NTCIR-14
Short Text Conversation Task (STC-3)

Sosuke Kato Zhaohao Zeng Tetsuya Sakai

(Waseda University)

Minlie Huang

(Tsinghua University)

stc3org@list.waseda.jp

http://sakailab.com/ntcir14stc3/

Twitter: @ntcirstc

Version 20180606

STC-1, -2, -3.

	Japanese	Chinese	English
NTCIR-12 STC-1 22 active participants	Twitter, Retrieval	Weibo, Retrieval	
NTCIR-13 STC-2 27 active participants	Yahoo! News, Retrieval+ Generation	Weibo, Retrieval+ Generation	
NTCIR-14 STC-3		Weibo, Generation for given emotion categories	
		Weibo+English translations, distribution estimation for subjective annotations	

Chinese Emotional Conversation Generation (CECG) subtask

Dialogue Quality (DQ) and Nugget Detection (ND) subtasks

Single-round, Non task-oriented

Multi-round, task-oriented (helpdesk)

STC-3 subtasks

- Chinese Emotional Conversation Generation (CECG): for details, please visit

http://coai.cs.tsinghua.edu.cn/hml/challenge/

- Dialogue Quality (DQ):

please read this slide deck

- Nugget Detection (ND):

please read this slide deck

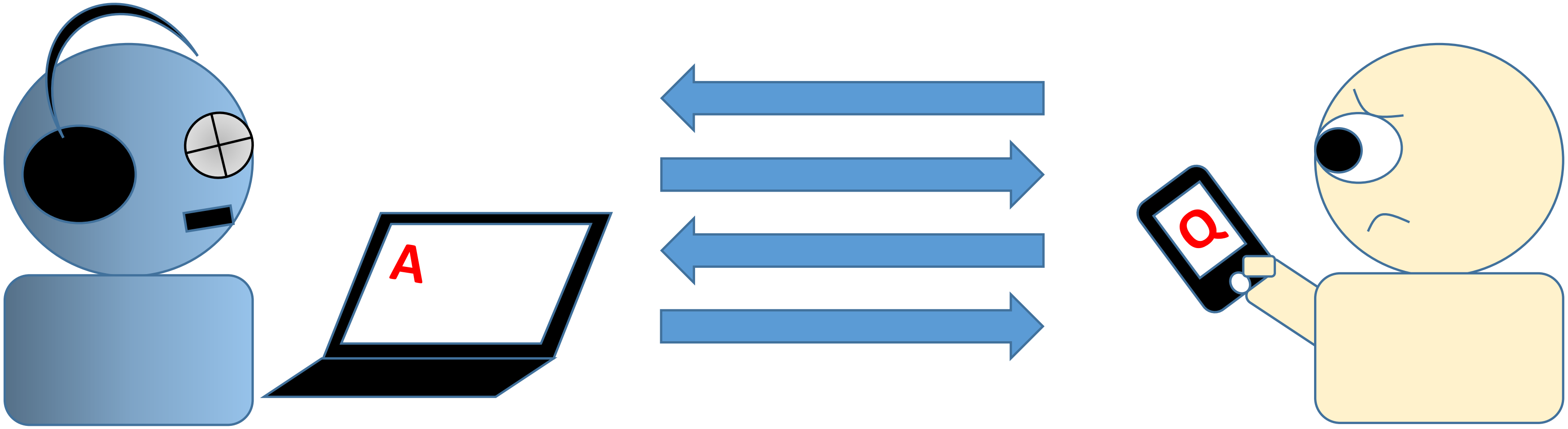
Dialogue Quality (DQ) and Nugget Detection (ND) subtasks of STC-3@NTCIR-14

Sosuke Kato Zhaohao Zeng Tetsuya Sakai

(Waseda University)

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

Evaluator
A
a
Your overall score is 20%.

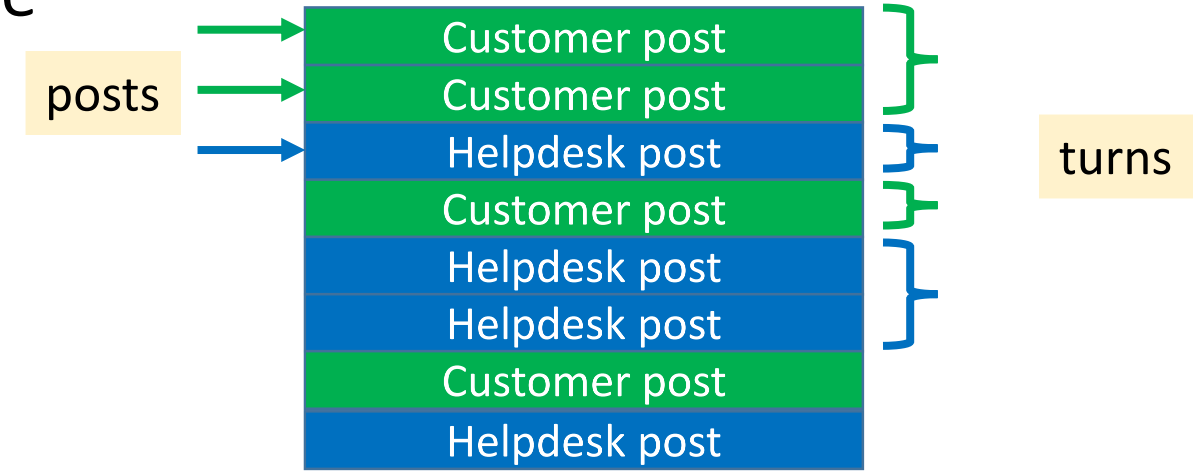
Posts, Turns

- Post

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

- Turns (or utterance blocks)

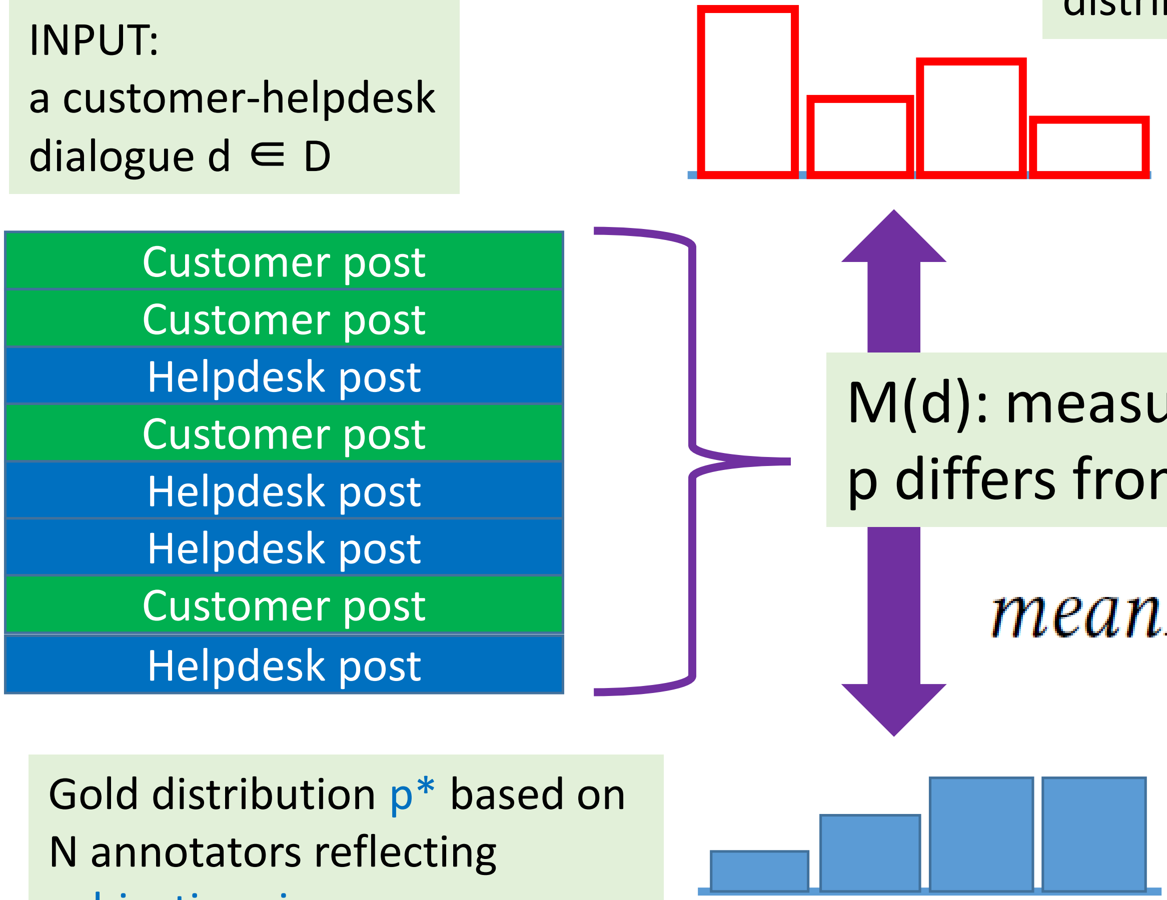
Maximal consecutive posts by the same utterer in a dialogue



NTCIR-14 STC-3 (Chinese and English)
Dialogue Quality subtask

OUTPUT: an estimated probability distribution p of dialogue quality score

INPUT:
a customer-helpdesk dialogue d ∈ D



Customer post
Customer post
Helpdesk post
Customer post
Helpdesk post
Helpdesk post
Customer post
Helpdesk post

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

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

Gold distribution p* based on N annotators reflecting subjective views

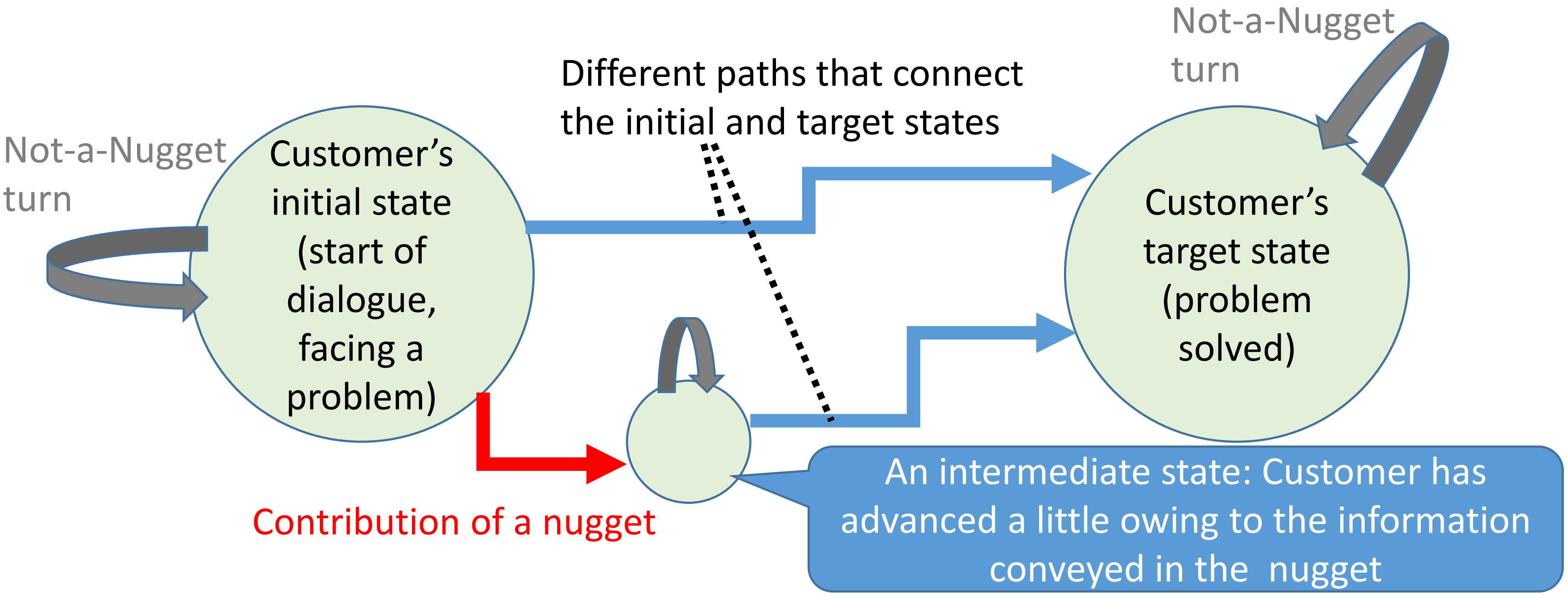
ordinal bins (dialogue quality scores)

Dialogue Quality: target scores

- 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

Nuggets

- A nugget is an 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).



Not-a-Nugget turn

Customer's initial state (start of dialogue, facing a problem)

Different paths that connect the initial and target states

Not-a-Nugget turn

Customer's target state (problem solved)

Contribution of a nugget

An intermediate state: Customer has advanced a little owing to the information conveyed in the nugget

Nugget types

- CNUG0: Customer trigger (problem stated)
- CNUG*: Customer goal (solution confirmed)
- HNUG*: Helpdesk goal (solution stated)
- CNUG: Customer regular
- HNUG: Helpdesk regular

Contains info that leads to solution

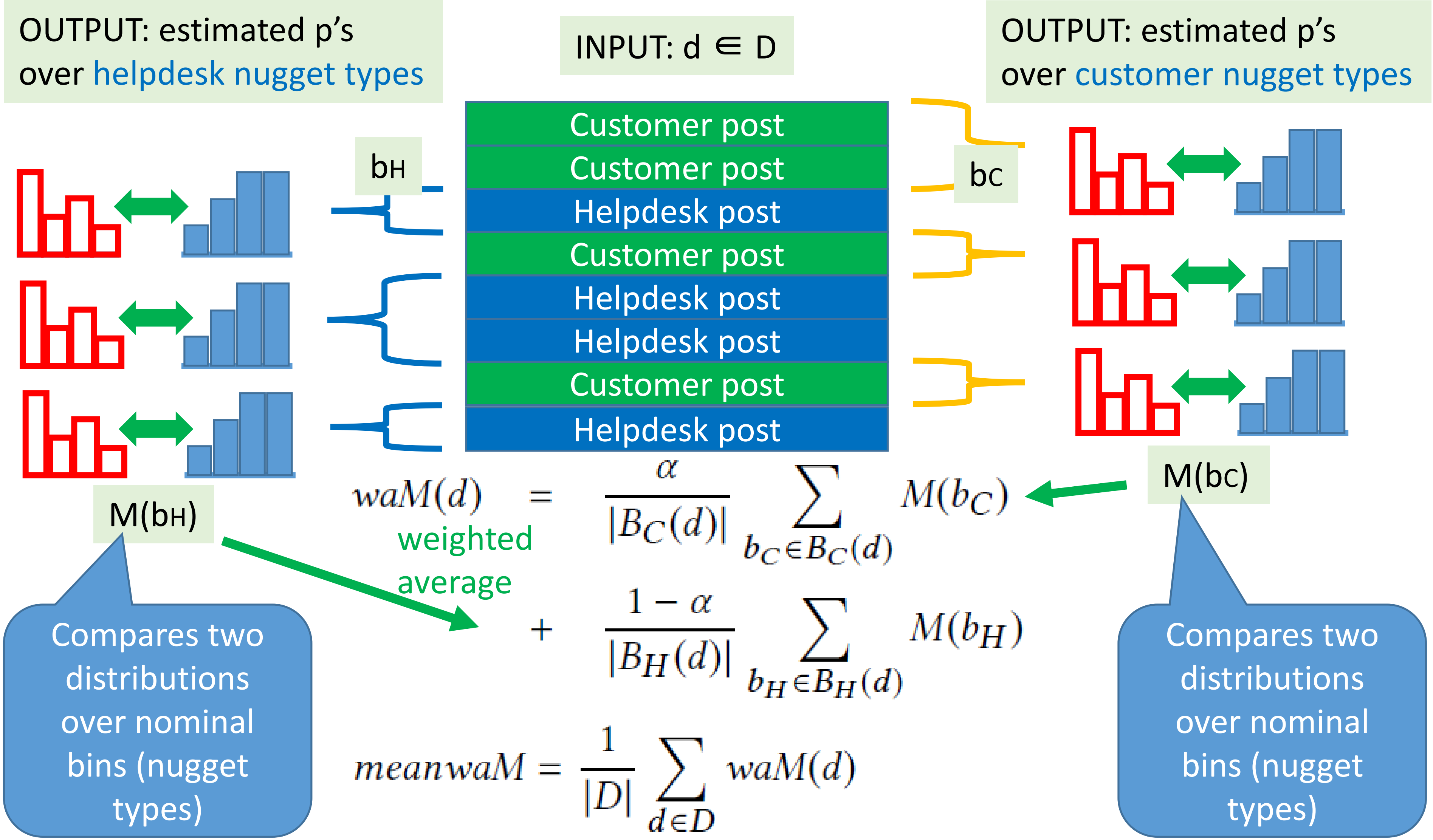
- CNaN: Customer Not-a-Nugget
- HNaN: Helpdesk Not-a-Nugget

Does not contain info that leads to solution

Nugget types: 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.	CNUG0 (problem stated)
H: Please synchronise your PC and phone using iTunes first, and then upload your picture.	HNUG* (solution stated)
C: I’d done the synchronization 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!	CNUG* (solution confirmed)
H: You are very welcome. If you have any problems using XXX Moble Phone Software, please contact us again, or visit XXX.com.	HNaN (Not-a-Nugget)

NTCIR-14 STC-3 (Chinese and English)
Nugget Detection subtask



OUTPUT: estimated p's over helpdesk nugget types

INPUT: d ∈ D

OUTPUT: estimated p's over customer nugget types

Customer post

Customer post

Helpdesk post

Customer post

Helpdesk post

Helpdesk post

Customer post

Helpdesk post

b_H

b_C

M(b_H)

M(b_C)

$$waM(d) = \frac{\alpha}{|B_C(d)|}\sum_{b_C\in B_C(d)} M(b_C) + \frac{1-\alpha}{|B_H(d)|}\sum_{b_H\in B_H(d)} M(b_H)$$

weighted average

$$meanwaM = \frac{1}{|D|}\sum_{d\in D} waM(d)$$

Compares two distributions over nominal bins (nugget types)

Compares two distributions over nominal bins (nugget types)

Why nuggets?

- If nuggets can be detected automatically, they may serve as useful features for automatically estimating the dialogue quality.
- Automatic nugget detection may help us diagnose a dialogue closely (why it failed, where it failed).
- Ultimately, experiences from the nugget detection subtask may help us design Helpdesk systems that provide the solution to a given problem effectively and efficiently.

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

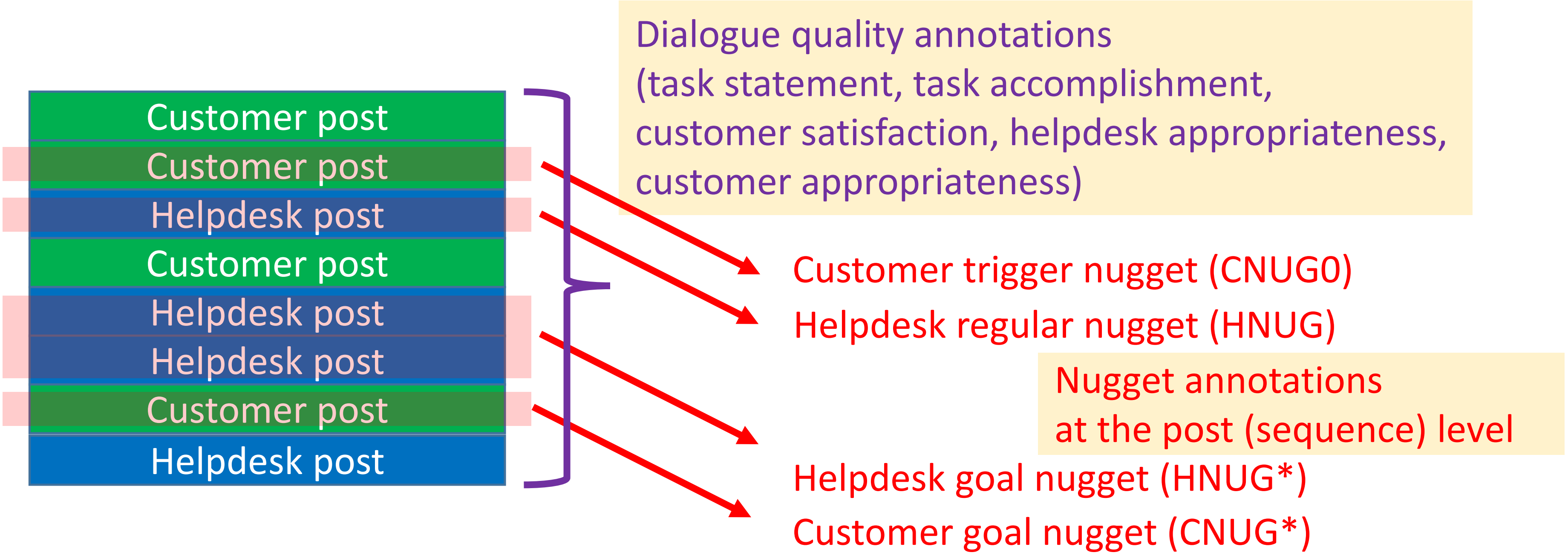
as M(b) for each turn b.

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

DCH-1 Chinese dialogue test collection [Zeng+17]

http://waseda.box.com/DCH-0-1
http://ceur-ws.org/Vol-2008/paper_1.pdf

- 3700 Chinese customer-helpdesk dialogues mined from Weibo, with annotations
- English translation available for 1264 (34%) of DCH-1 (more will be translated May-June)



Dialogue quality annotations
(task statement, task accomplishment, customer satisfaction, helpdesk appropriateness, customer appropriateness)

Customer trigger nugget (CNUG0)

Helpdesk regular nugget (HNUG)

Nugget annotations
at the post (sequence) level

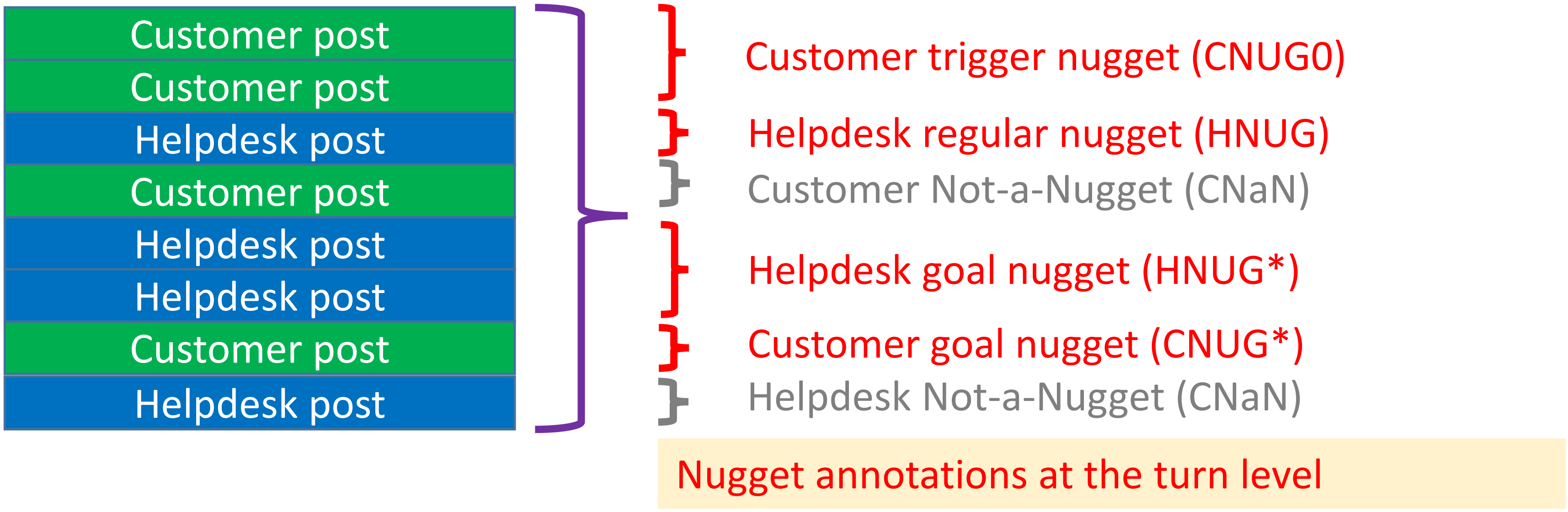
Helpdesk goal nugget (HNUG*)

Customer goal nugget (CNUG*)

Constructing STC-3 training data from DCH-1 (May-Aug)

- DCH-1 will be re-annotated for the Dialogue Quality (A-score, S-score, E-score) and the Nugget Detection (CNUG0, CNUG*, HNUG*, CNUG, HNUG, CNaN, HNaN) subtasks by 10-20 annotators per dialogue

Dialogue quality annotations
(A-score, S-score, E-score distributions)



Customer post
Customer post
Helpdesk post
Customer post
Helpdesk post
Helpdesk post
Customer post
Helpdesk post

Customer trigger nugget (CNUG0)
Helpdesk regular nugget (HNUG)
Customer Not-a-Nugget (CNaN)
Helpdesk goal nugget (HNUG*)
Customer goal nugget (CNUG*)
Helpdesk Not-a-Nugget (CNaN)

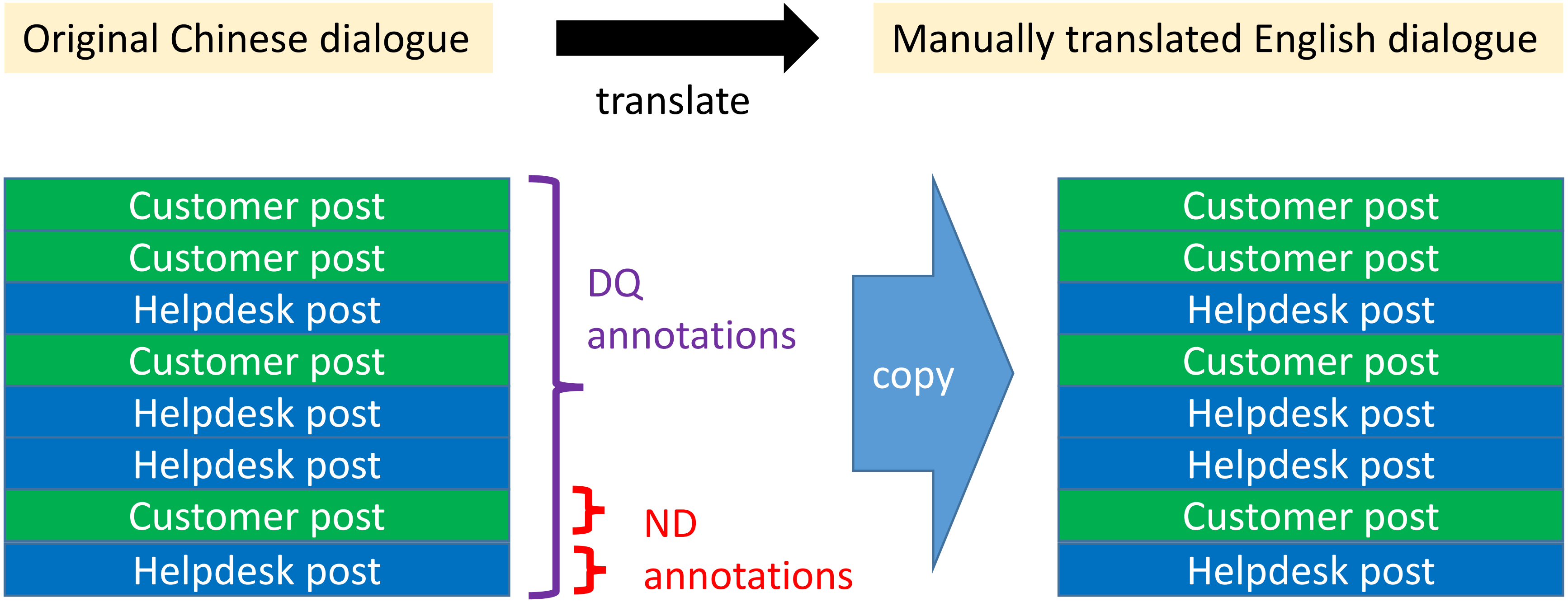
Nugget annotations at the turn level

Test data

- To be crawled in April (about 300 Chinese Weibo helpdesk/customer dialogues)
- To be annotated in May-Aug
- To be translated into English in May-June

On annotations

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



Original Chinese dialogue

translate

Manually translated English dialogue

Customer post
Customer post
Helpdesk post
Customer post
Helpdesk post
Helpdesk post
Customer post
Helpdesk post

DQ annotations

ND annotations

copy

Customer post
Customer post
Helpdesk post
Customer post
Helpdesk post
Helpdesk post
Customer post
Helpdesk post

Schedule for DQ and ND subtasks (incl. generic NTCIR-14 schedule)

Oct-Dec, 2017 Training data translation into English

1264 out of 3700 (34%) done

April 2018 Crawling Chinese test data from Weibo, develop an annotation tool

May-Jun, 2018 Test data + additional training data translation into English

May-Aug, Training/test Chinese data annotation

Aug 31, 2018 STC-3 task registrations due (CECG, DQ, ND)

Sep 1, 2018 Training data with annotations released

Nov 1, 2018 Test data released

Nov 30, 2018 Run submissions due

Feb 1, 2019 Results summary and draft overview released

Mar 15, 2019 Participant paper submissions due

May 1, 2019 All camera-ready papers due

Jun 2019 NTCIR-14 Conference@NII, Tokyo