The NTCIR-16 Dialogue Evaluation Task (DialEval-2)

dialeval2org@list.waseda.jp

Tetsuya Sakai, Sijie Tao (Waseda University)

Inho Kang (Naver Corporation)

http://sakailab.com/dialeval2/

Version 20210325

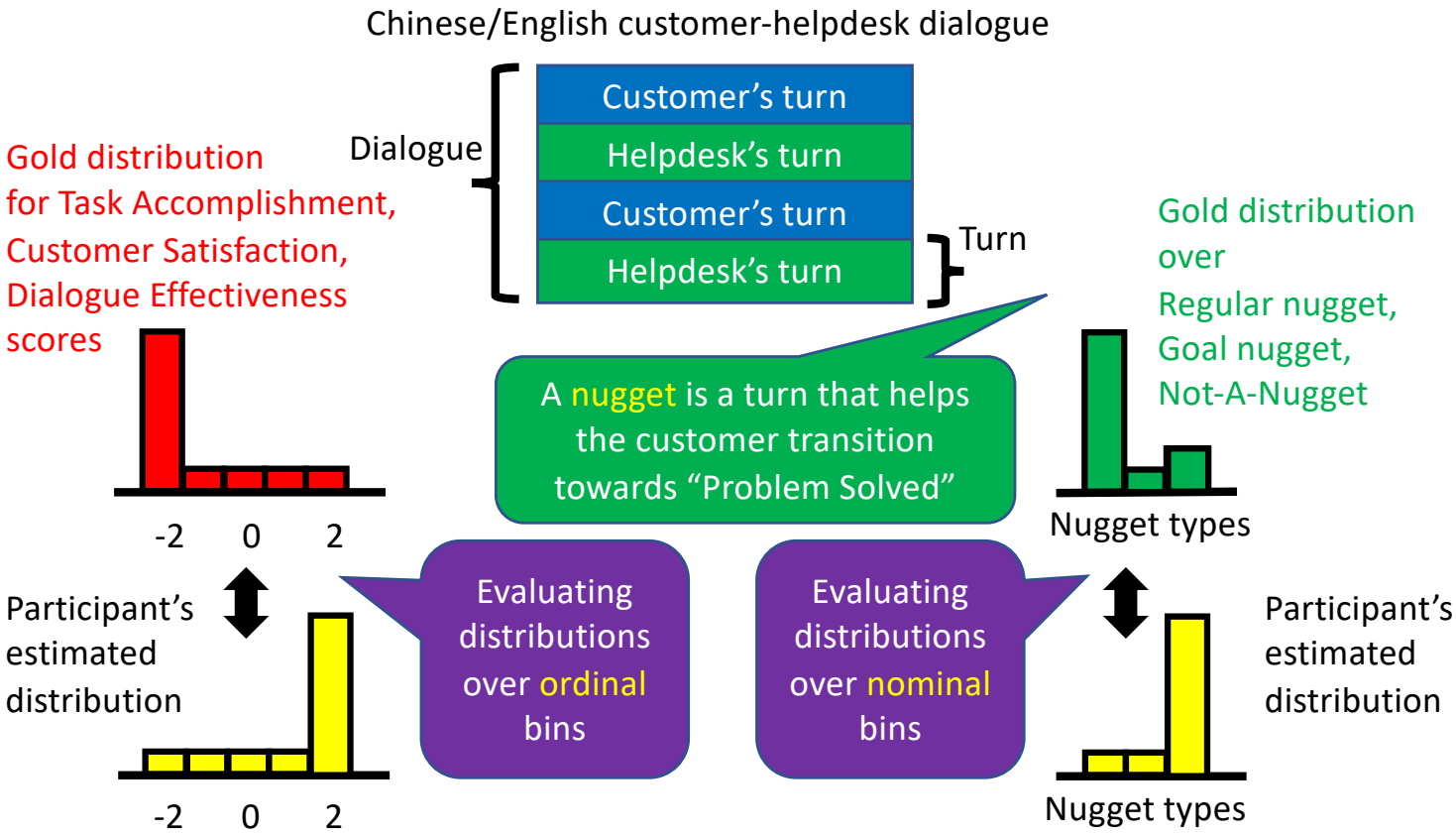
Introduction

- The NTCIR-16 Dialogue Evaluation Task (DialEval-2) hosts two subtasks, Dialogue Quality (DQ) and Nugget Detection (ND), which are exactly the same as those from NTCIR-14 STC-3 and NTCIR-15 DialEval-1.
- 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 (similar to dialogue acts) for each turn.
- Data: Chinese and English

Task overview (same as DialEval-1)

Dialogue Quality Subtask

Nugget Detection Subtask



Customer-Helpdesk dialogue: an example

C: I copied a picture from my PC to my mobile phone, but it kind of looks fuzzy on the phone. How can I solve this? P.S. I'm no good at computers and mobile phones.

Trigger

H: Please synchronise your PC and phone using iTunes first, and then upload your picture.

Solution

C: I'd done the synchronisation but did not upload it with XXX Mobile Assistant. I managed to do so by following your advice. You are a real expert, thank you!

Confirmation

H: You are very welcome. If you have any problems using XXX Mobile Phone Software, please contact us again, or visit XXX.com.

Dialogue Quality Subtask (1)

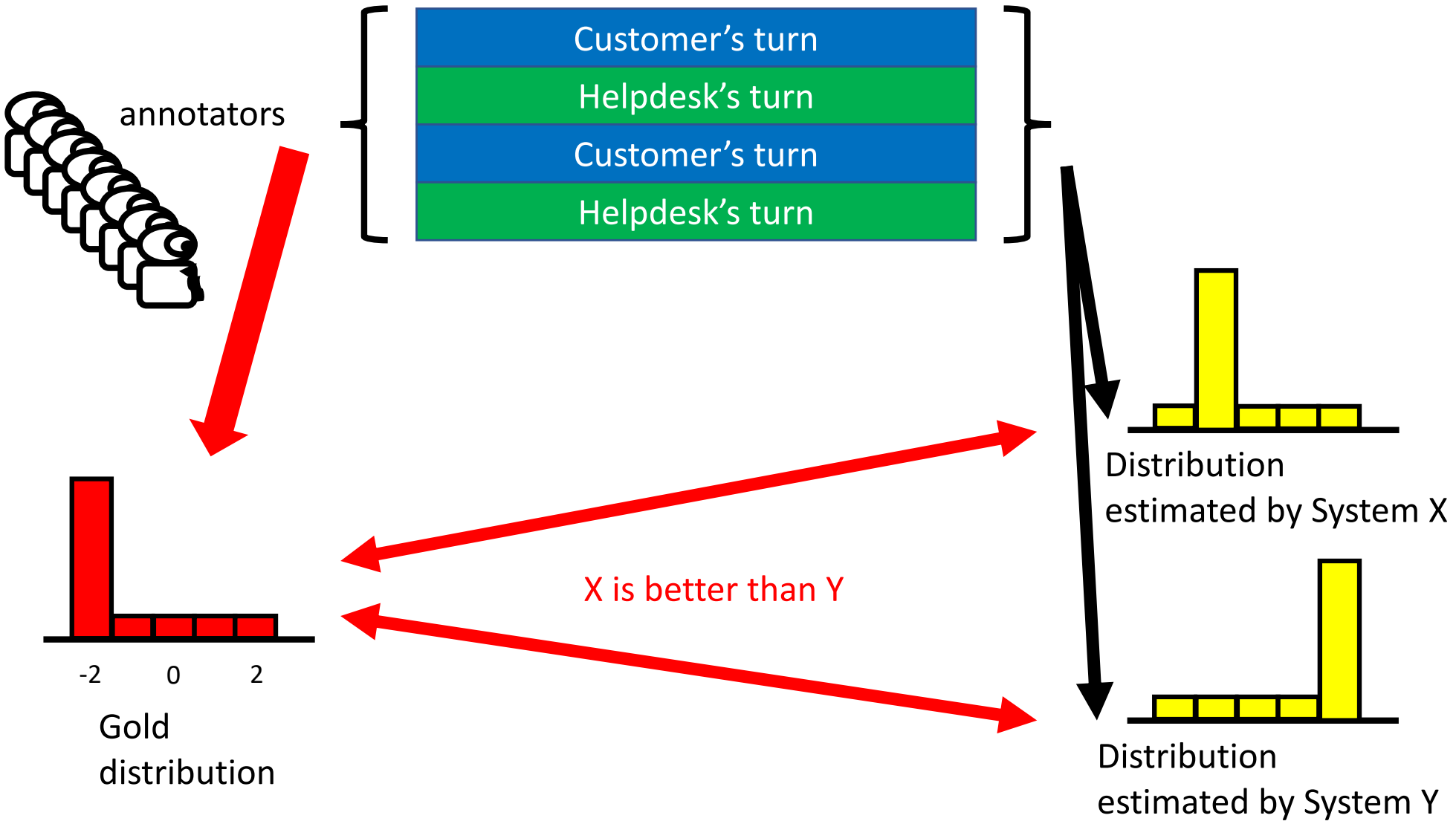
- Given a customer-helpdesk dialogue, return an estimated distribution of dialogue quality ratings for the entire dialogue.
- Three types of dialogue quality ratings (Likert scale -2 to 2):

A-score: Task Accomplishment

S-score: Customer Satisfaction (about the dialogue itself, not about the product/service)

E-score: Dialogue Effectiveness

Dialogue Quality Subtask (2)



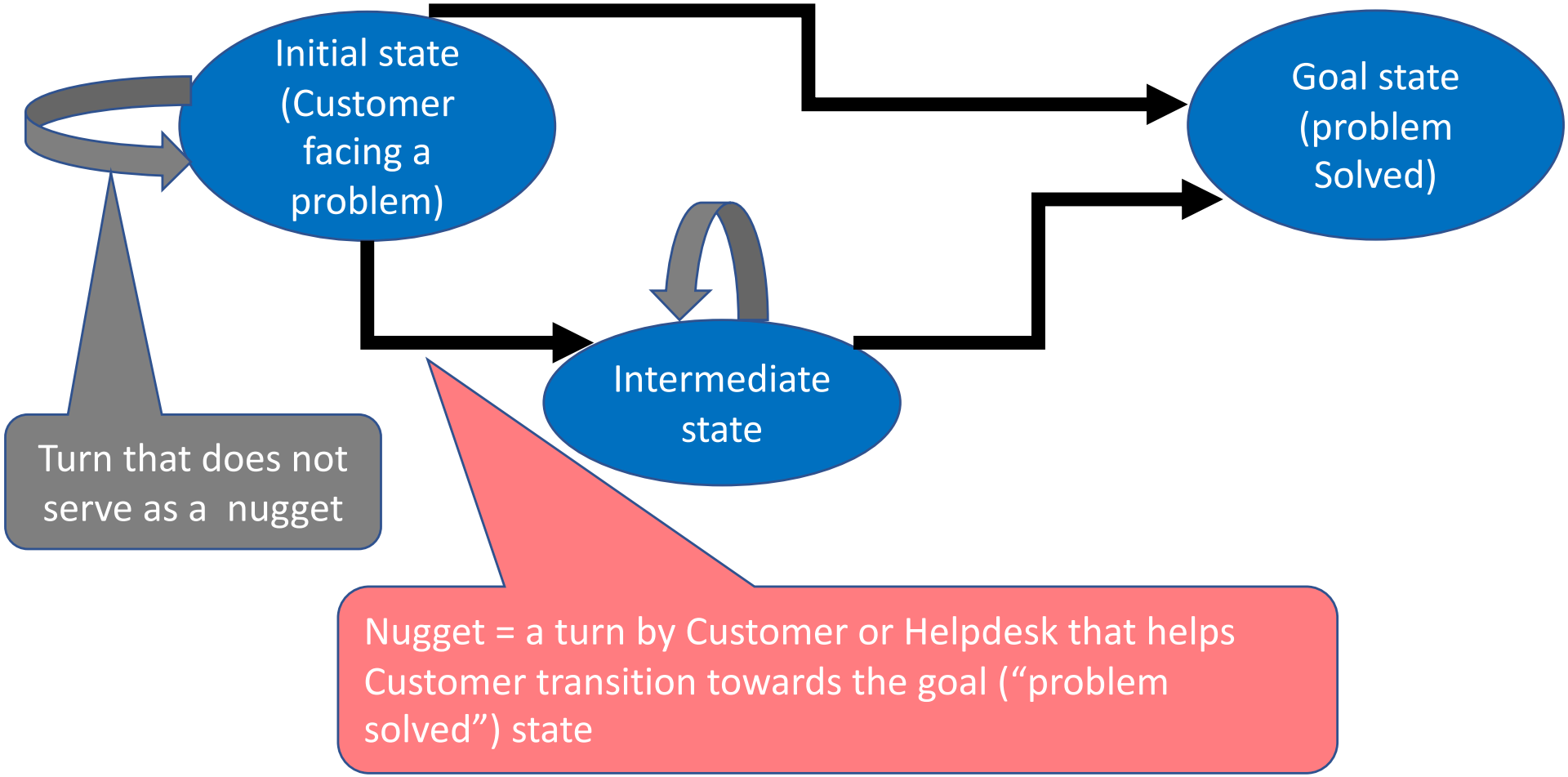
DQ evaluation measures for comparing gold and estimated distributions

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

For more info on the evaluation measures, see

https://waseda.box.com/SIGIR2018preprint

What is a nugget?

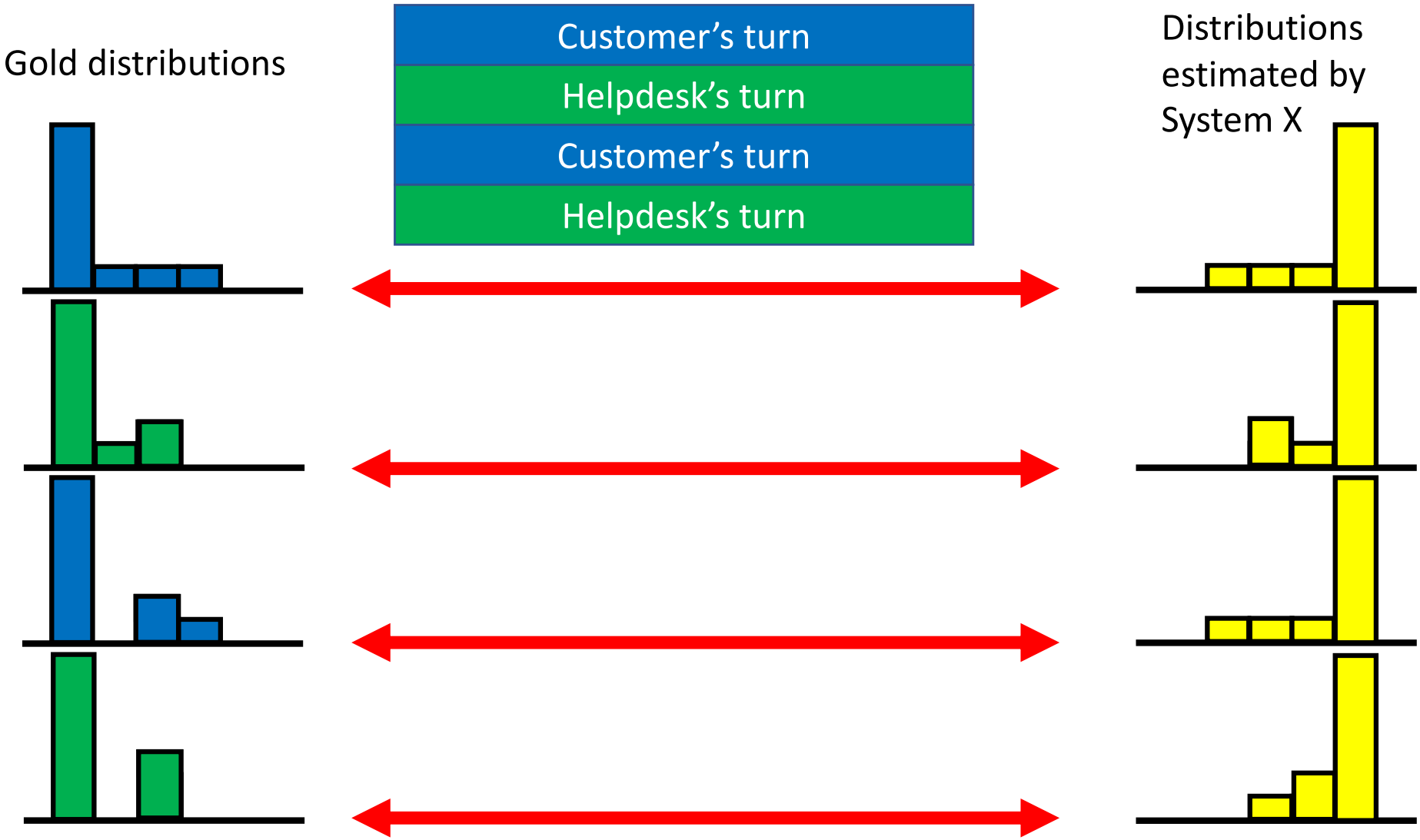


Nugget Detection Subtask (1)

- 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 (2)



ND evaluation measures for comparing gold and estimated distributions

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

For more info on the evaluation measures, see

https://waseda.box.com/SIGIR2018preprint

Why the task is important

- 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.

Training data: DCH-2 (Chinese-English parallel corpus with DQ and ND labels)

- Participants can obtain the data by sending an application form to us

Table 1: DCH-1 and DCH-2 dialogue corpus statistics.

	DCH-1 [11]	DCH-2
Data timestamps	Jan. 2013 - Sep. 2016	Jan. 2013 - Sep. 2016 (DCH-1 corpus) Oct. 2016 - Apr. 2018 (NTCIR-14 STC-3 NDDQ test dialogues [9]) Apr. 2018 - Jul. 2019 (NTCIR-15 DialEval-1 test dialogues [10])
#Chinese dialogues	3,700	4,390 (DCH-1 + 390 STC-3 + 300 DialEval-1)
#English translations	1,264 (34%)	4,390 (100%)
#Helpdesk accounts	161	161
Avg. #turns/dialogue	4.162	4.201
Avg. turn length (#chars)	48.31	54.541
# Annotators per dialogue	19	20

Test data sample size

Topic set size design:

http://link.springer.com/content/pdf/10.1007%2Fs10791-015-9273-z.pdf (open access)

Based on the DQ results from DialEval-1 with RNOD and NMD, we determined the number of test dialogues for DialEval-2. The most unstable measure was RNOD for the DQ-A subtask, whose residual variance was 0.00495. By plugging in this value to http://www.f.waseda.jp/tetsuya/samplesizeANOVA2.xlsx with m=10 systems and a minimum detectable difference minD=0.05 under Cohen's five-eighty convention, we obtained n=62. That is, 62 topics are enough for achieving 80% power for differences in mean RNOD of 0.05 or higher in terms ANOVA for 10 systems. Based on this result, we plan to prepare about 66 test dialogues, 11 for each i, where i is the number of turns in a dialogue (i=2,…,7).

Important Dates (Timezone: Japan (UTC+9))

April 2021	Task registrations open: training data given to participants
April-May 2021	Annotating the test dialogues
December 1 2021	Test dialogues released/Task registrations due
January 15 2022	Run submissions due
February 1, 2022	Evaluation results released
February 1, 2022	Draft task overview paper released
March 1, 2022	Draft participant paper submissions due
May 1, 2022	All camera-ready paper submissions due
June 2022	NTCIR-16 Conference in NII, Tokyo, Japan