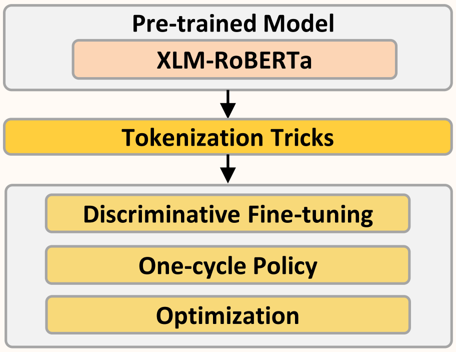
IMNTPU Dialogue System Evaluation at NTCIR-16 DialEval-2 Dialogue Quality and Nugget Detection

Ting-Yun Hsiao[1], Yung-Wei Teng[1], Pei-Tz Chiu[1], Mike Tian-Jian Jiang[2] and Min-Yuh Day[1, *]
*myday@gm.ntpu.edu.tw
[1]Information Management, National Taipei University, New Taipei City, Taiwan
[2]Zeals Co., Ltd. Tokyo, Japan

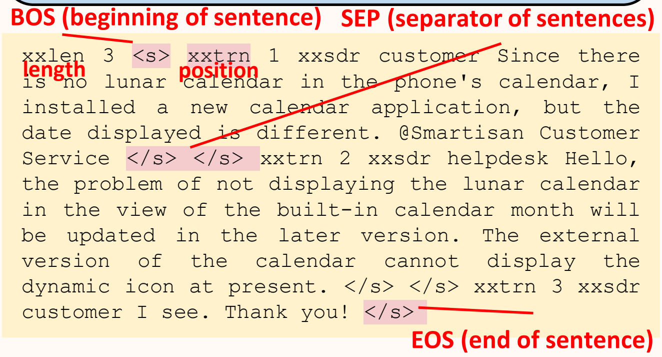
A surge in interest in the evaluation of the quality of chatbot conversation has been observed in recent years. We performed Dialogue Quality (DQ) and Nugget Detection (ND) subtasks in Chinese and English. However, the majority of existing conventional approaches are based on the long short-term memory (LSTM) model. The paper suggests a method for assisting customers in resolving problems. This subtask aims to automatically determine the status of dialogue sentences in a dialogue system's logs. In conversation tasks, we developed fine-tuning methodologies for the transformer model. To evaluate and show the concept, we created a wide framework for testing and displaying the XLM-RoBERTa model's performance on conversational texts. Finally, the experimental findings of the two subtasks demonstrated the efficacy of our strategy. The experimental findings for the DialEval-2 task showed that the suggested method's performance is reasonably equal to that of the LSTM-based baseline model. The main contribution of our study is our suggestion of two crucial elements, namely, tokenization methods and fine-tuning procedures, to increase the conversation quality and nugget identification subtasks in dialogue assessment.

IMNTPU Research Architecture



Pre-trained Model
XLM-RoBERTa
Tokenization Tricks
Discriminative Fine-tuning
One-cycle Policy
Optimization

Tokenization Tricks



BOS (beginning of sentence) SEP (separator of sentences)

```
xxlen 3 <s> xxtrn 1 xxsdr customer Since there
is no lunar calendar in the phone's calendar, I
installed a new calendar application, but the
date displayed is different. @Smartisan Customer
Service </s> </s> xxtrn 2 xxsdr helpdesk Hello,
the problem of not displaying the lunar calendar
in the view of the built-in calendar month will
be updated in the later version. The external
version of the calendar cannot display the
dynamic icon at present. </s> </s> xxtrn 3 xxsdr
customer I see. Thank you! </s>
```

EOS (end of sentence)

Fine-tuning Techniques

Discriminative

- Different layers capture different types of information. They should be fine-tuned to different extents.
- The amount of fine-tuning required increases gradually as we move towards the last layer.

One-cycle Policy

- Slanted triangular learning rates
  - Intuition for adapting parameters to task-specific features.
  - The model should converge quickly to a suitable region and then refine its paramaters.

Performance

NTCIR-16 DialEval-2 Chinese Dialogue Quality (DQ) Test set

	A-score		S-score		E-score	
Model	RSNOD	NMD	RSNOD	NMD	RSNOD	NMD
IMNTPU-run0	0.2479	0.1618	0.2032	**0.1315**	0.1860	0.1427
Baseline-run0	0.2301	0.1772	**0.1998**	0.1523	0.1854	0.1579

NTCIR-16 DialEval-2 Chinese Dialogue Quality (DQ) Development set

	A-score		S-score		E-score	
Model	RSNOD	NMD	RSNOD	NMD	RSNOD	NMD
IMNTPU-run0	**0.2262**	**0.1495**	0.2076	0.1344	**0.1694**	**0.1251**

NTCIR-16 DialEval-2 English Dialogue Quality (DQ) Test set

	A-score		S-score		E-score	
Model	RSNOD	NMD	RSNOD	NMD	RSNOD	NMD
IMNTPU-run0	0.2535	0.1654	0.2020	0.1312	0.1826	0.1400
Baseline-run0	0.2321	0.1780	0.1986	0.1467	0.1745	0.1431

NTCIR-16 DialEval-2 English Dialogue Quality (DQ) Development set

	A-score		S-score		E-score	
Model	RSNOD	NMD	RSNOD	NMD	RSNOD	NMD
IMNTPU-run0	**0.2102**	**0.1397**	**0.1879**	**0.1216**	**0.1617**	**0.1184**

NTCIR-16 DialEval-2 Chinese Nugget Detection (ND) Test set

Model	JSD	RNSS
Baseline-run0	**0.0585**	**0.1651**

NTCIR-16 DialEval-2 Chinese Nugget Detection (ND) Development set

Model	JSD	RNSS
IMNTPU-run0	2.0670	1.3969

NTCIR-16 DialEval-2 English Nugget Detection (ND) Test set

Model	JSD	RNSS
IMNTPU-run0	**0.0601**	**0.1574**
Baseline-run0	0.0625	0.1722

NTCIR-16 DialEval-2 English Nugget Detection (ND) Development set

Model	JSD	RNSS
IMNTPU-run0	0.0752	0.1727

Conclusions and Contributions

- Most of our runs outperform the baselines.
- XLM-RoBERTa performs relatively well for both Chinese and English data sets.
- We proposed two critical elements, namely, Tokenization procedures and Fine-Tuning Approaches, to improve the DQ and ND subtasks in dialogue analysis.

ACKNOWLEDGMENTS

This research was supported in part by the Ministry of Science and Technology (MOST), Taiwan under grant number 110-2410-H-305-013-MY2, and National Taipei University (NTPU) under grant number 110-NTPU-ORDA-F-001, 111-NTPU-ORDA-F-001, and 111-NTPU-ORDA-F-003.

Information Management, National Taipei University
NTCIR-16 Conference, June 14-17, 2022, Tokyo, Japan