Overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) Task

Zhaohao Zeng Sosuke Kato Tetsuya Sakai Inho Kang
Waseda University Naver Corporation

dialeval1org@list.waseda.jp

http://sakailab.com/dialeval1/

Dec 8-11, 2020@NTCIR-15, NII, Japan.

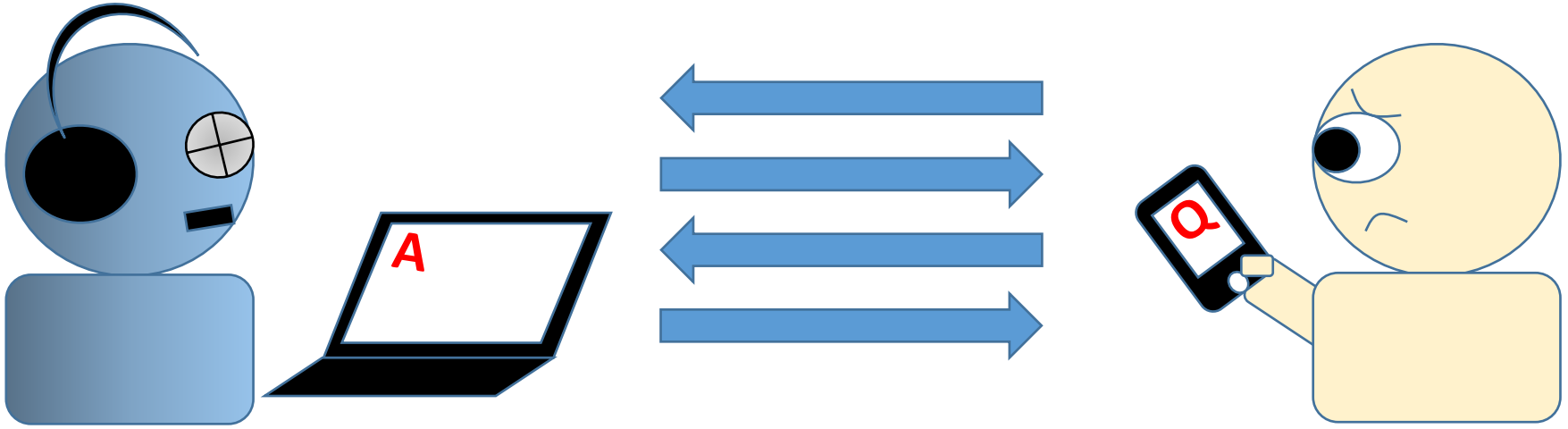
DialEval-1:
The Successor of STC-3 NDDQ at NTCIR-14

- Previous Short Text Conversation (STC) tasks mainly investigate retrieving/generating responses for chit-chat dialogues.

- DialEval focus on how to evaluate task-oriented dialogues
  - Customer-helpdesk dialogues

Motivation

- You cannot improve what you cannot measure.
- ⇒ To build good task-oriented, multi-round, textual dialogue systems, we need good ways to evaluate them.



Online evaluation is important but

- Costly and does not scale
- Difficult to compare different systems
- Not repeatable even for the same system

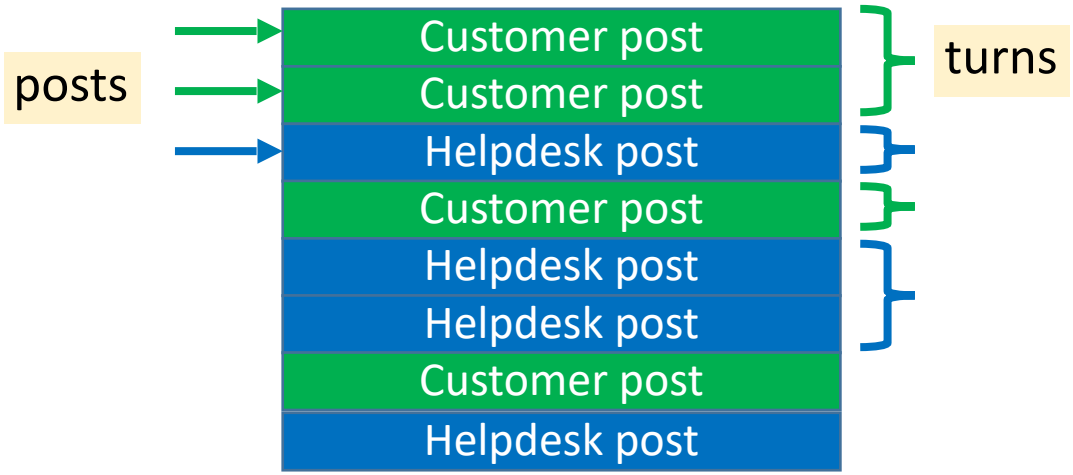
Customer-Helpdesk Dialogues

- Textual dialogues
- To solve some certain problems for the customer
- Post

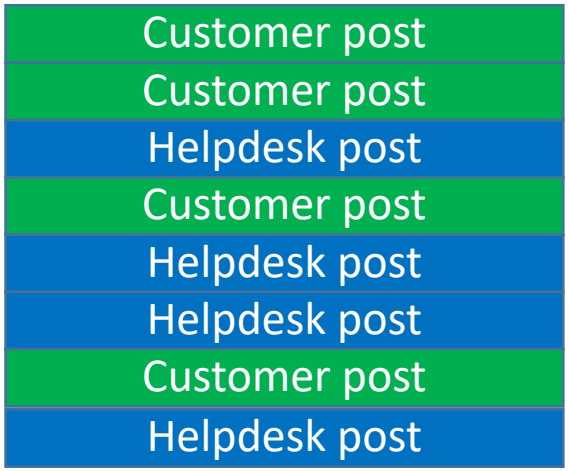
Text entered by utterer in a dialogue, each with a timestamp

- Turn

Maximal consecutive posts by the same utterer



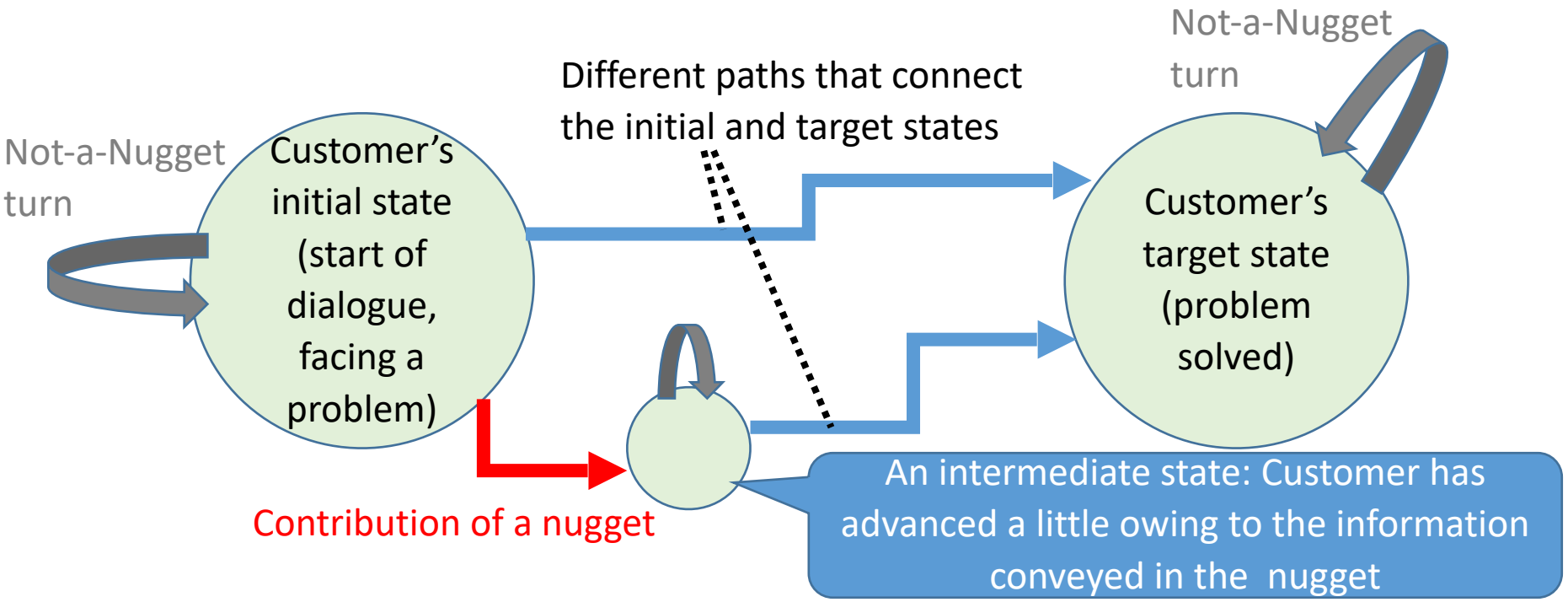
Dialogue Quality (DQ) subtask (Chinese and English)



- Estimate quality scores for customer-helpdesk dialogues
  - A-score: Task Accomplishment (Has the problem been solved? To what extent?)
  - S-score: Customer Satisfaction of the dialogue (not of the product/service or the company)
  - E-score: Dialogue Effectiveness (Do the utterers interact effectively to solve the problem efficiently?)
  - Scale: -2, -1, 0, 1, 2

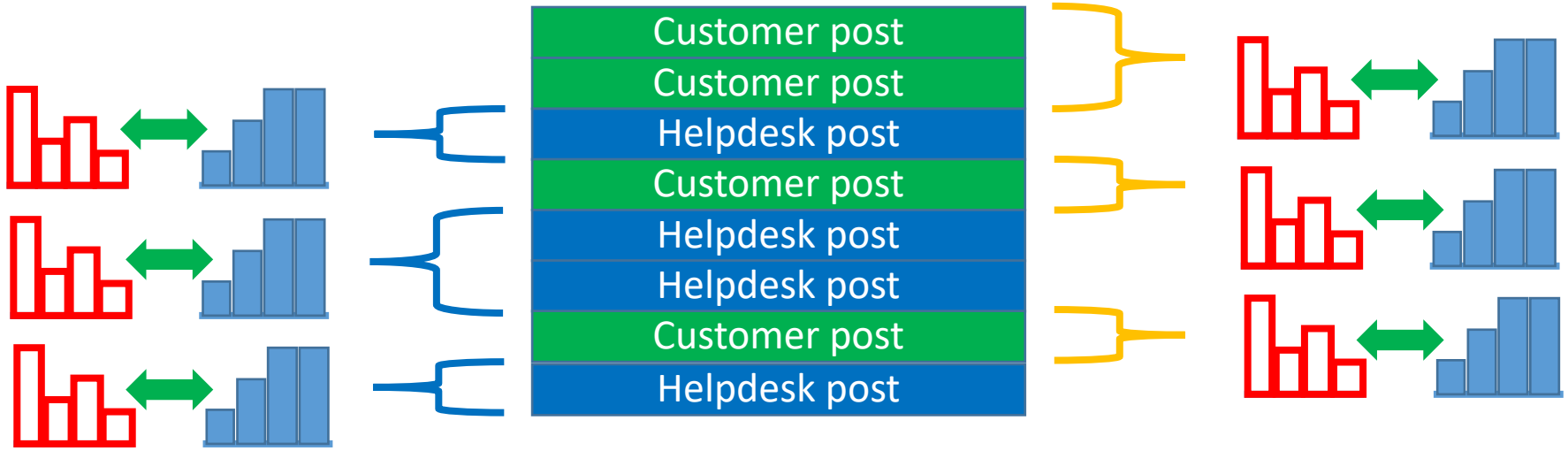
Nugget Detection (ND) subtask (Chinese and English) cont'd

- A nugget is a turn that helps the Customer transition from the current state (where the problem is yet to be solved) towards the target state (where the problem has been solved).



Nugget Detection (ND) subtask (Chinese and English) cont'd

INPUT: customer-helpdesk dialogue

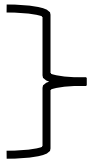


OUTPUT: estimated nugget types for
each helpdesk turn and each customer turn

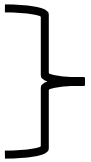
- Trigger nugget
- Goal nugget
- Regular nugget
- Not-a-nugget

Nugget types

- Customer trigger (problem stated)
- Helpdesk goal (solution stated)
- Customer goal (solution confirmed)
- Customer regular
- Helpdesk regular
- Customer Not-a-Nugget
- Helpdesk Not-a-Nugget



Contains information that leads to solution



Does not contain information that leads to solution

Nugget types: an example

C: Customer H: Helpdesk

Customer Trigger → C: I cannot log in your website

H: Could you **refresh** the browser and have a try again?

C: I tried, but it did not help

Helpdesk Regular → H: Could you tell me which browser you are using?

Customer Regular → C: OK, it is Internet Explorer 6

Helpdesk Goal → H: Could you use Chrome instead? We don't support IE6.

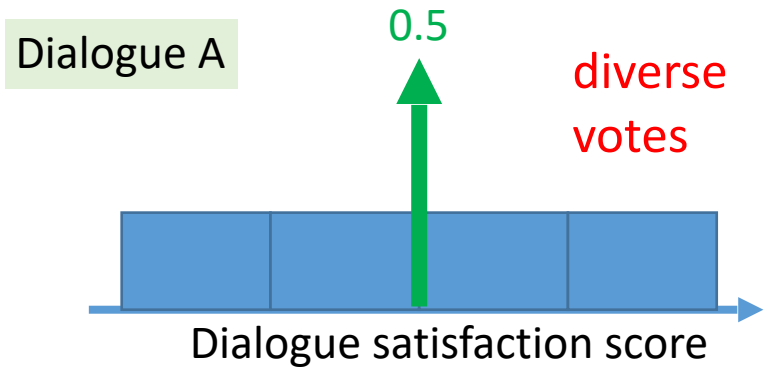
Customer Goal → C: Okay, it works now with Chrome! Thank you.

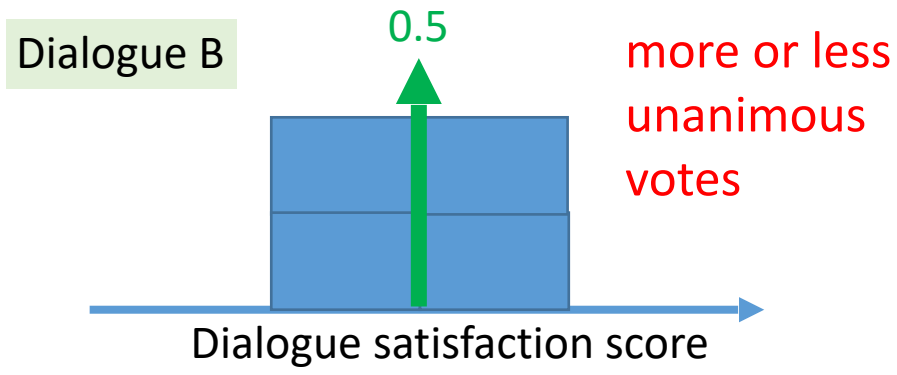
Why nuggets?

- Useful features for automatically estimating the dialogue quality.
- Diagnose a dialogue closely (why it failed, where it failed).
- Design Helpdesk systems that provide the solution to a given problem effectively and efficiently.

Evaluation based on distributions (1)

- People can have diverse views for the same dialogue.
- Traditionally, raw assessor votes are thrown away after settling on a final label. It does not preserve the diversity/unanimity of the votes.





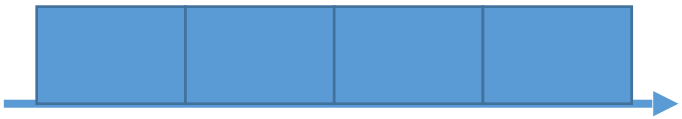
cf. Dialogue Breakdown Detection Challenge (DBDC)

Evaluation based on distributions (2)

- We use the gold distribution directly for system evaluation: we let the systems estimate it.
- Such systems can be used as a component of a dialogue system that can accommodate diverse people.

Dialogue A

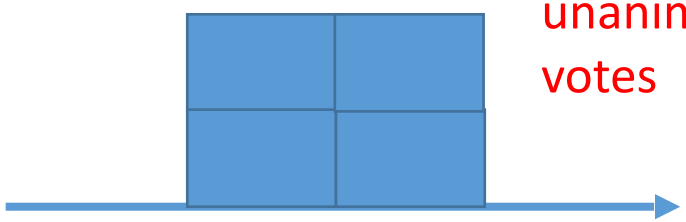
diverse votes



Dialogue satisfaction score

Dialogue B

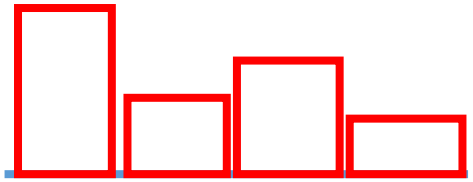
more or less unanimous votes



Dialogue satisfaction score

cf. Dialogue Breakdown Detection Challenge (DBDC)

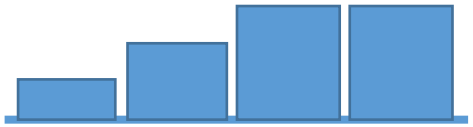
Evaluation based on distributions (3)



System estimated distribution

M(d): measure quantifying how p differs from p* (see later slide)

$$meanM = \frac{1}{|D|} \sum_{d \in D} M(d)$$



Golden distribution

Evaluation measures (comparing system and gold distributions)

- Dialogue Quality (ordinal bins):

- NMD: Normalised Match Distance - a special case of Earth Mover's Distance

- RSNOD: Root Symmetric Normalised Order-aware Divergence

as M(d) for each dialogue d.

- Nugget Detection (nominal bins):

- RNSS: Root Normalised Sum of Squared errors

- JSD: Jensen-Shannon divergence

as M(b) for each turn b.

See: Sakai, .T:
Comparing Two Binned Probability Distributions for Information Access Evaluation
https://waseda.box.com/SIGIR2018preprint

Training and Dev Data

- NTCIR-14 STC-3 NDDQ subtasks dataset
- Mined Chinese Dialogues from Weibo (a Chinese microblog)

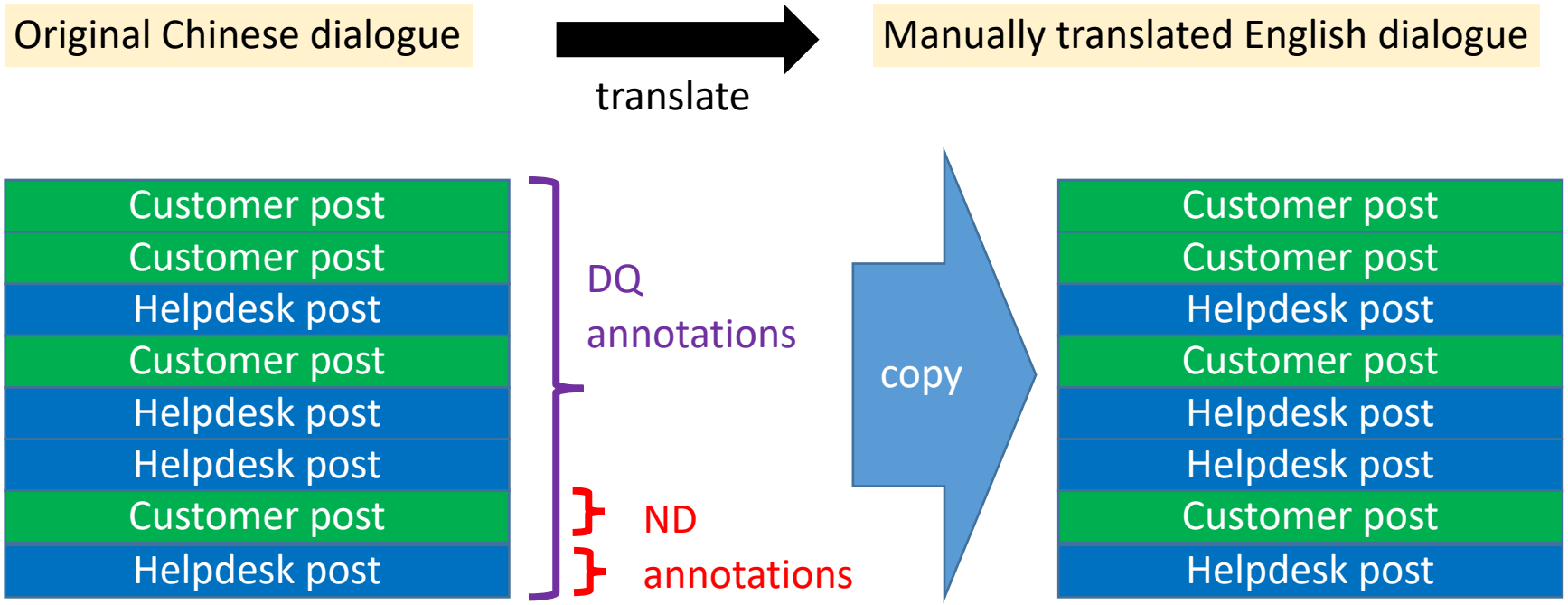
- Training: 3700 (2256 has English translation)
- Dev: 390 (all have English translation)
- Annotated by 19-20 annotators per dialogue
  - Quality scores (A-score, S-score, E-score)
  - Nugget types (CNUG0, CNUG*, HNUG*, CNUG, HNUG, CNaN, HNaN)

Test data

- 300 new Chinese Weibo helpdesk/customer dialogues crawled in Jul 2019
- Annotated in the same way as the training data
- All test dialogues manually translated into English

On annotations

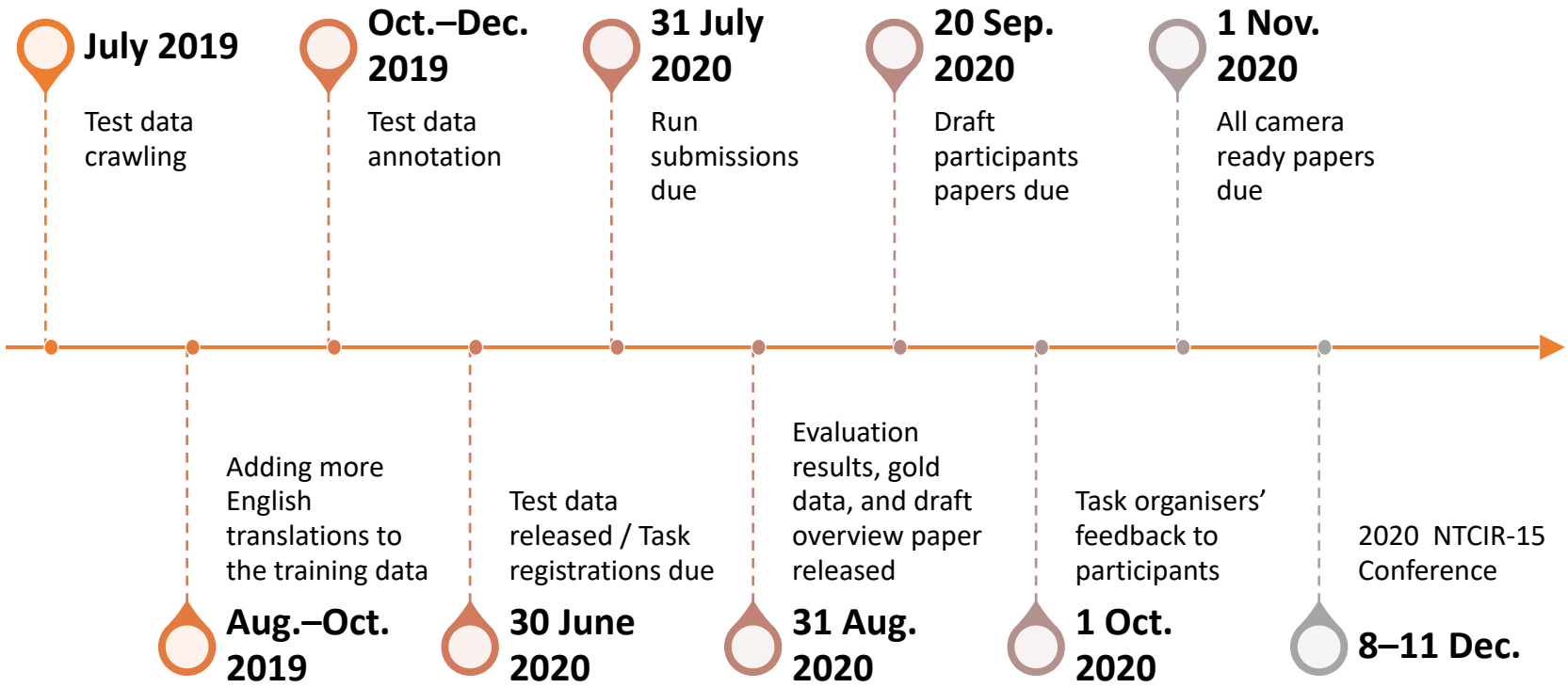
- For both training and test data, only the Chinese portions were annotated. These annotations will then be copied onto the English portions.



Supplementary Resources

- A LSTM baseline (tensorflow) as a start point on Github: https://github.com/DialEval-1/LSTM-baseline
- A Python Script for evaluation
- A web tool to evaluate test data during the development period
  - Limited times
  - Hidden data

Timeline (incl. generic NTCIR-15 schedule)



Participant Teams

- 7 teams
- 16 participant runs for DQ
- 12 participant runs for ND

Table 2: The Number of Runs

Team	Chinese		English	
	DQ	ND	DQ	ND
Baseline	3	3	3	3
IMTKU	3	3	1	3
NKUST	2	2	1	1
RSLNV	1	2	1	2
SKYMN	0	0	3	0
TMUDS	0	3	0	0
TUA1	3	1	0	0
WUST	1	1	0	0
Total	13	15	9	9

Results (a part)

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

Run	Mean RSNOD	Run	Mean NMD
TUA1-run2	0.2102	IMTKU-run2	0.1392
IMTKU-run2	0.2130	TUA1-run0	0.1396
TUA1-run0	0.2136	IMTKU-run0	0.1406
IMTKU-run0	0.2165	TUA1-run2	0.1412
IMTKU-run1	0.2204	IMTKU-run1	0.1442
Baseline-run0	0.2305	TUA1-run1	0.1510
RSLNV-run0	0.2345	NKUST-run1	0.1594
WUST-run0	0.2427	Baseline-run0	0.1598
NKUST-run1	0.2430	RSLNV-run0	0.1606
Baseline-run2	0.2473	Baseline-run2	0.1643
TUA1-run1	0.2484	WUST-run0	0.1724
NKUST-run0	0.2696	NKUST-run0	0.2384
Baseline-run1	0.2706	Baseline-run1	0.2522

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

Run	Mean RSNOD	Run	Mean NMD
IMTKU-run2	0.1918	IMTKU-run2	0.1254
IMTKU-run1	0.1964	IMTKU-run0	0.1284
IMTKU-run0	0.1977	IMTKU-run1	0.1290
TUA1-run2	0.2024	TUA1-run2	0.1310
TUA1-run0	0.2053	TUA1-run0	0.1322
NKUST-run1	0.2057	NKUST-run1	0.1363
Baseline-run0	0.2088	TUA1-run1	0.1397
WUST-run0	0.2131	Baseline-run2	0.1442
RSLNV-run0	0.2141	Baseline-run0	0.1455
Baseline-run2	0.2288	RSLNV-run0	0.1483
TUA1-run1	0.2302	WUST-run0	0.1540
NKUST-run0	0.2653	NKUST-run0	0.2289
Baseline-run1	0.2811	Baseline-run1	0.2497

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

Run	Mean RSNOD	Run	Mean NMD
TUA1-run0	0.1615	TUA1-run0	0.1144
TUA1-run2	0.1617	IMTKU-run1	0.1165
IMTKU-run1	0.1631	IMTKU-run0	0.1181
IMTKU-run0	0.1648	TUA1-run2	0.1187
IMTKU-run2	0.1655	IMTKU-run2	0.1194
Baseline-run0	0.1782	TUA1-run1	0.1253
WUST-run0	0.1795	WUST-run0	0.1386
TUA1-run1	0.1810	Baseline-run0	0.1386
RSLNV-run0	0.1811	RSLNV-run0	0.1393
NKUST-run0	0.2222	NKUST-run1	0.1508
NKUST-run1	0.2295	Baseline-run2	0.1781
Baseline-run1	0.2425	NKUST-run0	0.1973
Baseline-run2	0.2614	Baseline-run1	0.2110

Table 7: Chinese Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
IMTKU-run0	0.0674	WUST-run0	0.1633
WUST-run0	0.0695	IMTKU-run0	0.1636
Baseline-run0	0.0709	Baseline-run0	0.1673
IMTKU-run1	0.0726	IMTKU-run1	0.1700
RSLNV-run0	0.0746	RSLNV-run0	0.1749
IMTKU-run2	0.0752	IMTKU-run2	0.1754
RSLNV-run2	0.0768	RSLNV-run2	0.1760
TUA1-run0	0.0859	TUA1-run0	0.1892
TMUDS-run1	0.0883	TMUDS-run2	0.1948
TMUDS-run2	0.0887	TMUDS-run1	0.1953
TMUDS-run0	0.0906	TMUDS-run0	0.1995
Baseline-run2	0.1301	Baseline-run2	0.2068
NKUST-run1	0.1905	NKUST-run1	0.3036
Baseline-run1	0.2858	NKUST-run0	0.4169
NKUST-run0	0.3116	Baseline-run1	0.4190

Result Highlights

- Two approaches outperform the LSTM baseline with statistical significance in Chinese subtasks
  - First time since STC-3 NDDQ
- However, the superiority is not statistically significant in English subtasks.
  - Current English dataset have fewer dialogues

DialEval will be back
(If our task proposal is accepted)

- New test data
- Fully bilingual dataset
  - English translation for all dialogues
  - Will be released before Dec 2021 (NTCIR-15 registration due)
- Please follow @ntcirdialeval on Twitter

Thank you!

Any question?