RSLDE at NTCIR-16 DialEval-2 Task

Fan Li
Waseda University
Japan
lif@akane.waseda.jp

Tetsuya Sakai
Waseda University
Japan
tetsuyasakai@acm.org

ABSTRACT

In this paper, we report the work of the RSLDE team at the dialogue evaluation (DialEval-2) task of NTCIR-16, including the Chinese and English dialogue quality (DQ) and nugget detection (ND) subtasks. We implemented two sentence-level baselines that fine-tune BERT and XLNet along with a linear layer for the ND subtask. In addition, we propose a model based on BERT to capture the structure and context information of a customer-helpdesk dialogue. This dialogue-level model modifies the input and embeddings of BERT. We add a Transformer encoder layer over this model as our third model for the ND subtask and first model for the DQ subtask. The second model for the DQ subtask is the same dialogue-level model but without the Transformer encoder layer. Our XLNet model generated the best run for both English and Chinese ND subtasks. Our dialogue-level model outperformed the baselines for the Chinese DQ subtask but not for the English DQ subtask.

TEAM NAME

RSLDE

SUBTASKS

Nugget Detection (Chinese, English)
Dialogue Quality (Chinese, English)

KEYWORDS

Pre-trained language model, Transformer, Dialogue representation

1 INTRODUCTION

The RSLDE team participated in the English and Chinese dialogue quality (DQ) and nugget detection (ND) subtasks of DialEval-2 [11]. The task details of the two subtasks can be found in the overview paper [11]. We propose three models for the ND subtask and two models for the DQ subtask.

In the dialogue quality subtask, our proposed models include a modified pre-trained BERT model. By modifying the input sequence and position embeddings, this BERT model can capture the dialogue structure and extract dialogue context information. We also adopt this dialogue-level model in the nugget detection subtask, and implement a vanilla BERT model for comparison. Meanwhile, we use another pre-trained Transformer-based language model, XLNet [15], to generate customer and helpdesk nuggets.

The rest of this paper is organized as follows. Section 2 presents previous related studies, whereas section 3 describes our proposed approaches. Our submission and an analysis of results are reported in section 4. Finally, section 5 provides some concluding remarks.

2 RELATED WORK

The developments in the field of natural language processing (NLP) and machine learning (ML) have recently led to an increasing interest in task-oriented dialogue (TOD) systems. Such system assists users in accomplishing their goals by providing meaningful system responses. Compared to a traditional help desk, an automatic dialogue agent enhanced through the TOD system is able to handle customer inquires at any time of the day, which is more economic and efficient. However, evaluating such systems is usually dependent on human annotators, and is thus time-intensive and expensive. It is therefore of significant interest to propose an automatic dialogue evaluation system to alleviate this problem.

For the traditional closed-domain spoken dialogue system (such as a slot-filling task), most automatic evaluation studies follow the PARADISE framework [13]. However, evaluating customer-helpdesk dialogues is related to but extremely different from evaluating slot-filling dialogues. Customer-helpdesk dialogues discuss diverse problems about products and services, as opposed to slot-filling dialogues, which list up all the required slot in advance [17]. Therefore, one of the most significant discussions occurring in customer-helpdesk dialogue systems is the creation of an appropriate automatic method, which is the aim of the DialEval-2 task.

Several methods were proposed in the NTCIR-15 DialEval-1 task [16]. Most of the participants utilized a pretrained language model (e.g., BERT [4]) in their networks. These Transformer-based pretrained language models were pretrained on a large corpus of unlabeled data and are able to transfer knowledge efficiently to downstream tasks. For example, the IMTKU team used XLM-RoBERTa, a variant of BERT, by fine-tuning various transfer learning recipes for both DQ and ND subtasks [5]. Most of their runs outperformed the baselines. Similarly, the TUA1 team employed a pretrained BERT network for extracting features of dialogues before feeding them to a Bi-LSTM network [6]. Furthermore, they applied a self-attention network and several feed-forward neural network layers over this Bi-LSTM network. As a result, the proposed network can generate predictions that are very close to human annotations.

In addition to BERT and its variants, XLNet is another successful and representative Transformer-based pretrained language model. In contrast to BERT, which uses autoencoder (AE) language modeling for its training, XLNet applies a generalized autoregressive (AR) language modeling, and outperformed BERT on 20 tasks, often by a large margin, including question answering, natural language inference, sentiment analysis, and document ranking [15]. Therefore, one objective of this study is to investigate BERT and XLNet, the two most successful Transformer-based language models, in terms of their performance on DQ and ND subtasks.

Another important part of the DialEval-1 task is an approach to representing the structure of a dialogue. At NTCIR-15, the IMTKU

team utilized some tokenization tricks to better represent the structure [5]. the RSLNV team used two different encoders for the sentence and its context and applied one attention layer over the RNNs to understand the dialogue context [1]. The TUA1 team embeded not only each token but also its speaker identity and added this speaker embedding into the BERT output as a compound input to the Bi-LSTM network [6]. Inspired by Liu [8], who obtained the representation of each sentence of a document by modifying the input sequence and the embeddings of BERT, we propose a similar approach to better obtain the representation of each utterance of a customer-helpdesk dialogue.

3 APPROACHES

3.1 Sentence-level Baseline

Considering the fact that some nuggets can involve several typical patterns of words, one of our attempts is to tackle an ND subtask as a sentence-level classification problem, which is similar to the sentiment analysis task. Based on this attempt, we apply transfer learning by fine-tuning a pretrained BERT and XLNet, as our baselines for the ND subtask.

3.1.1 BERT. BERT [4] is a widely used Transformer-based language model that uses autoencoder (AE) language modeling for its pre-training. Rather than applying a left-to-right language model, the authors adopted masked language model (MLM) for the BERT pretraining. MLM masks 15% of all WordPiece tokens in each sequence at random and then requires the model to predict them, which thus enabled BERT to fuse the left and right contexts of a sentence and provide a bidirectional representation. In addition, Devlin et al. [4] also use a next sentence prediction (NSP) task that jointly pretrains text-pair representations. This enables BERT to transfer the pre-trained knowlodge on a wider range of downstream tasks. These novel procedures allow BERT to outperform the other similar Transformer-based model, such as GPT and GPT-2 [9] [10].

The BERT model we used for our sentence-level baseline is based on the pre-trained uncased base model of a HuggingFace's Transformers [14]. There are 12 layers, 12 attention heads, 768 neurons, and 110M parameters in total. First, the dialogues are split by turn, and each turn contains one or more utterances from either a customer or a helpdesk. Second, as the input of the BERT model, we rebuild two datasets that contain only utterances of the customer and helpdesk respectively. In this way, we transferred the ND subtask to a classification task. As the only difference between them, for classification task, the target label is a one-hot vector, whereas for the ND subtask, it has a probability distribution.

3.1.2 XLNet. Like BERT, XLNet [15] is also based on Transformer, but is pre-trained using a generalized autoregressive (AR) language model. Traditional AR language modeling seeks to estimate the probability distribution of a text corpus using an autoregressive model. Specifically, given a text sequence $\mathbf{x}$, AR language modeling factorizes the likelihood into either a forward or a backward product, and thus has the drawback of being a unidirectional network, i.e., it can only reach either the left or right context of the evaluated token. However, as demonstrated with BERT, bidirectional pre-training is very important for language representations, which

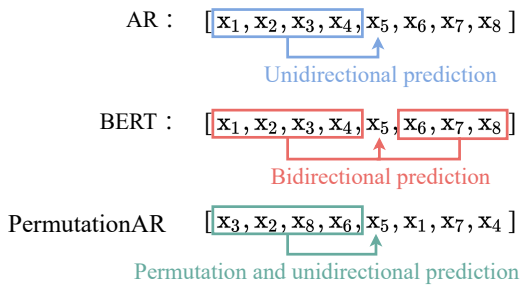

Figure 1: Comparison of AR, AE, and permutation AR

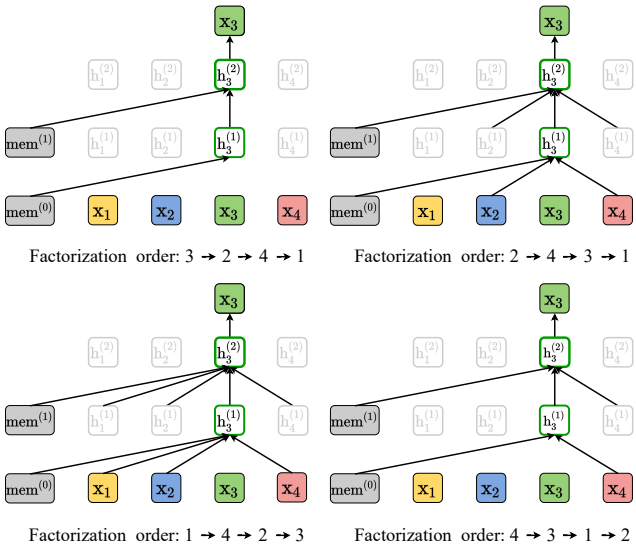

Figure 2: Illustration of the permutation language modeling objective for predicting x_3 given the same input sequence $\mathbf{x}$ but with different factorization orders (duplicated from [15])

motivated Yang et al. [15] to propose a generalized autoregressive method called permutation language modeling.

Permutation language modeling not only retains the benefits of AR models but also allows models to capture bidirectional contexts [15]. We compare AR, BERT, and permutation language modeling in Figure 1. During the training process, the permutation method, as shown in Figure 2, first shuffles the input tokens, and then predicts words for training according to a traditional autoregressive method. Therefore, for a sequence $\mathbf{x}$ of length T, there are $T!$ different orders for conducting a valid autoregressive factorization, which makes XLNet able to gather information from all positions on both sides. In addition, as another major contribution of XLNet, it follows the key ideas of Transformer-XL [3], i.e., the segment recurrence mechanism and relative encoding scheme for effectively learning long sentences. For further details, we refer the reader to the original study [15]. We use XLNet base models both for English and Chinese tasks. Similar to BERT, the base model has 12 layers, 12 attention heads, 768 neurons, and 110M parameters in

total. We use the XLNet-base-cased model of HuggingFace's Transformers [14] for the English subtask and HFL'XLNet-base chinese model [2] for the Chinese subtask. The training details are the same as those of BERT.

3.2 Dialog-level Model

For the baseline ND models above, we assumed that all utterances in a dialogue are independent to each other, i.e., are context-free. However, this assumption may lead to a loss of context information and thus limit the performance of our automatic evaluating systems. To address this limitation, inspired by Liu [8], who proposed a method to represent the sentence in a document for an extractive summarization, we present a BERT-based dialogue-level model to better represent the dialogue structure and better capture the context information in a dialogue. Because we need all information of a dialogue to estimate the quality, this BERT-based dialogue-level model is also apparently suitable for a DQ subtask.

3.2.1 Encoding Dialogue. Extracting context information to obtain the representation of each utterance is one of the most important procedures for both ND and DQ tasks. However, as described above, the pre-training objective of BERT is to obtain the representation of a masked token rather than that of a sentence. In addition, although BERT has segmentation embeddings for indicating different sentences, it only has two labels (sentences A and B), which means that it is inappropriate for dealing with a dialogue that has more than two turns. Therefore, following the key idea of Liu [8], we modify the input sequence and embeddings of BERT such that it can obtain the representation of each utterance in a dialogue.

As illustrated in Figure 3, we insert [CLS] and [SEP] tokens at the beginning and end of each utterance, respectively, and the embeddings of [CLS] are used as the representation of this utterance. Similar to using BERT in a classification task, the [CLS] token is considered to be able to aggregate features from the corresponding sentence. Hence, this modification makes BERT able to intuitively obtain the respective representation of each utterance in a dialogue.

Meanwhile, we use interval segment embeddings to distinguish multiple utterances within a dialogue. Specifically, for $sent_i$, we assign a segment embedding E_A or E_B conditioned on i as odd or even. Moreover, as described in the task details, a dialogue is formed by merging all consecutive posts by either a customer or helpdesk, which means that customer and helpdesk turns alternately appear. For example, using b_C and b_H to indicate customer and helpdesk turns respectively, a dialogue $[b_C, b_H, b_C, b_H, b_C]$ will be assigned as $[E_A, E_B, E_A, E_B, E_A]$. In other words, for the segment embeddings of BERT, all customer turns are assigned as E_A, and all helpdesk turns are assigned as E_B, which informs the model of dialogue structure.

In the following, T_i denotes the vector of the i-th [CLS] symbol from the top BERT layer and will be used as the representation for utterances $sent_i$.

3.2.2 Fine-tuning. After obtaining the utterance vectors of a dialogue, it is necessary to implement a layer to capture the relationship between these utterances. In our model, we employ a Transformer encoder layer to extract dialogue-level features focusing on the ND or DQ subtask from the BERT outputs:

$$\tilde{h}^l = \mathrm{LN}(h^{l-1} + \mathrm{MHAtt}(h^{l-1})), \quad (1)$$

$$h^l = \mathrm{LN}(\tilde{h}^l + \mathrm{FFN}(\tilde{h}^l)), \quad (2)$$

where $h^0 = \mathrm{PosEmb}(T)$ and T are the utterance vectors output by BERT. PosEmb represents positional encoding, a function that indicates the information about the relative or absolute position of the utterance in the dialogue. LN is the layer normalization operation, and MHAtt is the multi-head attention operation. The superscript l indicates the depth of the stacked layer.

For the ND subtask, we applied two densely connected NN layers over this Transformer encoder for generating customer and helpdesk nuggets. Meanwhile, for the DQ subtask, we first employed a global average pooling layer [7] to obtain the representation of the entire dialogue, and then applied three densely connected NN layers over it to generate A, E, and S score, respectively.

In addition, in order to verify the effect of this Transformer layer, we also implemented a baseline model for the DQ subtask, i.e., a dialogue embedding BERT model with a simple dense layer placed over it. In other words, the Transformer encoder layer was excluded.

Table 1: Official Runs Details

Task	Language	Run	Model	Batch size
DQ	English	0	Dialogue-BERT+Transformer	12
		1	Dialogue-BERT	
	Chinese	0	Dialogue-BERT+Transformer	
		1	Dialogue-BERT	
ND	English	0	XLNet Baseline	8
		1	BERT Baseline	
		2	Dialogue-BERT+Transformer	
	Chinese	0	XLNet Baseline	
		1	BERT Baseline	
		2	Dialogue-BERT+Transformer	

4 EXPERIMENTS

4.1 Run description

The details of our official runs are shown in Table 1. For the dialogue-level model, BERT and the other layers (Transformer encoder and dense layer) are fine-tuned jointly. We utilize Adam with $\beta_1 = 0.9, \beta_2 = 0.98$ for the optimization and cross-entropy as the loss function. All optimizers of our official run models follow the learning rate schedule of Vaswani et al. [12], with warming-up on first 10,000 steps:

$$lr = 2e^{-3} \cdot \min(step^{-0.5}, step \cdot warmup^{-1.5})$$

All models are trained for 100 epochs. Model checkpoints are saved once a better dev score is obtained. We chose the JSD score for the ND subtask and the average NMD score for the DQ subtask as our dev score.

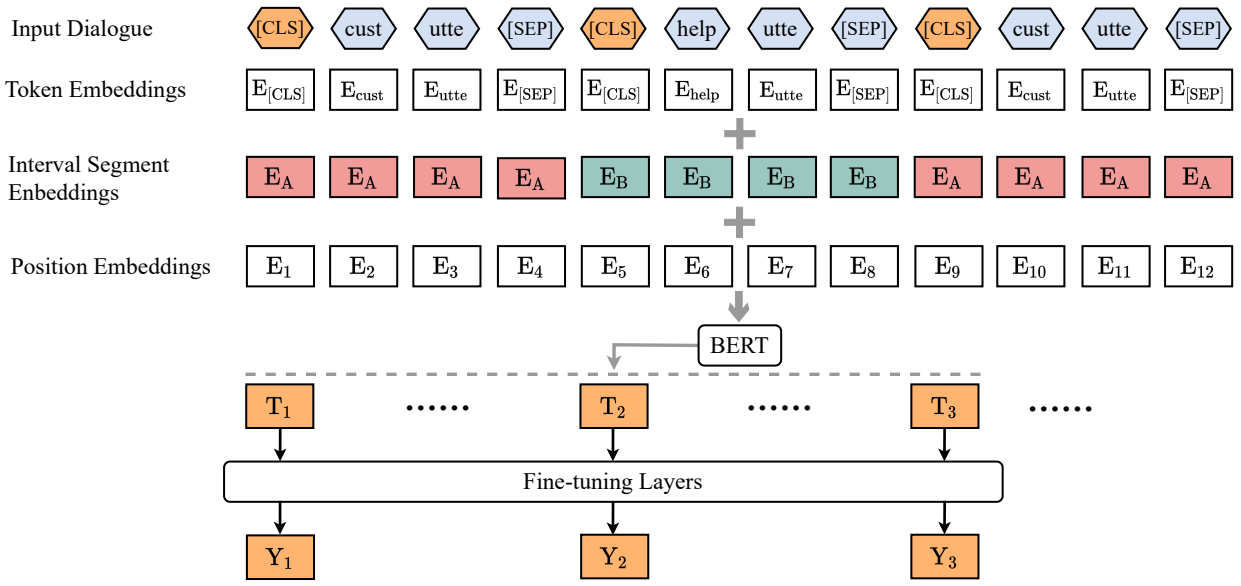

Figure 3: Overview architecture of the dialogue-level model. Here “cust" and “helpdesk" are abbreviations for customer and helpdesk, respectively, and “utte" indicates an utterance.

Table 2: Chinese Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
RSLDE-run0	0.0560$^{(1)}$	RSLDE-run0	0.1604$^{(1)}$
BL-LSTM	0.0585	BL-LSTM	0.1651
RSLDE-run2	0.0607	RSLDE-run1	0.1712
RSLDE-run1	0.0634	RSLDE-run2	0.1720
BL-popularity	0.1864	BL-popularity	0.2901
BL-uniform	0.2042	BL-uniform	0.3371

Table 3: Chinese Dialogue Quality (A-score) Results

Run	Mean RSNOD	Run	Mean NMD
BL-LSTM	0.2301	RSLDE-run0	0.1537
BL-popularity	0.2320	RSLDE-run1	0.1551
RSLDE-run0	0.2438	BL-popularity	0.1577
RSLDE-run1	0.2446	BL-LSTM	0.1772
BL-uniform	0.2767	BL-uniform	0.2500

Table 4: Chinese Dialogue Quality (S-score) Results

Run	Mean RSNOD	Run	Mean NMD
RSLDE-run0	0.1938	RSLDE-run1	0.1229
RSLDE-run1	0.1964	RSLDE-run0	0.1243
BL-LSTM	0.1998	BL-popularity	0.1288
BL-popularity	0.2062	BL-LSTM	0.1523
BL-uniform	0.2959	BL-uniform	0.2565

Table 5: Chinese Dialogue Quality (E-score) Results

Run	Mean RSNOD	Run	Mean NMD
RSLDE-run0	0.1660	RSLDE-run0	0.1222$^{(2)}$
RSLDE-run1	0.1725	RSLDE-run1	0.1286
BL-LSTM	0.1854	BL-LSTM	0.1579
BL-uniform	0.2496	BL-popularity	0.1710
BL-popularity	0.2569	BL-uniform	0.2106

Table 6: English Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
RSLDE-run0	0.0557$^{(1)}$	RSLDE-run0	0.1615$^{(2)}$
BL-LSTM	0.0625	BL-LSTM	0.1722
RSLDE-run2	0.0676	RSLDE-run2	0.1778
RSLDE-run1	0.0691	RSLDE-run1	0.1853
BL-popularity	0.1864	BL-popularity	0.2901
BL-uniform	0.2042	BL-uniform	0.3371

Table 7: English Dialogue Quality (A-score) Results

Run	Mean RSNOD	Run	Mean NMD
BL-popularity	0.2320	BL-popularity	0.1577
BL-LSTM	0.2321	BL-LSTM	0.1780
RSLDE-run0	0.2615	RSLDE-run1	0.1896
RSLDE-run1	0.2725	RSLDE-run0	0.1957
BL-uniform	0.2767	BL-uniform	0.2500

4.2 Results and Discussion

We list our experiment results of our official runs for the DialEval-2 task in Tables 2 to 9. The numbers in parentheses indicate the ranking of our runs according to the official results [11]. The organisers

Table 8: English Dialogue Quality (S-score) Results

Run	Mean RSNOD	Run	Mean NMD
BL-LSTM	0.1986	BL-popularity	0.1288
BL-popularity	0.2062	RSLDE-run0	0.1381
RSLDE-run0	0.2078	RSLDE-run1	0.1438
RSLDE-run1	0.2154	BL-LSTM	0.1467
BL-uniform	0.2959	BL-uniform	0.2565

Table 9: English Dialogue Quality (E-score) Results

Run	Mean RSNOD	Run	Mean NMD
BL-LSTM	0.1745	RSLDE-run0	0.1429
RSLDE-run0	0.1832	BL-LSTM	0.1431
RSLDE-run1	0.1889	RSLDE-run1	0.1444
BL-uniform	0.2496	BL-popularity	0.1710
BL-popularity	0.2569	BL-uniform	0.2106

of DialEval-2 provided three baseline models: a LSTM baseline, a uniform baseline, and a popularity baseline. The details of these three baselines can be found in the overview paper [11].

As shown in Table 2 and Table 6, although our XLNet baseline model is simply a sentence-level model, it still outperformed all other participant runs for both English and Chinese ND tasks. In other words, our XLNet baseline model outperformed our BERT baseline model for the sequence labeling task, which is in agreement with the results found by the author of XLNet [15]. In addition, by comparing the results of RSLDE-run2 and RSLDE-run1, as shown in Table 2 and Table 6, it can be observed that the dialogue-level model outperformed the BERT baseline model although they both used BERT, which means that it is meaningful to consider the structure and context information of a customer-helpdesk dialogue to detect its nugget.

According to Tables 3 to 5, except for the mean RSNOD of the A-score, both RSLDE-run0 and RSLDE-run1 outperformed the official baselines for the Chinese DQ subtask. Moreover, for most situations, the score of RSLDE-run0 is higher than that of RSLDE-run1, which means that our Transformer encoder layer was able to extract the features of the whole dialogue and evaluate its quality. However, as shown in Tables 7 to 9, neither model outperformed the official baselines for the English DQ subtask. This could be because the hyperparameters we used for the transfer learning are unsuitable for the English task.

5 CONCLUSION

In this paper, we proposed several models for automatically evaluating customer-helpdesk dialogues, i.e., two sentence-level models that respectively fine-tune BERT and XLNet as our baseline models, and two dialogue-level models. The two dialogue-level models modify the input sequence and embeddings of BERT to better capture the structure and context of a dialogue. One also uses a Transformer encoder to extract the dialogue-level feature, and the other, which lacks this Transformer encoder, is our baseline model for the DQ subtask.

Based on the official results, the following can be concluded:

- The XLNet model has an outstanding language understanding capability for customer-helpdesk dialogues.
- Considering the structure and context information of a dialogue is important for the dialogue nugget detection.

In future studies, it will be worth searching for technical aspects that make the hyperparameters suitable for the transfer learning of a Transformer-based model's.

ACKNOWLEDGEMENT

We thank the DialEval-2 organisers and the NTCIR chairs for giving us the opportunity to participate in this task.

REFERENCES

[1] Ting Cao, Fan Zhang, Haoxiang Shi, Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai, Injae Lee, Kyungduk Kim, and Inho Kang. 2020. RSLNV at the NTCIR-15 Dialogue Evaluation (DialEval-1) Task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*. 57–61.

[2] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. Association for Computational Linguistics, Online, 657–668. https://www.aclweb.org/anthology/2020.findings-emnlp.58

[3] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019).

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Mike Tian-Jian Jiang, Yueh-Chia Wu, Sheng-Ru Shaw, Zhao-Xian Gu, Yu-Chen Huang, Min-Yuh Day, Cheng-Jhe Chiang, and Cheng-Han Chiu. 2020. IMTKU Multi-Turn Dialogue System Evaluation at the NTCIR-15 DialEval-1 Dialogue Quality and Nugget Detection. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*. 68–74.

[6] Xin Kang, Yunong Wu, and Fuji Ren. 2020. TUA1 at the NTCIR-15 DialEval Task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*. 53–56.

[7] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. 1989. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems* 2 (1989).

[8] Yang Liu. 2019. Fine-tune BERT for extractive summarization. *arXiv preprint arXiv:1903.10318* (2019).

[9] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[11] Sijie Tao and Tetsuya Sakai. 2022. Overview of the NTCIR-16 Dialogue Evaluation (DialEval-2) Task. In *Proceedings of NTCIR-16*. to appear.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[13] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Madrid, Spain, 271–280. https://doi.org/10.3115/976909.979652

[14] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[15] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[16] Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai, and Inho Kang. 2020. Overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) Task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*. 13–34.

[17] Zhaohao Zeng, Cheng Luo, Lifeng Shang, Hang Li, and Tetsuya Sakai. 2018. Towards Automatic Evaluation of Customer-Helpdesk Dialogues. *Journal of Information Processing* 26 (2018), 768–778. https://doi.org/10.2197/ipsjjip.26.768