TUA1 at the NTCIR-15 DialEval-1 Task

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

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

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

ABSTRACT

In this paper we report the work of TUA1 team in the dialogue evaluation (DialEval-1) task of NTCIR-15, for the Chinese dialogue quality (DQ) and nugget detection (ND) subtasks. In the proposed method, first we employ a pre-trained BERT network for feature extraction from a dialogue sequence, and feed these feature vectors to a Bi-LSTM network together with a speaker embedding which is learned to separate the customer and helpdesk semantically. Then we feed the output, which is a sequence of semantic vectors, into a self-attention network, where a few attention heads are learned to assign evaluation weights over the sequence of vectors and summarize them into several high-level semantic vectors. Finally we concatenate these high-level semantic vectors and put them through several feed-forward neural network layers to finally predict the dialogue quality scores. We train the networks based on the criteria of mean squared error and Sinkhorn divergence respectively for dialogue quality prediction and on that of mean sequared error for nugget detection. The results suggest that the proposed method is promising in learning a dialogue quality prediction system for generating very close predictions to the human annotators.

TEAM NAME

TUA1

SUBTASKS

Dialogue Quality (Chinese)
Nugget Detection (Chinese)

1 INTRODUCTION

Learning to evaluate the dialogue quality is a new research topic, which could provide useful information for tuning the dialogue systems in an efficient manner. The TUA1 team participates in the Chinese language dialogue quality (DQ) and nugget detection (ND) subtasks of DialEval-1. The detailed task descriptions for dialogue quality and nugget detection can be found in the overview paper [6]. This work is a continuation of our previous works in text conversation [3, 5, 7] and reading comprehension [8].

In the dialogue quality subtask, we incorporate the pre-trained BERT network, a Bi-LSTM network, a self-attention network, and a Feed-forward network into our dialogue quality prediction (DQP) network, which takes the interactive dialogue sequence from a customer and a helpdesk and learns to evaluate the dialogue quality of the helpdesk. Our network simultaneously evaluates three types of the dialogue quality, which are 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 four components, which are mentioned in the dialogue quality subtask, to construct our nugget detection network. In each turn, we predict the nugget labels corresponding to sender and give the probability distribution with respect to each label. Four probability scores related to {CNUG0, CNUG, CNUG*, CNaN} are given for the customer turn while three including {HNUG, HNUG*, HNaN} are given for the helpdesk turn.

Both the dialogue quality predictions and the nugget detection predictions are evaluated with two cross-bin metrics, which are the root symmetric normalised order-aware divergence (RSNOD) and normalised match distance (NMD) [6]. These metrics gives the measurement of similarity of two probabilities, which in the case of dialogue quality prediction measures the similarity of the ground truth probability of the dialogue quality and the model probability. Higher similarities of the ground truth probabilities and the predicted probabilities correspond to smaller values of in RSNOD and NMD. Among all participants of the Chinese dialogue evaluation subtasks, the TUA1 team achieves the highest similarities in both RSNOD and NMD metrics for the dialogue effective prediction, the highest similarity in RSNOD and the second highest similarity in NMD for the task accomplishment prediction. For the TUA1 customer satisfaction prediction and nugget detection prediction, the RSNOD and NMD metrics also indicate reasonable results.

The rest of this paper is organized as follows. Section 2 and 3 describe the proposed dialogue quality prediction and nugget detection networks of TUA1, respectively. Section 4 reports our submissions and analyses the results. Section 5 concludes our work.

2 DIALOGUE QUALITY PREDICTION NETWORK

We incorporate four major components to construct a neural network for dialogue quality prediction, that is, the pre-trained BERT network, the Bi-LSTM network, the self-attention network, and the feed-forward network. As shown in Fig. 1, the pre-trained BERT network locates at the bottom of the dialogue quality prediction (DQP) network. BERT is a deep neural network for natural language understanding proposed by [1], in which the raw text pieces can be transformed into sequences of feature vectors by a multi-layer bidirectional Transformer network. The BERT model that we use is the pre-trained base Chinese model based on the HuggingFace's Transormers library [4].

As the input of the BERT network, the dialogues are split by turns, each of which contains one or more utterances from either a customer or a helpdesk. For each turn we combine all the utterances into a long utterance with the space character in between, which corresponds to all the words of a speaker in one dialogue turn. The speaker identity, that is either the customer or the helpdesk, is recorded in each turn for the speaker embedding. We tokenize the combined utterances and concatenate these tokenized utterances into an even longer token sequence, with a [SEP] token in between of the combined utterances and a [CLS] token at the very beginning of the sequence. The pre-trained BERT network takes in a sequence of these tokens and generates a sequence of feature vectors, each

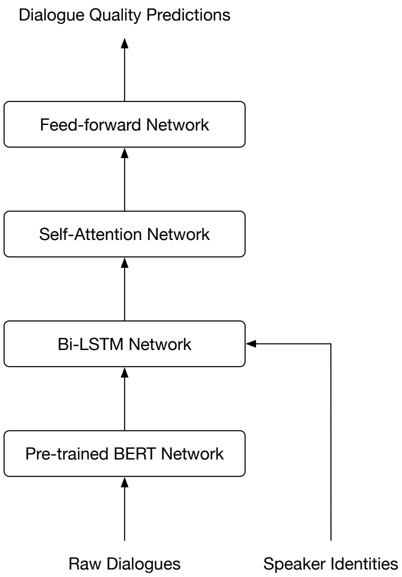

Figure 1: The structure of dialogue quality prediction (DQP) network.

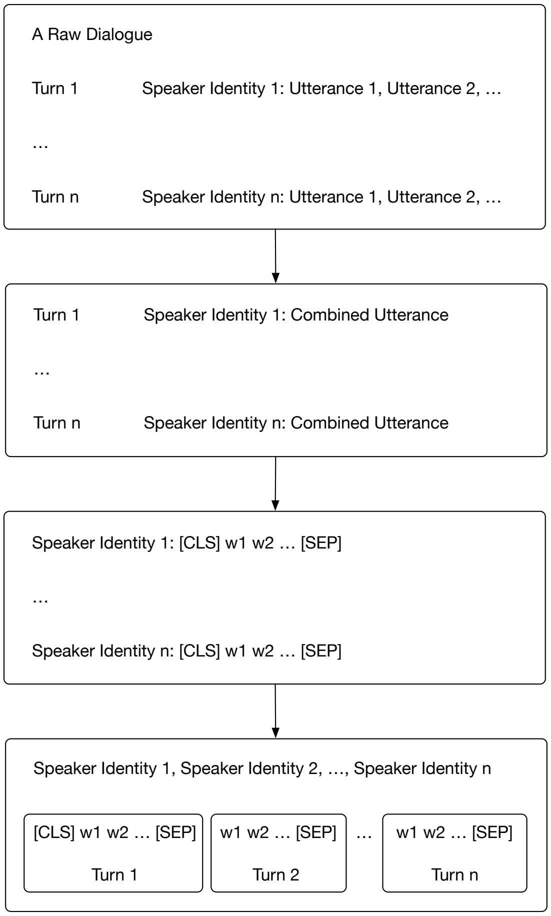
A Raw Dialogue
Turn 1
Speaker Identity 1: Utterance 1, Utterance 2, ...
...
Turn n
Speaker Identity n: Utterance 1, Utterance 2, ...
Turn 1
Speaker Identity 1: Combined Utterance
...
Turn n
Speaker Identity n: Combined Utterance
Speaker Identity 1: [CLS] w1 w2 ... [SEP]
...
Speaker Identity n: [CLS] w1 w2 ... [SEP]
Speaker Identity 1, Speaker Identity 2, ..., Speaker Identity n
[CLS] w1 w2 ... [SEP]
Turn 1
w1 w2 ... [SEP]
Turn 2
...
w1 w2 ... [SEP]
Turn n

Figure 2: The process of model input for a raw dialogue.

of which represents the meaning of a token in the dialogue. The detailed process are illustrated in Fig. 2.

The feature vectors generated by BERT is then fed into a Bi-LSTM network as the second component from the bottom of Fig. 1. For each token we also embed its speaker identity with a feature vector of the same size of dimension as the BERT output, and we add the speaker embedding to the BERT output as a compound input to the Bi-LSTM network. The Bi-LSTM network sequentially scans the input feature vectors in both the forward and the backward manners and integrates them to its output.

The self-attention network takes the sequential features from Bi-LSTM, as shown in Fig. 1, learns to assign attentional weights on these features, and summarizes them into a compact feature vector. We take a scaled dot-product attention to summarize an input sequence. And since the self-attention network consists of multiple attention heads, we can learn to summarize the input feature vectors into a compact feature vector by different attentions separately.

The compact feature vectors from different attention heads are concatenated and fed to several forward neural network layers as the top component of Fig. 1, which generate the predictions over the dialogue qualities. For each type of dialogue quality, that is, the type A-score of the task accomplishment, the type S-score of the customer satisfaction, and the type E-score of the dialogue effectiveness, we use a separate set of feed-forward layers for the dialogue quality prediction. The network generates probabilities over the quality label set $\Gamma = \{-2, -1, 0, 1, 2\}$ for every quality type. More illustration of the quality labels can be found in [6]. In this way, the network can distinguish different dialogue quality predictions with separate sets of parameters.

We employ two criteria for evaluating the training loss, which are the mean squared error and the Sinkhorn divergence [2]. Given a batch of n training examples (x_i, y_i), where x_i is a dialogue and y_i is the ground truth quality scores given by l human annotators to the dialogue, the model generates $\hat{y}_i$ as the dialogue quality predictions.

For the training loss of mean squared error, the model generates three distributions $\hat{y}_i^A$, $\hat{y}_i^S$, and $\hat{y}_i^E$ in $\hat{y}_i$ 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

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

For the training loss based on Sinkhorn divergence, the model generates a set of score samples for each of the A, S, E quality types of the dialogues. We take the Sinkhorn divergence between the sample distribution and the human annotation distributions, for each quality type, and the training loss can be given by

$$\mathrm{loss}(y, \tilde{y}) = \sum_{\kappa \in \{A,S,E\}} \frac{1}{n} \sum_{i=1}^{n} \mathrm{Div}(y_i^\kappa, \tilde{y}_i^\kappa) \qquad (2)$$

in which y_i^κ indicates the ground truth annotations for a dialogue i on quality type κ, $\tilde{y}_i^\kappa$ indicates the corresponding samples generated by the model, and Div calculates the Sinkhorn divergence of these two sets of samples.

3 NUGGET DETECTION NETWORK

The similar network architecture to dialogue quality subtask is built for dealing with the nugget detection subtask. Except speaker identity part, we integrate the pre-trained BERT network, the Bi-LSTM network, the self-attention network and the feed-forward network into our nugget detection network for predicting the nugget label with the corresponding probability score.

We divide all the utterances into two parts named sender part and helpdesk part in each turn, that is, utterances extracted from either sender or helpdesk are trained separately. We feed the utterances to the pre-trained BERT network to get the tokenized sequences and construct the feature vector. Then, the feature vectors are used as input of Bi-LSTM network. The attention weights are learned from output and hidden states given by Bi-LSTM. Finally, we calculate the probability distribution by using the feed-forward network for nugget labels including four labels with respect to sender utterances while three labels with respect to helpdesk utterances in each turn.

4 EXPERIMENT

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

- RUN0: DQP network with the mean squared error loss.
- RUN1: DQP network with the Sinkhorn divergence loss for single-label probabilities.
- RUN2: DQP network with the Sinkhorn divergence loss for multi-label probabilities.

The difference in RUN1 and RUN2 is that the generated score samples in RUN1 are considered as sets of Γ probabilities, each of which represents the probability for a single quality label γ, while those generated in RUN2 are considered as sets of probabilities, each of which represents the probability distribution over multiple ($\|\Gamma\|$) quality labels.

We report the results of three types of dialogue quality predictions in Table 1 to 3, the evaluation of which are based on the mean Root Symmetric Normalised Order-aware Divergence (RSNOD) metric and the mean Normalised Match Distance (NMD) metric. The hyper script numbers of (1) and (2) indicate the ranking of our runs among all participant runs in the task. Both metrics calculate the distance of probability distributions of the model generations and the ground truth probabilities, over the quality labels in Γ. The NMD metric calculates the difference of probabilities over the incrementally constructed Γ subsets, that is, $\{-2\}, \{-2, -1\}, \cdots \Gamma$, from the model generation and the ground truth. The RSNOD metric considers the difference between every combination of two probability bins, where each bin represents the probability of a $\gamma \in \Gamma$, from the model prediction and the ground truth as well as the ordinal distance of the two bins in Γ. The overview paper [6] by running both mean RSNOD and mean NMD on all submissions of the dialogue quality subtask suggests that the difference of the two evaluation metrics is not statistically significant.

RUN	Mean RSNOD	Mean NMD
0	0.2136	**0.1396**$^{(2)}$
1	0.2484	0.1510
2	**0.2102**$^{(1)}$	0.1412

Table 1: The A-score Results for Chinese Dialogue Quality Prediction.

RUN	Mean RSNOD	Mean NMD
0	0.2053	0.1322
1	0.2302	0.1397
2	**0.2024**	**0.1310**

Table 2: The S-score Results for Chinese Dialogue Quality Prediction.

RUN	Mean RSNOD	Mean NMD
0	**0.1615**$^{(1)}$	**0.1144**$^{(1)}$
1	0.1810	0.1253
2	0.1617	0.1187

Table 3: The E-score Results for Chinese Dialogue Quality Prediction.

Run	Mean JSD	Mean RNSS
0	0.0859	0.1892

Table 4: The Results for Chinese Dialogue Nugget Detection.

Among all participant runs, our model generates the type A dialogue quality predictions of the highest similarity to the ground truths based on the mean RSNOD metric (0.2102) and of the second highest similarity to the ground truths based on the mean NMD metric (0.1396), as is shown in Table 1. For the type S dialogue quality prediction, RUN2 results receive the highest similarities based on the mean RSNOD and the mean NMD metrics among our three runs. Finally, for the type E dialogue quality, our model generates the most similar predictions to the ground truth probabilities among all participant runs. These results suggest that the proposed dialogue quality prediction network is well competent among all the participants for automatically evaluating how good the dialogues are in the customer-helpdesk conversations.

Next, we report the result of our nugget detection in Table 4, which is evaluated based on the mean Jensen-Shannon Divergence (JSD) metric and the mean Root Normalised Sum of Squares (RNSS) metric, respectively. Both metrics evaluate the bin-by-bin difference between the model generations and the ground truth probabilities, over the nugget labels. Since the label sets for the customer dialogues and for the helpdesk dialogues are different, the reported scores reflect the mean of these evaluations over the customer dialogues and the helpdesk dialogues. Our nugget detection run receives the scores of 0.0859 and 0.1892, respectively for the mean JSD evaluation and the mean RNSS evaluation. As discussed in the overview paper [6], there are no significant differences between these scores and the best scores, that is 0.0674 in terms of mean JSD and 0.1633 in terms of mean RNSS. This result suggests that the proposed nugget detection network is suitable to predict if and in which way the dialogue turns are helpful towards the problem solving in the customer-helpdesk conversations.

5 CONCLUSIONS

In this paper, we report the details of the dialogue evaluation method proposed by the TUA1 team and discuss the results at the NTCIR-15 DialEval-1 task. Both the dialogue quality prediction network and the nugget detection network consist of four functionally and structurally distinct networks. Specifically, the pre-trained BERT network extract feature sequentially from a raw dialogue, the Bi-LSTM network scans and integrates the features and the speaker embeddings in two directions, the self-attention network learns to summarize the vector of features with several attention heads, and the feed-forward network maps the summarized features into the dialogue quality predictions. We submit three runs for the dialogue quality prediction subtask, wwhich are trained on the mean squared error loss and two versions of the Sinkhorn divergence loss. Meanwhile, we submit one run for the nugget detection subtask, which is 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 generate three types of predictions which are very close to the human annotations, while the evaluation results based on the mean JSD and the mean RNSS metrics suggest that our nugget prediction network is able to reasonably detect if and how the dialogue turns are helpful in customer-helpdesk conversations.

ACKNOWLEDGMENT

This research has been supported by JSPS KAKENHI Grant Number 19K20345.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[2] Aude Genevay, Gabriel Peyré, and Marco Cuturi. 2018. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*. 1608–1617.

[3] Tianhao She, Xin Kang, Shun Nishide, and Fuji Ren. 2018. Improving LEO robot conversational ability via deep learning algorithms for children with autism. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS2018)*. IEEE, 416–420.

[4] 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. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

[5] Yunong Wu, Xin Kang, Kenji Kita, and Fuji Ren. 2017. TUA1 at NTCIR-13 Short Text Conversation 2 Task. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-13)*. 211–214.

[6] Zeng Zhaohao, Kato Sosuke, Sakai Tetsuya, and Kang Inho. 2020. Overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) Task. In *Proceedings of NTCIR-15.*

[7] 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 (NTCIR-14)*. 338–345.

[8] Zheng Zhou, Xin Kang, Nishide Shun, and Fuji Ren. 2019. Capture Important Information for Reading and Reasoning by A Two-Pass Attention Mechanism in Dynamic Memory Network. In *Proceedings of 2019 6th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS2019)*. IEEE.