Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task

Tetsuya Sakai, Sijie Tao, Zhaohao Zeng
Waseda University, Japan
tetsuyasakai@acm.org

Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu
Tsinghua University, P.R.C.
yiqunliu@tsinghua.edu.cn

Maria Maistro
University of Copenhagen, Denmark
mm@di.ku.dk

Zhicheng Dou
Renmin University of China, P.R.C.
dou@ruc.edu.cn

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

ABSTRACT

This is an overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) task. The task features the Chinese subtask (adhoc web search) and the English subtask (adhoc web search, replicability and reproducibility), and received 48 runs from 9 teams. We describe the subtasks, data, evaluation measures, and the official evaluation results.

1 INTRODUCTION

This paper presents an overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task.[1] The task is basically adhoc web search (i.e., ranked retrievel of web pages for a given search topic); it features Chinese and English subtasks. The goal of the task is to quantify technological advances in web search (in terms of the quality of the Search Engine Result Page "above the fold"), and to address the problems of replicability (whether a result reported by Group X can also be obtained by Group Y on the same data) and reproducibility (whether a result reported by Group X can be obtained by Group Y on some different data) in the same context.

The We Want Web task (WWW-1) was launched at NTCIR-13 in 2017 [8], in response to the termination of the web track at TREC 2014.[2] WWW-1 received 32 runs from five teams. The NTCIR-14 WWW-2 task was held in 2019 [9], and received 31 runs from five teams. Also, NTCIR-14 hosted the first CENTRE (CLEF/NTCIR/-TREC Rreproducibility) task [15], which is a "metatask" that spans CLEF, TREC, and NTCIR [4–6, 19]. As NTCIR-14 CENTRE attracted only one participating team, for NTCIR-15 the WWW and CENTRE joined forces, to organise the We Want Web with CENTRE (WWW-3) task.

The WWW-3 Chinese subtask is a traditional ad hoc web search task. The WWW-3 English subtask features the CENTRE replicability and reproducibility experiments in addition to traditional ad hoc web search. Both subtasks have 80 new topics. All participating runs were required to process 160 topics (80 topics from WWW-2 and the new 80 topics) so that, at least in the English subtask, replicability and reproducibility can be studied. At the NTCIR-14 WWW-2 Task, we decided to create 80 topics based on *topic set size design* [11, 13], using the residual variances of the evaluation measures from the NTCIR-13 WWW-1 task. More specifically, according to the largest variance obtained from WWW-1 (which represents the most statistically unstable measure among the ones we use, namely, normalised Expected Reciprocal Rank), having 80 topic was found to be sufficient for ensuring 80% statistical power for comparing any pair of systems using the t-test at the 5% significance level [9, Table 3]. For the NTCIR-15 WWW-3 Task, we decided to create another 80 topics to ensure the same statistical power.

Table 1: WWW-3 timeline (time zone: UTC+9).

March 25, 2020	Topics released
April 13, 2020	Baseline runs released
May 31, 2020	Run submissions due
June-August 2020	Relevance assessments
August 31, 2020	Evaluation results released

Table 2: WWW-3 run statistics.

Team	Chinese	English	Total
ESTUCeng	-	3	3
KASYS	-	5	5
NAUIR	-	5	5
RUCIR	5	5	10
SLWWW	-	5	5
THUIR	5	5	10
Technion	-	5	5
baseline	1	1	2
mpii	-	3	3
total	11 (3 teams)	37 (9 teams)	48 (9 teams)

Table 1 shows the timeline of the WWW-3 task. Table 2 shows names of the participating teams and the number of runs submitted to the task.

The remainder of this paper is organised as follows. Sections 2 and 3 describe the Chinese and English subtasks, respectively. Sections 4 and 5 describe the evaluation measures we use to quantify retrieval effectiveness and replicability/reproducibility, respectively. Sections 6 and 7 report on the Chinese and English subtask results in terms of retrieval effectiveness. Section 8 discusses the

[1] http://sakailab.com/www3/
[2] https://twitter.com/djoerd/status/536128465276530688

Table 3: Chinese query set (Int. indicates the intent types: we only point out the navigational queries while the remaining ones are informational or transactional; Trans. indicates whether the query is translated to English and shared by the English subtask)

qid	Query	Int.	Trans.	qid	Query	Int.	Trans.	qid	Query	Int.	Trans.
0001	全球军力排名		Y	0028	身份证号码			0055	办理深圳户口		
0002	哔哩哔哩	NAV		0029	携程网	NAV		0056	丰田卡罗拉		Y
0003	小猪佩奇		Y	0030	中秋节			0057	防水密封胶		
0004	奔跑吧兄弟			0031	成龙电影大全			0058	多点触摸屏		
0005	易烊千玺			0032	老友记		Y	0059	挖泥船		Y
0006	快递哪家好			0033	奥迪		Y	0060	大红袍的价格		
0007	韵达快递单号查询			0034	肩周炎		Y	0061	大学生创业项目		
0008	如何治疗癫痫		Y	0035	网络游戏热门排行榜			0062	沃尔沃		Y
0009	阿里妈妈			0036	怎样快速减肥		Y	0063	秋季养生粥		
0010	冷饮店加盟			0037	教师资格证			0064	太极拳		
0011	UC 浏览器			0038	智能家居		Y	0065	龙的传人		
0012	魅族官网	NAV		0039	房屋租赁合同		Y	0066	糖尿病治疗		Y
0013	西游记			0040	甲骨文			0067	霸王别姬		
0014	腾讯视频官网	NAV		0041	电信宽带			0068	毕业生自我评价		
0015	千与千寻		Y	0042	九价 hpv 疫苗			0069	纸的由来		Y
0016	股票入门		Y	0043	小美人鱼		Y	0070	邢台市邮编		
0017	武汉大学			0044	教师节的来历		Y	0071	三国无双		
0018	精灵宝可梦		Y	0045	人脸识别		Y	0072	查询驾驶证扣分记录		
0019	呼伦贝尔旅游攻略			0046	治疗癌症的药		Y	0073	红薯怎么水养		
0020	观后感怎么写			0047	燕窝的作用		Y	0074	李四光简介		
0021	烟台特色小吃			0048	左旋肉碱的危害			0075	杭州有几个区		
0022	美国移民		Y	0049	烟雾净化器			0076	眼冒金星是什么原因		
0023	住房公积金贷款			0050	时间校准		Y	0077	枇杷叶的功效与作用		
0024	民间小调			0051	人工智能		Y	0078	关于动物的电影		Y
0025	大连理工大学			0052	加勒比海盗		Y	0079	互联网的利与弊		Y
0026	澳门金沙酒店			0053	美元汇率			0080	阿胶糕的做法		
0027	德州扑克		Y	0054	wifi 密码怎么改		Y				

CENTRE replicability/reproducibilty aspects of the English subtask results. Finally, Section 9 concludes this paper.

2 CHINESE SUBTASK

2.1 Topics

As we did in the previous WWW tasks, we sampled 80 queries for WWW-3 Chinese subtask from Sogou's query logs in one day of August 2018. Among these 80 queries, 54 are torso queries, whose frequencies are between 10 to 1,000 per day. 13 queries are tail queries which appeared less than 10 times in one day's log and the remaining 13 queries are hot queries which have a frequency larger than 1,000. Compared to the previous WWW task, we selected fewer navigational queries and factoid questions (4 in WWW-3 v.s. 9 in WWW-2). Also, we included more torso queries in the topic set because we believed that they are most appropriate for an ad hoc task. The content of the queries, the intent types (navigational/information & transactional), and whether the queries are shared by English subtask are presented in Table 3.

2.2 Target Document Collection

Following the previous WWW tasks, we adopted SogouT-16 as the document collection. SogouT-16 contains about 1.17 billion Web pages, which were sampled from the index of Sogou.com, the second largest commercial search engine in China. Considering that the original SogouT might be difficult to handle for some research groups (almost 80TB after decompression), we prepared a "Category B" version of SogouT-16, which is denoted as "SogouT16 B". This subset contains about 15% of the webpages of SogouT-16. This Chinese corpus is free for research purpose[3].

[3]http://tiangong.sogou.com/dataset

2.3 Additional Materials

2.3.1 SogouQCL[21]. Besides the target document collection, we provide an extra training set, Sogou-QCL, in this round of WWW-3 Chinese subtask. The Sogou-QCL contains two kinds of training sets:

- The first set contains traditional relevance assessments for 1,000 Chinese queries, and for each query, Sogou-QCL contains about 20 query-doc relevance judgments. Each pair was annotated by three trained assessors. Sogou-QCL also provides the title and content extracted from raw HTMLs for each document.
- The second set consists of click labels generated by click models. Releasing the original click logs could possibly harm user's privacy because it may contain personally identifiable information. Therefore we provide the relevance scores estimated based on the behavior of a large group of users. More specifically, for each query-doc pair, we provide five kinds of weak relevance label computed by five popular click models: UBM, DBN, TCM, PSCM, and TACM. These click models utilize rich user behavior in search logs, such as click, skip, and dwell time, to estimate the relevance of the query-document pairs in click logs. Sogou-QCL contains more than half a million queries and more than 9 millions of documents. To the best of our knowledge, this is so far the largest free training collection for Chinese ranking problems.

Handling the raw HTML content can be difficult. Therefore, we also provide the extracted content, including the title and text content of each document, with professional tools of Sogou.com. We hope it will reduce the effort for the participants and help them focus on the design of ranking models.

2.3.2 Topics and Qrels from WWW-1 and WWW-2. The Chinese topics and qrels from WWW-1 and WWW-2 are accessible to all

Table 4: Relevance assessment statistics for Chinese qrels

	WWW-2 qrels	WWW-3 qrels
#topics	80	80
#assessors/topic	3	3
Pool depth	20	30
Total #docs pooled	12,271	11,712
Total L3-relevant	1,961	784
Total L2-relevant	1,524	1,427
Total L1-relevant	2,401	2,389
Total L0	6,385	7,112

participants. Additionally, the participants can use the qrels from previous WWW tasks as the validation set for their models. Everyone can visit http://www.thuir.cn/ntcirwww3/[21] to download them.

2.3.3 Baseline Run. We provided a basic BM25 baseline run to all the participants. This baseline run was generated by our own free online retrieval/page rendering service. The online retrieval system is based on Solr[4], with the default parameter settings.

2.4 Chinese Run Type

Unlike the English subtask, replicability and reproducibility were not specifically addressed in the Chinese subask. Hence, standard ad hoc retrieval runs were submitted to the subtasks. These runs are called NEW runs, to distinguish them from the runs submitted to address replicability and reproducibility (See Section 3.4).

2.5 Relevance Assessments

The Chinese relevance assessments were organized by Tsinghua University. We used a pooling depth of 30, and collected relevance assessments for the 80 new WWW-3 topics. In total, we collected relevance assessments for 11,712 query-document pairs.

We contacted an annotation company named 小牛雅智. The relevance assessments were conducted in their company from June 12th to July 20th, 2020. For each query in the query set, we provided a task description field in addition to the query content field, to help the assessors understand the search intent more specifically. We also provided the relevance assessment criteria to them, which are the same as the ones used in WWW-2. The specific criteria are shown as follows:

GARBLED Garbled - The HTML page is shown to user with the *garbled* state.

NONREL Nonrelevant - It is *unlikely* that the user who entered this search query will find this page relevant.

MARGREL Marginally relevant - the user will get some relevant information from this page. However, she needs to browse more pages to satisfy her information needs.

REL Relevant - it is *possible* that the user who entered this search query will find this page relevant.

HIGHREL Highly relevant - it is *likely* that the user who entered this search query will find this page relevant.

[4]http://lucene.apache.org/solr/

Finally, NONREL and GARBLED labels were mapped to zero; MARGREL labels were mapped to one; REL labels were mapped to two and HIGHREL labels were mapped to three for defining the gain values. Each query-document pair was annotated by 3 assessors. When using an ordinal metric difference function, the Krippendorff's α of the 4-level relevance annotation is 0.8128, which indicates a high agreement between the different assessors. Table 4 shows the statistics for the qrels of WWW-2 and WWW-3 topics.

3 ENGLISH SUBTASK

3.1 Topics

We followed the approach we took at the NTCIR-14 WWW-2 task for constructing the 80 new English topics. That is, we manually translated 30 of the WWW-3 Chinese topics into English, and sampled 50 topics from the AOL query log. Participants were asked to process both the 80 WWW-2 and 80 WWW-3 topics, available at https://waseda.box.com/www2www3topics-E. Just like the Chinese topics, each topic has a qid (or "topic ID"), a content (or "title") and a description.

3.2 Target Document Collection

Following WWW-1 and WWW-2, we used as our target corpus clueweb12-B13, which contains about 50 million web pages.[5]

3.3 Additional Materials

3.3.1 Topics and Qrels from WWW-1 and WWW-2. English topics and qrels from WWW-1 and WWW-2 were made available to participants for tuning their systems.[6] For the WWW-1 English subtask, two versions of qrels exist: the original qrels from the NTCIR-13 WWW-1 task [8], and the updated version from the NTCIR-14 CENTRE task [15]. They were both made publicly available.

3.3.2 Baseline Run. We also provided a vanilla BM25 baseline run (with the HTML files of the retrieved documents) to participants, by using the ClueWeb12 Batch Service[7]. This gives them the option to participate by just reranking the top 1,000 documents in some way, without indexing the clueweb12-B13 corpus.

3.4 English Run Types

We had three run types in the English subtask, although not all of the runs submitted actually follow our run naming schemes.

3.4.1 NEW runs. These are the usual ad hoc runs, based on whatever algorithms the participants chose to experiment with.

3.4.2 REV (revived) runs. In the CENTRE effort within the WWW-3 English subtask, the target runs for replicability and reproducibility were **THUIR-E-CO-MAN-Base-2** (A-run, based on LambdaMART, representing the state-of-the-art from WWW-2) and THUIR-E-CO-PU-Base-4 (B-run, the baseline run based on BM25). The main research question for this round of CENTRE was: *can the gain of LambdaMART over BM25 as reported by THUIR (Tsinghua University) at WWW-2 be successfully replicated and reproduced*?

[5]http://lemurproject.org/clueweb12/

[6]http://sakailab.com/www3english/

[7]http://boston.lti.cs.cmu.edu/Services/clueweb12_batch/

Table 5: Distribution of pooled documents over the relevance levels in the English qrels.

relevance level	original WWW-2 qrels (pool depth: 50)	new WWW-2 qrels (pool depth: 15)	WWW-3 qrels (pool depth: 15)	total
L0	13,305	2,820	4,116	6,936
L1	6,469	3,952	4,001	7,953
L2	4,664	5,342	5,513	10,855
L3	2,332	3,569	3,030	6,599
L4	857	15	17	32
total	27,627	15,698	16,677	32,375

We asked THUIR to provide a pair of REV (revived) runs: these were meant to represent exactly the same systems as the original A-run and B-run, so that the only difference is that the REV runs cover not only the WWW-2 topics but also the new WWW-3 topics. However, as it turned out, THUIR could not provide such runs for WWW-3: their new LambdaMART run (THUIR-E-CO-REV-2) substantially underperforms **THUIR-E-CO-MAN-Base-2** when compared on the WWW-2 topic set, regardless of which version of the qrels is used (see Section 3.5). Similarly, their new BM25 run (THUIR-E-CO-REP-5) substantially underperforms THUIR-E-CO-PU-Base-4 when evaluated with the official WWW-2 qrels.

In summary, our reliance on the REV runs was unsuccessful this time.

3.4.3 REP (replicated/reproduced) runs. These runs specifically address the CENTRE questions by trying to replicate/reproduce the original A-run and B-run. We shall refer to these as REP A-run and B-runs, and evaluate them from the following viewpoints.

- **Replicability of ordering of documents** Can another group replicate the exact ranking of documents?
- **Replicability of absolute per-topic effectiveness** Can the per-topic scores of the original run (A-run or B-run) be replicated by another group?
- **Replicability with statistical testing** Are the per-topic scores of the original and replicated runs significantly different?
- **Replicability of an effect over a baseline** Can the deltas of per-topic scores between the original A-run and B-run be replicated by another group? Can the sample effect size (i.e., the overall gain of the A-run over the B-run) be replicated by another group?
- **Reproducibility with statistical testing** Are the per-topic scores of the original and reproduced run significantly different?
- **Reproducibility of an effect over a baseline** Can the sample effect size be reproduced by another group on the new WWW-3 test collection?

Section 5 describes how we quantify these aspects of the REP runs.

3.5 Relevance Assessments

Recall that all WWW-3 runs processed the 160 topics: 80 from WWW-2 and another 80 new WWW-3 topics. Although the qrels are already available for the WWW-2 topics, we collected new relevance assessments for all 160 topics, where each topic was independently assessed by eight assessors, as described below.

Two depth-15 pool files were prepared for each topic: one is the usual pool file created with NTCIRPOOL[8]: the pooled documents were sorted by the number of runs that returned that document (first key, descending order) and the sum of ranks of that document in those runs (second key, ascending order). This approach aims to present "popular" documents to the assessors first, and has been used in many NTCIR tasks since the NTCIR-7 ACLIA IR4QA task first implemented this particular method [16]. The other type of pool file follows the TREC approach of randomising the order of the pooled documents. These two types of pool files (prioritised and randomised) were used for the purpose of a new study (outside the scope of this overview paper), which follows up on the work of Sakai and Xiao [17]: their research questions were: (1) Which type of pool files enables more efficient assessments? (2) Which enables higher inter-assessor agreements?

A total of 24 assessors were hired from the international course of the computer science and communications engineering department, Waseda University, Japan.[9] 14 of them were master students and the other 10 were undergraduates. For each topic, 4 assessors processed the prioritised pools, and another 4 assessors processed the randomised pools. Having 160 topics, each with 8 assessors as described above, satisfies the sample size requirements given in Sakai and Xiao [17]: the above two research questions shall be addressed elsewhere. For providing the official results for WWW-3, we simply consolidate all eight sets of assessments for each topic to form graded relevance assessments. As each set of assessments contains labels on a 3-point scale (highly relevant (2), relevant (1), nonrelevant/error (0) [8][10]), we obtained the *official WWW-3 English qrels* with a 5-point scale (L0-L4) by taking the integer part of $\log_2(S+1)$, where S is the sum of the eight raw labels. We also obtained the new WWW-3 version of the qrels for the WWW-2 topics using the same procedure.

The total number of depth-15 pooled documents (topic-document pairs) was 32,375. Hence, the total number of judgements amounted to $8 * 32,375 = 259,000$. Table 5 shows the distribution of pooled documents over the relevance levels, for the original WWW-2 qrels, the new WWW-2 qrels, and the WWW-3 qrels. While the original WWW-2 qrels used depth-50 pools based on 20 runs from 5 teams [9], the WWW-3 version used depth-15 pools based on 37 runs from 9 teams.

[8] http://research.nii.ac.jp/ntcir/tools/ntcirpool-en.html

[9] One student is actually from the Japanese course, but is a Chinese student who is proficient in English.

[10] Of the 259,000 raw assessor labels, 28,144 were highly relevant (2), 61,512 were relevant (1), 163,090 were nonrelevant (0), and 6,254 were error (0).

4 RANKED RETRIEVAL EVALUATION MEASURES

Following the previous rounds of We Want Web [8, 9], we use nDCG@10 (MSnDCG@10), Q@10, and nERR@10 [12] to quantify retrieval effectiveness, using the NTCIREVAL tool with a linear gain value setting.[11] In addition, we compute *intentwise Rank-Biased Utility* (iRBU) introduced by Sakai and Zeng [18], as nDCG and iRBU (with the transition probability parameter $p = 0.99$) were the top two measures in terms of agreement with users' SERP preferences in their experiment. iRBU is based on ideas from *Rank-Biased Utility* (RBU) [1], *Rank-Biased Precison* [10], and *Expected Reciprocal Rank* (ERR) [3].

iRBU at cutoff $l(= 10)$ is defined as follows. Let $g(r)$ denote the gain value of a document at rank r, and let $\mathrm{gv}_{\max}$ denote the maximum gain value for the entire test collection (under a linear gain value setting, $\mathrm{gv}_{\max}$ is equal to the highest relevance level for the collection). The probability that the user is satisfied with the document at rank r is obtained as:

$$P_{\mathrm{sat}}(r) = \frac{g(r)}{\mathrm{gv}_{\max} + 1} \,. \tag{1}$$

Hence, the probability that the user will reach as far as rank r and finally gets satisfied is obtained as:

$$P_{\mathrm{ERR}}(r) = P_{\mathrm{sat}}(r) \prod_{k=1}^{r-1} (1 - P_{\mathrm{sat}}(k)) \,. \tag{2}$$

iRBU simply replaces the reciprocal rank in the formula of ERR with p^r ($p = 0.99$), which is a function of the user effort for examining the ranked list of size r.

$$iRBU@l = \sum_{r=1}^{l} P_{\mathrm{ERR}}(r) p^r \,. \tag{3}$$

5 CENTRE EVALUATION MEASURES

In this round of CENTRE, we quantify replicability and reproducibility as suggested in Breuer et al. [2]. Details are presented in the following.

5.1 Replicability

5.1.1 Replicating the Ordering of Documents. First, we evaluate whether the replicating run can retrieve the same exact ranking of documents retrieved by the original run (i.e. A-run or B-run). Note that the following measures will be instantiated with the A-run as an example.

We compute Kandall's τ union [5, 6], which compares the relative order of documents by computing Kendall's τ with respect to the union of the original and replicated rankings. Observe that this is necessary since Kendall's τ is defined for permutations of items from the same list [7], while replicated runs can rank documents that were not retrieved by the original run.

Let A be the original A-run and A' the replicated A-run. A_j denotes the ranked list of document ids for topic j for the original A-run and similarly A'_j is the ranked list of documents for the REP A-run. Kendall's τ union is computed as follows:

(1) consider the union of A_j and A'_j by removing duplicate entries;
(2) consider the rank position of documents from the union in A_j and A'_j;
(3) compute Kendall's τ between these two lists of rank positions.

Kendall's τ at step 3 is computed as follows:

$$\tau_j(r, r') = \frac{P - Q}{\sqrt{(P + Q + U)(P + Q + V)}} \tag{4}$$

where r and r' are the list of rank positions obtained at step 2, P is the total number of concordant pairs, Q the total number of discordant pairs, U and V are the number of ties, in r and r' respectively.

As reported in previous work [2, 5, 6], Kendall's τ can be too strict when comparing 2 list of documents and is not top heavy. Therefore, in addition to Kendall's τ we also compute Rank-Biased Overlap (RBO) [20]. RBO for the j-th topic is computed as follows:

$$\mathrm{RBO}_j(A, A') = (1 - \phi) \sum_{i=1}^{\infty} \phi^{i-1} \cdot O_i \tag{5}$$

where $\phi \in [0, 1]$ is a parameter to adjust the measure top-heaviness: the smaller ϕ, the more top-weighted the measure; and O_i is the proportion of overlap up to rank i, which is defined as the cardinality of the intersection between A_j and A'_j up to i divided by i. Therefore, RBO accounts for the overlap of two rankings and discounts the overlap while moving towards the end of the ranking, since it is more likely for two rankings to have a greater overlap when many rank positions are considered.

5.1.2 Replicating Absolute Per-Topic Effectiveness. Consider the problem of replicating the original A-run (i.e., state-of-the-art) from NTCIR-14 WWW-2. Individual REP A-runs can be evaluated in terms of whether per-topic effectiveness scores are faithful to the original ones as follows. Let $M_j^C(A)$ denote the effectiveness score for topic j of the original A-run with test collection C (i.e., the WWW-2 test collection). Similarly, let $M_j^C(A')$ denote the corresponding score of a REP A-run. Following CENTRE@CLEF [5, 6], the root mean square error for replicating absolute per-topic differences is computed as follows.

$$RMSE_{\mathrm{abs}} = \sqrt{\frac{1}{n_C} \sum_{j=1}^{n_C} (M_j^C(A') - M_j^C(A))^2} \,, \tag{6}$$

where n_C is the topic set size of C. Note that $RMSE_{\mathrm{abs}}$ focuses on the per-topic measure scores rather than the actual documents retrieved.

5.1.3 Replicating with t-Test. We compare the original and replicated runs from a statistical point of view [2]. We run a two tailed paired t-test between $M_j^C(A)$ and $M_j^C(A')$. The p-value returned by the t-test informs on the success of the replicability experiment: the smaller the p-value, the stronger the evidence that A and A' are statistically significantly different.

5.1.4 Replicating an Effect over a Baseline. Consider the problem of replicating the *effect* of the A-run from WWW-2 over the B-run (i.e., baseline) which is also from WWW-2. To be consistent

[11] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html

with the previous notations, we denote the score of a REP B-run by $M_j^C(B')$, and so on. Following the NTCIR-14 CENTRE approach, we first compute the effectiveness score *deltas* as follows [15]:

$$\Delta M_j^C = M_j^C(A) - M_j^C(B)\,, \quad \Delta' M_j^C = M_j^C(A') - M_j^C(B')\,. \tag{7}$$

Then the root mean squared error for the score deltas, which reflects the faithfulness of the per-topic deltas, can be computed as follows.

$$RMSE_\Delta = \sqrt{\frac{1}{n_C}\sum_{j=1}^{n_C}(\Delta' M_j^C - \Delta M_j^C)^2}\,. \tag{8}$$

Also following NTCIR-14 CENTRE, we also evaluate replicability based on the sample effect sizes in terms of the Effect Ratio (ER):

$$\mathrm{ER}_{\mathrm{repli}} = \frac{\frac{1}{n_C}\sum_{j=1}^{n_C}\Delta' M_j^C}{\frac{1}{n_C}\sum_{j=1}^{n_C}\Delta M_j^C} = \frac{\sum_{j=1}^{n_C}\Delta' M_j^C}{\sum_{j=1}^{n_C}\Delta M_j^C}\,. \tag{9}$$

The implications of ER scores are as follows [15].

- $\mathrm{ER} \le 0$**:** The REP A-run failed to outperform the REP B-run and therefore the effect replication is a complete failure;
- $0 < \mathrm{ER} < 1$**:** The effect replication is somewhat successful, but the replicated effect is smaller compared to the original effect;
- $\mathrm{ER} = 1$**:** The effect replication is perfect in the sense that the original effect has been recovered as is;
- $\mathrm{ER} > 1$**:** The effect replication is successful, and the replicated effect is larger compared to the original effect.

ER considers only relative improvements. For example, if the mean scores of the original A-run and B-run are 1.0 and 0.9, respectively, and if the mean scores of the REP A-run and B-run are 0.2 and 0.1, respectively, then $\sum_{j=1}^{n_C}\Delta' M_j^C/n_C = \sum_{j=1}^{n_C}\Delta M_j^C/n_C = 0.1$ and therefore $\mathrm{ER}_{\mathrm{repli}} = 1$. However, these REP runs are clearly not what we want. Hence, to take into account absolute scores, we emply *Delta Relative Improvement* [2] to complement ER:

$$\Delta RI_{\mathrm{repli}} = \frac{\frac{1}{n_C}\sum_{j=1}^{n_C}\Delta M_j^C}{\frac{1}{n_C}\sum_{j=1}^{n_C}M_j^C(B)} - \frac{\frac{1}{n_C}\sum_{j=1}^{n_C}\Delta' M_j^C}{\frac{1}{n_C}\sum_{j=1}^{n_C}M_j^C(B')} \tag{10}$$

$$= \frac{\sum_{j=1}^{n_C}\Delta M_j^C}{\sum_{j=1}^{n_C}M_j^C(B)} - \frac{\sum_{j=1}^{n_C}\Delta' M_j^C}{\sum_{j=1}^{n_C}M_j^C(B')\,.} \tag{11}$$

The implications of ΔRI scores are as follows.

- $-1 \le \Delta RI \le 0$**:** The relative improvement of the REP A-run over the REP B-run is larger than the original relative improvement;
- $\Delta RI = 0$**:** The relative improvements are the same;
- $0 \le \Delta RI \le 1$**:** The relative improvement of the REP A-run over the REP B-run is smaller than the original relative improvement.

As described in Breuer *et al.* [2], we plot ΔRI against *ER*. From the above, it is clear that if $\mathrm{ER} = 1$ and $\Delta RI = 0$, that represents a perfect "clone", and that $\mathrm{ER} > 1$ and $\Delta RI < 0$ represents the most positive result in terms of whether the effect originally observed can be verified by other groups (on the same data).

Table 6: 11 WWW-3 Chinese runs.

Run name	Remarks
RUCIR-C-CD-NEW-1	
RUCIR-C-CD-NEW-2	
RUCIR-C-CD-NEW-3	
RUCIR-C-CD-NEW-4	
RUCIR-C-CO-NEW-5	
THUIR-C-CO-NEW-1	
THUIR-C-CO-NEW-2	
THUIR-C-CO-NEW-3	
THUIR-C-CO-NEW-4	
THUIR-C-CO-REV-5	
baselineChn	organisers' vanilla BM25 baseline

5.2 Reproducibility

5.2.1 Reproducing with t-Test. We compare the original and reproduced runs from a statistical point of view [2]. Instead of a paired t-test, we run a two tailed unpaired t-test between $M_j^C(A)$ and $M_j^C(A')$, since the scores are computed with respect to different topics (WWW-2 and WWW-3). The p-value returned by the t-test informs on the success of the reproducibility experiment: the smaller the p-value, the stronger the evidence that A and A' are statistically significantly different.

5.2.2 Reproducing an Effect over a Baseline. Consider the problem of reproducing the *effect* of the A-run from WWW-2 over the B-run (i.e., baseline) which is also from WWW-2. The new test collection is denoted by D (the WWW-3 test collection in our case), with topic set size n_D. We quantify reproducibility also with ER and DeltaRI.

$$\mathrm{ER}_{\mathrm{repro}} = \frac{\frac{1}{n_D}\sum_{j=1}^{n_D}\Delta' M_j^D}{\frac{1}{n_C}\sum_{j=1}^{n_C}\Delta M_j^C}\,. \tag{12}$$

In our case, $n_C = n_D = 80$ and therefore they cancel out in the above equation.

Since ER considers relative improvements only, we can complement it using Delta Relative Improvement for the case of reproducibility as well.

$$\Delta RI_{\mathrm{repro}} = \frac{\frac{1}{n_C}\sum_{j=1}^{n_C}\Delta M_j^C}{\frac{1}{n_C}\sum_{j=1}^{n_C}M_j^C(B)} - \frac{\frac{1}{n_D}\sum_{j=1}^{n_D}\Delta' M_j^D}{\frac{1}{n_D}\sum_{j=1}^{n_D}M_j^D(B')} \tag{13}$$

$$= \frac{\sum_{j=1}^{n_C}\Delta M_j^C}{\sum_{j=1}^{n_C}M_j^C(B)} - \frac{\sum_{j=1}^{n_D}\Delta' M_j^D}{\sum_{j=1}^{n_D}M_j^D(B')\,.} \tag{14}$$

Finally, ΔRI is plotted against *ER*.

6 CHINESE SUBTASK RESULTS

Table 6 shows the list of 11 runs submitted to the WWW-3 Chinese subtask.

6.1 Official Results

Although the submitted runs contain the ranking results for all 160 queries of the WWW-2 and WWW-3, the evaluation is based on the 80 WWW-3 queries. We removed the query "携程网" (qid:

Table 7: WWW-3 Chinese subtask official results with nDCG and Q (mean over the 80 WWW-3 test topics).

Run name	Mean nDCG	Run name	Mean Q
RUCIR-C-CD-NEW-4	0.5296	RUCIR-C-CD-NEW-4	0.4787
RUCIR-C-CD-NEW-3	0.5136	RUCIR-C-CD-NEW-3	0.4700
RUCIR-C-CD-NEW-1	0.4923	RUCIR-C-CD-NEW-1	0.4510
RUCIR-C-CO-NEW-5	0.4543	RUCIR-C-CO-NEW-5	0.4094
RUCIR-C-CD-NEW-2	0.4314	RUCIR-C-CD-NEW-2	0.3887
THUIR-C-CO-NEW-2	0.4112	THUIR-C-CO-NEW-2	0.3525
THUIR-C-CO-NEW-1	0.4051	THUIR-C-CO-NEW-1	0.3464
THUIR-C-CO-NEW-3	0.3940	THUIR-C-CO-NEW-3	0.3325
baselineChn	0.2848	baselineChn	0.2317
THUIR-C-CO-REV-5	0.2705	THUIR-C-CO-REV-5	0.2093
THUIR-C-CO-NEW-4	0.2329	THUIR-C-CO-NEW-4	0.1728

Table 8: WWW-3 Chinese subtask official results with nERR and iRBU (mean over the 80 WWW-3 test topics).

Run name	Mean nERR	Run name	Mean iRBU
RUCIR-C-CD-NEW-4	0.6442	RUCIR-C-CD-NEW-4	0.8798
RUCIR-C-CD-NEW-3	0.6200	RUCIR-C-CD-NEW-3	0.8621
RUCIR-C-CD-NEW-1	0.6029	RUCIR-C-CO-NEW-5	0.8525
THUIR-C-CO-NEW-2	0.5706	RUCIR-C-CD-NEW-1	0.8299
THUIR-C-CO-NEW-1	0.5489	RUCIR-C-CD-NEW-2	0.8245
RUCIR-C-CO-NEW-5	0.5456	THUIR-C-CO-NEW-2	0.7751
RUCIR-C-CD-NEW-2	0.5412	THUIR-C-CO-NEW-1	0.7493
THUIR-C-CO-NEW-3	0.5169	THUIR-C-CO-NEW-3	0.7356
THUIR-C-CO-REV-5	0.4065	baselineChn	0.6491
baselineChn	0.3800	THUIR-C-CO-REV-5	0.6384
THUIR-C-CO-NEW-4	0.3489	THUIR-C-CO-NEW-4	0.6011

0029) from the evaluation set, because all the pooled documents are nonrelevant.

Tables 7-8 show the mean effectiveness scores for all Chinese runs over the 80 WWW-3 topics. Tables 22 and 23 in the Appendix summarize the Randomized Tukey HSD tests results with $B = 10,000$ trails [13]. The results indicate that RUCIR-C-CD-NEW-4 performs the best among all the runs. It outperforms other runs in four evaluation metrics. The performance of RUCIR-C-CD-NEW-3 and RUCIR-C-CD-NEW-1, is close to the best performing run, RUCIR-C-CD-NEW-4, as the differences in nDCG and Mean Q between these three runs are not statistically significant. In Table 9, we compare the system rankings according to the four evaluation measures in terms of Kendall's τ, as well as their 95% confidence intervals. It can be observed that the nDCG, mean Q and iRBU rankings are very similar, resulting to $\tau > 0.9$, but the nERR behaves relatively differently when compared to other three metrics.

7 ENGLISH SUBTASK RESULTS

Table 10 shows the list of 37 runs submitted to the WWW-3 English subtask. As **baselineEng** simply relied on the ClueWeb12 batch service, its WWW-2 topic set portion is identical to **baseline_eng_v1** at the WWW-2 task [9].

7.1 Official Results

Here we discuss the mean effectiveness scores over the 80 WWW-3 topics. Note that participants had access to the original qrels of the WWW-2 topics and therefore the run performances for these topics do not represent their effectiveness for new topics.

Tables 11-12 show the official WWW-3 English results, averaged over the 80 WWW-3 test topics. Tables 24 and 25 in the Appendix show the accompanying significance test results based on paired Tukey HSD tests with $B = 10,000$ trials [13]. For example, it can be observed from Table 24(a) that in terms of nDCG, KASYS-E-CO-NEW-1, mpii-E-CO-NEW-1, and KASYS-E-CO-NEW-4, are the top performers in that these three runs are the only runs that outperformed as many as 20 runs (20 runs counted from the bottom of Table 11 left column). Q agrees with the above results (Table **??**(b)). In contrast, Table 25(c) shows that mpii-E-CO-NEW-1 is the top performer in terms of significance test results with nERR, which suggests that this run is particularly good at navigational searches. Table 13 shows the SYSDESC (system description) fields [14, p.73] of these three runs.

On the other hand, the high scores shown in Table 11-12 (especially with Q which is the most recall-oriented among the four measures) suggest that the WWW-3 English test collection is *not reusable*, in the sense that it is probably not suitable for evaluating new runs. Many of the document returned by the participating systems were judged relevant, but it is likely that there are many unidentified relevant documents in the target corpus.

Table 14 compares the official system rankings according to different evaluation measures in terms of Kendall's τ. It can be observed that iRBU behaves relatively differently compared to the other three measures. Moreover, by comparing Tables 24 and 25, it can be observed that iRBU has the lowest *discriminative power* [12].

7.2 Official WWW-2 Qrels vs. The New Qrels for the WWW-2 Topics

We are in a unique situation where we have two independently-constructed versions of qrels for the WWW-2 topics: the official WWW-2 qrels constructed by pooling the WWW-2 runs, and the new qrels constructed by pooling the WWW-2 topic set portion of the WWW-3 runs. In this section, we compare the outcome of evaluating the WWW-2 runs using these two versions to see how/whether they are interchangeable. Note that, although the WWW-2 topic set portion of the WWW-3 runs can also be evaluated using these two versions, this is less interesting as the WWW-3 runs were allowed to tune themselves with the official WWW-2 qrels.

Tables 26-28 in the Appendix compare the ranking of the 20 WWW-2 runs based on the official WWW-2 qrels with that based on our new qrels. It can be observed that the rankings are quite different, and the score ranges based on the new qrels are much wider compared to the official results. Table 15 quantifies the discrepancies in terms of Kendall's τ ($n = 20$ runs).

Note that evaluating the WWW-2 runs with our new qrels is similar to evaluating new runs that did not contribute to the pools. The results shown in Tables 26-28 and Table 15 suggest that the WWW test collections are *not reusable*, in the sense that runs that did not contribute to the pools cannot be ranked reliably. (See also the discussion in Section 7.1.) The wide score ranges based on the new qrels probably mean the following. The WWW-2 runs that are given high scores are the ones that are very similar to the WWW-3

Table 9: Run ranking correlations in terms of Kendall's τ with 95%CIs ($n = 11$ Chinese runs evaluated on the WWW-3 topic set).

	Q	nERR	iRBU
nDCG	1.000 [1.000, 1.000]	0.818 [0.579, 0.928]	0.964 [0.906, 0.986]
Q	-	0.818 [0.579, 0.928]	0.964 [0.906, 0.986]
nERR	-	-	0.782 [0.508, 0.912]

Table 10: 37 WWW-3 English runs.

Run name	Remarks
ESTUCeng-E-CO-NEW-{1,2,3}	
KASYS-E-CO-NEW-{1,4,5}	
KASYS-E-CO-REP-2	REP A-run (LambdaMART)
KASYS-E-CO-REP-3	REP B-run (BM25)
NAUIR-E-CO-NEW-{1,2,3,4,5}	
RUCIR-E-CO-NEW-{1,2,3,4,5}	
SLWWW-E-CD-NEW-5	
SLWWW-E-CO-REP-{1,2,3}	
SLWWW-E-CO-REP-4	REP A-run (LambdaMART)
THUIR-E-CO-NEW-4	
THUIR-E-CO-REP-5	
THUIR-E-CO-REV-{1,3}	
THUIR-E-CO-REV-2	(LambdaMART)
Technion-E-CO-NEW-{1,2,3,4,5}	
baselineEng	organisers' vanilla BM25 baseline
mpii-E-CO-NEW-{1,2,3}	

runs: if the documents returned by the WWW-3 runs are judged relevant by the WWW-3 assessors, then the WWW-2 runs that returned similar search results are rated high. In contrast, WWW-2 runs that are not similar to any of the WWW-3 runs are basically completely new runs from the viewpoint of the WWW-3 version of the qrels, and therefore are underrated, due to some unjudged relevant documents in the search results.

8 CENTRE: REPLICABILITY AND REPRODUCIBILITY RESULTS

In the following, we evaluate the two REP A-runs (contributed by KASYS and SLWWW) and the REP B-run (contributed by KASYS) shown in Table 10. Note that for some runs there are missing topics and documents, which affected replicability and reproducibility scores. Specifically, KASYS-E-CO-REP-2 and KASYS-E-CO-REP-3 do not retrieve any document for topic 0044, retrieve just 5 documents for topics 0063 and 0080, 80 documents for topic 0074, and 440 documents for topic 0036. Moreover, KASYS-E-CO-REP-2 do not retrieve any document for topic 0119.

For replicability, we evaluate both WWW-2 and WWW-3 runs with the original qrels from WWW-2. Indeed, Section 7.2 shows that using WWW-3 qrels substantially changes the ranking of runs from WWW-2, while the aim of replicability is to replicate the exact results from WWW-2.

For reproducibility we used WWW-2 for the original runs with WWW-2 topics and WWW-3 qrels for REP runs with WWW-3 topics.

8.1 Replicability

8.1.1 Results: Replicating the Ordering of Documents. Table 16 reports the average Kendall's τ and RBO between the original and replicated runs. All replicability runs have low Kendall's τ and RBO scores, meaning that none of the participating group could replicate the original list of topics. Among A-runs, KASYS-E-CO-REP-2 has the highest Kendall's τ and RBO scores.

8.1.2 Results: Replicating Absolute Per-Topic Effectiveness. Table 17 reports RMSE_{abs} scores, where the lower the score the better the replication outcome. Again, none of the REP runs could successfully replicate the effectiveness scores of the original runs from WWW-2. KASYS-E-CO-REP-2 has the best RMSE_{abs} scores with respect to all the evaluation measures.

It is interesting to observe that RMSE_{abs} scores are always lower for nDCG and Q-measure than nERR and iRBU. As shown by Breuer et al. [2], it is harder to replicate more top heavy measures as nERR and iRBU than more recall oriented measures as nDCG and Q-measure.

8.1.3 Results: Replicating with t-Test. Table 18 reports p-values, resulted by running a two tailed paired t-test between the effectiveness scores of the original run and the replicated run. Accordingly to previous results, SLWWW-E-CO-REP-4 is statistically different from the original A-run with respect to all measures. On the other hand, KASYS-E-CO-REP-3 has very high p-values with respect to all measures, thus we cannot conclude that this run is statistically different from the original B-run. Finally, KASYS-E-CO-REP-2 has p-values lower than 0.05 for all measures except for nERR.

8.1.4 Results: Replicating an Effect over a Baseline. In this section, we evaluate the REP A-run and REP B-run contributed by KASYS in terms how how successfully they replicate the difference between the original A-run and B-run observed at WWW-2, using RMSE_{Δ}, ER_{repli}, and ΔRI_{repli}. Recall that while RMSE_{Δ} examine the per-topic deltas, ER_{repli}, and ΔRI_{repli} examine the overall effect (See Section 5.1.4). As shown in Table 10, KASIS was the only team that submitted a pair of REP runs so we evaluate this pair only.

Table 19 reports replicability scores for RMSE_{Δ}, ER_{repli} and ΔRI_{repli} for all the evaluation measures. RMSE_{Δ} scores are aligned with $RMSE_{\text{abs}}$ scores (see Table 17), showing that KASYS REP runs could not replicate nor the original effectiveness scores neither the effect over a baseline. This is further corroborated by ER_{repli} scores which are close to 0 and lower than 0 for 3 effectiveness measures out of 4. In terms of ΔRI_{repli}, the relative improvement of the REP A-run over the REP B-run is smaller than the original relative improvement, except for iRBU which shows similar relative improvements.

Table 11: WWW-3 English subtask official results with nDCG and Q (mean over the 80 WWW-3 test topics).

Run name	Mean nDCG	Run name	Mean Q
KASYS-E-CO-NEW-1	0.6935	KASYS-E-CO-NEW-1	0.7123
mpii-E-CO-NEW-1	0.6897	KASYS-E-CO-NEW-4	0.7073
KASYS-E-CO-NEW-4	0.6893	mpii-E-CO-NEW-1	0.7016
KASYS-E-CO-NEW-5	0.6812	KASYS-E-CO-NEW-5	0.6990
mpii-E-CO-NEW-2	0.6743	mpii-E-CO-NEW-2	0.6905
Technion-E-CO-NEW-1	0.6581	Technion-E-CO-NEW-1	0.6815
Technion-E-CO-NEW-2	0.6560	Technion-E-CO-NEW-2	0.6739
ESTUCeng-E-CO-NEW-3	0.6537	Technion-E-CO-NEW-4	0.6717
ESTUCeng-E-CO-NEW-1	0.6508	ESTUCeng-E-CO-NEW-3	0.6644
Technion-E-CO-NEW-4	0.6505	ESTUCeng-E-CO-NEW-1	0.6638
mpii-E-CO-NEW-3	0.6337	mpii-E-CO-NEW-3	0.6556
Technion-E-CO-NEW-3	0.6315	Technion-E-CO-NEW-3	0.6509
SLWWW-E-CD-NEW-5	0.6291	SLWWW-E-CD-NEW-5	0.6445
KASYS-E-CO-REP-3	0.6275	Technion-E-CO-NEW-5	0.6426
SLWWW-E-CO-REP-2	0.6227	KASYS-E-CO-REP-3	0.6402
Technion-E-CO-NEW-5	0.6163	SLWWW-E-CO-REP-2	0.6359
KASYS-E-CO-REP-2	0.6131	KASYS-E-CO-REP-2	0.6256
THUIR-E-CO-REV-3	0.6049	THUIR-E-CO-REV-3	0.6241
THUIR-E-CO-REV-1	0.5994	NAUIR-E-CO-NEW-2	0.6095
NAUIR-E-CO-NEW-1	0.5989	NAUIR-E-CO-NEW-1	0.6089
NAUIR-E-CO-NEW-5	0.5982	NAUIR-E-CO-NEW-5	0.6083
NAUIR-E-CO-NEW-2	0.5980	THUIR-E-CO-REV-1	0.6068
THUIR-E-CO-REV-2	0.5876	THUIR-E-CO-REV-2	0.6010
NAUIR-E-CO-NEW-3	0.5851	NAUIR-E-CO-NEW-3	0.5977
baselineEng	0.5748	**baselineEng**	0.5850
SLWWW-E-CO-REP-3	0.5719	SLWWW-E-CO-REP-3	0.5822
RUCIR-E-CO-NEW-5	0.5611	RUCIR-E-CO-NEW-5	0.5755
NAUIR-E-CO-NEW-4	0.5557	NAUIR-E-CO-NEW-4	0.5712
RUCIR-E-CO-NEW-2	0.5418	RUCIR-E-CO-NEW-2	0.5594
RUCIR-E-CO-NEW-3	0.5363	RUCIR-E-CO-NEW-3	0.5569
SLWWW-E-CO-REP-1	0.5189	SLWWW-E-CO-REP-1	0.5366
RUCIR-E-CO-NEW-1	0.5158	RUCIR-E-CO-NEW-1	0.5276
THUIR-E-CO-NEW-4	0.5112	THUIR-E-CO-NEW-4	0.5250
ESTUCeng-E-CO-NEW-2	0.4991	ESTUCeng-E-CO-NEW-2	0.5051
THUIR-E-CO-REP-5	0.4767	THUIR-E-CO-REP-5	0.4899
SLWWW-E-CO-REP-4	0.4465	SLWWW-E-CO-REP-4	0.4531
RUCIR-E-CO-NEW-4	0.4251	RUCIR-E-CO-NEW-4	0.4207

8.2 Reproducibility

8.2.1 Results: Reproducing with t-Test. Table 20 reports p-values, resulted by running a two tailed unpaired t-test between the effectiveness scores of the original run and the reproduced run. Recall that original runs are evaluated with respect to WWW-2 topics and qrels, while reproduced runs are evaluated with respect to WWW-3 topics and qrels.

Differently from replicability results in Table 18, SLWWW-E-CO-REP-4 has the greatest p-values and with a level $\alpha = 0.05$ it is not statistically different from the original A-run for nERR and iRBU. On the other hand, KASYS-E-CO-REP-2 and KASYS-E-CO-REP-3 have very low p-values with respect to all measures: they are statistically significantly different from the corresponding original runs with $\alpha = 0.05$, therefore they are not successful in terms of reproducibility.

8.2.2 Results: Reproducing an Effect over a Baseline. In this section, we evaluate the REP A-run and REP B-run contributed by KASYS in terms how how successfully they *reproduce* the difference between the original A-run and B-run observed at WWW-2, using $\mathrm{ER}_{\mathrm{repro}}$, and $\Delta RI_{\mathrm{repro}}$ (See Section 5.2.2). As shown in Table 10, KASIS was the only team that submitted a pair of REP runs so we evaluate this pair only.

Table 21 reports reproducibility scores for $\mathrm{ER}_{\mathrm{repro}}$ and $\Delta RI_{\mathrm{repro}}$ for all the evaluation measures. $\mathrm{ER}_{\mathrm{repro}}$ scores are aligned with $\mathrm{ER}_{\mathrm{repli}}$ scores in Table 19: they are all lower than 0, denoting a failure in the reproducibility experiments, especially fo iRBU. In

Table 12: WWW-3 English subtask official results with nERR and iRBU (mean over the 80 WWW-3 test topics).

Run name	Mean nERR	Run name	Mean iRBU
mpii-E-CO-NEW-1	0.8090	KASYS-E-CO-NEW-4	0.9389
KASYS-E-CO-NEW-1	0.7959	KASYS-E-CO-NEW-1	0.9382
KASYS-E-CO-NEW-4	0.7893	KASYS-E-CO-NEW-5	0.9298
Technion-E-CO-NEW-1	0.7791	Technion-E-CO-NEW-1	0.9217
mpii-E-CO-NEW-2	0.7787	Technion-E-CO-NEW-2	0.9187
KASYS-E-CO-NEW-5	0.7768	mpii-E-CO-NEW-2	0.9175
ESTUCeng-E-CO-NEW-1	0.7597	mpii-E-CO-NEW-1	0.9165
ESTUCeng-E-CO-NEW-3	0.7561	ESTUCeng-E-CO-NEW-1	0.9163
Technion-E-CO-NEW-2	0.7502	ESTUCeng-E-CO-NEW-3	0.9161
Technion-E-CO-NEW-4	0.7471	THUIR-E-CO-REV-1	0.9112
KASYS-E-CO-REP-3	0.7410	mpii-E-CO-NEW-3	0.9041
mpii-E-CO-NEW-3	0.7395	SLWWW-E-CO-REP-2	0.9040
Technion-E-CO-NEW-3	0.7366	KASYS-E-CO-REP-3	0.9037
SLWWW-E-CD-NEW-5	0.7362	Technion-E-CO-NEW-4	0.9011
SLWWW-E-CO-REP-2	0.7345	THUIR-E-CO-REV-3	0.9007
Technion-E-CO-NEW-5	0.7286	SLWWW-E-CD-NEW-5	0.8981
KASYS-E-CO-REP-2	0.7213	NAUIR-E-CO-NEW-5	0.8973
THUIR-E-CO-REV-1	0.7190	SLWWW-E-CO-REP-3	0.8919
NAUIR-E-CO-NEW-2	0.7190	Technion-E-CO-NEW-3	0.8897
NAUIR-E-CO-NEW-1	0.7144	Technion-E-CO-NEW-5	0.8895
THUIR-E-CO-REV-2	0.7133	KASYS-E-CO-REP-2	0.8876
NAUIR-E-CO-NEW-5	0.7124	NAUIR-E-CO-NEW-3	0.8859
THUIR-E-CO-REV-3	0.7102	RUCIR-E-CO-NEW-5	0.8855
SLWWW-E-CO-REP-3	0.7061	NAUIR-E-CO-NEW-4	0.8852
NAUIR-E-CO-NEW-3	0.6915	NAUIR-E-CO-NEW-1	0.8846
RUCIR-E-CO-NEW-5	0.6789	NAUIR-E-CO-NEW-2	0.8844
NAUIR-E-CO-NEW-4	0.6786	RUCIR-E-CO-NEW-2	0.8839
baselineEng	0.6757	THUIR-E-CO-REV-2	0.8829
RUCIR-E-CO-NEW-3	0.6575	**baselineEng**	0.8802
RUCIR-E-CO-NEW-2	0.6550	SLWWW-E-CO-REP-1	0.8773
ESTUCeng-E-CO-NEW-2	0.6524	RUCIR-E-CO-NEW-3	0.8695
THUIR-E-CO-NEW-4	0.6481	ESTUCeng-E-CO-NEW-2	0.8677
SLWWW-E-CO-REP-1	0.6397	RUCIR-E-CO-NEW-1	0.8641
RUCIR-E-CO-NEW-1	0.6311	THUIR-E-CO-NEW-4	0.8579
THUIR-E-CO-REP-5	0.6021	THUIR-E-CO-REP-5	0.8259
SLWWW-E-CO-REP-4	0.5804	SLWWW-E-CO-REP-4	0.8220
RUCIR-E-CO-NEW-4	0.5596	RUCIR-E-CO-NEW-4	0.8194

Table 13: SYSDESC fields of the top 3 runs acccording to nDCG and Q.

KASYS-E-CO-NEW-1	Document ranking via sentence modeling using BERT(MS MARCO -> MB, k=3)
KASYS-E-CO-NEW-4	Document ranking via sentence modeling using BERT(MS MARCO -> MB, k=2)
mpii-E-CO-NEW-1	we re-rank top 1000 documents from the official baseline. For the re-ranking method, we use the ELECTRA-Base model fine-tuned on the MSMARCO passage dataset. The model is later fine-tuned on the www1 and www2 content queries. The utilized method requires a relative larger number of additional parameters.

terms of ΔRI_{repro}, the relative improvement of the REP A-run over the REP B-run are similar across all evaluation measures.

9 CONCLUSIONS

This paper provided an overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) task. Although our original plan was to discuss technological advances by comparing the REV A-run with the top NEW run on the WWW-3 topics, this was not possible

Table 14: Run ranking correlations in terms of Kendall's τ with 95%CIs ($n = 37$ English runs evaluated on the WWW-3 topic set).

	Q	nERR	iRBU
nDCG	0.970 [0.953, 0.981]	0.916 [0.871, 0.946]	0.823 [0.735, 0.884]
Q	-	0.898 [0.844, 0.934]	0.799 [0.702, 0.867]
nERR	-	-	0.805 [0.710, 0.871]

Table 15: Kendall's τ: 20 WWW-2 runs ranked with the Official WWW-2 qrels vs. those ranked with the new qrels.

Measure	τ	95%CI
nDCG	0.611	[0.368, 0.776]
Q	0.495	[0.215, 0.700]
nERR	0.716	[0.519, 0.841]

Table 16: Kendall's τ union and RBO for replicability run averaged across topics.

Run Type	Run Name	Kendall's τ	RBO $p = 0.9$
REP A-run	KASYS-E-CO-REP-2	0.1152	0.2202
REP A-run	SLWWW-E-CO-REP-4	0.0446	0.0171
REP B-run	KASYS-E-CO-REP-3	−0.0183	0.0980

Table 17: RMSE$_{abs}$ scores for each replicability runs and measures.

		RMSE$_{abs}$			
Run Type	Run Name	nDCG	Q	nERR	iRBU
REP A-run	KASYS-E-CO-REP-2	0.1822	0.2003	0.3000	0.2448
REP A-run	SLWWW-E-CO-REP-4	0.2770	0.3321	0.3833	0.4317
REP B-run	KASYS-E-CO-REP-3	0.2240	0.2312	0.3452	0.3382

as we failed to obtain a reliable REV A-run this time. Also, our CENTRE analysis suggests that replicability and reproducibility are very tough problems. Furthermore, our analysis of the English subtask results based on multiple versions of qrels suggests that our qrels sets are not reusable.

Given the above somewhat depressing findings, we will propose to continue the WWW task at NTCIR-16 to tackle the unsolved problems. As we never had many participating teams in the Chinese subtask, we will terminate this subtask and focus on English web search, and try to measure technological advances, replicability and reproducibility. As our next target A-run, we plan to use KASYS-E-CO-NEW-1 (a BERT-based run) as this was one of the most successful runs in the WWW-3 English subtask. Makoto P. Kato of KASYS (University of Tsukuba) has kindly agreed to provide a REV A-run for the next round of the WWW task , so we should be able to discuss technological advances, replicability and reproducibility next time (if the task proposal is accepted for NTCIR-16!). We also plan to introduce a new target web corpus, which should be yet another challenge in terms of reproducibility.

ACKNOWLEDGEMENTS

The WWW-3 participants have made our work possible. Thank you! We also thank the NTCIR organisers and staff for their constant support.

This paper was partially supported by the the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 893667.

DISCLAIMER

Certain companies and products are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the products or companies identified are necessarily the best available for the purpose.

REFERENCES

[1] Enrique Amigó, Damiano Spina, and Jorge Carrillo de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducting the Rank-Biased Utility Metric. In *Proceedings of ACM SIGIR 2018*. 625–634.

[2] Timo Breuer, Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, Philipp Shaer, and Ian Soboroff. 2020. How to Measure the Reproducibility of System-oriented IR Experiments. In *Proceedings of ACM SIGIR 2020*. 349–358.

[3] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of ACM CIKM 2009*. 621–630.

[4] Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. 2019. CENTRE@CLEF 2019. In *Proceedings of ECIR 2019 Part II (LNCS 11438)*. 283–290.

[5] Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. 2019. CENTRE@CLEF: Sequel in the Systematic Reproducibility Realm. In *Proceedings of CLEF 2019 (LNCS 11696)*. 287–300.

[6] Nicola Ferro, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. 2018. Overview of CENTRE@CLEF 2018: a First Tale in the Systematic Reproducibility Realm. In *Proceedings of CLEF 2018 (LNCS 11018)*. 239–246.

[7] M. G. Kendall. 1948. *Rank correlation methods*. Griffin, Oxford, England.

[8] Cheng Luo, Tetsuya Sakai, Yiqun Liu, Zhicheng Dou, Chenyan Xiong, and Jingfang Xu. 2017. Overview of the NTCIR-13 We Want Web Task. In *Proceedings of NTCIR-13*. 394–401. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/pdf/ntcir/01-NTCIR13-OV-WWW-LuoC.pdf

[9] Jiaxin Mao, Tetsuya Sakai, Cheng Luo, Peng Xiao, Yiqun Liu, and Zhicheng Dou. 2019. Overview of the NTCIR-14 We Want Web Task. In *Proceedings of NTCIR-14*. 455–467. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/01-NTCIR14-OV-WWW-MaoJ.pdf

[10] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM TOIS* 27, 1 (2008).

[11] Tetsuya Sakai. 2013. Topic Set Size Design. *Information Retrieval Journal* 19, 3 (2013), 256–283. http://link.springer.com/content/pdf/10.1007%2Fs10791-015-9273-z.pdf

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

[13] 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

[14] Tetsuya Sakai. 2019. How to Run an Evaluation Task: with a Primary Focus on Ad Hoc Information Retrieval. In *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, Nicola Ferro and Carol Peters (Eds.). Springer, 71–102.

[15] Tetsuya Sakai, Nicola Ferro, Ian Soboroff, Zhahao Zeng, Peng Xiao, and Maria Maistro. 2019. Overview of the NTCIR-14 CENTRE Task. In *Proceedings of*

Table 18: p-value returned by a two tailed paired t-test run between the original and replicated runs.

		p-value			
Run Type	Run Name	nDCG	Q	nERR	iRBU
REP A-run	KASYS-E-CO-REP-2	0.0017	1.8291×10^{-4}	0.1757	0.0174
REP A-run	SLWWW-E-CO-REP-4	1.0237×10^{-17}	1.6874×10^{-16}	1.7870×10^{-11}	1.4702×10^{-13}
REP B-run	KASYS-E-CO-REP-3	0.3560	0.1943	0.4403	0.1350

Table 19: Results for replicability of effects over a baseline.

		RMSE_Δ				ER_{repli}				ΔRI_{repli}			
A-run	B-run	nDCG	Q	nERR	iRBU	nDCG	Q	nERR	iRBU	nDCG	Q	nERR	iRBU
KASYS-E-CO-REP-2	KASYS-E-CO-REP-3	0.2150	0.2208	0.3759	0.3105	-0.8051	-1.0480	0.5340	-0.6508	0.1233	0.1587	0.0306	0.0107

Table 20: p-value returned by a two tailed unpaired t-test run between the original and reproduced runs.

		p-value			
Run Type	Run Name	nDCG	Q	nERR	iRBU
REP A-run	KASYS-E-CO-REP-2	7.6649×10^{-11}	6.3367×10^{-10}	9.2738×10^{-7}	0.0015
REP A-run	SLWWW-E-CO-REP-4	0.0104	0.0079	0.0798	0.1583
REP B-run	KASYS-E-CO-REP-3	6.7566×10^{-15}	1.5768×10^{-13}	9.5898×10^{-12}	1.8840×10^{-4}

Table 21: Results for reproducibility of effects over a baseline.

		ER_{repro}				ΔRI_{repro}			
A-run	B-run	nDCG	Q	nERR	iRBU	nDCG	Q	nERR	iRBU
KASYS-E-CO-REP-2	KASYS-E-CO-REP-3	-0.6638	-0.6344	-0.5885	-3.3709	0.0891	0.0959	0.0978	0.0241

NTCIR-14. 494–509. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/01-NTCIR14-OV-CENTRE-SakaiT.pdf

[16] Tetsuya Sakai, Noriko Kando, Chuan-Jie Lin, Teruko Mitamura, Hideki Shima, Donghong Ji, Kuang-Hua Chen, and Eric Nyberg. 2008. Overview of the NTCIR-7 ACLIA IR4QA Task. In *Proceedings of NTCIR-7*. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C1/IR4QA/01-NTCIR7-OV-IR4QA-SakaiT.pdf

[17] Tetsuya Sakai and Peng Xiao. 2020. Randomised vs. Prioritised Pools for Relevance Assessments: Sample Size Considerations. In *Proceedings of AIRS 2019 (LNCS 12004)*. 94–105.

[18] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures are "Good"?. In *Proceedings of ACM SIGIR 2019*. 595–604.

[19] Ian Soboroff, Nicola Ferro, Maria Maistro, and Tetsuya Sakai. 2020. Overview of the TREC 2018 CENTRE Track. In *Proceedings of TREC 2018*.

[20] W. Webber, A. Moffat, and J. Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Transactions on Information Systems (TOIS)* 4, 28 (November 2010), 20:1–20:38.

[21] Yukun Zheng, Zhen Fan, Yiqun Liu, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Sogou-qcl: A new dataset with click relevance label. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1117–1120.

APPENDIX

Table 22: Randomised Tukey HSD test results ($B = 10,000$ trials) for the mean nDCG and Q scores in Table 7. The runs in the left column are statistically significantly better than those in the right column at the 5% significance level.

(a) Mean nDCG	
RUCIR-C-CD-NEW-4	RUCIR-C-CO-NEW-5 . . . THUIR-C-CO-NEW-4 (8 runs)
RUCIR-C-CD-NEW-3	ditto
RUCIR-C-CD-NEW-1	RUCIR-C-CD-NEW-2 . . . THUIR-C-CO-NEW-4 (7 runs)
RUCIR-C-CO-NEW-5	THUIR-C-CO-NEW-3 . . . THUIR-C-CO-NEW-4 (4 runs)
RUCIR-C-CD-NEW-2	**baselineChn** . . . THUIR-C-CO-NEW-4 (3 runs)
THUIR-C-CO-NEW-2	ditto
THUIR-C-CO-NEW-1	ditto
THUIR-C-CO-NEW-3	ditto
baselineChn	THUIR-C-CO-NEW-4 (1 run)
(b) Mean Q	
RUCIR-C-CD-NEW-4	RUCIR-C-CO-NEW-5 . . . THUIR-C-CO-NEW-4 (8 runs)
RUCIR-C-CD-NEW-3	RUCIR-C-CD-NEW-2 . . . THUIR-C-CO-NEW-4 (7 runs)
RUCIR-C-CD-NEW-1	ditto
RUCIR-C-CO-NEW-5	THUIR-C-CO-NEW-1 . . . THUIR-C-CO-NEW-4 (4 runs)
RUCIR-C-CD-NEW-2	**baselineChn** . . . THUIR-C-CO-NEW-4 (3 runs)
THUIR-C-CO-NEW-2	ditto
THUIR-C-CO-NEW-1	ditto
THUIR-C-CO-NEW-3	ditto
baselineChn	THUIR-C-CO-NEW-4 (1 run)

Table 23: Randomised Tukey HSD test results ($B = 10,000$ trials) for the mean nERR and iRBU scores in Table 8. The runs in the left column are statistically significantly better than those in the right column at the 5% significance level.

(a) Mean nERR	
RUCIR-C-CD-NEW-4	THUIR-C-CO-NEW-1 . . . THUIR-C-CO-NEW-4 (7 runs)
RUCIR-C-CD-NEW-3	RUCIR-C-CO-NEW-5 . . . THUIR-C-CO-NEW-4 (6 runs)
RUCIR-C-CD-NEW-1	RUCIR-C-CD-NEW-2 . . . THUIR-C-CO-NEW-4 (5 runs)
THUIR-C-CO-NEW-2	THUIR-C-CO-NEW-3 . . . THUIR-C-CO-NEW-4 (4 runs)
THUIR-C-CO-NEW-1	ditto
RUCIR-C-CO-NEW-5	THUIR-C-CO-REV-5 . . . THUIR-C-CO-NEW-4 (3 runs)
RUCIR-C-CD-NEW-2	ditto
THUIR-C-CO-NEW-3	ditto
(b) Mean iRBU	
RUCIR-C-CD-NEW-4	RUCIR-C-CD-NEW-2 . . . THUIR-C-CO-NEW-4 (7 runs)
RUCIR-C-CD-NEW-3	THUIR-C-CO-NEW-2 . . . THUIR-C-CO-NEW-4 (6 runs)
RUCIR-C-CO-NEW-5	ditto
RUCIR-C-CD-NEW-1	THUIR-C-CO-NEW-1 . . . THUIR-C-CO-NEW-4 (5 runs)
RUCIR-C-CD-NEW-2	ditto
THUIR-C-CO-NEW-2	THUIR-C-CO-NEW-3 . . . THUIR-C-CO-NEW-4 (4 runs)
THUIR-C-CO-NEW-1	**baselineChn** . . . THUIR-C-CO-NEW-4 (3 runs)
THUIR-C-CO-NEW-3	ditto

Table 24: Randomised Tukey HSD test results ($B = 10,000$ trials) for the mean nDCG and Q scores in Table 11. The runs in the left column are statistically significantly better than those in the right column at the 5% significance level.

(a) Mean nDCG	
KASYS-E-CO-NEW-1	THUIR-E-CO-REV-3 . . . RUCIR-E-CO-NEW-4 (20 runs)
mpii-E-CO-NEW-1	ditto
KASYS-E-CO-NEW-4	ditto
KASYS-E-CO-NEW-5	NAUIR-E-CO-NEW-1 . . . RUCIR-E-CO-NEW-4 (18 runs)
mpii-E-CO-NEW-2	THUIR-E-CO-REV-2 . . . RUCIR-E-CO-NEW-4 (15 runs)
Technion-E-CO-NEW-1	**baselineEng** . . . RUCIR-E-CO-NEW-4 (13 runs)
Technion-E-CO-NEW-2	SLWWW-E-CO-REP-3 . . . RUCIR-E-CO-NEW-4 (12 runs)
ESTUCeng-E-CO-NEW-3	RUCIR-E-CO-NEW-5 . . . RUCIR-E-CO-NEW-4 (11 runs)
ESTUCeng-E-CO-NEW-1	ditto
Technion-E-CO-NEW-4	ditto
mpii-E-CO-NEW-3	RUCIR-E-CO-NEW-2 . . . RUCIR-E-CO-NEW-4 (9 runs)
Technion-E-CO-NEW-3	ditto
SLWWW-E-CD-NEW-5	ditto
KASYS-E-CO-REP-3	ditto
SLWWW-E-CO-REP-2	RUCIR-E-CO-NEW-3 . . . RUCIR-E-CO-NEW-4 (8 runs)
Technion-E-CO-NEW-5	SLWWW-E-CO-REP-1 . . . RUCIR-E-CO-NEW-4 (7 runs)
KASYS-E-CO-REP-2	ditto
THUIR-E-CO-REV-3	ditto
THUIR-E-CO-REV-1	RUCIR-E-CO-NEW-1 . . . RUCIR-E-CO-NEW-4 (6 runs)
NAUIR-E-CO-NEW-1	ditto
NAUIR-E-CO-NEW-5	ditto
NAUIR-E-CO-NEW-2	ditto
THUIR-E-CO-REV-2	ESTUCeng-E-CO-NEW-2 . . . RUCIR-E-CO-NEW-4 (4 runs)
NAUIR-E-CO-NEW-3	ditto
baselineEng	THUIR-E-CO-REP-5 . . . RUCIR-E-CO-NEW-4 (3 runs)
SLWWW-E-CO-REP-3	ditto
RUCIR-E-CO-NEW-5	ditto
NAUIR-E-CO-NEW-4	SLWWW-E-CO-REP-4, RUCIR-E-CO-NEW-4 (2 runs)
RUCIR-E-CO-NEW-2	ditto
RUCIR-E-CO-NEW-3	ditto
SLWWW-E-CO-REP-1	RUCIR-E-CO-NEW-4 (1 run)
RUCIR-E-CO-NEW-1	ditto
THUIR-E-CO-NEW-4	ditto
(b) Mean Q	
KASYS-E-CO-NEW-1	NAUIR-E-CO-NEW-2 . . . RUCIR-E-CO-NEW-4 (19 runs)
KASYS-E-CO-NEW-4	ditto
mpii-E-CO-NEW-1	ditto
KASYS-E-CO-NEW-5	NAUIR-E-CO-NEW-5 . . . RUCIR-E-CO-NEW-4 (17 runs)
mpii-E-CO-NEW-2	NAUIR-E-CO-NEW-3 . . . RUCIR-E-CO-NEW-4 (14 runs)
Technion-E-CO-NEW-1	**baselineEng** . . . RUCIR-E-CO-NEW-4 (13 runs)
Technion-E-CO-NEW-2	SLWWW-E-CO-REP-3 . . . RUCIR-E-CO-NEW-4 (12 runs)
Technion-E-CO-NEW-4	RUCIR-E-CO-NEW-5 . . . RUCIR-E-CO-NEW-4 (11 runs)
ESTUCeng-E-CO-NEW-3	NAUIR-E-CO-NEW-4 . . . RUCIR-E-CO-NEW-4 (10 runs)
ESTUCeng-E-CO-NEW-1	ditto
mpii-E-CO-NEW-3	RUCIR-E-CO-NEW-2 . . . RUCIR-E-CO-NEW-4 (9 runs)
Technion-E-CO-NEW-3	ditto
SLWWW-E-CD-NEW-5	SLWWW-E-CO-REP-1 . . . RUCIR-E-CO-NEW-4 (7 runs)
Technion-E-CO-NEW-5	ditto
KASYS-E-CO-REP-3	ditto
SLWWW-E-CO-REP-2	ditto
KASYS-E-CO-REP-2	RUCIR-E-CO-NEW-1 . . . RUCIR-E-CO-NEW-4 (6 runs)
THUIR-E-CO-REV-3	ditto
NAUIR-E-CO-NEW-2	ESTUCeng-E-CO-NEW-2 . . . RUCIR-E-CO-NEW-4 (4 runs)
NAUIR-E-CO-NEW-1	ditto
NAUIR-E-CO-NEW-5	ditto
THUIR-E-CO-REV-1	ditto
THUIR-E-CO-REV-2	ditto
NAUIR-E-CO-NEW-3	ditto
baselineEng	THUIR-E-CO-REP-5 . . . RUCIR-E-CO-NEW-4 (3 runs)
SLWWW-E-CO-REP-3	ditto
RUCIR-E-CO-NEW-5	SLWWW-E-CO-REP-4, RUCIR-E-CO-NEW-4 (2 runs)
NAUIR-E-CO-NEW-4	ditto
RUCIR-E-CO-NEW-2	ditto
RUCIR-E-CO-NEW-3	ditto
SLWWW-E-CO-REP-1	RUCIR-E-CO-NEW-4 (1 run)
RUCIR-E-CO-NEW-1	ditto
THUIR-E-CO-NEW-4	ditto

Table 25: Randomised Tukey HSD test results ($B = 10,000$ trials) for the mean nERR and iRBU scores in Table 12. The runs in the left column are statistically significantly better than those in the right column at the 5% significance level.

(a) Mean nERR	
mpii-E-CO-NEW-1	SLWWW-E-CO-REP-3 . . . RUCIR-E-CO-NEW-4 (14 runs)
KASYS-E-CO-NEW-1	NAUIR-E-CO-NEW-3 . . . RUCIR-E-CO-NEW-4 (13 runs)
KASYS-E-CO-NEW-4	RUCIR-E-CO-NEW-5 . . . RUCIR-E-CO-NEW-4 (12 runs)
Technion-E-CO-NEW-1	ditto
mpii-E-CO-NEW-2	ditto
KASYS-E-CO-NEW-5	**baselineEng** . . . RUCIR-E-CO-NEW-4 (10 runs)
ESTUCeng-E-CO-NEW-1	RUCIR-E-CO-NEW-3 . . . RUCIR-E-CO-NEW-4 (9 runs)
ESTUCeng-E-CO-NEW-3	RUCIR-E-CO-NEW-2 . . . RUCIR-E-CO-NEW-4 (8 runs)
Technion-E-CO-NEW-2	THUIR-E-CO-NEW-4 . . . RUCIR-E-CO-NEW-4 (6 runs)
Technion-E-CO-NEW-4	SLWWW-E-CO-REP-1 . . . RUCIR-E-CO-NEW-4 (5 runs)
KASYS-E-CO-REP-3	ditto
mpii-E-CO-NEW-3	ditto
Technion-E-CO-NEW-3	RUCIR-E-CO-NEW-1 . . . RUCIR-E-CO-NEW-4 (4 runs)
SLWWW-E-CD-NEW-5	ditto
SLWWW-E-CO-REP-2	ditto
Technion-E-CO-NEW-5	THUIR-E-CO-REP-5 . . . RUCIR-E-CO-NEW-4 (3 runs)
KASYS-E-CO-REP-2	ditto
THUIR-E-CO-REV-1	ditto
NAUIR-E-CO-NEW-2	ditto
NAUIR-E-CO-NEW-1	ditto
THUIR-E-CO-REV-2	ditto
NAUIR-E-CO-NEW-5	ditto
THUIR-E-CO-REV-3	ditto
SLWWW-E-CO-REP-3	ditto
NAUIR-E-CO-NEW-3	SLWWW-E-CO-REP-4, RUCIR-E-CO-NEW-4 (2 runs)
RUCIR-E-CO-NEW-5	RUCIR-E-CO-NEW-4 (1 run)
NAUIR-E-CO-NEW-4	ditto
baselineEng	ditto
(b) Mean iRBU	
KASYS-E-CO-NEW-4	SLWWW-E-CO-REP-1 . . . RUCIR-E-CO-NEW-4 (8 runs)
KASYS-E-CO-NEW-1	ditto
KASYS-E-CO-NEW-5	ESTUCeng-E-CO-NEW-2 . . . RUCIR-E-CO-NEW-4 (6 runs)
Technion-E-CO-NEW-1	THUIR-E-CO-NEW-4 . . . RUCIR-E-CO-NEW-4 (4 runs)
Technion-E-CO-NEW-2	ditto
mpii-E-CO-NEW-2	THUIR-E-CO-REP-5 . . . RUCIR-E-CO-NEW-4 (3 runs)
mpii-E-CO-NEW-1	ditto
ESTUCeng-E-CO-NEW-1	ditto
ESTUCeng-E-CO-NEW-3	ditto
THUIR-E-CO-REV-1	ditto
mpii-E-CO-NEW-3	ditto
SLWWW-E-CO-REP-2	ditto
KASYS-E-CO-REP-3	ditto
Technion-E-CO-NEW-4	ditto
THUIR-E-CO-REV-3	ditto
SLWWW-E-CD-NEW-5	ditto
NAUIR-E-CO-NEW-5	ditto
SLWWW-E-CO-REP-3	ditto
Technion-E-CO-NEW-3	ditto
Technion-E-CO-NEW-5	ditto
KASYS-E-CO-REP-2	ditto
NAUIR-E-CO-NEW-3	SLWWW-E-CO-REP-4, RUCIR-E-CO-NEW-4 (2 runs)
RUCIR-E-CO-NEW-5	ditto
NAUIR-E-CO-NEW-4	ditto
NAUIR-E-CO-NEW-1	ditto
NAUIR-E-CO-NEW-2	ditto
RUCIR-E-CO-NEW-2	ditto
THUIR-E-CO-REV-2	ditto
baselineEng	RUCIR-E-CO-NEW-4 (1 run)

Table 26: WWW-2 runs evaluated with the Official WWW-2 qrels and with the new qrels (mean nDCG). The results in the left column are the same as those shown in the WWW-2 overview paper [9, Table 12].

Run name	Official	Run name	New
THUIR-E-CO-MAN-Base-3	0.3536	THUIR-E-CO-MAN-Base-3	0.7108
THUIR-E-CO-MAN-Base-2	0.3512	**THUIR-E-CO-MAN-Base-2**	0.7000
RUCIR-E-CO-PU-Base-2	0.3489	THUIR-E-CO-MAN-Base-1	0.6940
THUIR-E-CO-MAN-Base-1	0.3444	**baseline_eng_v1**	0.6517
MPII-E-CO-NU-Base-3	0.3413	THUIR-E-CO-PU-Base-5	0.6517
MPII-E-CO-NU-Base-2	0.3394	RUCIR-E-CO-PU-Base-2	0.6207
MPII-E-CO-NU-Base-4	0.3336	RUCIR-E-DE-PU-Base-4	0.5915
THUIR-E-CO-PU-Base-4	0.3294	MPII-E-CO-NU-Base-3	0.5081
RUCIR-E-DE-PU-Base-4	0.3293	MPII-E-CO-NU-Base-4	0.5033
MPII-E-CO-NU-Base-5	0.3293	MPII-E-CO-NU-Base-1	0.5001
baseline_eng_v1	0.3258	RUCIR-E-DE-PU-Base-3	0.4903
THUIR-E-CO-PU-Base-5	0.3258	RUCIR-E-DE-PU-Base-1	0.4903
MPII-E-CO-NU-Base-1	0.3204	MPII-E-CO-NU-Base-2	0.4880
RUCIR-E-DE-PU-Base-3	0.3137	RUCIR-E-DE-PU-Base-5	0.4694
RUCIR-E-DE-PU-Base-1	0.3137	MPII-E-CO-NU-Base-5	0.4687
RUCIR-E-DE-PU-Base-5	0.2876	THUIR-E-CO-PU-Base-4	0.4396
SLWWW-E-CO-NU-Base-1	0.2860	SLWWW-E-CD-NU-Base-3	0.3544
ORG-MANUAL	0.2844	SLWWW-E-CO-NU-Base-4	0.3533
SLWWW-E-CO-NU-Base-4	0.2775	SLWWW-E-CO-NU-Base-1	0.3532
SLWWW-E-CD-NU-Base-3	0.2767	ORG-MANUAL	0.2386

Table 27: WWW-2 runs evaluated with the Official WWW-2 qrels and with the new qrels (mean Q). The results in the left column are the same as those shown in the WWW-2 overview paper [9, Table 12].

Run name	Offcial	Run name	New
THUIR-E-CO-MAN-Base-2	0.3391	THUIR-E-CO-MAN-Base-3	0.7349
RUCIR-E-CO-PU-Base-2	0.3352	**THUIR-E-CO-MAN-Base-2**	0.7251
MPII-E-CO-NU-Base-4	0.3265	THUIR-E-CO-MAN-Base-1	0.7115
THUIR-E-CO-MAN-Base-3	0.3256	**baseline_eng_v1**	0.6727
MPII-E-CO-NU-Base-2	0.3255	THUIR-E-CO-PU-Base-5	0.6727
THUIR-E-CO-MAN-Base-1	0.3249	RUCIR-E-CO-PU-Base-2	0.6128
MPII-E-CO-NU-Base-3	0.3183	RUCIR-E-DE-PU-Base-4	0.5684
THUIR-E-CO-PU-Base-4	0.3161	MPII-E-CO-NU-Base-3	0.4630
MPII-E-CO-NU-Base-5	0.3110	MPII-E-CO-NU-Base-4	0.4528
RUCIR-E-DE-PU-Base-4	0.3094	MPII-E-CO-NU-Base-1	0.4469
baseline_eng_v1	0.3043	RUCIR-E-DE-PU-Base-5	0.4357
THUIR-E-CO-PU-Base-5	0.3043	RUCIR-E-DE-PU-Base-3	0.4349
MPII-E-CO-NU-Base-1	0.3009	RUCIR-E-DE-PU-Base-1	0.4349
RUCIR-E-DE-PU-Base-3	0.2973	MPII-E-CO-NU-Base-2	0.4311
RUCIR-E-DE-PU-Base-1	0.2973	MPII-E-CO-NU-Base-5	0.3974
ORG-MANUAL	0.2685	THUIR-E-CO-PU-Base-4	0.3710
SLWWW-E-CO-NU-Base-1	0.2665	SLWWW-E-CO-NU-Base-1	0.2964
RUCIR-E-DE-PU-Base-5	0.2659	SLWWW-E-CO-NU-Base-4	0.2794
SLWWW-E-CD-NU-Base-3	0.2499	SLWWW-E-CD-NU-Base-3	0.2791
SLWWW-E-CO-NU-Base-4	0.2498	ORG-MANUAL	0.1616

Table 28: WWW-2 runs evaluated with the Official WWW-2 qrels and with the new qrels (mean nERR). The results in the left column are the same as those shown in the WWW-2 overview paper [9, Table 12].

Run name	Official	Run name	New
THUIR-E-CO-MAN-Base-1	0.5048	THUIR-E-CO-MAN-Base-3	0.8216
THUIR-E-CO-MAN-Base-2	0.5026	**THUIR-E-CO-MAN-Base-2**	0.8161
RUCIR-E-CO-PU-Base-2	0.4917	THUIR-E-CO-MAN-Base-1	0.8157
THUIR-E-CO-MAN-Base-3	0.4805	RUCIR-E-CO-PU-Base-2	0.7752
baseline_eng_v1	0.4779	**baseline_eng_v1**	0.7557
THUIR-E-CO-PU-Base-5	0.4779	THUIR-E-CO-PU-Base-5	0.7557
MPII-E-CO-NU-Base-4	0.4723	RUCIR-E-DE-PU-Base-4	0.7335
THUIR-E-CO-PU-Base-4	0.4692	MPII-E-CO-NU-Base-4	0.6783
MPII-E-CO-NU-Base-3	0.4658	MPII-E-CO-NU-Base-1	0.6646
RUCIR-E-DE-PU-Base-4	0.4602	RUCIR-E-DE-PU-Base-3	0.6489
MPII-E-CO-NU-Base-2	0.4590	RUCIR-E-DE-PU-Base-1	0.6489
MPII-E-CO-NU-Base-5	0.4584	MPII-E-CO-NU-Base-2	0.6303
MPII-E-CO-NU-Base-1	0.4541	THUIR-E-CO-PU-Base-4	0.6223
RUCIR-E-DE-PU-Base-3	0.4469	MPII-E-CO-NU-Base-3	0.6167
RUCIR-E-DE-PU-Base-1	0.4469	MPII-E-CO-NU-Base-5	0.6138
ORG-MANUAL	0.4294	RUCIR-E-DE-PU-Base-5	0.5499
RUCIR-E-DE-PU-Base-5	0.4188	SLWWW-E-CD-NU-Base-3	0.5155
SLWWW-E-CO-NU-Base-1	0.4071	SLWWW-E-CO-NU-Base-4	0.5150
SLWWW-E-CD-NU-Base-3	0.4034	SLWWW-E-CO-NU-Base-1	0.4883
SLWWW-E-CO-NU-Base-4	0.4015	ORG-MANUAL	0.4027