Utterance Position-Aware Dialogue Act Recognition

Yuki Yano[1] **Akihiro Tamura**[2] **Takashi Ninomiya**[3] **Hiroaki Obayashi**[4]
[1]Nara Institute of Science and Technology [2]Doshisha University
[3]Ehime University [4]transcosmos inc.
yano.yuki.yt0@is.naist.jp
aktamura@mail.doshisha.ac.jp ninomiya@cs.ehime-u.ac.jp
oobayashi.hiroaki@trans-cosmos.co.jp

Abstract

This study proposes an utterance position-aware approach for a neural network-based dialogue act recognition (DAR) model, which incorporates positional encoding for utterance's absolute or relative position. The proposed approach is inspired by the observation that some dialogue acts have tendencies of occurrence positions. The evaluations on the Switchboard corpus show that the proposed positional encoding of utterances statistically significantly improves the performance of DAR.

1 Introduction

The recognition of dialogue acts (DAs), which represent the intention or function of each utterance in a dialogue, is useful for various dialogue applications, such as dialogue systems and dialogue summarization. Recently, neural network- (NN-)based approaches have become dominant in dialogue act recognition (DAR) because NN-based models obtained higher performance than other approaches. Most existing NN-based DAR models treat DAR as a sequence labeling problem, where utterances are first encoded with a hierarchical recurrent neural network (RNN), and subsequently, the sequence of DAR labels is identified from the encoded representations using the Conditional Random Field (CRF) (Kumar et al., 2018; Chen et al., 2018; Li et al., 2019; Raheja and Tetreault, 2019). These CRF-based models can capture the local dependencies of DA sequences; however, they cannot model global dependencies due to the first-order Markov assumption. To alleviate this problem, Colombo et al. (2020) proposed a sequence-to-sequence (seq2seq) architecture for DAR, which could learn global dependencies, and achieve the state-of-the-art performance. We focus on this work and employ it as the basis of our proposed model.

Existing NN-based DAR models focus on the context of a dialogue and dependencies between DAs but do not focus on the position of an utterance. However, we have found that some DAs have tendencies of occurrence positions. For example, the "Open-Question" DA in the Switchboard corpus tends to appear at the beginning of a telephonic conversation.

Inspired by this observation, this study proposes an utterance position-aware approach for a NN-based DAR model, which explicitly encodes the position of an utterance by positional encoding and generates DA label sequences based on the hidden vectors augmented with the positional encoding of utterance. In particular, we implement two types of positional encodings of utterances: (1) sinusoidal positional encoding used in Transformer neural machine translation (Vaswani et al., 2017), which represents the absolute position of each utterance; (2) length-ratio positional encoding, which is a positional encoding based on the ratio of the position of an utterance to the dialogue length (i.e., the total number of utterances in the dialogue). The second encoding aims to encode relative positional information to alleviate the variation of dialogue length.

The evaluations on the Switchboard corpus (Stolcke et al., 2000), which is one of the most popular benchmark datasets in DAR, show that the performance of DAR is statistically significantly improved by incorporating the proposed positional encoding of utterances (up to +0.31 precision). Our analysis demonstrates that the proposed model statistically significantly improves the performance of long dialogues.

2 Proposed DAR Model

Figure 1 shows an overview of the proposed model. The proposed model incorporates positional encoding of utterances (Section 2.2) into the baseline seq2seq DAR model (Section 2.1).

Proceedings of Recent Advances in Natural Language Processing, pages 1567–1574
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_176

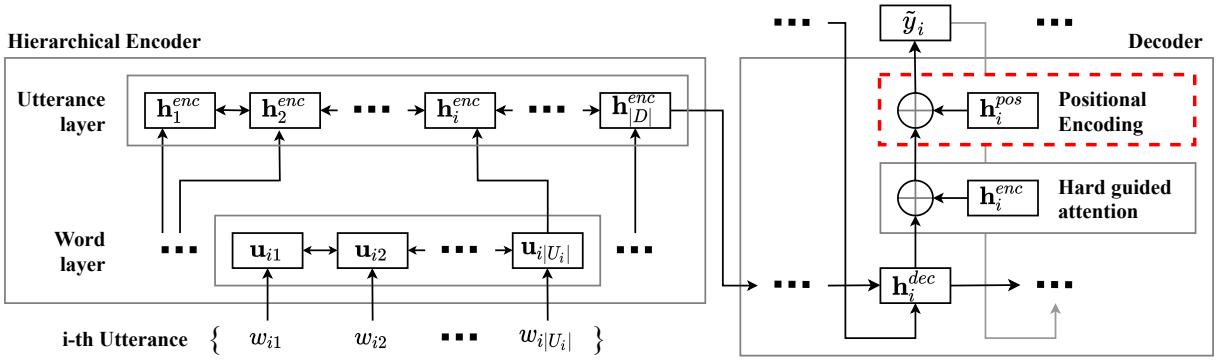


Figure 1: Overview of the proposed model. The red dashed rectangle represents the proposed positional encoding of utterances.

2.1 Baseline DAR Model

Our DAR model takes a dialogue document $D = (U_1, \cdots, U_{|D|})$, which is the sequence of utterances U_i as an input and predicts the sequence of DA labels $\tilde{Y} = (\tilde{y}_1, \cdots, \tilde{y}_{|D|})$ from D, where y_i is the DA label of U_i and each utterance is the sequence of words (i.e., $U_i = (w_{i1}, \cdots, w_{i|U_i|})$).

The baseline model is a seq2seq model that consists of a hierarchical encoder and a decoder with guided attention proposed by Colombo et al. (2020). The hierarchical encoder hierarchically encodes a dialogue document using two types of RNN layers, a word layer and an utterance layer. Particularly, the word layer first generates the sequence of intermediate word vectors $(\mathbf{u}_{i1}, \cdots, \mathbf{u}_{i|U_i|})$ of each utterance U_i from the sequence of word embedding vectors $(\mathbf{w}_{i1}, \cdots, \mathbf{w}_{i|U_i|})$. Subsequently, the utterance layer generates the sequence of intermediate utterance vectors $(\mathbf{h}^{enc}_1, \cdots, \mathbf{h}^{enc}_{|D|})$ from the outputs of the word layer. We use bi-directional gated recurrent unit (Cho et al., 2014) as RNN. The computations in the encoder are as follows:

$$\mathbf{u}_{ij} = \mathrm{BiGRU}^{word}(\mathbf{u}_{ij-1}, \mathbf{w}_{ij}), \quad (1)$$

$$\mathbf{h}^{enc}_i = \mathrm{BiGRU}^{utt}(\mathbf{h}^{enc}_{i-1}, \mathbf{u}_{i|U_i|}). \quad (2)$$

The decoder autoregressively generates the sequence of DA labels after receiving the previous hidden state and the previous output as inputs. In each timestep i, the previous label $\tilde{y}_{i-1}$ is first converted into an embedding vector $f_{embed}(\tilde{y}_{i-1})$ in the same way as word embedding, following which the hidden vector $\mathbf{h}^{dec}_i$ is generated as follows:

$$\mathbf{e}_i = \begin{cases} f_{embed}(\langle SOS \rangle) & (i = 1) \\ f_{embed}(\tilde{y}_{i-1}) & (otherwise), \end{cases} \quad (3)$$

$$\mathbf{h}^{dec}_i = \begin{cases} \mathrm{GRU}(\mathbf{h}^{enc}_{|D|}, \mathbf{e}_1) & (i = 1) \\ \mathrm{GRU}(\mathbf{h}^{dec}_{i-1}, \mathbf{e}_i) & (otherwise), \end{cases} \quad (4)$$

where $\mathbf{h}^{enc}_{|D|}$ denotes the encoder's final hidden state and $\langle SOS \rangle$ denotes the special label. Finally, the i-th DA label is predicted by applying hard guided attention (Colombo et al., 2020), which attends only to the corresponding encoder's hidden state, as follows:

$$\mathbf{z}_i = [\mathbf{h}^{dec}_i; \mathbf{h}^{enc}_i], \quad (5)$$

$$\tilde{y}_i = \mathrm{LogSoftmax}(\mathrm{ReLU}(\mathbf{W}\mathbf{z}_i)), \quad (6)$$

where [;] indicates a concatenation operation, and $\mathbf{W}$ denotes a parameter matrix.

2.2 Positional Encodings of Utterances

The proposed model incorporates positional encoding of utterances into the baseline DAR model described in Section 2.1 to explicitly consider the position of an utterance in the inference of its DA label. Specifically, the proposed model predicts the i-th DA label from the following $\mathbf{z}_i$, which is the concatenation of the original one and the positional encoding of the i-th utterance $\mathbf{h}^{pos}_i$ rather than Equation (5).

$$\mathbf{z}_i = [\mathbf{h}^{dec}_i; \mathbf{h}^{enc}_i; \mathbf{h}^{pos}_i]. \quad (7)$$

In this work, we implement two types of positional encodings of utterances as $\mathbf{h}^{pos}_i$ in Equation (7): (1) absolute positional encoding ($\mathrm{PE}_{\mathrm{abs}}$), which encodes the absolute positional information of each utterance, and (2) relative positional encoding ($\mathrm{PE}_{\mathrm{rel}}$), which encodes the relative positional information of each utterance considering the dialogue length.

Absolute Positional Encoding ($\mathrm{PE}_{\mathrm{abs}}$)

We incorporate the absolute positional information of utterances by applying sinusoidal positional encoding used in the Transformer (Vaswani et al.,

	Train	Validation	Test
# of conversations	1,003	112	19
# of utterances	174,257	18,135	4,078

Table 1: Statistics of our experiment data.

Hyperparameters	Search Range	Optimized Value		
		s2s	s2s + PE_{abs}	s2s + PE_{rel}
Word embedding size	16 - 512	64	512	512
Encoder GRU size	16 - 512	300	300	300
Decoder GRU size	16 - 512	64	16	128
PE size	16 - 512	-	32	32
Clip gradient value	1.0 - 5.0	3.0	3.5	4.5
Dropout	0.1 - 0.7	0.4	0.5	0.5
Learning rate	5×10^{-6} - 1×10^{-2}	1×10^{-3}	1×10^{-3}	1×10^{-3}
Weight decay	5×10^{-6} - 1×10^{-2}	1×10^{-5}	1×10^{-6}	5×10^{-6}

Table 2: Hyperparameters of each model.

2017) to positional encoding of utterances in DAR. Let d be the dimension of the positional encoding. The k-th element of the positional encoding of the i-th utterance, $PE_{abs(i,k)}$, is calculated as follows:

$$PE_{abs(i,2k)} = \sin\left(\frac{i}{10000^{2k/d}}\right), \quad (8)$$

$$PE_{abs(i,2k+1)} = \cos\left(\frac{i}{10000^{2k/d}}\right). \quad (9)$$

Relative Positional Encoding (PE_{rel})

We incorporate the relative positional information of utterances by length-ratio positional encoding, which encodes the ratio of the position of an utterance to the dialogue length. PE_{abs} could not capture an occurrence phase (e.g., the beginning, middle, or last part) in a dialogue because the same absolute position can appear at different phases. For example, the 10th utterance belongs to the beginning part of a long dialogue (e.g., dialogue length = 100) whereas it belongs to the last part of a short dialogue (e.g., dialogue length = 10). To alleviate the variation of dialogue length, PE_{rel} encodes positional information normalized by the dialogue length $|D|$ as follows:

$$PE_{rel(i,2k,|D|)} = \sin\left(\frac{i}{|D|^{2k/d}}\right), \quad (10)$$

$$PE_{rel(i,2k+1,|D|)} = \cos\left(\frac{i}{|D|^{2k/d}}\right). \quad (11)$$

Note that our PE_{rel} is the same formulation as the one of Takase and Okazaki (2019). However, Takase and Okazaki (2019) have used the positional encoding for controlling the output length in document summarization. The purpose is different from ours, and our work is the first attempt to introduce a relative positional encoding to DAR.

3 Experiments

3.1 Settings

We evaluated our proposed method on the DAR task with the Switchboard corpus (SwDA) (Stolcke et al., 2000), consisting of 1,155 telephonic conversations using the standard splitting of the corpus (Lee and Dernoncourt, 2016). Table 1 shows the statistics of the SwDA dataset. The corpus comprises 43 different kinds of DA labels, and the vocabulary size is 19K.

We compared our two types of proposed models (s2s+PE_{abs} and s2s+PE_{rel}), each of which incorporates the positional encoding of utterances, described in Section 2.2, into the baseline DAR model (s2s), described in Section 2.1. We measured the performance of DAR as precision. We used Adam optimizer (Kingma and Ba, 2015) to train each DAR model, which is updated using a scheduler with a patience of 20 epochs and a reduced rate of 0.5. We used weight decay, gradient norm clipping, and dropouts (Srivastava et al., 2014). Each model was implemented using PyTorch and trained on a single NVIDIA GeForce GTX 1080 Ti. The hyperparameters of each model were optimized using

Model	Precision
Baseline Model	
s2s	77.84
Proposed Model	
s2s + PE_{abs}	**78.15***
s2s + PE_{rel}	78.06*

Table 3: Precision (%) on the SwDA corpus.

Dialogue Act	s2s	s2s + PE_{abs}	s2s + PE_{rel}
Statement-non-opinion	87.30	87.50 (+0.20)	**87.69** (+0.39)
Backchannel	90.26	89.95 (−0.31)	**90.37** (+0.12)
Statement-opinion	65.95	**67.31** (+1.36)	66.67 (+0.71)
Uninterpretable	81.27	**81.78** (+0.51)	81.23 (−0.03)
Agree or Accept	63.05	**64.86** (+1.81)	63.69 (+0.64)
Appreciation	81.95	**82.58** (+0.63)	81.95 (±0.00)
Yes-No-Question	81.50	**81.83** (+0.33)	81.67 (+0.17)
Yes Answers	72.77	75.64 (+2.88)	**76.27** (+3.51)
Conventional-closing	93.41	**93.48** (+0.07)	93.19 (−0.22)
Wh-Question	76.25	74.33 (−1.93)	**76.98** (+0.73)

Table 4: Precision (%) for the top 10 most frequent DA labels.

Optuna (Akiba et al., 2019) with 100 trials. Table 2 shows the hyperparameters of each model, including search range and the values optimized on the validation set.

3.2 Results

Table 3 shows the results of the experiment, wherein the numbers in bold represent the best scores, and * indicates that the improvement over the baseline "s2s" is statistically significant according to the Wilcoxon signed-rank test ($p \leq 0.01$). The results show that the performance of a seq2seq-based DAR model can be significantly improved by incorporating PE_{abs} or PE_{rel}. In particular, PE_{abs} and PE_{rel} increase precision by 0.31 and 0.22, respectively. This demonstrates the effectiveness of our proposed method.

Dialogue Act	# of Occurrences (Percentage)
Statement-non-opinion	65.7K (37.7 %)
Backchannel	33.3K (19.1 %)
Statement-opinion	22.7K (13.0 %)
Uninterpretable	13.2K (7.6 %)
Agree or Accept	9.8K (5.6 %)
Appreciation	4.1K (2.4 %)
Yes-No-Question	4.1K (2.3 %)
Yes Answers	2.5K (1.4 %)
Conventional-closing	2.1K (1.2 %)
Wh-Question	1.7K (0.9 %)
33 Other DAs	14.4K (8.2 %)
TOTAL	174.2K (100 %)

Table 5: The number of occurrences of the top 10 most frequent DA labels.

4 Analysis

4.1 Precision for Each DA Label

To analyze the effectiveness of the proposed method, we examined the precision of each model for each DA label. For simplicity, we evaluated the top 10 most frequent DA labels on the SwDA corpus. Table 5 shows the top 10 DA labels, accompanied by the number of occurrences. For the top 10 labels, Figure 2 shows the histograms of the label's absolute and relative positions in the training data, where the vertical axis is the number of occurrences, and the horizontal axis is the relative or absolute position. The relative positions are normalized by the dialogue length.

Table 4 shows the results, where the difference from "s2s" is shown in a parenthesis. As can be seen in the table, either of our proposed models, "s2s+PE_{abs}" or "s2s+PE_{rel}," achieves the best precision for all the labels, and our proposed models tend to obtain larger gains for DA labels with

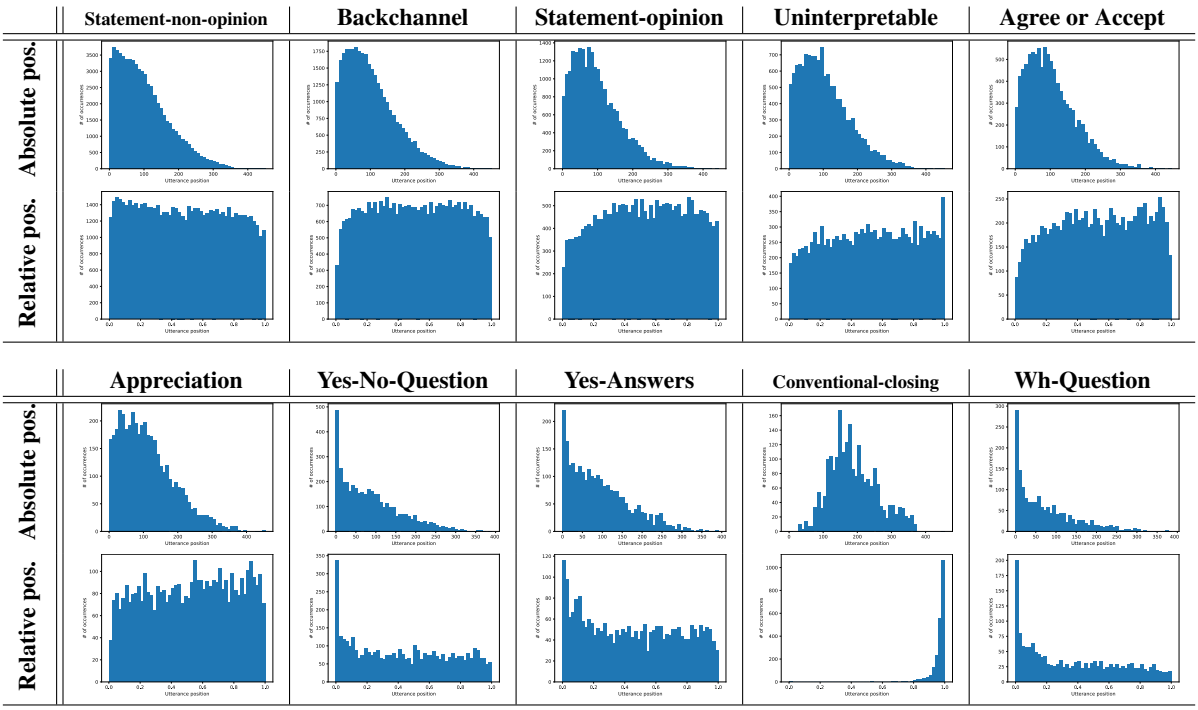


Figure 2: Histograms of absolute and relative positions for the top 10 most frequent DA labels. The relative positions are scaled from 0 to 1 according to the dialogue length.

		$len \leq 100$	$100 < len \leq 200$	$200 < len \leq 300$	$300 < len$	TOTAL
# of Utterances	Train	10,951	80,258	60,546	22,502	174,257
	Test	190	1,052	2,197	639	4,078

Table 6: Distribution of dialogue length.

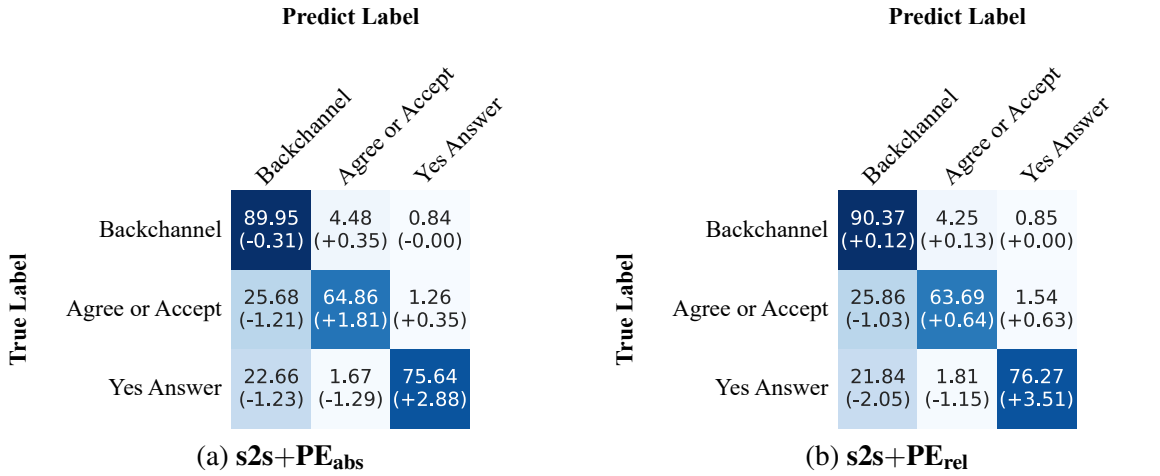
low precision in "s2s" (e.g., "Statement-opinion," "Agree or Accept," and "Yes Answer"). In particular, Table 4 and Figure 2 show that the gain tends to be larger for DA labels with stronger tendencies of occurrence positions (e.g., "Yes Answer" (+3.50) and "Wh-Question" (+0.73)). This indicates that our proposed positional encodings of utterances are useful for the prediction of the DAs that have a tendency to appear at a certain part of a dialogue.

4.2 Analysis of Similar DA Labels in Appearance

One of the difficulties in DAR is the recognition of DA labels in a confused group (Kumar et al., 2018; Bothe et al., 2018). In SwDA, "Backchannel," "Agree or Accept," and "Yes Answer" are easily confused with each other, and so are "Wh-Question," "Open-Question," and "Rhetorical-Question." This is because these labels have common expressions (e.g., *"Yes."* and *"Yeah."* may belong to "Backchannel," "Agree or Accept," and "Yes Answer"). In this section, we analyze the performance of the proposed models for DA labels in a confused group.

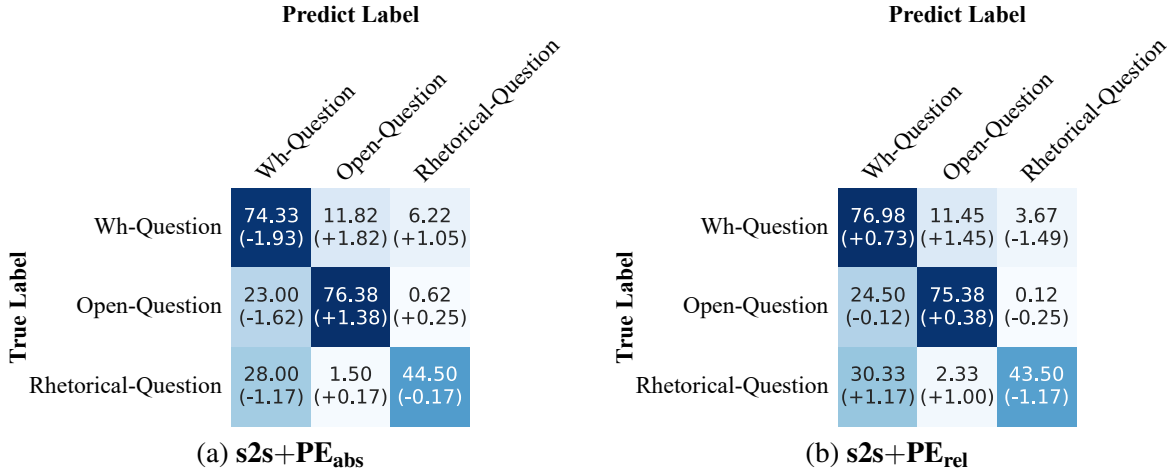
Figure 3 shows the confusion matrix of "Backchannel," "Agree or Accept," and "Yes Answer." As can be seen in Figure 3, the misrecognition as "Yes Answer" is widely reduced and the precision of "Yes Answer" is highly improved. This might be because "Yes Answer" has a tendency to appear at the beginning part, which is different from the appearance patterns of the other two labels (see Figure 2). This demonstrates that the proposed methods are effective to DA labels in a confused group as well.

Figure 4 shows the confusion matrix of "Wh-Question," "Open-Question," and "Rhetorical-Question," and Figure 5 shows the histograms of the absolute positions for the three labels in addition to those of the relative positions in the training data. As presented in Figure 4, for PE_{rel}, the precision of "Wh-Question" is improved, but the misrecognition of "Wh-Question" as "Open-Question" increases. This might be because "Wh-Question" has a tendency of relative occurrence positions, and so does "Open-Question." As for PE_{abs}, the



(a) **s2s+PE$_{abs}$** (b) **s2s+PE$_{rel}$**

Figure 3: Confusion matrix of "Backchannel," "Agree or Accept," and "Yes Answer." The difference from "s2s" is shown in a parenthesis.



(a) **s2s+PE$_{abs}$** (b) **s2s+PE$_{rel}$**

Figure 4: Confusion matrix of "Wh-Question," "Open-Question," and "Rhetorical-Question." The difference from "s2s" is shown in a parenthesis.

precision of "Wh-Question" decreases, and the misrecognition of "Wh-Question" as "Open-Question" and "Rhetorical-Question" increases. This might be because the three labels have the similar tendency of absolute occurrence positions.

4.3 Impact of Dialogue Length

We analyze the effectiveness of the proposed method on various dialogue lengths. We divide the test data into four groups according to the dialogue length and measure precision on each group. Tables 6 and 7 show the statistics of each group and the results, respectively. In Table 7, the numbers in bold represent the best scores, and * indicates that the improvement over the baseline "s2s" is statistically significant according to the Wilcoxon signed-rank test ($p \leq 0.01$).

Table 7 shows that the precision of "s2s" tends to decrease as the dialogue length increases. In contrast, both proposed models alleviate the tendency of "s2s" and preserve precision for long dialogues. Additionally, Table 7 shows that the proposed models statistically significantly outperform "s2s" on the group with the dialogue lengths of 300 or more. This indicates that our proposed models are effective for long dialogues.

In Table 7, the improvement of "s2s+PE$_{rel}$" over "s2s" is greater than that of "s2s+PE$_{abs}$." This indicates that "s2s+PE$_{rel}$" could successfully encode positional information of utterances with large absolute position by normalizing their positions.

5 Conclusions

In this paper, we have proposed an utterance position-aware approach for a seq2seq-based DAR

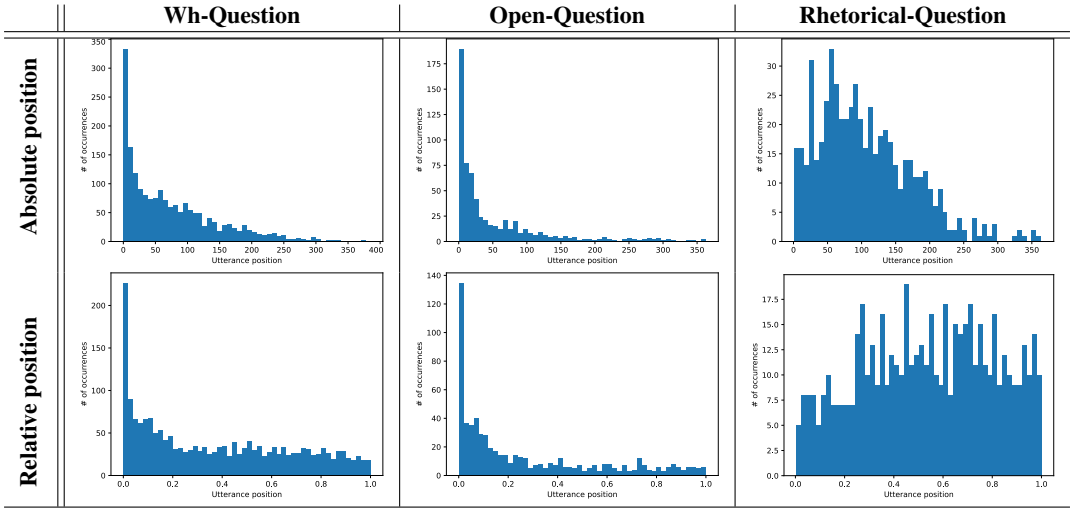


Figure 5: Histograms of absolute and relative positions for "Wh-Question," "Open-Question," and "Rhetorical-Question." The relative positions are scaled from 0 to 1 according to the dialogue length.

Model		$len \leq 100$	$100 < len \leq 200$	$200 < len \leq 300$	$300 < len$
Baseline Model	s2s	**79.56**	78.56	77.29	76.39
Proposed Model	s2s + PE_{abs}	79.37	**78.86**	**77.53***	77.58*
	s2s + PE_{rel}	79.37	78.68	77.40	**77.80***

Table 7: Precision (%) according to dialogue length.

model, which uses positional encoding for absolute or relative positions of utterances. The experiments showed that our sinusoidal absolute positional encoding and length-ratio relative positional encoding significantly improve the performance of DAR. Through the analysis, we confirmed that the proposed approach contributes to the prediction of DAs with strong tendencies of occurrence positions. Moreover, the analysis shows that the proposed approach works well for long dialogues.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number JP21K12031.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 2623–2631. ACM.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 225–234. ACM.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloé Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7594–7601. AAAI Press.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with CRF. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3440–3447. AAAI Press.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520, San Diego, California. Association for Computational Linguistics.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2019. A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392, Hong Kong, China. Association for Computational Linguistics.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15*, 15(1):1929–1958.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3999–4004. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.