Evaluating Helpdesk Dialogues: Initial Considerations from An Information Access Perspective

Tetsuya Sakai[1,a)] Zhaohao Zeng[1,b)] Cheng Luo[2,1,c)]

Abstract: Whenever a user of a commercial product or a service encounters a problem, an effective and efficient way to solve it is to contact the helpdesk. In the present study, we address the problem of evaluating textual dialogues between the customer and the helpdesk, such as those that take the form of online chats. Thus, we consider textual, dyadic and task-oriented dialogues. Successful dialogues are desirable both for the customer and for the company that sells the product/service. We are currently building a small-scale Chinese helpdesk dialogue test collection (with English translations) and planning to design a suite of language-independent evaluation measures for quantifying the success of each dialogue. Establishing reliable automatic evaluation measures for helpdesk dialogues should ultimately enable researchers and engineers to build and tune intelligent helpdesk systems efficiently. This paper reports on our initial step towards this direction to seek feedback from the Japanese natural language processing research community.

1. Introduction

Whenever a user of a commercial product or a service encounters a problem, an effective and efficient way to solve it is to contact the helpdesk. In the present study, we address the problem of evaluating textual dialogues between the customer and the helpdesk, such as those that take the form of online chats. Successful dialogues are desirable both for the customer and for the company that sells the product/service. We are currently building a small-scale Chinese helpdesk dialogue test collection (with English translations) and planning to design a suite of language-independent evaluation measures for quantifying the success of each dialogue. Establishing reliable automatic evaluation measures for helpdesk dialogues should ultimately enable researchers and engineers to build and tune intelligent helpdesk systems efficiently. This paper reports on our initial step towards this direction to seek feedback from the Japanese natural language processing research community.

While some types of dialogues are just for fun or just "idle talks," helpdesk dialogues are usually *task-oriented* (or *goal-driven*): the customer is facing a specific problem and she wants to solve it. This is similar to situations in which a search engine user has an *information need* which she wants to satisfy. While a search engine user interacts with the search engine to obtain the information she seeks, a customer interacts with the helpdesk by exchanging textual posts. While a concise search engine snippet (or summary) may provide relevant information to the user, a concise and successful textual dialogue between the customer and the helpdesk is beneficial to both parties. Based on this view, we approach the problem of evaluating helpdesk dialogues from an *information access* [16] perspective.

2. Related Work

This section surveys prior art on dialogue evaluation, including those that primarily target *spoken* dialogues. Section 2.1 discusses evaluation efforts that primarily target non-task-oriented dialogues; Section 2.2 discusses those that target task-oriented dialogues. We note that most studies in the latter category handle tasks that involve a relatively rigid pre-defined structure, or *slot filling*, with the exception of the unstructured Ubuntu dialogue corpus discussed in Section 2.2.3. Section 2.3 discusses studies on evaluation methods for textual information access, including query-focussed summsarisation and question answering, some of which directly inspired our work.

2.1 Evaluating Non-Task-Oriented Dialogues

2.1.1 Evaluating Conversational Responses

Evaluating generated responses in non-task-oriented dialogues is a difficult problem. In 2015, Galley *et al.* [4] proposed *Discriminative* BLEU, which generalises BLEU [14], a machine translation evaluation measure that compares the system output with multiple reference translations at the *n*-gram level. Discriminative BLEU introduces positive and negative weights to human references (i.e., gold standard responses) in the computation of *n*-gram-based precision, which is the primary component of BLEU. Because it is difficult to obtain mutiple hand-crafted references for conversational data, they propose to automatically mine candidate responses from a corpora of conversations and then have the annotators rate the quality of the candidates. The reference weights reflect the result of the quality annotations.

[1] Waseda University, Japan
[2] Tsinghua University, P.R.C.
[a)] tetsuyasakai@acm.org
[b)] zhaohao@fuji.waseda.jp
[c)] luochengleo@gmail.com

2.1.2 Dialogue Breakdown Detection Challenge

Also in 2015, Higashinaka *et al.* [6] ran the first *Dialogue Breakdown Detection Challenge* using Japanese human-system chat corpora, to evaluate the system's ability to detect the point in a given dialogue where it becomes difficult to continue due to the system's inappropriate response. This effort used 1,146 text chat dialogues for training and another 100 for development and testing. After each system utterance in the dialogue, participating systems were required to provide a diagnosis: **NB** (not a breakdown), **PB** (possible breakdown), or **B** (breakdown). They were also required to submit a probability distribution over the three labels. To define the gold standard data for this task, multiple assessors were hired, so that a gold probability distribution can be constructed for each utterance. By comparing the best gold label with the system's output, accuracy, presision, recall and F-measure were computed. Morever, by comparing the gold distribution over the three labels with the system's distribution, Jensen-Shannon Divergence and Mean Squared Error were computed. Using a distribution as the gold standard probably reflects the view that there can be multiple acceptable choices within a dialogue, as suggested also by other studies [2], [4].

The second Diaologue Breakdown Detection Challenge is currently underway [*1].

2.1.3 Evaluating the Short Text Conversation Task

At NTCIR-12 (January 2015 - June 2016), the Short Text Conversation (STC) task was run using Weibo[*2] data (for the Chinese subtask) and Twitter data (for the Japanese subtask), and attracted 22 participating teams [23]. The STC task required participating systems to return a valid comment in response to an input tweet (given without any prior context). Instead of relying on natural language generation, systems were required to search a repository of past tweets and return a ranked list as possible responses. Information retrieval evaluation measures, namely, nG@1 (normalised gain at rank 1), nERR (normalised expected reciprocal rank) and P+ were used to evaluate the participating systems. Gold standard labels were created manually by hiring multiple assessors who used the following axes to decide on a single graded label (L0, L1 or L2): *coherence*, *topical relevance*, *context-independence*, and *non-repetitiveness*.

Recently, the second STC task (STC-2) has been launched for NTCIR-13 (June 2016 - December 2017); this time systems are allowed to generate their own responses[*3].

2.2 Evaluating Task-Oriented Dialogues

2.2.1 PARADISE

In 1997, Walker *et al.* [25] proposed the PARADISE (PARAdigm for DIalogue System Evaluation) framework for evaluating task-oriented spoken dialogue systems. The basic idea is to collect a variety of real human-system dialogues for a specific task (e.g., train timetable lookup) as well as subjective measures of *user satisfaction* for each dialogue, and use *task success* and *cost* as explanatory variables so that the user satisfaction measures for new dialogues can be estimated by means of linear regression. PARADISE requires an *attribute-value matrix* that represents the task: for example, for the train timetable domain, attributes such as "depart-city," "arrival-city" and "depart-time" must be specified in advance. This is contrast to our helpdesk case because while it is task-oriented, the required attributes depend on the customer's problem and cannot be listed up exhaustively in advance. In this respect, helpdesk dialogues probably lie somewhere in between non-task-oriented dialogues and the slot-filling dialogues that PARADISE deals with.

The PARADISE framework was subsequently used in the DARPA COMMUNICATOR Programme that evaluated spoken dialogue systems in the travel planning domain [26]. The effort produced the Communicator 2000 Corpus consisting of 662 dialogues based on nine different systems, with per-call survey results on dialogue efficiency, dialogue quality, task success and user satisfaction. Here, a new utterance tagging scheme called DATE (Dialogue Act Tagging for Evaluation) was introduced, which enables three orthogonal annotations along the axes of *speech-act* (e.g., "request-info," "apology"), *task-subtask* (e.g., "origin," "destination," "date") and *conversational-domain* ("about-task," "about-communication," or "situation-frame"). Again, unlike our case, their task-subtask annotation scheme needs to be defined in advance.

2.2.2 Spoken Dialogue Challenge and Dialogue State Tracking Challenge

In 2009, Black and Eskenazi [1] launched the *Spoken Dialogue Challenge* (SDC) by leveraging an existing Pittsburgh bus timetable information system. They proposed live evaluation of the submitted spoken dialogue systems by actually calling them on the phone, thereby conducting both subjective evaluation (e.g., user satisfaction) and objective evaluation (e.g., task completion and number of turns). As possible future work, they mention automatic evaluation by means of a user simulator.

In 2013, the Dialogue State Tracking Challenge (DSTC) was launched by leveraging human-system spoken dialogues (about bus timetable information) collected through the effort at the aforementioned SDC [27]. The entire corpus consisted of over 15,000 dialogues. For each time t in a given dialogue, participating systems (i.e., "trackers") are given a set of N_t possible dialog state hypotheses (plus a meta-hypothesis REST which indicates that none of them are correct), where a hypothesis is an assignment of values to slots in the system. The trackers are expected to output a probability distribution over these states. For example, at the point in a dialogue where the user specifies a bus route she wants to take, the tracker is asked to output a probability distribution over several bus routes. The gold standard data was constructed by consolidating labels obtained through crowdsourcing. Several evaluation measures were used to compare the system's probability distribution with the gold standard label; for example, accuracy measures whether the top ranked hypothesis in the distribution equals the gold standard label. Other measures used were mean probability of the first correct hypothesis, mean reciprocal rank of the first correct hypothesis and variants of receiver-operating characteristic curves.

The fourth DSTC (DSTC4) used 35 *human-human* dialogues, not *human-system* dialogues as in previous years [8]. The dia-

[*1] https://sites.google.com/site/dialoguebreakdowndetection2/
[*2] http://www.weibo.com/
[*3] http://ntcirstc.noahlab.com.hk/STC2/stc-cn.htm

logues discussed tourist information in Singapore. A single session could contain multiple topics, and dialogue tracking was done at each sub-dialogue level. In this challenge, dialogue states were slot-value pairs such as "FROM: Changi Airport," "CUISINE: Singaporean," and "NAME: Little India." While the past DSTCs had teams submit a file with tracker output, DSTC4 had teams provide trackers as a web service. For the main task, the gold-standard states were defined at the sub-dialog level rather than the utterance level, and the evaluation measures used were frame structure-level accuracy, slot-level precision, recall and F-measure. These were computed for every turn ("Schedule 1") and for the last turn of every sub-dialogue ("Schedule 2").

The results of the latest Dialog State Tracking Challenge (DSTC5) will be presented at the 2016 IEEE Workshop on Spoken Language Technology[*4].

2.2.3 Response Selection with the Ubuntu Dialogue Corpus

In 2015, Lowe *et al.* [11] released the Ubuntu Dialogue Corpus, which contains 930,000 *human-human* dialogues extracted from Ubuntu chats. Chats involving more than two parties are automatically disentangled to form dyadic dialogues. Their effort is more similar to ours than the aforementioned studies on task-oriented dialogue evaluation in that they focus primarily on unstructured dialogues rather than slot-filling. They formed a response selection test data set by setting aside 2% of the corpus and forming (*context*, *response*, *flag*) triplets based on this set. Here, context is the sequence of utterances that appear prior to the response in the dialogue; response is either the actual correct response from the dialogue or a randomly chosen utterance from outside the dialogue (but within the test set); *flag* is one for the correct response and zero for incorrect responses. For each correct response they generated nine additional triplets containing different incorrect responses. Thus, response selection systems would be given a context and ten choices of responses, and required to select one or more responses. They use recall at k as the evaluation measure, where k is the size of the set of responses selected by the system and therefore "recall at 1" reduces to accuracy. Note that this evaluation setting does not require annotations for defining the gold standard. They do not consider *ranked* lists of responses as is done at STC (See Section 2.1.3).

2.2.4 Subjective Evaluation of Task-Oriented Dialogues

The most straightforward approach to evaluating dialogues is to collect subjective assessments from the user who actually experienced the dialogue, in the form of a questionnaire. Hone and Graham [7] used a large questionnaire to evaluate an in-car speech interface and identified *system response accuracy*, *likeability*, *cognitive demand*, *annoyance*, *habitability* and *speed* as the key factors in subjective evaluation by means of factor analysis; their approach is known as SASSI (Subjective Assessment of Speech System Interfaces). Hartikainen *et al.* [5] applied a service quality assessment from marketing to the evaluation of telephone-based email application; their method is known as SERVQUAL.

We regard subjective evaluation scores as the gold standard. However, it is not practical to conduct subjective evaluation for every dialogue that we want to evaluate. Subjective evaluation does not necessarily tell us *why* a dialogue has been rated successful, which may be key to improvement; neither does it tell us anything about an unrated dialogues. Hence, we regard subjective evaluation scores as the target variables, and aim to explore effective explanatory variables as well as accurate functions for predicting the subjective scores given a new dialogue. This is similar in spirit to PARADISE, but we would like to evaluate helpdesk dialogues where no slot-filling schemes are predefined. Hence our approach is to leverage some ideas from nugget-based information access evaluation, as discussed below.

2.3 Evaluating Textual Information Access

2.3.1 ROUGE, Nugget Pyramids, POURPRE

While BLEU [14], a measure originally designed for machine translation evaluation, is basically an n-gram-based precision, ROUGE [9], a BLEU-inspired measure designed for text summarisation evaluation, is basically a suite of measures including n-gram-based (or skip-gram-based) recall and F-measure. Just as BLEU requires multiple reference translations, ROUGE requires multiple reference summaries. Note that the basic unit of comparison, namely n-grams etc., are automatically extracted from both the references and the system output.

In contrast to the above automatically extracted units of comparison, manually-devised *nuggets* have been used in both summarisation evaluation [13] and question answering evaluation. In the TREC Question Answering (QA) tracks, a nugget is defined as "*a fact for which the assessor could make a binary decision as to whether a response contained that nugget*" [10]. Having constructed nuggets, (weighted) recall, precision and F-measure scores can be computed, except that the precision computation requires special handling: while we can count the number of nuggets present or missing in the system output, we cannot count the number of "non-nuggets" (i.e., irrelevant pieces of information) in the same output. To work around this problem, a fixed-length "allowance" was introduced at the TREC QA tracks so that nugget precision could be defined based soley on the system output length. The TREC QA tracks also used a measure called POURPRE, which replaces the manual nugget matching step with automatic nugget matching based on unigrams. The NTCIR ACLIA (Advanced Crosslingual Information Access) Task adapted these methods for evaluating QA with Asian languages [12].

2.3.2 S-measure, T-measure and U-measure

As was discussed above, traditional evaluation measures for summarisation and question answering employ variants of recall, precision and F-measure based on small textual units. Hence, they regard the system output as a *set* of n-grams, nuggets, and so on. In contrast, Sakai, Kato and Song [22] introduced a nugget-based evaluation measure called *S-measure* for evaluating textual summaries for mobile search, by incorporating a *decay factor* for nugget weights based on nugget *positions*. Just like information retrieval for ranked retrieval defines a decay function over ranks of documents [16], S-measure defines a linear decay function over the text, using offset positions of the nuggets. This reflects the view that important nuggets should be presented first and that

[*4] `http://workshop.colips.org/dstc5/`

we should minimise the amount of text that the user has to read. Sakai and Kato [21] complements S-measure with a precision-like measure called *T-measure*, which, unlike the aforementioned allowance-based precision used at the TREC QA track, takes into account the fact that different pieces of information require different textual lengths. They define an "iUnit" (information unit) as "*an atomic piece of information that stands alone and is useful to the user*."

Sakai and Dou [20] generalised the idea of S-measure to handle various textual information access tasks, including web search. Their measure, known as *U-measure*, contructs a string called *trailtext*, which is a concatenation of all the texts that the user has read (obtained by observation or by assuming a user model). Then, over the trailtext, a linear decay function is defined. This measure is similar to *Time-Biased Gain* of Smucker and Clarke [24], but uses the trailtext rather than elapsed time to define the decay function, and can handle nonlinear traversals (i.e., violation of the user model which says that the user scans the ranked list from top to bottom) in web search [16].

In the present study, we tentatively propose to regard a dyadic dialogue as a trailtext to define our pilot evaluation measure.

3. An Overview of Our Project

Our project employs a two-phase approach that involves constructing a pilot helpdesk dialogue test collection followed by constructing a larger test collection that reflects the lessons learnt from the pilot collection. Both collections will be based on real human-human dyadic textual dialogues. We plan to take the following steps:

(1) Construct a pilot Chinese test collection that includes *subjective labels* and *nuggets* as well as an English translation for each dialogue.

(2) Design pilot evaluation measures based on nuggets and investigate its usefulness and limitations by examining the correlation between subjective labels and the evaluation measure scores.

(3) Revise the criteria for subjective annotation and nugget annotation, as well as the evaluation measures.

(4) Construct a larger Chinese test collection with subjective labels, nuggets and English translations, and re-investigate the correlation between subjective labels and the revised evaluation measures.

(5) Release the finalised test collection to the research community, with code for computing our finalised measures.

At the time of this writing, we have partially covered steps (1) and (2).

Sections 3.1 and 3.2 provide more details on subjective labels and nuggets, respectively. Section 3.3 discusses the requirements for the evaluation measures that we hope to design. Section 3.4 discusses the next steps once we have completed this project.

3.1 Subjective Labels

We define a *subjective label* as an evaluation score assigned to a given dialogue by a humal annotator who reads that entire dialogue. Possible axes of evaluation may include:

- Whether the task (i.e., the problem to be solved) is clearly stated and whether it is actually accomplished;
- Efficiency of the dialogue in accomplishing the task;
- Whether Customer is likely to have been satisfied with the dialogue, and to what degree.

Subjective labels could also accommodate two different viewpoints:

Customer's viewpoint Does Helpdesk solve Customer's problem efficiently? Customer may want a solution quickly while providing minimal information to Helpdesk.

Helpdesk's viewpoint Does Customer provide accurate and sufficient information so that Helpdesk can provide the right solution? Helpdesk also wants to solve Customer's problem through minimal interactions, as these interactions translate directly into cost for the company.

Depending on situations, one might want to prioritise one of these viewpoints over the other.

Our actual pilot criteria for subjective annotation is discussed in Section 4.2.

3.2 Nuggets

The subjective labels are our target variables. While it might be possible to some extent to design a black-box machine learning algorithm that takes a corpus with subjective labels as training data and a new dialogue as input and predicts a subjective label for that dialogue, we are more interested in *why* a particular dialogue receives a set of certain subjective labels. In other words, we would like to understand the mechanisms of successful and unsuccessful helpdesk dialogues. To this end, we would like to identify some explicit "explanatory variables" for predicting the target variables.

Since we are dealing with textual dialogues without a predefined slot filling scheme, one natural choice for the explicit explanatory variables would be *nuggets* (See Section 2.3), that serve as building blocks of the dialogues. Just as the quality of a summary or an answer output a question answering system is evaluated based on nuggets, *n*-grams, and so on, it may be possible to explain the quality of a dialogue based on nuggets, or possibly automatically extracted surrogates of nuggets. Thus, our fundamental hypothesis is:

Parts-Make-The-Whole Hypothesis *The overall quality of a helpdesk dialogue is governed by the quality of its parts*.

If the above hypothesis turns out to be true, then it may be possible to even replace subjective annotation with nugget annotation.

If we introduce nugget annotation, another interesting hypothesis would be:

Consistency Hypothesis *Nugget annotation achieves higher inter-annotator consistency than subjective annotation.*

We argue that this is a reasonable hypothesis, as nugget annotation involves judgments for smaller units than the entire dialogue and therefore may reduce subjectivity and/or variations in the annotation procedure.

Besides consistency, other possible advantages of nugget annotation over subjective annotation include *sensitivity* and *reusability*. In the form of hypotheses, they can be stated as follows:

Sensitivity Hypothesis *Compared to subjective annotation, nugget annotation enables finer distinctions among different* *dialogues.*

While subjective annotation concerns the entire dialogue and therefore reflects the “all-or-nothing” view (even if the scores are graded), nugget annotation enables the assignment of partial scores to specific parts within the dialogue, which may enable more fine-grained distinctions among different dialogues by means of nugget-based evaluation measures. In other words, while subjective annotation can only give us crude scores, nugget annotation may enable us to design continuous and discriminative measures. Note, on the other hand, that this property may conflict with the consistency hypothesis: if we have a measure with high degrees of freedom, that may possibly translate into low inter-annotator consistency, unless we collapse the nugget-based measures back into some coarse scores.

Reusability Hypothesis *Compared to subjective annotation, nugget annotation enables us to predict the quality of unannotated dialogues more accurately.*

Because subjective labels concern the entire dialogue, and does not tell us why those labels were chosen, they may provide little information for new dialogues. In contrast, nugget annotation may be able to provide information for a new dialogue that discusses an already known task. This is because some of the nuggets in the annotated dialogue may be reusable for that new dialogue. For example, if an annotated dialogue that solves a particular task and a *goal nugget* (e.g., Helpdesk’s utterance that actually solves Customer’s problem) is annotated in that dialogue, then this goal nugget may be useful for evaluating an unannotated dialogue that tries to solve the same task.

Compared to traditional nugget-based information access evaluation that was discussed in Section 2.3, there are two unique features in nugget-based helpdesk dialogue evaluation:

- A dialogue involves two parties, Customer and Helpdesk. Accordingly, there are at least two types of nuggets;
- Even within each type of nuggets, nuggets are not homogeneous, by which we mean that some nuggets may play special roles. In particular, since the dialogues we consider are task-oriented (though lacking in slot filling schemes), there must be some nuggets that represent the state of *identifying* the task and those that represent the state of *accomplishing* it.

We view a nugget as a constituent of a dialogue that helps the Customer’s current state advance towards his target state, namely, the state of having found the solution to the problem (i.e., of task accomplishment). **Fig. 1** depicts this concept.

Our pilot definition of a nugget is discussed in Section 4.3.

3.3 Evaluation Measure Requirements

We would like to design evaluation measures that takes as input a dialogue and a set of nuggets extracted from it and computes a score for it. Ideally, each of these measures should possess at least some of the following properties:

(a) It is highly correlated with one or more of the subjective labels. In other words, it should validate the Parts-Make-The-Whole Hypothesis.
(b) It is easy to compute and to interpret.
(c) It can accommodate both the Customer’s and Helpdesk’s viewpoints and can prioritise one over the other if required.
(d) It can accomodate nugget weights, i.e., the importance of each piece of information.
(e) For a particular task, the measure should prefer a dialogue that contains task accomplishment over one that does not.
(f) Given a pair of dialogues that contain the same set of nuggets for the same task, the measure should prefer the shorter one over the longer one, since the interactions should be minimal.
(g) Given a pair of dialogues that contain task accomplishment for the same task (where the sets of nuggets covered may be different), the measure should prefer the dialogue that reaches task accomplishment more quickly than the other.

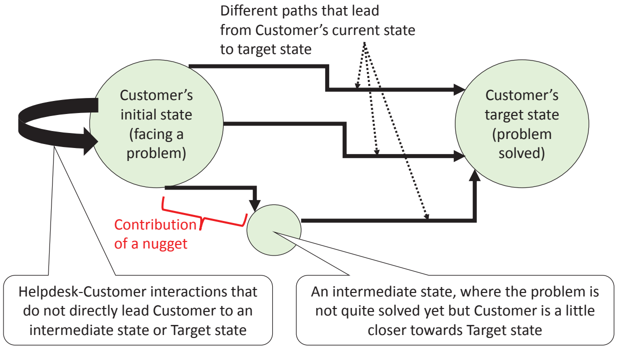

Different paths that lead from Customer’s current state to target state
Customer’s initial state (facing a problem)
Customer’s target state (problem solved)
Contribution of a nugget
Helpdesk-Customer interactions that do not directly lead Customer to an intermediate state or Target state
An intermediate state, where the problem is not quite solved yet but Customer is a little closer towards Target state

Fig. 1 Task accomplishment as state transitions, and the role of a nugget.

3.4 Future Prospects

Having completed the steps discussed at the beginning of Section 3, the test collections with subjective and nugget annotations and the evaluation measures may be utilised by the research community for various purposes. One problem we would like to tackle is to construct and evaluate an automatic helpdesk system that utilises the test collection as an unstructured FAQ knowledge base to answer human inquiries. For example, we could randomly pick a successful dialogue from the test collection and give it to a subject. After reading the human-human dialogue, the subject initiates a dialogue with the auto-Helpdesk, perhaps starting with an expression that is different from the one in the human-human dialogue. The new human-system dialogue may take a path that is different from the one in the original dialogue. Will the system still manage to provide the goal nuggets, while avoiding breakdowns (See Section 2.1.2)? How will the human-system dialogue compare to the original human-human dialogue in terms of the nugget-based evaluation measures?

4. A Pilot Test Collection

This section describes our current status in developing the pilot test collection.

4.1 Dialogue Mining

Weibo, one of the most influential Social Network websites in China, has been utilised extensively by both consumers and companies. Many companies have one or more official Weibo accounts to promote their products and services, and to maintain connections with the customers. When a customer has a problem with a product or a service, posting a message that mentions the company's official account can often solve it. The customer can not only obtain some suggestions from their social network friends but also receive responses from customer service staff that operates the official account.

Based on the above observation, we crawled Weibo in April 2016 to contruct our pilot customer-helpdesk dialogue data as follows.

Step 1 We collected an initial set of Weibo accounts which we denote by A_0, by searching Weibo account names that contained keywords such as "assistant" and "helper" (in Chinese).

Step 2 For each account in A_0, we first crawled the 200 most recent posts[*5] that contain a mention of the official account using "@." We then filtered out accounts that did not respond to more than half of these posts, to obtain "active" official accounts[*6]. We denote the filtered set of active accounts as A.

Step 3 For each account in A, we crawled the 2,000 most recent posts that contains a mention of the official acccount, and extracted customer-helpdesk dyadic dialogues from them. We then kept those that consist of at least five posts by Customer *and* at least five posts by Helpdesk. Thus 234 dialogues were obtained[*7].

4.2 Subjective Annotation

We discussed the general ideas behind subjective annotation in Section 3.1. Here, we describe the steps we actually took for subjective annotation in the pilot test collection construction phrase. Since our pilot test collection comprises Chinese weibo data, we hired four Chinese master students (male master students in Computer Science) and assigned three students to each dialogue at random for the subjective annotation.

4.2.1 Subjective Annotation Criteria

Fig. 2 shows the annotation criteria that we provided to the annotators. In addition, Fig. 1 (with the mention of the nugget removed) was shown to them to help them understand the concept of state transitions that we described in Section 3.2.

Each dialogue was annotated individually by three different assessors, using a web browser-based tool that we designed for this purpose. As there are five questions for subjective annotation, each dialogue is associated with 15 labels in total. This was done for all 234 dialogues that we described in Section 4.1, resulting in a total of 3,510 labels, or 702 labels for each of the five subjective annotation criteria.

Table 1 summarises the result of subjective annotation. *Fleiss' kappa* [3] values with 95% confidence invervals (CIs) [16] are shown to indicate inter-annotator agreements where the numerical labels are treated as *nominal* categories. However, since the distribution over the categories is extremely biased for some of the subjective questions, we also report Randolph's κ_{free} [15], which is more suitable in such situations [16]. For Q3, where the raw labels take five values (−2 to 2), they were collapsed into three categories −1, 0 and 1 as indicated in the caption of Table 1.

Table 1 Distributions and inter-annotator agreements of subjective labels. Three annotators were assigned to each of the 234 dialogues and therefore there are 702 labels per question. Note that Fleiss' κ and Randolph's κ_{free} treat the labels as *nominal* categories. For Q3, 2 and 1 were collapsed into 1, and −2 and −1 were collapsed into −1.

	#labels/category	Fleiss' κ [95% CI]	κ_{free}
Q1 (2/1/0)	692/9/1	−0.013 [−0.082, 0.056]	0.957
Q2 (2/10)	154/219/329	0.417 [0.363, 0.470]	0.444
Q3 (1/0/-1)	191/284/227	0.480 [0.428, 0.533]	0.487
Q4 (2/1/0)	420/142/140	0.041 [−0.014, 0.095]	0.192
Q5 (2/1/0)	631/49/22	0.243 [0.183, 0.303]	0.788

It can be observed from Table 1 that Q1 received 692 "Yes" labels, 9 "Partially" label and only "1" No label. Due to this extreme bias, Fleiss' kappa says that the inter-annotator agreement is statistically nonexistent. However, note that Randolph's κ_{free}, which does not depend on this biased distribution, correctly reflects the fact that the assessors agree most of the time. It can be observed that while the κ_{free} for Q5 (Customer Utterance Quality) is moderately high (0.788), that for Q4 (Helpdesk utterance quality) is low (0.192). Note also that the distribution over categories for Q5 is more heavily biased towards "Yes" (631 labels out of 702). Thus, while Q4 and Q5 are semantically symmetric, it is possible that Q4 is a more difficult or ill-defined question for annotators than Q5. Will nugget annotation be able to remedy this kind of problem, showing that our Consistency Hypothesis (Section 3.2) holds true?

4.3 Nugget Annotation

We hired the same four Chinese students for nugget annotation, and the work is currently in progress. For the construction of a large-scale test collection, we plan to hire different assessors across the two annotation tasks for the majority of the dialogues, while hiring the same set of assessors across the two annotation tasks for the rest, so that we can quantify the effect of using different annotators on the correlation between subjective labels and nugget-based evaluation scores. According to our preliminary results, many of the dialogues in our pilot data are *unsuccessful* dialogues, and therefore it is difficult to identify nuggets that contribute to the state transitions as depicted in Figure 1. Hence, to build a larger test collection with a sufficient number of successful dialogues with nuggets, we will probably have to obtain an even larger pool of dialogues first. We would like our finalised set of dialogues with nuggets to offer sufficiently high *statistical power* [19].

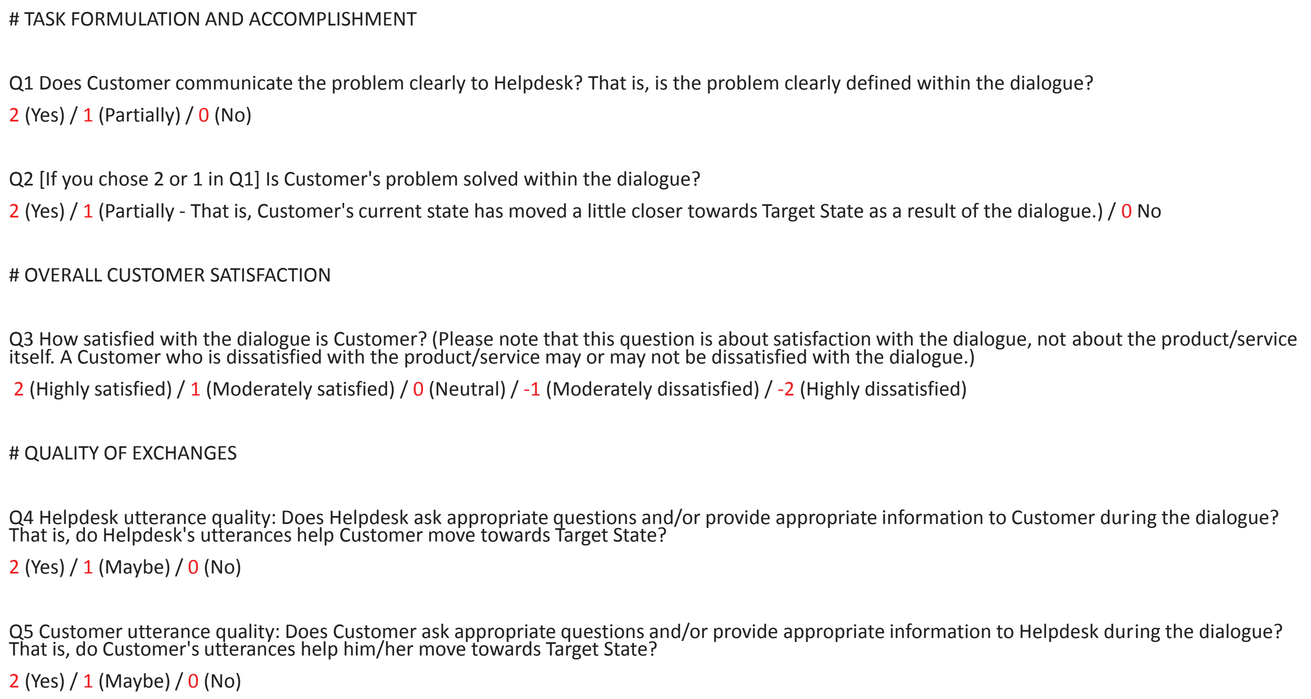
First, we define a *post* as a piece of text input by either Customer or Helpdesk, who presses ENTER to upload it on Weibo. Hence a post can be a sentence, a part of a sentence or even multiple sentences. We then define a nugget as follows.

(I) It is a post, or a sequence of consecutive posts by the same

[*5] See 4.3 for our definition of a post.

[*6] Many official Weibo accounts are just for promotion, and do not respond to customer inquiries.

[*7] Let C denote a post by Customer, and H denote a post by Helpdesk. Then, according to our current filtering method, a dialogue of the form $CCCCCHHHHH$ qualifies in theory, even though this is actually a *single-turn* dialogue. In future, when filtering dialogues by length, we will first merge consective posts by the same interlocutors to form an *utterance block*, and then count the number of utterance blocks rather than utterances. In this way, $CCCCCHHHHH$ will be converted into $C'H'$, where C' and H' denote utterance blocks, and will not qualify as a multi-turn dialogue.



# TASK FORMULATION AND ACCOMPLISHMENT

Q1 Does Customer communicate the problem clearly to Helpdesk? That is, is the problem clearly defined within the dialogue?
2 (Yes) / 1 (Partially) / 0 (No)

Q2 [If you chose 2 or 1 in Q1] Is Customer's problem solved within the dialogue?
2 (Yes) / 1 (Partially - That is, Customer's current state has moved a little closer towards Target State as a result of the dialogue.) / 0 No

# OVERALL CUSTOMER SATISFACTION

Q3 How satisfied with the dialogue is Customer? (Please note that this question is about satisfaction with the dialogue, not about the product/service itself. A Customer who is dissatisfied with the product/service may or may not be dissatisfied with the dialogue.)
2 (Highly satisfied) / 1 (Moderately satisfied) / 0 (Neutral) / -1 (Moderately dissatisfied) / -2 (Highly dissatisfied)

# QUALITY OF EXCHANGES

Q4 Helpdesk utterance quality: Does Helpdesk ask appropriate questions and/or provide appropriate information to Customer during the dialogue? That is, do Helpdesk's utterances help Customer move towards Target State?
2 (Yes) / 1 (Maybe) / 0 (No)

Q5 Customer utterance quality: Does Customer ask appropriate questions and/or provide appropriate information to Helpdesk during the dialogue? That is, do Customer's utterances help him/her move towards Target State?
2 (Yes) / 1 (Maybe) / 0 (No)

Fig. 2 Pilot subjective annotation criteria.

interlocutor (i.e., either Customer or Helpdesk).

(II) It can neither partially nor wholly overlap with another nugget.

(III) It should be minimal: that is, it should not contain irrelevant posts at the start, the end or in the middle. An irrelevant post is one that does not contribute to the Customer transition mentioned in 5.

(IV) It helps Customer transition from Current State (including Initial State) towards Target State (i.e., when the problem is solved).

Moreover, we tentatively define the following four mutually exclusive nugget *types*:

CNUG0 Customer's "trigger" nuggets. These are nuggets that define Customer's initial problem, which directly caused Customer to contact Helpdesk.

CNUG Customer's "regular" nuggets. These are nuggets in Customer's utterances that are useful from Helpdesk's point of view.

HNUG Helpdesk's "regular" nuggets. These are nuggets in Helpdesk's utterances that are useful from Customer's point of view.

CNUG$*$ Customer's "goal" nuggets. These are nuggets in Customer's utterances which tell Helpdesk that Customer's problem has been solved.

HNUG$*$ Helpdesk's "goal" nuggets. These are nuggets in Helpdesk's utterances which provide the Customer with a solution to the problem.

Each nugget type may or may not be present in a dialogue. Multiple nuggets of the same type may be present in a dialogue.

The above definitions, together with Fig. 1, were given to the annotators. The annotators used the aforementioned web-based tool to identify nuggets, but this task was done independently from the subjective annotation task. Two annotators were randomly assigned to each dialogue for nugget annotation.

The top half of **Fig. 3** shows an example of a dialogue (translated into English) with nuggets annotated by the first author based on the English translation. Rectangles with dotted lines represent regular nuggets; those with solid lines represent goal nuggets. Lines that start with a "#" are comments.

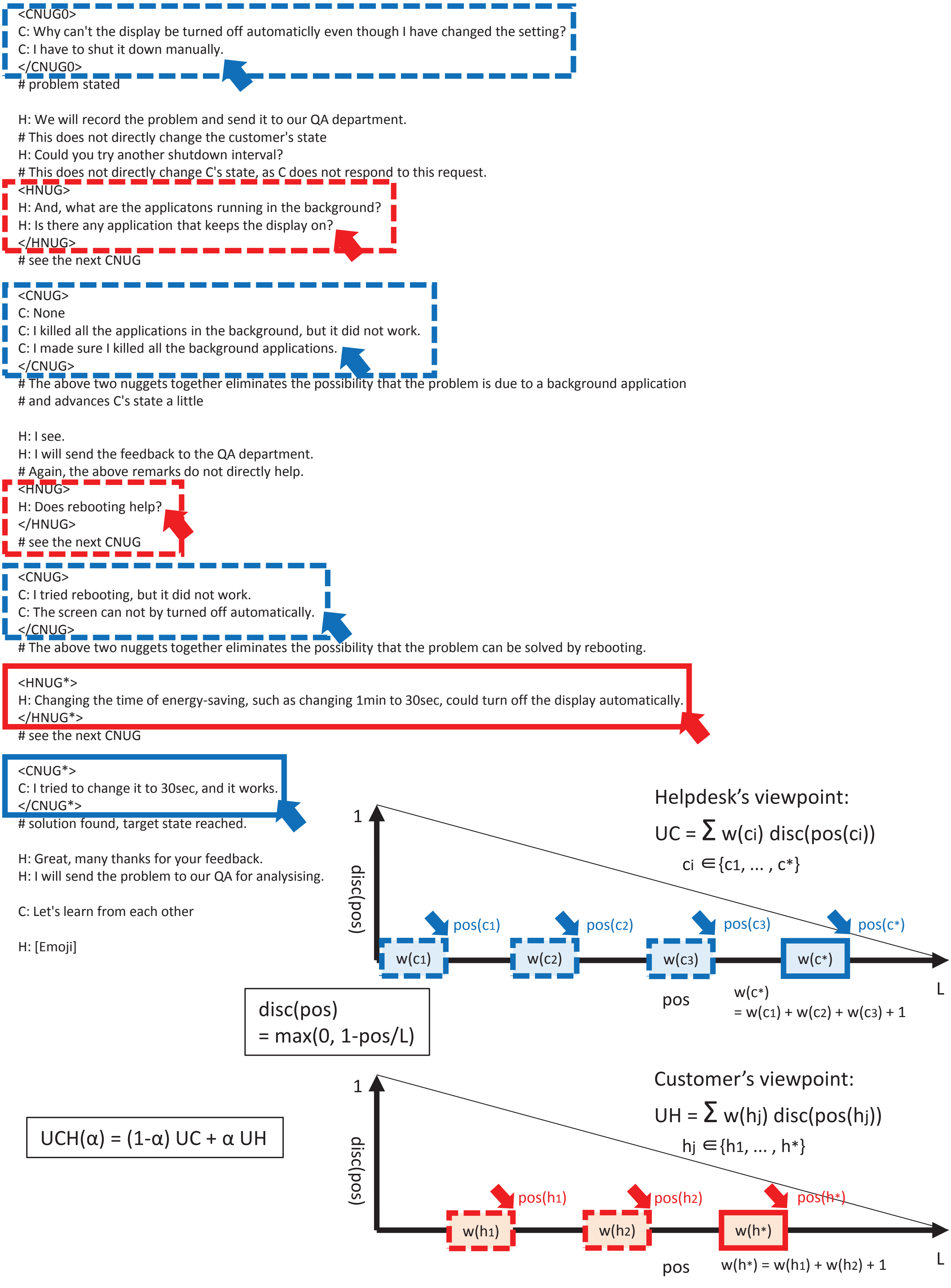
5. Pilot Evaluation Measures

In this section, we describe our pilot evaluation measure that hopefully satisfies at least some of the requirements discussed in Section 3.3. They are based on U-measure [20] and its predecessor S-measure [22].

5.1 UCH

We regard a dialogue as a piece of text which may or may not contain nuggets. We also define a *maximum tolerable dialogue length* L: this may be obtained, for example, as the maximum dialogue length within a corpus. Alternatively, if we have a requirement that each dialogue has to terminate within T minutes, and if we have an estimate of the average typing speed s (characters per minute) in an online chat, we may obtain L as $T \times s$. Then we can define a decay function over the piece of text, which starts as 100% at the beginning of the text, and reaches 0% when the text length reaches L. The latter represents the point where every piece of information becomes worthless, and the decaying curve implies that a nugget is rewarded more heavily if it appears early in the text.

The bottom half of Fig. 3 presents the above concept using the aforementioned dialogue annotated with nuggets. Let N and M denote the number of Customer's non-goal nuggets and Helpdesk's non-goal nuggets identified within a dialogue, re-



```
<CNUG0>
C: Why can't the display be turned off automaticlly even though I have changed the setting?
C: I have to shut it down manually.
</CNUG0>
# problem stated

H: We will record the problem and send it to our QA department.
# This does not directly change the customer's state
H: Could you try another shutdown interval?
# This does not directly change C's state, as C does not respond to this request.
<HNUG>
H: And, what are the applicatons running in the background?
H: Is there any application that keeps the display on?
</HNUG>
# see the next CNUG

<CNUG>
C: None
C: I killed all the applications in the background, but it did not work.
C: I made sure I killed all the background applications.
</CNUG>
# The above two nuggets together eliminates the possibility that the problem is due to a background application
# and advances C's state a little

H: I see.
H: I will send the feedback to the QA department.
# Again, the above remarks do not directly help.
<HNUG>
H: Does rebooting help?
</HNUG>
# see the next CNUG

<CNUG>
C: I tried rebooting, but it did not work.
C: The screen can not by turned off automatically.
</CNUG>
# The above two nuggets together eliminates the possibility that the problem can be solved by rebooting.

<HNUG*>
H: Changing the time of energy-saving, such as changing 1min to 30sec, could turn off the display automatically.
</HNUG*>
# see the next CNUG

<CNUG*>
C: I tried to change it to 30sec, and it works.
</CNUG*>
# solution found, target state reached.

H: Great, many thanks for your feedback.
H: I will send the problem to our QA for analysising.

C: Let's learn from each other

H: [Emoji]
```

Fig. 3 Computing UCH: an example

spectively; for simplicity, let us assume that there is at most one Customer's goal nugget and at most one Helpdesk's goald nugget in a dialogue. Let $\{c_1, \ldots, c_N, c_*\}$ denote the set of nuggets from Customer's posts, and let $\{h_1, \ldots, h_M, h_*\}$ denote that from Helpdesk's posts. Let $pos(c_i)$ ($i \in \{1, \ldots, N, *\}$) be the offset position of nugget c_i within the dialogue; for languages such as Chinese and Japanese, we use the number of characters to define the offset. $pos(h_j)$ ($j \in \{1, \ldots, M, *\}$) is defined similarly. Given a nugget position (pos), the discount value can be given by:

$$disc(pos) = \max(0, 1 - \frac{pos}{L}) \ . \tag{1}$$

Note that the above functions satisfies the property discussed above: having nuggets earlier in the text is rewarded.

As shown in Fig. 3, given the weight of each regular nugget ($w(c_i), w(h_j)$), a simple evaluation measure that reflects Helpdesk's viewpoint can be computed based on the nuggets from Customer's utterances as:

$$UC = \sum_{c_i \in \{c_1, \ldots, c_N, c_*\}} w(c_i) \ disc(pos(c_i)) \ . \tag{2}$$

Here, let us define the weight of CNUG∗ as:

$$w(c_*) = 1 + \sum_{i=1}^{N} w(c_i) \ . \tag{3}$$

This is an attempt at reflecting Property (d) discussed in Section 3.3: task accomplishment is what matters most. To be more specific, when the discounting function is ignored and dialogues are regarded as sets of nuggets, then having only the goal nugget is better than having all the regular nuggets.

Similarly, a measure that reflects Customer's viewpoint can be computed based on the nuggets from Helpdesk's utterances as:

$$UH = \sum_{h_j \in \{h_1, \ldots, h_M, h_*\}} w(h_j) \ disc(pos(h_j)) \ , \tag{4}$$

where

$$w(c_*) = 1 + \sum_{j=1}^{M} w(h_j) \ . \tag{5}$$

Finally, for a given parameter α ($0 \leq \alpha \leq 1$) that specifies the importance of the Customer's viewpoint relative to Helpdesk's, we can define the following combined measure:

$$UCH(\alpha) = (1 - \alpha)UC + \alpha UH \ . \tag{6}$$

Note that $UCH(0.5)$ is equivalent to simply putting the two graphs in Fig. 3 on top of each other and then computing a single U-measure score.

Like U-measure, UCH is an unnormalised measure: it does not have a score range such as $[0, 1]$. However, it is known that if *score standardisation* is applied to evaluation measure scores, then normalisation becomes unnecessary [17], [18].

In summary, UCH means "U-measure computed based on Customer's and Helpdesk's nuggets," where the trailtext (i.e., all the text that a search engine user has read) is replaced by the dialogue text (i.e., all the textual exchanges between Customer and Helpdesk). UH means "U-measure computed based on Helpdesk's nuggets, from Customer's viewpoint." UC means "U-measure computed based on Customer's nuggets, from Helpdesk's viewpoint." It is hoped that the latter two may provide additional insights into dialogue evaluation. At the time of this writing, we are preparing to investigate the correlation between the subjective labels and UCH.

Of course, variants of UCH are possible: for example, while UH and UC employs a common discount function, the two may utilise different functions, reflecting different levels of patience; another possibility would be to make the gradient of the discount function vary according to whether the current interlocutor is Customer or Helpdesk. On the other hand, we should bear in mind the importance of Property (b) discussed in Section 3.3: evaluation measures should be easy to compute and to interpret.

Since each post has a timestamp, another possibility would be to define a decay function over time rather than the dialogue text, as in Time-Biased Gain [24]. For example, given a parameter T which represents the maximum dialogue duration allowed, the value of a nugget at time t could be discounted using the following decay function:

$$disc(t) = \max(0, 1 - \frac{t}{T}) \ . \tag{7}$$

Time-based measures may be able to reflect the time gap between posts, while they may not be able to accurately reflect the informativeness of each post, or the differences across languages. These are also questions to be addressed in our future work.

6. Conclusions and Future Work

We have outlined our ongoing effort in building a Chinese helpdesk dialogue test collection and designing evaluation measures for automatically quantifying the effectiveness of a given dialogue from the viewpoints of both parties involved, i.e., Customer and Helpdesk. We plan to eventually make the finalised test collections with English translations publicly available to the research community.

Our future plan is as follows:

(1) Complete the steps described in Section 3, while examining the Parts-Make-The-Whole, Consistency, Sensitivity, and Reusability Hypotheses for nugget annotation described in Section 3.2;

(2) As the next project, evaluate human-system helpdesk dialogues by employing our test collection and the approach described in Section 3.4. One possible approach to implement this would be to run a task at collaborative venues such as NTCIR.

We hope that these efforts will make some contributions to the research in task-oriented dialogue systems.

Acknowledgments We thank Hang Li and Lifeng Shang (Huawei Noah's Ark Lab) for helpful discussions and continued support; and Guan Jun, Lingtao Li and Yimeng Fan (Waseda University) for helping us construct the pilot test collection.

We also thank Ryuichiro Higashinaka (NTT Media Intelligence Laboratories) for providing us with valuable information related to the evaluation of non-task-oriented dialogues.

References

[1] Black, A. W. and Eskenazi, M.: The Spoken Dialogue Challenge, *Proceedings of SIGDIAL 2009*, pp. 337–340 (2009).

[2] DeVault, D., Leuski, A. and Sagae, K.: Toward Learning and Evaluation of Dialogue Policies with Text Examples, *Proceedings of SIGDIAL 2011*, pp. 39–48 (2011).

[3] Fleiss, J. L.: Measuring Nominal Scale Agreement among Many Raters, *Psychological Bulletin*, Vol. 76, No. 5, pp. 378–382 (1971).

[4] Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J. and Dolan, B.: ΔBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets, *Proceedings of ACL 2015*, pp. 445–450 (2015).

[5] Hartikainen, M., Salonen, E.-P. and Turunen, M.: Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method, *Proceedings of INTERSPEECH 2004-ICSLP* (2004).

[6] Higashinaka, R., Funakoshi, K., Kobayashi, Y. and Inaba, M.: The Dialogue Breakdown Detection Challenge: Task Description, Datasets, and Evaluation Metrics, *Proceedings of LREC 2016* (2016).

[7] Hone, K. S. and Graham, R.: Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI), *Natural Language Engineering*, Vol. 6, No. 3-4, pp. 287–303 (2000).

[8] Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D. and Henderson, M.: The Fourth Dialog State Tracking Challenge, *Proceedings of IWSDS 2016* (2016).

[9] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 74–81 (2004).

[10] Lin, J. and Demner-Fushman, D.: Will Pyramids Built of Nuggets Topple Over?, *Proceedings of HLT/NAACL 2006*, pp. 383–390 (2006).

[11] Lowe, R., Row, N., Serban, I. V. and Pineau, J.: The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems, *Proceedings of SIGDIAL 2015*, pp. 285–294 (2015).

[12] Mitamura, T., Shima, H., Sakai, T., Kando, N., Mori, T., Takeda, K., Lin, C.-Y., Song, R., Lin, C.-J. and Lee, C.-W.: Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access, *Proceedings of NTCIR-8*, pp. 15–24 (2010).

[13] Nenkova, A., Passonneau, R. and McKeown, K.: The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation, *ACM Transactions on Speech and Language Processing*, Vol. 4, No. 2 (2007).

[14] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation, *Proceedings of ACL 2002*, pp. 311–318 (2002).

[15] Randolph, J. J.: Free-marginal Multirater Kappa (Multirater κ_{free}): An Alternative to Fleiss' Fixed Marginal Multirater Kappa, *Joensuu Learning and Instruction Symposium 2005* (2005).

[16] Sakai, T.: *Information Access Evaluation Methodology: For the Progress of Search Engines (in Japanese)*, Coronasha (2015).

[17] Sakai, T.: The Effect of Score Standardisation on Topic Set Size Design, *Proceedings of AIRS 2016* (2016).

[18] Sakai, T.: A Simple and Effective Approach to Score Standardisation, *Proceedings of ACM ICTIR 2016* (2016).

[19] Sakai, T.: Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015, *Proceedings of ACM SIGIR 2016* (2016).

[20] Sakai, T. and Dou, Z.: Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation, *Proceedings of ACM SIGIR 2013*, pp. 473–482 (2013).

[21] Sakai, T. and Kato, M. P.: One Click One Revisited: Enhancing Evaluation based on Information Units, *Proceedings of AIRS 2012* (2012).

[22] Sakai, T., Kato, M. P. and Song, Y.-I.: Click the Search Button and Be Happy: Evaluating Direct and Immediate Information Access, *Proceedings of ACM CIKM 2011*, pp. 621–630 (2011).

[23] Shang, L., Sakai, T., Lu, Z., Li, H., Higashinaka, R. and Miyao, Y.: Overview of the NTCIR-12 Short Text Conversation Task, *Proceedings of NTCIR-12*, pp. 473–484 (2016).

[24] Smucker, M. D. and Clarke, C. L. A.: Time-Based Calibration of Effectiveness Measures, *Proceedings of ACM SIGIR 2012*, pp. 95–104 (2012).

[25] Walker, M. A., Litman, D. J., Kamm, C. A. and Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents, *Proceedings of ACL 1997*, pp. 271–280 (1997).

[26] Walker, M. A., Passoneau, R. and Boland, J. E.: Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems, *Proceedings of ACL 2001*, pp. 515–522 (2001).

[27] Williams, J., Raux, A., Ramachandran, D. and Black, A.: The Dialog State Tracking Challenge, pp. 404–413 (2013).