CUIS at the NTCIR-14 STC-3 DQ Subtask[⋆]

Kai Cong and Wai Lam

The Chinese University of Hong Kong, Hong Kong, China
{congkai,wlam}@se.cuhk.edu.hk

Abstract. We present our model participated in the NTCIR-14 Short Text Conversation-3 [13] Dialogue Quality (DQ) subtask. The DQ subtask is to assign quality scores to each customer-helpdesk dialogue in terms of three criteria: task accomplishment (A), customer satisfaction (S), and efficiency (E). Each dialogue is composed of posts by customer or helpdesk and each post consists of turns of sentences, which naturally forms a hierarchical structure. In this work, we consider this problem as a document classification task and have attempted various ways until we finalize our model into a hierarchical attention network with bidirectional GRUs and Google's BERT sentence embedding, which stands for Bidirectional Encoder Representations from Transformers. Also, our model is augmented with a sender-aware encoding to differentiate the contributions from the customer side and the helpdesk side. According to the official STC evaluation in the test dataset, our proposed system achieves among the top performance teams in the English dataset.

Keywords: Document Classification · Hierarchical Attention · Bidirectional GRU · BERT Embedding.

Team Name: CUIS.

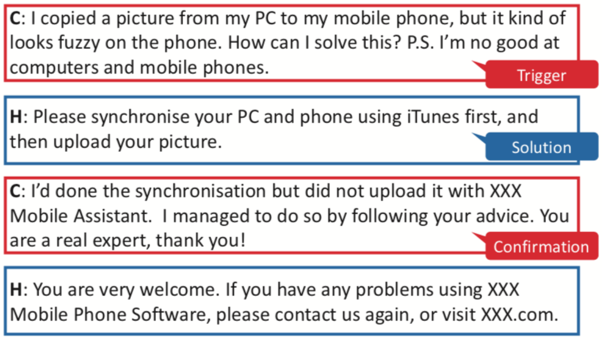
Subtasks: Dialogue Quality (DQ) Subtask (for Chinese and English).

1 Intoduction

The motivation behind the DQ subtask is to build a reliable automatic evaluation method for task-oriented dialogues [10] so that we can construct and efficiently adjust the automatic helpdesk system [5]. Figure 1 shows an example of a customer-helpdesk conversation. It can be observed that it was initiated by the customer's report on a specific problem she faced, which we call a trigger. This is an example of a successful dialogue, because Helpdesk provides a practical solution to the problem, and the Customer acknowledges that the problem has been solved. Unlike traditional closed-domain dialogues, the automatic

[⋆] The work described in this paper is substantially supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Codes: 14203414) and the Direct Grant of the Faculty of Engineering, CUHK (Project Code: 4055093).

helpdesk may have to handle requests from different areas, making it impossible to solve problems through predefined slot filling schemes. Instead, many researchers have pursued deep learning methods to build such general chatbots, which in turn requires more flexible and appropriate evaluation tools [6].



C: I copied a picture from my PC to my mobile phone, but it kind of looks fuzzy on the phone. How can I solve this? P.S. I'm no good at computers and mobile phones.

H: Please synchronise your PC and phone using iTunes first, and then upload your picture.

C: I'd done the synchronisation but did not upload it with XXX Mobile Assistant. I managed to do so by following your advice. You are a real expert, thank you!

H: You are very welcome. If you have any problems using XXX Mobile Phone Software, please contact us again, or visit XXX.com.

Fig. 1. An example of a dialogue between the customer(C) and helpdesk(H) [10]

In the STC3 DQ subtask, we aim to evaluate the quality of task-oriented, multi-round, textual dialogue systems in human-like manner. Hence, the quality scores our model predicts should be as close as possible to the human annotations in terms of task accomplishment, customer satisfaction, and efficiency for each dialogue. We tackle this problem as a multi-label document classification task as we aim at mapping the document representation [9] to some predicted quality scores.

We could apply the attention mechanism [1] in document classification. In the DQ dataset, dialogues are composed of turns (either by the customer or helpdesk), while turns are composed of sentences. Thus it has a natural hierarchical structure. Therefore, we adopt a two-level hierarchical neural network to encode the document: (i) the sentences are used as the granularity input network to extract sentence-level features to get the feature vectors representing turns i.e. turn-level vectors; (ii) then turn-level vectors are used to project the turn-level features from the input network to get the final dialogue-level representation. Precisely the dialogue representation is expressed through a fully-connected linear layer and a softmax function for the final prediction. In the turn-level encoding, we adopt the sender-aware encoding to project the turns of customer and helpdesk into separate space. In order to assign different weights to different sentences and turns, an attention mechanism with hierarchical architecture is designed to improve the performance. In fact, the importance of sentences and turns are highly context dependent, i.e. the same sentence or turn may be differentially important in different context. To incorporate such consideration, our model includes two levels of attention mechanisms [11] — one at the sentence level and one at the turn level — which let the model learn to pay different attention to individual sentences and turns when constructing the representation of the dialogue. In addition, we adopt the BERT sentence embedding as the sentence-level input to reduce the imbalance in the dataset and bring external information. In our experiment, all the techniques demonstrate improved performance in our model and the final version has achieved good results in the STC3 official evaluation, especially in the English test dataset.

2 Model Description

The overall architecture of our model is shown in Fig. 2. We employ the Hierarchical Attention Networks (HAN) [12] as our basic model. In addition, we employ the BERT pre-training for providing the initial sentence embeddings and the sender-aware encoding scheme to enrich the information in turn embeddings. We describe the details of different components in the following sections.

2.1 BERT Preprocessing

Recently, the BERT model [4] released by Google AI team has aroused great response in the NLP community, which is regarded as a milestone in the text representation. By training language models on 3.3 billion text corpus and fine-tuning them on different downstream tasks, the authors have achieved state-of-the-art results in 11 NLP tasks, and some of the results have been greatly improved compared with the previous best results. There are many studies using pre-trained linguistic representations to complete downstream NLP tasks. It can be summarized as two types, namely, feature-based and fine-tuning. As a fine-tuning method, the authors propose an improved scheme: Bidirectional Encoder Representations from Transformers (BERT) with a new objective function and increased sentence-level information.

One point we wish to convey is that models with huge amount of parameters in large-scale tasks can bring about substantial increase in effect, but this is the first time we have seen such models can also bring significant increase in small tasks as long as they are adequately pre-trained. Thus, in our model, we apply the pre-trained BERT sentence-embedding to produce the sentence-level embeddings for our task.

2.2 Hierarchical Attention Network

Our Hierarchical Attention Networks (HAN) consists of several parts: a sentence sequence encoder, a sentence-level attention layer, a turn sequence encoder and a turn-level attention layer.

GRU-based Sequence Encoder The encoder of HAN is GRU-based. For every input word, the encoder outputs a vector and a hidden state, and uses the

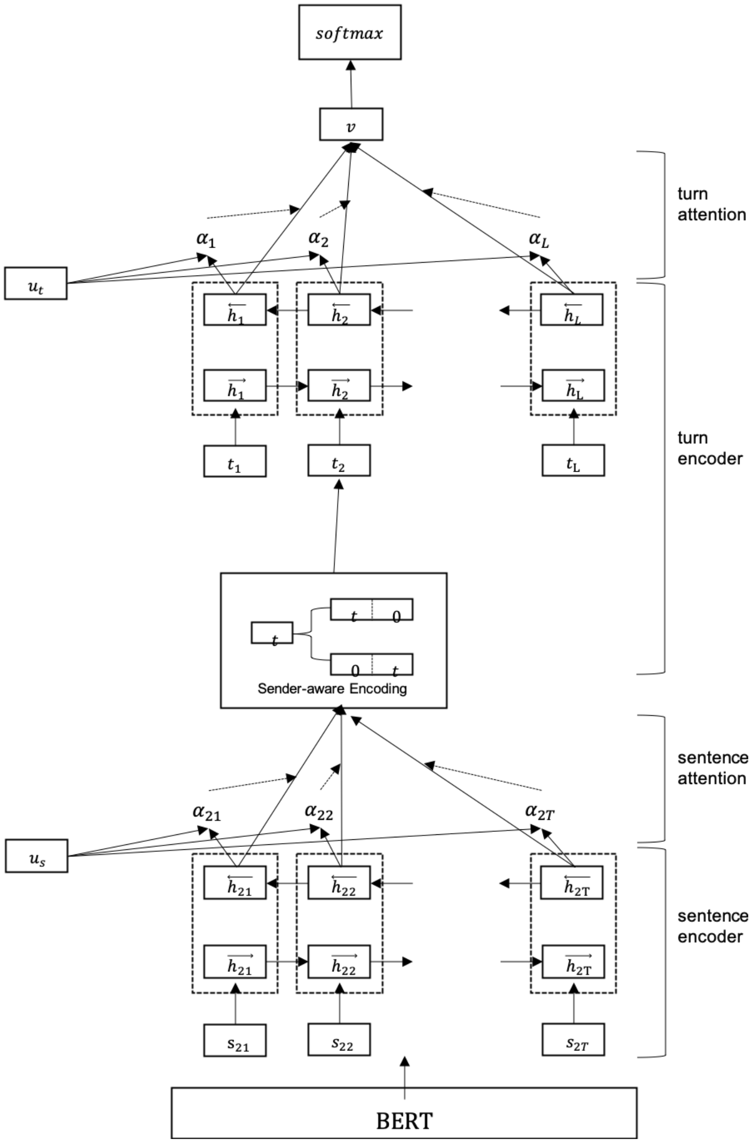


Fig. 2. The architecture of our model.

hidden state for the next input word. There are two types of gates: the reset gate r_t and the update gate z_t. They together control how information is updated to the state. The formulation for GRU [3] is as follows:

$$r_t = \sigma(W_r s_{t-1} + U_r x_t + b_r), \tag{1}$$

$$z_t = \sigma(W_z s_{t-1} + U_z x_t + b_z), \tag{2}$$

$$\tilde{s_t} = \phi(W(r_t \odot s_{t-1}) + Ux_t + b), \tag{3}$$

$$s_t = z_t \odot s_{t-1} + (1 - z_t) \odot \tilde{s}_t. \tag{4}$$

Here r_t is the reset gate which controls how much the past state contributes to the candidate state. If r_t is zero, then it forgets the previous state.

Hierarchical Attention In the DQ subtask, every dialogue to be classified can be divided into several turns and every turn can be segmented into several sentences. So the first part of the hierarchical attention model is to encode each sentence embedding with bidirectional GRUs [8] and the formulae are given below:

$$\begin{aligned} x_{it} &= S_e s_{it}, t \in [1, T], \\ \overrightarrow{h_{it}} &= \overrightarrow{GRU}(x_{it}), t \in [1, T], \\ \overleftarrow{h_{it}} &= \overleftarrow{GRU}(x_{it}), t \in [1, T]. \end{aligned} \tag{5}$$

But for sentences in a turn, not every sentence is useful for forming the turn representation and the final prediction. Sometimes we pay more attention to ‘good’ and ‘sad’ words with emotional meaning [2]. In order to enable the recurrent neural network to automatically focus ‘attention’ on these words, we adopt a token-based attention model, whose formulae are given as follows:

$$\begin{aligned} u_{it} &= tanh(W_w h_{it} + b_w), \\ \alpha_{it} &= \frac{exp(u_{it}^T u_w)}{\sum_t exp(u_{it}^T u_w)}, \\ t_i &= \sum_t \alpha_{it} h_{it}. \end{aligned} \tag{6}$$

Firstly, the output of bidirectional RNN is transformed by a linear layer, then the importance of each word is calculated by the softmax function. Finally the expression of each sentence is obtained by weighted average of the output of bidirectional RNNs.

The turn-level attention model has the exactly same structure as the sentence-level one so we omit some details. After this two-level attention network, we obtain a more precise representation for each dialogue to be evaluated through different linear classifiers for different criteria.

2.3 Sender-aware Encoding

In the DQ dataset, utterance blocks are the basis of the dialogue, formed by merging all consecutive posts by the same utterer. Thus we use the sender-aware encoding during the turn-level to enrich the turn-level embedding and differentiate the input from the customer side or the helpdesk side. Here we simply adopt the way in the STC-3 baseline model which concatenates the input with an array of zeros either at the head or the tail as follows:

$$t = \begin{cases} [\mathbf{0}|\mathbf{t}] & \text{if} \quad \text{Customer}, \\ [\mathbf{t}|\mathbf{0}] & \text{if} \quad \text{Helpdesk}. \end{cases} \tag{7}$$

In the experiments, the sender-aware encoding scheme is able to bring clear improvement.

2.4 Document Classification

After HAN, we are able to obtain the document vector v, which is a high level representation of the dialogue and can be used as features for dialogue classification via different linear classifiers:

$$p = softmax(W_c v + b_c). \tag{8}$$

We use the negative log likelihood of the correct labels as the training loss:

$$L = -\sum_{d} \log \ p_{dj}, \tag{9}$$

where j is the corresponding score (A,E,S score respectively) for the dialogue d.

3 Experiments

3.1 Datasets

We evaluate the effectiveness of our model on both English dataset and Chinese dataset of dialogues provided by NTCIR.

The Chinese dataset contains 4,090 (3,700 for training + 390 for testing) customer-helpdesk dialogues which are crawled from Weibo, a major Chinese social media. All of these dialogues are labeled with A,E,S scores separately by 19 annotators.

The English dataset contains 2,062 dialogues (1,672 for training + 390 for testing), which are manually translated from a subset of the Chinese dataset. The English dataset shares the same annotations with the Chinese dataset.

3.2 Evaluation

We adopt the NTCIR official evaluation metrics for measuring the performance. Through investigation using artificial distributions as well as real ones from a dialogue evaluation task, it has been demonstrated that the two cross-bin measures, namely, the Normalised Match Distance (NMD; a special case of the Earth Mover's Distance) and the Root Symmetric Normalised Order-aware Divergence (RSNOD), are indeed substantially different from the bin-by-bin measures and better for this DQ subtask. Furthermore, RSNOD lies between the popular bin-by-bin measures and NMD in terms of how it behaves [7]. Therefore both RSNOD and NMD are used as the evaluation metrics.

3.3 Results and Analysis

We have investigated the prediction quality in the English dataset during training of following models with respect to the baseline Bag-of-Words (BoW) LSTM model monitored by Tensorboard:

- orange curve: Baseline BoW+LSTM
- blue curve: HAN
- red curve: HAN+BERT

From the graph of the RSNOD score for different models, we observe that our model has achieved a significant improvement. For all the scores: A-score(Task accomplishment), S-score(Customer Satisfaction) and E-score(Dialogue Efficiency), we can see clear gaps between different models. Both the HAN architecture and the BERT pre-training contribute a lot to the improvement of the final prediction.

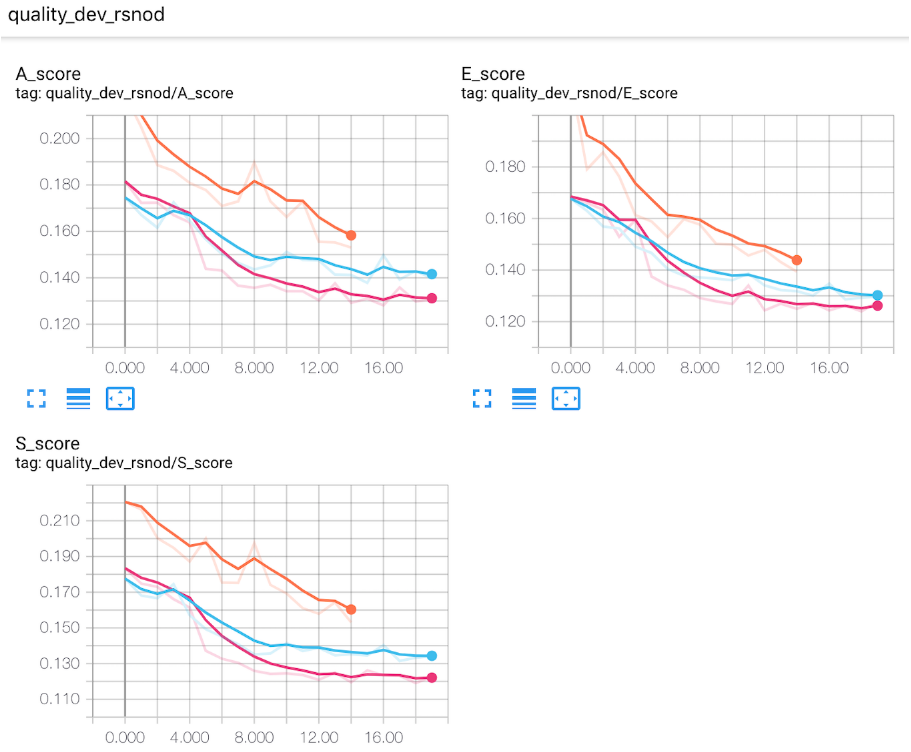


Fig. 3. RSNOD score for different models during training

Besides, we notice that the BERT pre-training performs extremely well on the English dataset. One reason is that the English dataset contains less training data and the imbalance is more serious. Then BERT is able to bring external information and has stronger capability of dealing with this issue. In contrast, since BERT is mainly trained by Google on the English dataset, it might lack the precise granularity we need in the Chinese sentence embeddings.

Table 1. English Dialogue Quality (A-score) Results.

Run	Mean RSNOD	Run	Mean NMD
BL-lstm	0.1320	BL-lstm	0.0896
CUIS-run0	0.1360	CUIS-run0	0.0901
SLSTC-run2	0.1370	SLSTC-run1	0.0908
SLSTC-run1	0.1391	SLSTC-run2	0.0933
WIDM-run0	0.1411	WIDM-run0	0.0939
WIDM-run1	0.1411	WIDM-run1	0.0939
SLSTC-run0	0.1493	SLSTC-run0	0.1017
BL-uniform	0.2478	BL-popurlarity	0.1677
BL-popurlarity	0.2532	BL-uniform	0.1855

We submit the BERT+HAN model as our final CUIS version with the hidden size of GRU as 150 and attention size as 300. According to the overview result from official NTCIR-STC3 statistics, our former analysis has been verified. Our model achieves among the top performance teams in the DQ English dataset. According to the official STC-3 evaluation for English DQ dataset, we achieve a very high ranking among all the teams for A-score estimated by both RSNOD and NMD metrics as show in Table 1 [13]. In addition we achieve a high ranking in S-score and E-score using the NMD metric.

4 Conclusion

In this DQ subtask, we design a modified hierarchical attention networks (HAN) for classifying customer-helpdesk dialogue quality. The sentence-level embedding comes from the BERT pre-training, which brings the external information. Our model progressively builds a dialogue vector by aggregating important sentences into turn vectors and then aggregating important turn vectors into dialogue vectors. Experimental results demonstrate that our model performs well.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proc. ICLR (2015)
2. Chen, H., Sun, M., Tu, C., Lin, Y., Liu, Z.: Neural sentiment classification with user and product attention. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1650–1659 (2016)

3. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: Proc. NIPS, Montreal, QC, Cannada. pp. 1–9 (2014)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Lowe, R., Noseworthy, M., Serban, I.V., Angelard-Gontier, N., Bengio, Y., Pineau, J.: Towards an automatic turing test: Learning to evaluate dialogue responses. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1116–1126. Vancouver, Canada (Jul 2017)
6. Noseworthy, M., Cheung, J.C.K., Pineau, J.: Predicting success in goal-driven human-human dialogues. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. pp. 253–262 (2017)
7. Sakai, T.: Comparing two binned probability distributions for information access evaluation. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1073–1076. ACM (2018)
8. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997)
9. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1422–1432 (2015)
10. Tetsuya, S., Zhaohao, Z., Luo, C.: Evaluating helpdesk dialogues: Initial considerations from an information access perspective. NTCIR-13(NL) pp. 1–10 (Jan 2016)
11. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)
12. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
13. Zeng, Z., Kato, S., Sakai, T.: Overview of the ntcir-14 short text conversation task: Dialogue quality and nuggest detection subtasks. In: Online Proceedings of NTCIR-14 (to appear,2019)