# NKUST at the NTCIR-15 DialEval-1 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

Chi-Chia Chen
Department of Computer Science
and Information Engineering
National Kaohsiung University of
Science and Technology
Kaohsiung, Taiwan, R.O.C
1106108103@nkust.edu.tw

## ABSTRACT

Chatbot dialogue quality evaluation is an important topic. Most existing automatic evaluation methods are based on models that can handle time series (for example, the long short-term memory (LSTM) model). However, this research adopted another approach to directly convert a complete dialogue into a semantic vector through Bidirectional Encoder Representations from Transformers (BERT). Subsequently, the vector was entered into a simple classification model for training and prediction. The experimental results for the DialEval-1 task reveal that the performance of the proposed method is reasonably comparable to that of a LSTM-based baseline model.

## KEYWORDS

Dialogue quality evaluation, Nugget detection, BERT, DialEval-1.

## TEAMNAME

NKUST

## SUBTASKS

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

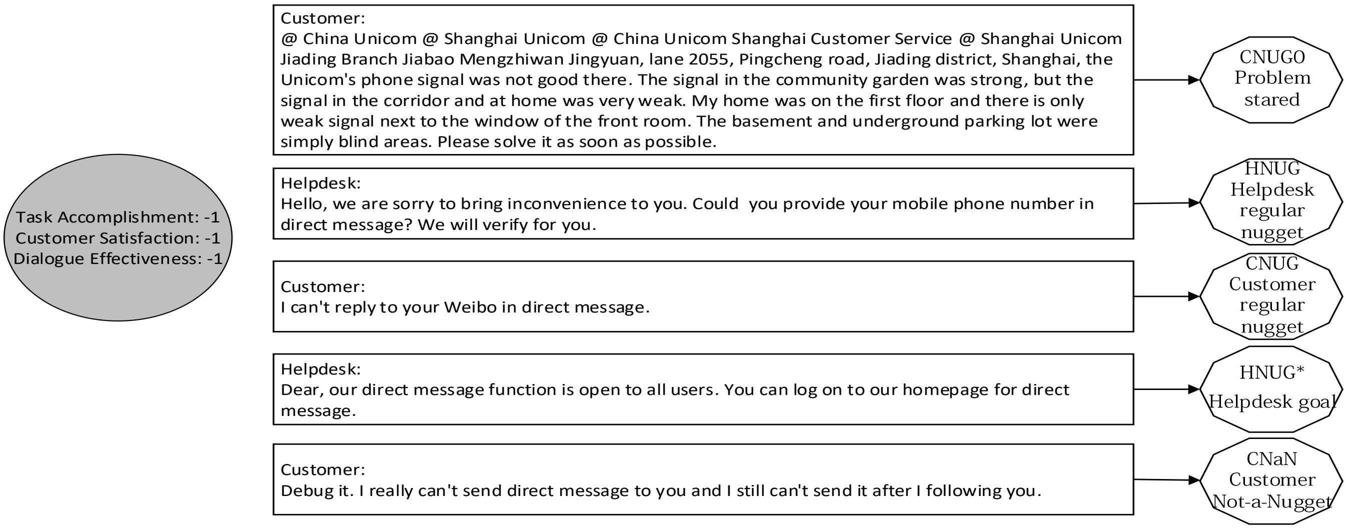
## 1 Introduction

Chatbots have been extensively studied recently while dialogue quality evaluation and dialogue focus extraction have become relevant research topics. The method proposed for these two issues is called Dialogue Assessment Model (after this referred to as DAM) in this paper. According to the STC-3 task at the 14th NTCIR conference (NTCIR-14) [1] and the DialEval-1 task at NTCIR-15 [12], two tasks should be used to compare the different DAMs. They are dialogue quality (DQ) and nugget detection (ND). Figure 1 shows examples of DQ and ND. It also shows that there is a dialogue between a consumer and the helpdesk. There are three indices to assessing the quality of dialogue. They are task accomplishment (A-score), customer satisfaction with the dialogue (S-score), and dialogue effectiveness (E-score). Each index has five levels. The meaning thereof is listed in Table 1. The DQ task of a DAM is to accurately estimate the scores of the three indices for each dialogue. The STC-3 uses a cross-bin metric to evaluate the performance of a DAM for a DQ task..

Each dialogue consists of several posts. Each post contains the conversation content of the consumer of the helpdesk. Furthermore, the consumer's posts can be classified and marked in four categories. Firstly, the "trigger" denotes the starting content, and it is marked as CNUG0. Secondly, the "goal" refers to the content of the question. It is marked as CNUG*. Thirdly, the "regular nugget" is the focus of the consumer and is marked as CNUG. Finally, representing the unimportant content of the consumer, the "not-a-nugget" is marked as CNaN. Similarly, the posts by the helpdesk are classified into three categories. They are identical to the client's posts, except there are no triggers. They are marked as HNUG*, HNUG, and HNaN, correspondingly. The ND task of a DAM is to identify the category of each post. The STC-3 adopts a bin-by-bin metric to measure the performance of a DAM for the ND task.

**Table 1 Meaning of each level in the three aspects for dialogue**

| levels | A-score | S-score | E-score |
|---|---|---|---|
| 2 | Exceptionally high task accomplishment | High customer satisfaction | Highly effective dialogue |
| 1 | High task accomplishment | Fair customer satisfaction | Sufficiently effective dialogue |
| 0 | Acceptable level of task accomplishment | Neutral satisfaction | Acceptably effective dialogue |
| -1 | Low task accomplishment | Slight customer dissatisfaction | Ineffective dialogue |
| -2 | Exceptionally low task accomplishment | High customer dissatisfaction | Exceptionally ineffective dialogue |



Customer:
@ China Unicom @ Shanghai Unicom @ China Unicom Shanghai Customer Service @ Shanghai Unicom Jiading Branch Jiabao Mengzhiwan Jingyuan, lane 2055, Pingcheng road, Jiading district, Shanghai, the Unicom's phone signal was not good there. The signal in the community garden was strong, but the signal in the corridor and at home was very weak. My home was on the first floor and there is only weak signal next to the window of the front room. The basement and underground parking lot were simply blind areas. Please solve it as soon as possible.

Helpdesk:
Hello, we are sorry to bring inconvenience to you. Could you provide your mobile phone number in direct message? We will verify for you.

Customer:
I can't reply to your Weibo in direct message.

Helpdesk:
Dear, our direct message function is open to all users. You can log on to our homepage for direct message.

Customer:
Debug it. I really can't send direct message to you and I still can't send it after I following you.

**Figure 1: Illustration of DQ levels and ND categories using a dialogue**

Most of the DAMs proposed in the STC-3 use bag-of-words (BoW) and long short-term memory (LSTM) models [2] as bases. BoW models can convert dialogues directly into semantic vectors, while LSTM models can learn to predict the category of dialogue according to the semantics of the dialogue. However, experimental results show that these methods still need improvement. Two pertinent issues have been discussed. First, the semantic representations of the BoW models deviate from the real semantics. Secondly, semantic and classification learning by the LSTM models can still be improved. Hence, in this paper, a model designed for Single Sentence Classification (SSC) tasks was integrated with Bidirectional Encoder Representations from Transformers (BERT) [3] to predict DQ and ND.

BERT is a language model to construct semantic spaces and to generate context-dependency semantic vectors. It adopts the method at the encoder stage of the transformer [4] to construct semantic spaces while the transformer uses self-attention as the technical core. Self-attention modifies semantic vector for each word according to its preceding and succeeding words in a sentence, and trains models by word prediction. Moreover, models can further modify the semantic space according to prediction accuracy. Therefore, for the same word, BERT provides different semantic vectors as the preceding and succeeding words are not the same.

In addition, Devlin et al [3]. proposed the use of BERT to perform SSC models. Because BERT can produce the semantic vectors of sentences, a complete dialogue can be converted into a semantic presentation in a vector. Next, this vector can be inserted as a classification model or an SSC task model to learn and predict how to classify according to the dialogue's semantics. Hence, this paper proposed two dialogue quality prediction models. BERT was integrated with DNN and SSC tasks separately. The paper is arranged as follows. Section 2 provides a comprehensive review of the relevant studies on dialogue quality assessment. Section 3 introduces the two dialogue quality prediction models proposed. Section 4 illustrates the experiments to validate the proposed models. Data for the experiments, assessment indicators and experimental results are dealt with in this section. Finally, according to the experimental results, the characteristics and limitations of the proposed models are discussed. Suggestions are made for future research.

## 2 Related Works

In the early stages of chatbot research, the focus had been on how to construct chatbots. At the same time, the Bilingual Evaluation Understudy (BLEU) [5], used to evaluate the performance of machine translation models, or manual inspections were adopted to assess dialogue effectiveness. For example, Xu et al. [6] employed these two methods to assess the proposed chatbot's performance. However, when BLEU is adopted to validate dialogue effectiveness, the actual outcome of such dialogue has to be known to determine the differences between machine-made dialogues and real dialogues. As a consequence, this method does not apply to automatic dialogue quality assessments.

In recent years, several automatic dialogue quality evaluation models have been proposed. They mainly use word embedding together with LSTM models. This architecture is relatively intuitive. Because words in these models are often represented as vectors, one-hot encoding cannot be used with limited training data. With word embedding, the number of dimensions of the input vectors can be reduced. However, a dialogue is a time series-related process. Hence, it is reasonable to use LSTM models suitable for handling sequential data. For instance, Zeng et al. [12] proposed a basic architecture in which a BoW model was integrated with an LSTM model. This model was adopted as the baseline model for the STC3 task.

Similarly, CUIS [7] also used this architecture. BERT was utilized as the word embedding tool, and a classifier with GRU was used as the core. For the classifier, a hierarchical attention network (HAN) was adopted by CUIS. The experimental results demonstrated that the model performance is satisfactory. Meanwhile, WUST [8] used a dialogue quality prediction model formed by combining a modified BoW model with a three-layer Bi-LSTM model.

On the contrary, SLSTC [9] used BERT to replace the BoW model. The Bi-LSTM [10] model was employed to address the problem that the semantic information on words at the end of a sentence cannot be obtained using earlier words in the LSTM model. SLSTC also designed a new loss function. The original loss function calculates the differences between the predicted and actual values directly. These differences are then used to train models. The new loss function uses the overall probability distributions of different scores for a dialogue as the prediction targets. The differences between the predicted and actual probability distributions are adopted for training. In addition, this study discovered that the BERT fine-tune method is insignificant for addressing the issue concerned. This might be due to limited training data.

## 3 Model Design

The proposed model uses word embedding with a classifier. For the word embedding model, BERT is adopted. Unlike previous studies, this research explores whether dialogue quality assessment is possible when the text of a completed dialogue is used directly for prediction. Therefore, two simple classifier designs are adopted. One is a multi-layer, fully-connected classifier, while the other is a BERT SSC task model. They are described as follows.

BERT consists of multiple transformer encoders (TE), as shown in Figure 2. The figure indicates that n is the maximum length of the input sentence, while $E_1$ to $E_n$ represents the input word's results after embedding, and $T_1$ to $T_n$ denote the output semantic vectors of $E_1$ to $E_n$, generated according to the context before and after the input sentence. In the hidden layer between the input and output, 12 layers of TE are used. Self-attention is adopted for internal computation (Vaswani et al., 2017). It helps to integrate the semantic of the preceding and succeeding words in the sentence. This enables the model to understand the whole sentence. Each word has to pass through 12 heads in each TE layer, while each head denotes a self-attention computation process. The computation results of the 12 heads will be input to the next TE layer. Finally, the semantic vector of the word is obtained.

The proposed method is based on the assumptions explained in section 1. Dialogue is considered a part of the text. BERT is adopted to extract the text's semantic and convert it into a semantic vector, subsequently input to a classifier. In this study, two classifier designs are adopted. One is a multi-layer, fully-connected neural network, giving the probabilities of prediction indicators at different levels. It consists of three layers. There are 768 and 1536 neurons in the input and hidden layers, respectively. Different designs with 5, 4, or 3 neurons are used for the output layer according to the number of levels predicted by the model. For the neurons in the output layer, the softmax function is adopted as an activation function. For all other neurons in the model, the ReLU function is employed. The mean-square error (MSE) is used as the loss function during model training. In addition, the dropout, learning rate, and the batch size during training are assigned to 0.25, 0.05, and 125, respectively. In this paper, the model employs this classifier is denoted as Model 0.

The other classifier is used in a BERT SSC task model, which is a fine-tuned BERT. When the linear classifier is being trained, BERT is fine-tuned to produce the prediction indicator's probabilities at different levels. The linear classifier network is made up of only two layers. The input layer contains 768 neurons, while the output layer has 3 to 5 neurons according to the number of levels predicted by the model. Similarly, the activation functions of the neurons in the output layer are softmax functions. Besides, MSE is used as the loss function during training. The batch size and learning rate are 8 and 0.01, correspondingly. In this paper, the model employs this classifier is denoted as Model 1.

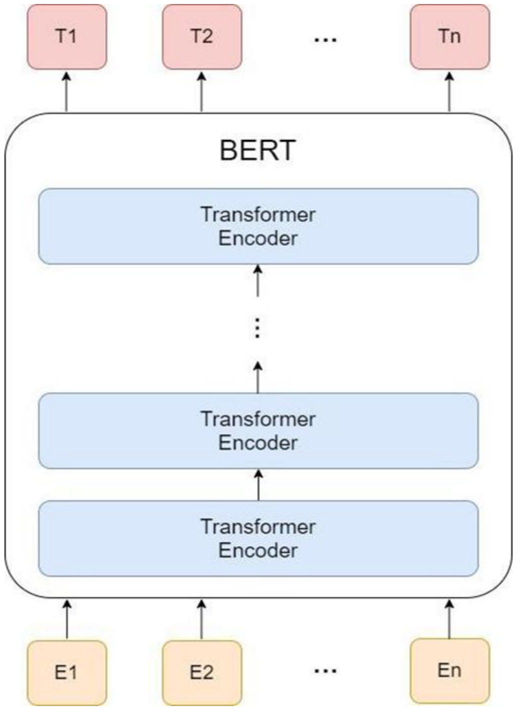


**Figure 2: BERT framework**

## 4 Experiments

The open-source code of Bert For Sequence Classification implemented by Huggingface [11] was adopted to design the proposed method. The dataset provided by the NTCIR-15 DialEval-1 task was used to train and test the method. The Chinese training data provided by the DialEval-1 task contain 3700 dialogues, which include a total of 15400 posts. The English training data contain 2251 dialogues, which are equivalent to 9211 posts in total.

Apart from the DialEval-1 task dataset and the trained BERT model, no other external data were used in the experiments. Data have to be treated before they can be used to train the models. The preprocessing procedures are illustrated in section 4.1, whereas the equations adopted in the DialEval-1 task for model performance evaluation are given in section 4.2. Finally, the proposed method's performance for the DialEval-1 task is shown and discussed in section 4.3.

## 4.1 Data Preprocessing

Each dialogue provided by the DialEval-1 task contains data remarked on by 19 experts. There are two types of marks. The first type gives the levels of the three DQ indicators. It should be emphasized that, because the 19 experts may have different judgments on the same indicator of the same dialogue, the DialEval-1 task provides 19 results for each indicator for each dialogue. When the proposed method is applied for model training, one model will be trained independently for each dialogue indicator. This model requires the probability of the indicator at each level so that it can be trained to learn the relationships between semantic vectors and levels. Therefore, Equation (1) is used to convert each indicator's 19 remarks into probabilities at different levels. For dialogue *D*, the probability *PoL<sub>i</sub>(D)* of an indicator at level *i* is

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

where $h_i$ is the number of experts who graded dialogue *D* as level *i*, and *G* denotes the set of levels.

The remarks of the second type are the ND categories of different posts in a dialogue. As mentioned in section 1, there are two subjects for ND: the consumer and the helpdesk, and they have four and three categories, respectively. For training and test data, the subject of each post is marked. Similarly, there are 19 results for the category of each post. Hence, Equation (2) is employed to convert each post's 19 remarks into probabilities of different categories. For post *P*, contributed by subject *S*, the probability $PoC_j(P)$ at where it belongs to category *j* is

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

where $h_j$ denotes the number of experts who classified post *P* into category *j*, and *C* is the set of the categories.

## 4.2 Performance Evaluation

This paper adopted bin-by-bin and cross-bin metrics [12] to evaluate the model performance. There are two indicators for the bin-by-bin method: *Root Normalized Sum of Squares* (*RNSS*) and *Jensen-Shannon Divergence* (*JSD*). *RNSS* is calculated, as illustrated in Equation (3)

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

where $p$ is the predicted level; $p^*$ is the actual level, and A represents the set of levels. The function *SS* is the sum of squares, which is obtained using Equation (4).

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

*JSD* is calculated, as shown in Equation (5).

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

*KLD* is acquired, as shown in Equation (6).

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

Similarly, the cross-bin metric also includes two indicators: the *Normalized Match Distance (NMD)* and the *Root Symmetric Normalized Order-Aware Divergence (RSNOD)*. The calculation equation for the *NMD* is given in Equation (7).

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

where the *Match Distance* (*MD*) is calculated using Equation (8).

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

The equation for *RSNOD* is provided below (Equation (9)).

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

where the *Symmetric Order-Aware Divergence* (*SOD*) is obtained with the help of Equation (10).

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

## 4.3 Experimental Results

Tables 1 to 3 list the experimental results of the two proposed models for the DialEval-1 task. NKUST Run 0 and NKUST Run 1 give the results obtained using the Models 0 and 1 mentioned in section 3, respectively. BL-LSTM, BL-popularity, and BL-uniform are the baseline models proposed by Zeng et al. [12]. With the DQ indicator experiments for Chinese dialogues, all indicators of all aspects demonstrate that Run 1 outperforms BL-popularity, BL-uniform, and Run 0. In terms of task accomplishment, the performance of Run 1 is similar to that of BL-LSTM. Meanwhile, Customer Satisfaction of the dialogue for Run 1 is higher than that of BL-LSTM. Lastly, the dialogue effectiveness of Run 1 is lower than that of BL-LSTM. For the English dialogues, experiments were only performed using Run 0. For both Chinese and English dialogues, Run 0 can only outperform BL-uniform.

**Table 1. Chinese Dialogue Quality Results**

| | A-score | | S-score | | E-score | |
|---|---|---|---|---|---|---|
| Model | RSNOD | NMD | RSNOD | NMD | RSNOD | NMD |
| BL-LSTM | 0.2305 | 0.1598 | 0.2088 | 0.1455 | 0.1782 | 0.1386 |
| BL-popularity | 0.2473 | 0.1643 | 0.2288 | 0.1442 | 0.2614 | 0.1781 |
| BL-uniform | 0.2706 | 0.2522 | 0.2811 | 0.2497 | 0.2425 | 0.2110 |
| NKUST Run 0 | 0.2696 | 0.2384 | 0.2653 | 0.2289 | 0.2222 | 0.1973 |
| NKUST Run 1 | 0.2430 | 0.1594 | 0.2057 | 0.1363 | 0.2295 | 0.1508 |

**Table 2. English Dialogue Quality Results**

| | A-score | | S-score | | E-score | |
|---|---|---|---|---|---|---|
| Model | RSNOD | NMD | RSNOD | NMD | RSNOD | NMD |
| BL-LSTM | 0.2271 | 0.1591 | 0.2111 | 0.1413 | 0.1687 | 0.1248 |
| BL-popularity | 0.2473 | 0.1643 | 0.2288 | 0.1442 | 0.2614 | 0.1781 |
| BL-uniform | 0.2706 | 0.2522 | 0.2811 | 0.2497 | 0.2425 | 0.2110 |
| NKUST Run 0 | 0.2801 | 0.2345 | 0.2637 | 0.2189 | 0.2248 | 0.1963 |
| NKUST Run 1 | N/A | N/A | N/A | N/A | N/A | N/A |

Tables 1 to 2 reveal that the dialogue quality assessment performance of Run 0, to which all posts in a dialogue are input, is comparable to that of the LSTM model, in which time series are emphasized. However, this may be because BERT replaces the BoW model. Hence, further analysis is needed to prove whether a prediction model using the complete text of dialogue as the processing unit is feasible. Moreover, Run 0 differs from Run 1, mainly because the fine-tune method is adopted for the latter. Hence, from the experimental results, it is found that the use of the fine-tune method is more important than the differences in the classifier designs. This finding is different from the conclusions by SLSTC. Further research on this should be conducted.

Table 3 reflects the ND category classification results predicted by the models. The performances of Run 0 and Run 1 are not satisfactory. Run 1 outperforms BL-uniform and Run 0. The two models adopting time-series generate more satisfactory results than Run 1. This suggests, for the ND category prediction, information of the preceding and succeeding texts must be included. Otherwise, it will be challenging to determine whether a paragraph is essential merely based on its semantic meaning.

**Table 3. Chinese & English Nugget Detection Results**

| | Chinese | | English | |
|---|---|---|---|---|
| Model | JSD | RNSS | JSD | RNSS |
| BL-LSTM | 0.0709 | 0.1673 | 0.0762 | 0.1781 |
| BL-popularity | 0.1301 | 0.2068 | 0.1301 | 0.2068 |
| BL-uniform | 0.2858 | 0.4190 | 0.2858 | 0.4190 |
| NKUST Run 0 | 0.3116 | 0.4169 | 0.3157 | 0.4172 |
| NKUST Run 1 | 0.1905 | 0.3036 | N/A | N/A |

## 5 Conclusions and Future Research

This study proposes two dialogue quality prediction models. The method’s characteristic is to make predictions directly using the overall semantic meanings of dialogues, by not inserting the paragraphs of the dialogue separately. The experimental results demonstrate that the model’s performances are similar to those of the LSTM models proposed by Zeng et al. However, more experiments with different model designs are required to validate this conclusion.

It is believed that the following items are worth further investigation. First, BERT can be used with classifiers using LSTM as cores, and model structures can be adjusted with the fine-tune method. Future research should focus on this. Secondly, the model’s loss functions can be modified according to the loss function suggested by SLSTC. This way, the loss function fits better into the dialogue quality assessment model. As a result, the model can learn the correct prediction method more quickly.

## ACKNOWLEDGMENTS

This work is supported in part by the Ministry of Science and Technology, Taiwan, R.O.C. under the Grant MOST 107-2511-H-992-001-MY3.

## REFERENCES

[1] Zeng, Z., Kato, S., & Sakai, T. (2019). Overview of the NTCIR-14 short text conversation task: Dialogue quality and nugget detection subtasks. In Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies (pp. 289-315).

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

[3] 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.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 5998-6008.

[5] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318)

[6] Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 3506-3510).

[7] Cong, K., & Lam, W. (2019). CUIS at the NTCIR-14 STC-3 DQ Subtask⋆. In Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies (pp. 309-398)

[8] Yan, M., Liu, M., & Xiang, J. (2019). WUST at the NTCIR-14 STC-3 Dialogue Quality and Nugget Detection Subtask. In Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies (pp. 376-382).

[9] Kato, S., Suzuki, R., Zeng, Z., & Sakai, T. (2019). SLSTC at the NTCIR-14 STC-3 dialogue quality and nugget detection subtasks. In Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies (pp. 355-361).

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

[11] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shlerifer, S.,von Platen, P., Ma, C., Jernite, Y., Plu, Julien., Xu, Canwen., Scao, L. T., Gugger, S., Drame, M., Lhoest, Q., Rush, M. A., Hugging Face & Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv, arXiv-1910.

[12] 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.