TUA1 at the NTCIR-16 DialEval-2 Task

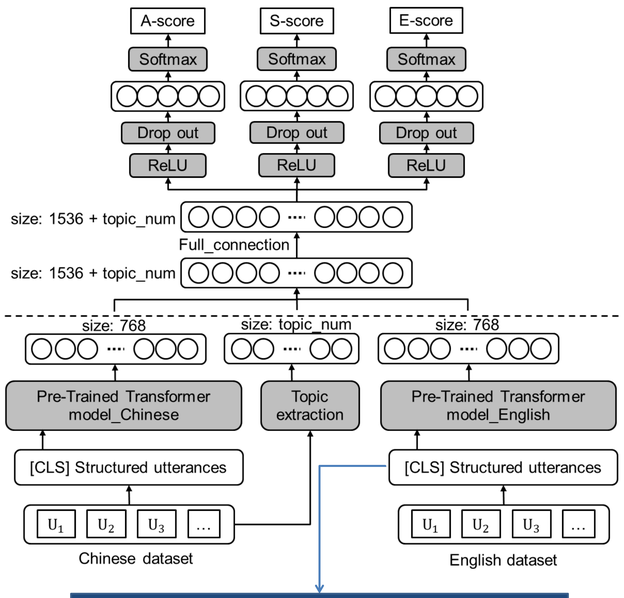
Fei Ding[1], Xin Kang[1], Yunong Wu[2], and Fuji Ren[1]
[1] Faculty of Engineering, Tokushima University, Japan
[2] CDXT Medical Technology, China

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. 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.

The TUA1 team participates in the Dialogue Quality and Nugget Detection subtasks of NTCIR-16 DialEval-2 task. This work is a continuation of our previous works in text conversation and text emotion analysis.

Methodology



An example of the structured utterance

```
"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]
```

The dialogue quality prediction network mainly consisted of a feature extractor and a feedforward network. The two parts are shown on the left, separated by a dotted line in between.
Firstly, in order to better represent the structure of a set of dialogue, we preprocess the input dialogue sequence.
Secondly, we employ pre-trained Transformer networks to extract the hidden representations of the structured dialogue input. Topic information is also extracted as a part of the hidden representation of the input dialogue.
Thirdly, the feedforward network is located above the dotted line, which is arranged as follows: a full connection layer, an activation function, a dropout layer, a linear dimension reduction layer, and a softmax function.
Finnally, the network get the probabilities over the quality label set for every quality type. We employ the mean squared error (MSE) loss for evaluating the training loss.

In nugget detection subtask, we used the same model as the dialogue quality task. the differnet is, 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.

Experiments

Results for different topic numbers.

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

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	-

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**

TUA1 team submits three runs for the Chinese DQ subtask, one run for the English DQ subtask, two runs for the Chinese ND subtask, and one run for the English ND subtask.

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.

Case study

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

"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!"

What does LDA exactly learn from the Topic clustering?

Topic number = 10
Max topic words =5

When ``China Unicom'' (gray blocks) and ``Signal'' (the white block with boxe) appear at the same dialogue, the customer satisfaction and task accomplishment scores for dialogue quality always tend to be low because China Unicom's signal is admittedly poor.

Conclusions

Our contributions:
✓A novel and effective architecture for Dialogue Quality and Nugget Detection tasks.
✓The best scores for RSNOD and NMD metrics in both Chinese and English Dialogue Quality subtasks.

Future works:
➢Extend stop words library, such as "123456789".
➢Further study the effect of different Transformer levels.

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

Poster maker: Ding Fei, E-mail: 248362831@qq.com