Test Collections and Measures for Evaluating Customer-Helpdesk Dialogues

Anonymous Author(s)

ABSTRACT

We address the problem of evaluating textual, task-oriented dialogues between the customer and the helpdesk, such as those that take the form of online chats. As an initial step towards evaluating automatic helpdesk dialogue systems, we have constructed a test collection comprising 3,700 real Customer-Helpdesk multi-turn dialogues by mining Weibo, a major Chinese social media. We have annotated each dialogue with multiple subjective quality annotations and nugget annotations, where a nugget is a minimal sequence of posts by the same utterer that helps towards problem solving. In addition, 10% of the dialogues have been manually translated into English. Our test collection, DCH-1, will be made publicly available for research purposes. We also propose a simple nugget-based evaluation measure for task-oriented dialogue evaluation, which we call UCH, and explore its usefulness and limitations.

KEYWORDS

dialogues; evaluation; helpdesk; measures; nuggets; test collections

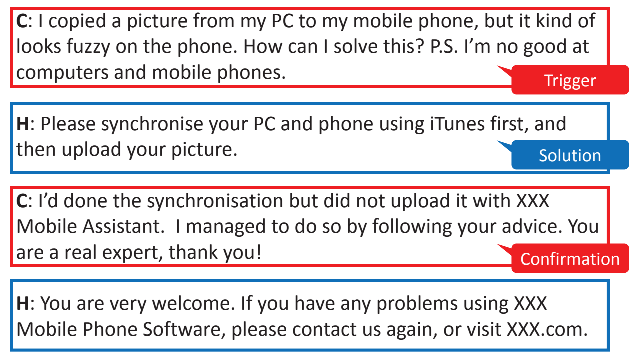

Figure 1: An example of a dialogue between Customer (C) and Helpdesk (H).

1 INTRODUCTION

Whenever a user of a commercial product or a service encounters a problem, an effective way to solve it would be to contact the helpdesk. Efficient and successful dialogues are desirable both for the customer and the company that sells the product/service. Recent advances in artificial intelligence suggest that, in the not-too-distant future, these *human-human* Customer-Helpdesk dialogues will be replaced by *human-machine* ones. In order to build and efficiently tune *automatic helpdesk systems*, reliable automatic evaluation methods for task-oriented dialogues are required.

Figure 1 shows an example of a Customer-Helpdesk dialogue. It can be observed that it is initiated by Customer's report of a particular problem she is facing, which we call a *trigger*. This is an example of a successful dialogue, for Helpdesk provides an actual *solution* to the problem and Customer acknowledges that the problem has been solved. Unlike the classical *closed-domain* task-oriented dialogues, Helpdesk may have to handle diverse requests, which makes it impossible for us to solve the problems by pre-defined *slot filling* schemes that are required by many existing evaluation measures for task-oriented dialogues (See Section 2.2).

In the present study, we address the problem of evaluating textual Customer-Helpdesk dialogues, such as those that take the form of online chats. As an initial step towards evaluating automatic helpdesk dialogue systems, we have constructed a test collection comprising 3,700 real customer-helpdesk multi-turn dialogues by mining Weibo[1], a major Chinese social media. We have annotated each dialogue with *subjective quality annotations* (*task statement*, *task accomplishment*, *customer satisfaction*, *helpdesk appropriateness*, *customer appropriateness*) as well as *nugget annotations*, where a nugget is a minimal sequence of posts by the same utterer that helps towards problem solving. In addition, 10% of the dialogues have been manually translated into English. Our test collection, *DCH-1* (Dialogues between Customer and Helpdesk) will be made publicly available for research purposes, along with a smaller pilot collection *DCH-0*, which contains 234 dialogues.

We also propose a simple nugget-based evaluation measure for task-oriented dialogue evaluation, which we call UCH (Utility for Customer and Helpdesk), and explore its usefulness and limitations. We believe that, while subjective dialogue evaluation can evaluate the dialogue as a whole, automatic evaluation methods will eventually require more local pieces of evidence from the dialogue text for close diagnosis. For this reason, we collected both subjective annotations and nugget annotations for each dialogue, in the hope that automatic evaluation measures defined as a function of nuggets will eventually be able to predict subjective scores with reasonable accuracy. While our test collections contain *manually* identified nuggets, a possible next step would be to devise ways to extract them automatically. Another possible benefit of constructing nuggets is that a set of nuggets collected from a dialogue may also be useful for evaluating a different dialogue that discusses a similar problem (i.e., reusability).

2 RELATED WORK

2.1 Evaluating Non-Task-Oriented Dialogues

Evaluating generated responses in non-task-oriented dialogues is a difficult problem. Galley *et al.* [4] proposed *Discriminative* BLEU,

[1] http://www.weibo.com

EVIA'17, Dec 2017, Tokyo, Japan
2017. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 multiple hand-crafted references for conversational data, they 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.

Higashinaka *et al.* [6] ran the first *Dialogue Breakdown Detection Challenge* using Japanese human-machine 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 annotators 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, precision, recall and F-measure were computed. Moreover, 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 third Dialogue Breakdown Detection Challenge is currently in preparation[2].

At NTCIR-12 , the first Short Text Conversation (STC) task was run using Weibo data (for the Chinese subtask) and Twitter data (for the Japanese subtask), attracting 22 participating teams [19]. 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 were used to evaluate the participating systems. Gold standard labels were created manually by hiring multiple annotators 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; this time, systems are allowed to generate their own responses[3].

2.2 Evaluating Task-Oriented Dialogues

Two decades ago, Walker *et al.* [20] 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-machine dialogues for a specific task (e.g., train timetable lookup) as well as subjective ratings 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 [21]. 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.

Lowe *et al.* [10] released the Ubuntu Dialogue Corpus, which contains 930,000 *human-human* dialogues extracted from Ubuntu chats. 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. However, while they automatically disentangled the chats to form dyadic dialogues, their original chat logs usually involve more than two parties, which makes it different from our dyadic customer-helpdesk DCH-1 dataset. 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 are 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.

The most straightforward approach to evaluating dialogues is to collect subjective assessments from the user who actually experienced the dialogue. 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*

[2] https://dbd-challenge.github.io/dbdc3/
[3] http://ntcirstc.noahlab.com.hk/STC2/stc-cn.htm

al. [5] applied a service quality assessment from marketing to the evaluation of telephone-based email application; their method is known as SERVQUAL. Paek [13] discusses SASSI, SERVQUAL and PARADISE in a survey paper that discusses spoken dialogue evaluation, along with his Wizard-of-Oz approach of using human performance to replace a system component in order to define a gold standard.

2.3 Evaluating Textual Information Access

While the aforementioned BLEU [14] is basically equivalent to an *n*-gram-based precision, ROUGE [8], 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 [12] and question answering evaluation. In the TREC Question Answering (QA) tracks, a nugget is defined as "*a fact for which the annotator could make a binary decision as to whether a response contained that nugget*" [9]. Having constructed nuggets, (weighted) recall, precision and F-measure scores can be computed, except that the precision computation requires special handling: while one can count the number of nuggets present or missing in the system output, one 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 Cross-lingual Information Access) Task adapted these methods for evaluating QA with Asian languages [11].

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 [18] 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, 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 we should minimise the amount of text that the user has to read. Sakai and Kato [17] 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.*"

Table 1: Test collection statistics. ∗Only 40 dialogues from DCH-0 were annotated with nuggets.

	DCH-0	DCH-1
Source	www.weibo.com	
Language	Chinese	
Data timestamps	Jan. 2013 - Sep. 2016	
#Dialogues	234	3,700
#English translations	40	370
#Helpdesk accounts	16	161
Avg. #posts/dialogue	13.402	4.512
Avg. #utterance blocks/dialogue	12.021	4.162
Avg. post length (#chars)	35.011	44.568
Avg. utterance block length length (#chars)	39.031	48.313
#annotators/dialogue	2	3
Subjective annotation criteria	TS, TA, CS, HA, CA (See Section 3.4)	
Nugget types	CNUG0, CNUG, HNUG, CNUG∗, HNUG∗ (See Section 3.5)	
Triggerless dialogues	1∗	184

Sakai and Dou [16] generalised the idea of S-measure to handle various textual information access tasks, including web search. Their measure, known as *U-measure*, constructs 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 (See Section 4).

3 DESIGNING AND BUILDING DCH-1

3.1 Overview

Our ultimate goal is automatic evaluation of human-machine Customer-Helpdesk dialogues. As a first step towards it, we built two test collections based on *real* (i.e., human-human) Customer-Helpdesk dialogues, which we call DCH-0 and DCH-1.

DCH-0, our smaller collection, was used to establish an efficient and reliable test collection construction procedure. For example, although we started constructing DCH-0 by using the number of *posts* in each dialogue for sampling dialogues of different lengths, where a post refers to a piece of timestamped text entered by either Customer or Helpdesk, we quickly realised that posts are often a mere artifact of the Weibo users' arbitrary hits of the ENTER key, and that they are not suitable as the basic semantic unit. Based on this experience, we used the *utterance block* as the basis for measuring the length of a dialogue in DCH-1, formed by merging all consecutive posts by the same utterer.

Table 1 provides some statistics of DCH-0 and DCH-1. As shown in the table, 184 of the 3,700 DCH-1 dialogues are "triggerless," by which we mean that Customer and Helpdesk exchange remarks even though Customer does not seem to be facing any problem (*cf.* Figure 1)[4]. Below, we discuss the construction and validation of DCH-1.

[4] We tried filtering out these triggerless dialogues for the analyses reported in Section 5, but the effect of this on our results was not substantial.

3.2 Dialogue Mining

The 3,700 Helpdesk dialogues contained in the DCH-1 test collection were mined from Weibo in September 2016 as follows. (1) We collected an initial set of Weibo accounts by searching Weibo account names that contained keywords such as "assistant" and "helper" (in Chinese). We denote this set by A_0. (2) For each account name a in A_0, we added a prefix "@" to a and used the string as a query for searching up to 40 conversational threads (i.e., initial post plus comments on it) that contain a mention of the official account[5]. We then filtered out accounts that did not respond to over one half of these threads. We denote the filtered set of "active" accounts as A. (3) For each account a in A, we retrieved all threads that contain a mention of a from January 2013 to September 2016, and extracted Customer-Helpdesk dyadic dialogues from them. We then kept those that consist of at least one utterance block by Customer *and* one by Helpdesk. As a result, 21,669 dialogues were obtained. This collection is denoted as D_0. (4) As D_0 is too large for annotation, we sampled 3,700 dialogues from it as follows. For $i = 2, 3, \ldots, 6$, we randomly sampled 700 dialogues that contained i utterance blocks. In addition, we randomly sampled 200 that contained $i = 7$ utterance blocks; we could not sample 700 dialogues for $i = 7$ as D_0 did not contain enough dialogues that are very long.

10% (370) of the Chinese Dialogues in DCH-1 were manually translated English by a professional translation company for research purposes.

3.3 Annotators

We hired 16 Chinese undergraduate students from the Faculty of Science and Engineering at XXXXX University so that each Chinese dialogue was annotated independently by three annotators. The assignment of dialogues to annotators was randomised; given a dialogue, each annotator first read the entire dialogue carefully, and then gave it ratings according to the five subjective annotation criteria described in Section 3.4; finally, he/she identified nuggets within the same dialogue, where nuggets were defined as described in Section 3.5. An initial face-to-face instruction and training session for the annotators was organised by the first author of this paper at XXXXX University; subsequently, the annotators were allowed to do their annotation work online using a web-browser-based tool at their convenient location and time. The number of dialogues assigned to each annotator was $3,700 * 3/16 = 693.75$ on average; all of them completed their work within two weeks as they were initially asked to do. The actual annotation time spent by each annotator was 18-20 hours.

3.4 Subjective Annotation

By subjective annotation, we mean manual quantification of the quality of a dialogue as a whole. As there are two players involved in a Customer-Helpdesk dialogue, we wanted to accommodate the following two 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.

Moreover, we wanted to assess *customer satisfaction* as this is of utmost importance for both parties. While customer satisfaction ratings should ideally be collected from the *real* customer at the time of dialogue termination, we had no choice but to collect surrogate, post-hoc ratings by the annotators instead.

By considering the above points as well as our results from the smaller DCH-0 collection, we finally devised the following five subjective annotation criteria:

Task Statement Whether the task (i.e., the problem to be solved) is clearly stated by Customer (denoted by **TS**);

Task Accomplishment Whether the task is actually accomplished (denoted by **TA**);

Customer Satisfaction Whether Customer is likely to have been satisfied with the dialogue, and to what degree (denoted by **CS**);

Helpdesk Appropriateness Whether Helpdesk provided appropriate information (denoted by **HA**);

Customer Appropriateness Whether Customer provided appropriate information (denoted by **CA**).

Figure 2 shows the actual instructions for annotators: note that **CS** is on a five-point scale (−2 to 2), while the other four are on a three-point scale (−1 to 1). Table 2 shows the inter-rater agreement (for three assessors) of the subjective labels in terms of Fleiss' κ [3] and Randolph's κ_{free} [15]; κ_{free} is known to be more suitable when the labels are heavily skewed across the categories, which is indeed the case here. "2+ agree" means the proportion of dialogues for which at least two annotators agree, e.g., $(-1, -1)$; "3 agree" means the proportion of dialogues for which all three annotators agree, e.g., $(-1, -1, -1)$.

It can be observed that the agreement among the three assessors is low, except perhaps for **TS**, which reflects the highly subjective nature of this labelling task. While it may be possible to improve the inter-assessor agreement a little in our future work by revising the labelling instructions, it should be stressed that our labelling task is not document relevance assessments, and that it is inherently highly subjective. We believe that, as our future work, hiring more than three assessors and preserving their different viewpoints in the test collection, is more important than trying to force them into reaching an agreement.

3.5 Nugget Annotation

We had three annotators independently identified nuggets for each dialogue as follows. At the instruction and training session, annotators were given the diagram shown in Figure 3, which reflects our view that accumulating nuggets will eventually solve Customer's problem, together with a written definition of nuggets, as described below. (1) A nugget is a post, or a sequence of consecutive posts by the same utterer (i.e., either Customer or Helpdesk). (2) It can neither partially nor wholly overlap with another nugget. (3) 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

[5] Weibo's interface for conversational threads is somewhat different from Twitter's: comments to a post are not displayed on the main timeline; they are displayed under each post only if the "comments" button is clicked.

TS: Does Customer communicate the problem clearly to HelpDesk?
1: Yes, 0: Partially, -1: No

TA: Is Customer's problem solved?
1: Yes, 0: Partially, -1: No

CS: How satisfied with the dialogue is Customer?
2: Highly satisfied, 1: Moderately satisfied, 0: Neutral
-1: Moderately dissatisfied, -2: Moderately dissatisfied

HA: Helpdesk utterance quality: Does Helpdesk ask appropriate questions and/or provide appropriate information to Customer during the dialogue?
1: Yes, 0: Maybe, -1: No

CA: Customer utterance quality: Does Customer provide appropriate information to Helpdesk during the dialogue?
1: Yes, 0: Maybe, -1: No

Figure 2: Subjective annotation criteria.

Table 2: Inter-annotator agreement of the subjective annotations for DCH-1 (3,700 dialogues, 3 annotators per dialogue). Note that Fleiss' κ and Randolph's κ_{free} treat the ratings as *nominal* categories. 2+ agree means the proportion of dialogues for which at least two annotators agree; 3 agree means the proportion of dialogues for which all three annotators agree. For CS, 2 and 1 were treated as 1, and −2 and −1 were treated as −1.

	2+ agree	3 agree	Fleiss' κ	κ_{free}
TS	.981	.729	.301	.719
TA	.925	.361	.273	.324
CS	.938	.349	.276	.318
HA	.873	.309	.197	.245
CA	.857	.288	.141	.216

that does not contribute to the Customer transition (See Figure 3). (4) It helps Customer transition from Current State (including Initial State) towards Target State (i.e., when the problem is solved).

Note that we utilise Weibo *posts* as the atomic building blocks for forming nuggets; This takes into account the remark by Wang *et al.* [22]: "*Experience from question answering evaluations has shown that users disagree about the granularity of nuggets—for example, whether a piece of text encodes one or more nuggets and how to treat partial semantic overlap between two pieces of text.*" Note also that according to our definition, an utterance block (i.e., maximal consecutive posts by the same utterer) generally subsumes one or more nuggets.

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: (1) A dialogue involves two parties, Customer and Helpdesk; (2) Even within the same utterer, 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 (but not *closed-domain*, which makes slot filling approaches infeasible), there must be some nuggets that represent the state of *identifying* the task and those that represent the state of *accomplishing* it.

Based on the above considerations, we defined 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.

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.

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

Using a pull-down menu on our web-browser-based tool, assessors categorised each *post* into CNUG0, CNUG, HNUG, CNUG$*$, HNUG$*$, or NAN (not a nugget). Then, consecutive posts with the same label (e.g., CNUG followed by CNUG) were automatically merged to form a nugget.

Table 3 shows the inter-annotator agreement of the nugget annotations, where the posts are used as the basis for comparison. The 3,700 dialogues in DCH-1 contains a total of 7,155 Helpdesk posts, all of which were annotated independently by three annotators, producing a total of 21,465 annotations, A direct comparison with the subjective annotation agreement shown in Table 2 would be difficult, since both the *annotation unit* (dialogues vs. nuggets) and the *annotation schemes* (numerical ratings vs. nugget types) are different. However, it can be observed that the agreement for Customer nuggets is substantially higher than for the Helpdesk nuggets. A possible explanation for this would be that it is easier for annotators to judge the contribution of Customer's utterances for reaching his/her target state than to judge that of Helpdesk, at least for regular nuggets: while Helpdesk often *asks* Customer for more information regarding the problem context, it is Customer's utterances that actually *provide* that information.

While directly comparing the *inter-annotator* agreement of subjective annotation and nugget annotation seems difficult, we would like to compare the *intra-annotator* consistency by making each annotator process the same dialogue multiple times in our future work.

4 UCH: A DIALOGUE EVALUATION MEASURE

We now propose an evaluation measure that leverages nuggets for quantifying the quality of Customer-Helpdesk dialogues. We regard a Customer-Helpdesk dialogue as a *trailtext* of U-measure, which may or may not contain nuggets. Let *pos* denote the position (i.e., offset from the beginning of the dialogue) of a nugget; for ideographic languages such as Chinese and Japanese, we use the number of *characters* to define the offset position. Given a *patience*

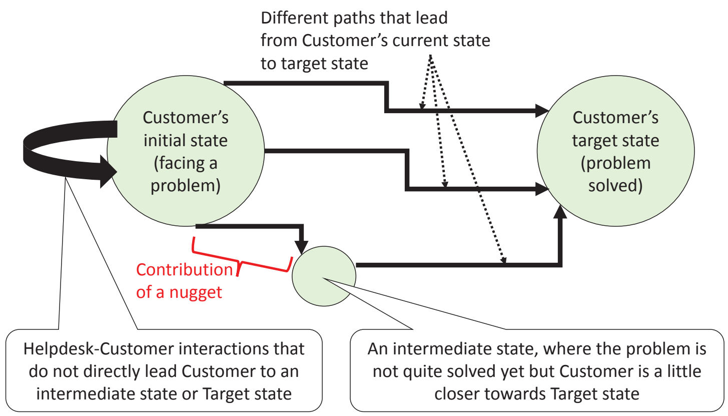

Figure 3: Task accomplishment as state transitions, and the role of a nugget.

Table 3: Inter-annotator agreement of the nugget annotations for DCH-1 (3,700 dialogues, 3 annotators per dialogue). 2+ **agree means the proportion of nuggets for which at least two annotators agree; 3 agree means the proportion of dialogues for which all three annotators agree. NAN means "not a nugget." 95% CI for κ are also shown.**

	2+ agree	3 agree	Fleiss'κ	κ_{free}
Helpdesk (#total posts) (HNUG/HNUG* /NAN)	.907	.299	.174 [.165, .184]	.253
Customer (#total posts) (CNUG0/CNUG /CNUG*/NAN)	.959	.491	.488 [.481, .496]	.529

parameter L, we define a *decay function* over the trailtext as [16]:

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

This is for discounting the value of a nugget that appear later in the dialogue; at position L, the value of any nugget wears out completely. In our experiments, we let $L = L_{max} = 916$ as this is the number of (Chinese) characters in the longest dialogue from the DCH-1 collection. The benefit of introducing L is discussed in Section 5.2.

Let N and M denote the number of Customer's non-goal nuggets and Helpdesk's non-goal nuggets identified within a dialogue, respectively; for simplicity, let us assume that there is at most one Customer's goal nugget (c_*) and at most one Helpdesk's goal nugget (h_*) 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 position of nugget c_i; $pos(h_j)$ ($j \in \{1, \ldots, M, *\}$) is defined similarly.

Given the *gain value* of each non-goal nugget ($g(c_i)$), a simple evaluation measure based solely on Customer's utterances can be computed as:

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

In the present study, we define the gain value of CNUG* as $g(c_*) = 1 + \sum_{i=1}^{N} g(c_i)$. This is an attempt at reflecting the view that *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, given the gain value of each non-goal nugget ($g(h_j)$), a measure solely based on Helpdesk's utterances can be computed as:

$$UH = \sum_{h_j \in \{h_1, \ldots, h_M, h_*\}} g(h_j) \ D(pos(h_j)) \ , \tag{3}$$

where $g(h_*) = 1 + \sum_{j=1}^{M} g(h_j)$. Finally, for a given parameter α ($0 \leq \alpha \leq 1$) that specifies the *contribution* of Helpdesk's utterances relative to Customer's, we can define the following combined measure:

$$UCH_{\alpha} = (1 - \alpha) UC + \alpha UH \ . \tag{4}$$

By default, we use $\alpha = 0.5$. Note that $UCH_{0.5}$ is equivalent to computing a single U-measure score without distinguishing between Customer's and Helpdesk's nuggets. The choice of α is discussed in Section 5.3.

Since we have three independent nugget annotations per dialogue, We tried two approaches to computing a single score for a given dialogue: *Average UCH* (AUCH) simply computes a UCH score each annotator and then takes the average for that dialogue; *Consolidated UCH* (CUCH) merges the nuggets from multiple annotators first and then computes a single UCH score. Due to lack of space, we only report on results with AUCH, which consistently outperformed CUCH in our experiments.

5 ANALYSIS WITH UCH

This section addresses the following questions: *How does UCH correlate with subjective ratings?* (Section 5.1); *Is the patience parameter L useful for estimating subjective ratings?* (Section 5.2); and *Which utterer plays the major role when estimating subjective ratings with UCH?* (Section 5.3).

In the analysis reported below, we use the z-score of each subjective rating before averaging them over the three annotators. That is, for each annotator and subjective criterion, we first compute the

Table 4: Kendall's τ between AUCH and average subjective ratings for DCH-1 (3,700 dialogues), with 95% CIs.

	AUCH
TS	.267 [.237, .277]
TA	.256 [.244, .289]
CS	.118 [.097, .141]
HA	.414 [.398, .432]
CA	.434 [.417, .450]

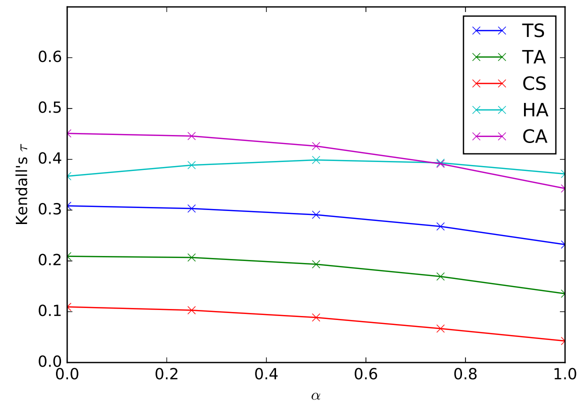

Figure 5: Effect of α on the τ between average subjective ratings and AUCH.

mean and standard deviation of the raw ratings, and then process each raw rating by subtracting the mean and then dividing by the standard deviation. This is to remove each annotator's inherent scoring tendency.

5.1 Correlation with Subjective Annotations

Table 4 shows the Kendall's τ values between AUCH and the average subjective ratings for the DCH-1 collection, with 95% confidence intervals. It can be observed that AUCH is reasonably highly correlated with **HA** (.414, 95% CI[.398, .432]) and **CA** (.434, 95% CI[.417, .450]). That is, even though the inter-annotator agreement for *appropriateness* is relatively low (Table 2), AUCH manages to estimate the *average* appropriateness with reasonable accuracy. On the other hand, the table shows that the τ between AUCH and **CS** is very low, albeit statistically significant (.118, 95% CI[.097, .141]). One possible explanation for this might be that the **CS** ratings themselves are not as reliable as we would like. First, as we have discussed in Section 3.4, the annotators are not the actual customers; second, our manual inspection of some of the dialogues from DCH-0 and DCH-1 suggest that the annotator's ratings may be influenced by his/her prior impression of the product/service or the company, rather than the contents of the particular dialogue in question.

5.2 The Patience Parameter L

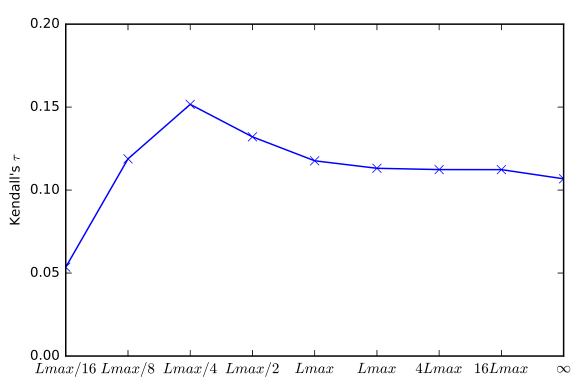

Figure 4: Effect of L on the τ between average customer satisfaction and AUCH.

As was explained in Section 4, UCH inherits the patience parameter L from S-measure [18] and U-measure [16], to discount the value of a nugget based on its position within the dialogue. As we have mentioned earlier, we let $L = L_{max} = 916$ by default, as this is the length of the longest dialogue within DCH-1. Using a small L means that the decay function becomes steep and that we do not tolerate long dialogues; using an extremely large L is equivalent to switching off the decay function, thereby treating the dialogue as a *set* of nuggets (See Eq. 1).

Figure 4 shows the effect of L on the τ between average **CS** and AUCH. It can be observed that, at least for DCH-1, $L = L_{max}/4 = 229$ seems to be a good choice if AUCH is to be used for estimating customer satisfaction. This suggests that user satisfaction may be linked to user patience, and that considering nugget positions as UCH does is of some use. However, as was discussed earlier, the reliability of the **CS** ratings deserves a closer investigation in our future work.

5.3 The Contribution Parameter α

As Eq. 4 shows, UCH can decide on a balance between Customer's utterances and Helpdesk's; a small α means that we rely more on Customer nuggets for computing UCH. Figure 5 shows the effect of α on the τ between AUCH and different average subjective ratings. The trends are the same for **TS**, **TA**, **CS**, and **CA**: the smaller the α, the higher the rank correlation. That is, to achieve the highest τ, it is best to rely entirely on Customer utterances, i.e., to completely ignore Helpdesk utterances.

Interestingly, however, the trend is different for **HA**: the curve for **HA** suggests that $\alpha = 0.5$, our default value, is in fact the best choice. That is, to achieve the highest τ with Helpdesk Appropriateness, treating Customer's and Helpdesk's nuggets equally appears to be a good choice. While it is obvious that Helpdesk's utterances need to be taken into account in order to estimate Helpdesk Appropriateness, the curve implies that Customer's utterances also play an important part in the estimation. These results suggest that different subjective annotation criteria requires different balances between Customer's and Helpdesk's utterances.

6 CONCLUSIONS

As an initial step towards evaluating automatic dialogue systems, we constructed DCH-1, which contains 3,700 real Customer-Helpdesk multi-turn dialogues mined from Weibo. We have annotated each dialogue with subjective quality annotations (**TS**, **TA**, **CS**, **HA**, and **CA**) and nugget annotations, with three annotators per dialogue. In addition, 10% of the dialogues have been manually translated into English. We described how we constructed the test collection and the philosophy behind it. We also proposed UCH, a simple nugget-based evaluation measure for task-oriented dialogue evaluation, and explored its usefulness and limitations. Our main findings on UCH based on the DCH-1 collection are as follows.

(1) UCH correlates better with subjective ratings that reflect the appropriateness of utterances (**HA** and **CA**) than with customer satisfaction (**CS**);
(2) The patience parameter L of UCH, which considers the positions of nuggets within a dialogue, may be a useful feature for enhancing the correlation with customer satisfaction;
(3) For the majority of our subjective annotation criteria, customer utterances seem to play a much more important role for UCH to achieve high correlations with subjective ratings than helpdesk utterances do, according to our analysis on the parameter α.

Our future work includes the following:

- Comparing subjective annotation and nugget annotation in terms of *intra*-annotator agreement;
- Investigating the reliability of offline customer satisfaction ratings by comparing them with real customer ratings collected right after the termination of a helpdesk dialogue;
- Collecting subjective and nugget annotations for the English subcollection of DCH-1, and comparing across Chinese and English;
- Devising ways for automatic nugget identification and automatic categorisation of nuggets into different nugget types;
- Running a shared task (e.g., at NTCIR) by treating the human-human dialogues in DCH-1 as ideal cases
- Running a shared task (e.g., at NTCIR) by leveraging DCH-1 as a small training data set: given a dialogue, participating systems are required to estimate the distribution of subjective scores such as user satisfaction over multiple annotators, as well as the distribution of nugget types (e.g. trigger, regular, goal, not-a-nugget) over multiple assessors for each utterance [1].

REFERENCES

[1] Anonymous: Towards Automatic Evaluation of Multi-Turn Dialogues: A Task Design that Leverages Inherently Subjective Annotations, *Submitted to EVIA 2017*.
[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] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 74–81 (2004).
[9] Lin, J. and Demner-Fushman, D.: Will Pyramids Built of Nuggets Topple Over?, *Proceedings of HLT/NAACL 2006*, pp. 383–390 (2006).
[10] 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).
[11] 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).
[12] 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).
[13] Paek, T.: Toward Evaluation that Leads to Best Practices: Reconciling Dialog Evaluation in Research and Industry, *Bridging the Gap: Academic and Industrial Research in Dialogue Technologies Workshop Proceedings*, pp. 40–47 (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. and Dou, Z.: Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation, *Proceedings of ACM SIGIR 2013*, pp. 473–482 (2013).
[17] Sakai, T. and Kato, M. P.: One Click One Revisited: Enhancing Evaluation based on Information Units, *Proceedings of AIRS 2012* (2012).
[18] 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).
[19] 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).
[20] 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).
[21] 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).
[22] Wang, Y., Sherman, G., Lin, J. and Efron, M.: Assessor Differences and User Preferences in Tweet Timeline Generation, *Proceedings of ACM SIGIR 2015*, pp. 615–624 (2015).