NKUST at the NTCIR-16 DialEval-2 Task

Tao-Hsing Chang
Department of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C
changth@nkust.edu.tw

Jian-He Chen
Department of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C
C107151129@nkust.edu.tw

ABSTRACT

It is important to evaluate the quality of dialogues generated by chatbots. Most previous automatic evaluation methods have been based on models (e.g., LSTM [1]) that are capable of processing time series. This study presents three models for dialogue quality and two nugget detection subtasks, respectively. Specifically, the first model is a Pegasus [2] model that can transform dialogues into short summaries; the second model is a Bi-LSTM [3] that merely adjusts the internal model structure; and the third model is a multi-agent model simulating situations in which multiple annotators generate different evaluation results for the same text. The experimental results show that certain opinions may need to be corroborated by more refined experimental design and the testing of more model parameters before they are applicable to this issue.

KEYWORDS

Dialogue quality evaluation, nugget detection, automatic summarization, BERT, Bi-LSTM, multi-agent model.

TEAM NAME

NKUST

SUBTASKS

Nugget Detection (Chinese, English)

Dialogue Quality (Chinese)

1 INTRODUCTION

The DialEval-2 task held by NTCIR-16 [4] needs to address the following issues. The participants of the DialEval-2 task must design a model that reads a dialogue between the helpdesk and consumers and then performs two subtasks: 1) judge the content quality of the dialogue and 2) determine whether each post in the dialogue is important. In this study, this model is referred to as a dialogue evaluation model (DAM). The subtask for judging the dialogue quality (DQ) is referred to as the DQ subtask by DialEval-2. The DQ subtask requires that the DAM assess the DQ in three aspects, including task accomplishment (TA), customer satisfaction (CS), and dialogue effectiveness (DE). The DialEval-2 uses A score, S score, and E score to indicate the evaluation scores of dialogues in TA, CS, and DE, respectively. Table 1 to Table 3 presents the meaning of different evaluation scores.

Table 1 Meanings of TA scores for a dialogue

	TA
2	The completeness of the dialogue is very high.
1	The completeness of the dialogue is high.
0	The completeness of the dialogue is acceptable.
-1	The completeness of the dialogue is not high.
-2	The completeness of the dialogue is low.

Table 2 Meanings of CS scores for a dialogue

	CS
2	The customer is very satisfied with the dialogue.
1	The customer is satisfied with the dialogue.
0	The customer has no special reaction to the dialogue.
-1	The customer is a little unsatisfied with the dialogue.
-2	The customer is very unsatisfied with the dialogue.

Table 3 Meanings of DE scores for a dialogue

	DE
2	The dialogue is very efficient.
1	The dialogue is sufficiently efficient.
0	The dialogue is efficient.
-1	The dialogue is inefficient.
-2	The dialogue is very inefficient.

Dialogues are generated between consumers and the helpdesk and the subtask for judging the importance of a post is referred to as the nugget detection (ND) subtask by DialEval-2. The ND subtask has labeled whether each post in a dialogue belongs to a consumer or the helpdesk and then requires the DAM to identify the category of each post in the dialogue. A consumer's post falls under one of the following four categories: "trigger," "goal," "regular nugget," and "not-a-nugget." The category "trigger" comprises posts the starting content of which is expressed by a consumer, and the DAM should mark as "CNUG0"; "goal" indicates that the post is the question content expressed by a consumer, and the DAM should mark it as "CNUG*"; "regular nugget" indicates that the post is the

key content expressed by a consumer, and the DAM should mark it as "CNUG"; "not-a-nugget" indicates that the post is unimportant content expressed by a consumer, and the DAM should mark it as "CNaN." Posts from the helpdesk also fall under one of three categories including "goal," "regular nugget," and "not-a-nugget," and the DAM should mark them as "HNUG*," "HNUG," and "HNaN," respectively. Figure 1 shows an example for different categories of posts in a dialogue.

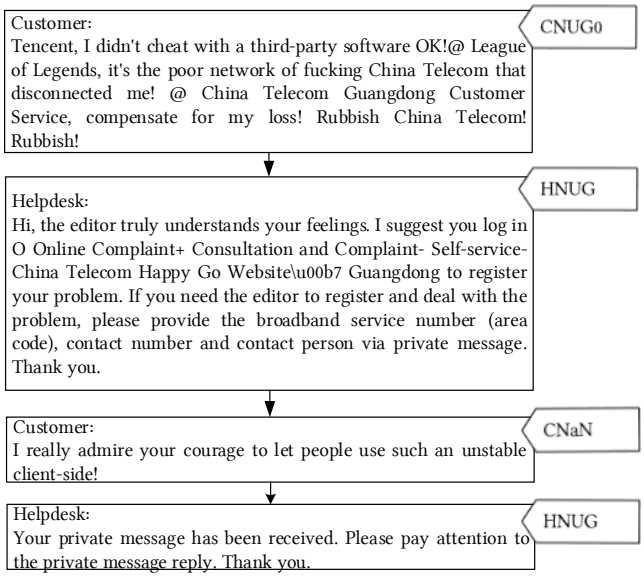

Figure 1 An example for describing different posts of a dialogue

The remainder of this paper is organized as follows. Section 2 reviews the studies concerning DQ evaluation; Section 3 presents the three DA prediction models; Section 4 describes the methods and results of verification of model performance, including experimental data and evaluation indices; and the last section discusses the characteristics and limitations of the method presented herein according to the experimental results and suggests the possible orientations of subsequent studies.

2 RELATED WORK

The NTCIR 15 DialEval-1 Task [5] is the previous edition of DialEval-2. Many teams have proposed models from different perspectives, the designs of which share many common features. In most of these models, semantic vectors are generated through word embedding combined with the context enhancement model, and then a classifier generates the evaluation results. Bi-LSTM is the most commonly used design for processing the context information. TUA1 [6] converts a complete dialogue into a semantic vector through a pre-trained BERT [7] model and then learns the semantic vector and determines whether it belongs to a customer or the helpdesk through the Bi-LSTM. Subsequently, the result learned by the bi-directional LSTM enters an attention [8] layer, which captures more abstract semantic vectors from the learning results. Finally, the output results of the attention layer enter a full connection layer for the purpose of prediction. The IMTKU [9] follows a similar design but uses the XLM-RoBERTa [10] language model, because the model is pre-trained simultaneously using the texts of different languages. This model also uses the transfer learning method. The tokenization and fine-tune techniques of some transformers are used in their model design. The SKYMN [11] has tried to use several different language model architectures such as CNN, LGC, DistilRoBERTa [12], and ALBERT [13]. The NKSUT [14] proposed using the overall dialogue semantic vector to predict text quality.

3 METHODS

The DAM herein is an improvement on the design proposed by Chang et al. (2020), which is based on BERT [7] combined with single sentence tasks. BERT is a language model that generates semantic vectors after creating a semantic space based on the context-dependency vocabulary. Using the encoder model generated in a transformer [8], BERT generates a semantic space, and self-attention is the core of the transformer. In the self-attention training, the model outputs vectors based on the word before and after each word inputted to the model and is trained by word guessing, thus enabling the model to further adjust the structure of the semantic space according to the correctness of word guessing.

Devlin et al. (2018) proposed the use of BERT to create an SSC task model. BERT can output a full-sentence semantic vector, so we can convert a complete dialogue into a semantic representation in a vector, and then input the vector to a classification model or SSC task model to learn and predict classification according to the semantic meanings of dialogues. Therefore, this study presents two DQ prediction models, including DNN combined with BERT and SSC task.

In recent years, many improved versions (e.g., ELECTRA [15]) of models have been proposed for converting texts to semantic vectors. To verify whether the conception herein can effectively improve the evaluation accuracy of the DAM, only BERT is used in this study as a model for converting texts to semantic vectors within a limited time. The DAM can be designed differently because different strategies are adopted to process the DQ subtask and ND subtask. These design details are described in the following subsections.

3.1 DAM for DQ subtask

In this study, we modify the DAM proposed by Chang et al. (2020) based on the finding that a dialogue usually contains many posts that are not directly related to the topic of the dialogue, and may disturb the conversion of the dialogue into semantic vectors. Hence, we propose that the important content in a dialogue be extracted through the abstracting process, thus reducing the influence of such topic-unrelated posts. In this study, a new abstract is generated for each dialogue in question using Pegasus.[2]

Pegasus is an abstractive rather than extractive automatic summarization model. The main principle of Pegasus is to mask the

important sentences of the training text during model pre-training and to require the model to predict the masked sentences according to the context of the text. In this way, Pegasus can learn to transform the input text into a few representative and significant new sentences. Therefore, the DAM herein first delivers a dialogue to Pegasus to generate a summary, and then the DAM converts the summary into semantic vectors and assesses the quality of the dialogue. This approach has an additional advantage: there is a limit to the length of the text that is inputted into a semantic vector model (e.g., BERT) for processing. Pegasus offers an effective solution for excessively lengthy texts by rewriting the dialogue into a less wordy summary.

Figure 2 presents the architecture of the DAM designed for the DQ subtask in this study. As described above, each dialogue, whether at the training or test stage, outputs a summary text through Pegasus and then provides it to the BERT model for processing. The BERT model converts the summary text into semantic vectors and inputs them to a classifier. This classifier is a full connection neural network with three layers and there are 768, 1,536, and 10 neurons at the three layers, respectively. The classifier can output scores as described in Table 1. At the training stage, the scores outputted by the classifier are compared with the real scores provided by DialEval-2 to calculate the loss values using the MSE function, thus allowing the classifier to learn according to the loss values and allowing the BERT model to perform the fine-tune procedure. The DAM herein is labeled as a run0 model in the DQ subtask result report of DialEval-2.

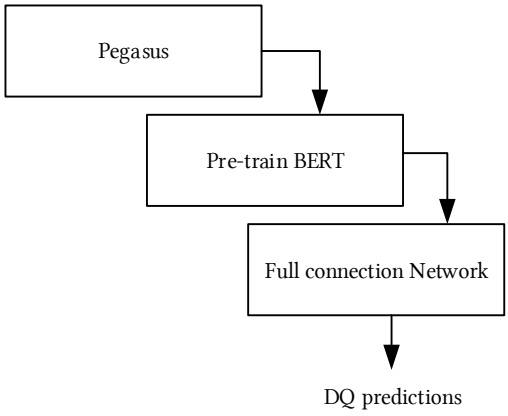

Figure 2. Architecture of the NKUST DQ run0 model

3.2 DAM for ND subtask

In this study, the DAM proposed by [14] for the ND subtask is modified based on two concepts. In the original design, the DAM directly outputs the semantic vector of each single post to the classifier for category labeling. However, it is difficult to determine the category of a post using its own semantic meaning alone because the category of the post is influenced by the multiple posts that appear before and after it. Some previous studies present a similar view. Therefore, to determine the category of a post, the vectors of previous and subsequent posts are fed to the LSTM first. The LSTM then adjusts the semantic vector of the post accordingly, and the classifier finally evaluates the probability that the adjusted semantic vector falls under each category. The LSTM model used by the DAM herein is Bi-LSTM, which has 768 dimensions in the input vector and 256 dimensions in the output vector. In the classifier part, this study adopts a full connection neural network with three layers, which comprise 512, 1,024, and 10 neurons, respectively. This DAM, shown in Figure 3, is labeled as a run0 model in the ND subtask result report of DialEval-2.

The second idea conceived in this study is that the evaluation of DQ and post category is actually very subjective, so each dialogue or post contains the evaluation results of 19 annotators in the dataset provided by DialEval-2. If the DAM follows a multi-agent design (i.e., there is an exclusive model for each annotator's evaluation results and each exclusive model is only trained using the evaluation results of the corresponding annotator), the 19 exclusive models can be regarded as 19 evaluators with their own opinions. Finally, the results outputted by the 19 models are integrated by a classifier, which outputs the probability value of the post under each category.

Figure 4 presents the architecture of the DAM designed according to this concept. After receiving a post, the DAM feeds it to 19 agents at the same time. Each agent has the same architecture, where the BERT model converts the post into vectors and sends them to a full connection neural network with three layers comprising 768, 256, and 7 neurons, respectively. The output results of the 19 agents are fed to a full connection neural network classifier with three layers comprising 133, 50, and 7 neurons, respectively. Finally, the classifier outputs a seven-dimensional vector, and the value of each dimension denotes the probability that the post falls under the corresponding category of the dimension. This DAM, shown in Figure 4, is labeled as a run1 model in the ND subtask result report of DialEval-2.

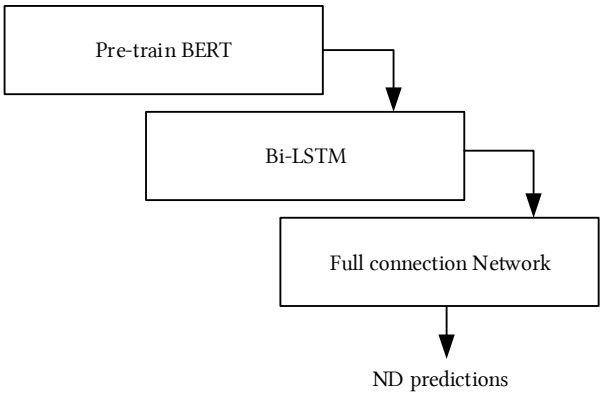

Figure 3. Architecture of the NKUST ND run0 model

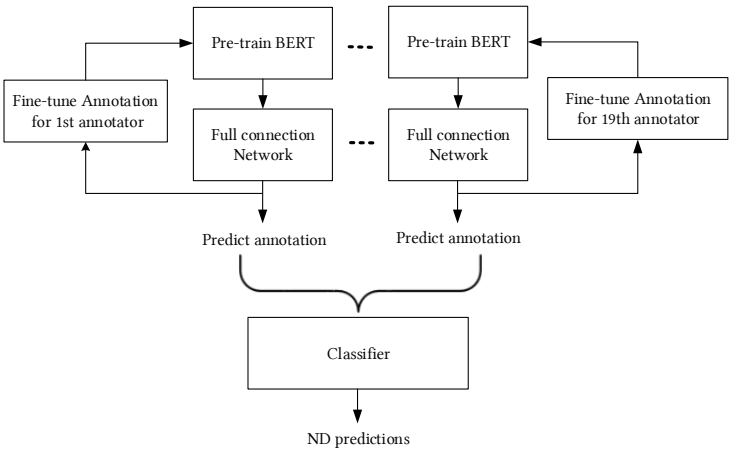

Figure 4 Architecture of the NKUST ND run1 model

4 EXPERIMENTS

The DAM herein uses the dataset provided by the DialEval-2 task of NTCIR-16 to train and test a model. This dataset is provided in both Chinese and English. The training data in Chinese comprise 3,700 dialogues with a total of 15,400 posts; the training data in English comprise 2,251 dialogues with a total of 9,211 posts. Each dialogue or post involves the results of manual evaluations conducted by 19 annotators.

Apart from the dataset provided by the DialEval-2 task and pre-trained BERT model, no other external data are used in this experiment. Data need to be pre-processed before they are used to train a model, so the method specified in the study by Chang et al. (2020) is used in this study to pre-process the data (as detailed in Section 4.1). Section 4.2 describes the equation that is used by the DialEval-2 task to measure the model performance. Section 4.3 demonstrates and discusses the performance of the DAM in the DialEval-2 task.

4.1 Data Preprocessing

Each dialogue and post provided by DialEval-2 contains data labeled by 19 annotators. The 19 annotators may give inconsistent evaluation results for the same dialogue, and the evaluation results given by the 19 annotators need to be integrated into a ground truth when the DAM proposed by the DQ subtask is trained and tested. Therefore, the evaluation results given by the 19 annotators are integrated into probability values of five quality levels by Equation (1). The probability value PoLi(D) of the dialogue D at that level i is calculated using the following equation:

$$\mathrm{PoL_i(D)} = {h_i}\Big/{\sum_{i\in G} h_i} \tag{1}$$

where, hi denotes the number of annotators who assess the dialogue D as Level i; and G denotes the set of five levels.

Likewise, each post provided by DialEval-2 also causes the same problem for the DAM herein. Therefore, the 19 evaluation results of each post are converted into category probability values for use as the ground truth through Equation (2). The probability PoCj(P) that the post P falls under the category j is calculated using the following equation:

$$PoC_j(P) = {h_j}\Big/{\sum_{j\in C} h_j} \tag{2}$$

where, hj denotes the number of annotators who assess the dialogue P as Level j; and C denotes the set of seven categories.

4.2 Performance Evaluation

DialEval-2 uses Bin-by-Bin and Cross-Bin to assess the model performance. The Bin-by-Bin method involves two indices including the RNSS and Jensen-Shannon Divergence (JSD). The RNSS is calculated using Equation (3).

$$\mathrm{RNSS}(p,p^*) = \sqrt{\frac{\mathrm{SS}(p,p^*)}{2}} \tag{3}$$

where, p denotes the predicted level; p* denotes the real level; and A denotes the set of levels. The Sum of Squares (SS) function is calculated using the following equation:

$$\mathrm{SS}(p,p^*) = \sum_{i\in A} (p(i)-p^*(i))^2 \tag{4}$$

The JSD is calculated using the following equation:

$$\mathrm{JSD}(p,p^*) = \frac{\mathrm{KLD}(p\|pm)+\mathrm{KLD}(p^*\|pm)}{2} \tag{5}$$

where, pm(i)=(p(i)+p*(i))/2 for i=1,.....L; The function KLD is expressed as follows:

$$\mathrm{KLD}(p_1\|p_2) = \sum_{i \text{ s.t. } p_{1(i)>0}} p_1(i)\log_2\frac{p_1(i)}{p_2(i)} \tag{6}$$

Cross-bin involves two indices including the NMD and Root Symmetric Normalized Order-Aware Divergence (RSNOD). The NMD is calculated using Equation (7):

$$\mathrm{NMD}(p,p^*) = \frac{\mathrm{MD}(p,p^*)}{2} \tag{7}$$

where MD is calculated using Equation (8):

$$MD(p,p^*)=\sum_{i\in A}|cp(i)-cp^*(i)| \quad (8)$$

where, $cp(i)=\sum_{k=1}^{i}p(k)$ and $cp^*=\sum_{k=1}^{i}p^*(i)$

RSNOD is calculated using Equation (9):

$$RSNOD(p,p^*)=\sqrt{\frac{SOD(p,p^*)}{L-1}} \quad (9)$$

The SOD function and the related function it uses are expressed as follows:

$$SOD(p,p^*)=\frac{OD(p//p^*)+OD(p^*//p)}{2} \quad (10)$$

$$OD(p//p^*)=\frac{1}{|B^*|}\sum_{i\in B^*}DW(i) \quad (11)$$

$$DW(i)=\sum_{j\in A}|i-j|\left(p(j)-p^*(j)\right)^2 \quad (12)$$

4.3 Experimental results

Table 4 and 5 describes the performance of the DAM in the DQ subtask. We can see that the performance of Run 0 is very unsatisfactory. To determine whether the design of "auto-generating the summary of a dialogue and feeding it to the DAM" does not perform as expected, we examined the summary results of Pegasus. Figure 4 shows an example in which the summary of a dialogue is auto-generated by Pegasus. In this example, the summary generated by Pegasus contains the request from the customer but ignores the solution offered by the helpdesk. This may account for the poor performance of the DAM in the DQ task.

Table 4. Chinese Dialogue Quality Results (RSNOD)

indices / model	TA	CS	DE
TUA1-run2	0.1992	0.1758	0.1671
Baseline-run0	0.2301	0.1998	0.1854
NKUST-run0	0.2774	0.2732	0.2253

Table 5. Chinese Dialogue Quality Results (NMD)

indices / model	TA	CS	DE
TUA1-run2	0.1325	0.1166	0.1310
Baseline-run0	0.1772	0.1523	0.1579
NKUST-run0	0.2453	0.2293	0.1897

Table 6 describes the performance of the DAM in the ND subtask. We can see that the bidirectional Bi-LSTM indeed serves to capture the relationship between different posts and judge the post category accordingly. By contrast, the multi-agent Run 1 model does not perform as expected. This may be because the inconsistency between annotators is replicated in each agent and the classifier structure is too simple to effectively address this inconsistency. We did not try other classifier designs due to time constraints.

Table 6. Chinese & English Nugget Detection Results

indices / model	Chinese		English	
	JSD	RNSS	JSD	RNSS
RSLDE-run0	0.0560	0.1604	0.0557	0.1574
Baseline-run0	0.0585	0.1651	0.0625	0.1722
NKUST Run 0	0.0670	0.1761	0.0641	0.1744
BL-uniform	0.2042	0.3371	0.1864	0.2901
BL-popularity	0.1864	0.2901	0.2042	0.3371
NKUST Run 1	0.2432	0.3774		

Customer:
I arrived at the Unicom Business Hall before 9am today and it was the eighth time! @ China Unicom Beijing Customer Service I lined up till 11 o'clock and there were 26 people in front of me. During the transaction, I was told the fixed line number was transferred, so the transfer of broadband failed! I was asked to wait! At least half an hour... Do you think your business volume is too small? I unbound it last month and it can be transferred this month. I have to pay the phone bills of this month next month, which can't be paid online. I must go to the business hall! Are you nuts?
Helpdesk:
Hello! We are sorry for your unpleasant feeling. What specific business do you want to transact?
Customer:
Transfer the fixed line phone and renew the broadband package.
Helpdesk:
Do you renew the annual WO family package?
Customer:
I just want to know I have the annual broadband package and I only need to transfer it, why I have to unbind it, pay an extra month of broadband package fees and go to the business hall again next month? Why can't these businesses be transacted one time or online? Who have time to spend half a day in lining up?
Helpdesk:
If it is WO family package, you shall cancel it first and then the transfer can be transacted. If it is still within the annual period, it can't be transacted till the annual package expires. The transfer concerns the change of the user information. To protect the users' interest, self-service transaction is not supported at present. You have to go to the business hall. We are really sorry for your inconvenience.

Figure 5(a) An example for dialogue

I arrived at the Unicom Business Hall before 9am today and it was the eighth time! . During the transaction, I was told the fixed line number was transferred, so the transfer of broadband failed!, What specific business do you want to transact? Transfer the fixed line phone and renew the broadband package. I just want to know I have the annual broadband package and I only need to transfer it, why I have to unbind it, pay an extra month of broadband package fees. If it is WO family package, you shall cancel it first and then the transfer can be transacted..

Figure 5(b) Manual summary

Hello! We are sorry for your unpleasant feeling. What specific business do you want to transact? Transfer the fixed line phone and renew the broadband package.Do you renew the annual WO family package? I just want to know I have the annual broadband package and I only need to transfer it, why I have to unbind it, pay an extra month of broadband package fees

Figure 5(c) Summary generated by Pegasus

4.4 Error Analysis

In order to explore why the proposed method for DQ did not work as expected, we performed an additional experiment. This experiment extracted 100 dialogues from the validation data provided by DialEva-1 as Dataset A. For each dialogue in Dataset A, we summarized it manually. These manual summaries form Dataset B. In addition, these dialogues are also input to Pegasus to generate machine summaries, called Dataset C. We use the DAM designed by Chang et al. (2020) to predict the E score of the dialogues. The accurate rate of this DAM for the Datasets A, B and C is 0.39, 0.51, and 0.36, respectively. It is clear that the DAM with manual summaries have better accuracy than that with original dialogues. However, using summaries generated by Pegasus degrade the prediction accuracy of the DAM.

After observing the summaries generated by Pegasus, we found that the quality of the summaries was not good, leading to a decrease in the prediction accuracy of the DAM. The original dialogue in Figure 5(a) is used as an example to illustrate this problem. Figure 5(b) is the manual summary of the dialogue in Figure 5(a) while Figure 5(c) is the summary of the dialogue generated by Pegasus. The E score of this dialogue is 1. The E score of the text in Figure 5(a) and that in Figure 5(c) are -1 while that in Figure 5(b) is 1. Obviously, the dialogue has more irrelevant posts to the topic, while the manual summary effectively removes the posts. However, Pegasus's summary is not sound. Therefore, we believe that summarizing the dialogue can improve the accuracy rate of DAM, but only if the dialogue is summarized correctly.

5 CONCLUSION

DA remains an important issue. The DAM herein does not surpass the baseline, but the experiment in this study is not sufficiently complete due to time constraints, and some models have not been completely tested or reported. We believe that if such models receive improved testing (e.g., using language models with better performance and testing more parameter combinations), the DAM designed based on the view herein may have a chance to perform better.

REFERENCES

[1]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780

[2]. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning pp. 11328-11339.

[3]. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks, 18(5-6), 602-610

[4]. Tao. S., Sakai. T. (2022). Overview of the NTCIR-16 Dialogue Evaluation (DialEval-2) Task,

[5]. Zeng, Z., Kato, S., Sakai, T., & Kang, I. (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

[6]. Kang, X., Wu, Y., & Ren, F. TUA1 at the NTCIR-15 DialEval Task. In Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies, 53-56.

[7]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

[8]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[9]. Jiang, M. T. J., Wu, Y. C., Shaw, S. R., Gu, Z. X., Huang, Y. C., Day, M. Y., ... & Chiu, C. H. 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.

[10]. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[11]. Wang, J., Zhang, Y., Sakai, T., & Yamana, H. SKYMN at the NTCIR-15 DialEval-1 Task. In Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies, 40-46.

[12]. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[13]. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[14]. Chang, T. H., Chen, J. H., & Chen, C. C. NKUST at the NTCIR-15 DialEval-1 Task. In Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies, 62-67.

[15]. Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.