Towards Automatic Evaluation of Multi-Turn Dialogues: A Task Design that Leverages Inherently Subjective Annotations

Tetsuya Sakai
Waseda University, Tokyo, Japan
tetsuyasakai@acm.org

ABSTRACT

This paper proposes a design of a shared task whose ultimate goal is automatic evaluation of multi-turn, dyadic, textual helpdesk dialogues. The proposed task takes the form of an offline evaluation, where participating systems are given a dialogue as input, and output at least one of the following: (1) an estimated distribution of the annotators' quality ratings for that dialogue; and (2) an estimated distribution of the annotators' nugget type labels for each utterance block (i.e., a maximal sequence of consecutive posts by the same utterer) in that dialogue. This shared task should help researchers build automatic helpdesk dialogue systems that respond appropriately to inquiries by considering the diverse views of customers. The proposed task has been accepted as part of the NTCIR-14 Short Text Conversation (STC-3) task. While estimated and gold distributions are traditionally compared by means of root mean squared error, Jensen-Shannon divergence and the like, we propose a pilot measure that considers the order of the probability bins for the dialogue quality subtask, which we call Symmetric Normalised Order-aware Divergence (SNOD).

KEYWORDS

dialogues; divergence; evaluation; nuggets; probability distributions; test collections

1 INTRODUCTION

More and more companies are providing online customer services where a customer can exchange realtime textual messages about the company's services and products with a (probably human) helpdesk operator. This means more convenience for the customers, but more burden on the companies. Hence, research in automatic helpdesk dialogue systems is highly practical as a means to reduce the cost for the companies. To design and tune automatic dialogue systems efficiently and in a costly manner, automatic evaluation of dialogue quality is desirable.

As an initial step towards automatic evaluation of helpdesk dialogue systems, this paper proposes a design of a shared task. The proposed task takes the form of an offline evaluation, where participating systems are given a dialogue as input, and output at least one of the following: (1) an estimated distribution of the annotators' quality ratings for that dialogue; and (2) an estimated distribution of the annotators' nugget type labels for each *utterance block* (i.e., a maximal sequence of consecutive posts by the same utterer) in that dialogue. This shared task should help researchers build automatic helpdesk dialogue systems that respond appropriately to inquiries by considering the diverse views of customers. The proposed task has been accepted as part of the NTCIR-14 Short Text Conversation (STC-3) task.

While estimated and gold distributions are traditionally compared by means of *root mean squared error*, *Jensen-Shannon divergence* [6] and the like, we propose a pilot measure that considers the *order* of the probability bins for the dialogue quality subtask, which we call *Symmetric Normalised Order-aware Divergence (SNOD)*.

EVIA 2017, co-located with NTCIR-13, Tokyo, Japan.

2 RELATED WORK

2.1 Dialogue Evaluation in Brief

While to our knowledge the task proposed in the present paper is novel, dialogue evaluation is not a new problem. For example, it was in 1997 that Walker *et al.* [13] proposed the PARADISE (PARAdigm for Dialogue System Evaluation) framework for evaluating spoken dialogue systems in the train timetable domain. In 2000, Hone and Graham [5] proposed the questionnaire-based SASSI (Subjective Assessment of Speech System Interfaces) for evaluating an in-car speech interface. However, existing studies along these lines of research mostly focus on closed-domain applications. The topics that helpdesks need to deal with are far more diverse.

Recently, Lowe *et al.* [7] released the Ubuntu dialogue corpus and proposed a *response selection* task: systems are given a dialogue context, one correct response immediately following the context plus nine "fake" responses sampled from outside the dialogue, and are required to select one or more appropriate responses from them. Their effort is more similar to ours in that the topics discussed in the dialogues are more diverse than those dealt with by traditional dialogue evaluation. However, since their "correct" response is the original response from the dialogue in their task, their task does not involve manual annotations at all. In contrast, the present study addresses the problem of annotators' subjective decisions that may be unanimous in some cases but contradictory in others. In fact, our proposal is to preserve the diverse views in the annotations "as is" and leverage them at the step of evaluation measure calculation, as we shall describe in Section 3.

There are also a few recent efforts in evaluating *non-task-oriented* dialogues, or dialogues without a specific purpose (e.g. [2]). The *Dialogue Breakdown Detection Challenge* [3, 4] (Section 2.2) and the NTCIR *Short Text Conversation* task [12] (Section 2.3) are also non-task-oriented. However, we are more interested in helpdesk dialogues that try to solve a specific problem that the customer is facing.

2.2 Dialogue Breakdown Detection

The Dialogue Breakdown Detection Challenge (DBDC) [4] provides human-machine non-task-oriented chats to participating systems.

Participating systems are required to examine each machine utterance, and determine the likelihood that the utterance caused a dialogue breakdown (i.e., a point where it becomes difficult to continue a proper conversation any further due to an inappropriate utterance). More specifically, the system is required, for each machine utterance, to output an estimated distribution of multiple annotators over three categories: **NB** (not a breakdown), **PB** (possible breakdown), and **B** (breakdown). This enabled the task to evaluate systems by comparing the system's estimated distribution with the gold distribtution of the annotators in terms *Mean Squared Error* and *Jensen-Shannon divergence* (See Section 3.2.1). The Third DBDC [3] will be concluded at *Dialog System Technology Challenges* (DSTC6) on December 10, 2017[1].

Our proposed task was directly inspired by DBDC, which reflects the view that the annotations by different people can be inherently different, and that systems should be aware of that. We believe that this is particularly important for dialogue systems that need to face diverse customers, often in the absence of absolute truths. Thus, instead of trying to consolidate multiple annotations to form a single gold label, we represent the gold data as a distribution of annotators; we also require systems to produce estimated distributions, rather than an estimated judgement of an "average" person[2]. One important point to note is that while the probability bins (i.e., the categories) of DBDC are *ordered* (e.g., **PB** is closer to **NB** than **B** is), the aforementioned measures do not take this into account. In the present study, we introduce a pilot measure called Symmetric Normalised Order-aware Divergence (SNOD) as an attempt to solve this issue.

2.3 Short Text Conversation

The NTCIR Short Text Conversation (STC) task [11, 12], the largest task in NTCIR-12 and -13, also handles non-task-oriented dialogues. However, their task setting has so far considered *single-turn* dialogues only: given a Chinese Weibo[3] post (in the Chinese subtask), can participating systems either retrieve or generate an appropriate response?

While the STC task also hires multiple assessors and require them to label tweets based on four criteria (fluent, coherent, self-sufficient, substantial[4]), they consolidate the labels of the multiple assessors to form the final graded relevance level (e.g., relevant and highly relevant). While Sakai's *unanimity-aware* gains [9] were applied for the NTCIR-13 STC-2 Chinese subtask to weight unanimous ratings more heavily compared to controversial ones, the task did not involve direct comparisons of gold and system distributions.

As was mentioned earlier, the framework proposed in the present study has been accepted as part of the NTCIR-14 STC-3 task.

2.4 DCH-1 Test Collection

Recently, Zeng *et al.* [14] reported on a Chinese helpdesk-customer dialogue test collection and proposed a nugget-based evaluation measure called *UCH*, which was adapted from an information retrieval evaluation measure called *U-measure* [10]. They hired three annotators per dialogue (helpdesk-customer interactions mined from Weibo) and obtained dialogue-level *quality* annotations as well as *nugget* annotations, where a nugget is a minimal sequence of consecutive posts by the same utterer that helps towards problem solving. In essence, a nugget is a "relevant" portion within an utterance block.

Each of the three annotators independently provided the following dialogue-level quality labels for each dialogue [14]:

- **TS** *Task Statement*: whether the task (i.e., the problem to be solved) is clearly stated by Customer (scores: $\{-1, 0, 1\}$);
- **TA** *Task Accomplishment*: whether the task is actually accomplished (scores: $\{-1, 0, 1\}$);
- **CS** *Customer Satisfaction*: whether Customer is likely to have been satisfied with the dialogue, and to what degree (scores: $\{-2, -1, 0, 1, 2\}$);
- **HA** *Helpdesk Appropriateness*: whether Helpdesk provided appropriate information (scores: $\{-1, 0, 1\}$);
- **CA** *Customer Appropriateness*: whether Customer provided appropriate information (scores: $\{-1, 0, 1\}$).

Moreover, they independently identified the following types of nuggets within each utterance block [14]:

- **CNUG0** Customer's *trigger nuggets*. These are nuggets that define Customer's initial problem, which directly caused Customer to contact Helpdesk.
- **HNUG** Helpdesk's *regular nuggets*. These are nuggets in Helpdesk's utterances that are useful from Customer's point of view.
- **CNUG** Customer's *regular nuggets*. These are nuggets in Customer's utterances that are useful from Helpdesk's point of view.
- **HNUG∗** Helpdesk's *goal nuggets*. These are nuggets in Helpdesk's utterances which provide the Customer with a solution to the problem.
- **CNUG∗** Customer's *goal nuggets*. These are nuggets in Customer's utterances which tell Helpdesk that Customer's problem has been solved.

In our proposed task design, we tentatively use the aforementioned annotation scheme of DCH-1, so that we can discuss our ideas with concrete examples. However, it should be noted that our proposal does not require that the dialogue-level and nugget annotations are done in exactly the same way as above. If we do use the above schema in a new task, however, it would enable us to directly utilise the DCH-1 test collection as training data for the participants, as we shall describe in the next section.

3 PROPOSED TASK DESIGN

Our ultimate goal is the automatic evaluation of Helpdesk-Customer (be it human-human or human-machine) dialogues; as a first step, we propose the following shared task.

3.1 Task Definition

Participating teams are provided with training data, for example, the aforementioned DCH-1 test collection with multiple dialogue-level and nugget annotations per dialogue. Then, in the test phase,

[1] http://workshop.colips.org/dstc6/
[2] See Maddalena *et al.* [8] and Sakai [9] for related discussions in the context of information retrieval evaluation.
[3] http://weibo.com
[4] http://ntcirstc.noahlab.com.hk/STC2/submission_evaluation/EvaluationCriteriaCN.pdf

each team is given a new set of dialogues as input. Let D be the test of dialogues in the test set. Two subtasks are described below. It is hoped that these offline (i.e., laboratory-based) tasks will serve as initial steps towards evaluating real customer-helpdesk dialogue systems.

3.1.1 Dialogue Quality Subtask. First, participating systems are given a list of possible *dialogue quality levels* $\{1, 2, \ldots, L\}$ and the number of annotators a. Then, for each $d \in D$, participating systems are required to return an estimated distribution of annotators over the quality levels. For example, if $L = 5$ (five levels) for Customer Satisfaction (See Section 2.4) and $a = 10$, a participating system might return $(2, 2, 2, 2, 2)$ (i.e., two annotators for each quality level). Note that the gold distribution can also be represented similarly, e.g., $(0, 0, 1, 4, 5)$. Thus, the probability bins (i.e., dialogue quality levels) are *ordered*, just like those in the Dialogue Breakdown Detection Challenge (See Section 2.2).

If a system can thus accurately estimate the dialogue quality (e.g., customer satisfaction, task accomplishment, etc.) *from different people's viewpoints*, that system can potentially serve as a component of a dialogue for self-diagnosis and self-improvement for satisfying diverse customers.

3.1.2 Nugget Detection Subtask. First, participating systems are given a list of Customer nugget types (e.g., {**CNUG0**, **CNUG**, **CNUG**∗, **NaN**}) and a list of Helpdesk nugget types (e.g., {**HNUG**, **HNUG**∗, **NaN**}. For each $d \in D$, participating systems are required to return, for every utterance block in the dialogue, an estimated distribution of the annotators over nugget types. For example, if $a = 10$ and we have the nugget types from DCH-1, a participating system may return, for a particular *Customer* utterance block, an estimated distribution $(3, 4, 3, 0)$, which means "Three annotators said **CNUG0**; four said **CNUG**; three said **CNUG**∗; none said **NaN**." Similarly, for a particular *Helpdesk* utterance block, the same system may return $(4, 4, 2)$, which means "Four annotators said **HNUG**; four said **HNUG**∗; two said **NaN**." Note that the gold distribution for each utterance block can be represented similarly[5], and that the probability bins (i.e., nugget types) are *nominal* (i.e., unordered).

If a system can accurately detect nuggets and their types, that will help researchers utilise nugget-based evaluation measures without having to manually construct nuggets. Nugget-based evaluation measures may provide more fine-grained diagnoses of systems' failures than dialogue-level annotations: for example, if designed appropriately, they may be able to tell us exactly *where* in the dialogue a problem occurred, and *why*.

3.2 Evaluation Measures

3.2.1 Comparing Two Distributions with Existing Measures. Both of the aforementioned subtasks require a comparison of the system's estimated probability distribution over the gold distribution. Figure 1 shows two examples where the estimated distribution is compared with the gold distribution when there are five bins (i.e., dialogue quality levels or nugget types). One might consider *variational distance* [6], which forms the basis of *mean absolute error*

[5] In the DCH-1 collection, nuggets were generally identified as "relevant" parts of within an utterance block. However, treating entire utterance blocks as nuggets may facilitate both the annotation and evaluation steps.

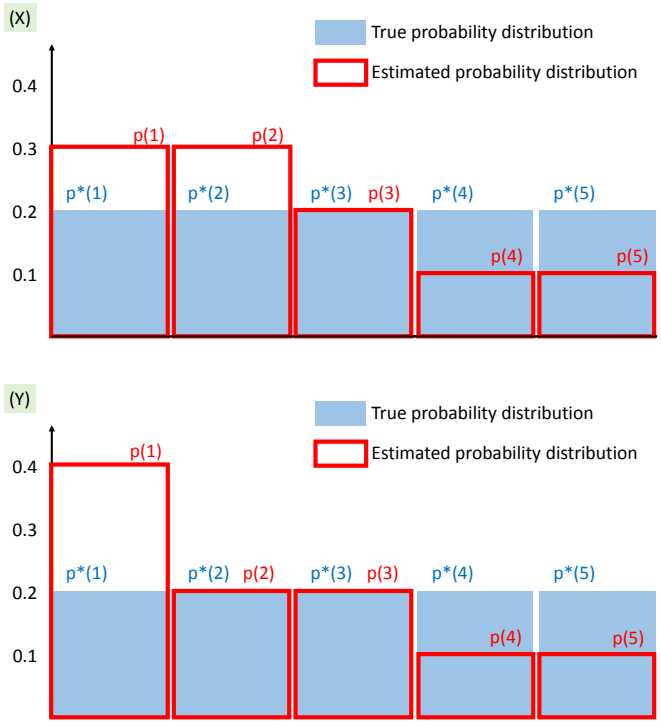

Figure 1: Examples of true and estimated probability distributions.

(MAE) [1], as a candidate measure for comparing the estimated distribution p with the gold distribution p^*:

$$V(p, p^*) = \sum_i |p(i) - p^*(i)| \,, \tag{1}$$

where $p(i), p^*(i)$ are the estimated and true probabilities for the i-th bin. Dividing it by two (representing the case with a complete lack of overlap) would ensure the $[0, 1]$ range. However, accumulating the per-bin errors in this way is not ideal for our purpose, because variational distance cannot penalise "outlier" probabilities. For example, we argue that Figure 1(X) should be rated higher than (Y), because the latter distribution is too skewed compared to the gold distribution; the system is *falsely confident* that Bin 1 has a very high probability. However, the variational distance is clearly 0.4 (0.2 when normalised) for both (X) and (Y): the two systems are treated as equivalent according to this measure. For this reason, we prefer the measures discussed below over variational distance or MAE.

Root mean squared error (RMSE) is often used along with MAE in the research community. This approach is more suitable for our purpose because of its ability to penalise outliers. In our case, we can define a measure based on *Sum of Squares* (SS) first:

$$SS(p, p^*) = \sum_i (p(i) - p^*(i))^2 \,. \tag{2}$$

Since the largest possible value of SS is $1^2 + 1^2 = 2$, we can use *Root Normalised Sum of Squares* (RNSS), which has the [0, 1] range:

$$RNSS(p, p^*) = \sqrt{\frac{SS(p, p^*)}{2}} . \quad (3)$$

For the examples in Figure 1, the RNSS of (X) is 0.1414 while that of (Y) is 0.1732; hence (X) outperforms (Y). The reader is referred to Chai and Draxler [1] for a discussion of the advantages of RMSE (which is similar to RNSS) over MAE.

Another measure that can distinguish the difference between Figure 1(X) and (Y) is the (normalised, symmetric version of) *Jensen-Shannon divergence* (JSD) [6], which we denote as $\mathcal{JSD}(p, p^*)$[6]. First, for probability distributions p_1 and p_2, the *Kullback-Leibler divergence* (KLD), which is not symmetric, is defined as:

$$KLD(p_1, p_2) = \sum_{i \text{ s.t. } p_1(i)>0} p_1(i) \log_2 \frac{p_1(i)}{p_2(i)} . \quad (4)$$

Note that the above is undefined if $p_2(i) = 0$: JSD avoids this limitation as described below.

For a given pair of distributions p and p^*, let p_M be a probability distribution such that, for every bin i, $p_M(i) = (p(i)+p^*(i))/2$. Then, JSD, which is symmetric, is defined as:

$$\mathcal{JSD}(p, p^*) = \frac{KL(p, p_M) + KL(p^*, p_M)}{2} . \quad (5)$$

Thus, by introducing p_M, we can avoid the aforementioned limitation of KLD, since $p1(i) > 0$ implies that $p_M(i) > 0$ also. Moreover, provided that the logarithm base in Eq. 4 is two, the above JSD has the [0, 1] range. Lin [6] proves that the above form of JSD is bounded above by the normalised variational distance (See Eq. 1):

$$\mathcal{JSD}(p, p^*) \leq \frac{V(p, p^*)}{2} . \quad (6)$$

For the examples shown in Figure 1, $\mathcal{JSD}(p, p^*) = (0.0408+0.0372)/2 = 0.0390$ for (X), and $\mathcal{JSD}(p, p^*) = (0.0490 + 0.0490)/2 = 0.0490$ for (X). Again, (X) is considered to be superior.

3.2.2 Comparing Two Distributions with Order-Aware Measures. For the dialogue quality subtask, the probability bins are ordinal, but the aforementioned measures do not take that into account. For example, compare Figure 2(a) with (d), and (b) with (c) (the left half in each figure): where we have $L = 3$ ordinal bins and the true and the estimated distributions are represented in blue and red, respectively. Because RNSS and JSD are summations of differences across the bins, they give the same score to (a) and (d) (RNSS=1, JSD=1), and to (b) with (c) (RNSS=0.8819, JSD=1). However, for ordinal bins, it is clear that (d) is better than (a), and (c) is better than (b). The problem is that there is no notion of *distance* between different bins. Hence we propose a new measure for comparing two distributions where bins are ordinal.

Let A be the set of bins used in the task, where $|A| = L(> 1)$. First, we define sets of bins of nonzero probabilities $B^* = \{i|p^*(i) > 0\}(\subseteq A)$ and $B = \{i|p(i) > 0\}(\subseteq A)$. Then, given estimated and gold distributions p and p^*, we define *Order-aware Divergence* as:

$$OD(p, p^*) = \frac{1}{|B^*|} \sum_{i \in B^*} \sum_{j \in A,\ j \neq i} |i - j|(p(j) - p^*(j))^2 . \quad (7)$$

It can be observed that OD is not symmetric: for every nonzero bin i of p^*, it computes a sum of *weighted* squares for the other bins, where the weight is given as the distance between i and every other bin j. Hence, $B^* = B$ is a sufficient condition that implies the symmetry of OD. We will come back to this point later with a few examples.

Symmetric Order-aware Divergence (SOD) can easily be defined as:

$$SOD = \frac{OD(p, p^*) + OD(p^*, p)}{2} . \quad (8)$$

To ensure that the measure has the [0, 1] range, we should consider the maximum possible value of OD for a given L: it is clear from the definition of OD that in situations such as if $p(1) = 1$ and $p^*(L) = 1$, that is, when both estimated and gold distributions occupy exactly one bin and the two bins are as far apart as possible from each other, the worst-case OD is given by $(L - 1) * 1^2 = L - 1$. Hence, *Normalised Order-aware Divergence* (NOD) and *Symmetric Normalised Order-aware Divergence* (SNOD) may be defined as:

$$NOD(p, p^*) = \frac{OD(p, p^*)}{L - 1} , \quad (9)$$

$$SNOD(p, p^*) = \frac{SOD(p, p^*)}{L - 1} . \quad (10)$$

Note that SNOD is symmetric, but NOD is generally not.

Figure 2, which we have mentioned earlier, contains the NOD and SNOD scores for (a)-(d). The right half of the figures (a)'-(d)', which swaps the estimated and gold distributions, are used for computing SNOD. It can be observed that the SNOD score goes down as we move from (a) to (d). Hence (d) is considered better than (a), and (c) is considered better than (b). In particular, note that while the (S)NOD for (a) is 1, the maximum possible value, that for (d) is 0.5, reflecting the linear weighting scheme of OD.

Figure 3 provides a few other examples with $L = 3$: this time, the gold distribution is uniform. While RS and JS give the same score to (I) and (II) (RNSS=0.5774, JSD=0.4591), and to (III) and (IV) (RNSS=0.3333, JSD=0.2075), it can be observed that the SNOD score goes down as we move from (I) to (IV).

Finally, we compute the SNOD scores for the examples given in Figure 1, where $L = 5$: the results are shown in Figure 4. It can be observed that SNOD prefers (X) to the more skewed (Y). Moreover, note that $B^* = B$ holds for these examples, since both probability distributions cover all the bins. Hence $NOD(p, p^*) = NOD(p^*, p) = SNOD(p, p^*)$ holds[7].

To sum up, we propose to use RNSS, JSD, and SNOD for comparing the probability distributions in the dialogue quality subtask (since the bins are ordered), to use RNSS and JSD for comparing the probability distributions in the nugget detection subtask (since the bins are nominal).

3.2.3 Dialogue Quality Measures. The Dialogue Quality subtask needs to compare, for each dialogue, the system's estimated

[6]The original definition of the Jensen-Shannon divergence assigns a weight to each probability distribution. Our definition of JSD equals the "L divergence" of Lin [6] divided by two.

[7] Another sufficient condition for the symmetry of (N)OD is: $|B^*| = |B| = 1$ *and* $B^* \neq B$. That is, $p^*(i) = 1$ for a particular i and $p(j) = 1$ for a particular $j(\neq i)$. See Figure 2(a) and (d).

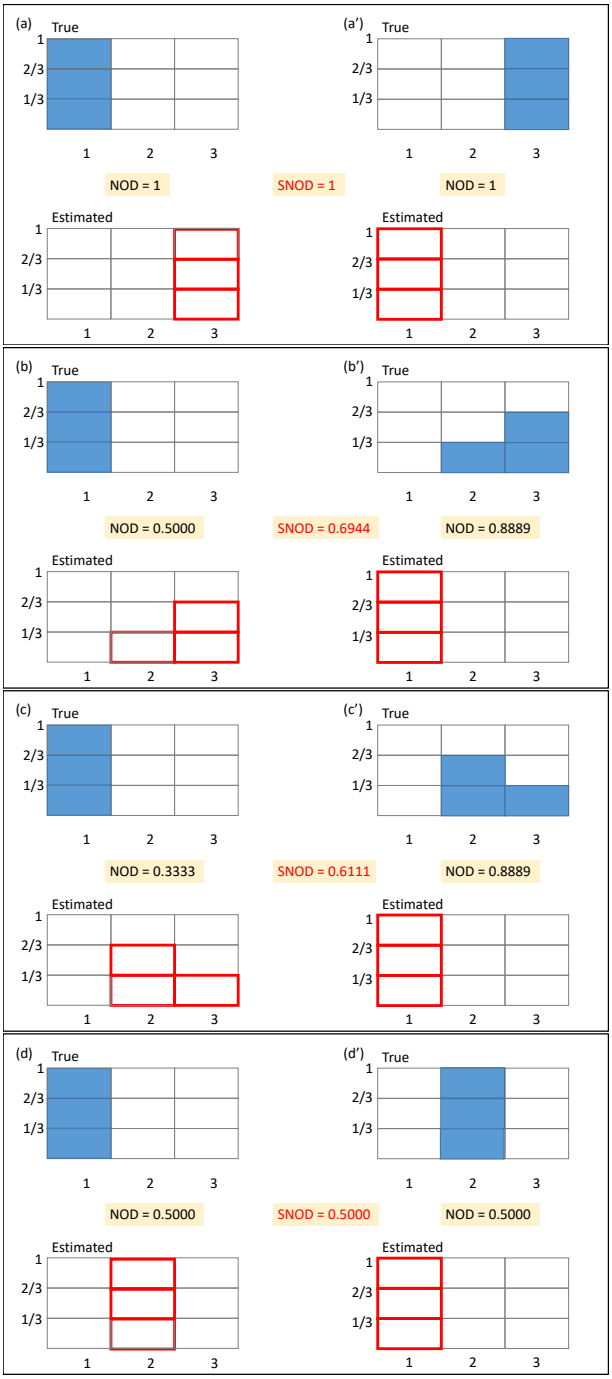

Figure 2: Examples of SNOD scores where $L = 3, p^*(1) = 1$.

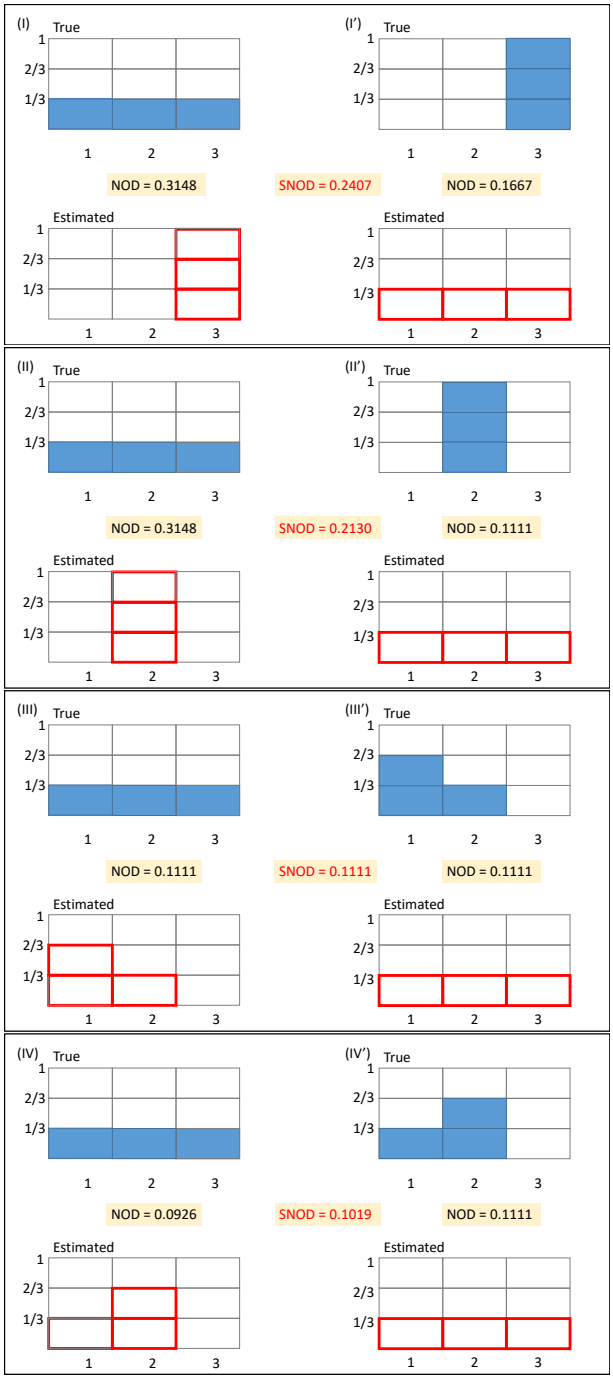

Figure 3: Examples of SNOD scores where $L = 3, p^*(1) = p^*(2) = p^*(3) = 1/3$.

distribution of a annotators over L quality levels with the gold distribution. Let $a(i)$ be the system's estimated number of annotators who chose Level i for a dialogue, and let $a^*(i)$ be the corresponding true number, so that $\sum_{i=1}^{L} a(i) = \sum_{i=1}^{L} a^*(i) = a$. Hence, for each dialogue d, we can construct probability distributions p, p^* by letting $p(i) = a(i)/a, p^*(i) = a^*(i)/a$ for $i = 1, \ldots, L$, and compute $M(d) = M(p, p^*)$, where $M \in \{RNSS, \overline{JSD}, SNOD\}$. Figure 5(a) shows a conceptual diagram of how these measures are computed.

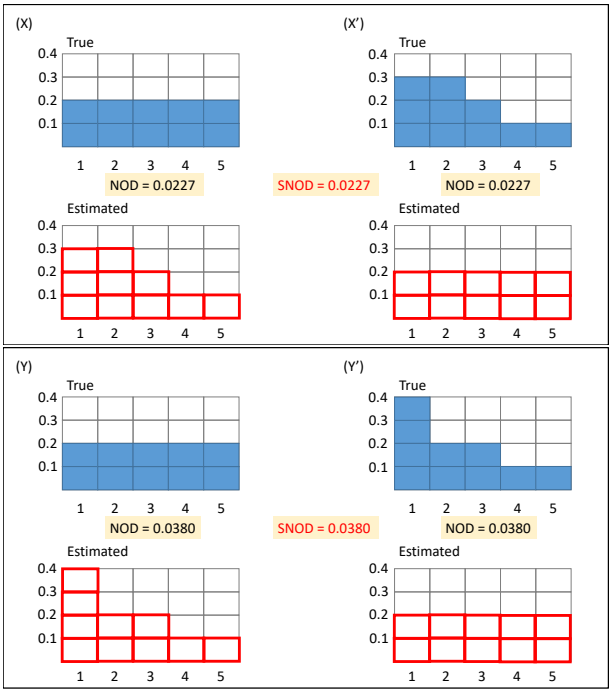

Figure 4: Examples of SNOD scores for $L = 5$, with probability distributions discussed in Figure 1.

The participating systems can then be compared in terms of *mean* RNSS, *mean* JSD, and *mean* SNOD:

$$meanM = \frac{1}{|D|} \sum_{d \in D} M(d) \,, \tag{11}$$

where $M \in \{RNSS, JSD, SNOD\}$.

3.2.4 Nugget Detection Measures. The Dialogue Quality subtask first needs to evaluate, for each utterance block, the accuracy of the system's estimated distribution of annotators over nugget types; then consolidate the results for the entire dialogue[8].

Let T_C be the number of possible Customer nugget types (including **NaN**), and let T_H be the number of possible Helpdesk nugget types (including **NaN**). For example, if the Customer nugget types are **CNUG0**, **CNUG**, **CNUG**$*$, and **NaN**, then $T_C = 4$; if the Helpdesk nugget types are **HNUG**, **HNUG**$*$, and **NaN**, then $T_H = 3$. Let $B_C(d)$ be the set of Customer utterance blocks of a given test dialogue d, and let $B_H(d)$ be the set of Helpdesk utterance blocks for d.

For each Customer block $b_C \in B_C(d)$, let $a(i)$ be the system's estimated number of annotators who chose the i-th Customer nugget type ($1 \le i \le T_C$) for b_C; let $a^*(i)$ be the corresponding true number of annotators. Note that for any block b_C, $\sum_{i=1}^{T_C} a(i) = \sum_{i=1}^{T_C} a^*(i) = a$, since we have a total of a annotators. Hence, for each Customer utterance block b_C, we can construct probability distributions p, p^* by letting $p(i) = a(i)/a, p^*(i) = a^*(i)/a$ for $i = 1, \ldots, T_C$, and compute $M(b_C) = M(p, p^*)$, where $M \in \{RNSS, JSD\}$. Figure 5(b) shows a conceptual diagram of how the measures are computed for a Customer utterance block.

Similarly, for each Helpdesk block $b_H \in B_H(d)$, we can compute $M(b_H)$ where $M \in \{RNSS, JSD\}$.

The entire dialogue d can then be evaluated by *(weighted) average* RNSS and *(weighted) average* JSD:

$$\begin{aligned} waM(d) &= \frac{\alpha}{|B_C(d)|} \sum_{b_C \in B_C(d)} M(b_C) \\ &+ \frac{1-\alpha}{|B_H(d)|} \sum_{b_H \in B_H(d)} M(b_H) \,, \end{aligned} \tag{12}$$

where α ($0 \le \alpha \le 1$) is a parameter for emphasising Customer or Helpdesk results and where $M \in \{RNSS, JSD\}$.

Finally, the participating systems can be compared in terms of *mean (weighted) Average* RNSS and *mean (weighted) Average* JSD:

$$meanwaM = \frac{1}{|D|} \sum_{d \in D} waM(d) \,. \tag{13}$$

4 CONCLUSIONS

This paper proposed a design of a shared task whose ultimate goal is automatic evaluation of multi-turn, dyadic, textual helpdesk dialogues. The proposed task takes the form of an offline evaluation, where participating systems are given a dialogue as input, and output at least one of the following: (1) an estimated distribution of the annotators' quality ratings for that dialogue; and (2) an estimated distribution of the annotators' nugget type labels for each utterance block in that dialogue. This shared task should help researchers build automatic helpdesk dialogue systems that respond appropriately to inquiries by considering the diverse views of customers. The proposed framework has been accepted as part of the NTCIR-14 Short Text Conversation task; we plan to provide the proposed tasks for both Chinese and English dialogues.

We also proposed SNOD, a pilot measure that considers the *order* of the probability bins for the dialogue quality subtask. In our future work, the properties of the measures considered in this paper will be examined with real dialogue data.

ACKNOWLEDGEMENTS

I thank the EVIA reviewers who gave me constructive comments, especially Reviewer 1 who pointed out the limitation of RNSS and JSD for the purpose of comparing two distributions where the categories are ordered. This led me to my proposal of SNOD.

REFERENCES

[1] T. Chai and R.R. Draxler. 2014. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against avoiding RMSE in the Literature. *Geoscientific Model Development* 7 (2014), 1247–1250.
[2] Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. ΔBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets. In *Proceedings of ACL 2015*. 445–450.
[3] Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. 2017. Overview of Dialogue Breakdown Detection Challenge 3. In *Proceedings of Dialog System Technology Challenge 6 (DSTC6) Workshop*.

[8] This is the *macroaveraging* approach, where we assume that each *dialogue* is as important as any other, as it represents a particular customer experience. An alternative would be the *microaveraging* approach, which views each *utterance block* to be as important as any other. The latter implies that longer dialogues impact the overall system performance more heavily, which is not necessarily what we want in the present study.

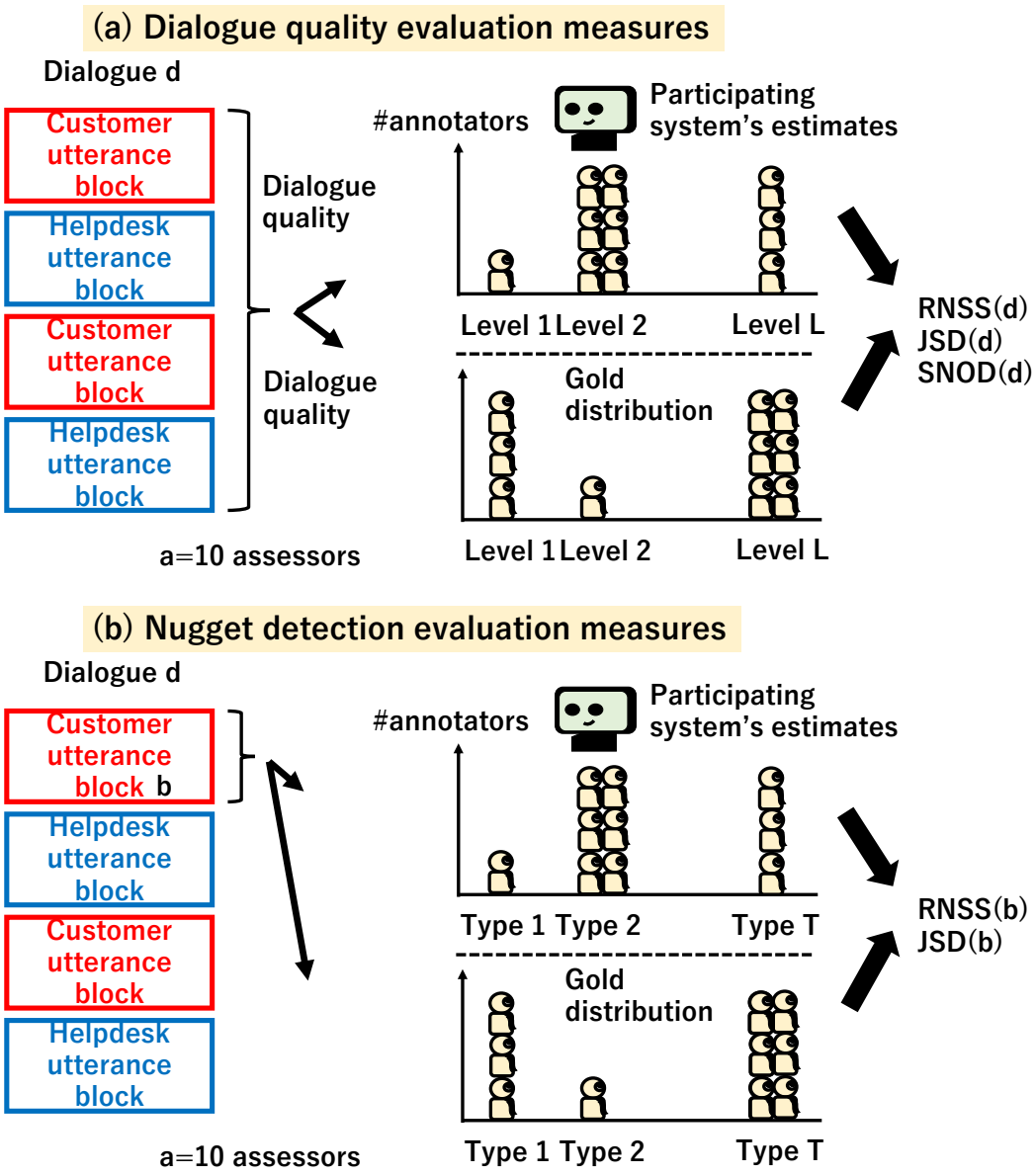

Figure 5: Conceptual diagrams of the proopsed subtasks and the evaluation measures.

[4] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The Dialogue Breakdown Detection Challenge: Task Description, Datasets, and Evaluation Metrics. In *Proceedings of LREC 2016.*

[5] Kate S. Hone and Robert Graham. 2000. Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering* 6, 3-4 (2000), 287–303.

[6] Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151.

[7] Ryan Lowe, Nissan Row, Iulian V. Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of SIGDIAL 2015.* 285–294.

[8] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering Assessor Agreement in IR Evaluation. In *Proceedings of ACM ICTIR 2017.* 75–82.

[9] Tetsuya Sakai. 2017. Unanimity-Aware Gain for Highly Subjective Assessments. In *Proceedings of EVIA 2017.*

[10] Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation. In *Proceedings of ACM SIGIR 2013.* 473–482.

[11] Lifeng Shang, Tetsuya Sakai, Hang Li, Ryuichiro Higashinaka, Yusuke Miyao, Yuki Arase, and Masako Nomoto. 2017. Overview of the NTCIR-13 Short Text Conversation Task. In *Proceedings of NTCIR-13.*

[12] Lifeng Shang, Tetsuya Sakai, Zhengdong Lu, Hang Li, Ryuichiro Higashinaka, and Yusuke Miyao. 2016. Overview of the NTCIR-12 Short Text Conversation Task. In *Proceedings of NTCIR-12.* 473–484.

[13] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings of ACL 1997.* 271–280.

[14] Zhaohao Zeng, Cheng Luo, Lifeng Shang, Hang Li, and Tetsuya Sakai. 2017. Test Collections and Measures for Evaluating Customer-Helpdesk Dialogues. In *Proceedings of EVIA 2017.*