RSLDE at the NTCIR-16 Dialogue Evaluation

Fan Li, Tetsuya Sakai

Waseda University

Abstract

- RSLDE at DialEval-2
  - English and Chinese Dialogue Quality (DQ) subtask
  - English and Chinese Nugget Detection (ND) subtask
- Key Challenges
  - DQ task:
    - Representation of the structure of dialogue
  - ND task:
    - Representation of Dialogue Structure
    - Does Dialogue Context Matters?

Introduction

The RSLDE team participated in the English and Chinese dialogue quality (DQ) and nugget detection (ND) subtasks of DialEval-2 [1].
Our proposed model:

- For ND Task:
  - Sentence Level Model based on BERT [2] and XLNet[3]
  - Dialogue Level Model: Dialogue Embedding + Transformer Encoder + Feed-Forward
  - Dialogue Level Baseline: Dialogue Embedding + Feed-Forward
- For DQ Task:
  - Dialogue Level Model: Dialogue Embedding + Transformer Encoder + Feed-Forward
  - Dialogue Level Baseline: Dialogue Embedding + Feed-Forward

Idea: Nugget has pattern of words

- CNUG0:
  - "@ Ctrip Customer Service Please read the details in the picture."
  - "@ Smartisan Technology Customer Service ... of the wireless ring of mobile phone ?"
- CNUG*:
  - "Thank you."
  - "A customer service staff called to explain the problem this morning. I'm satisfied with this reply. The staff's attitude was sincere. I think Unicom is quite good."
- HNUG*:
  - "We will redouble our efforts to do better service."
  - "You're welcome. That's our job."

Sentence Level Model

Language Model

BERT and XLNet (Permutation AR)

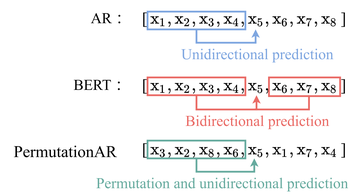


Figure 1:Comparison between BERT and XLNet

Approach

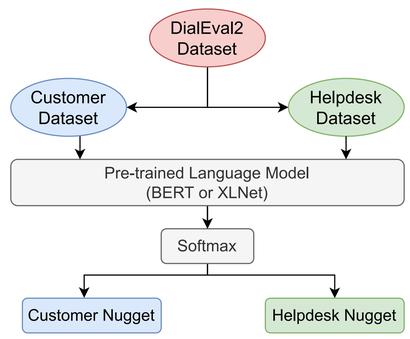
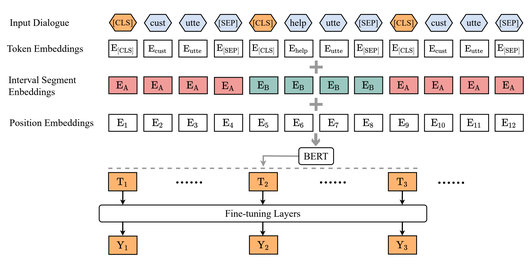


Figure 2:RSLDE sentence model

1. Split the dialogues by turn, and each turn contains one or more utterances from either a customer or a helpdesk.
2. Rebuild two datasets that contain only utterances of the customer and helpdesk respectively.
3. Feed this 2 datasets into sentence-level model.

Dialogue Model

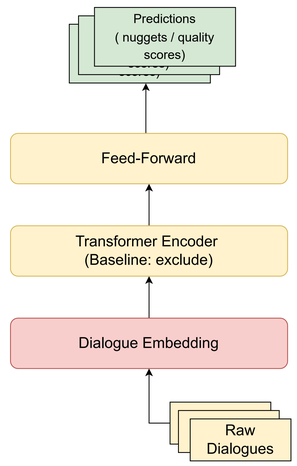


Modified BERT's Interval Segment Embedding to represent the dialogue structure.

Figure 3:Dialogue Embedding

1. Insert [CLS] and [SEP] tokens at the beginning and end of each utterance.
2. Modify BERT's interval segment embeddings.
3. E_A represents customer's utterance and E_B represents helpdesk's utterance.
4. Thus, the model knows where this utterance comes from.

Dialogue Model Structure



Experiments

Loss Function (for all tasks):
$\text{loss} = -\sum_{i=1}^{\text{size}} y_i \cdot \log \hat{y}_i$
Metrics:
Dialogue Quality: Normalised Match Distance
Nugget Detection: Jensen-Shannon Divergence
Submitted Runs:

Task	Language	Run	Model	Batch size
DQ	English	0	Dialogue-BERT+Transformer	12
		1	Dialogue-BERT	
	Chinese	0	Dialogue-BERT+Transformer	
		1	Dialogue-BERT	
ND	English	0	XLNet Baseline	8
		1	BERT Baseline	
		2	Dialogue-BERT+Transformer	
	Chinese	0	XLNet Baseline	
		1	BERT Baseline	
		2	Dialogue-BERT+Transformer	

Results

Table 1:Chinese ND

Run	Mean JSD	Run	Mean RNSS
RSLDE-run0	0.0560(1)	RSLDE-run0	0.1604(1)
BL-LSTM	0.0585	BL-LSTM	0.1651
RSLDE-run2	0.0607	RSLDE-run1	0.1712
RSLDE-run1	0.0634	RSLDE-run2	0.1720
BL-popularity	0.1864	BL-popularity	0.2901
BL-uniform	0.2042	BL-uniform	0.3371

Table 2:Chinese A-score

Run	Mean RSNOD	Run	Mean NMD
BL-LSTM	0.2301	RSLDE-run0	0.1537
BL-popularity	0.2320	RSLDE-run1	0.1551
RSLDE-run0	0.2438	BL-popularity	0.1577
RSLDE-run1	0.2446	BL-LSTM	0.1772
BL-uniform	0.2767	BL-uniform	0.2500

Table 3:Chinese S-score

Run	Mean RSNOD	Run	Mean NMD
RSLDE-run0	0.1938	RSLDE-run1	0.1229
RSLDE-run1	0.1964	RSLDE-run0	0.1243
BL-LSTM	0.1998	BL-popularity	0.1288
BL-popularity	0.2062	BL-LSTM	0.1523
BL-uniform	0.2959	BL-uniform	0.2565

Table 4:Chinese E-score

Run	Mean RSNOD	Run	Mean NMD
RSLDE-run0	0.1660	RSLDE-run0	0.1222(2)
RSLDE-run1	0.1725	RSLDE-run1	0.1286
BL-LSTM	0.1854	BL-LSTM	0.1579
BL-uniform	0.2496	BL-popularity	0.1710
BL-popularity	0.2569	BL-uniform	0.2106

Table 5:English ND

Run	Mean JSD	Run	Mean RNSS
RSLDE-run0	0.0557(1)	RSLDE-run0	0.1615(2)
BL-LSTM	0.0625	BL-LSTM	0.1722
RSLDE-run2	0.0676	RSLDE-run2	0.1778
RSLDE-run1	0.0691	RSLDE-run1	0.1853
BL-popularity	0.1864	BL-popularity	0.2901
BL-uniform	0.2042	BL-uniform	0.3371

Table 6:English A-score

Run	Mean RSNOD	Run	Mean NMD
BL-popularity	0.2320	BL-popularity	0.1577
BL-LSTM	0.2321	BL-LSTM	0.1780
RSLDE-run0	0.2615	RSLDE-run1	0.1896
RSLDE-run1	0.2725	RSLDE-run0	0.1957
BL-uniform	0.2767	BL-uniform	0.2500

Table 7:English S-score

Run	Mean RSNOD	Run	Mean NMD
BL-LSTM	0.1986	BL-popularity	0.1288
BL-popularity	0.2062	RSLDE-run0	0.1381
RSLDE-run0	0.2078	RSLDE-run1	0.1438
RSLDE-run1	0.2154	BL-LSTM	0.1467
BL-uniform	0.2959	BL-uniform	0.2565

Table 8:English E-score

Run	Mean RSNOD	Run	Mean NMD
BL-popularity	0.1745	RSLDE-run0	0.1429
RSLDE-run0	0.1832	BL-LSTM	0.1431
RSLDE-run1	0.1889	RSLDE-run1	0.1444
BL-uniform	0.2496	BL-popularity	0.1710
BL-popularity	0.2569	BL-uniform	0.2106

Conclusion

- The XLNet model has an outstanding language understanding capability for customer-helpdesk dialogues.
- Considering the structure and context information of a dialogue is important for the dialogue nugget detection.

References

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

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv:1810.04805, 2018.

[3] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le.
Xlnet: Generalized autoregressive pretraining for language understanding.
Advances in neural information processing systems, 32, 2019.