

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

Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai (Waseda University), Inho Kang (Naver)
dialeval1org@list.waseda.jp
 Homepage: <http://sakailab.com/dialeval1/>

Task as One Sentence

DQ: The **D**ialogue **Q**uality subtask
ND: The **N**ugget **D**etection subtask

requires the systems to **predict distributions** of

dialogue quality scores
 nugget types.

customer-helpdesk

dialogue: task-oriented, multi-round.

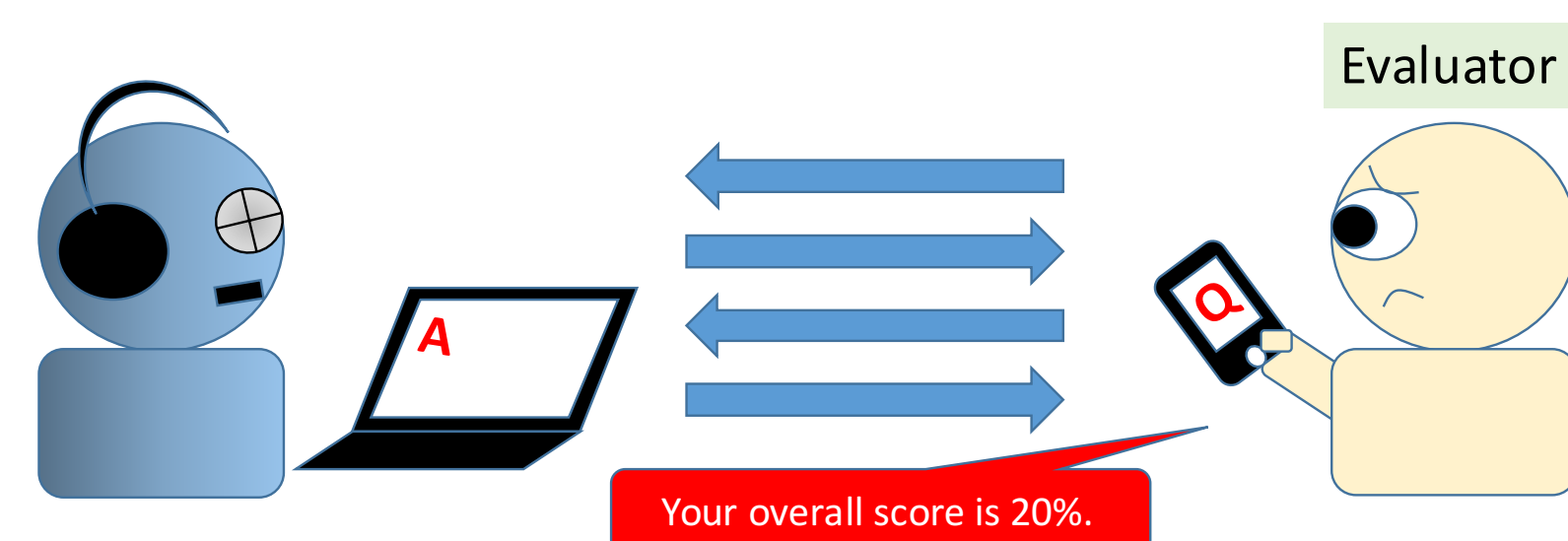
Motivation

Why prediction?

To build good dialogue system, we need good ways to evaluate them.

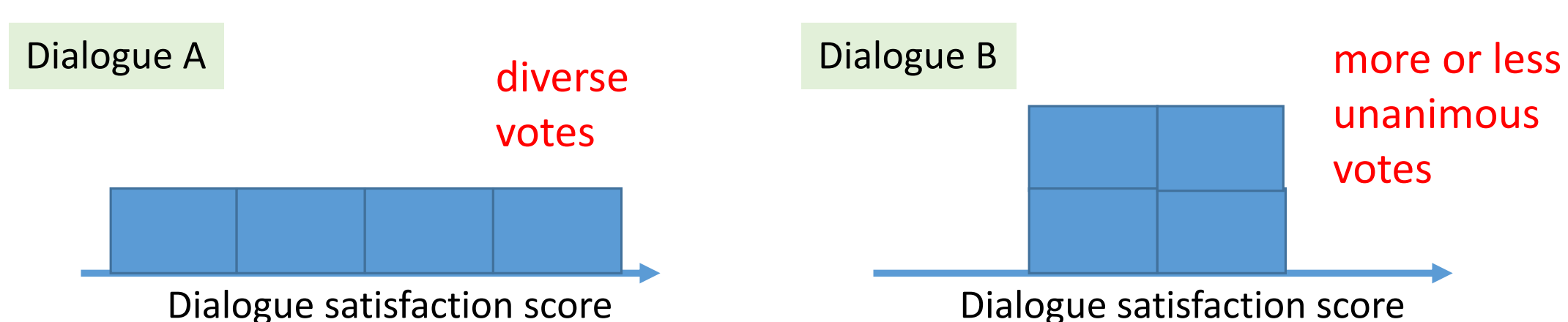
↔ Online evaluation :

- Expensive and does not scale
- Difficult to compare different systems
- Not repeatable



Why distributions?

People can have diverse views for the same dialogue.



3 quality types

- **A-score:** Task **A**ccomplishment (Has the problem been solved? To what extent?)
- **S-score:** Customer **S**atisfaction of the dialogue (not of the product/service or the company)
- **E-score:** Dialogue **E**ffectiveness (Do the utterers interact effectively to solve the problem efficiently?)

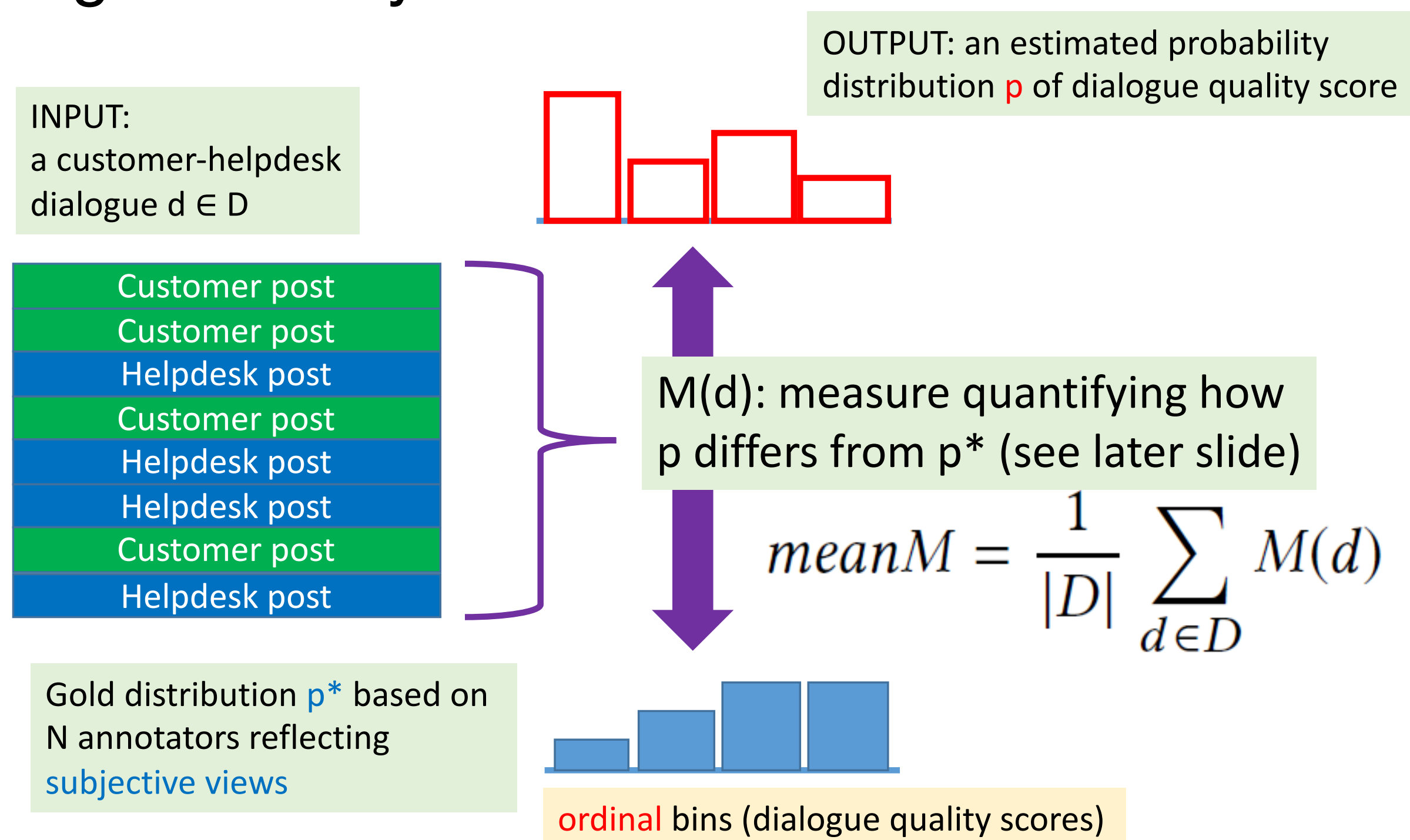
- **CNUG0:** Customer trigger (problem stated)
- **CNUG*:** Customer goal (solution confirmed)
- **HNUG*:** Helpdesk goal (solution stated)
- **CNUG:** Customer regular
- **HNUG:** Helpdesk regular
- **CNaN:** Customer Not-a-Nugget
- **HNaN:** Helpdesk Not-a-Nugget

Contains info that leads to solution

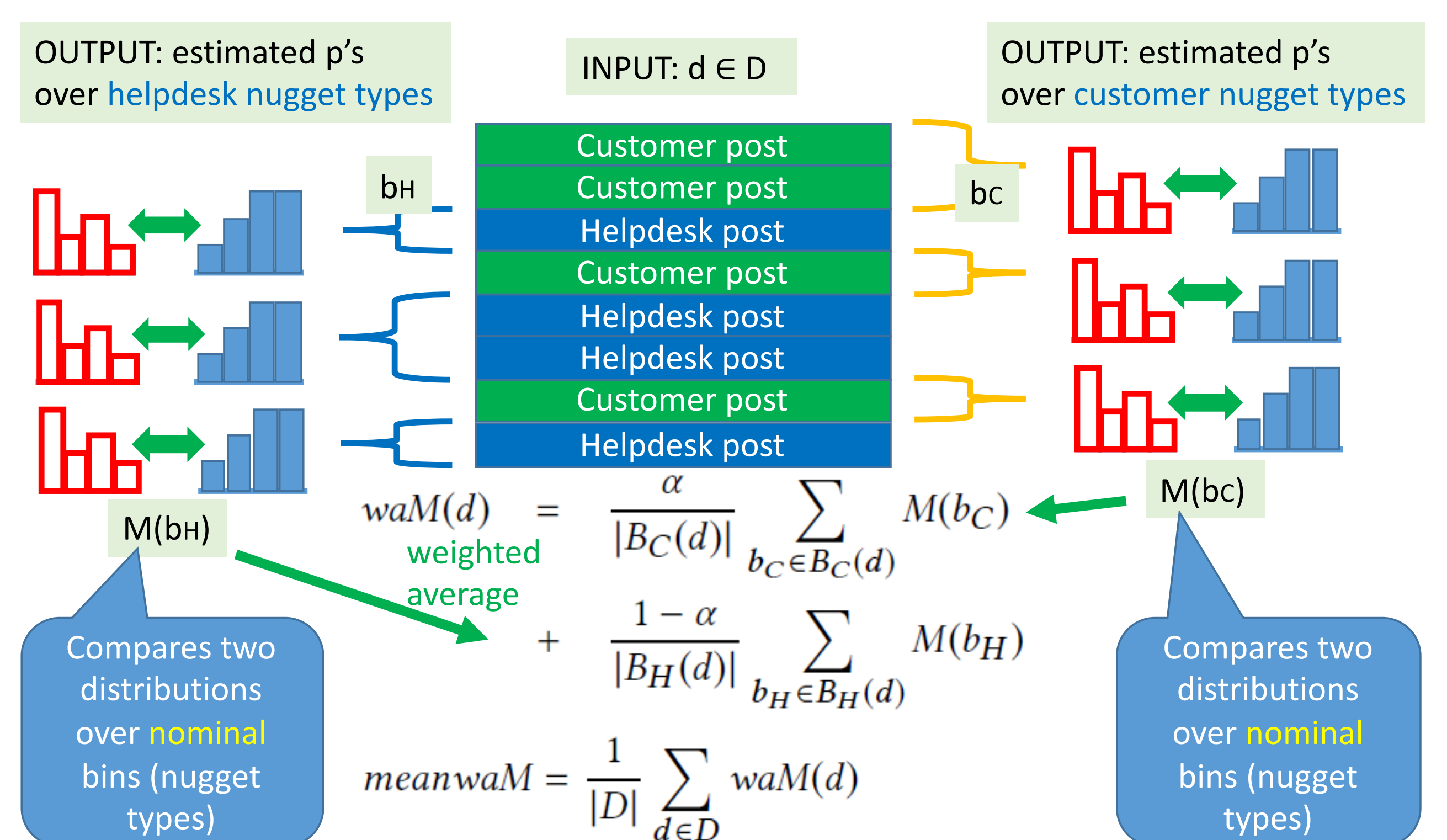
Does not contain info that leads to solution

Evaluation

Dialogue Quality subtask



Nugget Detection subtask



Dataset

- Chinese customer-helpdesk dialogues mined from Weibo, with annotations
- 3700 Training + 390 Dev + 300 test
- English translation partially available
 - 2252 Training + 390 Dev + 300 test

Summary

For the Chinese DQ and ND subtasks,
 - Two BERT-based models outperformed the BiLSTM (Bidirectional Long Short-term Memory) baseline model (BL-Istm) with statistical significance

For the English DQ and ND subtasks,
 - None of the models outperformed the BiLSTM baseline.

Results

Please see the overview paper.

Future

The English translation for all the dialogues will be available to form a fully bilingual dataset at next DialEval