Dialogue-Act Prediction of Future Responses based on Conversation History

Koji Tanaka[†], **Junya Takayama**[†], **Yuki Arase**[†‡]
[†]Graduate School of Information Science and Technology, Osaka University
[‡]Artificial Intelligence Research Center (AIRC), AIST
{tanaka.koji, takayama.junya, arase}@ist.osaka-u.ac.jp

Abstract

Sequence-to-sequence models are a common approach to develop a chatbot. They can train a conversational model in an end-to-end manner. One significant drawback of such a neural network based approach is that the response generation process is a black-box, and how a specific response is generated is unclear. To tackle this problem, an interpretable response generation mechanism is desired. As a step toward this direction, we focus on dialogue-acts (DAs) that may provide insight to understand the response generation process. In particular, we propose a method to predict a DA of the next response based on the history of previous utterances and their DAs. Experiments using a Switch Board Dialogue Act corpus show that compared to the baseline considering only a single utterance, our model achieves 10.8% higher F1-score and 3.0% higher accuracy on DA prediction.

1 Introduction

Dialogue systems adopt neural networks (NNs) (Vinyals and Le, 2015) because they allow a model to be developed in an end-to-end manner without manually designed rules and patterns for response generation. However, in a NN-based approach, the response generation process is hidden in the model, which makes it difficult to understand why the model generates a specific response. This is a significant problem in commercially produced chatbots because the model outputs cannot be controlled. To tackle this problem, Zhao et al. (2018) argued that interpretable response generation models are important. As the first step toward this direction, we focus on dialogue-acts (DAs) as clues to understand the response generation process. We speculate that the predicted DAs indicates which types of response the model tries to generate.

	Utterance (DA)
1	Oh, I've only, I've only skied in Utah once. (Statement)
2	Oh, really? (Question)
3	I only skied once my whole life. (Statement)
4	Uh-huh. (Uninterpretable)
5	But, do you do a lot of skiing there? (Question)

Table 1: Example of utterances and their DAs (in parenthesis) sampled from the SwDA corpus.

Specifically, we propose a method to predict the DA of the next response. This problem was proposed by Reithinger et al. (1996). A conversation consists of a sequence of utterances and responses, where the next response depends on the history of utterances and responses. Table 1 shows an example of a conversation with utterances and their DAs sampled from the Switch Board Dialogue Act (SwDA) corpus. The DA of the last response, "But, do you do a lot of skiing there? (Question)" is not predictable using the previous utterance of "Uh-huh." nor using its DA of "Uninterpretable". To correctly predict the DA, we need to refer to the entire sequence starting from first utterance of "Oh, I've only skied in Utah once." when the speaker is talking about skiing experience.

Our model considers the conversation history for DA prediction. It independently encodes sequences of text surfaces and DAs of utterances using a recurrent neural network (RNN). Then it predicts the most likely DA of the next response based on the outputs of RNNs. Cervone et al. (2018) showed that a DA is useful to improve the coherency of response. The predicted DAs can be used to generate a future response, which adds controllability and interpretability into a neural di-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 197–202
Florence, Italy, July 28 - August 2, 2019.

alogue system.

We used a SwDA corpus for the evaluation, in which telephone conversations are transcribed and annotated with DAs. The macro Precision, Recall, F1, and overall Accuracy measure the performance compared the baseline. The results show that our model, which considers the history of utterances and their DAs, outperforms the baseline, which only considers the input utterance by 10.8% F1 and 3.0% Accuracy.

2 Related Works

Previous studies on DA prediction aimed to predict the current DA from the corresponding utterance text. Kalchbrenner and Blunsom (2013) proposed a method using Convolutional Neural Network (CNN) to obtain a representation capturing the local features of utterance and RNN to obtain the context representation of the utterance. Experiments using the SwDA corpus showed that their method outperformed previous methods for DA prediction using non-neural machine learning models. Khanpour et al. (2016) proposed a method based on multi-layer RNN that uses an utterance as an input. Their method achieved an 80.1% prediction accuracy of the SwDA corpus, and is the current state-of-the-art method.

Unlike these previous studies, we focus on DA prediction of the next (*i.e.*, unseen) response. Reithinger et al. (1996) proposed a statistical method using a Markov chain. Using their original corpus, their method achieved of the 76.1% top_3 accuracy. We tackled the problem of DA prediction of the next utterance considering the history of utterances and previous DAs using a NN. We anticipate that the predicted DA is useful for understanding the response generation process and improving the quality of the response generation.

3 Proposed Model

Figure 1 illustrates the design of our model, which consists of three encoders with different purposes. The Utterance Encoder encodes the utterance text into a vector, which is then inputted into the Context Encoder that handles the history of utterance texts. The Dialogue-act (DA) Encoder encodes and handles the sequence of DAs. Finally, outputs of the Context and DA Encoders are concatenated and input to a classifier that predicts the DA of the next response. Note that our model does not peek into the text of next response to predict the DA.

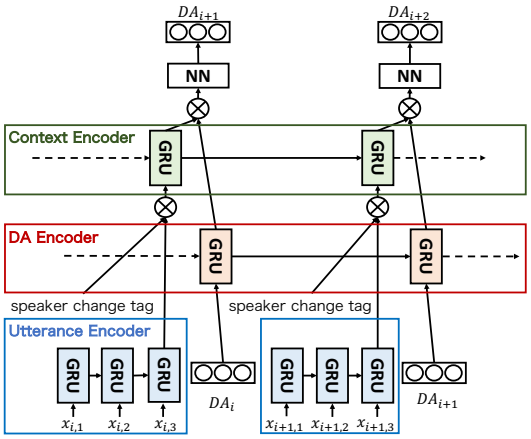

Figure 1: Design of our model consisting of three encoders that encode 1) text surfaces, 2) DAs, and 3) utterance history. ($\otimes$ concatenates vectors)

Consequently, the predicted DA is used to generate the response text in the future.

3.1 Utterance Encoder & Context Encoder

The Utterance Encoder vectorizes an input utterance. It is an RNN that takes each word in the utterance in a forward direction by applying padding in order to realize a uniform input size. Then, the Context Encoder, which is another RNN, takes the final output of the Utterance Encoder to generate a context vector that handles the history of utterances. While our model takes a single sentence as an input to the Utterance Encoder, the speakers do not necessarily change at every single sentence in a natural conversation. Hence, our model allows cases where the same speaker continuously speaks. Specifically, a speaker change tag, which is inputted into the Context Encoder, is used to indicate when the speaker changes.

3.2 Dialogue-act (DA) Encoder

The DA Encoder plays the role of handling the history of DAs. A DA is represented as a one-hot vector and encoded by RNN. During the training, we use teacher forcing to avoid error propagation. That is, the gold DA of the current utterance is inputted into the model instead of the predicted one.

3.3 Dialogue-act Prediction

Finally, the classifier determines the DA of the next response. It is a single fully-connected layer culminating in the soft-max layer. Given a concatenation of outputs from the Context Encoder

Tag	# of tags in the corpus
Statement	576,005
Uninterpretable	93,238
Understanding	241,008
Agreement	55,375
Directive	3,685
Greeting	6,618
Question	54,498
Apology	11,446
Other	19,882

Table 2: Distribution of DA tags in the preprocessed SwDA corpus.

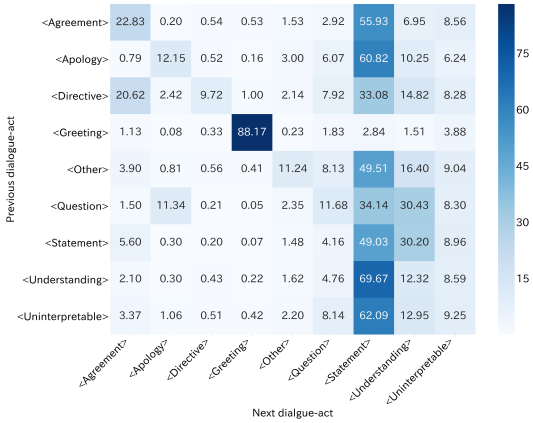

Figure 2: Conditional Probabilities of DA transitions. "Greeting" has a clear pattern, which is followed by a "Greeting". Other DAs tend to be followed by a "Statement".

and DA Encoder, the classifier conducts a multi-class classification and identifies the most likely DA of the next response.

4 Experiment

4.1 Switch Board Dialogue Act Corpus (SwDA)

We evaluate the accuracy of our model to predict the DA of the next response using the SwDA corpus, which transcribes telephone conversation and annotates DAs of utterances. The SwDA corpus conforms to the damsl tag schema.[1] We assembled the tag sets referring to easy damsl (Isomura et al., 2009) into 9 tags (Table 2) in order to consolidate tags with a significantly low frequency. The SwDA corpus provides transcriptions of 1,155 conversations with 219,297 utterances. One conversation contains 189 utterances on average. Because the average length of utterance sequences is large, we use a sliding window with a size of 5 to cut a sequence into several conversations.

The number of conversations increases to 212,367 with 1,061,835 utterances. Table 2 shows the distribution of DAs in the processed corpus. We randomly divide the conversations in the corpus into 80%, 10%, and 10% for training, development, and testing, respectively.

4.2 Model Settings

We apply a Gated Recurrent Unit (GRU) (Cho et al., 2014) to each RNN in our model. We set the dimensions of word embedding to 300 and those of the DA embedding to 100. The dimensions of the GRU hidden unit of the Utterance Encoder are set to 512, while those the Context Encoder are set to 513 (one element is for the speaker change tag) and those of the DA Encoder are set to 128. Hence, the dimensions of an input into the classifier are 641. The dimensions of the hidden unit of the fully-connected layer are set to 100. The cross-entropy error is used for the loss function, and the Adam (Kingma and Ba, 2014) optimizer with a learning rate of $5e-5$ is used for optimization. The number of epochs is set to 30. We use the model with the lowest development loss for testing.

We use teacher forcing for training and similar setting for testing by inputting the gold DA of the previous time step into the DA Encoder. This means that the evaluation results here show the performance when the predictions of the previous time steps are all correct.

As Table 2 shows, the numbers of DAs are highly diverse. To avoid frequent tags dominating the results, we measure the macro averages of precision, recall, and F1-score of each DA. We also measure the overall accuracy.

4.3 Baselines

To investigate the effects of each encoder in our model, we compare our model to the baseline (Table 3). The second and third rows are simple methods. Max-Probability is another non-neural baseline that outputs the DA with highest conditional probability from the input DA. Figure 2 shows the conditional probability of DA transitions computed in our training set. "Greeting" has a no-

[1] https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

	Utterance Encoder	Context Encoder	DA Encoder	Precision	Recall	F1-score	Accuracy
Proposed model	✓	✓	✓	52.7	**32.5**	**32.4**	**69.7**
Max-Probability				15.9	19.6	16.9	54.8
Utterance-only	✓			24.4	21.6	21.6	66.7
Utterance-seq	✓	✓		30.9	25.1	23.8	68.5
DA+Utterance-seq	✓	✓	✓(single-turn)	**53.1**	29.9	30.3	68.7
DAseq-only			✓	44.7	28.7	27.9	67.1
DAseq+Utterance	✓		✓	45.8	29.0	29.3	68.2

Table 3: Macro averages of precision, recall, F1-score, and overall accuracy

ticeable pattern in which it is followed by "Greeting". This is natural considering human communication. On the other hand, other DAs are mostly followed by "Statement". This implies that only a previous DA is insufficient to predict the next DA.

The rest of Table 3 shows NN-based baselines. The Utterance-only is the model that only has the Utterance Encoder (*i.e.*, it predicts the DA of the next response based only on the input utterance). The Utterance-seq, which has the Utterance Encoder and Context Encoder, predicts the DA based on a sequence of utterances. On the other hand, the DAseq-only has only the DA Encoder and predicts the DA of the next response based on the sequence of previous DAs. The DAseq+Utterance has the Utterance Encoder and DA Encoder, which considers the sequence of DAs and the single utterance. The DA+Utterance-seq contains the Utterance Encoder and Context Encoder. It considers only the DA of the input utterance and not the sequence.

4.4 Results

Table 3 shows the macro averages of the precision, recall, and F1-score, as well as overall accuracies for each model. For all evaluation model, our model exhibits the best performances; recall, F1, and accuracy 32.5%, 32.4%, and 69.7%, respectively. As discussed in Section 1, Khanpour et al. (2016) achieved 80.1% prediction accuracy for the same SwDA corpus. Their method predicts the DAs of the current utterance given in text. Although their accuracy is not directly comparable to ours due to differences in data splits, 80.1% can be regarded as the upper-bounds of our task. Our method achieves 87.0% of this upper-bound. Below we investigate which encoders contribute to prediction.

Max-Probability performs quite poorly rather than other neural network based model. This may be because of the imbalanced transition patterns of DAs as shown in Figure 2, which shows that

Tag	# of tags in the corpus	Proposed model	Utterance-seq
Statement	576,005	80.8	80.4
Uninterpretable	93,238	4.7	2.6
Understanding	241,008	69.5	67.6
Agreement	55,375	23.1	15.3
Directive	3,685	2.7	0.0
Greeting	6,618	81.3	46.7
Question	54,498	8.1	2.0
Apology	11,446	22.7	11.3
Other	19,882	3.6	0.0

Table 4: F1-score per DA

the next DA prediction requires more features to achieve precise prediction.

Utterance-seq achieves 1.8% higher accuracy than Utterance-only, demonstrating the effectiveness of considering the history of utterances rather than a single utterance.

The DA+Utterance-seq outperforms Utterance-seq on F1 by 6.5%. This result implies that a previous DA is an effective hint for DA prediction of next responses. In addition, the sequence of DAs is also effective for the next DA prediction, which is shown by the superior performance of the DAseq-only to the Utterance-seq. Specifically, DAseq-only performs 4.1% higher macro-F1 than Utterance-seq, but has 1.4% lower accuracy than Utterance-seq. Similarly, DAseq+Utterance achieves 5.5% higher F1 than Utterance-seq. Overall, DAs of either single-turn or a sequence largely boost precision, recall, and F1. On the other hand, a sequence of utterances contributes to accuracy. These results imply that the sequence of DAs is effective to predict infrequent DAs and the sequence of utterances is effective to predict common DAs. This may be because the DA Encoder is more robust against the data sparseness issue due to its much smaller vocabulary size compared to that of the Utterance Encoder. These analyses show that our model achieves the best performance considering both sequence of utterances and DAs.

Table 4 shows the F1-scores per DA of the

	Utterance (DA)	Gold DA	Proposed model	Utterance-seq
1	What are they , (Uninterpretable)	Statement	Statement	Statement
2	the , (Statement)	Statement	Statement	Statement
3	I know , (Statement)	Statement	Statement	Statement
4	a Rabbit 's one , diesel (Statement)	Agreement	Understanding	Understanding
5	Uh-huh , (Agreement)	**Agreement**	**Agreement**	**Statement**
1	I hope so too . (Statement)	Statement	Statement	Statement
2	You know . Right now there 's a lot on the market for sale because of people having lost Yes . (Statement)	Understanding	Understanding	Understanding
3	Yes . (Understanding)	Statement	Statement	Statement
4	and everything (Statement)	Statement	Statement	Statement
5	so that 's , you know , that keeps prices down (Statement)	**Understanding**	**Understanding**	**Understanding**
1	It does n't seem like , (Statement)	Statement	Statement	Statement
2	but I guess when you think of it everybody has some sort of aerosol in their home (Statement)	Understanding	Understanding	Understanding
3	Yeah . (Understanding)	Statement	Statement	Statement
4	You know , (Statement)	Statement	Statement	Statement
5	and it 's kind of dangerous . (Statement)	**Agreement**	**Understanding**	**Understanding**

Table 5: Examples of predicted DAs by the proposed model and Utterance-seq. DA in a parenthesis shows that of the input utterance, while "Gold DA" shows the DAs of the next responses. The column of "Proposed Model" column shows the predicted DAs of the next responses by the proposed model, and the column of "Utterance-seq" shows the predicted DAs of the next responses by Utterance-seq.

proposed model and Utterance-seq. The proposed model outperforms Utterance-seq on all the DAs. In particular, infrequent tags of "Agreement", "Greeting", "Question" and "Apology" show significant improvements between 6.1% and 34.6%. Furthermore, the proposed model correctly predicts "Directive" and "Other" even though Utterance-seq does not predict any of these correctly.

4.5 Examples of Predicted DAs

Table 5 shows examples of the predicted DAs by the proposed model and Utterance-seq. The first example shows that the proposed model correctly predicts "Agreement", which only has 5.2% occurrence in the training set, whereas Utterance-seq most frequently predicts it as "Statement".

The second and third examples demonstrate the difficulty of DA prediction of the next response. The input utterances of these examples have the same DA sequences, but the DAs of the final responses differ ("Understanding" and "Agreement"). While both the proposed model and Utterance-seq correctly predict the final DA of the second example, both fail in the third example. The third conversation is about an aerosol, and the response to the final utterance of "and it's kind of dangerous." depends on if one of the speakers understands the danger of the aerosol. To correctly predict DAs in such a case, a much longer conversation sequence and/or personalize the prediction model must be considered based on profiles or knowledge of speakers. This is the direction of our future work.

5 Conclusion

We propose a method to predict a DA of the next response considering the sequences of utterances and DAs. The evaluation results using the SwDA corpus show that the proposed model achieves 69.7% accuracy and 32.4% macro-F1. Additionally, the results show that the sequence of DAs significantly helps the prediction of infrequent DAs.

In the future, we plan to develop a response generation model using the predicted DAs.

Acknowledgements

This project is funded by Microsoft Research Asia, Microsoft Japan Co., Ltd., and JSPS KAKENHI Grant Number JP18K11435.

References

Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. 2018. Coherence models for dialogue. In *Proceedings of the Interspeech*, pages 1011–1015.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the EMNLP*, pages 1724–1734.

Naoki Isomura, Fujio Toriumi, and Kenichiro Ishii. 2009. Evaluation method of non-task-oriented dialogue system by HMM. *The IEICE transactions on information and systems (Japanese edition)*, 92(4):542–551.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on CVSC*, pages 119–126.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of the COLING*, pages 2012–2021.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the ICLR*.

Norbert Reithinger, Ralf Engel, Michael Kipp, and Martin Klesen. 1996. Predicting dialogue acts for a speech-to-speech translation system. In *Proceeding of the ICSLP*, pages 654–657 vol.2.

Oriol Vinyals and Quoc V Le. 2015. A Neural Conversational Model. In *Proceedings of the ICML*.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the ACL*, pages 1098–1107.