Overview of the NTCIR-16 Dialogue Evaluation (DialEval-2) Task

Sijie Tao and Tetsuya Sakai

Waseda University

Outline

1. History, definition, and motivation of DialEval
2. The new data collection for DialEval-2
3. Participants
4. Results
5. Conclusions

Outline

1. **History, definition, and motivation of DialEval**
2. The new data collection for DialEval-2
3. Participants
4. Results
5. Conclusions

History of the task

- NTCIR-14, Jun 2019, Short Text Conversation Task (STC-3) [Zeng+19]
  - DCH-1 Dataset used as training and test sets
  - 3,700 + 390 for Chinese, 1,672 + 390 for English
- NTCIR-15, Dec 2020, Dialogue Evaluation Task (DialEval-1) [Zeng+20]
  - DCH-1 used as training and development sets, new test set built
  - 3,700 + 390 + 300 for Chinese, 2,251 + 390 + 300 for English
- NTCIR-16, Jun 2022, Dialogue Evaluation Task (DialEval-2)
  - DCH-2 [Zeng+21] as training and development sets, new test set built
  - 4,090 + 300 + 65 for both Chinese and English

Task Definition

- DialEval-2 hosts two subtasks:
  - Dialogue Quality (**DQ**)
  - Nugget Detection (**ND**)
- **DQ**: Given a customer-helpdesk dialogue, return an estimated distribution of dialogue quality ratings for the entire dialogue
- **ND**: Given a customer-helpdesk dialogue, return an estimated distribution of labels over nugget types for each turn

An Example of a customer-helpdesk dialogue [Zeng+20]

C: The Smartisan App Store of my mobile phone has been disabled for nearly half a month and the system couldn't be updated. The network was normal. Please give me an explanation.
2016-5-22 13:45

Trigger

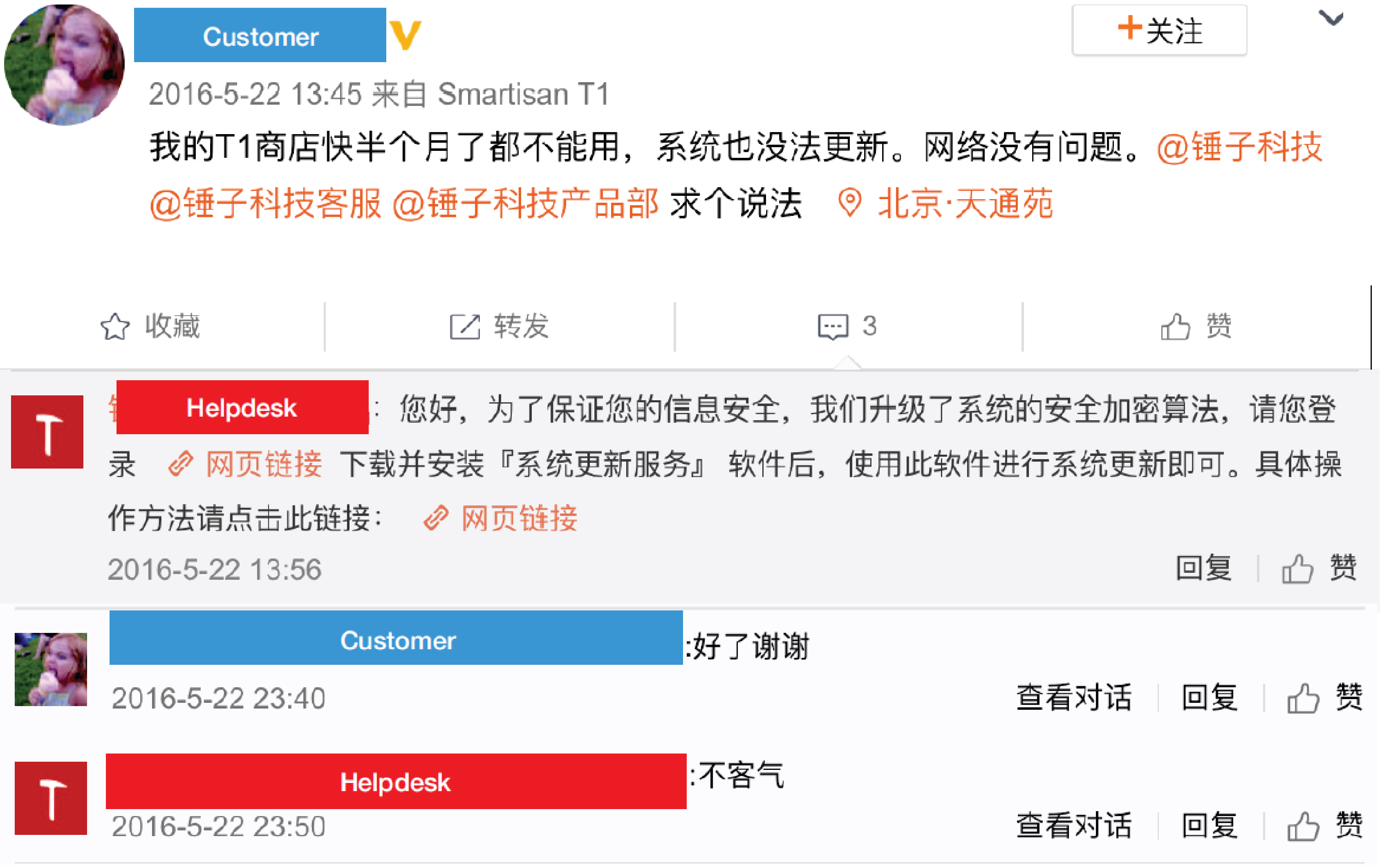
H: To ensure information security, we updated the system security encryption algorithm. Please visit the website, and download and install "System Update Service" to update your system. For detailed operations, please visit the link
2016-5-22 13:56

Solution

C: It worked properly. Thank you!
2016-5-22 23:40

Confirmation

H: You are welcome
2016-5-22 23:50



Customer

2016-5-22 13:45 来自 Smartisan T1

我的T1商店快半个月了都不能用，系统也没法更新。网络没有问题。@锤子科技 @锤子科技客服 @锤子科技产品部 求个说法 北京·天通苑

Helpdesk：您好，为了保证您的信息安全，我们升级了系统的安全加密算法，请您登录 网页链接 下载并安装『系统更新服务』软件后，使用此软件进行系统更新即可。具体操作方法请点击此链接： 网页链接

2016-5-22 13:56

Customer:好了谢谢

2016-5-22 23:40

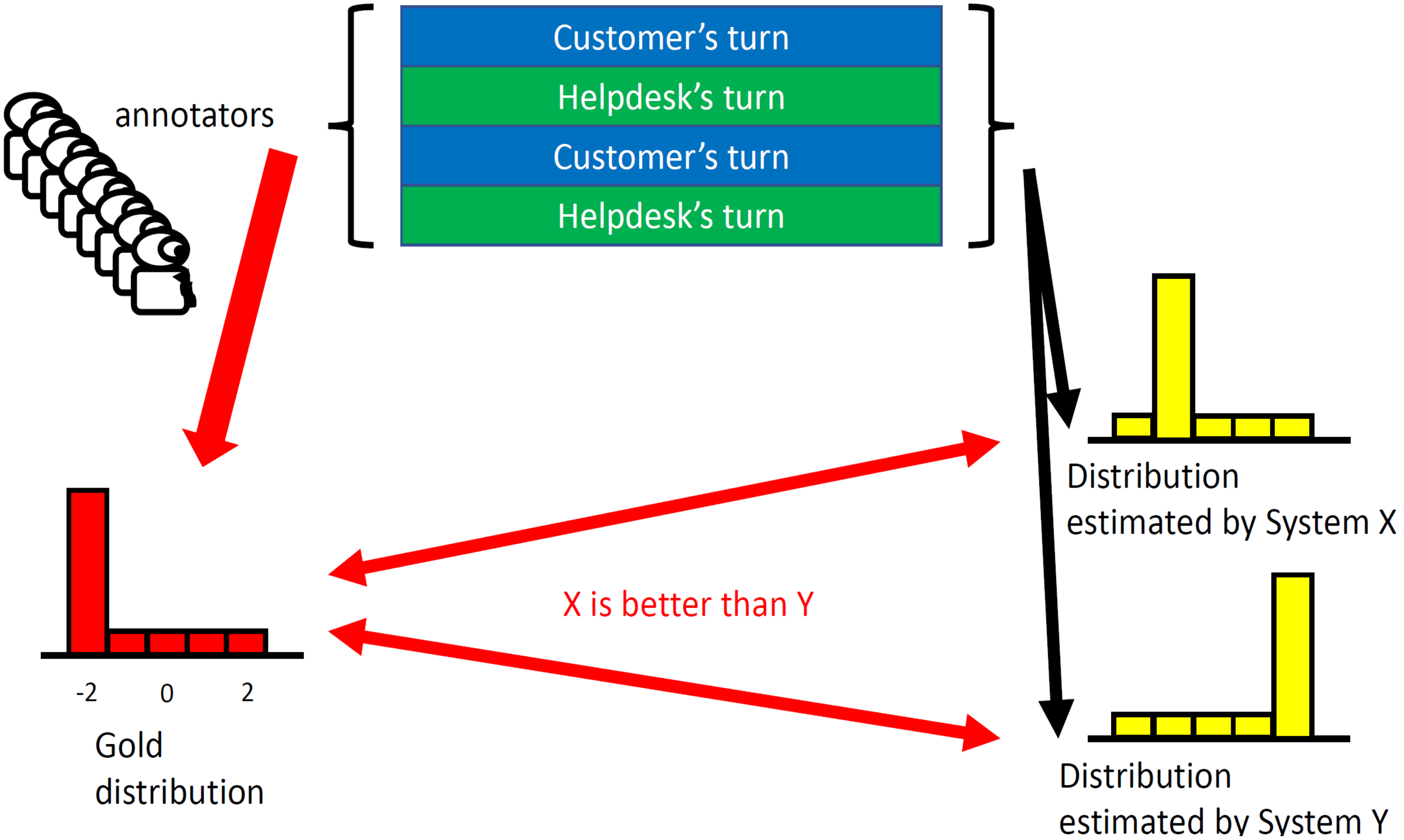
Helpdesk:不客气

2016-5-22 23:50

Dialogue Quality Subtask (DQ)

- Given a customer-helpdesk dialogue, return an estimated distribution of dialogue quality ratings for the entire dialogue.
- Three types of dialogue quality ratings (Likert scale -2 to 2):
  - **A**-score: Task **A**ccomplishment
  - **S**-score: Customer **S**atisfaction (about the dialogue itself, not about the product/service)
  - **E**-score: Dialogue **E**ffectiveness

Dialogue Quality Subtask (DQ)



Customer's turn

Helpdesk's turn

Customer's turn

Helpdesk's turn

annotators

-2 0 2

Gold distribution

X is better than Y

Distribution estimated by System X

Distribution estimated by System Y

Dialogue Quality Subtask (DQ)

- Evaluation metrics
  - NMD (Normalised Match Distance)
  - RSNOD (Root Symmetric Normalised Order-aware Divergence) [Sakai18]
- Both measures take into account the distance between two bins, to make sure X is rated higher than Y in the previous slide.

Nugget Detection Subtask (ND)

- What is a nugget?

An Example of a customer-helpdesk dialogue [Zeng+20]

C: The Smartisan App Store of my mobile phone has been disabled for nearly half a month and the system couldn't be updated. The network was normal. Please give me an explanation.
2016-5-22 13:45

Trigger

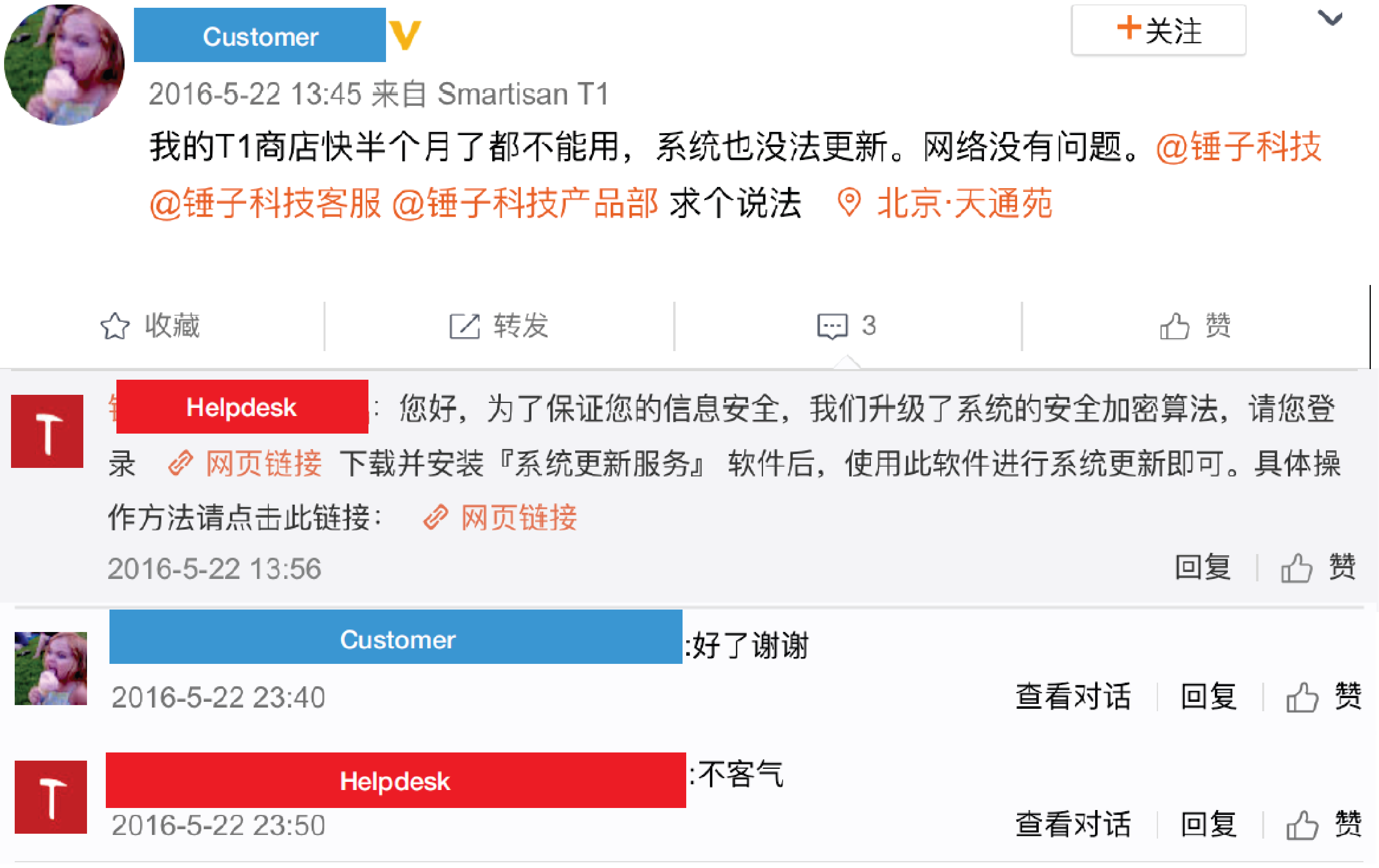
H: To ensure information security, we updated the system security encryption algorithm. Please visit the website, and download and install "System Update Service" to update your system. For detailed operations, please visit the link
2016-5-22 13:56

Solution

C: It worked properly. Thank you!
2016-5-22 23:40

Confirmation

H: You are welcome
2016-5-22 23:50



Customer V +关注

2016-5-22 13:45 来自 Smartisan T1

我的T1商店快半个月了都不能用，系统也没法更新。网络没有问题。@锤子科技 @锤子科技客服 @锤子科技产品部 求个说法 北京·天通苑

收藏 | 转发 | 3 | 赞

Helpdesk：您好，为了保证您的信息安全，我们升级了系统的安全加密算法，请您登录 网页链接 下载并安装『系统更新服务』软件后，使用此软件进行系统更新即可。具体操作方法请点击此链接： 网页链接

2016-5-22 13:56 回复 | 赞

Customer :好了谢谢

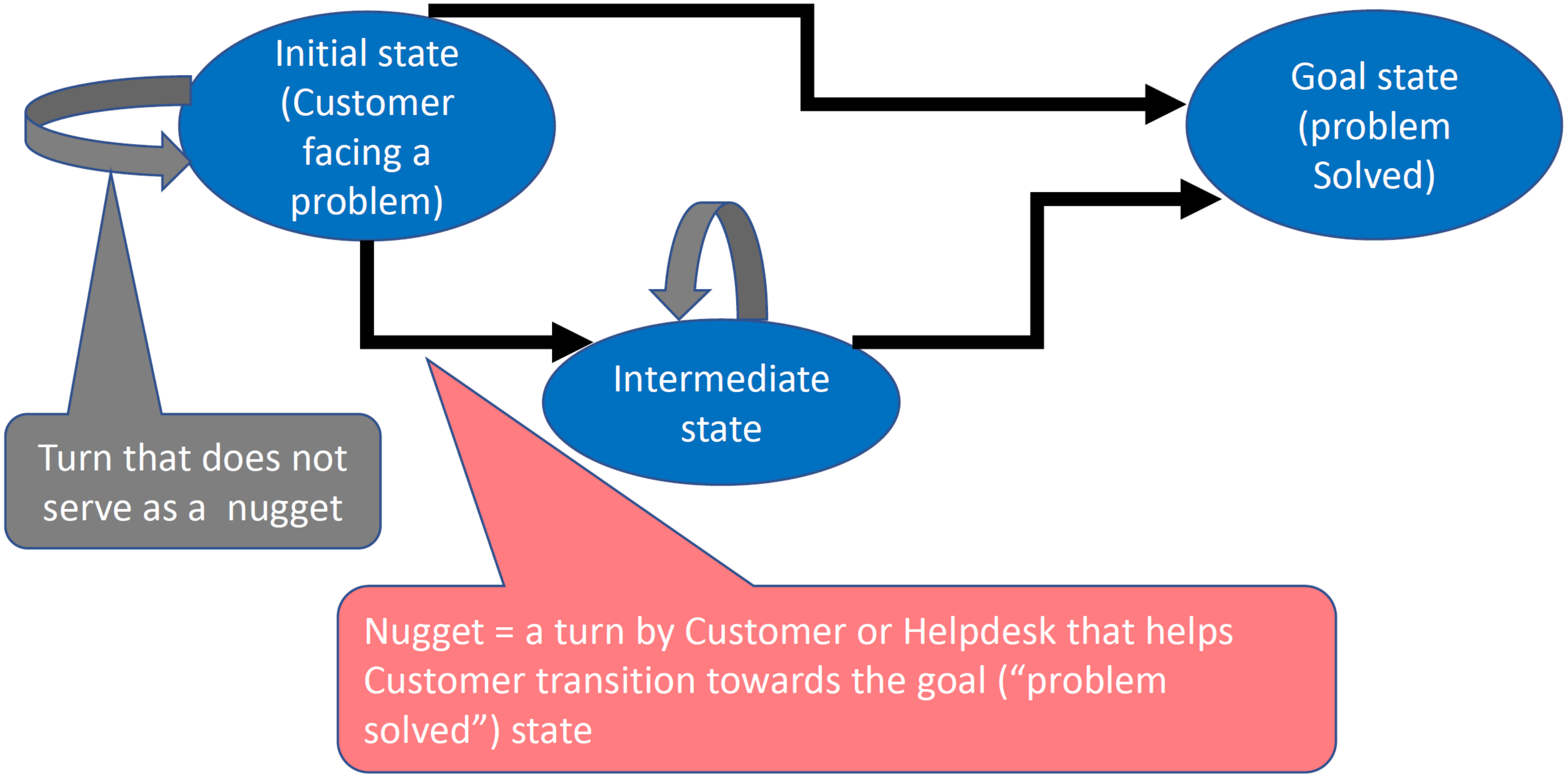
2016-5-22 23:40 查看对话 | 回复 | 赞

Helpdesk :不客气

2016-5-22 23:50 查看对话 | 回复 | 赞

Nugget Detection Subtask (ND)

- What is a nugget?


Initial state (Customer facing a problem)
Goal state (problem Solved)
Intermediate state

Turn that does not serve as a nugget

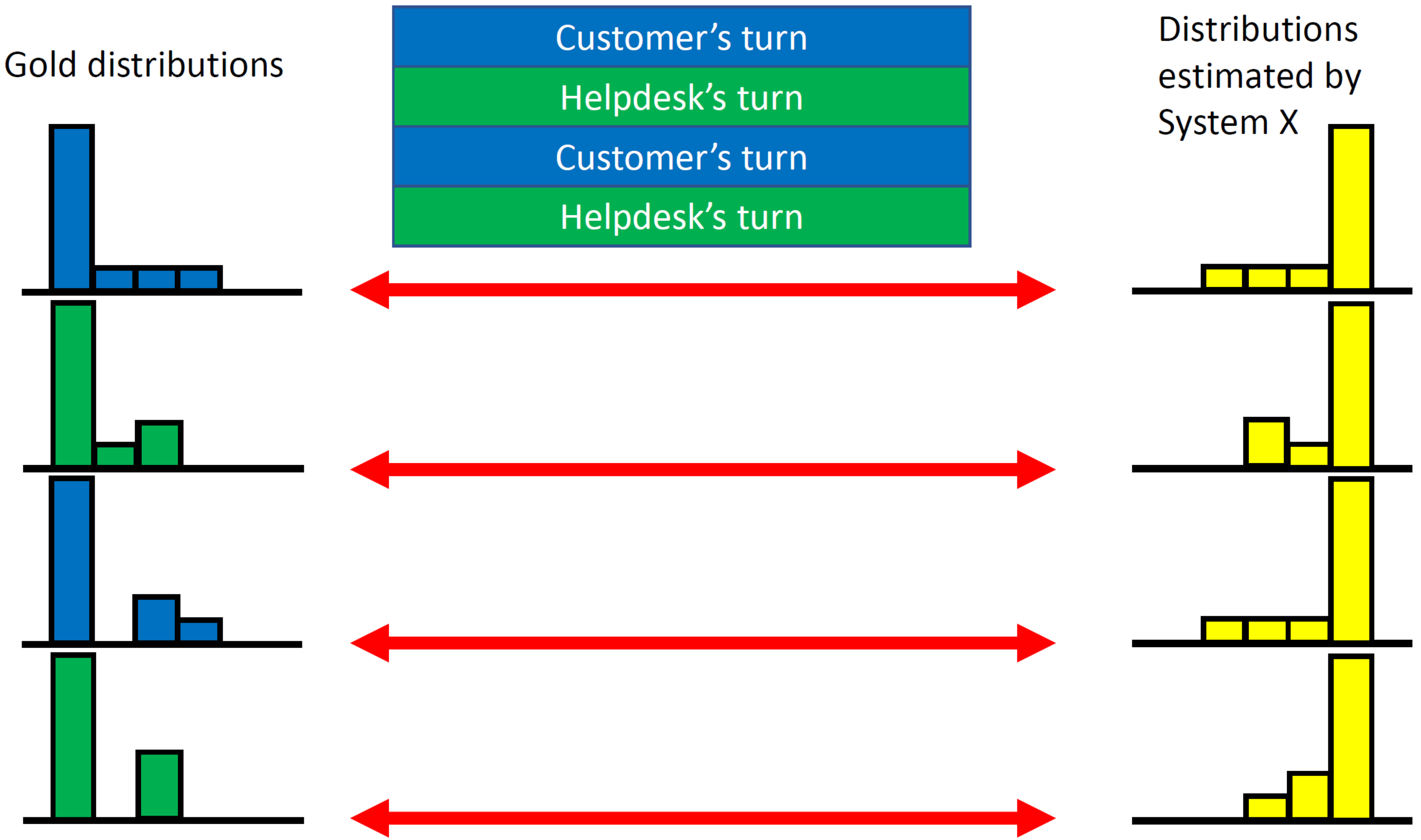
Nugget = a turn by Customer or Helpdesk that helps Customer transition towards the goal (“problem solved”) state

Nugget Detection Subtask (ND)

- Given a customer-helpdesk dialogue, return an estimated distribution of labels over nugget types for each turn

Nugget type	Customer	Helpdesk
Trigger	CNUG0: tell the problem to Helpdesk	
Regular	CNUG	HNUG
Goal	CNUG*: tell Helpdesk that the problem has been solved	HNUG*: tell Customer the solution to the problem
Not-a-nugget	CNaN	HNaN

Nugget Detection Subtask (ND)



Nugget Detection Subtask (ND)

- Evaluation metrics
  - RNSS (Root Normalised Sum of Squares)
  - JSD (Jensen-Shannon Divergence) [Sakai18]
- No need to use NMD or RSNOD, as the bins in the ND subtask are nominal (e.g. HNUG, HNUG*, HNaN), not ordinal

Motivation of the task

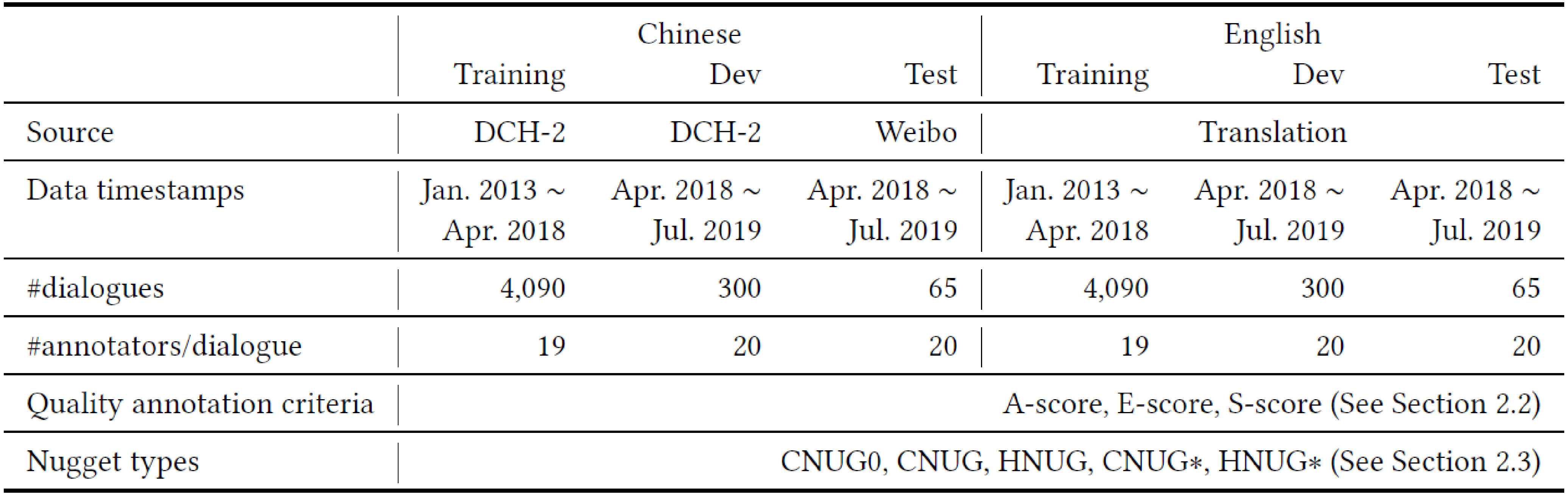
- Evaluate customer-helpdesk dialogues automatically
- DQ: An effective DQ system is useful for building helpdesk systems that can generate effective utterances for diverse users.
- ND: An effective ND system is useful for building effective helpdesk systems that can self diagnose at the dialogue turn level to improve themselves.

Outline

1. History, definition, and motivation of DialEval
2. **The new data collection for DialEval-2**
3. Participants
4. Results
5. Conclusions

The new data collection for DialEval-2

- For DialEval-2, we use DCH-2 dataset [Zeng+21] as training and development sets
- A new test set which contains 65 dialogues is additionally built



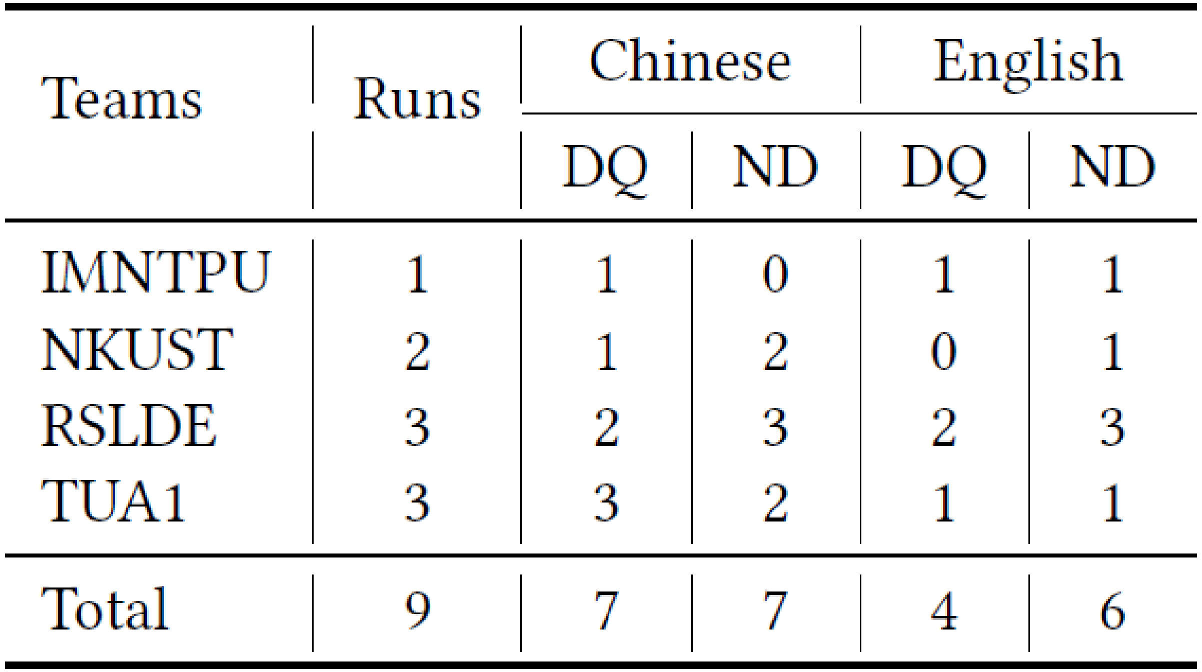
	Chinese			English		
	Training	Dev	Test	Training	Dev	Test
Source	DCH-2	DCH-2	Weibo	Translation		
Data timestamps	Jan. 2013 ~ Apr. 2018	Apr. 2018 ~ Jul. 2019	Apr. 2018 ~ Jul. 2019	Jan. 2013 ~ Apr. 2018	Apr. 2018 ~ Jul. 2019	Apr. 2018 ~ Jul. 2019
#dialogues	4,090	300	65	4,090	300	65
#annotators/dialogue	19	20	20	19	20	20
Quality annotation criteria	A-score, E-score, S-score (See Section 2.2)					
Nugget types	CNUG0, CNUG, HNUG, CNUG∗, HNUG∗ (See Section 2.3)					

Outline

1. History, definition, and motivation of DialEval
2. The new data collection for DialEval-2
3. **Participants**
4. Results
5. Conclusions

Participant teams (only four, last time we had seven)

- IMNTPU (National Taipei University) [Hsiao+22]
- NKUST (National Kaohsiung University of Science and Technology) [Chang+22]
- RSLDE (Waseda University) [Li+22]
- TUA1 (Tokushima University) [Ding+22]



Teams	Runs	Chinese		English	
		DQ	ND	DQ	ND
IMNTPU	1	1	0	1	1
NKUST	2	1	2	0	1
RSLDE	3	2	3	2	3
TUA1	3	3	2	1	1
Total	9	7	7	4	6

Outline

1. History, definition, and motivation of DialEval
2. The new data collection for DialEval-2
3. Participants
4. **Results**
5. Conclusions

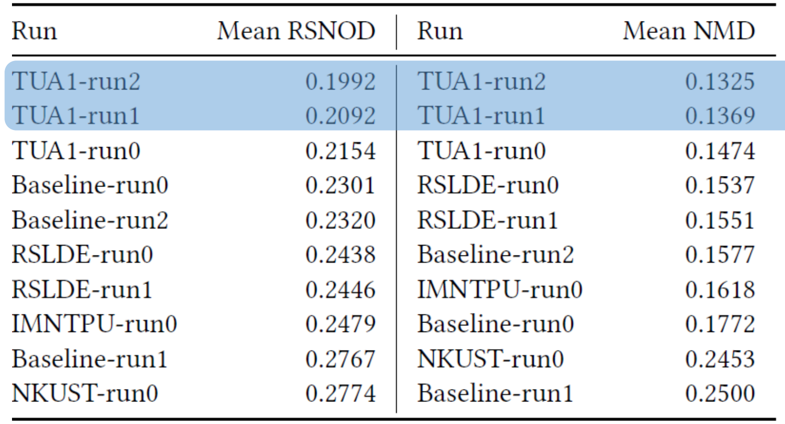
Results

- Baselines (exactly the same as the baselines in DialEval-1) [Zeng+20]
  - BL-lstm (Baseline-run0): A baseline model which leverages Bidirectional Long Short-term Memory;
  - BL-uniform (Baseline-run1): A baseline model which always predict the uniform distribution;
  - BL-popularity (Baseline-run2): A baseline model which predicts the probability of the most popular label as one, and predicts other labels as 0.

Results (DQ, Chinese)

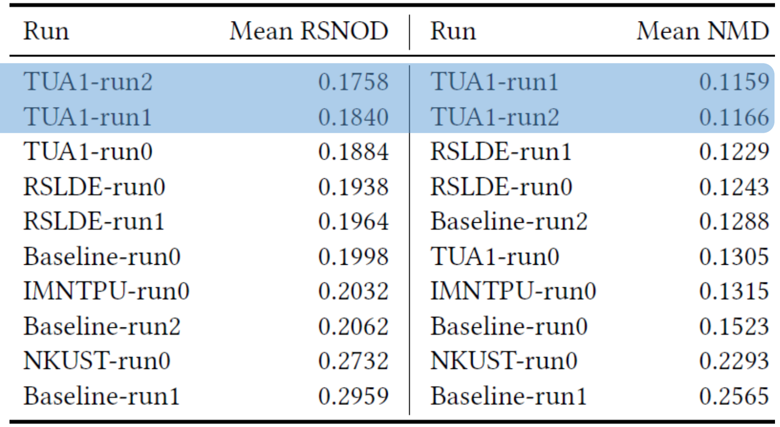
- TUA1-run1, 2 are the top runs in terms of RSNOD and NMD for A and S-score
- Only TUA-run0 outperforms Baseline-run0 statistically significantly in terms of NMD for E-score

Table 4: Chinese Dialogue Quality (A-score) Results



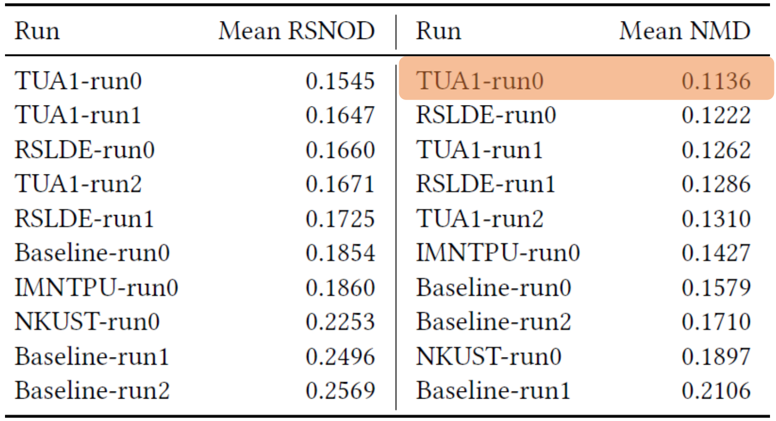
Run	Mean RSNOD	Run	Mean NMD
TUA1-run2	0.1992	TUA1-run2	0.1325
TUA1-run1	0.2092	TUA1-run1	0.1369
TUA1-run0	0.2154	TUA1-run0	0.1474
Baseline-run0	0.2301	RSLDE-run0	0.1537
Baseline-run2	0.2320	RSLDE-run1	0.1551
RSLDE-run0	0.2438	Baseline-run2	0.1577
RSLDE-run1	0.2446	IMNTPU-run0	0.1618
IMNTPU-run0	0.2479	Baseline-run0	0.1772
Baseline-run1	0.2767	NKUST-run0	0.2453
NKUST-run0	0.2774	Baseline-run1	0.2500

Table 5: Chinese Dialogue Quality (S-score) Results



Run	Mean RSNOD	Run	Mean NMD
TUA1-run2	0.1758	TUA1-run1	0.1159
TUA1-run1	0.1840	TUA1-run2	0.1166
TUA1-run0	0.1884	RSLDE-run1	0.1229
RSLDE-run0	0.1938	RSLDE-run0	0.1243
RSLDE-run1	0.1964	Baseline-run2	0.1288
Baseline-run0	0.1998	TUA1-run0	0.1305
IMNTPU-run0	0.2032	IMNTPU-run0	0.1315
Baseline-run2	0.2062	Baseline-run0	0.1523
NKUST-run0	0.2732	NKUST-run0	0.2293
Baseline-run1	0.2959	Baseline-run1	0.2565

Table 6: Chinese Dialogue Quality (E-score) Results

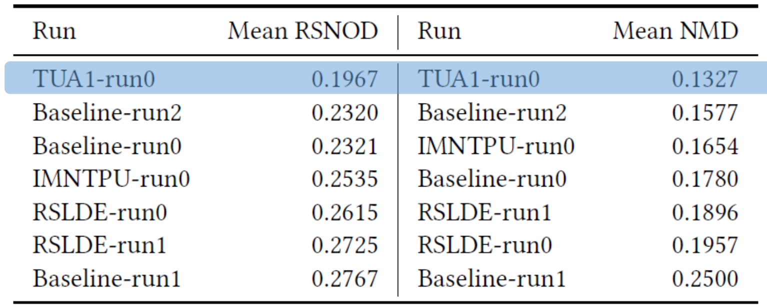


Run	Mean RSNOD	Run	Mean NMD
TUA1-run0	0.1545	TUA1-run0	0.1136
TUA1-run1	0.1647	RSLDE-run0	0.1222
RSLDE-run0	0.1660	TUA1-run1	0.1262
TUA1-run2	0.1671	RSLDE-run1	0.1286
RSLDE-run1	0.1725	TUA1-run2	0.1310
Baseline-run0	0.1854	IMNTPU-run0	0.1427
IMNTPU-run0	0.1860	Baseline-run0	0.1579
NKUST-run0	0.2253	Baseline-run2	0.1710
Baseline-run1	0.2496	NKUST-run0	0.1897
Baseline-run2	0.2569	Baseline-run1	0.2106

Results (DQ, English)

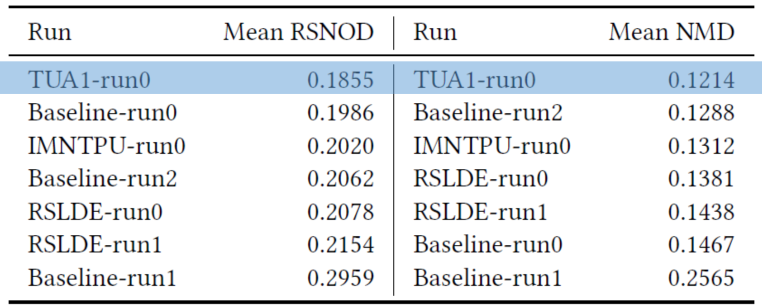
- TUA1-run0 is the top run and the only run that outperforms the baseline systems
- But the differences between TUA1-run0 and the top baselines are not statistically significant

Table 9: English Dialogue Quality (A-score) Results



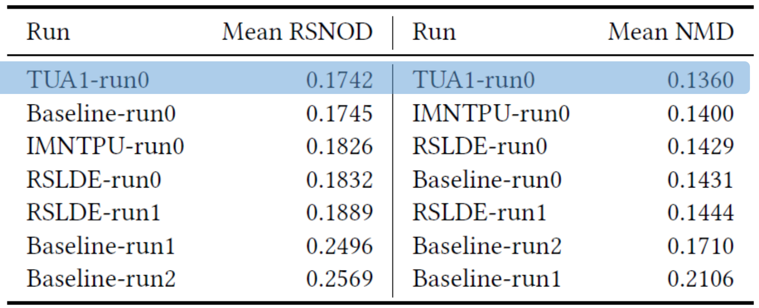
Run	Mean RSNOD	Run	Mean NMD
TUA1-run0	0.1967	TUA1-run0	0.1327
Baseline-run2	0.2320	Baseline-run2	0.1577
Baseline-run0	0.2321	IMNTPU-run0	0.1654
IMNTPU-run0	0.2535	Baseline-run0	0.1780
RSLDE-run0	0.2615	RSLDE-run1	0.1896
RSLDE-run1	0.2725	RSLDE-run0	0.1957
Baseline-run1	0.2767	Baseline-run1	0.2500

Table 10: English Dialogue Quality (S-score) Results



Run	Mean RSNOD	Run	Mean NMD
TUA1-run0	0.1855	TUA1-run0	0.1214
Baseline-run0	0.1986	Baseline-run2	0.1288
IMNTPU-run0	0.2020	IMNTPU-run0	0.1312
Baseline-run2	0.2062	RSLDE-run0	0.1381
RSLDE-run0	0.2078	RSLDE-run1	0.1438
RSLDE-run1	0.2154	Baseline-run0	0.1467
Baseline-run1	0.2959	Baseline-run1	0.2565

Table 11: English Dialogue Quality (E-score) Results



Run	Mean RSNOD	Run	Mean NMD
TUA1-run0	0.1742	TUA1-run0	0.1360
Baseline-run0	0.1745	IMNTPU-run0	0.1400
IMNTPU-run0	0.1826	RSLDE-run0	0.1429
RSLDE-run0	0.1832	Baseline-run0	0.1431
RSLDE-run1	0.1889	RSLDE-run1	0.1444
Baseline-run1	0.2496	Baseline-run2	0.1710
Baseline-run2	0.2569	Baseline-run1	0.2106

Results (ND, Chinese)

- RSLDE-run0 is the top run and the only run that can outperform Baseline-run0 in terms of both JSD and RNSS
- But the difference between them is not statistically significant

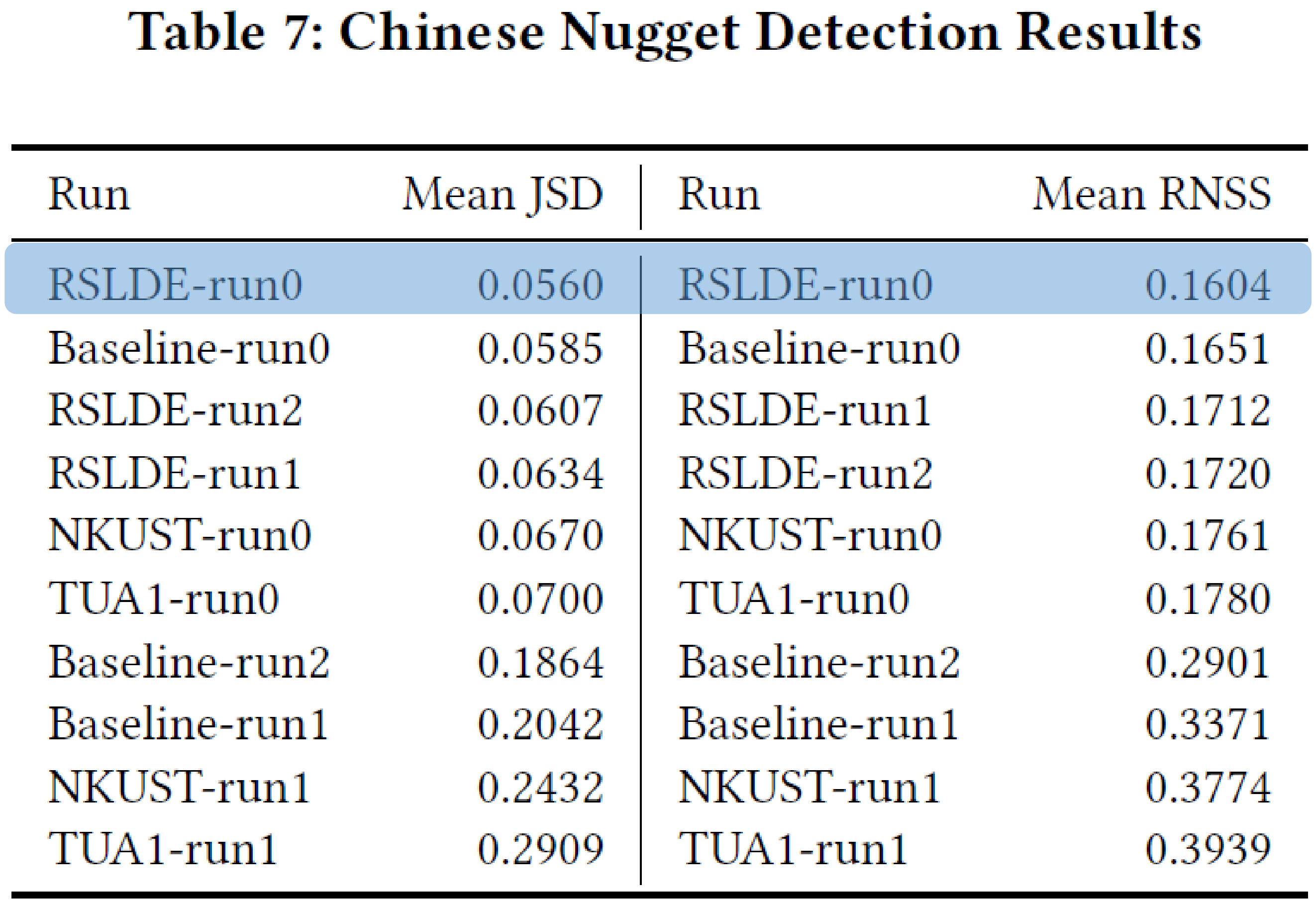


Table 7: Chinese Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
RSLDE-run0	0.0560	RSLDE-run0	0.1604
Baseline-run0	0.0585	Baseline-run0	0.1651
RSLDE-run2	0.0607	RSLDE-run1	0.1712
RSLDE-run1	0.0634	RSLDE-run2	0.1720
NKUST-run0	0.0670	NKUST-run0	0.1761
TUA1-run0	0.0700	TUA1-run0	0.1780
Baseline-run2	0.1864	Baseline-run2	0.2901
Baseline-run1	0.2042	Baseline-run1	0.3371
NKUST-run1	0.2432	NKUST-run1	0.3774
TUA1-run1	0.2909	TUA1-run1	0.3939

Results (ND, English)

- RSLDE-run0 and IMNTPU-run0 are the runs can outperform Baseline-run0
- But their differences between the baseline are not statistically significant

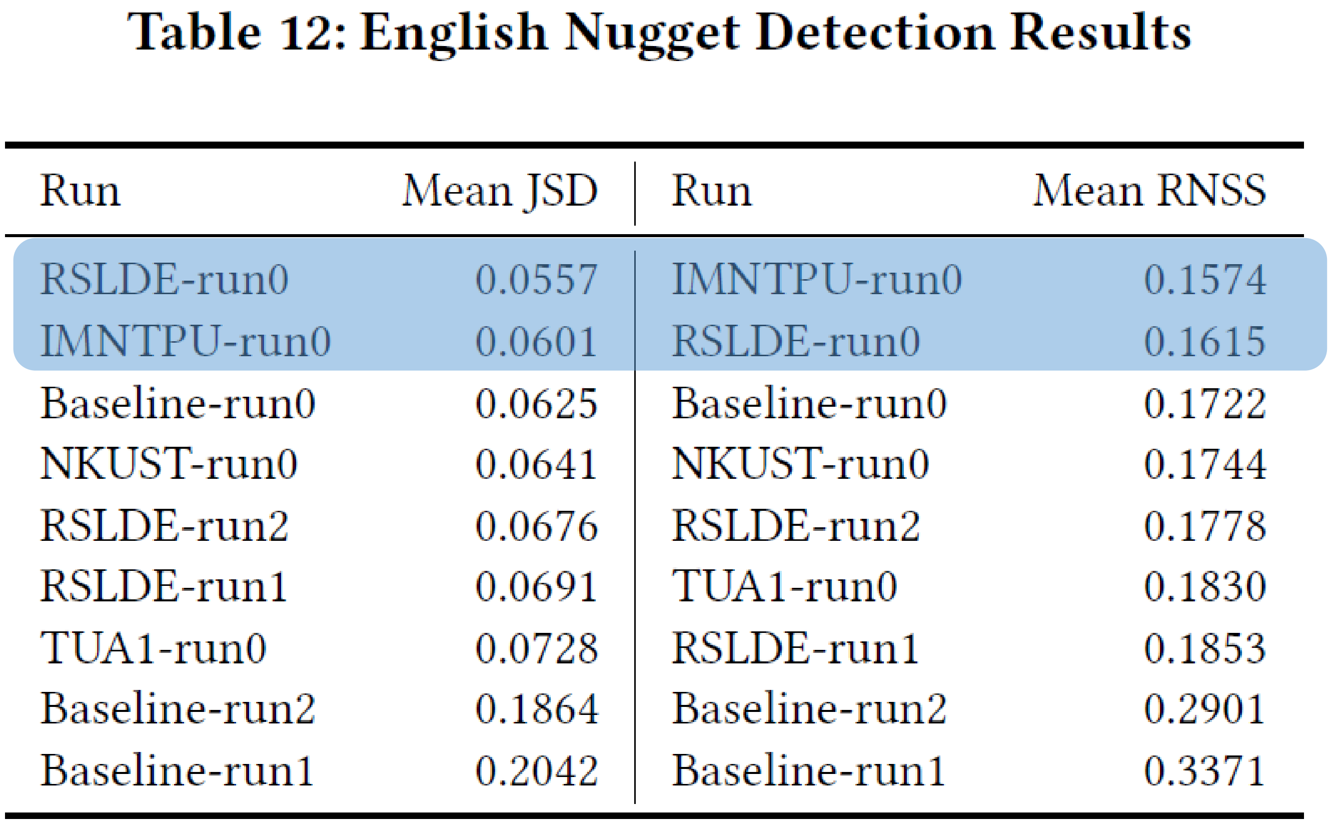


Table 12: English Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
RSLDE-run0	0.0557	IMNTPU-run0	0.1574
IMNTPU-run0	0.0601	RSLDE-run0	0.1615
Baseline-run0	0.0625	Baseline-run0	0.1722
NKUST-run0	0.0641	NKUST-run0	0.1744
RSLDE-run2	0.0676	RSLDE-run2	0.1778
RSLDE-run1	0.0691	TUA1-run0	0.1830
TUA1-run0	0.0728	RSLDE-run1	0.1853
Baseline-run2	0.1864	Baseline-run2	0.2901
Baseline-run1	0.2042	Baseline-run1	0.3371

Results

(Differences between metrics)

- The difference between different metrics are not statistically significant for both ND and DQ subtasks
- Consistent with what we observed at DialEval-1 and STC-3. [Zeng+19][Zeng+20]

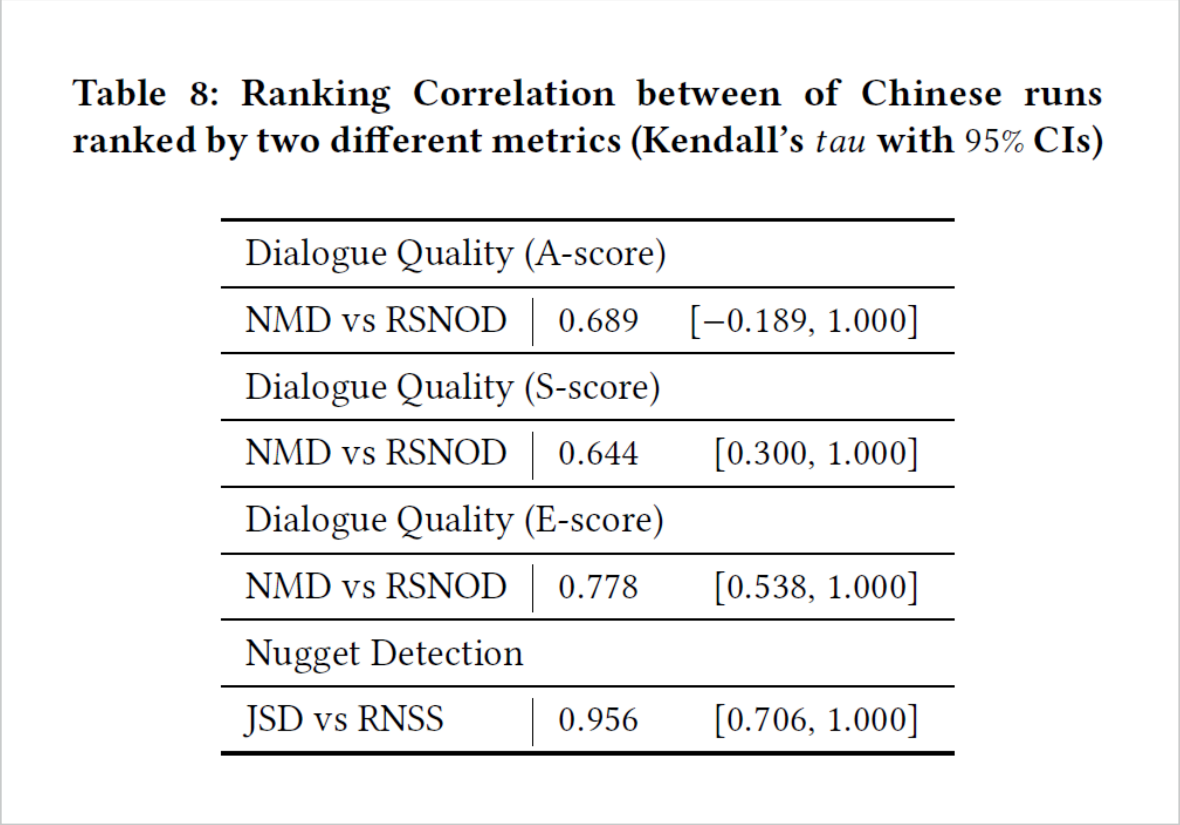


Table 8: Ranking Correlation between of Chinese runs ranked by two different metrics (Kendall's *tau* with 95% CIs)

Dialogue Quality (A-score)		
NMD vs RSNOD	0.689	[−0.189, 1.000]
Dialogue Quality (S-score)		
NMD vs RSNOD	0.644	[0.300, 1.000]
Dialogue Quality (E-score)		
NMD vs RSNOD	0.778	[0.538, 1.000]
Nugget Detection		
JSD vs RNSS	0.956	[0.706, 1.000]

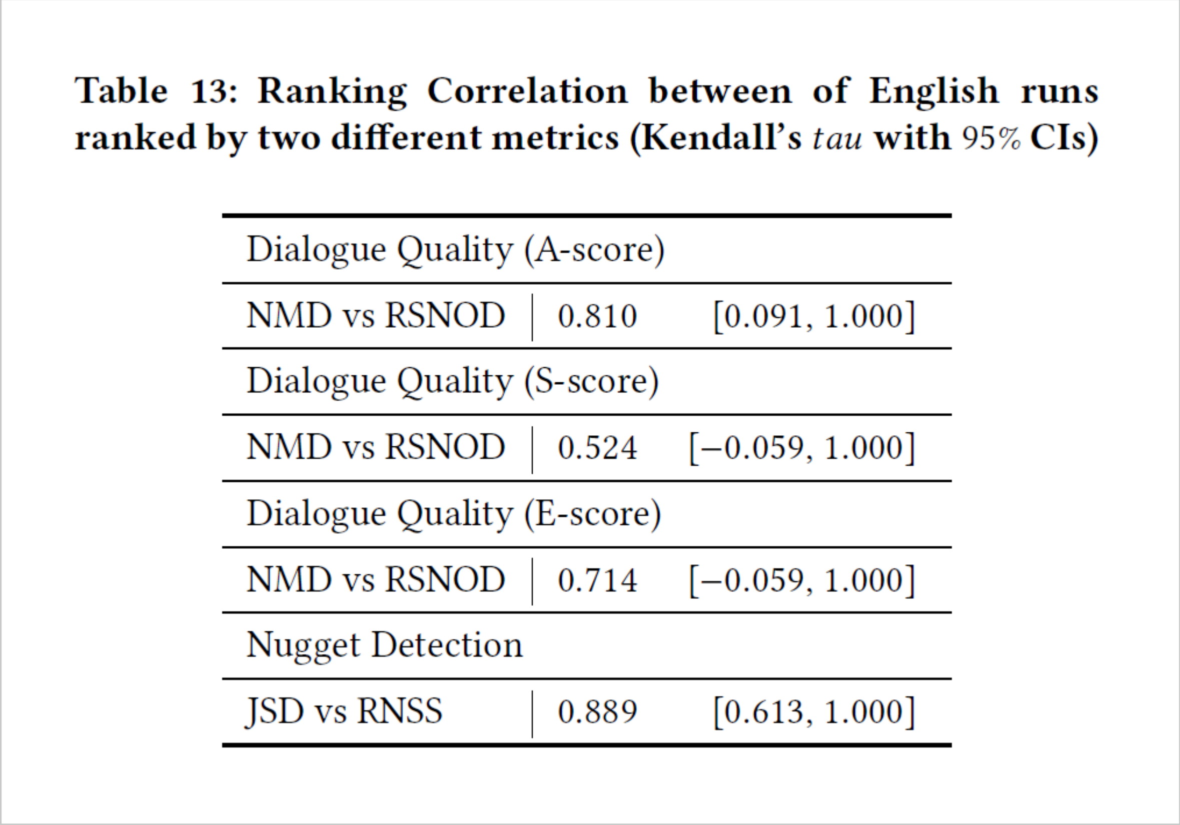


Table 13: Ranking Correlation between of English runs ranked by two different metrics (Kendall's *tau* with 95% CIs)

Dialogue Quality (A-score)		
NMD vs RSNOD	0.810	[0.091, 1.000]
Dialogue Quality (S-score)		
NMD vs RSNOD	0.524	[−0.059, 1.000]
Dialogue Quality (E-score)		
NMD vs RSNOD	0.714	[−0.059, 1.000]
Nugget Detection		
JSD vs RNSS	0.889	[0.613, 1.000]

Outline

1. History, definition, and motivation of DialEval
2. The new data collection for DialEval-2
3. Participants
4. Results
5. **Conclusions**

Conclusions

- Overview of DialEval-2:
  - Task definition
  - Data collection
  - Evaluation results
- From the evaluation results, we observe that
  - Only one run from TUA1 outperform the LSTM baseline significantly in Chinese DQ task in terms of NMD for E-score.
  - In other subtasks, none of the runs can outperform the LSTM baseline significantly.
  - No substantial difference is observed between the evaluation metrics for each subtasks.

References

[Zeng+19] Zhaohao Zeng, Sosuke Kato, and Tetsuya Sakai. 2019. Overview of the NTCIR-14 Short Text Conversation Task: Dialogue Quality and Nugget Detection Subtasks. In *Proceedings of NTCIR-14*. 290–315.

[Zeng+20] Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai, and Inho Kang. 2020. Overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) Task. In *Proceedings of NTCIR-15*. 13–34.

[Zeng+21] Zhaohao Zeng and Tetsuya Sakai. 2021. DCH-2: A Parallel Customer-Helpdesk Dialogue Corpus with Distributions of Annotators' Labels. *CoRR* abs/2104.08755 (2021). arXiv:2104.08755 https://arxiv.org/abs/2104.08755

[Sakai18] Tetsuya Sakai. 2018. Comparing Two Binned Probability Distributions for Information Access Evaluation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR'18)*. ACM, New York, NY, USA, 1073–1076.

References

[Chang+22] Tao-Hsing Chang and Jian-He Chen. 2022. NKUST at the NTCIR-16 DialEval-2 Task. In *Proceedings of NTCIR-16*. to appear.

[Ding+22] Fei Ding, Kang Xin, Yunong Wu, and Fuji Ren. 2022. TUA1 at the NTCIR-16 DialEval-2 Task. In *Proceedings of NTCIR-16*. to appear.

[Hsiao+22] Ting-Yun Hsiao, Yung-Wei Teng, Pei-Tz Chiu, Mike Tian-Jian Jiang, and Min-Yuh Day. 2022. IMNTPU Dialogue System Evaluation at the NTCIR-16 DialEval-2 Dialogue Quality and Nugget Detection. In *Proceedings of NTCIR-16*. to appear.

[Li+22] Fan Li and Tetsuya Sakai. 2022. RSLDE at the NTCIR-16 DialEval-2 Task. In *Proceedings of NTCIR-16*. to appear.