speakers was used for training data and the data of the one remaining speaker was used as evaluation data.

5.2. Experimental results

5.2.1. Experiment 1: Comparison of models using acoustic features and models using linguistic features

The results of our first experiment are shown in Table 3. When only acoustic features were used, the baseline SVM method (No. 1) achieved better performance than the proposed method using LSTM, even when attention was applied (No. 3). We believe that this is because the baseline method uses richer information, acquired from the Interspeech 2010 paralinguistic challenge feature set, than the proposed method using only MFCC-based features.

The confusion matrices of methods No. 1 and No. 3 are shown in Table 4 and Table 5, respectively. Precision, recall, and F-measure for each class are also shown at the bottom of these matrices. We can see from this comparison that by using rich acoustic features, classification performance for “p-back”, “exp” and “gin-up” were particularly improved. This suggests that it is difficult to capture the context characteristics of such low frequency labels using only MFCC features, even when a sophisticated classifier is being used. On the other hand, when only linguistic features were used, our proposed method (No. 5) achieved better performance than the baseline method (No. 4). This suggests that an LSTM-based classifier enables us to more effectively capture time sequence information, such as the order of the words in each utterance. Additionally, the classification performance of the proposed method was further improved by introducing attention (No. 6). The confusion matrices for methods No. 4 and No. 6 are shown in Table 6 and Table 7, respectively. By comparing these results, we can see that by using LSTM, classification accuracy for “p-back” and “smile” are particularly improved. This suggests that time series information is important when determining when to actively engage with users.

Table 3
Comparison of performance using acoustic features, linguistic features or Naïve classifiers.

No.	Model	Attention	Accuracy	Precision	Recall	F-Measure
1	SVM (Acoustic)	n/a	0.534	0.418	0.282	0.337
2	LSTM (Acoustic)	x	0.493	0.259	0.195	0.223
3		o	0.495	0.278	0.202	0.234
4	SVM(Linguistic)	n/a	0.497	0.307	0.193	0.237
5	LSTM (Linguistic)	x	0.536	0.328	0.282	0.303
6		o	0.539	0.458	0.269	0.339
7	Random (Uniform)		0.111	0.111	0.111	0.111
8	Random (Weighted)	n/a	0.258	0.112	0.112	0.112
9	Majority		0.438	0.091	0.111	0.100

The performance of naive classifiers, such as random classifiers based on uniform distribution (No. 7), or on the frequency of distribution in the training data (No. 8) was clearly low, according to all of the evaluation metrics. Although the majority guess classifier (No. 9) achieved better accuracy than the random classifiers, its precision, recall and F-measure performance were all low. These results highlight the advantage of using trained classifiers which consider input features when addressing the response-type classification problem.

5.2.2. Experiment 2: Comparison of models using multi-task learning and models using single-task learning

The results of our second experiment are shown in Table 8. We can see how classification performance is improved for both single-task and multi-task learning approaches by introducing attention. The best performance was achieved when both attention and multi-task learning were used, and when $\alpha = 0.8$ (No. 6). This result suggests that the proposed method using multi-task learning with an appropriate weight can more effectively utilize the information extracted from the response utterances in the training corpus than standard, single-task learning.

5.2.3. Experiment 3: Interpolation of multiple models

The results of our third experiment are shown in Fig. 4. Here, the output label is determined using the interpolated likelihood of our proposed classifiers, one of which uses acoustic features while the other uses linguistic features, which can be represented by the following equation:

$$L = \beta L_{acoustic} + (1 - \beta) L_{linguistic} \quad (2)$$

As shown in Fig. 4, the performance of our proposed method is improved by using both acoustic and linguistic features for classification. The best performance observed in our experiments for this study was a classification rate of 0.549, which was achieved when using the interpolated likelihood of two classifiers, using acoustic or linguistic features, respectively, while applying multi-task learning, with interpolation weight β set to 0.4 or 0.5.

6. Conclusions

In this study, a novel method of selecting the correct type of response by a spoken dialog system was proposed. Since the target application of our proposed method is a reminiscence therapy system for

Table 4
Confusion matrix for SVM using acoustic features (Method No. 1).

	Classification results								
True class	**back**	**p-back**	**n-back**	**exp**	**gin-up**	**change**	**smile**	**emp**	**non**
back	1235	90	0	10	14	3	11	0	159
p-back	352	71	0	3	17	2	16	0	36
n-back	102	17	0	3	3	0	2	0	9
exp	75	15	0	15	12	7	11	0	28
gin-up	40	10	0	1	66	1	23	0	1
change	27	4	0	4	5	3	9	0	22
smile	58	20	0	4	35	2	70	0	7
emp	63	10	0	0	1	0	0	3	6
non	222	23	0	7	3	8	7	0	395
Precision	0.568	0.273	0.000	0.319	0.423	0.115	0.470	1.000	0.596
Recall	0.811	0.143	0.000	0.092	0.465	0.041	0.357	0.036	0.594
F-Measure	0.668	0.188	0.000	0.143	0.443	0.060	0.406	0.069	0.595

Table 5
Confusion matrix for LSTM using acoustic features (Method No. 3).

	Classification results								
True class	**back**	**p-back**	**n-back**	**exp**	**gin-up**	**change**	**smile**	**emp**	**non**
back	1276	32	0	6	6	0	14	0	188
p-back	337	35	0	5	5	0	23	0	92
n-back	98	5	0	2	0	0	5	0	26
exp	87	6	0	7	2	2	8	0	51
gin-up	87	11	0	3	8	1	14	0	18
change	40	2	0	2	2	0	5	0	23
smile	109	13	0	2	4	0	60	0	8
emp	58	8	0	0	2	0	2	0	13
non	302	13	0	6	2	0	5	0	337
Precision	0.533	0.280	0.000	0.212	0.258	0.333	0.441	0.000	0.446
Recall	0.838	0.070	0.000	0.043	0.056	0.000	0.306	0.000	0.507
F-Measure	0.652	0.112	0.000	0.071	0.092	0.000	0.361	0.000	0.475

Table 6
Confusion matrix for SVM using linguistic features (Method No. 4).

	Classification results								
True class	**back**	**p-back**	**n-back**	**exp**	**gin-up**	**change**	**smile**	**emp**	**non**
back	1428	12	0	0	0	1	1	0	80
p-back	471	12	0	0	0	0	0	1	13
n-back	132	4	0	0	0	0	0	0	0
exp	131	7	0	1	3	0	7	1	13
gin-up	124	4	0	3	1	0	3	0	7
change	46	7	0	1	1	0	1	0	18
smile	166	5	0	6	2	0	8	0	9
emp	45	5	0	0	0	0	0	29	4
non	391	13	0	3	1	6	2	1	248
Precision	0.487	0.174	0.000	0.071	0.125	0.000	0.364	0.906	0.633
Recall	0.938	0.024	0.000	0.006	0.007	0.000	0.041	0.349	0.373
F-Measure	0.641	0.042	0.000	0.011	0.013	0.000	0.074	0.504	0.469

Table 7
Confusion matrix for LSTM using linguistic features (Method No. 6).

	Classification results								
True class	**back**	**p-back**	**n-back**	**exp**	**gin-up**	**change**	**smile**	**emp**	**non**
back	1249	126	0	1	1	0	7	0	138
p-back	290	149	0	0	0	0	17	2	39
n-back	91	40	0	0	1	0	1	0	3
exp	72	41	0	3	0	0	21	0	26
gin-up	70	39	0	0	0	0	23	0	10
change	35	12	0	0	0	0	9	0	18
smile	81	45	0	1	0	0	56	0	13
emp	21	35	0	0	2	0	1	15	9
non	223	31	0	0	0	0	5	0	406
Precision	0.564	0.351	0.333	0.167	0.318	0.500	0.420	0.911	0.561
Recall	0.846	0.235	0.000	0.025	0.049	0.014	0.189	0.494	0.573
F-Measure	0.677	0.282	0.000	0.043	0.085	0.027	0.261	0.641	0.567

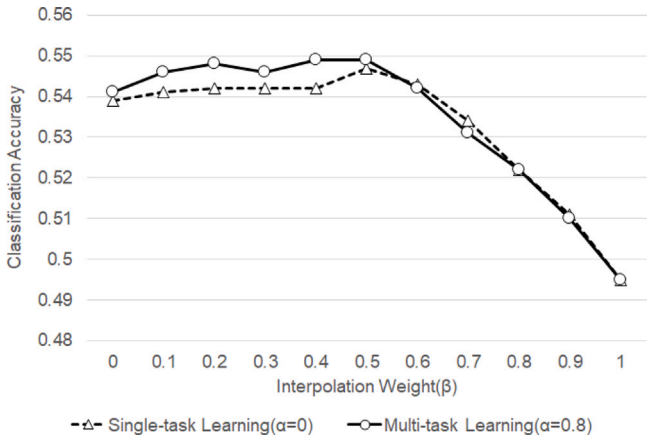

Fig. 4. Results of Experiment 3: Interpolation of acoustic and linguistic models.

Table 8
Comparison of method No. 6 models using single-task learning versus multi-task learning.

No.	Model	α	Attention	Accuracy	Precision	Recall	F-Measure
1	LSTM (Linguistic)	0[a]	x	0.536	0.328	0.282	0.303
2			o	0.539	0.458	0.269	0.339
3		0.7	x	0.532	0.342	0.268	0.300
4			o	0.536	0.428	0.273	0.333
5		0.8	x	0.535	0.401	0.263	0.318
6			o	0.541	0.434	0.287	0.346
7		0.9	x	0.536	0.361	0.280	0.315
8			o	0.537	0.373	0.285	0.323

[a]= Single-task Learning.

the elderly, we used transcription data from conversations with elderly people being questioned by an interviewer to train our response-type selector.

The contributions of this work are as follows:

- We proposed an LSTM-based response-type selector which handles linguistic information as a time series, allowing it to achieved better performance than a conventional, SVM-based point-wise classifier.
- We also proposed the use of multi-task learning with multiple decoders in our response-type selector, utilizing not only input utterances but also response utterances in the training corpus, allowing it to achieved better performance than a standard LSTM-based classifier employing single-task learning.
- We presented the results of detailed performance comparison experiments, which included comparisons between the use of acoustic and linguistic features, with and without attention, using various performance metrics such as classification accuracy, precision, recall and F-measure for each class, as well as for the entire set of evaluation data.
- We demonstrated that the best performance was achieved when the likelihood results of our model using acoustic features and the results of our model using linguistic features were combined using linear interpolation.
- Our proposed response selection framework for dialog systems, which explicitly classifies responses by type of response, enables dialog systems to avoid dull responses, such as "I don't know" and "I'm OK", in a natural manner. This improves the quality of the conversation, as these dull responses are frequently generated by conventional end-to-end dialog systems (Vinyals and Le, 2015; Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015).
- Our proposed dialog system framework and response-type selector can be applied not only to dialog systems for the elderly (Shitaoka et al., 2017; Su et al., 2017), but to any kind of non-task oriented dialog system.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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, 20H05562) 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 help with the experiments cited in this research.

References

Acartürk, C., Freitas, J., Fal, M., Dias, M.S., 2015. Elderly speech-gaze interaction: State of the art and challenges for interaction design. In: Universal Access in Human-Computer Interaction: Access to Today's Technologies. In: Series Lecture Notes in Computer Science, vol. 9175, pp. 3–12.

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: Proceedings of the International Conference on Learning Representations, ICLR.

Butler, R.N., 1963. The life review: An interpretation of reminiscence in the aged. Psychiatry 26 (1), 65–76.

Caruana, R., 1993. Multitask learning: A knowledge-based source of inductive bias. In: Proceedings of the International Conference on Machine Learning. ICML, Morgan Kaufmann, pp. 41–48.

Caruana, R., 1997. Multitask learning. Mach. Learn. 28 (1), 41–75.

Fratiglioni, L., Wang, H., Ericsson, K., Maytan, M., Winblad, B., 2000. Influence of social network on occurrence of dementia: A community-based longitudinal study. Lancet 1315–1319.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., Matsuo, Y., 2014. Towards an open-domain conversational system fully based on natural language processing. In: The International Conference on Computational Linguistics, COLING, pp. 928–939.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Inaba, M., Takahashi, K., 2016. 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.

Inoue, K., Milhorat, P., Lala, D., Zhao, T., Kawahara, T., 2016. Talking with ERICA, an autonomous android. In: The Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, pp. 212–215.

Kalchbrenner, N., Blunsom, P., 2013. Recurrent continuous translation models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, Vol. 3, No. 39, pp. 1700–1709.

Kim, S., Hori, T., Watanabe, S., 2017. 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.

Kitaoka, N., Segawa, S., Nishimura, R., Takeda, K., 2018. Recognizing emotions from speech using a physical model. Acoust. Sci. Technol. 39 (2), 167–170.

Kitaoka, N., Takeuchi, M., Nishimura, R., Nakagawa, S., 2005. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. Trans. Jpn. Soc. Artif. Intell. 20 (3), 220–228.

Koiso, H., Tsuchiya, T., Watanabe, R., Yokomori, D., Aizawa, M., Den, Y., 2017. 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.

Kudo, T., Yamamoto, K., Matsumoto, Y., 2004. Applying conditional random fields to Japanese morphological analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 230–237.

Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B., 2015. A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055.

Luong, M.-T., Le, Q.V., Sutskever, I., Vinyals, O., Kaiser, L., 2016. Multi-task sequence to sequence learning. In: International Conference on Learning Representations, ICLR.

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

Nisimura, R., Lee, A., Yamada, M., Shikano, K., 2005. Operating a public spoken guidance system in real environment. In: The Annual Conference of the International Speech Communication Association, Interspeech.

Ohsuga, T., Nishida, M., Horiuchi, Y., Ichikawa, A., 2005. Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue. In: The Annual Conference of the International Speech Communication Association, Interspeech, pp. 33–36.

Ohta, K., Marumotoa, R., Nishimura, R., Kitaoka, N., 2017. 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.

Ohta, K., Nishimura, R., Kitaoka, N., 2019. Type of response selection utilizing user utterance word sequence, LSTM and multi-task learning for chat-like spoken dialog systems. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC.

Ritter, A., Cherry, C., Dolan, W.B., 2011. Data-driven response generation in social media. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP, Association for Computational Linguistics, pp. 583–593.

Robinson, A.J., Fallside, F., 1988. Static and dynamic error propagation networks with application to speech coding. In: Neural Information Processing Systems, NIPS, pp. 632–641.

Roy, N., Baltus, G., Fox, D., Gemperle, F., Goetz, J., Hirsch, T., Margaritis, D., Montemerlo, M., Pineau, J., Schulte, J., et al., 2000. Towards personal service robots for the elderly. In: Workshop on Interactive Robots and Entertainment, WIRE, Vol. 25, p. 184.

Saczynski, J., Pfeifer, L., Masaki, K., Korf, E., Laurin, D., White, L., Launer, L., 2006. The effect of social engagement on incident dementia: the Honolulu-Asia aging study. Am. J. Epidemiol. 433–440.

Sato, T., Hashimoto, T., Okumura, M., 2017. Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval. In: Proceedings of the Annual Meeting of the Association for Natural Language Processing. The Association for Natural Language Processing (in Japanese).

Schuller, B.W., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C.A., Narayanan, S.S., et al., 2010. The INTERSPEECH 2010 paralinguistic challenge. In: The Annual Conference of the International Speech Communication Association, Interspeech, pp. 2795–2798.

Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J., 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI Conference on Artificial Intelligence.

Seto, E., Nishimura, R., Kitaoka, N., 2018. Customization of an example-based dialog system with user data and distributed word representations. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC, pp. 1718–1724.

Shang, L., Lu, Z., Li, H., 2015. Neural responding machine for short-text conversation. In: Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, ACL-IJCNLP.

Shitaoka, K., Tokuhisa, R., Yoshimura, T., Hoshino, H., Watanabe, N., 2017. Active listening system for a conversation robot. J. Nat. Lang. Process. 24 (1), 3–47 (in Japanese).

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., Dolan, B., 2015. A neural network approach to context-sensitive generation of conversational responses. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, pp. 196–205.

Su, M., Wu, C., Huang, K., Hong, Q., Wang, H., 2017. A chatbot using LSTM-based multi-layer embedding for elderly care. In: Proceedings of the International Conference on Orange Technologies, ICOT, pp. 70–74.

Vapnik, V., 2013. The Nature of Statistical Learning Theory. Springer Science & Business Media.

Vinyals, O., Le, Q., 2015. A neural conversational model. In: International Conference on Machine Learning (ICML) Deep Learning Workshop.