Overview of the NTCIR-16 Dialogue Evaluation (DialEval-2) Task

Sijie Tao
Waseda University, Japan
tsjmailbox@ruri.waseda.jp

Tetsuya Sakai
Waseda University, Japan
tetsuyasakai@acm.org

ABSTRACT

This paper provides an overview of the NTCIR-16 Dialogue Evaluation (DialEval-2) task. DialEval-2 is the successor of the NTCIR-15 DialEval-1 task and the NTCIR-14 Short Text Conversation STC-3 task. DialEval-2 consists of two subtasks: the Dialogue Quality (DQ) subtask and the Nugget Detection (ND) subtask. Both of the subtasks are designed to aim automatical evaluation of customer-helpdesk dialogues. The DQ subtask requires our participants to estimate three kinds of quality score for each dialogue: task accomplishment, customer satisfaction, and dialogue effectiveness. The ND subtask is set as a classification task, where participants are asked to classify every turn of a dialogue to detect nugget turns. A nugget stands for a turn being helpful for problem solving in the dialogue. In this paper, we introduce the task definition, data collection, evaluation measures, and the official evaluation results on the runs from the participant teams.

1 INTRODUCTION

The DialEval-1 task was lauched at NTCIR-15. The motivation of the task is to explore ideas of automatically evaluating customer-helpdesk dialogues. Nowadays, manual evaluation by human annotators can be expensive and time assuming. If we can automatically evaluate the quality of a dialogue and find out which turns are helpful, it will be more efficient and economical than evaluating by human annotators. Therefore, we continue this task at NTCIR-16. The important dates of DialEval-2 are shown in Table 1.

DialEval-2 is the succesor of DialEval-1 at NTCIR-15 in 2020 [16]. We continue running DQ and ND subtasks. At DialEval-2, we use the DCH-2 dataset [17]. The part of this dataset was first provided at DialEval-1. With the translation of the Chinese collection being fully completed, the DCH-2 dataset was released as a Chinese-English parallel corpus. The DCH-2 dataset is used as training and development set, and a new test set is built for DialEval-2. There are four teams participating in DialEval-2: IMNTPU [4], NKUST [1], RSLDE [5], and TUA1 [2]. The statistics of participant runs in each subtask are shown in Table 2.

Table 1: Schedule of DialEval-2 at NTCIR-16

Time	Content
Dec 1 2021	Test data released
Jan 15 2022	Run submissions due
Feb 1 2022	Evaluation Results and draft overview released
March 1 2022	Draft participant paper submissions due
May 1 2022	All camera-ready paper submissions due
June 14-17 2022	NTCIR-16 Conference

Table 2: The Statistics of Participant Runs in Each Subtask.

Teams	Runs	Chinese		English	
		DQ	ND	DQ	ND
IMNTPU	1	1	0	1	1
NKUST	2	1	2	0	1
RSLDE	3	2	3	2	3
TUA1	3	3	2	1	1
Total	9	7	7	4	6

This paper is organised as follows. Section 2 and Section 3 describe task definition and evaluation methods, respectively. Section 4 introduces the data collection. Section 5 presents the official evaluation results. Finally, Section 6 concludes this paper.

2 TASK DEFINITION

The task definition of DialEval-2 is identical to that of DialEval-1. Hence this section is largely a duplicate of the task definition section of the DialEval-1 overview paper [16].

The goal of DialEval-2 is to explore approaches to evaluating task-oriented, multi-round, textual helpdesk-customer dialogue systems automatically. Identical to DialEval-1, there are 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 dialogue effectiveness; 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 Customer post is denoted by p_C and

Figure 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 [16].

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 a 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;

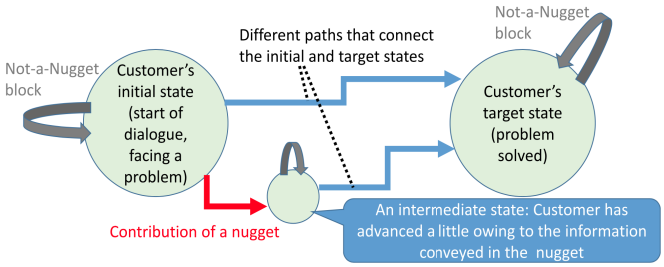

Figure 2: Task accomplishment as state transitions, and the role of a nugget [16].

(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.

2.4 Chinese and English Subtasks

The dialogues are originally in Chinese, and part of them are manually translated into English for DialEval-1. Thanks to the release of DCH-2 corpus, every Chinese dialogue has an English translation at DialEval-2. Thus, each subtask has a Chinese version and an English version, and the participants can choose the language to build their runs.

2.5 Baselines

The baseline models are exactly the same as the baselines at DialEval-1. There are three baseline models for each language and each subtask as follows;

BL-lstm (Baseline-run0) A baseline model[1] which leverages Bidirectional Long Short-term Memory [3, 14];

BL-uniform (Baseline-run1) A baseline model which always predict the uniform distribution;

BL-popularity (Baseline-run2) A baseline model which predicts the probability of the most popular label as one, and predicts other labels as 0. Note that the it accesses the golden truth to find the most popular label. This baseline is to show the upper bound of a single label.

3 EVALUATION METHODS

The evaluation measures of DialEval-2 are the same as those of DialEval-1. Hence this setion is largely a duplicate of the Evaluation Methods section of the DialEval-1 overview paper [16].

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]. We evaluate these subtasks by comparing the probability distribution estimated 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 the metrics 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 $p(i)(i = 1, \ldots, L)$ denote the system 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 Metrics for Dialogue Quality Subtask

Since the classes of DQ subtask are non-nominal, cross-bin metrics are more suitable than bin-by-bin metrics. As discussed by Sakai [9][12][11], bin-by-bin metrics 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 metrics: *Normalised Match Distance* (NMD) and *Root Symmetric Normalised Order-aware Divergence* (RSNOD).

3.1.1 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 are a 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:

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

3.1.2 Root Symmetric Normalised Order-aware Divergence (RSNOD). is a metric 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:

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

Let $B^* = \{i \mid p^*(i) > 0\}$, 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:

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

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

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

Finally, we define the Root Symmetric Normalised OD:

$$RSNOD(p, p^*) = \sqrt{\frac{SOD(p, p^*)}{L - 1}} \quad (6)$$

[1] https://github.com/DialEval-2/LSTM-baseline

In the DQ subtask, we use both NMD and RSNOD as metrics to evaluate participants' runs. Extensive experiments on *ordinal quantification* measures such as NMD and RSNOD have been reported in Sakai [11–13].

3.2 Evaluation Metrics for Nugget Detection Subtask

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

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

$$RNSS = \sqrt{\frac{\sum_{i \in A}(p(i) - p^*(i))^2}{2}} \tag{7}$$

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

$$JSD(p \parallel p^*) = \frac{KLD(p \parallel p_M) + KLD(p_M \parallel p^*)}{2} \tag{8}$$

$$\text{where } KLD(p_1 \parallel p_2) = \sum_{i \ s.t. \ p_1(i)>0} p_1(i) \log_2 \frac{p_1(i)}{p_2(i)} \tag{9}$$

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 combine the two evaluation scores into a single one for each dialogue. Specifically, we calculate the average metric 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. We let $\alpha = 0.5$ throughout this paper.

4 DATA COLLECTION

4.1 Training and Development Data

The statistics of the DialEval-2 data collection is shown in Table 3. We use the DCH-2 data collection [15] for training and development, as part of DCH-2 was utilised at NTCIR-15 DialEval-1 task. The DCH-2 data collection consists of *real* (i.e., human-human) customer-helpdesk dialogues collected from Weibo, and there are 4,390 Chinese-English parallel dialogues. At DialEval-2, 4,090 pairs of dialogues from DCH-2, which was also partially used as training and development set at DialEval-1, are used as training set. The other 300 pairs of dialogues which was the test set at DialEval-1, are used as develeopment set.

4.2 Test Data

In order to construct the test set, we sampled 65 dialogues from the data we have crawled from Weibo. According to *topic size design* [10] and the residual variances obtained from the previous DialEval-1 task, 62 dialogues is enough to ensure a statistical power of 80% with a *minimum detectable range* of 0.05 for ANOVA with 10 systems at the 5% significance level. We hired 20 Chinese students from the Faculty of Science and Engineering at Waseda University to annotate the dialogues. Following the same annotation instructions as DialEval-1, each dialogue is annotated by each annotator independently. Besides the annotation, all the test dialogues are manually translated into English. We hired a professional company to complete the translation. As the translation does not change the semantic information of the dialogues, the English dialogues share the same annotation with the Chinese ones. Thus, we have a Chinese-English parallel test set for DialEval-2.

5 RESULTS

5.1 Chinese Subtasks

First, in order to identify the differences between the evaluation metrics, for each subtask, all the runs are ranked by the two evaluation metrics respectively, and then we calculate the ranking correlation using Kendall's *tau* between the two rankings, as well as their 95% confidence intervals.[2] The results are shown in In Table 8. It can be observed that the difference between different metrics are not statistically significant for both ND and DQ subtaks. This finding is consistent with what we observed at DialEval-1 and STC-3.

Tables 4 to 6 shows the mean evaluation scores of the DQ subtask in terms of A-score, S-score, and E-score respectively, and Table 7 shows the mean evaluation scores of the ND subtask. We conducted randomised Tukey HSD tests using the Discpower tool [3] with $B = 5,000$ trials [7]. Tables 14 to 21 summarise the statistical significance test results, p-values, and effect sizes computed by Randomised Tukey HSD (i.e., standardised mean differences) based on one-way ANOVA (without replication) [10]. From the result shown in Table 18, it can be observed that only TUA1-run0 statistically significantly outperforms all the baseline systems in Chinese DQ task in terms of NMD for E-score. We also find that in Chinese ND task, RSLDE-run0 is the only system which can outperform Baseline-run0 in terms of both JSD and RNSS, but the difference between them is not statistically significant.

5.2 English Subtasks

Tables 9 to 11 shows the mean evaluation scores of the DQ subtask in terms of A-score, S-score, and E-score respectively, and Table 12 shows the mean evaluation scores of the ND subtask.

Following the Chinese subtasks, we also conduct randomised Tukey HSD tests for English runs, and Tables 22 to 29 summarise the significance test results, along with p-values and effect sizes. From the results of the DQ subtask, it can be observed that TUA1-run0 is the only system that outperformed all the baseline systems on average, but the differences are not statistically significant. From the results of the ND subtask, we find that RSLDE-run0 and IMNTPU-run0 outperform Baseline-run0, but the their gains are not statistically significant either. In Table 13, we also rank all the participant systems according to the two evaluation metrics of each subtasks, and then compute Kendall's *tau* of the rankings.

[2]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=10,000.
[3]http://research.nii.ac.jp/ntcir/tools/discpower-en.html

Table 3: Statistics of DialEval-2 Data collection. The unit of post/turn length is char for Chinese and token for English.

	Chinese			English		
	Training	Dev	Test	Training	Dev	Test
Source	DCH-2	DCH-2	Weibo	Translation		
Data timestamps	Jan. 2013 ∼ Apr. 2018	Apr. 2018 ∼ Jul. 2019	Apr. 2018 ∼ Jul. 2019	Jan. 2013 ∼ Apr. 2018	Apr. 2018 ∼ Jul. 2019	Apr. 2018 ∼ Jul. 2019
#dialogues	4,090	300	65	4,090	300	65
#annotators/dialogue	19	20	20	19	20	20
Quality annotation criteria	A-score, E-score, S-score (See Section 2.2)					
Nugget types	CNUG0, CNUG, HNUG, CNUG∗, HNUG∗ (See Section 2.3)					

Table 4: Chinese Dialogue Quality (A-score) Results

Run	Mean RSNOD	Run	Mean NMD
TUA1-run2	0.1992	TUA1-run2	0.1325
TUA1-run1	0.2092	TUA1-run1	0.1369
TUA1-run0	0.2154	TUA1-run0	0.1474
Baseline-run0	0.2301	RSLDE-run0	0.1537
Baseline-run2	0.2320	RSLDE-run1	0.1551
RSLDE-run0	0.2438	Baseline-run2	0.1577
RSLDE-run1	0.2446	IMNTPU-run0	0.1618
IMNTPU-run0	0.2479	Baseline-run0	0.1772
Baseline-run1	0.2767	NKUST-run0	0.2453
NKUST-run0	0.2774	Baseline-run1	0.2500

Table 5: Chinese Dialogue Quality (S-score) Results

Run	Mean RSNOD	Run	Mean NMD
TUA1-run2	0.1758	TUA1-run1	0.1159
TUA1-run1	0.1840	TUA1-run2	0.1166
TUA1-run0	0.1884	RSLDE-run1	0.1229
RSLDE-run0	0.1938	RSLDE-run0	0.1243
RSLDE-run1	0.1964	Baseline-run2	0.1288
Baseline-run0	0.1998	TUA1-run0	0.1305
IMNTPU-run0	0.2032	IMNTPU-run0	0.1315
Baseline-run2	0.2062	Baseline-run0	0.1523
NKUST-run0	0.2732	NKUST-run0	0.2293
Baseline-run1	0.2959	Baseline-run1	0.2565

The results show that the measures are highly positively correlated. This is also consistent with what we observed at previous tasks.

5.3 Top Runs

TUA1's runs are the top runs of the DQ subtask, and RSLDE and IMNTPU are the top runs of the ND subtask. The common feature of the top runs is that large scale pretrained language models play important roles in their systems. TUA1 conducted their experiments with several pre-trained Transformer models, such as BERT and RoBERTa [2]. RSLDE uses XLNet to build their top run [5], and IMNTPU chooses XLM-RoBERTa to tackle our task [4].

Table 6: Chinese Dialogue Quality (E-score) Results

Run	Mean RSNOD	Run	Mean NMD
TUA1-run0	0.1545	TUA1-run0	0.1136
TUA1-run1	0.1647	RSLDE-run0	0.1222
RSLDE-run0	0.1660	TUA1-run1	0.1262
TUA1-run2	0.1671	RSLDE-run1	0.1286
RSLDE-run1	0.1725	TUA1-run2	0.1310
Baseline-run0	0.1854	IMNTPU-run0	0.1427
IMNTPU-run0	0.1860	Baseline-run0	0.1579
NKUST-run0	0.2253	Baseline-run2	0.1710
Baseline-run1	0.2496	NKUST-run0	0.1897
Baseline-run2	0.2569	Baseline-run1	0.2106

Table 7: Chinese Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
RSLDE-run0	0.0560	RSLDE-run0	0.1604
Baseline-run0	0.0585	Baseline-run0	0.1651
RSLDE-run2	0.0607	RSLDE-run1	0.1712
RSLDE-run1	0.0634	RSLDE-run2	0.1720
NKUST-run0	0.0670	NKUST-run0	0.1761
TUA1-run0	0.0700	TUA1-run0	0.1780
Baseline-run2	0.1864	Baseline-run2	0.2901
Baseline-run1	0.2042	Baseline-run1	0.3371
NKUST-run1	0.2432	NKUST-run1	0.3774
TUA1-run1	0.2909	TUA1-run1	0.3939

6 CONCLUSION

This paper provides an overview of the NTCIR-16 DialEval-2 task. The overview describes the task definition, data collection, evaluation metrics, and the official evaluation results of DialEval-2. From the evaluation results, we observe that only one run from TUA1 outperform the LSTM baseline significantly in Chinese DQ task in terms of NMD for E-score. In other subtasks, none of the runs can outperform the LSTM baseline significantly. Moreover, no substantial difference is observed between the evaluation mectrics for each subtasks.

Table 8: Ranking Correlation between of Chinese runs ranked by two different metrics (Kendall's *tau* with 95% CIs)

Dialogue Quality (A-score)		
NMD vs RSNOD	0.689	[−0.189, 1.000]
Dialogue Quality (S-score)		
NMD vs RSNOD	0.644	[0.300, 1.000]
Dialogue Quality (E-score)		
NMD vs RSNOD	0.778	[0.538, 1.000]
Nugget Detection		
JSD vs RNSS	0.956	[0.706, 1.000]

Table 9: English Dialogue Quality (A-score) Results

Run	Mean RSNOD	Run	Mean NMD
TUA1-run0	0.1967	TUA1-run0	0.1327
Baseline-run2	0.2320	Baseline-run2	0.1577
Baseline-run0	0.2321	IMNTPU-run0	0.1654
IMNTPU-run0	0.2535	Baseline-run0	0.1780
RSLDE-run0	0.2615	RSLDE-run1	0.1896
RSLDE-run1	0.2725	RSLDE-run0	0.1957
Baseline-run1	0.2767	Baseline-run1	0.2500

Table 10: English Dialogue Quality (S-score) Results

Run	Mean RSNOD	Run	Mean NMD
TUA1-run0	0.1855	TUA1-run0	0.1214
Baseline-run0	0.1986	Baseline-run2	0.1288
IMNTPU-run0	0.2020	IMNTPU-run0	0.1312
Baseline-run2	0.2062	RSLDE-run0	0.1381
RSLDE-run0	0.2078	RSLDE-run1	0.1438
RSLDE-run1	0.2154	Baseline-run0	0.1467
Baseline-run1	0.2959	Baseline-run1	0.2565

Table 11: English Dialogue Quality (E-score) Results

Run	Mean RSNOD	Run	Mean NMD
TUA1-run0	0.1742	TUA1-run0	0.1360
Baseline-run0	0.1745	IMNTPU-run0	0.1400
IMNTPU-run0	0.1826	RSLDE-run0	0.1429
RSLDE-run0	0.1832	Baseline-run0	0.1431
RSLDE-run1	0.1889	RSLDE-run1	0.1444
Baseline-run1	0.2496	Baseline-run2	0.1710
Baseline-run2	0.2569	Baseline-run1	0.2106

ACKNOWLEDGEMENT

We thank the DialEval-2 participants and the NTCIR chairs for making this task happen.

Table 12: English Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
RSLDE-run0	0.0557	IMNTPU-run0	0.1574
IMNTPU-run0	0.0601	RSLDE-run0	0.1615
Baseline-run0	0.0625	Baseline-run0	0.1722
NKUST-run0	0.0641	NKUST-run0	0.1744
RSLDE-run2	0.0676	RSLDE-run2	0.1778
RSLDE-run1	0.0691	TUA1-run0	0.1830
TUA1-run0	0.0728	RSLDE-run1	0.1853
Baseline-run2	0.1864	Baseline-run2	0.2901
Baseline-run1	0.2042	Baseline-run1	0.3371

Table 13: Ranking Correlation between of English runs ranked by two different metrics (Kendall's *tau* with 95% CIs)

Dialogue Quality (A-score)		
NMD vs RSNOD	0.810	[0.091, 1.000]
Dialogue Quality (S-score)		
NMD vs RSNOD	0.524	[−0.059, 1.000]
Dialogue Quality (E-score)		
NMD vs RSNOD	0.714	[−0.059, 1.000]
Nugget Detection		
JSD vs RNSS	0.889	[0.613, 1.000]

REFERENCES

[1] Tao-Hsing Chang and Jian-He Chen. 2022. NKUST at the NTCIR-16 DialEval-2 Task. In *Proceedings of NTCIR-16.* to appear.

[2] Fei Ding, Kang Xin, Yunong Wu, and Fuji Ren. 2022. TUA1 at the NTCIR-16 DialEval-2 Task. In *Proceedings of NTCIR-16.* to appear.

[3] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9, 8 (1997), 1735–1780.

[4] Ting-Yun Hsiao, Yung-Wei Teng, Pei-Tz Chiu, Mike Tian-Jian Jiang, and Min-Yuh Day. 2022. IMNTPU Dialogue System Evaluation at the NTCIR-16 DialEval-2 Dialogue Quality and Nugget Detection. In *Proceedings of NTCIR-16.* to appear.

[5] Fan Li and Tetsuya Sakai. 2022. RSLDE at the NTCIR-16 DialEval-2 Task. In *Proceedings of NTCIR-16.* to appear.

[6] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40, 2 (2000), 99–121.

[7] Tetsuya Sakai. 2014. Metrics, Statistics, Tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173).* 116–163.

[8] Tetsuya Sakai. 2017. Towards Automatic Evaluation of Multi-Turn Dialogues: A Task Design that Leverages Inherently Subjective Annotations. In *Proceedings of EVIA 2017.*

[9] Tetsuya Sakai. 2018. Comparing Two Binned Probability Distributions for Information Access Evaluation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18).* ACM, New York, NY, USA, 1073–1076.

[10] Tetsuya Sakai. 2018. *Laboratory experiments in information retrieval: Sample sizes, effect sizes, and statistical power.* Springer. https://link.springer.com/book/10.1007/978-981-13-1199-4.

[11] Tetsuya Sakai. 2021. A Closer Look at Evaluation Measures for Ordinal Quantification. In *Proceedings of the CIKM 2021 Workshops.* http://ceur-ws.org/Vol-3052/paper21.pdf

[12] Tetsuya Sakai. 2021. Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification. In *Proceedings of ACL-IJCNLP 2021.* 2759–2769. https://aclanthology.org/2021.acl-long.214.pdf

[13] Tetsuya Sakai. 2022. On Variants of Root Normalised Order-aware Divergence and a Divergence based on Kendall ' s Tau. (2022). https://arxiv.org/abs/2204.07304

[14] Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Trans. Signal Processing* 45, 11 (1997), 2673–2681.

[15] Zhaohao Zeng, Sosuke Kato, and Tetsuya Sakai. 2019. Overview of the NTCIR-14 Short Text Conversation Task: Dialogue Quality and Nugget Detection Subtasks. In *Proceedings of NTCIR-14.* 290–315.

[16] Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai, and Inho Kang. 2020. Overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) Task. In *Proceedings of NTCIR-15*. 13–34.
[17] Zhaohao Zeng and Tetsuya Sakai. 2021. DCH-2: A Parallel Customer-Helpdesk Dialogue Corpus with Distributions of Annotators' Labels. *CoRR* abs/2104.08755 (2021). arXiv:2104.08755 https://arxiv.org/abs/2104.08755

APPENDIX A STATISTICAL SIGNIFICANCE TESTS

Table 14: Statistical significance in terms of NMD (Chinese DQ subtask, A-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run2	NKUST-run0	($p < 0.0001, ES_{E1} = 1.198$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.247$)
TUA1-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 1.151$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.200$)
TUA1-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 1.040$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.089$)
RSLDE-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.973$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.023$)
RSLDE-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.958$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.007$)
Baseline-run2	NKUST-run0	($p < 0.0001, ES_{E1} = 0.930$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.980$)
IMNTPU-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.887$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.937$)
Baseline-run0	NKUST-run0	($p = 0.0002, ES_{E1} = 0.724$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.773$)

Table 15: Statistical significance in terms of RSNOD (Chinese DQ subtask, A-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run2	RSLDE-run1	($p = 0.0442, ES_{E1} = 0.455$)
	IMNTPU-run0	($p = 0.0190, ES_{E1} = 0.488$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.777$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.783$)
TUA1-run1	Baseline-run1	($p < 0.0001, ES_{E1} = 0.677$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.683$)
TUA1-run0	Baseline-run1	($p = 0.0002, ES_{E1} = 0.615$)
	NKUST-run0	($p = 0.0002, ES_{E1} = 0.621$)
Baseline-run0	Baseline-run1	($p = 0.0326, ES_{E1} = 0.467$)
	NKUST-run0	($p = 0.0292, ES_{E1} = 0.474$)
Baseline-run2	NKUST-run0	($p = 0.0444, ES_{E1} = 0.454$)

Table 16: Statistical significance in terms of NMD (Chinese DQ subtask, S-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 1.293$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.603$)
TUA1-run2	NKUST-run0	($p < 0.0001, ES_{E1} = 1.285$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.595$)
RSLDE-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 1.214$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.524$)
RSLDE-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 1.197$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.507$)
Baseline-run2	NKUST-run0	($p < 0.0001, ES_{E1} = 1.146$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.456$)
TUA1-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 1.126$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.437$)
IMNTPU-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 1.116$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.426$)
Baseline-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.878$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.188$)

Table 17: Statistical significance in terms of RSNOD (Chinese DQ subtask, S-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run2	NKUST-run0	($p < 0.0001, ES_{E1} = 0.963$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.188$)
TUA1-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.882$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.108$)
TUA1-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.839$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.064$)
RSLDE-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.785$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.010$)
RSLDE-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.760$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.985$)
Baseline-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.725$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.950$)
IMNTPU-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.692$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.917$)
Baseline-run2	NKUST-run0	($p = 0.0004, ES_{E1} = 0.663$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.888$)

Table 18: Statistical significance in terms of NMD (Chinese DQ subtask, E-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run0	Baseline-run0	($p = 0.0148, ES_{E1} = 0.512$)
	Baseline-run2	($p = 0.0002, ES_{E1} = 0.662$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.879$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.120$)
RSLDE-run0	Baseline-run2	($p = 0.0040, ES_{E1} = 0.563$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.779$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.020$)
TUA1-run1	Baseline-run2	($p = 0.0128, ES_{E1} = 0.517$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.733$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.974$)
RSLDE-run1	Baseline-run2	($p = 0.0270, ES_{E1} = 0.489$)
	NKUST-run0	($p = 0.0002, ES_{E1} = 0.706$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.947$)
TUA1-run2	Baseline-run2	($p = 0.0484, ES_{E1} = 0.462$)
	NKUST-run0	($p = 0.0002, ES_{E1} = 0.678$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.919$)
IMNTPU-run0	NKUST-run0	($p = 0.0064, ES_{E1} = 0.542$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.783$)
Baseline-run0	Baseline-run1	($p = 0.0008, ES_{E1} = 0.608$)

Table 19: Statistical significance in terms of RSNOD (Chinese DQ subtask, E-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.850$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.142$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 1.230$)
TUA1-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.729$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.021$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 1.108$)
RSLDE-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.713$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.004$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 1.092$)
TUA1-run2	NKUST-run0	($p = 0.0002, ES_{E1} = 0.699$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.991$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 1.079$)
RSLDE-run1	NKUST-run0	($p = 0.0008, ES_{E1} = 0.635$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.927$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 1.014$)
Baseline-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.772$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 0.859$)
IMNTPU-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.765$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 0.852$)

Table 20: Statistical significance in terms of JSD (the Chinese ND subtask) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
RSLDE-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 2.041$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.320$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.930$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 3.677$)
Baseline-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 2.003$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.282$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.892$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 3.639$)
RSLDE-run2	Baseline-run2	($p < 0.0001, ES_{E1} = 1.968$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.247$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.858$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 3.605$)
RSLDE-run1	Baseline-run2	($p < 0.0001, ES_{E1} = 1.926$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.205$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.815$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 3.562$)
NKUST-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.870$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.149$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.759$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 3.507$)
TUA1-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.822$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.101$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.711$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 3.458$)
Baseline-run2	TUA1-run1	($p < 0.0001, ES_{E1} = 1.637$)
Baseline-run1	TUA1-run1	($p = 0.0002, ES_{E1} = 1.358$)

Table 21: Statistical significance in terms of RNSS (the Chinese ND subtask) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
RSLDE-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.540$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.098$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.577$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 2.773$)
Baseline-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.484$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.042$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.521$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 2.717$)
RSLDE-run1	Baseline-run2	($p < 0.0001, ES_{E1} = 1.412$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.970$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.449$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 2.645$)
RSLDE-run2	Baseline-run2	($p < 0.0001, ES_{E1} = 1.402$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.960$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.439$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 2.635$)
NKUST-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.354$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.912$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.391$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 2.587$)
TUA1-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.331$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.890$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.368$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 2.565$)
Baseline-run2	NKUST-run1	($p = 0.0004, ES_{E1} = 1.037$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 1.233$)

Table 22: Statistical significance in terms of NMD (English DQ subtask, A-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run0	RSLDE-run1	($p = 0.0042, ES_{E1} = 0.562$)
	RSLDE-run0	($p = 0.0006, ES_{E1} = 0.623$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.160$)
Baseline-run2	Baseline-run1	($p < 0.0001, ES_{E1} = 0.913$)
IMNTPU-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.837$)
Baseline-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.712$)
RSLDE-run1	Baseline-run1	($p = 0.0010, ES_{E1} = 0.598$)
RSLDE-run0	Baseline-run1	($p = 0.0078, ES_{E1} = 0.538$)

Table 23: Statistical significance in terms of RSNOD (English DQ subtask, A-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run0	IMNTPU-run0	($p = 0.0046, ES_{E1} = 0.535$)
	RSLDE-run0	($p = 0.0002, ES_{E1} = 0.610$)
	RSLDE-run1	($p < 0.0001, ES_{E1} = 0.713$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 0.753$)

Table 24: Statistical significance in terms of NMD (English DQ subtask, S-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 1.535$)
Baseline-run2	Baseline-run1	($p < 0.0001, ES_{E1} = 1.451$)
IMNTPU-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 1.424$)
RSLDE-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 1.346$)
RSLDE-run1	Baseline-run1	($p < 0.0001, ES_{E1} = 1.281$)
Baseline-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 1.248$)

Table 25: Statistical significance in terms of RSNOD (English DQ subtask, S-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 1.047$)
Baseline-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.923$)
IMNTPU-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.891$)
Baseline-run2	Baseline-run1	($p < 0.0001, ES_{E1} = 0.851$)
RSLDE-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.836$)
RSLDE-run1	Baseline-run1	($p < 0.0001, ES_{E1} = 0.764$)

Table 26: Statistical significance in terms of NMD (English DQ subtask, E-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.872$)
IMNTPU-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.825$)
RSLDE-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.790$)
Baseline-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.788$)
RSLDE-run1	Baseline-run1	($p < 0.0001, ES_{E1} = 0.773$)
Baseline-run2	Baseline-run1	($p = 0.0386, ES_{E1} = 0.463$)

Table 27: Statistical significance in terms of RSNOD (English DQ subtask, E-score) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
TUA1-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.910$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 0.998$)
Baseline-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.907$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 0.995$)
IMNTPU-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.808$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 0.896$)
RSLDE-run0	Baseline-run1	($p < 0.0001, ES_{E1} = 0.802$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 0.890$)
RSLDE-run1	Baseline-run1	($p = 0.0004, ES_{E1} = 0.732$)
	Baseline-run2	($p < 0.0001, ES_{E1} = 0.820$)

Table 28: Statistical significance in terms of JSD (English ND subtask) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
RSLDE-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 2.659$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 3.021$)
IMNTPU-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 2.571$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.934$)
Baseline-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 2.522$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.885$)
NKUST-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 2.488$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.850$)
RSLDE-run2	Baseline-run2	($p < 0.0001, ES_{E1} = 2.418$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.781$)
RSLDE-run1	Baseline-run2	($p < 0.0001, ES_{E1} = 2.387$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.750$)
TUA1-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 2.311$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.674$)

Table 29: Statistical significance in terms of RNSS (English ND subtask) calculated by Randomised Tukey HSD tests

Run	significantly better than these runs	
IMNTPU-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.720$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.330$)
RSLDE-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.667$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.277$)
Baseline-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.529$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.138$)
NKUST-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.500$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.109$)
RSLDE-run2	Baseline-run2	($p < 0.0001, ES_{E1} = 1.456$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 2.065$)
TUA1-run0	Baseline-run2	($p < 0.0001, ES_{E1} = 1.388$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.997$)
RSLDE-run1	Baseline-run2	($p < 0.0001, ES_{E1} = 1.358$)
	Baseline-run1	($p < 0.0001, ES_{E1} = 1.968$)
Baseline-run2	Baseline-run1	($p = 0.0398, ES_{E1} = 0.609$)