TUA1 at the NTCIR-16 DialEval-2 Task

Fei Ding
c502147003@tokushima-u.ac.jp
Tokushima University
Tokushima, Japan

Xin Kang
kang-xin@tokushima-u.ac.jp
Tokushima University
Tokushima, Japan

Yunong Wu
raino.wu@gmail.com
CDXT medical technology
China

Fuji Ren
ren@tokushima-u.ac.jp
Tokushima University
Tokushima, Japan

ABSTRACT

In this paper, we report the work of TUA1 team in the dialogue evaluation (DialEval-2) task of NTCIR-16, which consists of two subtasks: the Dialogue Quality (DQ) subtask and the Nugget Detection (ND) subtask. Our proposed method consists of two parts: a feature extractor and a feedforward network. The feature extractor employs pre-trained Transformer networks to extract the hidden representations of the dialogue utterances and employs a Latent Dirichlet Allocation (LDA) method to extract the topic information of these utterances. The feedforward network then concatenates the hidden representations and the topics extracted by the feature extractor, compresses them through several feedforward layers into a desired dimension, and finally predicts the quality scores and nugget types of the dialogues. Since the DialEval-2 task dataset was composed of the one-to-one translated Chinese and English dialogues, we employ the pre-trained Transformer networks for Chinese and English, respectively, to extract the hidden representations of the dialogues. This makes it possible to process the sub-tasks on the Chinese or English datasets simultaneously. We train the neural network models based on the mean squared error for both dialogue quality prediction and nugget detection subtasks. Our proposed method reaches the best scores for RSNOD and NMD metrics in both Chinese and English dialogue quality subtasks among all participants. The results indicate that the proposed method is promising in learning a dialogue quality prediction system for generating very close predictions to the human annotators.

KEYWORDS

Tokenization, Fine-tuning, Transformers, Dialogue evaluation, Dialogue quality.

TEAM NAME

TUA1

SUBTASKS

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

1 INTRODUCTION

With the development of natural language understanding and generation technologies, more and more commercial companies have been setting up intelligent dialogue systems, such as helpdesk robots [4]. These dialogue systems could provide wait-free and homogeneous services for their customers, but at the same time suffer from the problems of misunderstanding and generating nonsense or offensive utterances to the system users. Assessing dialogue quality and analyzing key points of dialogue have become two important research topics, which can provide useful feedback automatically and effectively for tuning the dialogue systems. The TUA1 team participates in the dialogue quality (DQ) and nugget detection (ND) subtasks of NTCIR-16 Dialogue Evaluation (DialEval-2) task. The detailed task descriptions and dataset definitions can be found in the overview paper [3]. This work is a continuation of our previous works in text conversation [2, 8] and text emotion analysis [1].

In the dialogue quality subtask, we employ multiple pre-trained Transformer models and a topic clustering model, as a feature extractor. The extractor extracts the hidden representations and topical information from the one-to-one translated Chinese and English dialogues respectively. Then we feed the extracted features into a feedforward network to finally predict the quality scores of the input dialogue sequences. Our network simultaneously evaluates three types of the dialogue quality, namely the task accomplishment (A-score), the customer satisfaction (S-score), and the dialogue effectiveness (E-score).

In the nugget detection subtask, we employ the same feature extractor to obtain the feature vectors from the dataset. The only difference is that, to fit the labels of the nugget detection subtask, we divide the dataset into two parts: the helpdesk turn and the customer turn, and train two different models respectively to predict the probabilities of the nugget labels. Four probability scores related to {CNUG0, CNUG, CNUG*, CNaN} are predicted for the customer turn while three scores including {HNUG, HNUG*, HNaN} are predicted for the helpdesk turns. The feedforward network is also modified accordingly to accommodate the nugget detection subtask.

The dialogue quality prediction results are evaluated with two cross-bin metrics, which are the Root Symmetric Normalized Order-aware Divergence (RSNOD) and the Normalized Match Distance (NMD). In contrast to DQ subtask, the classes in nugget detection subtask are nominal, so bin-by-bin metrics are more suitable. Specifically, two metrics are used in ND subtask: Root Normalised Sum of Squares (RNSS) and Jensen-Shannon Divergence (JSD). Higher similarities of the ground truth probabilities and the predicted probabilities correspond to smaller values in RSNOD, NMD, RNSS, and JSD. We submitted 9 runs in total for both Chinese and English DQ and ND subtasks. Among all participants of the dialogue quality subtask, the TUA1 team achieves the best results of RSNOD and NMD scores for all types of dialogue quality predictions, that is, the predictions of the A-score, S-score, and E-score, for both the Chinese and English dialogues. For the Chinese Nugget Detection subtask, the TUA1 team also achieves desirable RNSS and JSD scores.

The rest of this paper is organized as follows. Section 2 and section 3 elaborate the proposed dialogue quality prediction method and nugget detection method respectively. Section 4 reports our runs and analyses the results from different aspects. Section 5 concludes our work and elaborates our future works.

2 DIALOGUE QUALITY PREDICTION NETWORK

The dialogue quality prediction network mainly consisted of two parts: a feature extractor and a feedforward network. The two parts are shown in Figure 1, separated by a dotted line in between.

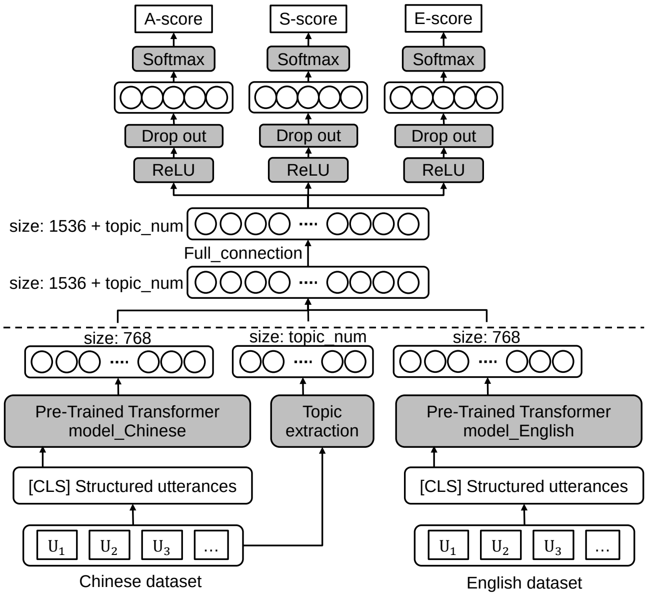


Figure 1: The structure of dialogue quality prediction network.

Firstly, in order to better represent the structure of a set of dialogue, we preprocess the input dialogue sequence. For example, consider a tokenized dialogue below:

```
"id": "3830772740080373"
[CLS] Customer(1/6): "What's going on with ... "
      Helpdesk(2/6): "Hi, I'm Little @ of ... "
      ...
      Helpdesk(6/6): "Dear, please choose ... " [SEP]
```

We add [CLS] and [SEP] special tokens at the beginning and end of each dialogue respectively to fit the training of Transformer models. Moreover, we append meta data at the beginning of every utterance in the dialogues. The meta data consists of both the sender information and the utterance position information in the entire dialogue. Customer/Helpdesk indicates the sender of the utterance, specifically, for the n-th utterance of a customer-helpdesk dialogue of m utterances in total, we append (n/m) as the utterance position information. Both the Chinese and English corpora are preprocessed by following the same procedure as depicted above.

Secondly, we employ pre-trained Transformer networks to extract the hidden representations of the structured dialogue input. The Transformer architecture [5] was first proposed by Google, which has achieved countless amazing results in Natural Language Processing (NLP) tasks in recent years and had a large number of model variants and a complete research ecology. We explored the HuggingFace [6] and experimented with a variety of the Transformer models, including BERT, Chinese-BERT, Roberta, and their fine-tuned models, for obtaining the proper hidden representations of dialogue utterances. We put the details of the model selection and experimental comparison in section 4. Unlike the strategy we proposed in NTCIR 15 [2], we did not use the feature vectors of each input token as the input of the downstream module, but only extracted the 768-dimensional feature vector of the [CLS] token as the entire dialogue hidden representation.

Topic information is also extracted as a part of the hidden representation of the input dialogue. Specifically, we employ a Latent Dirichlet Allocation (LDA) model to obtain the topic information of the input dialogue. The process of LDA extracting the topic words of the text is an unsupervised process and the number of topics K is a hyperparameter. The setting of K has a crucial influence on the result of prediction with different text lengths, which is discussed in section 4. The purpose of LDA is to infer the hidden topic structure using observable words which follows the following generative process (assuming the corpus has D dialogues and K topics):

(1) For each topic $k \in K$, calculate $\beta_k \sim$ Dirichlet(η). This draws a distribution of the words, which can be treated as the probability of a word appearing in topic k.
(2) For each document $d \in D$, calculate the topic proportions $\theta_d \sim$ Dirichlet(α).
(3) For each word i in document d:
   (a) Calculate the topic assignment $z_{di} \sim$ Multinomial(θ_d).
   (b) Calculate the observed word $w_{ij} \sim$ Multinomial($\beta_{z_{di}}$).

For parameter estimation, the posterior distribution is:

$$p(z,\theta,\beta|w,\alpha,\eta) = \frac{p(z,\theta,\beta|\alpha,\eta)}{p(w|\alpha,\eta)}. \tag{1}$$

Variational Bayesian method uses a simpler distribution $q(z,\theta,\beta|\lambda,\phi,\gamma)$ to approximate the posterior since it is intractable. The variational parameters λ,ϕ,γ are optimized to maximize the Evidence Lower Bound (ELBO):

$$\begin{aligned} &\log P(w|\alpha,\eta) \geq L(w,\theta,\gamma,\lambda) \triangleq \\ &E_q[\log p(w,z,\theta,\beta|\alpha,\eta)] - E_q[\log q(z,\theta,\beta)]. \end{aligned} \tag{2}$$

Maximizing ELBO is equivalent to minimizing the Kullback-Leibler (KL) divergence between $q(z,\theta,\beta)$ and the true posterior $p(z,\theta,\beta|w,\alpha,\eta)$. The topic probability vector of each dialogue is used as the input of the downstream module, together with the hidden representations extracted by the Transformer models. We remove stop_words to obtain more effective topic information.

Thirdly, the feedforward network is located above the dotted line, as shown in Figure 1. The architecture of the feedforward network is arranged as follows: a full connection layer, an activation function, a dropout layer, a linear dimension reduction layer, and a softmax function. The full connection layer can be regarded as a weight layer, whose input and output dimensions are the same. Due to the role of the full connection layer, the model can act on both Chinese and English datasets and only need to fine-tune the parameters of the feedforward network, learning the parameters of the extractor untouched. The linear layer takes the input of a 1546 to 1586 dimensions vector, and the output is a 5 dimension vector, which is the probability distribution of the dialogue quality scores. The model has three linear layers, which output the A-score, S-score, E-score respectively.

Finnally, the network get the probabilities over the quality label set $\Gamma = \{-2, -1, 0, 1, 2\}$ for every quality type. More illustrations of the quality labels can be found in [7]. We employ the mean squared error (MSE) loss for evaluating the training loss. The model generates three distributions $\hat{y}_i^A$, $\hat{y}_i^S$, and $\hat{y}_i^E$as the predictions for the A-score, S-score, and E-score of dialogue quality, respectively. We take the means of y_i over l human annotators for the A, S, and E scores as the targets, which are denoted by $\bar{y}_i^A$, $\bar{y}_i^S$, and $\bar{y}_i^E$ respectively. The training loss based on mean squared error is then given by:

$$\mathrm{loss}(\bar{y}, \hat{y}) = \sum_{\kappa \in \{A,S,E\}} \frac{1}{n} \sum_{i=1}^{n} \left(\bar{y}_i^{\kappa} - \hat{y}_i^{\kappa}\right)^2 . \tag{3}$$

3 NUGGET DETECTION NETWORK

In nugget detection subtask, in order to fit the nugget labels, we divide all the dialogue utterances into two parts, named the customer part and the helpdesk part, that is, utterances extracted from either customer or helpdesk are trained separately. Thus, all modules in the DQ subtask are carried over to the ND subtask except the LDA topic clustering module, since topic information corresponds to the entire dialogue. Similar to the DQ subtask, we feed the concatenated utterances to the pre-trained Transformer models to obtain the Chinese and English hidden representations, and then employ a feedforward network to get the probability distributions of each nugget label.

The mean squared error loss is also used for training in ND subtask. Detailed formula description can refer to Equation 3. A simple schematic diagram of the nugget detection network can refer to Figure 2.

4 EXPERIMENTS

4.1 Official Run Results and Discussions

TUA1 team submits three runs for the Chinese dialogue quality (DQ) subtask, one run for the English DQ subtask, two runs for the Chinese nugget detection (ND) subtask, and one run for the English ND subtask. The specifics of each run are detailed as follows.

(1) Chinese_DQ_run0: similar to the method from TUA1 team proposed in NTCIR 15 DialEval-1 Task, simply replace BERT with Roberta.
(2) Chinese_DQ_run1: proposed dialogue quality prediction network with topic nunber of 20 and dropout probability of 0.2.
(3) Chinese_DQ_run2: proposed dialogue quality prediction network with topic nunber of 10 and dropout probability of 0.1.

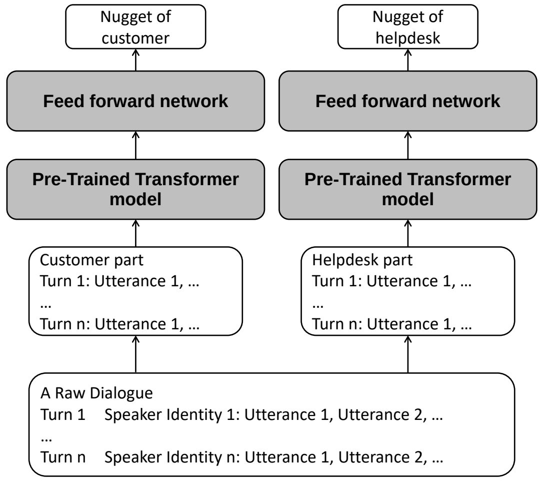


Figure 2: The structure of nugget detection network.

(4) English_DQ_run0: same model structure as Chinese_DQ_run2, just retrain the parameters with English dev dataset. Note that although the best results for the Chinese and English subtasks of DQ in the overview paper are run_2 and run_0 respectively, they are actually the same model structure with the same hyperparameters.
(5) Chinese_ND_run0: proposed nugget detection network.
(6) Chinese_ND_run1: similar to the method from TUA1 team proposed in NTCIR 15 DialEval-1 Task, simply replace BERT with Roberta.
(7) English_ND_run0: same model as Chinese_ND_run0.

We report the results of three types of dialogue quality predictions for Chinese and English subtasks in Table 1 and 2. The evaluation metrics are based on the mean Root Symmetric Normalised Order-aware Divergence (RSNOD) metric and the mean Normalised Match Distance (NMD) metric. The bold font in the table indicates the best result in the overview paper. Among all types of scores, the model of run_2 in Chinese subtask and run_0 in English subtask achieved the highest overall score. Note that they are the same model since we only submitted one run in the English subtask. The model structure can refer to Figure 1 with a topic number of 10 and dropout probability of 0.1, thus the dimension of the concatenated hidden representation is 1546. We employ "roberta-base-finetuned-jd" as the Chinese Pre-trained Transformer model and "roberta-base" as the English Pre-trained Transformer model. The last 2 layer parameters of the model are unfrozen for fine-tuning. A detailed discussion of Transformer model selection and unfreeze layers can be found in section 4.2.

Next, we report the result of nugget detection in Table 3, which is evaluated based on the mean Jensen-Shannon Divergence (JSD) metric and the mean Root Normalised Sum of Squares (RNSS) metric, respectively. The experimental results show that our proposed model, although not the best among the participants, still exceeds the baseline using BL-LSTM network.

Table 1: Results for Chinese Dialogue Quality Prediction.

Run	Score type	Mean RSNOD	Mean NMD
0	A-score	0.2154	0.1474
1	A-score	0.2092	0.1369
2	A-score	**0.1992**	**0.1325**
0	S-score	0.1884	0.1305
1	S-score	0.1840	**0.1159**
2	S-score	**0.1758**	0.1166
0	E-score	**0.1545**	**0.1136**
1	E-score	0.1647	0.1222
2	E-score	0.1671	0.1310

Table 2: Results for English Dialogue Quality Prediction.

Run	Score type	Mean RSNOD	Mean NMD
0	A-score	**0.1967**	**0.1327**
0	S-score	**0.1855**	**0.1214**
0	E-score	**0.1742**	**0.1360**

Table 3: Results for Nugget Detection.

Run	Language Type	Mean JSD	Mean RNSS
0	Chinese	0.0700	0.1780
1	Chinese	0.2909	0.3939
0	English	0.0728	0.1830

4.2 Pre-trained Transformer Model

In this section, we discuss the influence of different Transformer models. Thanks to the Hugging Face model library [1], we can easily try variety of pre-trained models. Several of the Chinese and English model pairs we tried and their experimental results are listed in Table4. Notation "-" represents English dialogue datasets not used in the model. "Score-sum" denotes the summation of RSNOD and NMD scores for all A, S, and E scores. The distance scores were transformed by $-log()$ for readability. Thus, the higher the transformed scores, the better the model's effectiveness. The best result is achieved by "roberta-base-finetuned-jd"[2] for Chinese dialogue and "roberta-base" for English dialogue. "roberta-base-finetuned-jd" is a Chinese RoBERTa-Base model fine-tuned with the "JD full" dataset, which consist of user reviews of different sentiment polarities. The experimental results show that using a pre-trained model which data type is closer to the certain task, or making appropriate domain adaption, can significantly improve the accuracy of model predictions.

We also tried multiple ways to unfreeze the parameter layers in the training of Transformer models. All models used in the experiments are based on the original BERT-BASE model with 12-layer bidirectional Transformer blocks and one pooler layer. The experimental results show that too many or too few unfreeze layers will reduce the accuracy of model prediction, and even destroy the final prediction result. The detailed comparison test is shown in Table 5. The best results were obtained when unfreeze last 2 Transformer layers and the pooler layer. Note that if we unfreeze all the parameters of the pre-trained model, GPU will be out of memory, and training will be disabled.

Table 4: Results for different pre-trained Transformer models.

Chinese_model	English_model	Score_sum
bert-base-chinese	-	13.37
Chinese-bert	-	14.59
bert-base-chinese	bert-base-cased	14.56
chinese-roberta-base	roberta-base	16.14
roberta-base-finetuned-jd	roberta-base	**16.27**

Table 5: Results for different unfreeze layers.

unfreeze layers	Score_sum
none	13.37
pooler	14.14
pooler & layer.11	16.14
pooler & layer.10-11	**16.27**
pooler & layer.9-11	15.96
pooler & layer.8-11	13.86
pooler & layer.6-11	13.77
all	-

4.3 Topic Number

In this section, we discuss the influence of topic numbers. According to [1], it will get a better result with the topic number K number between 10 and 20 when dealing with the emotion classification tasks. In this paper, we experimented with various K values and verified the aforementioned conclusions. The distance scores reaches the best when topic number K is 10. The detailed comparison test is shown in Table 6.

Table 6: Results for different topic numbers.

Topic Numbers	Score_sum
5	15.37
10	**16.27**
20	16.11
50	15.79

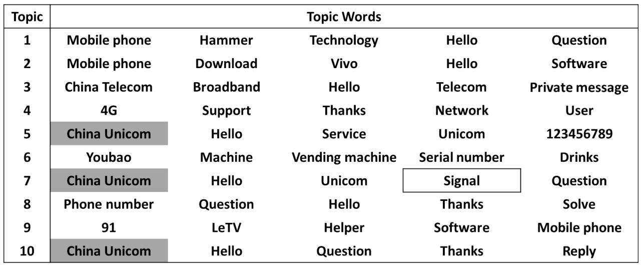
We performed some case studies to see what LDA learned during topic clustering with different topic numbers. Figure 3 is the topic clustering result with the total topic number of 10 on DCH-2. When "China Unicom" (gray blocks, a Chinese telecom company) and "Signal" (the white block with boxe) appear at the same dialogue, the customer satisfaction and task accomplishment scores for dialogue

[1] https://huggingface.co/models

[2] https://huggingface.co/uer/roberta-base-finetuned-jd-full-chinese

quality always tend to be low because China Unicom's signal is admittedly poor. Here are some examples:

```
"id": "4276198835786595",
"sender": "customer",
   "I newly applied mobile phone card of Unicom. It cannot
   be used on non-4g mobile phones and has no signal. The
   old card can be used..."
"sender": "helpdesk",
   "Hello, regarding the situation you have reported,
   please provide your Unicom number... Thank you!"
```



Topic	Topic Words				
1	Mobile phone	Hammer	Technology	Hello	Question
2	Mobile phone	Download	Vivo	Hello	Software
3	China Telecom	Broadband	Hello	Telecom	Private message
4	4G	Support	Thanks	Network	User
5	China Unicom	Hello	Service	Unicom	123456789
6	Youbao	Machine	Vending machine	Serial number	Drinks
7	China Unicom	Hello	Unicom	Signal	Question
8	Phone number	Question	Hello	Thanks	Solve
9	91	LeTV	Helper	Software	Mobile phone
10	China Unicom	Hello	Question	Thanks	Reply

Figure 3: Topic clustering with the total topic number of 10 on DCH-2.

5 CONCLUSIONS

In this paper, we report the details of the dialogue quality prediction network and nugget detection network proposed by the TUA1 team and discuss the results at the NTCIR-16 DialEval-2 task. Both the dialogue quality prediction network and the nugget detection network consist of a feature extractor and a feedforward network. Specifically, The feature extractor employs pre-trained Transformer networks to extract the hidden representations of the dialogue utterances. The feedforward network then concatenates the hidden representations and compresses them through several feedforward layers into the desired dimension, and finally predicts the quality scores and nugget types of the dialogues.

We submit four runs for the dialogue quality prediction subtask and three runs for the nugget detection subtask, which are trained on the mean squared error loss. The evaluation results based on the mean RSNOD and the mean NMD metrics indicate that the proposed dialogue quality prediction network could achieve the best three types of prediction scores among all the participants, while the evaluation results based on the mean JSD and the mean RNSS metrics suggest that our nugget prediction network is also able to reasonably detect if and how the dialogue turns in customer-helpdesk conversations.

After obtaining the evaluation results, we believe that some further work may improve the performance of the model, such as adding stop words to remove. In the DCH-2 dataset, "123456789" was used in place of the phone number that appeared. Removing these stop words may improve the results.

ACKNOWLEDGMENT

This work is supported by JSPS KAKENHI Grant Number 19K20345 and 19H04215.

REFERENCES

[1] Fei Ding, Xin Kang, Shun Nishide, Zhijin Guan, and Fuji Ren. 2020. A fusion model for multi-label emotion classification based on BERT and topic clustering. In *International Symposium on Artificial Intelligence and Robotics 2020*, Vol. 11574. International Society for Optics and Photonics, 115740D.

[2] 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.*

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

[4] Meg Tonkin, Jonathan Vitale, Sarita Herse, Mary-Anne Williams, William Judge, and Xun Wang. 2018. Design methodology for the ux of hri: A field study of a commercial social robot at an airport. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* 407–415.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[6] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[7] Zhaohao Zeng and Tetsuya Sakai. 2021. DCH-2: A Parallel Customer-Helpdesk Dialogue Corpus with Distributions of Annotators' Labels. *arXiv preprint arXiv:2104.08755* (2021).

[8] Yangyang Zhou, Zheng Liu, Xin Kang, Yunong Wu, and Fuji Ren. 2019. TUA1 at the NTCIR-14 STC-3 Task. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies.*