Overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) Task

Zhaohao Zeng
Waseda University, Japan
zhaohao@fuji.waseda.jp

Sosuke Kato
Waseda University, Japan
sow@suou.waseda.jp

Tetsuya Sakai
Waseda University, Japan
tetsuyasakai@acm.org

Inho Kang
Naver Corporation, Korea
once.ihkang@navercorp.com

ABSTRACT

In this paper, we provide an overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) task. DialEval-1 consists of two subtasks: Dialogue Quality (DQ) and Nugget Detection (ND). Both DQ and ND subtasks aim to evaluate customer-helpdesk dialogues automatically. The DQ subtask is to assign quality scores to each dialogue in terms of three criteria: task accomplishment, customer satisfaction, and efficiency; and the ND subtask is to classify whether a customer or helpdesk turn is a nugget, where being a nugget means that the dialogue turn helps towards problem solving. In this overview paper, we introduce the task setting, evaluation methods and data collection, and report the official evaluation results of 18 runs received from 7 teams.

1 INTRODUCTION

Recently, many researchers and engineers have been developing automatic dialogue agents to handle with customers' inquires in a more efficient and economic way. However, human annotators are usually employed to tune such dialogue systems, and such manual evaluation may be expensive and inefficient. To alleviate this problem, we propose to automatically evaluate customer-helpdesk dialogues [9]. Briefly, given a customer-helpdesk dialogue, we want to automatically know how good it is and which dialogue turns are helpful without human annotator. Thus, we launch DialEval-1 task at NTCIR-15 to explore ideas with researchers in this community.

DialEval-1 is the successor of ND and DQ subtasks of Short Text Conversation (STC-3) [8] at NTCIR-14 in 2019. At STC-3, we organised Dialogue Quality (DQ) and Nugget Detection (ND) subtasks along with Chinese Emotional Conversation Generation (CECG) subtask. At DialEval-1, we re-use the data collection provided at STC-3, but constructed a new test collection comprising 300 dialogues. In addition, by translating more dialogues, the English data collection of DialEval-1 comprising 2941 dialogues (2,251 training, 390 development, and 300 test).

The schedule of DialEval-1 is shown in Table 1. Table 2 shows the detailed number of runs submitted by each team for each subtask and each language.

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

Table 1: Schedule of DialEval-1 at NTCIR-15

Time	Content
July 1 2020	Test data released
July 31 2020	Run submissions due
Aug 31 2020	Results summary and draft overview released
Sep 20 2020	Participant paper submissions due
Nov 1 2020	All camera-ready papers due
Dec 2020	NTCIR-15 Conference

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

Team	Chinese		English	
	DQ	ND	DQ	ND
IMTKU	3	3	1	3
NKUST	2	2	1	1
RSLNV	1	2	1	2
SKYMN	0	0	3	0
TMUDS	0	3	0	0
TUA1	3	1	0	0
WUST	1	1	0	0
Total	10	12	6	6

2 TASK DEFINITION

The goal of DialEval-1 is to explore approaches to evaluating task-oriented, multi-round, textual helpdesk-customer dialogue systems automatically. Identical to STC-3, 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 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

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.

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 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.

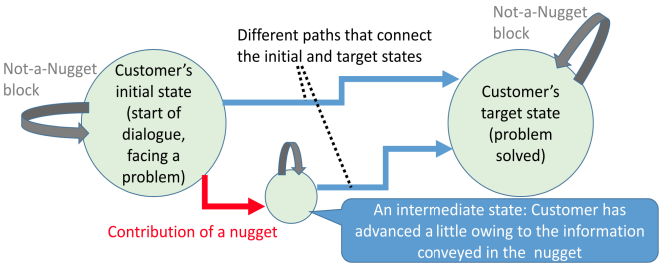

Figure 2: Task accomplishment as state transitions, and the role of a nugget.

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; (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 crawled from Weibo are originally in Chinese, but we manually translate a part of them into English (to be detailed in Section 4). Thus, each subtask has a Chinese version and a English version. That is, the participants must use Chinese training data only to build the Chinese runs and use English training data only to build the English runs.

2.5 Baselines

We prepared three baseline models for each language and each subtask as follows;

BL-lstm A baseline model[1] which leverages Bidirectional Long Short-term Memory [1, 7];

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

BL-popularity 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.

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

3 EVALUATION METHODS

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 [4]. Identical to STC-3 [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 [5], 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 [2]. 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 [5]. 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 $B^* = i|p^*(i) > 0$, 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:

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

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:

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

Finally, we define the Root Symmetric Normalised OD:

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

In the DQ subtask, we use both NMD and RSNOD as metrics 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 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:

$$\mathcal{J}SD(p||p^*) = \frac{KLD(p||p_M) + KLD(p_M||p^*)}{2} \tag{8}$$

$$\text{where } KLD(p1||p2) = \sum_{i\ s.t.\ p1(i)>0} p_1(i) \log_2 \frac{p_1(i)}{p_2(i)} \tag{9}$$

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. By default, $\alpha = 0.5$ at DialEval-1 task.

3.3 Online Evaluation

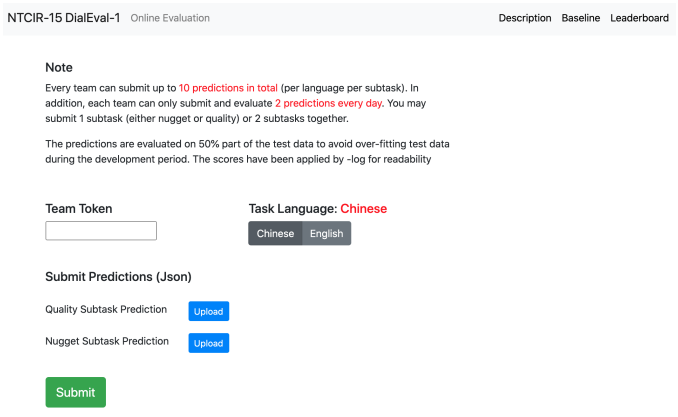

Figure 3: Screenshot of the online evaluation website

Before the due of participant run submission, we hosted an online evaluation website to allow participants to evaluate their predictions on the test data for tuning their models. To prevent overfitting the test data, only 50% of the test data are utilised for the online evaluation. Also, each team can only submit 10 predictions per language per subtask in total. Note that the online evaluation tool returns scores that are transformed by negative logarithm $-log_2$ for readability, but we do not apply it in this overview paper to be consistent with the STC-3 overview paper.

4 DATA COLLECTION

4.1 Training and Development Data

The statistics of the DialEval-1 data collection is shown in Table 3. We re-use the DCH-1 data collection [8] for training and development, as DCH-1 was utilised at NTCIR-14 STC-3 task for training and test. The DCH-1 data collection consists of *real* (i.e., human-human) customer-helpdesk dialogues collected from Weibo, and there are 3,700 training dialogues and 390 test dialogues. At DialEval-1, the DCH-1 training data is still utilised for training, but DCH-1 test data is used as development data to tune the model.

4.2 Test Data

For test, we collected dialogues from Weibo and annotated them to build a new test collection that consists of 300 annotated dialogues. The test dialogues of DialEval-1 are collected in the same manner as DCH-1. For annotation, we hired 20 Chinese students from the Faculty of Science and Engineering at Waseda University, and each dialogue was annotated by each annotator independently. The annotation instructions for the DialEval-1 annotators were the same as that used for annotating DCH-1.

4.3 English Data Collection

The original dialogues crawled from Weibo are in Chinese, so we hired a professional translation company to manually translate a part of the dialogues into English. Specifically, 2,251 of 3,700 training dialogues, all the development dialogues, and all the test dialogues have been translated in to English. Since the translation does not change the semantic information of the dialogues, we didn't annotate the English dialogues separately. Instead, the annotations of the English data collection are copied from the Chinese data collection.

5 RESULTS

5.1 Chinese Subtasks

First, we rank all the runs using the two evaluation metrics respectively for each subtask, and then calculate the ranking correlation using Kendall's τ between them, 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 subtasks when we rank the participant runs. This finding is consistent with what we observed at STC-3.

[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=10000.

Table 3: Statistics of DialEval-1 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-1	DCH-1	Weibo	Translation		
Data timestamps		Jan. 2013 ~ Apr. 2018	Apr. 2018 ~ Jul. 2019		Jan. 2013 ~ Apr. 2018	Apr. 2018 ~ Jul. 2019
#dialogues	3,700	390	300	2,251	390	300
#annotators/dialogue	19	19	20	19	19	20
Avg. #posts/dialogue	4.512	4.877	4.557	4.522	4.877	4.557
Avg. post length	44.568	47.988	52.198	31.986	30.890	39.769
Avg. turn length	48.313	52.008	55.314	34.964	33.478	42.143
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	0.2102	IMTKU-run2	0.1392
IMTKU-run2	0.2130	TUA1-run0	0.1396
TUA1-run0	0.2136	IMTKU-run0	0.1406
IMTKU-run0	0.2165	TUA1-run2	0.1412
IMTKU-run1	0.2204	IMTKU-run1	0.1442
BL-lstm	0.2305	TUA1-run1	0.1510
RSLNV-run0	0.2345	NKUST-run1	0.1594
WUST-run0	0.2427	BL-lstm	0.1598
NKUST-run1	0.2430	RSLNV-run0	0.1606
BL-popularity	0.2473	BL-popularity	0.1643
TUA1-run1	0.2484	WUST-run0	0.1724
NKUST-run0	0.2696	NKUST-run0	0.2384
BL-uniform	0.2706	BL-uniform	0.2522

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

Run	Mean RSNOD	Run	Mean NMD
IMTKU-run2	0.1918	IMTKU-run2	0.1254
IMTKU-run1	0.1964	IMTKU-run0	0.1284
IMTKU-run0	0.1977	IMTKU-run1	0.1290
TUA1-run2	0.2024	TUA1-run2	0.1310
TUA1-run0	0.2053	TUA1-run0	0.1322
NKUST-run1	0.2057	NKUST-run1	0.1363
BL-lstm	0.2088	TUA1-run1	0.1397
WUST-run0	0.2131	BL-popularity	0.1442
RSLNV-run0	0.2141	BL-lstm	0.1455
BL-popularity	0.2288	RSLNV-run0	0.1483
TUA1-run1	0.2302	WUST-run0	0.1540
NKUST-run0	0.2653	NKUST-run0	0.2289
BL-uniform	0.2811	BL-uniform	0.2497

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

Run	Mean RSNOD	Run	Mean NMD
TUA1-run0	0.1615	TUA1-run0	0.1144
TUA1-run2	0.1617	IMTKU-run1	0.1165
IMTKU-run1	0.1631	IMTKU-run0	0.1181
IMTKU-run0	0.1648	TUA1-run2	0.1187
IMTKU-run2	0.1655	IMTKU-run2	0.1194
BL-lstm	0.1782	TUA1-run1	0.1253
WUST-run0	0.1795	WUST-run0	0.1386
TUA1-run1	0.1810	BL-lstm	0.1386
RSLNV-run0	0.1811	RSLNV-run0	0.1393
NKUST-run0	0.2222	NKUST-run1	0.1508
NKUST-run1	0.2295	BL-popularity	0.1781
BL-uniform	0.2425	NKUST-run0	0.1973
BL-popularity	0.2614	BL-uniform	0.2110

Table 7: Chinese Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
IMTKU-run0	0.0674	WUST-run0	0.1633
WUST-run0	0.0695	IMTKU-run0	0.1636
BL-lstm	0.0709	BL-lstm	0.1673
IMTKU-run1	0.0726	IMTKU-run1	0.1700
RSLNV-run0	0.0746	RSLNV-run0	0.1749
IMTKU-run2	0.0752	IMTKU-run2	0.1754
RSLNV-run2	0.0768	RSLNV-run2	0.1760
TUA1-run0	0.0859	TUA1-run0	0.1892
TMUDS-run1	0.0883	TMUDS-run2	0.1948
TMUDS-run2	0.0887	TMUDS-run1	0.1953
TMUDS-run0	0.0906	TMUDS-run0	0.1995
BL-popularity	0.1301	BL-popularity	0.2068
NKUST-run1	0.1905	NKUST-run1	0.3036
BL-uniform	0.2858	NKUST-run0	0.4169
NKUST-run0	0.3116	BL-uniform	0.4190

Tables 4 to 6 shows the mean evaluation scores for the DQ subtask in terms of A-score, S-score, E-score, respectively, and Table 7 shows the mean evaluation scores for the ND subtask. We conduct randomised Tukey HSD tests using the Discpower tool[3] with

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

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

Dialogue Quality (A-score)		
NMD vs RSNOD	0.692	[0.257, 1.000]
Dialogue Quality (S-score)		
NMD vs RSNOD	0.769	[0.183, 1.000]
Dialogue Quality (E-score)		
NMD vs RSNOD	0.795	[0.562, 0.971]
Nugget Detection		
JSD vs RNSS	0.943	[0.783, 1.000]

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

Run	Mean RSNOD	Run	Mean NMD
IMTKU-run0	0.2197	IMTKU-run0	0.1437
BL-lstm	0.2271	BL-lstm	0.1591
RSLNV-run0	0.2311	RSLNV-run0	0.1603
SKYMN-run2	0.2410	SKYMN-run2	0.1608
SKYMN-run0	0.2471	SKYMN-run0	0.1626
BL-popularity	0.2473	BL-popularity	0.1643
SKYMN-run1	0.2555	SKYMN-run1	0.1663
BL-uniform	0.2706	NKUST-run0	0.2345
NKUST-run0	0.2801	BL-uniform	0.2522

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

Run	Mean RSNOD	Run	Mean NMD
IMTKU-run0	0.1892	IMTKU-run0	0.1250
BL-lstm	0.2111	BL-lstm	0.1413
RSLNV-run0	0.2169	BL-popularity	0.1442
SKYMN-run2	0.2177	RSLNV-run0	0.1454
SKYMN-run0	0.2223	SKYMN-run2	0.1468
BL-popularity	0.2288	SKYMN-run0	0.1480
SKYMN-run1	0.2305	SKYMN-run1	0.1515
NKUST-run0	0.2637	NKUST-run0	0.2198
BL-uniform	0.2811	BL-uniform	0.2497

$B = 10,000$ trials [3]. Tables 14 to 21 summarise the statistical significance test results and p-values and effect sizes computed by Randomised Tukey HSD (i.e., standardised mean differences) based on one-way ANOVA (without replication) [6]. It can be observed that show that which runs the target run are statistically significantly better than. At STC-3, none of the participant runs are statistically significantly better than the BL-LSTM model. However, at DialEval-1, IMTKU-run2 outperforms the baselines significantly ($p < 0.5$) in Chinese DQ subtask in terms of NMD. Also, some other runs also outperform the BL-LSTM for one of A, E, or S scores. For example, TUA1-run2 is significantly better than BL-LSTM in terms of NMD for A-score. Both IMTKU runs and TUA1 runs are based on BERT and its variants (e.g., XLM-RoBERTa).

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

Run	Mean RSNOD	Run	Mean NMD
IMTKU-run0	0.1657	IMTKU-run0	0.1221
BL-lstm	0.1687	BL-lstm	0.1248
SKYMN-run2	0.1783	SKYMN-run2	0.1321
RSLNV-run0	0.1789	SKYMN-run0	0.1322
SKYMN-run0	0.1803	SKYMN-run1	0.1343
SKYMN-run1	0.1842	RSLNV-run0	0.1354
NKUST-run0	0.2248	BL-popularity	0.1781
BL-uniform	0.2425	NKUST-run0	0.1963
BL-popularity	0.2614	BL-uniform	0.2110

Table 12: English Nugget Detection Results

Run	Mean JSD	Run	Mean RNSS
IMTKU-run0	0.0707	IMTKU-run0	0.1699
RSLNV-run0	0.0743	RSLNV-run0	0.1753
IMTKU-run2	0.0757	IMTKU-run2	0.1753
BL-lstm	0.0762	BL-lstm	0.1781
IMTKU-run1	0.0789	IMTKU-run1	0.1804
RSLNV-run1	0.0989	BL-popularity	0.2068
BL-popularity	0.1301	RSLNV-run1	0.2142
BL-uniform	0.2858	NKUST-run0	0.4172
NKUST-run0	0.3157	BL-uniform	0.4190

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

Dialogue Quality (A-score)		
NMD vs RSNOD	0.944	[0.733, 1.000]
Dialogue Quality (S-score)		
NMD vs RSNOD	0.833	[0.226, 1.000]
Dialogue Quality (E-score)		
NMD vs RSNOD	0.778	[0.355, 1.000]
Nugget Detection		
JSD vs RNSS	0.889	[0.562, 1.000]

5.2 English Subtasks

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

We also conduct randomised Tukey HSD tests for English runs and Tables 22 to 29 summarises the significance test results. Randomised Tukey HSD p-values and effect sizes (i.e., standardised mean differences) based on one-way ANOVA (without replication) [6] are also shown. From the English results with the evaluation metrics for ND and DQ subtasks, it can be observed that:

- Most participant runs underperform the LSTM baseline, while the differences are usually not statistically significant.

- IMTKU-run0 outperforms the LSTM baseline in most subtasks, but the differences are only statistically significant for DQ S-score subtask evaluated by RSNOD.
- The scores of the top run in the English subtasks are still worse than the scores of the top run in the Chinese subtasks.

In Table 8, we also compare the system rankings according to the two evaluation metrics of each subtasks in terms of Kendall's τ for English runs, and we also find that there is no statistically significant difference between the metrics.

5.3 Top Runs

IMTKU and TUA1 are the top runs in the Chinese and English subtasks. IMTKU adds more special tokens (such as the position of the current turn) to the input, and fine-tunes the pre-trained models directly. In contrast, TUA1 utilises the pre-trained BERT without fine-tuning, and trains the downstream networks (i.e., the followed LSTM and fully connected layers) only. The results may suggest that the large-scale pre-trained models may be more effective than the pure LSTM networks for DialEval-1.

6 CONCLUSION

This overview describe the task definition, data collection, evaluation metrics, and results of NTCIR-15 DialEval-1 task. From the results, we observed that (1) In Chinese subtasks, the top runs that utilise BERT statistically significantly outperform the Baseline LSTM models, which did not happen at the previous STC-3 task; (2) however, none of participant runs outperform the LSTM baseline significantly in all English subtasks except DQ S-score. (3) The results of the runs trained by the Chinese data are still slightly better than the runs trained by the English data. (4) There is no substantial difference between the evaluation metrics when we rank the participant runs.

ACKNOWLEDGEMENT

We thank the DialEval-1 participants and the NTCIR chairs for making this task happen. This work was partially supported by JSPS KAKENHI Grant Number 17H01830.

REFERENCES

[1] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9, 8 (1997), 1735–1780.
[2] 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.
[3] Tetsuya Sakai. 2014. Metrics, Statistics, Tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*. 116–163.
[4] Tetsuya Sakai. 2017. Towards Automatic Evaluation of Multi-Turn Dialogues: A Task Design that Leverages Inherently Subjective Annotations. In *Proceedings of EVIA 2017*.
[5] 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.
[6] 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.
[7] Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Trans. Signal Processing* 45, 11 (1997), 2673–2681.
[8] 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.
[9] Zhaohao Zeng, Cheng Luo, Lifeng Shang, Hang Li, and Tetsuya Sakai. 2018. Towards Automatic Evaluation of Customer-Helpdesk Dialogues. *Journal of Information Processing* 26 (2018), 768–778. https://www.jstage.jst.go.jp/article/ipsjjip/26/0/26_768/_pdf/-char/en

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	
IMTKU-run2	NKUST-run1	($p = 0.0344, ES_{E1} = 0.218$)
	BL-lstm	($p = 0.0273, ES_{E1} = 0.222$)
	RSLNV-run0	($p = 0.0166, ES_{E1} = 0.230$)
	BL-popularity	($p = 0.0018, ES_{E1} = 0.270$)
	WUST-run0	($p < 0.0001, ES_{E1} = 0.357$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.070$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.219$)
TUA1-run0	NKUST-run1	($p = 0.0434, ES_{E1} = 0.214$)
	BL-lstm	($p = 0.0344, ES_{E1} = 0.218$)
	RSLNV-run0	($p = 0.0210, ES_{E1} = 0.226$)
	BL-popularity	($p = 0.0021, ES_{E1} = 0.266$)
	WUST-run0	($p < 0.0001, ES_{E1} = 0.353$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.066$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.215$)
IMTKU-run0	RSLNV-run0	($p = 0.0394, ES_{E1} = 0.215$)
	BL-popularity	($p = 0.0035, ES_{E1} = 0.255$)
	WUST-run0	($p = 0.0001, ES_{E1} = 0.342$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.055$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.204$)
TUA1-run2	BL-popularity	($p = 0.0050, ES_{E1} = 0.249$)
	WUST-run0	($p = 0.0001, ES_{E1} = 0.336$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.049$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.197$)
IMTKU-run1	BL-popularity	($p = 0.0371, ES_{E1} = 0.216$)
	WUST-run0	($p = 0.0003, ES_{E1} = 0.303$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.016$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.165$)
TUA1-run1	WUST-run0	($p = 0.0164, ES_{E1} = 0.230$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.943$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.092$)
NKUST-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.852$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.001$)
BL-lstm	NKUST-run0	($p < 0.0001, ES_{E1} = 0.848$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.997$)
RSLNV-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.840$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.988$)
BL-popularity	NKUST-run0	($p < 0.0001, ES_{E1} = 0.800$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.948$)
WUST-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.713$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.861$)

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	BL-lstm	($p = 0.0297, ES_{E1} = 0.189$)
	RSLNV-run0	($p = 0.0024, ES_{E1} = 0.227$)
	WUST-run0	($p = 0.0001, ES_{E1} = 0.302$)
	NKUST-run1	($p = 0.0001, ES_{E1} = 0.305$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.345$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 0.356$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.552$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.562$)
IMTKU-run2	RSLNV-run0	($p = 0.0143, ES_{E1} = 0.201$)
	WUST-run0	($p = 0.0003, ES_{E1} = 0.276$)
	NKUST-run1	($p = 0.0003, ES_{E1} = 0.279$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.319$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 0.330$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.526$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.536$)
TUA1-run0	RSLNV-run0	($p = 0.0210, ES_{E1} = 0.195$)
	WUST-run0	($p = 0.0005, ES_{E1} = 0.270$)
	NKUST-run1	($p = 0.0004, ES_{E1} = 0.273$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.313$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 0.324$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.520$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.530$)
IMTKU-run0	WUST-run0	($p = 0.0009, ES_{E1} = 0.243$)
	NKUST-run1	($p = 0.0009, ES_{E1} = 0.246$)
	BL-popularity	($p = 0.0001, ES_{E1} = 0.286$)
	TUA1-run1	($p = 0.0001, ES_{E1} = 0.297$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.493$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.503$)
IMTKU-run1	WUST-run0	($p = 0.0093, ES_{E1} = 0.207$)
	NKUST-run1	($p = 0.0078, ES_{E1} = 0.210$)
	BL-popularity	($p = 0.0007, ES_{E1} = 0.250$)
	TUA1-run1	($p = 0.0005, ES_{E1} = 0.261$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.457$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.467$)
BL-lstm	NKUST-run0	($p < 0.0001, ES_{E1} = 0.364$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.373$)
RSLNV-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.326$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.335$)
WUST-run0	NKUST-run0	($p = 0.0007, ES_{E1} = 0.250$)
	BL-uniform	($p = 0.0005, ES_{E1} = 0.260$)
NKUST-run1	NKUST-run0	($p = 0.0008, ES_{E1} = 0.247$)
	BL-uniform	($p = 0.0006, ES_{E1} = 0.257$)
BL-popularity	NKUST-run0	($p = 0.0093, ES_{E1} = 0.207$)
	BL-uniform	($p = 0.0046, ES_{E1} = 0.217$)
TUA1-run1	NKUST-run0	($p = 0.0186, ES_{E1} = 0.197$)
	BL-uniform	($p = 0.0096, ES_{E1} = 0.206$)

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	
IMTKU-run2	BL-popularity	($p = 0.0384, ES_{E1} = 0.229$)
	BL-lstm	($p = 0.0166, ES_{E1} = 0.246$)
	RSLNV-run0	($p = 0.0027, ES_{E1} = 0.279$)
	WUST-run0	($p < 0.0001, ES_{E1} = 0.349$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.264$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.519$)
IMTKU-run0	RSLNV-run0	($p = 0.0195, ES_{E1} = 0.243$)
	WUST-run0	($p = 0.0003, ES_{E1} = 0.312$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.227$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.482$)
IMTKU-run1	RSLNV-run0	($p = 0.0264, ES_{E1} = 0.236$)
	WUST-run0	($p = 0.0003, ES_{E1} = 0.305$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.220$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.476$)
TUA1-run2	WUST-run0	($p = 0.0027, ES_{E1} = 0.280$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.195$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.450$)
TUA1-run0	WUST-run0	($p = 0.0058, ES_{E1} = 0.266$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.181$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.436$)
NKUST-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 1.131$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.386$)
TUA1-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 1.089$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.344$)
BL-popularity	NKUST-run0	($p < 0.0001, ES_{E1} = 1.035$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.290$)
BL-lstm	NKUST-run0	($p < 0.0001, ES_{E1} = 1.018$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.273$)
RSLNV-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.985$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.240$)
WUST-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.915$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.170$)
NKUST-run0	BL-uniform	($p = 0.0106, ES_{E1} = 0.255$)

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	
IMTKU-run2	WUST-run0	($p = 0.0080, ES_{E1} = 0.213$)
	RSLNV-run0	($p = 0.0042, ES_{E1} = 0.223$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.370$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 0.384$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.735$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.893$)
IMTKU-run1	BL-popularity	($p < 0.0001, ES_{E1} = 0.324$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 0.338$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.689$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.847$)
IMTKU-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.311$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 0.326$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.676$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.834$)
TUA1-run2	BL-popularity	($p = 0.0001, ES_{E1} = 0.264$)
	TUA1-run1	($p < 0.0001, ES_{E1} = 0.278$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.629$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.787$)
TUA1-run0	BL-popularity	($p = 0.0020, ES_{E1} = 0.235$)
	TUA1-run1	($p = 0.0007, ES_{E1} = 0.249$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.599$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.757$)
NKUST-run1	BL-popularity	($p = 0.0022, ES_{E1} = 0.231$)
	TUA1-run1	($p = 0.0010, ES_{E1} = 0.245$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.596$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.754$)
BL-lstm	BL-popularity	($p = 0.0199, ES_{E1} = 0.200$)
	TUA1-run1	($p = 0.0074, ES_{E1} = 0.215$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.565$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.723$)
WUST-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.522$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.680$)
RSLNV-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.512$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.670$)
BL-popularity	NKUST-run0	($p < 0.0001, ES_{E1} = 0.365$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.523$)
TUA1-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.351$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.509$)

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	WUST-run0	($p = 0.0006, ES_{E1} = 0.304$)
	BL-lstm	($p = 0.0006, ES_{E1} = 0.305$)
	RSLNV-run0	($p = 0.0001, ES_{E1} = 0.313$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.458$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.800$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.042$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.214$)
IMTKU-run1	WUST-run0	($p = 0.0035, ES_{E1} = 0.277$)
	BL-lstm	($p = 0.0034, ES_{E1} = 0.278$)
	RSLNV-run0	($p = 0.0020, ES_{E1} = 0.286$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.431$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.773$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.015$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.188$)
IMTKU-run0	WUST-run0	($p = 0.0110, ES_{E1} = 0.257$)
	BL-lstm	($p = 0.0108, ES_{E1} = 0.257$)
	RSLNV-run0	($p = 0.0067, ES_{E1} = 0.266$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.411$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.753$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.995$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.167$)
TUA1-run2	WUST-run0	($p = 0.0164, ES_{E1} = 0.250$)
	BL-lstm	($p = 0.0159, ES_{E1} = 0.250$)
	RSLNV-run0	($p = 0.0100, ES_{E1} = 0.259$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.404$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.746$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.988$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.160$)
IMTKU-run2	WUST-run0	($p = 0.0245, ES_{E1} = 0.241$)
	BL-lstm	($p = 0.0240, ES_{E1} = 0.241$)
	RSLNV-run0	($p = 0.0159, ES_{E1} = 0.250$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.395$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.737$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.979$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.151$)
TUA1-run1	NKUST-run1	($p < 0.0001, ES_{E1} = 0.321$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.664$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.906$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.078$)
WUST-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.496$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.738$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.910$)
BL-lstm	BL-popularity	($p < 0.0001, ES_{E1} = 0.496$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.738$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.910$)
RSLNV-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.487$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.729$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.901$)

Continued on next page

Table 18 – continued from previous page

Run	significantly better than these runs	
NKUST-run1	BL-popularity	($p < 0.0001, ES_{E1} = 0.342$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.584$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.757$)
BL-popularity	NKUST-run0	($p = 0.0235, ES_{E1} = 0.242$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.414$)

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	TUA1-run1	($p = 0.0233, ES_{E1} = 0.245$)
	RSLNV-run0	($p = 0.0214, ES_{E1} = 0.246$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.763$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.855$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.019$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.256$)
TUA1-run2	TUA1-run1	($p = 0.0270, ES_{E1} = 0.242$)
	RSLNV-run0	($p = 0.0252, ES_{E1} = 0.244$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.760$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.853$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.016$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.254$)
IMTKU-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.743$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.836$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.999$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.237$)
IMTKU-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.722$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.815$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.978$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.215$)
IMTKU-run2	NKUST-run0	($p < 0.0001, ES_{E1} = 0.713$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.805$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.968$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.206$)
BL-lstm	NKUST-run0	($p < 0.0001, ES_{E1} = 0.553$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.646$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.809$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.047$)
WUST-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.537$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.630$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.793$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.031$)
TUA1-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.518$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.610$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.774$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.011$)
RSLNV-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.516$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 0.609$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.772$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.010$)
NKUST-run0	BL-uniform	($p = 0.0135, ES_{E1} = 0.256$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.493$)
NKUST-run1	BL-popularity	($p < 0.0001, ES_{E1} = 0.401$)
BL-uniform	BL-popularity	($p = 0.0331, ES_{E1} = 0.237$)

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

Run	significantly better than these runs	
IMTKU-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.095$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.153$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.819$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.269$)
WUST-run0	BL-popularity	($p < 0.0001, ES_{E1} = 1.059$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.117$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.783$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.233$)
BL-lstm	BL-popularity	($p < 0.0001, ES_{E1} = 1.034$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.092$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.758$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.208$)
IMTKU-run1	BL-popularity	($p < 0.0001, ES_{E1} = 1.005$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.063$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.729$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.179$)
RSLNV-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.971$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.028$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.695$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.145$)
IMTKU-run2	BL-popularity	($p < 0.0001, ES_{E1} = 0.959$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 2.017$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.683$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.133$)
RSLNV-run2	BL-popularity	($p < 0.0001, ES_{E1} = 0.931$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.989$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.655$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.105$)
TUA1-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.772$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.830$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.496$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.946$)
TMUDS-run1	BL-popularity	($p < 0.0001, ES_{E1} = 0.731$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.789$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.455$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.905$)
TMUDS-run2	BL-popularity	($p < 0.0001, ES_{E1} = 0.724$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.782$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.448$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.898$)
TMUDS-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.691$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.749$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.415$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.865$)
BL-popularity	NKUST-run1	($p < 0.0001, ES_{E1} = 1.058$)
	BL-uniform	($p < 0.0001, ES_{E1} = 2.724$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.174$)

Continued on next page

Table 20 – continued from previous page

Run	significantly better than these runs	
NKUST-run1	BL-uniform	($p < 0.0001, ES_{E1} = 1.666$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 2.116$)
BL-uniform	NKUST-run0	($p = 0.0394, ES_{E1} = 0.450$)

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

Run	significantly better than these runs	
WUST-run0	TMUDS-run2	($p = 0.0167, ES_{E1} = 0.414$)
	TMUDS-run1	($p = 0.0132, ES_{E1} = 0.421$)
	TMUDS-run0	($p = 0.0013, ES_{E1} = 0.476$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.572$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.847$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.338$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.366$)
IMTKU-run0	TMUDS-run2	($p = 0.0182, ES_{E1} = 0.410$)
	TMUDS-run1	($p = 0.0153, ES_{E1} = 0.416$)
	TMUDS-run0	($p = 0.0014, ES_{E1} = 0.472$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.568$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.843$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.334$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.362$)
BL-lstm	TMUDS-run0	($p = 0.0119, ES_{E1} = 0.424$)
	BL-popularity	($p = 0.0002, ES_{E1} = 0.519$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.794$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.286$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.313$)
IMTKU-run1	TMUDS-run0	($p = 0.0362, ES_{E1} = 0.388$)
	BL-popularity	($p = 0.0009, ES_{E1} = 0.484$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.759$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.250$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.278$)
RSLNV-run0	BL-popularity	($p = 0.0134, ES_{E1} = 0.420$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.695$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.186$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.214$)
IMTKU-run2	BL-popularity	($p = 0.0174, ES_{E1} = 0.413$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.688$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.179$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.207$)
RSLNV-run2	BL-popularity	($p = 0.0207, ES_{E1} = 0.405$)
	NKUST-run1	($p < 0.0001, ES_{E1} = 1.680$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.172$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.199$)
TUA1-run0	NKUST-run1	($p < 0.0001, ES_{E1} = 1.506$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 2.997$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.025$)
TMUDS-run2	NKUST-run1	($p < 0.0001, ES_{E1} = 1.433$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 2.924$)

Continued on next page

Table 21 – continued from previous page

Run	significantly better than these runs	
	BL-uniform	($p < 0.0001, ES_{E1} = 2.952$)
TMUDS-run1	NKUST-run1	($p < 0.0001, ES_{E1} = 1.426$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 2.918$)
	BL-uniform	($p < 0.0001, ES_{E1} = 2.946$)
TMUDS-run0	NKUST-run1	($p < 0.0001, ES_{E1} = 1.371$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 2.862$)
	BL-uniform	($p < 0.0001, ES_{E1} = 2.890$)
BL-popularity	NKUST-run1	($p < 0.0001, ES_{E1} = 1.275$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 2.766$)
	BL-uniform	($p < 0.0001, ES_{E1} = 2.794$)
NKUST-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 1.491$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.519$)

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	
IMTKU-run0	BL-popularity	($p = 0.0360, ES_{E1} = 0.202$)
	SKYMN-run1	($p = 0.0114, ES_{E1} = 0.221$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.891$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.064$)
BL-lstm	NKUST-run0	($p < 0.0001, ES_{E1} = 0.740$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.913$)
RSLNV-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.728$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.902$)
SKYMN-run2	NKUST-run0	($p < 0.0001, ES_{E1} = 0.723$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.897$)
SKYMN-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.705$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.879$)
BL-popularity	NKUST-run0	($p < 0.0001, ES_{E1} = 0.689$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.862$)
SKYMN-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.670$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.843$)

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	
IMTKU-run0	SKYMN-run2	($p = 0.0210, ES_{E1} = 0.188$)
	SKYMN-run0	($p = 0.0002, ES_{E1} = 0.241$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.243$)
	SKYMN-run1	($p < 0.0001, ES_{E1} = 0.315$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.448$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.531$)
BL-lstm	SKYMN-run0	($p = 0.0420, ES_{E1} = 0.176$)
	BL-popularity	($p = 0.0371, ES_{E1} = 0.178$)
	SKYMN-run1	($p < 0.0001, ES_{E1} = 0.250$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.383$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.466$)
RSLNV-run0	SKYMN-run1	($p = 0.0031, ES_{E1} = 0.214$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.347$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.430$)
SKYMN-run2	BL-uniform	($p < 0.0001, ES_{E1} = 0.260$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.344$)
SKYMN-run0	BL-uniform	($p = 0.0052, ES_{E1} = 0.207$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.291$)
BL-popularity	BL-uniform	($p = 0.0061, ES_{E1} = 0.205$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.288$)
SKYMN-run1	NKUST-run0	($p = 0.0025, ES_{E1} = 0.217$)

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	
IMTKU-run0	BL-popularity	($p = 0.0401, ES_{E1} = 0.218$)
	RSLNV-run0	($p = 0.0215, ES_{E1} = 0.231$)
	SKYMN-run2	($p = 0.0095, ES_{E1} = 0.248$)
	SKYMN-run0	($p = 0.0037, ES_{E1} = 0.262$)
	SKYMN-run1	($p < 0.0001, ES_{E1} = 0.301$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 1.078$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.419$)
BL-lstm	NKUST-run0	($p < 0.0001, ES_{E1} = 0.893$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.234$)
BL-popularity	NKUST-run0	($p < 0.0001, ES_{E1} = 0.861$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.202$)
RSLNV-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.847$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.188$)
SKYMN-run2	NKUST-run0	($p < 0.0001, ES_{E1} = 0.831$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.172$)
SKYMN-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.817$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.158$)
SKYMN-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.777$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.118$)
NKUST-run0	BL-uniform	($p < 0.0001, ES_{E1} = 0.341$)

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	
IMTKU-run0	BL-lstm	($p = 0.0080, ES_{E1} = 0.212$)
	RSLNV-run0	($p = 0.0003, ES_{E1} = 0.268$)
	SKYMN-run2	($p = 0.0002, ES_{E1} = 0.276$)
	SKYMN-run0	($p < 0.0001, ES_{E1} = 0.320$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.383$)
	SKYMN-run1	($p < 0.0001, ES_{E1} = 0.399$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.721$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.889$)
BL-lstm	SKYMN-run1	($p = 0.0396, ES_{E1} = 0.187$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.508$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.677$)
RSLNV-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.453$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.621$)
SKYMN-run2	NKUST-run0	($p < 0.0001, ES_{E1} = 0.445$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.613$)
SKYMN-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.401$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.569$)
BL-popularity	NKUST-run0	($p < 0.0001, ES_{E1} = 0.338$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.506$)
SKYMN-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.321$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.490$)

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	
IMTKU-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.662$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.877$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.052$)
BL-lstm	BL-popularity	($p < 0.0001, ES_{E1} = 0.630$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.846$)
	BL-uniform	($p < 0.0001, ES_{E1} = 1.021$)
SKYMN-run2	BL-popularity	($p < 0.0001, ES_{E1} = 0.544$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.759$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.934$)
SKYMN-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.543$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.758$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.933$)
SKYMN-run1	BL-popularity	($p < 0.0001, ES_{E1} = 0.519$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.734$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.909$)
RSLNV-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.505$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.721$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.896$)
BL-popularity	NKUST-run0	($p = 0.0378, ES_{E1} = 0.215$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.390$)

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	
IMTKU-run0	SKYMN-run1	($p = 0.0265, ES_{E1} = 0.229$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 0.733$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.952$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.186$)
BL-lstm	NKUST-run0	($p < 0.0001, ES_{E1} = 0.696$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.915$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.149$)
SKYMN-run2	NKUST-run0	($p < 0.0001, ES_{E1} = 0.576$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.795$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.029$)
RSLNV-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.568$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.788$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.022$)
SKYMN-run0	NKUST-run0	($p < 0.0001, ES_{E1} = 0.552$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.771$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.005$)
SKYMN-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 0.504$)
	BL-uniform	($p < 0.0001, ES_{E1} = 0.723$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.957$)
NKUST-run0	BL-uniform	($p = 0.0399, ES_{E1} = 0.220$)
	BL-popularity	($p < 0.0001, ES_{E1} = 0.454$)
BL-uniform	BL-popularity	($p = 0.0210, ES_{E1} = 0.234$)

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

Run	significantly better than these runs	
IMTKU-run0	RSLNV-run1	($p = 0.0366, ES_{E1} = 0.491$)
	BL-popularity	($p < 0.0001, ES_{E1} = 1.033$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.741$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.260$)
RSLNV-run0	BL-popularity	($p < 0.0001, ES_{E1} = 0.970$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.677$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.197$)
IMTKU-run2	BL-popularity	($p < 0.0001, ES_{E1} = 0.945$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.653$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.172$)
BL-lstm	BL-popularity	($p < 0.0001, ES_{E1} = 0.936$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.644$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.163$)
IMTKU-run1	BL-popularity	($p < 0.0001, ES_{E1} = 0.891$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.598$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 4.118$)
RSLNV-run1	BL-popularity	($p = 0.0121, ES_{E1} = 0.542$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.250$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.769$)
BL-popularity	BL-uniform	($p < 0.0001, ES_{E1} = 2.708$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.227$)
BL-uniform	NKUST-run0	($p = 0.0215, ES_{E1} = 0.519$)

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

Run	significantly better than these runs	
IMTKU-run0	BL-popularity	($p = 0.0042, ES_{E1} = 0.507$)
	RSLNV-run1	($p = 0.0002, ES_{E1} = 0.609$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.399$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.424$)
RSLNV-run0	BL-popularity	($p = 0.0319, ES_{E1} = 0.433$)
	RSLNV-run1	($p = 0.0020, ES_{E1} = 0.535$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.325$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.350$)
IMTKU-run2	BL-popularity	($p = 0.0320, ES_{E1} = 0.432$)
	RSLNV-run1	($p = 0.0020, ES_{E1} = 0.535$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.325$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.350$)
BL-lstm	RSLNV-run1	($p = 0.0055, ES_{E1} = 0.496$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.286$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.311$)
IMTKU-run1	RSLNV-run1	($p = 0.0133, ES_{E1} = 0.466$)
	NKUST-run0	($p < 0.0001, ES_{E1} = 3.255$)
	BL-uniform	($p < 0.0001, ES_{E1} = 3.281$)
BL-popularity	NKUST-run0	($p < 0.0001, ES_{E1} = 2.892$)
	BL-uniform	($p < 0.0001, ES_{E1} = 2.918$)
RSLNV-run1	NKUST-run0	($p < 0.0001, ES_{E1} = 2.790$)
	BL-uniform	($p < 0.0001, ES_{E1} = 2.815$)