Overview of the NTCIR-14 Short Text Conversation Task: Dialogue Quality and Nugget Detection Subtasks

Zhaohao Zeng[1], Sosuke Kato[1], and Tetsuya Sakai[1]

Waseda University, Tokyo, Japan
zhaohao@fuji.waseda.jp,sow@suou.waseda.jp,tetsuyasakai@acm.org

Abstract. In this paper, we provide an overview of the NTCIR-14 Short Text Conversation-3 Dialogue Quality (DQ) and Nugget Detection (ND) subtasks. Both DQ and ND subtasks aim to evaluate customer-helpdesk dialogues automatically: (1) DQ subtask is to assign quality scores to each dialogue in terms of three criteria: task accomplishment, customer satisfaction, and efficiency; and (2) ND subtask is to classify whether a customer or helpdesk turn is a nugget, where being a nugget means that the turn helps towards problem solving. In this overview paper, we describe the task details, evaluation methods and dataset collection, and report the official results.

1 Introduction

As research of Natural Language Processing for Artificial Intelligence has progressed, many dialogue-related competitions have been held. One dialogue-related competition, Short Text Conversation (STC) task has been held in NTCIR. The first STC task in NTCIR-12 required the system of the participants to retrieve texts from the given repository as the IR problem. The second STC task in NTCIR-13 allowed the systems not only to retrieve texts but also to generate texts. While the STC-1 and STC2 tasks required the systems to retrieve/generate comments for microblogs, in NTCIR-14 the STC-3 task has three subtasks:

- Chinese Emotional Conversation Generation (CECG) subtask;
- Dialogue Quality (DQ) subtask;
- Nugget Detection (ND) subtask.

The Chinese Emotional Conversation Generation (CECG) subtask is discussed in a separate overview paper [17]. In this paper, we give an overview of two subtasks: Dialogue Quality (DQ) subtask and Nugget Detection (ND) for Chinese and English, where we attempt to tackle the problem of dialogue evaluation for helpdesk-customer dialogues. Recent advances in artificial intelligence suggest that, in the not-too-distant future, these *human-human* Customer-Helpdesk dialogues may soon be replaced by *human-machine* ones. To enable researchers and engineers to build and tune intelligent helpdesk agent systems efficiently, reliable automatic evaluation measures need to be established. However, as there

currently is no standard framework for evaluating textual, dyadic (i.e., between two utterers), multi-round, task-oriented dialogues, we propose two approaches to evaluating such dialogues in this paper. An example of a Customer-Helpdesk dialogue is shown in Figure 1: this is a two-round dialogue (i.e., there are two Customer-Helpdesk exchanges). 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.

Evaluating Customer-Helpdesk dialogues is related but very different from the traditional *closed-domain* spoken dialogue systems where the task boils down to *slot filling*. For example, for a ticket booking task, the system can utilise a predefined schema and converse with the customer to fill the necessary slots such as "depart-city" and "arrival-city" [13]. In contrast, Customer-Helpdesk dialogues discuss diverse problems about products and services and therefore required slots cannot be listed up exhaustively in advance. While it is possible to ask users to provide their feedback or hire human to evaluate customer-helpdesk dialogues manually, these methods are expensive and does not scale. Furthermore, these may lead to the difficulty in comparing different systems, and are not repeatable even for the same system. To advance the research in this field, it is necessary to build an approach to evaluating the systems we are building automatically [16].

As a first step towards automatically evaluating Customer-Helpdesk dialogues, we constructed a training and test collection comprising 4,090 (3,700 training + 390 test) real customer-helpdesk multi-round dialogues by mining Weibo[1], a major Chinese microblogging website. Furthermore, 2,062 (1,672 training + 390 test) dialogues have been translated into English to build an English version. Each dialogue has been annotated with *subjective quality annotations* as well as *nugget annotations*. Subjective quality annotations comprise *task accomplishment*, *customer satisfaction*, and *efficiency* scores, which can be regarded as the gold standard by researchers seeking ways to automatically evaluate the quality of a given dialogue. Nugget annotations comprise nugget type labels for each nugget within a dialogue, where a nugget is defined to be a turn that helps the customer advance towards problem solving. Participants are required to build systems to read a customer-helpdesk dialogue and predict (1) quality scores for each dialogue in Dialogue Quality (DQ) subtask (2) nugget types for each dialogue turn in Nugget Detection (ND) subtask. More details about the task definition and evaluation methods will be presented in Section 2 and Section 3. The dataset provided by ND and DQ subtasks will be detailed in Section 4, and Section 7 will present the official evaluation results.

The schedule of the STC-3 DQ and ND subtasks is shown in Table 1. We received 13 runs for the Chinese subtasks and 11 runs for the English runs. Table 2 shows the detailed number of runs submitted by each team for each subtask and each language.

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

Table 1. Schedule of the STC-3 DQ and ND subtasks at NTCIR-14

Time	Content
Sep, 2018	Test data released
Nov 30, 2018	Run submissions due
Feb 1, 2019	Results summary and draft overview released
Mar 15, 2019	Participant paper submissions due
May 1, 2019	All camera-ready papers due
Jun, 2019	NTCIR-14 Conference & EVIA 2019, Tokyo

Table 2. The Number of Runs

Team	Chinese		English	
	DQ	ND	DQ	ND
CUIS	1	0	1	0
SLSTC	3	3	3	3
WIDM	0	0	2	2
WUST	3	3	0	0
Total	7	6	6	5

2 Task Definition

Our goal is to explore methods to evaluate task-oriented, multi-round, textual helpdesk-customer dialogue systems automatically. Specifically, we designed two subtasks: (1) Dialogue Quality (DQ) subtask, which is to assign quality scores to each dialogue in terms of three subjective criteria: task accomplishment, customer satisfaction, and efficiency; and (2) Nugget Detection (ND) subtask is to classify whether a customer or helpdesk turn is a nugget, where being a nugget means that the turn helps towards problem solving. This section details what a customer-helpdesk dialogue is, followed by the definitions of the two subtasks.

2.1 Customer-Helpdesk Dialogue

In DQ and ND subtasks, a customer-helpdesk dialogue is a multi-round and textual dialogue that has two speakers: a Customer and a Helpdesk. The Customer usually comes with a problem and the helpdesk should try to help the customer to solve it. An example of a Customer-Helpdesk dialogue is shown in Figure 1: this is a two-round dialogue (i.e., there are two Customer-Helpdesk exchanges). 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.

We used the *turn* as the basis for measuring the length of a dialogue, formed by merging all consecutive posts by the same utterer. For example, if each Cus-

Fig. 1. An example of a dialogue between Customer (C) and Helpdesk (H). The left part is the translated dialogue and the right part is the screenshot of the original dialogue on Weibo.

tomer post is denoted by p_C and each helpdesk post is denoted by p_H, a dialogue of the form $[p_C, p_C, p_C, p_H, p_H, p_H, p_C, p_C]$ will be regarded as three turns, $[b_C, b_H, b_C]$, where b_C is a Customer turn and b_H is a Helpdesk one. This dialogue is considered as a three-turn dialogue.

2.2 Dialogue Quality (DQ) subtask

In Dialogue Quality (DQ) subtask, we want to obtain the subjective scores for each dialogue automatically to quantify the quality of a dialogue as a whole. Specifically, we introduce three quality scores for three different criteria:

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

For each of them, possible options are [2, 1, 0, −1, −2]. In other words, participants are required to assign a score from 2 to −2 for each of these criteria to each dialogue.

2.3 Nugget Detection (ND) subtask

In Nugget Detection (ND) subtask, participants are required to identify nuggets for each dialogue, where a nugget is an turn that helps the Customer transition from the current state (where the problem is yet to be solved) towards the

target state (where the problem has been solved). Figure 2 reflects our view that accumulating nuggets will eventually solve Customer's problem. The official definition of nuggets is (1) A nugget is a turn by either Helpdesk or Customer; (2) It can neither partially nor wholly overlap with another nugget; (3) It helps Customer transition from Current State (including Initial State) towards Target State (i.e., when the problem is solved).

Compared to traditional nugget-based information access evaluation, there are two unique features in nugget-based helpdesk dialogue evaluation:

– A dialogue involves two parties, Customer and Helpdesk;
– 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 turns that are useful from Customer's point of view.

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

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

CNUG∗ Customer's *goal nuggets*. These are nuggets in Customer's turns which tell Helpdesk that Customer's problem has been solved.

CNAN Customer's *not a nugget*. It means that the current customer turn does not help towards problem solving.

HNAN Helpdesk's *not a nugget*. It means that the current helpdesk turn does not help towards problem solving.

In the ND subtask, participants are required to predict a nugget type for each turn in dialogues. Note that each nugget type may or may not be present in a dialogue, and multiple nuggets of the same type may be present in a dialogue.

3 Evaluation Method

Evaluating such a customer-helpdesk dialogue is even subjective and difficult for human, and often there is no such thing as the ground truth: different people may have different opinions about the dialogue [8]. Instead of evaluating both ND and DQ subtasks as simple classification problems using metrics like accuracy, we evaluate these subtasks by comparing the probability distribution estimated

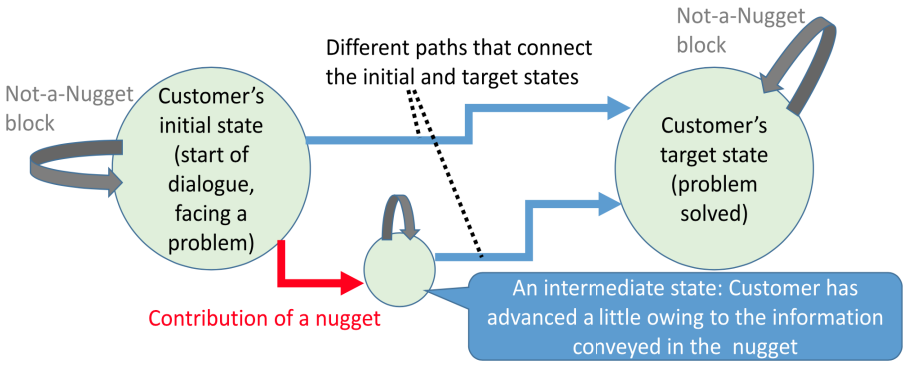

Fig. 2. Task accomplishment as state transitions, and the role of a nugget.

by the participants with the golden standard distribution, where the golden standard distribution is calculated by annotators' vote over the classes (i.e. 2 to −2 for DQ subtask and CNUG, HNUG, etc. for ND subtask).

We now formalise some measures for comparing two probability distributions. Let A denote a given set of classes, e.g., $A = 2, 1, 0, -1, -2$ for DQ subtask, and let $L = |A|$. Let $p(i)(i = 1, \ldots, L)$ denotes denote the system 's estimated probability for class i, so that $\sum_{i \in A} p(i) = 1$. Similarly, let $p^*(i)$ denote the corresponding true probability, where $\sum_{i \in A} p^*(i) = 1$.

3.1 Evaluation Measures for Dialogue Quality subtask

Since the classes of DQ subtask are non-nominal, cross-bin measures is more suitable than bin-by-bin measures. As discussed by Sakai [9], bin-by-bin measures such as Jensen-Shannon Divergence (See Section 3.2) are not adequate for this subtask as they do not consider the *distance* between classes. Thus, we utilise two cross-bin measures: *Normalised Match Distance* (NMD) and *Root Symmetric Normalised Order-aware Divergence* (RSNOD).

Normalised Match Distance (NMD) is a normalised version of Match Distance (MD), where MD is a special case of Earth Mover 's Distance where the probabilities add up to one and the number of bins area given [6]. Let $cp(i) = \sum_{k=1}^{i} p(k)$, and $cp^*(i) = \sum_{k=1}^{i} p^*(k)$. MD is just the sum of absolute errors compared from the cumulative probability distributions:

$$MD(p, p^*) = \sum_{i \in A} |cp(i) - cp^*(i)|. \quad (1)$$

Then, the normalised version NMD is calculated as follows:

$$NMD(p, p^*) = \frac{MD(p, p^*)}{L - 1} \quad (2)$$

Root Symmetric Normalised Order-aware Divergence (RSNOD) is a measure that considers the distance between a pair of bins more explicitly than NMD does [9]. First, a *distance-weighted* sum of squares (DW) is defined for each bin:

$$DW(i) = \sum_{j \in A} |i - j|(p(j) - p^*(j))^2. \tag{3}$$

Let $B^* = i|p^*(i) > 0$, that is, the set of bins where the gold probabilities are positive. *Order-Aware Divergence* (OD) is the DW averaged over these non-empty gold bins:

$$OD(p||p^*) = \frac{1}{|B^*|} \sum_{i \in B^*} DW(i) \tag{4}$$

Similarly, let $B = i|p(i) > 0$. Just as the symmetric JSD is obtained from KLD, *Symmetric* OD can be defined by swapping the system and gold distributions:

$$SOD(p, p^*) = \frac{OD(p, p^*) + OD(p^*, p)}{2} \tag{5}$$

Finally, we define the Root Symmetric Normalised OD:

In the DQ subtask, we use both NMD and RSNOD as M to evaluate participants' runs.

3.2 Evaluation Metrics for Nugget Detection subtask

In contrast to DQ subtask, the classes in ND subtask are nominal, so bin-by-bin measures are more suitable. Specifically, two measures are used in ND subtask: *Root Normalised Sum of Squares* (RNSS) and *Jensen-Shannon Divergence* (JSD).

Root Normalised Sum of Squares (RNSS) is defined as follows:

$$RNSS = \sqrt{\frac{\sum_{i \in A}(p(i) - p^*(i))^2}{2}} \tag{6}$$

Jensen-Shannon Divergence (JSD) Let $p_M(i) = \frac{p(i)+p^*(i)}{2}$, JSD is defined as:

$$JSD(p||p^*) = \frac{KLD(p||p_M) + KLD(p_M||p^*)}{2} \tag{7}$$

$$\text{where} \quad KLD(p1||p2) = \sum_{i \ s.t. \ p1(i)>0} p_1(i) \log_2 \frac{p_1(i)}{p_2(i)} \tag{8}$$

Since there are multiple turns in each dialogue and participants are required to predict a probability distribution for each turn in ND subtask, we need to find a way to combine the measure scores into a single one for each dialogue. Specifically, we calculate the average measure score for customer's turns S_C and helpdesk's turns S_H separately, and then a weighted sum $S_{ND} = \alpha S_C + (1 - \alpha)S_H$ will be used as the final evaluation score for each dialogue, where α is a parameter that controls the relatively importance between customers' nuggets and helpdesk' nuggets. By default, $\alpha = 0.5$.

Table 3. Statistics of Chinese Data collection.

	Training Set	Test Set
Source	`www.weibo.com`	
Language	Chinese	
Data timestamps	Jan. 2013 - Apr. 2018	
#annotators/dialogue	19	
#Dialogues	3,700	65
Avg. #posts/dialogue	4.512	4.877
Avg. post length (#chars)	44.568	47.988
Avg. turn length	48.313	52.008
Quality annotation criteria	A-score, E-score, S-score (See Section 2.2)	
Nugget types	CNUG0, CNUG, HNUG, CNUG∗, HNUG∗ (See Section 2.3)	

4 Test Collection

Our ultimate goal is automatic evaluation of human-machine Customer-Helpdesk dialogues. As a first step towards it, we built a dataset based on *real* (i.e., human-human) Customer-Helpdesk dialogues. Table 4 provides some statistics of the dataset. As shown in the table, DCH-1 consists of 3,700 Chinese dialogues for training and 390 dialogues for test. Furthermore, all the dialogues in the test set and 1,672 out of 3,700 dialogues in the training set have been translated into English to form an English version of the dataset. The annotations are obtained on the Chinese collection and the English version reused the same annotations.

4.1 Dialogue Mining

The 3,700 Chinese Helpdesk dialogues contained in the DCH-1 test collection were mined from Weibo in April 2018 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[2]. We then filtered out accounts that did not respond to over one half of these threads. As a result, we obtained 41 active account names. We denote the filtered set of "active" accounts as A.

[2] 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.

Table 4. Statistics of English Data collection.

	Training Set	Test Set
Source	`www.weibo.com`	
Language	English	
Data timestamps	Jan. 2013 - Apr. 2018	
#annotators/dialogue	19	
#Dialogues	1,672	390
Avg. #posts/dialogue	4.522	4.877
Avg. post length (#Tokens)	31.986	30.890
Avg. turn length	34.964	33.478
Quality annotation criteria	A-score, E-score, S-score (See Section 2.2)	
Nugget types	CNUG0, CNUG, HNUG, CNUG∗, HNUG∗ (See Section 2.3)	

3. For each account a in A, we retrieved all threads that contain a mention of a from January 2013 to Apr 2018, and extracted Customer-Helpdesk dyadic dialogues from them. We then kept those that consist of at least one turn by Customer *and* one by Helpdesk. As a result, 21,669 dialogues were obtained. This collection is denoted as D_0. Note that although we used account names in A as seeds for searching the dialogue corpus, we obtained dialogues involving not only these accounts but also *subaccounts* of these accounts. For example, when the customer mentions "ABCD Company Helpdesk," a subaccount called "ABCD Company Helpdesk Beijing" might actually respond to it. Such dialogues are also included in DCH-1; thus it actually covers helpdesk accounts that are outside A.
4. As D_0 is too large for annotation, we sampled 3,700 dialogues from it to build a training set as follows. For $i = 2, 3, \ldots, 6$, we randomly sampled 700 dialogues that contained i turns. In addition, we randomly sampled 200 that contained $i = 7$ turns; we could not sample 700 dialogues for $i = 7$ as D_0 did not contain enough dialogues that are very long.
5. We sampled 390 dialogues from it to build a test set. For $i = 2, 3, \ldots, 7$, we randomly sampled 65 dialogues that contained i turns.

To build the English training set, 1,672 of the Chinese Dialogues in the training set and all the dialogues in the test set were manually translated into English by a professional translation company.

4.2 Annotations

We hired 19 Chinese undergraduate students from the Faculty of Science and Engineering at Waseda University so that each Chinese dialogue was annotated

independently by each of the annotators. The English dialogues were not annotated as the same annotations are shared by both Chinese and English collections. 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 three dialogue quality annotation criteria described in Section 2.2; finally, he/she identified nuggets within the same dialogue, where nuggets were defined as described in Section 2.3. An initial face-to-face instruction and training session for the annotators was organised by the first author of this paper at Waseda University; subsequently, the annotators were allowed to do their annotation work online using a web-browser-based tool at their convenient location and time. All of them completed their work within two months as they were initially asked to do. The actual annotation time spent by each annotator was 72 hours. Each annotator was paid 1,200 Japanese Yen per hour.

5 Realtime Score Feedback

During the development period, we ran a server that provides realtime evaluation score feedback to participants. When a participant submits a run to our server, it returns a set of evaluation scores [3]. (Rescaling by $-\log_2$ was applied to the aforementioned measures so that larger scores meant higher performances.) Figure 3 shows an example of a response from our server. To prevent overfitting with test data, we used only 50% of the official test data for the score calculations, and a total of only 50 run submissions per team were allowed.

```
» »»» python submit.py --team_name "My Team" -s submission.json  --language English
{   'nugget': {'jsd': 0.025161759001242415, 'rnss': 0.0965867419061319 1},
    'quality': {   'nmd': {   'A': 0.08753139080222316,
                              'E': 0.08508171717323144,
                              'S': 0.08081326324593974},
                   'rsnod': {   'A': 0.1335972142240669,
                                'E': 0.1261157386588752,
                                'S': 0.13048224485651394}}}
```

Fig. 3. An example of a response from our server

6 Runs

6.1 Baseline models

We evaluated the submitted runs in Table 2 and the following baseline models. The organisers prepared three baseline models for each language and each subtask as follows;

[3] `https://sakai-lab.github.io/stc3-dataset/evaluation`

BL-lstm A baseline model [4] which leverages Bidirectional Long Short-term Memory [4, 11],

BL-uniform A baseline model which always predict the uniform distribution.

BL-popularity A baseline model which predicts the probability of the other labels except the most popular label as 0.

Note that the BL-popularity accesses the gold data labeled by multiple annotators.

6.2 Participants' runs

Four teams participated in STC3 DQ and ND subtasks: SLSTC [5], WIDM [1], WUST [14], and CUIS [2]. Here, we briefly summarise participants' approach. Since most participants implemented their models by modifying the BL-lstm baseline model, we focus on the difference between their methods and the baseline model.

SLSTC They made a few changes to the baseline BiLSTM model, and submitted three models, including (1) BiLSTM with multi-head attention [12], which utilises Transformer to encode the dialogue; (2) multi-task learning, which is a single model trained to predict both ND and DQ labels; (3) BiLSTM with BERT, which replaces the embedding layer of the baseline model with BERT [3].

WIDM Instead of Bag of Words (BoW), they utilised hierarchical CNN to encode sentences. For conversation representation, they submitted two runs which used LSTM and CNN respectively. Moreover]er, they proposed to add memory layer on gated CNN to obtain longer-term dependencies from long dialogues.

WUST To selectively obtain more context features, they adopted attention mechanism by inserting an extra attention layer into the LSTM baseline model before the output layer.

CUIS Different from other models which obtain turn-level representation directly from tokens, CUIS utilised BERT to obtain sentence-level representation first, and then apply Hierarchical Attention Networks (HAN) to obtain turn-level representation and conversation-level respectively.

7 Results

7.1 Chinese Results

Tables 5 to 7 show the mean evaluation scores for the DQ subtask in terms of A-score, S-score, E-score, respectively and Table 8 shows the mean evaluation

[4] `https://github.com/sakai-lab/stc3-baseline`

scores for the ND subtask. We conducted randomised Tukey HSD tests using the Discpower tool[5] with $B = 5,000$ trials [7] and Tables 15 to 22 summarises the statistical significance test results. Randomised Tukey HSD p-values and effect sizes (i.e., standardised mean differences) based on one-way ANOVA (without replication) [10] are also shown.

From the official Chinese results with the evaluation measures for the DQ and ND subtasks, it can be observed that:

- BL-lstm and all submitted runs are not statistically significantly different from each other for the DQ and ND subtasks;
- BL-lstm and all submitted runs outperform BL-popularity and BL-uniform significantly for the DQ and ND subtasks.

In Table 9, we compare the system rankings according to the two evaluation measures of each subtask in terms of Kendall's τ, as well as their 95% confidence intervals[6]. It can be observed that the JSD and RNSS rankings are statistically indistinguishable.

Table 5. Chinese Dialogue Quality (A-score) Results

Run	Mean RSNOD	Run	Mean NMD
SLSTC-run1	0.1235	SLSTC-run1	0.0819
SLSTC-run2	0.1249	SLSTC-run0	0.0831
WUST-run0	0.1251	WUST-run0	0.0836
WUST-run2	0.1263	SLSTC-run2	0.0843
BL-lstm	0.1263	WUST-run2	0.0845
CUIS-run0	0.1273	CUIS-run0	0.0859
WUST-run1	0.1274	WUST-run1	0.0860
SLSTC-run0	0.1306	BL-lstm	0.0863
BL-uniform	0.2478	BL-popularity	0.1677
BL-popularity	0.2532	BL-uniform	0.1855

[5] `http://research.nii.ac.jp/ntcir/tools/discpower-en.html`

[6] We calculate the confidence intervals using *kendall.ci* function of the NSM3 package (`https://www.rdocumentation.org/packages/NSM3/`) with the following options; alpha=0.05, bootstrap=T, B=10000.

Table 6. Chinese Dialogue Quality (S-score) Results

Run	Mean RSNOD	Run	Mean NMD
SLSTC-run2	0.1175	SLSTC-run2	0.0731
WUST-run0	0.1226	SLSTC-run1	0.0772
SLSTC-run1	0.1243	WUST-run0	0.0779
BL-lstm	0.1245	WUST-run2	0.0779
WUST-run2	0.1248	SLSTC-run0	0.0787
WUST-run1	0.1270	BL-lstm	0.0800
CUIS-run0	0.1281	WUST-run1	0.0808
SLSTC-run0	0.1290	CUIS-run0	0.0817
BL-popularity	0.2326	BL-popularity	0.1499
BL-uniform	0.2681	BL-uniform	0.1987

Table 7. Chinese Dialogue Quality (E-score) Results

Run	Mean RSNOD	Run	Mean NMD
SLSTC-run1	0.1159	SLSTC-run1	0.0754
WUST-run2	0.1167	WUST-run2	0.0774
SLSTC-run2	0.1178	SLSTC-run2	0.0779
BL-lstm	0.1182	WUST-run0	0.0780
CUIS-run0	0.1196	SLSTC-run0	0.0790
WUST-run0	0.1200	BL-lstm	0.0794
WUST-run1	0.1236	CUIS-run0	0.0795
SLSTC-run0	0.1238	WUST-run1	0.0828
BL-uniform	0.2162	BL-uniform	0.1580
BL-popularity	0.2774	BL-popularity	0.1950

Table 8. Chinese Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
SLSTC-run2	0.0217	SLSTC-run2	0.0876
BL-lstm	0.0220	BL-lstm	0.0899
WUST-run0	0.0223	WUST-run0	0.0909
SLSTC-run1	0.0225	SLSTC-run1	0.0913
WUST-run1	0.0233	WUST-run1	0.0931
SLSTC-run0	0.0241	SLSTC-run0	0.0946
WUST-run2	0.0250	WUST-run2	0.0980
BL-popularity	0.1665	BL-popularity	0.2653
BL-uniform	0.2304	BL-uniform	0.3708

Table 9. Kendall's τ with 95% CIs (Chinese)

Dialogue Quality (A-score)		
NMD vs RSNOD	0.556	[0.000, 0.951]
Dialogue Quality (S-score)		
NMD vs RSNOD	0.778	[0.400, 1.000]
Dialogue Quality (E-score)		
NMD vs RSNOD	0.778	[0.351, 1.000]
Nugget Detection		
JSD vs RNSS	1.000	[1.000, 1.000]

7.2 English Results

Tables 10 to 12 show the mean evaluation scores for the DQ subtask in terms of A-score, S-score, E-score, respectively and Table 13 shows the mean evaluation scores for the ND subtask. We conducted randomised Tukey HSD tests as is the case in the Chinese results and Tables 23 to 30 summarises the statistical significance test results. Randomised Tukey HSD p-values and effect sizes (i.e., standardised mean differences) based on one-way ANOVA (without replication) [10] are also shown.

From the official English results with the evaluation measures for the DQ and ND subtasks, it can be observed that:

- BL-lstm and all submitted runs except SLSTC-run0 are not statistically significantly different from each other for the DQ subtask;
- BL-lstm and all submitted runs except SLSTC-run0 outperform BL-popularity and BL-uniform significantly for the DQ subtask;
- BL-lstm and all submitted runs are not statistically significantly different from each other for the ND subtask;
- BL-lstm and all submitted runs outperform BL-popularity and BL-uniform significantly for the ND subtask.

In Table 9, we compare the system rankings according to the two evaluation measures of each subtask in terms of Kendall's , as well as their 95% confidence intervals. It can be observed that the JSD and RNSS rankings are statistically indistinguishable.

Table 10. English Dialogue Quality (A-score) Results

Run	Mean RSNOD	Run	Mean NMD
BL-lstm	0.1320	BL-lstm	0.0896
CUIS-run0	0.1360	CUIS-run0	0.0901
SLSTC-run2	0.1370	SLSTC-run1	0.0908
SLSTC-run1	0.1391	SLSTC-run2	0.0933
WIDM-run0	0.1411	WIDM-run0	0.0939
WIDM-run1	0.1411	WIDM-run1	0.0939
SLSTC-run0	0.1493	SLSTC-run0	0.1017
BL-uniform	0.2478	BL-popularity	0.1677
BL-popularity	0.2532	BL-uniform	0.1855

Table 11. English Dialogue Quality (S-score) Results

Run	Mean RSNOD	Run	Mean NMD
SLSTC-run2	0.1306	SLSTC-run1	0.0820
BL-lstm	0.1310	SLSTC-run2	0.0822
WIDM-run1	0.1318	BL-lstm	0.0838
SLSTC-run1	0.1340	CUIS-run0	0.0842
WIDM-run0	0.1344	WIDM-run0	0.0855
CUIS-run0	0.1353	WIDM-run1	0.0857
SLSTC-run0	0.1423	SLSTC-run0	0.0907
BL-popularity	0.2326	BL-popularity	0.1499
BL-uniform	0.2681	BL-uniform	0.1987

Table 12. English Dialogue Quality (E-score) Results

Run	Mean RSNOD	Run	Mean NMD
SLSTC-run2	0.1219	BL-lstm	0.0824
BL-lstm	0.1220	SLSTC-run2	0.0828
WIDM-run0	0.1274	CUIS-run0	0.0854
WIDM-run1	0.1274	SLSTC-run1	0.0859
CUIS-run0	0.1283	WIDM-run0	0.0875
SLSTC-run1	0.1321	WIDM-run1	0.0875
SLSTC-run0	0.1404	SLSTC-run0	0.0938
BL-uniform	0.2162	BL-uniform	0.1580
BL-popularity	0.2774	BL-popularity	0.1950

Table 13. English Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
BL-lstm	0.0248	BL-lstm	0.0952
SLSTC-run1	0.0252	SLSTC-run1	0.0973
SLSTC-run2	0.0263	SLSTC-run2	0.0979
WIDM-run0	0.0265	WIDM-run0	0.0990
WIDM-run1	0.0265	WIDM-run1	0.0990
SLSTC-run0	0.0289	SLSTC-run0	0.1037
BL-popularity	0.1665	BL-popularity	0.2653
BL-uniform	0.2304	BL-uniform	0.3708

Table 14. Kendall's with 95% CIs (English)

Dialogue Quality (A-score)		
NMD vs RSNOD	0.886	[0.613, 1.000]
Dialogue Quality (S-score)		
NMD vs RSNOD	0.667	[0.125, 1.000]
Dialogue Quality (E-score)		
NMD vs RSNOD	0.714	[0.200, 1.000]
Nugget Detection		
JSD vs RNSS	1.000	[1.000, 1.000]

8 Conclusions

The official results of the DQ and ND subtasks can be summarised as follows.

- In the Chinese DQ subtask, runs from CUIS, runs from SLSTC, runs from WUST and BL-lstm statistically significantly outperformed the BL-popularity and BL-uniform baselines in terms of both NMD and RSNOD for A-score, S-score and E-score, while BL-popularity statistically significantly outperformed BL-uniform for S-score and E-score. In terms of only NMD, BL-popularity statistically significantly outperformed BL-uniform for A-score.
- In the Chinese ND subtask, runs from SLSTC, runs from WUST and BL-lstm statistically significantly outperformed the BL-popularity and BL-uniform baselines in terms of both JSD and RNSS, while BL-popularity statistically significantly outperformed BL-uniform.
- In the English DQ subtask, runs from CUIS, runs from SLSTC, runs from WIDM and BL-lstm statistically significantly outperformed the BL-popularity and BL-uniform baselines in terms of both NMD and RSNOD for A-score, S-score and E-score, while BL-popularity statistically significantly outperformed BL-uniform for S-score and E-score. In terms of only NMD, BL-popularity statistically significantly outperformed BL-uniform for A-score. In terms of only RSNOD, BL-lstm statistically significantly outperformed SLSTC-run0 for A-score, SLSTC-run2 also statistically significantly outperformed SLSTC-run0 for E-score.
- In the English ND subtask, runs from SLSTC, runs from WIDM and BL-lstm statistically significantly outperformed the BL-popularity and BL-uniform baselines in terms of both JSD and RNSS, while BL-popularity statistically significantly outperformed BL-uniform.

Further discussions of the NTCIR-14 STC-3 Dialogue Quality and Nugget Detection subtasks will be given in our Final Report [15].

Acknowledgements

We thank the STC-3 participants and the NTCIR chairs for making this task happen. This work was supported by JSPS KAKENHI Gran Number 17H01830.

References

1. Cherng, H.E., Chang, C.H.: Dialogue quality and nugget detection for short text conversation (STC-3) based on hierarchical multi-stack model with memory enhance structure. In: NTCIR14. p. to appear (2019)
2. Cong, K., Lam, W.: CUIS at the ntcir-14 STC-3 DQ subtask. In: NTCIR14. p. to appear (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
5. Kato, S., Suzuki, R., Zeng, Z., Sakai, T.: SLSTC at the NTCIR-14 STC-3 dialogue quality and nugget detection subtasks. In: NTCIR14. p. to appear (2019)
6. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International journal of computer vision **40**(2), 99–121 (2000)
7. Sakai, T.: Metrics, statistics, tests. In: PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173). pp. 116–163 (2014)
8. Sakai, T.: Towards automatic evaluation of multi-turn dialogues: A task design that leverages inherently subjective annotations. In: Proceedings of EVIA 2017 (2017)
9. Sakai, T.: Comparing two binned probability distributions for information access evaluation. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1073–1076. SIGIR '18, ACM, New York, NY, USA (2018)
10. Sakai, T.: Laboratory experiments in information retrieval: Sample sizes, effect sizes, and statistical power. Springer (2018), https://link.springer.com/book/10.1007/978-981-13-1199-4
11. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Processing **45**(11), 2673–2681 (1997)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
13. Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A.: PARADISE: A framework for evaluating spoken dialogue agents. In: Proceedings of ACL 1997. pp. 271–280 (1997)
14. Yan, M., Liu, M., Xiang, J.: WUST at STC-3 dialogue quality and nugget detection subtask in NTCIR-14. In: NTCIR14. p. to appear (2019)
15. Zeng, Z., Kato, S., Sakai, T.: Final report of the NTCIR-14 short text conversation dialogue quality and nugget detection subtasks. In: LNCS. p. to appear (2019)
16. Zeng, Z., Luo, C., Shang, L., Li, H., Sakai, T.: Towards automatic evaluation of customer-helpdesk dialogues. Journal of Information Processing **26**, 768–778 (2018), https://www.jstage.jst.go.jp/article/ipsjjip/26/0/26_768/_pdf/-char/en
17. Zhang, Y., Huang, M.: Overview of the ntcir-14 short text generation subtask: Emotion generation challenge (2019)

A Appendix

Table 15. Statistical significance in terms of NMD (Chinese DQ subtask, A-score)

These runs are	significantly better than these runs	
SLSTC-run1	BL-popularity	($p < 0.0001, ES_{E1} = 1.362$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.646$)
SLSTC-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.343$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.627$)
WUST-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.335$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.620$)
SLSTC-run2	BL-popularity	($p < 0.0001, ES_{E1} = 1.324$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.609$)
WUST-run2	BL-popularity	($p < 0.0001, ES_{E1} = 1.320$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.605$)
CUIS-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.299$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.584$)
WUST-run1	BL-popularity	($p < 0.0001, ES_{E1} = 1.297$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.582$)
BL-lstm	BL-popularity	($p < 0.0001, ES_{E1} = 1.292$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.577$)
BL-popularity	BL-uniform	($p = 0.0008, ES_{E1} = 0.284$)

Table 16. Statistical significance in terms of RSNOD (Chinese DQ subtask, A-score)

These runs are	significantly better than these runs	
SLSTC-run1	BL-uniform	($p < 0.0001, ES_{E1} = 1.709$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.784$)
SLSTC-run2	BL-uniform	($p < 0.0001, ES_{E1} = 1.691$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.766$)
WUST-run0	BL-uniform	($p < 0.0001, ES_{E1} = 1.687$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.762$)
WUST-run2	BL-uniform	($p < 0.0001, ES_{E1} = 1.671$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.746$)
BL-lstm	BL-uniform	($p < 0.0001, ES_{E1} = 1.671$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.746$)
CUIS-run0	BL-uniform	($p < 0.0001, ES_{E1} = 1.657$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.732$)
WUST-run1	BL-uniform	($p < 0.0001, ES_{E1} = 1.656$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.731$)
SLSTC-run0	BL-uniform	($p < 0.0001, ES_{E1} = 1.612$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.687$)

Table 17. Statistical significance in terms of NMD (Chinese DQ subtask, S-score)

These runs are	significantly better than these runs	
SLSTC-run2	BL-popularity	$(p < 0.0001, ES_{E1} = 1.385)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.266)$
SLSTC-run1	BL-popularity	$(p < 0.0001, ES_{E1} = 1.310)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.191)$
WUST-run0	BL-popularity	$(p < 0.0001, ES_{E1} = 1.298)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.179)$
WUST-run2	BL-popularity	$(p < 0.0001, ES_{E1} = 1.298)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.179)$
SLSTC-run0	BL-popularity	$(p < 0.0001, ES_{E1} = 1.283)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.164)$
BL-lstm	BL-popularity	$(p < 0.0001, ES_{E1} = 1.261)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.142)$
WUST-run1	BL-popularity	$(p < 0.0001, ES_{E1} = 1.245)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.126)$
CUIS-run0	BL-popularity	$(p < 0.0001, ES_{E1} = 1.230)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.111)$
BL-popularity	BL-uniform	$(p < 0.0001, ES_{E1} = 0.881)$

Table 18. Statistical significance in terms of RSNOD (Chinese DQ subtask, S-score)

These runs are	significantly better than these runs	
SLSTC-run2	BL-popularity	$(p < 0.0001, ES_{E1} = 1.660)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.170)$
WUST-run0	BL-popularity	$(p < 0.0001, ES_{E1} = 1.586)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.096)$
SLSTC-run1	BL-popularity	$(p < 0.0001, ES_{E1} = 1.561)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.071)$
BL-lstm	BL-popularity	$(p < 0.0001, ES_{E1} = 1.558)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.069)$
WUST-run2	BL-popularity	$(p < 0.0001, ES_{E1} = 1.554)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.065)$
WUST-run1	BL-popularity	$(p < 0.0001, ES_{E1} = 1.522)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.033)$
CUIS-run0	BL-popularity	$(p < 0.0001, ES_{E1} = 1.506)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.017)$
SLSTC-run0	BL-popularity	$(p < 0.0001, ES_{E1} = 1.493)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 2.004)$
BL-popularity	BL-uniform	$(p < 0.0001, ES_{E1} = 0.511)$

Table 19. Statistical significance in terms of NMD (Chinese DQ subtask, E-score)

These runs are	significantly better than these runs	
SLSTC-run1	BL-uniform	$(p < 0.0001, ES_{E1} = 1.568)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.268)$
WUST-run2	BL-uniform	$(p < 0.0001, ES_{E1} = 1.530)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.230)$
SLSTC-run2	BL-uniform	$(p < 0.0001, ES_{E1} = 1.520)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.220)$
WUST-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.518)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.219)$
SLSTC-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.499)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.200)$
BL-lstm	BL-uniform	$(p < 0.0001, ES_{E1} = 1.491)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.192)$
CUIS-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.490)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.191)$
WUST-run1	BL-uniform	$(p < 0.0001, ES_{E1} = 1.427)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.127)$
BL-uniform	BL-popularity	$(p < 0.0001, ES_{E1} = 0.701)$

Table 20. Statistical significance in terms of RSNOD (Chinese DQ subtask, E-score)

These runs are	significantly better than these runs	
SLSTC-run1	BL-uniform	$(p < 0.0001, ES_{E1} = 1.698)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.732)$
WUST-run2	BL-uniform	$(p < 0.0001, ES_{E1} = 1.684)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.718)$
SLSTC-run2	BL-uniform	$(p < 0.0001, ES_{E1} = 1.665)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.700)$
BL-lstm	BL-uniform	$(p < 0.0001, ES_{E1} = 1.658)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.692)$
CUIS-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.635)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.669)$
WUST-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.627)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.662)$
WUST-run1	BL-uniform	$(p < 0.0001, ES_{E1} = 1.566)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.601)$
SLSTC-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.564)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.598)$
BL-uniform	BL-popularity	$(p < 0.0001, ES_{E1} = 1.035)$

Table 21. Statistical significance in terms of JSD (the Chinese ND subtask)

These runs are	significantly better than these runs	
SLSTC-run2	BL-popularity	($p < 0.0001, ES_{E1} = 5.254$)
	BL-uniform	($p < 0.0001, ES_{E1} = 7.573$)
BL-lstm	BL-popularity	($p < 0.0001, ES_{E1} = 5.245$)
	BL-uniform	($p < 0.0001, ES_{E1} = 7.563$)
WUST-run0	BL-popularity	($p < 0.0001, ES_{E1} = 5.232$)
	BL-uniform	($p < 0.0001, ES_{E1} = 7.550$)
SLSTC-run1	BL-popularity	($p < 0.0001, ES_{E1} = 5.224$)
	BL-uniform	($p < 0.0001, ES_{E1} = 7.542$)
WUST-run1	BL-popularity	($p < 0.0001, ES_{E1} = 5.196$)
	BL-uniform	($p < 0.0001, ES_{E1} = 7.515$)
SLSTC-run0	BL-popularity	($p < 0.0001, ES_{E1} = 5.168$)
	BL-uniform	($p < 0.0001, ES_{E1} = 7.486$)
WUST-run2	BL-popularity	($p < 0.0001, ES_{E1} = 5.134$)
	BL-uniform	($p < 0.0001, ES_{E1} = 7.453$)
BL-popularity	BL-uniform	($p < 0.0001, ES_{E1} = 2.319$)

Table 22. Statistical significance in terms of RNSS (the Chinese ND subtask)

These runs are	significantly better than these runs	
SLSTC-run2	BL-popularity	($p < 0.0001, ES_{E1} = 3.494$)
	BL-uniform	($p < 0.0001, ES_{E1} = 5.568$)
BL-lstm	BL-popularity	($p < 0.0001, ES_{E1} = 3.448$)
	BL-uniform	($p < 0.0001, ES_{E1} = 5.522$)
WUST-run0	BL-popularity	($p < 0.0001, ES_{E1} = 3.428$)
	BL-uniform	($p < 0.0001, ES_{E1} = 5.502$)
SLSTC-run1	BL-popularity	($p < 0.0001, ES_{E1} = 3.422$)
	BL-uniform	($p < 0.0001, ES_{E1} = 5.496$)
WUST-run1	BL-popularity	($p < 0.0001, ES_{E1} = 3.386$)
	BL-uniform	($p < 0.0001, ES_{E1} = 5.460$)
SLSTC-run0	BL-popularity	($p < 0.0001, ES_{E1} = 3.357$)
	BL-uniform	($p < 0.0001, ES_{E1} = 5.431$)
WUST-run2	BL-popularity	($p < 0.0001, ES_{E1} = 3.290$)
	BL-uniform	($p < 0.0001, ES_{E1} = 5.364$)
BL-popularity	BL-uniform	($p < 0.0001, ES_{E1} = 2.074$)

Table 23. Statistical significance in terms of NMD (English DQ subtask, A-score)

These runs are	significantly better than these runs	
BL-lstm	BL-popularity	($p < 0.0001, ES_{E1} = 1.174$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.443$)
CUIS-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.167$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.436$)
SLSTC-run1	BL-popularity	($p < 0.0001, ES_{E1} = 1.156$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.425$)
SLSTC-run2	BL-popularity	($p < 0.0001, ES_{E1} = 1.119$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.388$)
WIDM-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.110$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.380$)
WIDM-run1	BL-popularity	($p < 0.0001, ES_{E1} = 1.110$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.380$)
SLSTC-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.992$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.261$)
BL-popularity	BL-uniform	($p = 0.0034, ES_{E1} = 0.269$)

Table 24. Statistical significance in terms of RSNOD (English DQ subtask, A-score)

These runs are	significantly better than these runs	
BL-lstm	SLSTC-run0	($p = 0.0428, ES_{E1} = 0.228$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.532$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.604$)
CUIS-run0	BL-uniform	($p < 0.0001, ES_{E1} = 1.480$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.552$)
SLSTC-run2	BL-uniform	($p < 0.0001, ES_{E1} = 1.466$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.538$)
SLSTC-run1	BL-uniform	($p < 0.0001, ES_{E1} = 1.438$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.510$)
WIDM-run0	BL-uniform	($p < 0.0001, ES_{E1} = 1.412$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.484$)
WIDM-run1	BL-uniform	($p < 0.0001, ES_{E1} = 1.412$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.484$)
SLSTC-run0	BL-uniform	($p < 0.0001, ES_{E1} = 1.304$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.376$)

Table 25. Statistical significance in terms of NMD (English DQ subtask, S-score)

These runs are	significantly better than these runs	
SLSTC-run1	BL-popularity	($p < 0.0001, ES_{E1} = 1.166$)
	BL-uniform	($p < 0.0001, ES_{E1} = 2.005$)
SLSTC-run2	BL-popularity	($p < 0.0001, ES_{E1} = 1.163$)
	BL-uniform	($p < 0.0001, ES_{E1} = 2.002$)
BL-lstm	BL-popularity	($p < 0.0001, ES_{E1} = 1.136$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.975$)
CUIS-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.128$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.967$)
WIDM-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.106$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.945$)
WIDM-run1	BL-popularity	($p < 0.0001, ES_{E1} = 1.102$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.941$)
SLSTC-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.017$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.856$)
BL-popularity	BL-uniform	($p < 0.0001, ES_{E1} = 0.839$)

Table 26. Statistical significance in terms of RSNOD (English DQ subtask, S-score)

These runs are	significantly better than these runs	
SLSTC-run2	BL-popularity	($p < 0.0001, ES_{E1} = 1.395$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.880$)
BL-lstm	BL-popularity	($p < 0.0001, ES_{E1} = 1.391$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.876$)
WIDM-run1	BL-popularity	($p < 0.0001, ES_{E1} = 1.379$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.864$)
SLSTC-run1	BL-popularity	($p < 0.0001, ES_{E1} = 1.349$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.834$)
WIDM-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.344$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.829$)
CUIS-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.332$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.816$)
SLSTC-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.236$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.720$)
BL-popularity	BL-uniform	($p < 0.0001, ES_{E1} = 0.485$)

Table 27. Statistical significance in terms of NMD (English DQ subtask, E-score)

These runs are	significantly better than these runs	
BL-lstm	BL-uniform	$(p < 0.0001, ES_{E1} = 1.327)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 1.975)$
SLSTC-run2	BL-uniform	$(p < 0.0001, ES_{E1} = 1.321)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 1.969)$
CUIS-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.275)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 1.923)$
SLSTC-run1	BL-uniform	$(p < 0.0001, ES_{E1} = 1.265)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 1.913)$
WIDM-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.237)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 1.885)$
WIDM-run1	BL-uniform	$(p < 0.0001, ES_{E1} = 1.237)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 1.885)$
SLSTC-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.127)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 1.774)$
BL-uniform	BL-popularity	$(p < 0.0001, ES_{E1} = 0.648)$

Table 28. Statistical significance in terms of RSNOD (English DQ subtask, E-score)

These runs are	significantly better than these runs	
SLSTC-run2	SLSTC-run0	$(p = 0.0120, ES_{E1} = 0.295)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 1.504)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.479)$
BL-lstm	SLSTC-run0	$(p = 0.0124, ES_{E1} = 0.293)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 1.502)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.477)$
WIDM-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.416)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.391)$
WIDM-run1	BL-uniform	$(p < 0.0001, ES_{E1} = 1.416)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.391)$
CUIS-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.403)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.378)$
SLSTC-run1	BL-uniform	$(p < 0.0001, ES_{E1} = 1.342)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.316)$
SLSTC-run0	BL-uniform	$(p < 0.0001, ES_{E1} = 1.210)$
	BL-popularity	$(p < 0.0001, ES_{E1} = 2.184)$
BL-uniform	BL-popularity	$(p < 0.0001, ES_{E1} = 0.975)$

Table 29. Statistical significance in terms of JSD (English ND subtask)

These runs are	significantly better than these runs	
BL-lstm	BL-popularity	$(p < 0.0001, ES_{E1} = 4.649)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 6.745)$
SLSTC-run1	BL-popularity	$(p < 0.0001, ES_{E1} = 4.636)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 6.732)$
SLSTC-run2	BL-popularity	$(p < 0.0001, ES_{E1} = 4.598)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 6.694)$
WIDM-run0	BL-popularity	$(p < 0.0001, ES_{E1} = 4.593)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 6.690)$
WIDM-run1	BL-popularity	$(p < 0.0001, ES_{E1} = 4.593)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 6.690)$
SLSTC-run0	BL-popularity	$(p < 0.0001, ES_{E1} = 4.512)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 6.608)$
BL-popularity	BL-uniform	$(p < 0.0001, ES_{E1} = 2.096)$

Table 30. Statistical significance in terms of RNSS (English ND subtask)

These runs are	significantly better than these runs	
BL-lstm	BL-popularity	$(p < 0.0001, ES_{E1} = 3.042)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 4.929)$
SLSTC-run1	BL-popularity	$(p < 0.0001, ES_{E1} = 3.005)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 4.892)$
SLSTC-run2	BL-popularity	$(p < 0.0001, ES_{E1} = 2.994)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 4.881)$
WIDM-run0	BL-popularity	$(p < 0.0001, ES_{E1} = 2.974)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 4.861)$
WIDM-run1	BL-popularity	$(p < 0.0001, ES_{E1} = 2.974)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 4.861)$
SLSTC-run0	BL-popularity	$(p < 0.0001, ES_{E1} = 2.889)$
	BL-uniform	$(p < 0.0001, ES_{E1} = 4.776)$
BL-popularity	BL-uniform	$(p < 0.0001, ES_{E1} = 1.887)$