Overview of the NTCIR-13 We Want Web Task

Cheng Luo
Tsinghua University, P.R.C.
chengluo@tsinghua.edu.cn

Tetsuya Sakai
Waseda Univerisity, Japan
tetsuyasakai@acm.org

Yiqun Liu
Tsinghua University, P.R.C.
yiqunliu@tsinghua.edu.cn

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

Chenyan Xiong
Carnegie Mellon University, U.S.A.
cx@cs.cmu.edu

Jingfang Xu
Sogou Inc., P.R.C.
xujingfang@sogou-inc.com

ABSTRACT

In this paper, we provide an overview of the NTCIR We Want Web (WWW) task, which comprises the Chinese and the English subtasks. The WWW task is a classical ad-hoc textual retrieval task. This round of WWW received 19 runs from 4 teams for the Chinese subtask, and 13 runs from 3 teams for the English subtask. In this overview paper, we describe the task details, data and evaluation methods, as well as the report on the official results.

Keywords

ad hoc retrieval; click data; evaluation; information retrieval; test collections; web search

1. INTRODUCTION

Information access tasks have diversified: currently there are various novel tracks/tasks at NTCIR, TREC, CLEF etc. This is in sharp contrast to the early TRECs where there were only a few tracks, where the ad hoc track (a set of new topics run against a static document collection) was at the core. But is the ad hoc task a solved problem? It seems to us that researchers have moved on to new tasks not because they have completely solved the problem, but because they have reached a plateau. Ad hoc Web search, in particular, is still of utmost practical importance. Web search engines such as Baidu, Bing and Google are doing excellent jobs for users, but they are black boxes. We believe IR researchers should continue to study and understand the core problems of ranked retrieval and advance the state of the art. If we can improve the ad hoc IR performance, other tasks will also benefit from it.

Straight ad hoc web search tasks have disappeared from NTCIR and TREC. We believe that researchers still want to tackle basic web search problems and go beyond BM25F. Moreover, a “stable” evaluation forum, involving several rounds of NTCIR or TREC, to monitor the progress of IR algorithms seems to be in order. In addition, on the evaluation side, researchers (as well as search engine companies) want measures that really reflect the user’ s experience, rather than those that produce some numbers based on a ranked list of document IDs.

Recently, deep neural networks have already delivered great improvements in many machine learning tasks, such as speech recognition, computer vision, natural language processing, and etc. A number of studies have already been proposed to address the challenges in IR, in particular, ad hoc search. We believe that it is a necessity to provide an evaluation forum and monitor the development of neural IR models on time dimension.

Based on these considerations, we decided to run an ad-hoc evaluation task in NTCIR 12, which is named as We Want Web (WWW). The name of this task is inspired by the buzz in social media when the Web Track was terminated at TREC 2014: “We want Web”, “Web ad hoc now!”, and etc.

The main task of WWW is a traditional ad hoc task. The participants need to build their ranking systems on a given corpus. Then they are required to submit several runs for a given topic set. In this round of WWW (NTCIR-13), we have the Chinese subtask and the English subtask. The two subtasks adopt similar task setting with different data (see Section 3). There is some overlap between the two query sets, to support potential cross-language IR studies. In our plan, we will run another Japanese subtask in the future rounds of WWW. More details about the task definition will be presented in Section 2. The performance of retrieval systems will be evaluated in classical TREC ways. We presented the details of relevance judgments in Section 4.1, and official results in Section 6.

The schedule of WWW in NTCIR-13 is presented in Table 1. Although there are quite a few teams registered for our task, finally we only received 19 Chinese runs from 4 teams, and 13 English runs from 3 teams. We suspect that one of the potential reasons for the poor participation is the lack of training data for machine-learning-based approaches to web search. We discuss the further plan for WWW in the Section 7.

2. TASK DEFINITION

2.1 Main task definition

The main task of WWW is a classical ad hoc search task. The organizers will provide a corpus, which contains a large number of documents (web pages) and a query set. Then the participants need to construct their own ranking systems on the corpus. Retrieval results for each query will be submitted in the form of a ranked list. After receiving the runs from participants, the organizers will first construct a result pool by aggregating the top k results from all the runs. The depth of the pool determines how many results will be taken into consideration when comparing the performance of different submissions. For example, if we use 20, we can only calculate the metrics whose cutoff is smaller than 20.

Table 1: Schedule of WWW at NTCIR-13

Time	Content
Jul to Aug 2016	Corpora released to registered participants
Aug to Sep 2016	Designing and constructing topics
Oct 2016 to Jan 2017	User behavior data collected for the topics
Feb to Mar 2017	User behavior data released to registered participants
Apr 2017	Task registration due
May 2017	Topics released; runs received
July 16, 2017	runs received
July to Aug 2017	Relevance assessments
Sep 1, 2017	Results and Draft Task overview released to participants
Oct 1, 2017	Participants' draft papers due
Nov 1, 2017	All camera ready papers due
Nov 2017	Pre-NTCIR-13 WWW Workshop on Failure Analysis in Beijing
Dec 2017	NTCIR-13 Conference

The depth of pooling is also limited by the cost for relevance judgments, in terms of time and money. Relevance judgments are conducted on the result pool. We adopt the typical TREC relevance judgment setting in WWW. Once the relevance judgments are finished, the organizers are able to calculate various evaluation metrics (such as Precision, Recall, nDCG and etc.) to compare the performance of different submitted runs.

Considering that building an index system on a large corpus might be very challenging and time-consuming, we offer a much easier plan for the participants. We provide a baseline ranking so that the participants could directly use their own algorithm to rerank it. More specifically, for each query, we provide the top 1,000 retrieved results as well as corresponding relevance score and the original HTML.

2.2 Subtasks

In WWW of NTCIR-13, we have Chinese subtask and English subtask. Considering the fact that NTCIR INTENT/IMine have had relatively small number of Japanese subtask participants, we will save the Japanese subtask until NTCIR-14.

The Chinese subtask and the English subtask basically adopt same task settings. The major difference is the data we provided.

For Chinese subtask, we provide a training set containing 200 Chinese queries. These queries are sampled from a commercial search engine's query log. The training set has two parts of data. The first one is the click logs collected by the commercial search engine. The click logs are collected from March, 2017 to April 2017. The second part of the data is relevance judgments for queries in training set. Unfortunately, for English subtask, we have no data for training. This also prevents the participants to build more complex ranking system.

2.3 Long term plan for WWW

We plan to run WWW for at least three rounds at NTCIR, to track relatively long term development of ranking techniques. We also would like to introduce a Japanese subtask at NTCIR-14, if there are sufficient demands. At NTCIR-15, we will decide whether to continue for NTCIR-16 based on participants' demands.

3. DATA

3.1 Corpus

For the Chinese Subtask, we adopt the new SogouT-16 as the document collection [2]. SogouT-16 contains about 1.17B Web pages, which are sampled from the index of Sogou, which is the second largest commercial search engine in China. Considering that the original SogouT might be a little bit difficult to handle for some research groups (almost 80TB after decompression), we prepare a "Category B" version of SogouT-16, which is denoted as "SogouT-16 B". This subset contains about 15

For the English Subtask, we adopt the ClueWeb12-B13 as the document collection [1]. This corpus is also free for research purpose. You only need to pay for the disks and the shipment. More information can be found at Clueweb-12's homepage. ClueWeb-12 also has a free online retrieval/page rendering service, it can be utilized after the agreement is signed.

The retrieval system for Chinese system was constructed based on `Solr` [1], with the default parameter settings. For English, we use the retrieval system provided by ClueWeb12.

3.2 Query set

The queries for Chinese subtask are sampled from a commercial search engine's query logs in one day of March 2017. Almost all the queries are torso queries, which means that their frequencies are between 10 to 1000 one day. Although the head and tail queries also need investigation, we believe that the torso queries are most appropriate for such an evaluation 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 2.

The queries for English subtask come from two sources. The first part is the translations of some Chinese queries. Although WWW is not a task for cross language information retrieval (CLIR), the data (relevance judgments, runs etc.) may potentially benefit CLIR research in the future. The second part is the queries sampled from another international search engines (note it is different from the search engine used in Chinese subtask). This search engine's users are mainly located in English speaking countries. The query logs we used is a small subset of one day's records. Thus we randomly sampled some queries whose frequencies are between 1 and 100. The content of the queries, the intent

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

types and whether the queries are shared by Chinese subtask are presented in Table 3.

For both English and Chinese query set, we did not use a lot of navigational queries. Since both SogouT and Clueweb are small subsets of the entire Web, it is very likely that the perfect answer for a navigational query is not in the corpus.

It should be noted that during the relevance assessment process, we find that the 0014 query for English query set is misspelled as "equation edior". The correct spelling is "equation editor". We keep the original spelling as released to the participants.

3.3 Training data

For the Chinese Subtask, we provide a user behavior collection for training purpose. The behavior collection includes 2 parts.

For the training set, we have 200 queries which have no overlap with the query set of Chinese subtask. For each query, we provide users' clicks, the URLs of presented results, as well as the dwell time on each clicked results.

More specifically, for each entry in the training set. We have

`anonymized User ID query a list of URLs presented to the users clicked urltimestamps of actions`

We also provide some relevance annotations for each query. The relevance annotations were made by professional assessors from the search engine's quality evaluation department.

For the queries in query set of Chinese subtask, we provide similar behavior data, except for relevance judgments. All of these behavior data is collected by a commercial search engine from March 2016 to April 2016. Due to privacy concerns, the users' IDs are anonymized. For each query, at most 500 entries of behavior (500 sessions) are served, since we think 500 is enough for feature extraction and model training.

4. RUNS, POOLING AND RELEVANCE ASSESSMENTS

4.1 Received Runs

Table 4 summarises our run statistics.

4.2 Relevance assessments

The Chines relevance assessments were organised at Tsinghua University, China. The relevance judgments were conducted via a web-based system which was developed by an undergraduate student, Mr. Weixuan WU. All the documents were orginized as 25 annotation tasks. Each task contains about 800 documents which are belonging to at most two queries. There is no overlap between different task. We hired 51 assessors in the campus via posters, maillist as well as social networks. 37 of the 51 assessors have finished only one task while the remaining ones have finished multiple tasks (the most hard-working assessor have finished 5 tasks). Each task takes about two hours and the assessors will receive about 200 RMB (about 30 USD) for each task. We encourage the participants to take as many tasks as they can since we believe the more documents they have judged, the more stable their inner relevance models are.

The assessments were conducted in a lab-environment. Before entering the assessment session, the assessors will first take an instruction (about 15 minutes) about the relevance judgment criteria:

- **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.

Although the assessors we hired may not as stable as trained professional assessors, we found that it is much faster with acceptable quality. Finally, NONREL labels were mapped to zero; MARGREL labels were mapped to one; REL labels were mapped to two and HIGHREL labels were mapped to THREE.

The English relevance assessments were organised at Waseda University, Japan, using a web-based relevance assessment developed by the Sakai Laboratory of the same university, called PLY[2]. Nine main assessors were hired through a Japanese crowdsourcing service called Lancers; for 50 odd-numbered topics, we additionally hired five students for the purpose of studying inter-assessor consistency between crowd workers and students. The official qrels do not reflect the judgments of the students. Each assessor was shown only the queries on the judgment interface: no additional information such as description and narrative fields were provided. The relevance assessment criteria given to each assessor were as follows:

- **ERROR** The right panel does not show any contents at all, even after waiting for a few seconds for the content to load.
- **H.REL** Highly relevant - it is *likely* that the user who entered this search query will find this page relevant.
- **REL** Relevant - it is *possible* that the user who entered this search query will find this page relevant.
- **NONREL** Nonrelevant - it is *unlikely* that the user who entered this search query will find this page relevant.

Finally, ERROR and NONREL were mapped to zero, REL was mapped to one, and H.REL was mapped to two, and the relevance levels $L4$ through $L0$ were obtained by summing the judgments of the two assessors for each topic.

Table 5 summarises our relevance assessment statistics.

5. EVALUATION MEASURES AND TOOLS

We used the NTCIREVAL tool[3] to compute MSnDCG@10 (Microsoft version of nDCG at cutoff 10), Q@10 (Q-measure

[2] Authors: Xiao Peng, Lingtao Li, and Yimeng Fang.

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

Table 2: 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)

qid	Query	Int.	Trans.	qid	Query	Int.	Trans.	qid	Query	Int.	Tran
0001	ascii 码		Y	0035	囚徒健身		Y	0068	泰山简介		
0002	CAD		Y	0036	四川大学	NAV	Y	0069	济公		
0003	fifa	NAV	Y	0037	地暖价格			0070	海王星		Y
0004	nike	NAV	Y	0038	坦克大战经典版			0071	清远马拉松		
0005	pets 考试			0039	多米诺骨牌		Y	0072	港币对人民币汇率		Y
0006	vmware 虚拟机		Y	0040	姚文元			0073	湿疹是怎么引起的		
0007	万年历查询		Y	0041	孤岛危机			0074	炫酷网名		
0008	三星手机官网	NAV	Y	0042	学雷锋			0075	物权法	NAV	
0009	上海公交查询			0043	宝马 x7			0076	电影排行榜		Y
0010	世乒赛		Y	0044	家和万事兴			0077	登录路由器		Y
0011	东方时空			0045	对外经济贸易大学	NAV		0078	百香果		
0012	中华小当家			0046	少林寺		Y	0079	视频合并		Y
0013	书信格式		Y	0047	尤金		Y	0080	神探狄仁杰		
0014	传统节日			0048	工作总结开头		Y	0081	空腹吃苹果好吗		Y
0015	侏罗纪世界		Y	0049	巧克力的英文			0082	第九套广播体操		
0016	保定市招聘信息			0050	广东外语外贸大学	NAV		0083	羊驼		Y
0017	信用卡申请		Y	0051	开场舞			0084	联塑管业官网	NAV	
0018	儿童简笔画大全		Y	0052	张震岳经典歌曲			0085	脂肪肝的饮食禁忌		
0019	元素周期表	NAV	Y	0053	打字练习		Y	0086	芒果		
0020	公式编辑器	NAV	Y	0054	拍立得相机哪款好			0087	花瓣		Y
0021	养老金并轨			0055	文言文翻译			0088	英语名言		Y
0022	农用汽车		Y	0056	有关黄河的诗句			0089	藏头诗		
0023	出国移民		Y	0057	机动车违章		Y	0090	褒义词		Y
0024	动态图片大全		Y	0058	机器人		Y	0091	论语十二章		
0025	北京四合院			0059	梦里花落知多少			0092	走进春天		
0026	十二星座		Y	0060	植树节的来历			0093	道光皇帝		
0027	十面埋伏		Y	0061	樱桃		Y	0094	长安汽车		
0028	历任黑龙江省省长			0062	樱桃小丸子		Y	0095	青苹果乐园		
0029	口红怎么涂好看		Y	0063	欧冠决赛		Y	0096	音符		Y
0030	可可西里			0064	欧洲步		Y	0097	飞越疯人院		Y
0031	周杰伦演唱会			0065	比特币			0098	高汤的做法		
0032	哈尔滨旅游			0066	氢氧化镁		Y	0099	魔方还原步数		Y
0033	唐伯虎点秋香			0067	河北合并县			0100	墨尔本大学世界排名		Y
0034	唐老鸭		Y								

at cutoff 10), and nERR@10 (normalised expected reciprocal rank at cutoff 10) [3]. Linear gain values were used, e.g., 9 for $L9$-relevant, 1 for $L1$-relevant.

The Discpower tool[4] was used to conduct randomised Tukey HSD tests, each with $B = 10,000$ trials [3].

6. OFFICIAL RESULTS

6.1 Chinese Run Results

Table 6 shows the mean effectiveness scores for all Chinese runs. Table 7 summarises the statistical significance test results. Randomised Tukey HSD p-values and effect sizes (i.e., standardised mean differences) based on two-way ANOVA (without replication) residual variances (0.0279 for MSnDCG@10, 0.0315 for Q@10, and 0.0466 for nERR@10) are also shown [4]. For example, the effect size for the difference between RUCIR-C-NU-Base-1 and THUIR-C-CU-Base-1 in terms of MSnDCG@10 is given by $ES_{HSD} = (0.6323 - 0.4828)/\sqrt{0.0279} = 0.895$.

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

From the official Chinese results with the three evaluation measures, it can be observed that:

- RUCIR and CMUIR are the top performing teams, in that they both statistically significantly outperforms THUIR and SLWWW, and are not statistically significantly different from each other;
- THUIR statistically significantly outperforms SLWWW.

Table 8 compares the system rankings according to the three evaluation measures in terms of Kendall's τ, and their 95% confidence intervals. It can be observed that the three rankings are statistically equivalent.

6.2 English Run Results

Table 9 shows the mean effectiveness scores for all English runs. Table 10 summarises the statistical significance test results. Randomised Tukey HSD p-values and effect sizes (i.e., standardised mean differences) based on two-way ANOVA (without replication) residual variances (0.0297 for MSnDCG@10, 0.0360 for Q@10, and 0.0520 for nERR@10) are also shown [4].

Table 3: English 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 from Chinese)

qid	Query	Int.	Trans.	qid	Query	Int.	Trans.	qid	Query	Int.	Tra
0001	ascii code		Y	0035	Magnesium hydroxide		Y	0068	dell stock	NAV	
0002	CAD		Y	0036	Neptune		Y	0069	diwali		
0003	fifa	NAV	Y	0037	hkd rmb exchange rate		Y	0070	dna strand		
0004	nike	NAV	Y	0038	movie ranking	NAV	Y	0071	dog food for allergies		
0005	vmware virtual machine		Y	0039	router login		Y	0072	driving school		
0006	calendar		Y	0040	merge videos		Y	0073	drum		
0007	samsung official site	NAV	Y	0041	autumn			0074	EARNINGS CALENDAR		
0008	World Table Tennis Championships		Y	0042	Alpaca		Y	0075	famous black leaders		
0009	letter format		Y	0043	petal		Y	0076	financial engines		
0010	Jurassic World		Y	0044	English quotes		Y	0077	find part time job		
0011	credit card application		Y	0045	commendatory term		Y	0078	formal fallacy		
0012	child stick figures		Y	0046	musical note		Y	0079	grasslands		
0013	periodic table		Y	0047	rubik cube solution steps		Y	0080	hp printer offline		
0014	equation edior		Y	0048	melbourne university world ranking		Y	0081	ibm quote		
0015	agricultural machinery		Y	0049	yahoo finance	NAV		0082	itunes error		
0016	migrate abroad		Y	0050	dow jones			0083	jetstar airlines hong kong		
0017	gif collection		Y	0051	Volkswagen	NAV		0084	key man insurance		
0018	Astrological sign		Y	0052	1968 olympic coin value			0085	largest species of eel		
0019	House of Flying Daggers		Y	0053	absolute neutrophils			0086	low monocytes		
0020	Donald Duck		Y	0054	Anime pillow			0087	manila		
0021	Convict Conditioning		Y	0055	annual salary requirement			0088	mexico climate		
0022	Sichuan University	NAV	Y	0056	apologetic songs			0089	Mineral Element		
0023	Battle City		Y	0057	axle ratio			0090	native American Mexican		
0024	domino		Y	0058	best office software			0091	openwrt		
0025	Shaolin Monastery		Y	0059	bios setup			0092	pandora	NAV	
0026	Eugene		Y	0060	blueberry compote			0093	protecting embankment		
0027	Introduction of work report		Y	0061	boeing history			0094	Samosa Recipes		
0028	typing practice		Y	0062	brady motion			0095	soda water		
0029	Traffic Violation		Y	0063	candle in window meaning			0096	Star Wars Movies		
0030	robot		Y	0064	CaSe compound			0097	stomach disorder		
0031	cherry		Y	0065	cheap root canals			0098	tiffany keys	NAV	
0032	Chibi Maruko-chan		Y	0066	create website			0099	vegetable fermentation		
0033	UEFA Champions League final		Y	0067	recital themes			0100	weight loss		
0034	Euro Step		Y								

Table 4: Run statistics.

Team	Chinese	English	total
CMUIR	5	-	5
RMIT	-	4	4
RUCIR	5	5	10
SLWWW	4	-	4
THUIR	5	4	9
total	19 (4 teams)	13 (3 teams)	32

From the official English results with nDCG@10 and with Q@10, it can be observed that RMIT is the top performing team, in that it statistically significantly outperforms THUIR and RUCIR. On the other hand, the three teams are statistically equivalent in terms of nERR@10.

Table 11 compares the system rankings according to the three evaluation measures in terms of Kendall's τ, and their 95% confidence intervals. It can be observed that the three rankings are statistically equivalent.

7. FURTHER DISCUSSIONS

The original motivation for launching WWW contains two parts: (1) The Web track at TREC was terminated. However, we believe it is still a necessity to have a testbed to monitor the progress of searching techniques, especially given the rapid development of neural IR methods; (2) We wanted to quantify the progress of web search algorithms across several rounds of NTCIR, especially by leveraging score standardisation, a technique for making all topics comparable based on a known set of systems.

Unfortunately, though quite a few teams (20) registered for WWW, only 5 teams (including 4 teams from the organisers' institutions) participated in the end. This prevents us from conducting valid score standardisation experiments, because this technique relies on a large set of systems to ensure that a standardised score (e.g. standardised nDCG) of 0.5 means an "average" system. The pre-NTCIR-13 failure analysis workshop was also cancelled.

One of the main reasons for the poor participation might be the lack of training data for machine-learning-based approaches to Web search. Recently researchers are mainly focusing on methods based on neural networks, which are very data hungry approaches. In the future rounds of WWW, we plan to provide more training data to participants. We are also seeking cooperation with companies from industry.

8. REFERENCES

Table 5: Relevance assessment statistics.

	Chinese	English
#topics	100	100
#assessors/topic	3	2 (3 for odd-number topic IDs)
Pool depth	20	30
Total #docs pooled	20,400	22,912
Total $L9$-relevant	1,405	-
Total $L8$-relevant	1,608	-
Total $L7$-relevant	1,848	-
Total $L6$-relevant	2,052	-
Total $L5$-relevant	2,124	-
Total $L4$-relevant	2,017	1,583
Total $L3$-relevant	2,176	3,866
Total $L2$-relevant	1,822	4,329
Total $L1$-relevant	2,127	4,751
Total $L0$	3,221	8,383

[1] The clueweb12 dataset – the lemur project. http://www.lemurproject.org/clueweb12.php, 2012. Online; Accessed: 2017-02-01.

[2] C. Luo, Y. Zheng, Y. Liu, X. Wang, J. Xu, M. Zhang, and S. Ma. Sogout-16: A new web corpus to embrace ir research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1233–1236, New York, NY, USA, 2017. ACM.

[3] T. Sakai. Metrics, statistics, tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*, pages 116–163, 2014.

[4] T. Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, 2014.

Table 6: Official Chinese results.

Run	Mean nDCG@10	Run	Mean Q@10	Run	Mean nERR@10
RUCIR-C-NU-Base-1	0.6323	RUCIR-C-NU-Base-1	0.6449	RUCIR-C-NU-Base-1	0.7771
RUCIR-C-NU-Base-2	0.6241	RUCIR-C-NU-Base-2	0.6448	RUCIR-C-NU-Base-2	0.7597
CMUIR-C-NU-Base-1	0.6145	CMUIR-C-NU-Base-1	0.6294	CMUIR-C-NU-Base-1	0.7583
CMUIR-C-NU-Base-3	0.6059	CMUIR-C-NU-Base-3	0.6163	CMUIR-C-NU-Base-3	0.7406
CMUIR-C-NU-Base-5	0.5915	RUCIR-C-NU-Base-4	0.6049	CMUIR-C-NU-Base-5	0.7372
RUCIR-C-NU-Base-4	0.5873	CMUIR-C-NU-Base-5	0.5996	RUCIR-C-NU-Base-4	0.7217
CMUIR-C-NU-Base-2	0.5873	CMUIR-C-NU-Base-2	0.5955	RUCIR-C-NU-Base-5	0.7132
RUCIR-C-NU-Base-5	0.5827	RUCIR-C-NU-Base-5	0.5890	CMUIR-C-NU-Base-4	0.7086
CMUIR-C-NU-Base-4	0.5667	CMUIR-C-NU-Base-4	0.5780	CMUIR-C-NU-Base-2	0.7046
RUCIR-C-NU-Base-3	0.5361	RUCIR-C-NU-Base-3	0.5407	RUCIR-C-NU-Base-3	0.6767
THUIR-C-CU-Base-1	0.4828	THUIR-C-CU-Base-1	0.4942	THUIR-C-CU-Base-1	0.6443
THUIR-C-CU-Base-5	0.4258	THUIR-C-CU-Base-5	0.4335	THUIR-C-CU-Base-3	0.5717
THUIR-C-CU-Base-4	0.4258	THUIR-C-CU-Base-4	0.4335	THUIR-C-CU-Base-5	0.5695
THUIR-C-CU-Base-2	0.4179	THUIR-C-CU-Base-2	0.4235	THUIR-C-CU-Base-4	0.5695
THUIR-C-CU-Base-3	0.4137	THUIR-C-CU-Base-3	0.4144	THUIR-C-CU-Base-2	0.5626
SLWWW-C-NU-Base-2	0.3225	SLWWW-C-NU-Base-2	0.3099	SLWWW-C-NU-Base-1	0.4753
SLWWW-C-NU-Base-1	0.3206	SLWWW-C-NU-Base-1	0.3094	SLWWW-C-NU-Base-2	0.4723
SLWWW-C-NU-Base-4	0.2991	SLWWW-C-NU-Base-4	0.2949	SLWWW-C-NU-Base-4	0.4406
SLWWW-C-NU-Base-3	0.2909	SLWWW-C-NU-Base-3	0.2838	SLWWW-C-NU-Base-3	0.4327

Table 7: Statistical significance with the best Chinese run from each team (Randomised Tukey HSD test, $B = 10,000, \alpha = 0.05$).

These runs are	Significantly better than these runs in terms of mean nDCG@10
RUCIR-C-NU-Base-1	THUIR-C-CU-Base-1 ($p = 0.0001$, $ES_{HSD} = 0.895$), SLWWW-C-NU-Base-2 ($p = 0$, $ES_{HSD} = 1.855$)
CMUIR-C-NU-Base-1	THUIR-C-CU-Base-1 ($p = 0.0004$, $ES_{HSD} = 0.789$), SLWWW-C-NU-Base-2 ($p = 0$, $ES_{HSD} = 1.748$)
THUIR-C-CU-Base-1	SLWWW-C-NU-Base-2 ($p = 0$, $ES_{HSD} = 0.960$)
These runs are	Significantly better than these runs in terms of mean Q@10
RUCIR-C-NU-Base-1	THUIR-C-CU-Base-1 ($p = 0.0001$, $ES_{HSD} = 0.849$), SLWWW-C-NU-Base-2 ($p = 0$, $ES_{HSD} = 1.888$)
CMUIR-C-NU-Base-1	THUIR-C-CU-Base-1 ($p = 0.0005$, $ES_{HSD} = 0.761$), SLWWW-C-NU-Base-2 ($p = 0$, $ES_{HSD} = 1.800$)
THUIR-C-CU-Base-1	SLWWW-C-NU-Base-2 ($p = 0$, $ES_{HSD} = 1.039$)
These runs are	Significantly better than these runs in terms of mean nERR@10
RUCIR-C-NU-Base-1	THUIR-C-CU-Base-1 ($p = 0.0014$, $ES_{HSD} = 0.615$), SLWWW-C-NU-Base-1 ($p = 0$, $ES_{HSD} = 1.398$)
CMUIR-C-NU-Base-1	THUIR-C-CU-Base-1 ($p = 0.0112$, $ES_{HSD} = 0.528$), SLWWW-C-NU-Base-1 ($p = 0$, $ES_{HSD} = 1.311$)
THUIR-C-CU-Base-1	SLWWW-C-NU-Base-1 ($p = 0$, $ES_{HSD} = 0.783$)

Table 8: Kendall's τ values with 95% CIs (19 Chinese runs).

	Mean Q@10	Mean nERR@10
Mean nDCG@10	0.988 [0.630, 1.047]	0.930 [0.648, 1.044]
Mean Q@10	-	0.918 [0.599, 1.078]

Table 9: Official English results.

Run	Mean nDCG@10	Run	Mean Q@10	Run	Mean nERR@10
RMIT-E-NU-Own-1	0.6302	RMIT-E-NU-Own-1	0.6548	RMIT-E-NU-Own-1	0.7463
THUIR-E-PU-Base-3	0.5679	RMIT-E-NU-Own-4	0.5657	RMIT-E-NU-Own-4	0.7428
RMIT-E-NU-Own-4	0.5626	RMIT-E-NU-Own-3	0.5657	THUIR-E-PU-Base-3	0.7118
RMIT-E-NU-Own-2	0.5504	RMIT-E-NU-Own-2	0.5633	RMIT-E-NU-Own-2	0.7055
RMIT-E-NU-Own-3	0.5493	THUIR-E-PU-Base-3	0.5570	RUCIR-E-NU-Base-1	0.6988
THUIR-E-PU-Base-2	0.5360	THUIR-E-PU-Base-1	0.5369	RMIT-E-NU-Own-3	0.6977
THUIR-E-PU-Base-1	0.5323	THUIR-E-PU-Base-2	0.5304	THUIR-E-PU-Base-1	0.6754
RUCIR-E-NU-Base-1	0.5254	RUCIR-E-NU-Base-1	0.5135	THUIR-E-PU-Base-2	0.6744
RUCIR-E-NU-Base-3	0.4516	RUCIR-E-NU-Base-3	0.4402	RUCIR-E-NU-Base-3	0.5917
RUCIR-E-NU-Base-2	0.4207	RUCIR-E-NU-Base-2	0.4050	RUCIR-E-NU-Base-2	0.5795
RUCIR-E-NU-Base-5	0.3885	RUCIR-E-NU-Base-4	0.3859	RUCIR-E-NU-Base-4	0.5343
RUCIR-E-NU-Base-4	0.3843	RUCIR-E-NU-Base-5	0.3813	RUCIR-E-NU-Base-5	0.5292
THUIR-E-PU-Base-4	0.3157	THUIR-E-PU-Base-4	0.3018	THUIR-E-PU-Base-4	0.4648

Table 10: Statistical significance with the best English run from each team (Randomised Tukey HSD test, $B = 10,000, \alpha = 0.05$).

This run is	Significantly better than these runs in terms of mean nDCG@10
RMIT-E-NU-Own-1	THUIR-E-PU-Base-3 ($p = 0.045$, $ES_{HSD} = 0.361$), RUCIR-E-NU-Base-1 ($p = 0.0006$, $ES_{HSD} = 0.608$)
This run is	Significantly better than these runs in terms of mean Q@10
RMIT-E-NU-Own-1	THUIR-E-PU-Base-3 ($p = 0.0033$, $ES_{HSD} = 0.516$), RUCIR-E-NU-Base-1 ($p = 0$, $ES_{HSD} = 0.745$)
This run is	Significantly better than these runs in terms of mean nERR@10
N/A	N/A

Table 11: Kendall's τ values with 95% CIs (13 English runs).

	Mean Q@10	Mean nERR@10
Mean nDCG@10	0.846 [0.630, 1.047]	0.846 [0.648, 1.044]
Mean Q@10	-	0.846 [0.599, 1.078]