NTCIR-15 Dialogue Evaluation Task Definition

dialeval1org@list.waseda.jp

Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai (Waseda University)

Inho Kang (Naver Corporation)

Version 20190819

Introduction

- The NTCIR-15 Dialogue Evaluation Task (DialEval-1) hosts two subtasks, Dialogue Quality (DQ) and Nugget Detection (ND), which are exactly the same as those from NTCIR-14 STC-3.
- 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

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)

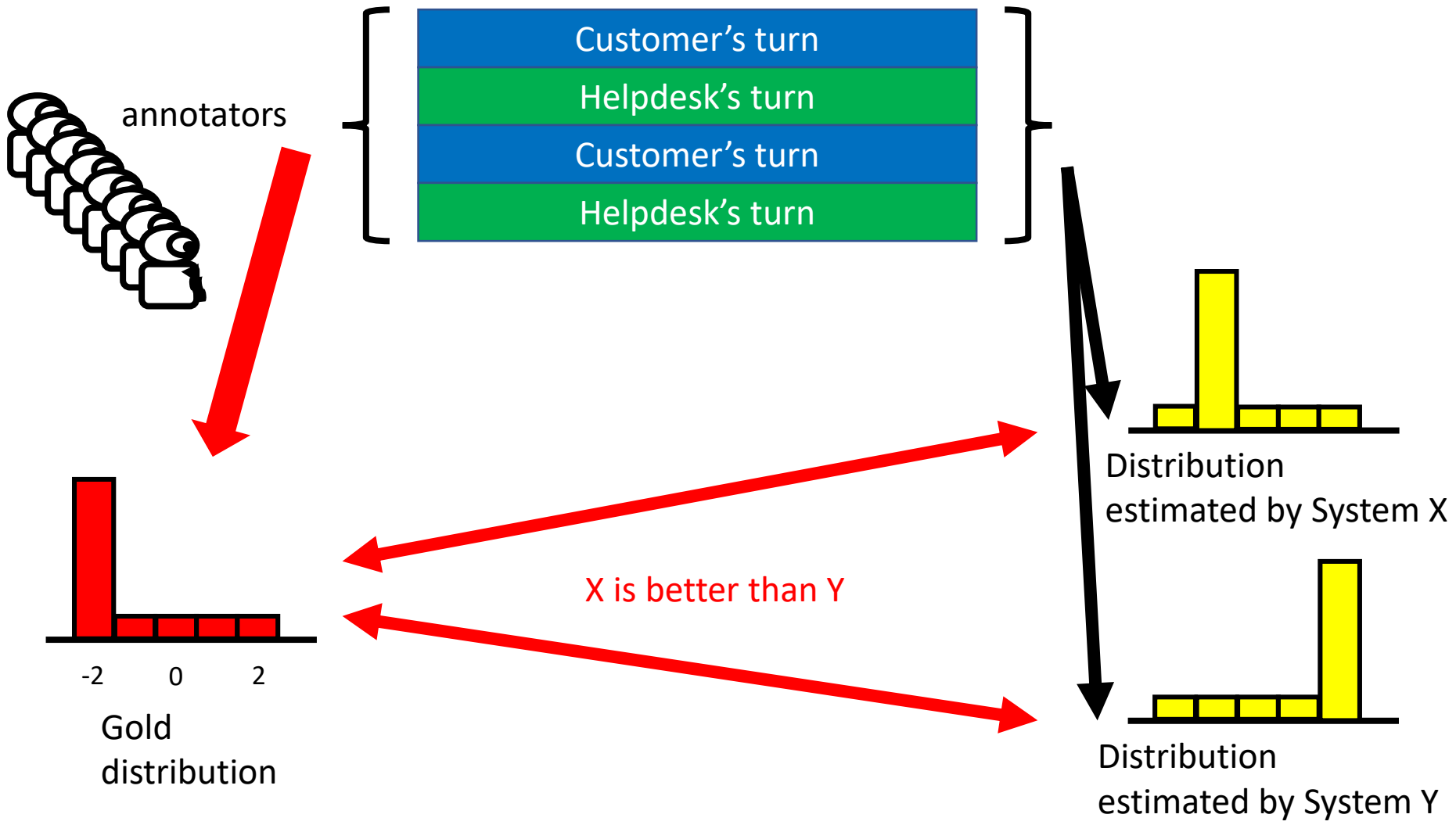
- Given a customer-helpdesk dialogue, return an estimated distribution of dialogue quality ratings for the entire dialogue.
- Three types of dialogue quality ratings (Likert scale -2 to 2):

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)



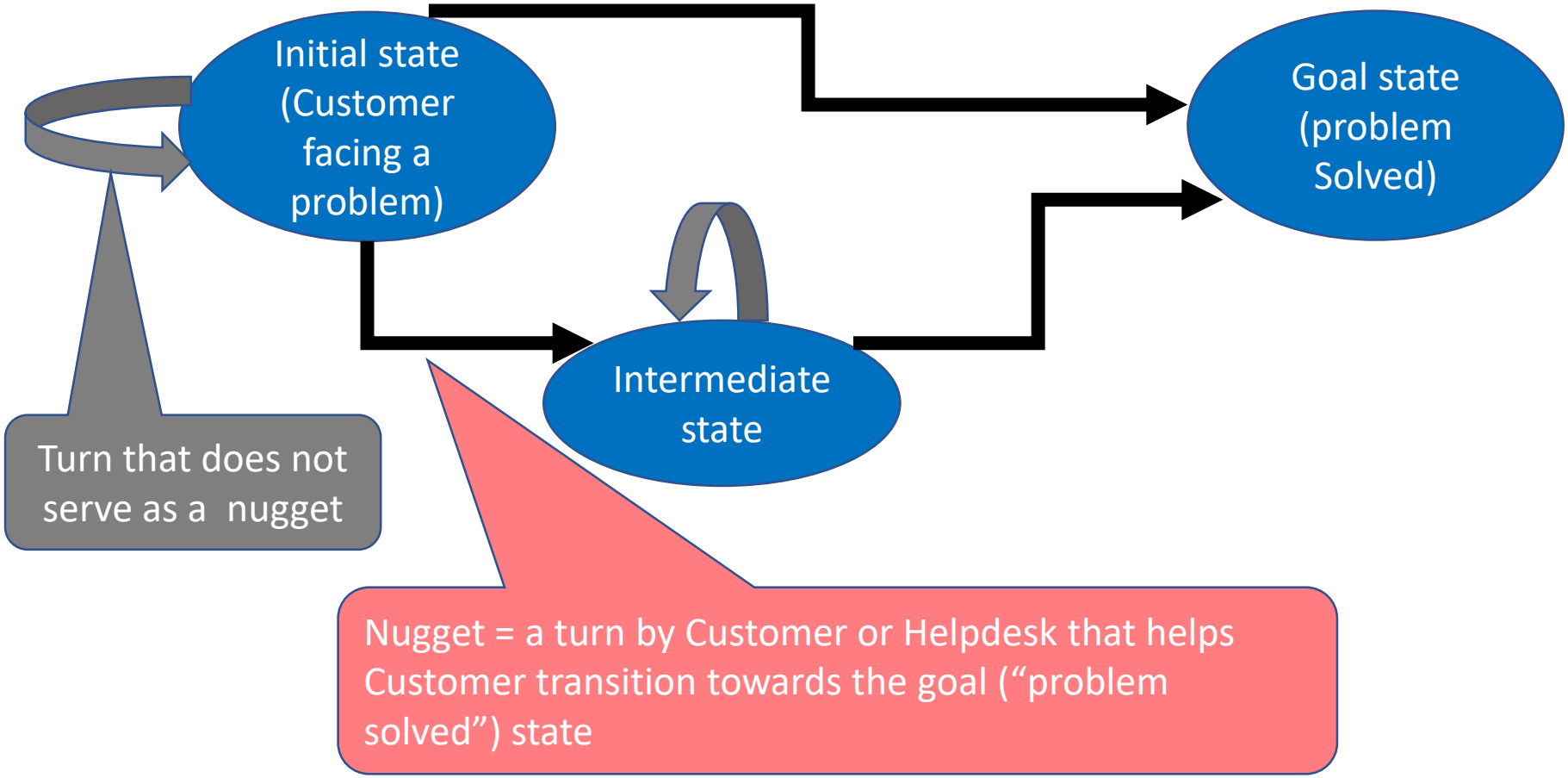
DQ evaluation measures for comparing gold and estimated distributions

- NMD (Normalised Match Distance)
- RSNOD (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?

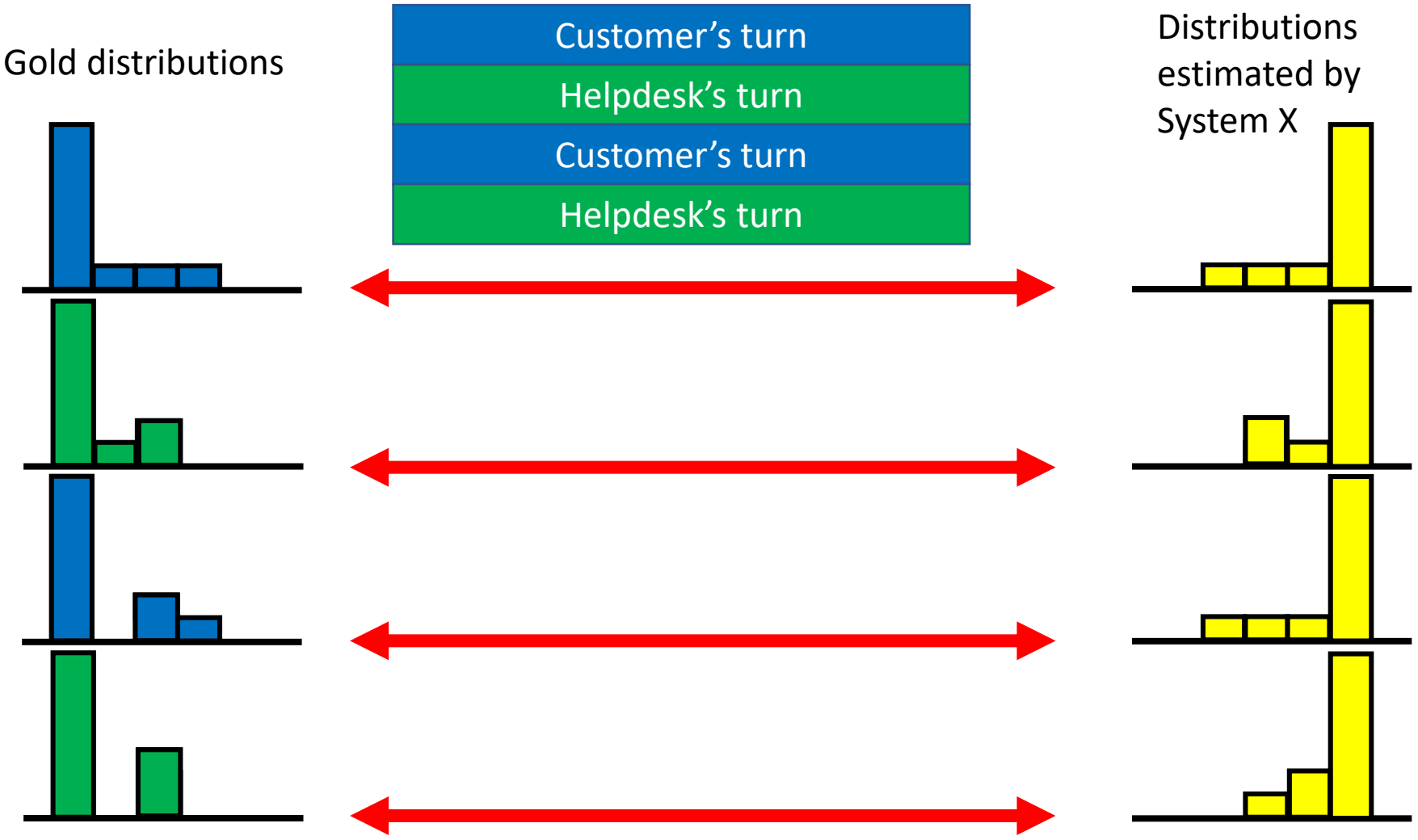


Nugget Detection Subtask (1)

- 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.