# Overview of NTCIR-8 ACLIA IR4QA

Tetsuya Sakai† Hideki Shima* Noriko Kando‡ Ruihua Song†
Chuan-Jie Lin⋆ Teruko Mitamura* Miho Sugimoto‡ Cheng-Wei Lee♭
†Microsoft Research Asia *Carnegie Mellon University
‡National Institute of Informatics ⋆National Taiwan Ocean University
♭Academia Sinica
tetsuyasakai@acm.org

## Abstract

*This paper presents an overview of NTCIR-8 ACLIA (Advanced Cross-lingual Information Access) IR4QA (Information Retrieval for Question Answering). Following the task definitions of the first IR4QA at NTCIR-7 [13, 15], IR4QA at NTCIR-8 evaluates cross-language IR using English topics and targetting documents in Simplified Chinese, Traditional Chinese or Japanese. The corresponding monolingual IR subtasks are also within its scope. The only difference between traditional "ad hoc" IR tasks and IR4QA is that the latter can optionally be seen as a component of a question answering system. This paper describes the task, how the organisers collaborated with 12 participating teams (who submitted a total of 84 runs) to obtain relevance assessments for our three IR4QA test collections, the formal evaluation results, and the "run ranking forecasts" that were provided to the participants right after the submission deadline. For the relationship between IR4QA and the entire ACLIA, we refer the reader to the overview papers of ACLIA [7, 8]. For details of the individual IR4QA systems, we refer the reader to the participants' reports.*
**Keywords:** *test collections, pooling, evaluation metrics, evaluation package, qrels, pseudo-qrels.*

## 1. Introduction

This paper presents an overview of NTCIR-8 ACLIA (Advanced Cross-lingual Information Access) IR4QA (Information Retrieval for Question Answering). Following the task definitions of the first IR4QA at NTCIR-7 [13, 15], IR4QA at NTCIR-8 evaluates cross-language IR using English topics and targetting documents in Simplified Chinese (CS), Traditional Chinese (CT) or Japanese (JA). The corresponding monolingual IR subtasks are also within its scope. This paper describes the task, how the organisers collaborated with 12 participating teams (who submitted a total of 84 runs) to obtain relevance assessments for our three IR4QA test collections, the formal evaluation results, and the "run ranking forecasts" that were provided to the participants right after the submission deadline. For the relationship between IR4QA and the entire ACLIA, we refer the reader to the overview paper of ACLIA [8]. For details of the individual IR4QA systems, we refer the reader to the participants' reports [1, 3, 4, 5, 6, 9, 17, 19, 20, 21, 23, 24]. Table 1 provides a list of IR4QA participants.

**Table 1. IR4QA participants.**

| team name | organisation |
|---|---|
| BRKLY | University of California, Berkeley |
| CYUT | Chaoyang University of Technology |
| DCU | Dublin City University |
| DLUT | Dalian University of Technology |
| IMU | Inner Mongolia University |
| KDEG | Trinity College Dublin |
| KECIR | Shenyang Institute of Aeronautical Engineering |
| LTI | Carnegie Mellon University |
| QUTIS | Queensland University of Technology |
| WHUCC | Computer School, Wuhan University |
| WHUQA | Wuhan University |
| WUST | Wuhan University of Science and Technology |

The important dates for IR4QA were as follows:

| | |
|---|---|
| *January 6* | *Topics released* |
| *January 20-22* | *Runs received* |
| *January 21-24* | *Run ranking forecasts released* |
| *March 3* | *CS and JA evaluation results released (before bug fix)* |
| *March 8* | *CT pool-depth-50 evaluation results released* |
| *March 24* | *CT pool-depth-100 evaluation results released* |
| *April 19* | *CS and JA results released (after bug fix)* |

The CS and JA evaluation results were first released on March 3, but it was later discovered that the "qrels" (relevance assessment data) contained duplicate documents per topic. We therefore fixed these qrels and released the correct CS and JA results on April 19. While the CT qrels file did not have this problem, the assessors went behind schedule so we had to release pool-depth-50 results first on March 8, and then pool-depth-100 results on March 24. (The CS and JA qrels were based on depth-100 pools from the very beginning.) The pooling procedure will be described in the next section.

What is unique about IR4QA at NTCIR-8 is that run ranking forecasts based on *pseudo-qrels* [13, 16, 18] were released to participants right after the organisers received all the runs, in the hope that the participants can do some useful experiments while they wait for the "true" qrels and evaluation results to arrive. A

*NTCIR-8 Workshop Meeting, 2010,* Tokyo, Japan.

pseudo-qrels file looks just like a regular qrels file, but is generated completely automatically based on "how popular each document is among the submitted runs." Details will follow.

The remainder of this paper is organised as follows. Section 2 describes how the runs were pooled and how the relevance assessments were obtained for each of our three IR4QA test collections. Section 3 defines the IR effectiveness metrics we use for evaluating runs, as well as the tool developed for this purpose. It also defines some additional statistics for examining the runs, as well as rank correlation metrics for comparing a pair of rankings. Section 4 reports on the IR4QA evaluation results based on the true qrels. For CS and JA, results both before and after the bug fix of the qrels are shown. For CT, both pool-depth-50 and pool-depth-100 results are shown. Section 5 reports on our run ranking forecasts based on pseudo-qrels, which were released to the participants prior to relevance assessments. Finally, Section 6 summarises our initial findings as the organisers of IR4QA at NTCIR-8.

## 2. Pooling and Relevance Assessments

Table 2 shows the number of runs submitted by each participating team for different language pairs: A *run* is a system output file containing a ranked list of documents for each topic (i.e. search request)[1]. For example, a total of 20 CS-CS monolingual runs (i.e., runs that used Simplified Chinese topics and retrieved Simplified Chinese documents) and a total of 28 EN-CS crosslingual runs (i.e. runs that used English topics and retrieved Simplified Chinese documents) were submitted. Hence a total of 48 CS runs (i.e. runs that retrieved Simplified Chinese documents) were used in pooling for relevance assessments.

Let $S$ be the set of systems (i.e. runs) that will contribute to the pool, and let $s \in S$. For a particular topic, let $D_X(s)$ denote the set of top $X$ documents from $s$. The depth-$X$ pool for this topic is defined as $\bigcup_{s \in S} D_X(s)$.

We created depth-100 pools for relevance assessments, but instead of assessing pooled documents after a sort by document IDs, we proceeded as follows:

1. Create a depth-$X$ pool for every topic ($X \in \{50, 100\}$).
2. Within each depth-$X$ pool, sort the documents by the number of runs containing the document at or above rank $X$ (the larger the better), and then break ties by the sum of ranks of that document within those runs (the smaller the better). That is, documents retrieved by many runs at high ranks are placed near the top of the *sorted pool* [13].
3. From each depth-100 sorted pool, remove all documents that occur in the corresponding depth-50 pool: for convenience, we call the resultant list a *sorted residual pool*.
4. For each topic, assign exactly one assessor[2]. Let the assessor assess the relevance of each document in the depth-50 sorted pool first, from top to bottom. When this is complete, let him/her assess the sorted residual pool.

The depth-100 pools were split as described above because the organisers were uncertain as to whether the assessors can finish in time. As it turned out, the CS and JA assessors did finish in time but the qrels files initially contained duplicate documents per topic. For CT, the progress was slower so we released the pool-depth-50 results first to the participants.

Tables 26-28 in the Appendix show the pool size for each topic for each document collection. Among these topics, ACLIA2-CS-{0001, 0003, 0004, 0007, 0008, 0010, 0011, 0013, 0014, 0017, 0020, 0021, 0022, 0024, 0025, 0027, 0031, 0034, 0037, 0050, 0055, 0060, 0079, 0081, 0086, 0088, 0089} and ACLIA2-JA-{0037, 0043, 0059, 0066, 0071, 0082} were removed from our evaluation data as the number of relevant documents found in the depth-100 pool was fewer than five for these topics. Hence CT and JA runs were evaluated using 73 and 94 topics, respectively. As for the CT topics, for the evaluation based on the depth-50 pools, we removed ACLIA2-CT-{0018, 0038, 0041, 0054, 0062, 0072, 0078, 0079, 0080, 0081, 0085, 0087, 0098, 0100} and therefore used 86 topics. Later, for the evaluation based on the depth-100 pools, we recovered ACLIA2-CT-0087 and therefore used 87 topics in the end.

The relevance assessments for the CS, CT and JA topics were separately conducted at Microsoft Research Asia, National Taiwan Ocean University and National Institute of Informatics, respectively. Following IR4QA at NTCIR-7, we used graded relevance:

$L2$ Relevant ($L2$-relevant). The document fully satisfies the information need expressed in the topic.

$L1$ Partially relevant ($L1$-relevant). The document only partially satisfies the information need expressed in the topic.

$L0$ Not relevant.

In our evaluation, both $L0$ (i.e. judged nonrelevant) documents and *unjudged* documents (i.e. documents that were not included in the pool) are both regarded as nonrelevant.

Tables 29-31 in the Appendix show the number of judged relevant/nonrelevant documents per topic for each of our three test collections (after the bug fix for CS and JA, and after completing the pool-depth-100 assessments for CT).

## 3. Metrics and Tools

### 3.1 Ranked Retrieval Effectiveness

Following the first round of IR4QA at NTCIR-7, we used three effectiveness metrics for evaluating the IR4QA runs: *Average Precision* (AP), Q-measure (Q) and a "Microsoft version" of normalised Discounted Cumulative Gain (nDCG) [13]. Among these three, only Q and nDCG can utilise graded relevance assessments.

For a particular topic, let $I(r)$ be a flag indicating whether the document retrieved at rank $r$ in a given run is relevant or not, and let $C(r) = \sum_{i=1}^{r} I(i)$. Let $R$ denote the number of known relevant documents for this topic, including partially relevant ones. Then, the AP of this run for this topic is given by:

$$AP = \frac{1}{R} \sum_{r} I(r) \frac{C(r)}{r} . \quad (1)$$

Let $\mathcal{L}$ be a relevance level, and let $gain(\mathcal{L})$ denote the *gain value* for retrieving an $\mathcal{L}$-relevant document. For the IR4QA data,

[1] The IR4QA run file format can be found in [13], Figure 4 (page 81).

[2] The assessors did not create the topics. The topic development procedure is described in the ACLIA2 overview paper [8].

**Table 2. Number of NTCIR-8 IR4QA runs submitted.**

| team | CS-CS | EN-CS | CT-CT | EN-CT | JA-JA | EN-JA |
|---|---|---|---|---|---|---|
| BRKLY | | | | | 5 | 2 |
| CYUT | | 4 | | 4 | | 4 |
| DCU | 2 | 3 | | | | |
| DLUT | | 3 | | | | |
| IMU | 3 | 1 | | | | |
| KDEG | 5 | 5 | 5 | 5 | | |
| KECIR | 5 | | | | | |
| LTI | | | | | 3 | 3 |
| QUTIS | | 5 | | 5 | | |
| WHUCC | 2 | 3 | | | | |
| WHUQA | 2 | 2 | | | | |
| WUST | 1 | 2 | | | | |
| total by lang. pair | 20 | 28 | 5 | 14 | 8 | 9 |
| total by document lang. | 48 | | 19 | | 17 | |

we have $L2$-relevant ("relevant") and $L1$-relevant ("partially relevant") documents, in addition to judged nonrelevant documents whose relevance level is denoted by $L0$. We let $gain(L2) = 2$ and $gain(L1) = 1$ throughout our analysis. Let $R(\mathcal{L})$ denote the number of known $\mathcal{L}$-relevant documents for a topic, so that $\sum_{\mathcal{L}} R(\mathcal{L}) = R$. Let $g(r) = gain(\mathcal{L})$ if the document at rank $r$ is $\mathcal{L}$-relevant and let $g(r) = 0$ otherwise. In particular, let $g^*(r)$ denote the gain at rank $r$ of an *ideal* ranked output, where an ideal ranked output for a particular topic is one that satisfies $I(r) = 1$ for $1 \leq r \leq R$ and $g(r) \leq g(r-1)$ for $r > 1$. For the IR4QA data, this can be achieved by listing up all $L2$-relevant documents, and then all $L1$-relevant documents.

The *cumulative gain* [2] at rank $r$ is defined as $cg(r) = \sum_{i=1}^{r} g(i)$. Similarly, let $cg^*(r) = \sum_{i=1}^{r} g^*(i)$. Let $\beta$ be a positive constant. Q is defined as:

$$Q\text{-}measure = \frac{1}{R} \sum_{r} I(r) \frac{C(r) + \beta cg(r)}{r + \beta cg^*(r)} \ . \qquad (2)$$

Letting $\beta = 0$ reduces Q to AP, and using a large $\beta$ makes Q more forgiving to relevant documents found near the bottom of the ranked list [11]. We let $\beta = 1$ throughout our analysis. Given flat gain values (i.e., binary relevance), $Q = AP$ holds iff there is no relevant document below rank $R$; $Q > AP$ holds iff there is at least one relevant document below rank $R$ [10]. Sakai and Robertson [14] have discussed how AP and Q can be interpreted from the viewpoint of a user population.

Let $l$ be a document cut-off value; we let $l = 1000$ throughtout our analysis. The version of nDCG we use is defined as:

$$nDCG = \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)} \ . \qquad (3)$$

See [13] for a discussion on different versions of nDCG.

We used the IR evaluation package `ir4qa_eval2.tar.gz` available at `http://research.nii.ac.jp/ntcir/tools/ir4qa_eval-en.html` for computing these evaluation metrics. The package is basically the same as `ir4qa_eval.tar.gz` that was used at NTCIR-7, and the version of nDCG described above is shown as "MSnDCG@1000" in its output files. For more information, we refer the reader to the README file contained in the package: it can also be used for evaluating runs from other tasks e.g. TREC ad hoc and NTCIR CLIR.

## 3.2 Coverage and Unique Relevants

In addition to measuring the effectiveness of ranked retrieval, we also examine the *coverage* of relevant documents and the number of *unique relevant* documents for each a run or a participating team as follows.

Let $REL$ be the set of relevant documents for a topic, so that $|REL| = R$. Let $D(s)$ denote the set of documents contained in run $s$ for the same topic. Now define $cvr(s) = D(s) \cap REL$: we refer to $|cvr(s)|$ as the coverage by run $s$ for that topic. (Note that coverage divided by $R$ gives recall.) Moreover, let $S(t)$ denote the set of runs submitted by team $t$, and let

$$CVR(t) = \bigcup_{s \in S(t)} cvr(s) \ .$$

We refer to $|CVR(t)|$ as the coverage by team $t$.

For a particular topic, the number of unique relevant documents found by run $s \in S(t)$ is given by:

$$ur(s) = |cvr(s) - \bigcup_{t' \neq t} CVR(t')| \ . \qquad (4)$$

Note that other runs from the same team $t$ do not hurt $ur(s)$. That is, for each run, we look at documents that were not found by any other *team*. Similarly, the number of unique relevant documents found by team $t$ is given by:

$$UR(t) = |CVR(t) - \bigcup_{t' \neq t} CVR(t')| \ . \qquad (5)$$

## 3.3 Rank Correlation

For measuring the similarity of a pair of run rankings or topic rankings (according to two different effectiveness metrics or qrels data), we use both *Kendall's $\tau$ rank correlation* and Yilmaz/Aslam/Robertson $\tau_{ap}$ [22].

Kendall's $\tau$ is a monotonic function of the probability that a *randomly chosen* pair of ranked items is ordered identically in the two rankings. Hence a swap near the top of a ranked list and that near the bottom of the same list has equal impact. Whereas, $\tau_{ap}$ is "top-heavy," in that it is a monotonic function of the probability that a randomly chosen item *and one ranked above it* are ordered identically in the two rankings. Like $\tau$, $\tau_{ap}$ lies between $-1$ and 1, but unlike $\tau$, it is not symmetrical: one of the input rankings is taken as the gold standard. When the errors (i.e. pairwise item swaps with respect to the gold standard) are uniformly distributed over the ranking being examined, $\tau_{ap}$ is equivalent to $\tau$.

Formally, let the size of the two ranked lists be $L$. Let $A$ be the number of item pairs that are ranked in the same order in both rankings, and let $B$ be the number of item pairs that are ranked in opposite order in the two rankings. Kendall's $\tau$ rank correlation is given by:

$$\tau = \frac{A - B}{L(L-1)/2} \ . \quad (6)$$

For a given ranked list to be examined, let $n(i)$ be the number of items *correctly* ranked above rank $i$ in the list with respect to a gold-standard ranked list. $\tau_{ap}$ is given by:

$$\tau_{ap} = \frac{2}{L-1} \sum_{i=2}^{L} \frac{n(i)}{i-1} - 1 \ . \quad (7)$$

## 4. Evaluation Results

This section presents the IR4QA evaluation results based on the "true" qrels, i.e. manual relevance assessment data. As mentioned earlier, for CS runs and JA runs, we present results both before and after the bug fix (based on depth-100 pools). For CT runs, we present both pool-depth-50 and pool-depth-100 results.

Tables 3-7 show the mean effectiveness values of all submitted CS, JA and CT runs, respectively. The runs have been sorted by each effectiveness metric. For CS and JA, the ranking after the bug fix has been compared with the one before the bug fix. Rank changes are indicated by arrows, where, for example, "↑ 2" means "up two ranks."

Table 8 compares each pair of run rankings according to two different metrics, in terms of $\tau$ and $\tau_{ap}$. For example, with the CS runs (before the bug fix), the $\tau$ between the AP-based ranking and the Q-based ranking is .970; the corresponding $\tau_{ap}$ is .961 when Q is taken as the gold standard and .960 when AP is taken as the gold standard. Since we have ample evidence that Q agrees well with AP and can handle graded relevance, we hereafter regard Q and nDCG as our primary metrics.

Table 9 presents our main results: For each metric, we selected the best *T-run*[3] from each team and sorted the runs by effectiveness; then, every adjacent pair in the sorted list was tested for statistical significance, using a two-sided *paired bootstrap test* [12][4]. Although we did not test statistical significance for *every* system pair, the table gives us a rough idea of how runs cluster together in terms of effectiveness. Rank changes before and after the bug fix are indicated by arrows: the only change is in the mean Q ranking, in which DLUT-EN-CS-03-T and QUTIS-EN-CS-04-T have been swapped. From this table, we can observe that:

- Among the 10 CS teams, KDEG, IMU, WHUCC, WHUQA, and DCU can be considered as the top performers (Note that Q and AP prefer KDEG to IMU, while nDCG prefers IMU);
- Among the 10 CS teams, KECIR, DLUT and QUTIS can be considered as middle performers; (Note that this middle group is not apparent in terms of AP, due to lack of a statistically significant difference with WUST. Also note that, after the bug fix, AP prefers DLUT to QUTIS, while Q and nDCG prefer QUTIS.)
- Among the 3 JA teams, LTI and BRKLY can be considered as the "top" performers (since the difference between these teams is not statistically significant);
- Similarly, among the 3 CT teams, KDEG is the top performer, and QUTIS is the middle performer.

Table 10 shows the *system descriptions* of the runs shown in the nDCG colums of Table 9. These texts were extracted from the submitted IR4QA run files[5].

We also examined the "hardness" of each topic, by averaging the per-topic metric values over all runs. Figures 1-3 visualise the average AP, Q and nDCG values per topic (using bug-fixed qrels for CS and JA, CT using pool-depth-100 qrels for CT). It can be observed that the cross-topic variances are very high.

Table 11 shows how two effectiveness metrics agree with each other when *topics* are ranked by the average effectiveness across runs. It can be observed that a topic ranking by Q is similar to that by AP. That is, a "hard" topic in terms of Q is also hard in terms of AP. Whereas, nDCG behaves a little differently. This is because Q and AP penalise low-recall ranked lists more heavily than nDCG does.

All of the above observations on correlations among the three metrics are consistent with previous findings (e.g. [13, 15]).

Tables 12-16 show the coverage of relevant documents summed across topics for each run and for each team (using bug-fixed qrels for CS and JA, CT using pool-depth-100 qrels for CT). "The best coverage awards" go to:

- KDEG for CS (for managing to cover 5180/5416=96% of known relevant documents listed up in Table 29);
- BRKLY for JA (for managing to cover 7463/8136=92% of know relevant documents listed up in Table 30);
- KDEG again for CT (for managing to cover 5258/5490=96% of known relevant documents listed up in Table 31).

The above percentages are extremely high, and suggest that the IR4QA test collections may be very incomplete: there may be many relevant documents not yet identified in the document collections.

Perhaps more interesting than coverage is the number of *unique* relevant documents found by a run or a team. Tables 17-21 show the statistics (all based on pool-depth-100 qrels). As can be seen, "the uniqueness awards" go to:

- QUTIS for CS;
- BRKLY for JA;
- KDEG for CT.

Table 17 deserves a particular attention: it shows that WHUQA, IMU, DLUT and CYUT did not contribute any unique relevant documents. Hence, even though the ACLIA2 IR4QA CS test collection was constructed based on run submissions from 10 teams, the relevant documents thus obtained are identical to those that

[3] A run that used only the information contained in the QUESTION field of each topic. The name of every such run has the "-T" suffix.

[4] Per-topic performance values of *all* submitted runs were released to the participants, so they can conduct a significance test of their choice for any of the runs.

[5] System descriptions of all runs were also released to the participants.

could have been obtained by relying on only 6 teams, namely, QUTIS, WHUCC, KECIR, WUST, KDEG and DCU.

Note, however, that the run contributions from the aforementioned 4 teams are still valuable: for NTCIR IR4QA, even these runs have the following effects:

- They affect the order of the relevance assessments, as assessors assess "popular" documents first (See Section 2);
- They affect the accuracy of our system ranking forecasts, as our pseudo-qrels also rely on popularity (See Section 5).

**Table 3. CS run results based on true qrels: mean effectiveness over 73 topics (BEFORE bug fix).**

| run | Mean AP | run | Mean Q | run | Mean nDCG |
|---|---|---|---|---|---|
| KDEG-CS-CS-02-DN | 0.4407 | KDEG-CS-CS-02-DN | 0.4779 | IMU-CS-CS-01-T | 0.6761 |
| KDEG-CS-CS-01-T | 0.4390 | KDEG-CS-CS-01-T | 0.4764 | KDEG-CS-CS-02-DN | 0.6727 |
| IMU-CS-CS-01-T | 0.4266 | IMU-CS-CS-01-T | 0.4628 | KDEG-CS-CS-01-T | 0.6674 |
| WHUCC-EN-CS-01-T | 0.4139 | WHUQA-CS-CS-01-T | 0.4528 | WHUQA-CS-CS-01-T | 0.6629 |
| WHUQA-CS-CS-01-T | 0.4128 | WHUCC-EN-CS-01-T | 0.4499 | IMU-CS-CS-02-T | 0.6580 |
| IMU-CS-CS-02-T | 0.4114 | IMU-CS-CS-02-T | 0.4480 | WHUQA-CS-CS-02-T | 0.6577 |
| DCU-CS-CS-01-T | 0.4111 | DCU-CS-CS-01-T | 0.4464 | IMU-CS-CS-03-T | 0.6575 |
| WHUQA-CS-CS-02-T | 0.4055 | WHUQA-CS-CS-02-T | 0.4458 | WHUCC-EN-CS-01-T | 0.6509 |
| IMU-CS-CS-03-T | 0.4032 | IMU-CS-CS-03-T | 0.4394 | DCU-CS-CS-01-T | 0.6466 |
| WHUCC-CS-CS-01-T | 0.3964 | WHUCC-CS-CS-01-T | 0.4323 | WHUCC-CS-CS-01-T | 0.6433 |
| WHUCC-EN-CS-03-T | 0.3912 | KDEG-CS-CS-03-T | 0.4244 | KDEG-CS-CS-03-T | 0.6346 |
| KDEG-CS-CS-03-T | 0.3867 | WHUCC-EN-CS-03-T | 0.4239 | WHUCC-EN-CS-02-T | 0.6345 |
| WHUCC-EN-CS-02-T | 0.3782 | WHUCC-EN-CS-02-T | 0.4163 | WHUCC-EN-CS-03-T | 0.6257 |
| KDEG-EN-CS-02-DN | 0.3776 | KDEG-EN-CS-02-DN | 0.4133 | WHUQA-EN-CS-01-T | 0.6139 |
| WHUQA-EN-CS-01-T | 0.3710 | WHUQA-EN-CS-01-T | 0.4085 | KDEG-EN-CS-02-DN | 0.6072 |
| WHUQA-EN-CS-02-T | 0.3555 | WHUQA-EN-CS-02-T | 0.3930 | WHUCC-CS-CS-02-T | 0.6060 |
| KDEG-CS-CS-04-T | 0.3530 | KDEG-CS-CS-05-T | 0.3889 | WHUQA-EN-CS-02-T | 0.5989 |
| KDEG-CS-CS-05-T | 0.3529 | KDEG-CS-CS-04-T | 0.3889 | KDEG-CS-CS-05-T | 0.5987 |
| WHUCC-CS-CS-02-T | 0.3502 | WHUCC-CS-CS-02-T | 0.3848 | KDEG-CS-CS-04-T | 0.5985 |
| KECIR-CS-CS-03-T | 0.3333 | KDEG-EN-CS-01-T | 0.3680 | KECIR-CS-CS-03-T | 0.5903 |
| DLUT-EN-CS-03-T | 0.3312 | KECIR-CS-CS-03-T | 0.3658 | QUTIS-EN-CS-04-T | 0.5882 |
| DLUT-EN-CS-02-T | 0.3312 | DCU-EN-CS-02-T | 0.3649 | KECIR-CS-CS-02-T | 0.5861 |
| KDEG-EN-CS-01-T | 0.3305 | DLUT-EN-CS-03-T | 0.3608 | KECIR-CS-CS-01-T | 0.5791 |
| DCU-EN-CS-02-T | 0.3289 | DLUT-EN-CS-02-T | 0.3608 | IMU-EN-CS-01-T | 0.5720 |
| QUTIS-EN-CS-04-T | 0.3198 | QUTIS-EN-CS-04-T | 0.3607 | KDEG-EN-CS-01-T | 0.5703 |
| DCU-CS-CS-02-T | 0.3190 | DCU-EN-CS-03-T | 0.3546 | DCU-EN-CS-02-T | 0.5630 |
| KECIR-CS-CS-02-T | 0.3188 | KECIR-CS-CS-02-T | 0.3540 | DCU-EN-CS-03-T | 0.5602 |
| IMU-EN-CS-01-T | 0.3184 | IMU-EN-CS-01-T | 0.3540 | DLUT-EN-CS-03-T | 0.5561 |
| DCU-EN-CS-03-T | 0.3158 | DCU-CS-CS-02-T | 0.3522 | DLUT-EN-CS-02-T | 0.5561 |
| DLUT-EN-CS-01-T | 0.3152 | DLUT-EN-CS-01-T | 0.3462 | KECIR-CS-CS-05-T | 0.5518 |
| KECIR-CS-CS-01-T | 0.3077 | KECIR-CS-CS-01-T | 0.3451 | DCU-CS-CS-02-T | 0.5489 |
| KDEG-EN-CS-03-T | 0.2777 | KDEG-EN-CS-03-T | 0.3171 | DLUT-EN-CS-01-T | 0.5400 |
| KECIR-CS-CS-04-T | 0.2767 | KECIR-CS-CS-04-T | 0.3106 | KDEG-EN-CS-03-T | 0.5284 |
| QUTIS-EN-CS-05-T | 0.2752 | KECIR-CS-CS-05-T | 0.3100 | KECIR-CS-CS-04-T | 0.5246 |
| KECIR-CS-CS-05-T | 0.2723 | QUTIS-EN-CS-05-T | 0.3086 | QUTIS-EN-CS-05-T | 0.5245 |
| KDEG-EN-CS-05-T | 0.2662 | KDEG-EN-CS-05-T | 0.3020 | QUTIS-EN-CS-03-T | 0.5127 |
| KDEG-EN-CS-04-T | 0.2661 | KDEG-EN-CS-04-T | 0.3020 | KDEG-EN-CS-05-T | 0.5087 |
| WUST-CS-CS-01-T | 0.2647 | QUTIS-EN-CS-03-T | 0.2886 | KDEG-EN-CS-04-T | 0.5083 |
| QUTIS-EN-CS-03-T | 0.2504 | WUST-CS-CS-01-T | 0.2871 | WUST-CS-CS-01-T | 0.4819 |
| DCU-EN-CS-01-T | 0.2280 | DCU-EN-CS-01-T | 0.2607 | DCU-EN-CS-01-T | 0.4578 |
| CYUT-EN-CS-02-T | 0.1996 | CYUT-EN-CS-02-T | 0.2263 | CYUT-EN-CS-02-T | 0.4290 |
| CYUT-EN-CS-01-T | 0.1955 | CYUT-EN-CS-01-T | 0.2225 | CYUT-EN-CS-01-T | 0.4152 |
| QUTIS-EN-CS-02-T | 0.1673 | QUTIS-EN-CS-02-T | 0.1967 | QUTIS-EN-CS-02-T | 0.4028 |
| CYUT-EN-CS-04-DN | 0.1562 | CYUT-EN-CS-04-DN | 0.1817 | CYUT-EN-CS-04-DN | 0.3933 |
| CYUT-EN-CS-03-D | 0.1445 | QUTIS-EN-CS-01-T | 0.1689 | CYUT-EN-CS-03-D | 0.3622 |
| WUST-EN-CS-02-T | 0.1434 | CYUT-EN-CS-03-D | 0.1674 | QUTIS-EN-CS-01-T | 0.3527 |
| QUTIS-EN-CS-01-T | 0.1420 | WUST-EN-CS-02-T | 0.1562 | WUST-EN-CS-02-T | 0.2916 |
| WUST-EN-CS-01-T | 0.1036 | WUST-EN-CS-01-T | 0.1204 | WUST-EN-CS-01-T | 0.2810 |

**Table 4. CS run results based on true qrels: mean effectiveness over 73 topics (AFTER bug fix). (Arrows indicate rank changes compared to the results BEFORE bug fix.)**

| run | Mean AP | run | Mean Q | run | Mean nDCG |
|---|---|---|---|---|---|
| KDEG-CS-CS-02-DN | 0.4488 | KDEG-CS-CS-02-DN | 0.4876 | IMU-CS-CS-01-T | 0.6835 |
| KDEG-CS-CS-01-T | 0.4471 | KDEG-CS-CS-01-T | 0.4865 | KDEG-CS-CS-02-DN | 0.6809 |
| IMU-CS-CS-01-T | 0.4333 | IMU-CS-CS-01-T | 0.4711 | KDEG-CS-CS-01-T | 0.6759 |
| WHUCC-EN-CS-01-T | 0.4209 | WHUQA-CS-CS-01-T | 0.4610 | WHUQA-CS-CS-01-T | 0.6706 |
| WHUQA-CS-CS-01-T | 0.4196 | WHUCC-EN-CS-01-T | 0.4583 | IMU-CS-CS-02-T | 0.6655 |
| DCU-CS-CS-01-T↑ 1 | 0.4187 | IMU-CS-CS-02-T | 0.4560 | WHUQA-CS-CS-02-T | 0.6653 |
| IMU-CS-CS-02-T↓ 1 | 0.4178 | DCU-CS-CS-01-T | 0.4557 | IMU-CS-CS-03-T | 0.6649 |
| WHUQA-CS-CS-02-T | 0.4120 | WHUQA-CS-CS-02-T | 0.4539 | WHUCC-EN-CS-01-T | 0.6588 |
| IMU-CS-CS-03-T | 0.4098 | IMU-CS-CS-03-T | 0.4477 | DCU-CS-CS-01-T | 0.6545 |
| WHUCC-CS-CS-01-T | 0.4028 | WHUCC-CS-CS-01-T | 0.4406 | WHUCC-CS-CS-01-T | 0.6519 |
| WHUCC-EN-CS-03-T | 0.3970 | KDEG-CS-CS-03-T | 0.4336 | KDEG-CS-CS-03-T | 0.6424 |
| KDEG-CS-CS-03-T | 0.3941 | WHUCC-EN-CS-03-T | 0.4307 | WHUCC-EN-CS-02-T | 0.6422 |
| WHUCC-EN-CS-02-T | 0.3847 | WHUCC-EN-CS-02-T | 0.4243 | WHUCC-EN-CS-03-T | 0.6327 |
| KDEG-EN-CS-02-DN | 0.3847 | KDEG-EN-CS-02-DN | 0.4220 | WHUQA-EN-CS-01-T | 0.6200 |
| WHUQA-EN-CS-01-T | 0.3760 | WHUQA-EN-CS-01-T | 0.4143 | KDEG-EN-CS-02-DN | 0.6147 |
| WHUQA-EN-CS-02-T | 0.3605 | WHUQA-EN-CS-02-T | 0.3987 | WHUCC-CS-CS-02-T | 0.6136 |
| KDEG-CS-CS-05-T↑ 1 | 0.3603 | KDEG-CS-CS-05-T | 0.3980 | KDEG-CS-CS-05-T↑ 1 | 0.6063 |
| KDEG-CS-CS-04-T↓ 1 | 0.3603 | KDEG-CS-CS-04-T | 0.3979 | KDEG-CS-CS-04-T↑ 1 | 0.6060 |
| WHUCC-CS-CS-02-T | 0.3569 | WHUCC-CS-CS-02-T | 0.3930 | WHUQA-EN-CS-02-T↓ 2 | 0.6050 |
| KECIR-CS-CS-03-T | 0.3411 | KDEG-EN-CS-01-T | 0.3766 | KECIR-CS-CS-03-T | 0.5981 |
| KDEG-EN-CS-01-T↑ 2 | 0.3378 | KECIR-CS-CS-03-T | 0.3749 | QUTIS-EN-CS-04-T | 0.5952 |
| DLUT-EN-CS-03-T↓ 1 | 0.3355 | DCU-EN-CS-02-T | 0.3719 | KECIR-CS-CS-02-T | 0.5941 |
| DLUT-EN-CS-02-T↓ 1 | 0.3355 | QUTIS-EN-CS-04-T↑ 2 | 0.3677 | KECIR-CS-CS-01-T | 0.5870 |
| DCU-EN-CS-02-T | 0.3347 | DLUT-EN-CS-03-T↓ 1 | 0.3667 | IMU-EN-CS-01-T | 0.5780 |
| KECIR-CS-CS-02-T↑ 2 | 0.3265 | DLUT-EN-CS-02-T↓ 1 | 0.3667 | KDEG-EN-CS-01-T | 0.5778 |
| DCU-CS-CS-02-T | 0.3260 | KECIR-CS-CS-02-T↑ 1 | 0.3635 | DCU-EN-CS-02-T | 0.5695 |
| QUTIS-EN-CS-04-T↓ 2 | 0.3255 | DCU-EN-CS-03-T↓ 1 | 0.3619 | DCU-EN-CS-03-T | 0.5671 |
| IMU-EN-CS-01-T | 0.3231 | DCU-CS-CS-02-T↑ 1 | 0.3614 | DLUT-EN-CS-03-T | 0.5626 |
| DCU-EN-CS-03-T | 0.3215 | IMU-EN-CS-01-T↓ 1 | 0.3596 | DLUT-EN-CS-02-T | 0.5626 |
| DLUT-EN-CS-01-T | 0.3195 | KECIR-CS-CS-01-T↑ 1 | 0.3546 | KECIR-CS-CS-05-T | 0.5586 |
| KECIR-CS-CS-01-T | 0.3154 | DLUT-EN-CS-01-T↓ 1 | 0.3519 | DCU-CS-CS-02-T | 0.5566 |
| KECIR-CS-CS-04-T↑ 1 | 0.2833 | KDEG-EN-CS-03-T | 0.3231 | DLUT-EN-CS-01-T | 0.5466 |
| KDEG-EN-CS-03-T↓ 1 | 0.2829 | KECIR-CS-CS-04-T | 0.3193 | KDEG-EN-CS-03-T | 0.5342 |
| QUTIS-EN-CS-05-T | 0.2800 | KECIR-CS-CS-05-T | 0.3177 | KECIR-CS-CS-04-T | 0.5322 |
| KECIR-CS-CS-05-T | 0.2782 | QUTIS-EN-CS-05-T | 0.3143 | QUTIS-EN-CS-05-T | 0.5301 |
| KDEG-EN-CS-05-T | 0.2709 | KDEG-EN-CS-05-T | 0.3073 | QUTIS-EN-CS-03-T | 0.5190 |
| KDEG-EN-CS-04-T | 0.2707 | KDEG-EN-CS-04-T | 0.3072 | KDEG-EN-CS-05-T | 0.5142 |
| WUST-CS-CS-01-T | 0.2694 | QUTIS-EN-CS-03-T | 0.2942 | KDEG-EN-CS-04-T | 0.5138 |
| QUTIS-EN-CS-03-T | 0.2546 | WUST-CS-CS-01-T | 0.2930 | WUST-CS-CS-01-T | 0.4881 |
| DCU-EN-CS-01-T | 0.2284 | DCU-EN-CS-01-T | 0.2617 | DCU-EN-CS-01-T | 0.4597 |
| CYUT-EN-CS-02-T | 0.2036 | CYUT-EN-CS-02-T | 0.2313 | CYUT-EN-CS-02-T | 0.4347 |
| CYUT-EN-CS-01-T | 0.1995 | CYUT-EN-CS-01-T | 0.2274 | CYUT-EN-CS-01-T | 0.4210 |
| QUTIS-EN-CS-02-T | 0.1722 | QUTIS-EN-CS-02-T | 0.2029 | QUTIS-EN-CS-02-T | 0.4091 |
| CYUT-EN-CS-04-DN | 0.1579 | CYUT-EN-CS-04-DN | 0.1841 | CYUT-EN-CS-04-DN | 0.3980 |
| CYUT-EN-CS-03-D | 0.1460 | QUTIS-EN-CS-01-T | 0.1741 | CYUT-EN-CS-03-D | 0.3665 |
| QUTIS-EN-CS-01-T↑ 1 | 0.1459 | CYUT-EN-CS-03-D | 0.1695 | QUTIS-EN-CS-01-T | 0.3581 |
| WUST-EN-CS-02-T↓ 1 | 0.1435 | WUST-EN-CS-02-T | 0.1564 | WUST-EN-CS-02-T | 0.2920 |
| WUST-EN-CS-01-T | 0.1037 | WUST-EN-CS-01-T | 0.1206 | WUST-EN-CS-01-T | 0.2815 |

**Table 5. JA run results based on true qrels: mean effectiveness over 94 topics (BEFORE bug fix).**

| run | Mean AP | run | Mean Q | run | Mean nDCG |
|---|---|---|---|---|---|
| LTI-JA-JA-01-T | 0.3893 | LTI-JA-JA-01-T | 0.4001 | LTI-JA-JA-01-T | 0.5977 |
| LTI-JA-JA-02-T | 0.3888 | LTI-JA-JA-02-T | 0.3997 | LTI-JA-JA-02-T | 0.5971 |
| LTI-JA-JA-03-T | 0.3837 | BRKLY-JA-JA-01-DN | 0.3965 | BRKLY-JA-JA-01-DN | 0.5954 |
| BRKLY-JA-JA-01-DN | 0.3829 | LTI-JA-JA-03-T | 0.3938 | LTI-JA-JA-03-T | 0.5895 |
| BRKLY-JA-JA-02-T | 0.3695 | BRKLY-JA-JA-02-T | 0.3832 | BRKLY-JA-JA-02-T | 0.5694 |
| BRKLY-EN-JA-01-DN | 0.3270 | BRKLY-EN-JA-01-DN | 0.3389 | BRKLY-EN-JA-01-DN | 0.5335 |
| BRKLY-EN-JA-02-T | 0.3040 | LTI-EN-JA-01-T | 0.3194 | BRKLY-JA-JA-04-DN | 0.5266 |
| LTI-EN-JA-01-T | 0.3013 | LTI-EN-JA-02-T | 0.3157 | LTI-EN-JA-01-T | 0.5082 |
| LTI-EN-JA-02-T | 0.2981 | BRKLY-EN-JA-02-T | 0.3156 | LTI-EN-JA-02-T | 0.5004 |
| BRKLY-JA-JA-04-DN | 0.2888 | BRKLY-JA-JA-04-DN | 0.2994 | BRKLY-JA-JA-05-T | 0.4962 |
| LTI-EN-JA-03-T | 0.2819 | LTI-EN-JA-03-T | 0.2991 | BRKLY-EN-JA-02-T | 0.4871 |
| BRKLY-JA-JA-05-T | 0.2732 | BRKLY-JA-JA-05-T | 0.2855 | LTI-EN-JA-03-T | 0.4781 |
| CYUT-EN-JA-02-T | 0.1719 | CYUT-EN-JA-02-T | 0.1788 | CYUT-EN-JA-02-T | 0.3638 |
| CYUT-EN-JA-01-T | 0.1708 | CYUT-EN-JA-01-T | 0.1776 | CYUT-EN-JA-01-T | 0.3613 |
| BRKLY-JA-JA-03-DN | 0.1413 | BRKLY-JA-JA-03-DN | 0.1438 | BRKLY-JA-JA-03-DN | 0.2909 |
| CYUT-EN-JA-03-D | 0.1023 | CYUT-EN-JA-03-D | 0.1027 | CYUT-EN-JA-03-D | 0.2565 |
| CYUT-EN-JA-04-DN | 0.0999 | CYUT-EN-JA-04-DN | 0.0985 | CYUT-EN-JA-04-DN | 0.2449 |

**Table 6. JA run results based on true qrels: mean effectiveness over 94 topics (AFTER bug fix). (Arrows indicate rank changes compared to the results BEFORE bug fix.)**

| run | Mean AP | run | Mean Q | run | Mean nDCG |
|---|---|---|---|---|---|
| LTI-JA-JA-01-T | 0.4356 | LTI-JA-JA-01-T | 0.4499 | LTI-JA-JA-01-T | 0.6571 |
| LTI-JA-JA-02-T | 0.4351 | LTI-JA-JA-02-T | 0.4494 | LTI-JA-JA-02-T | 0.6565 |
| LTI-JA-JA-03-T | 0.4293 | BRKLY-JA-JA-01-DN | 0.4468 | BRKLY-JA-JA-01-DN | 0.6544 |
| BRKLY-JA-JA-01-DN | 0.4277 | LTI-JA-JA-03-T | 0.4428 | LTI-JA-JA-03-T | 0.6483 |
| BRKLY-JA-JA-02-T | 0.4143 | BRKLY-JA-JA-02-T | 0.4334 | BRKLY-JA-JA-02-T | 0.6290 |
| BRKLY-EN-JA-01-DN | 0.3619 | BRKLY-EN-JA-01-DN | 0.3793 | BRKLY-EN-JA-01-DN | 0.5866 |
| BRKLY-EN-JA-02-T | 0.3458 | BRKLY-EN-JA-02-T↑ 2 | 0.3622 | BRKLY-JA-JA-04-DN | 0.5779 |
| LTI-EN-JA-01-T | 0.3327 | LTI-EN-JA-01-T↓ 1 | 0.3545 | LTI-EN-JA-01-T | 0.5571 |
| LTI-EN-JA-02-T | 0.3293 | LTI-EN-JA-02-T↓ 1 | 0.3506 | LTI-EN-JA-02-T | 0.5492 |
| BRKLY-JA-JA-04-DN | 0.3207 | BRKLY-JA-JA-04-DN | 0.3366 | BRKLY-EN-JA-02-T↑ 1 | 0.5442 |
| LTI-EN-JA-03-T | 0.3074 | LTI-EN-JA-03-T | 0.3282 | BRKLY-JA-JA-05-T↓ 1 | 0.5416 |
| BRKLY-JA-JA-05-T | 0.2990 | BRKLY-JA-JA-05-T | 0.3155 | LTI-EN-JA-03-T | 0.5219 |
| CYUT-EN-JA-02-T | 0.1905 | CYUT-EN-JA-02-T | 0.1993 | CYUT-EN-JA-02-T | 0.3982 |
| CYUT-EN-JA-01-T | 0.1894 | CYUT-EN-JA-01-T | 0.1982 | CYUT-EN-JA-01-T | 0.3957 |
| BRKLY-JA-JA-03-DN | 0.1587 | BRKLY-JA-JA-03-DN | 0.1624 | BRKLY-JA-JA-03-DN | 0.3132 |
| CYUT-EN-JA-03-D | 0.1119 | CYUT-EN-JA-03-D | 0.1121 | CYUT-EN-JA-03-D | 0.2775 |
| CYUT-EN-JA-04-DN | 0.1097 | CYUT-EN-JA-04-DN | 0.1079 | CYUT-EN-JA-04-DN | 0.2657 |

**Table 7. CT run results based on true qrels: mean effectiveness over (a) 86 topics (pool depth 50) and (b) 87 topics (pool depth 100).**

| | run | Mean AP | run | Mean Q | run | Mean nDCG |
|---|---|---|---|---|---|---|
| (a) pool depth 50 | KDEG-CT-CT-02-DN | 0.4900 | KDEG-CT-CT-02-DN | 0.5263 | KDEG-CT-CT-02-DN | 0.7175 |
| | KDEG-CT-CT-05-T | 0.4844 | KDEG-CT-CT-05-T | 0.5242 | KDEG-CT-CT-01-T | 0.7140 |
| | KDEG-CT-CT-01-T | 0.4818 | KDEG-CT-CT-01-T | 0.5227 | KDEG-CT-CT-05-T | 0.7129 |
| | KDEG-CT-CT-03-T | 0.4155 | KDEG-CT-CT-03-T | 0.4539 | KDEG-CT-CT-03-T | 0.6728 |
| | KDEG-EN-CT-02-DN | 0.3723 | KDEG-CT-CT-04-T | 0.4100 | KDEG-CT-CT-04-T | 0.6369 |
| | KDEG-CT-CT-04-T | 0.3713 | KDEG-EN-CT-02-DN | 0.4006 | KDEG-EN-CT-02-DN | 0.5689 |
| | KDEG-EN-CT-01-T | 0.3551 | KDEG-EN-CT-01-T | 0.3822 | KDEG-EN-CT-01-T | 0.5567 |
| | KDEG-EN-CT-05-T | 0.3500 | KDEG-EN-CT-05-T | 0.3765 | QUTIS-EN-CT-04-T | 0.5555 |
| | QUTIS-EN-CT-04-T | 0.3231 | QUTIS-EN-CT-04-T | 0.3569 | KDEG-EN-CT-05-T | 0.5414 |
| | KDEG-EN-CT-03-T | 0.2817 | KDEG-EN-CT-03-T | 0.3087 | KDEG-EN-CT-03-T | 0.4957 |
| | QUTIS-EN-CT-03-T | 0.2656 | QUTIS-EN-CT-03-T | 0.2957 | QUTIS-EN-CT-03-T | 0.4905 |
| | KDEG-EN-CT-04-T | 0.2460 | KDEG-EN-CT-04-T | 0.2728 | KDEG-EN-CT-04-T | 0.4578 |
| | QUTIS-EN-CT-02-T | 0.2161 | QUTIS-EN-CT-02-T | 0.2501 | QUTIS-EN-CT-02-T | 0.4374 |
| | QUTIS-EN-CT-01-T | 0.1943 | QUTIS-EN-CT-01-T | 0.2218 | QUTIS-EN-CT-01-T | 0.3997 |
| | CYUT-EN-CT-02-T | 0.1941 | CYUT-EN-CT-02-T | 0.2137 | CYUT-EN-CT-02-T | 0.3963 |
| | CYUT-EN-CT-01-T | 0.1733 | CYUT-EN-CT-01-T | 0.1923 | CYUT-EN-CT-01-T | 0.3672 |
| | CYUT-EN-CT-04-DN | 0.1486 | CYUT-EN-CT-04-DN | 0.1677 | CYUT-EN-CT-04-DN | 0.3516 |
| | CYUT-EN-CT-03-D | 0.1362 | CYUT-EN-CT-03-D | 0.1509 | CYUT-EN-CT-03-D | 0.3210 |
| | QUTIS-EN-CT-05-T | 0.1040 | QUTIS-EN-CT-05-T | 0.1167 | QUTIS-EN-CT-05-T | 0.2492 |
| | run | Mean AP | run | Mean Q | run | Mean nDCG |
| (b) pool depth 100 | KDEG-CT-CT-02-DN | 0.4635 | KDEG-CT-CT-02-DN | 0.5013 | KDEG-CT-CT-02-DN | 0.7077 |
| | KDEG-CT-CT-01-T | 0.4572 | KDEG-CT-CT-01-T | 0.4977 | KDEG-CT-CT-01-T | 0.7056 |
| | KDEG-CT-CT-05-T | 0.4561 | KDEG-CT-CT-05-T | 0.4952 | KDEG-CT-CT-05-T | 0.7021 |
| | KDEG-CT-CT-03-T | 0.3909 | KDEG-CT-CT-03-T | 0.4274 | KDEG-CT-CT-03-T | 0.6604 |
| | KDEG-EN-CT-02-DN | 0.3515 | KDEG-CT-CT-04-T | 0.3840 | KDEG-CT-CT-04-T | 0.6227 |
| | KDEG-CT-CT-04-T | 0.3478 | KDEG-EN-CT-02-DN | 0.3791 | KDEG-EN-CT-02-DN | 0.5579 |
| | KDEG-EN-CT-01-T | 0.3374 | KDEG-EN-CT-01-T | 0.3638 | KDEG-EN-CT-01-T | 0.5463 |
| | KDEG-EN-CT-05-T | 0.3304 | KDEG-EN-CT-05-T | 0.3568 | QUTIS-EN-CT-04-T | 0.5412 |
| | QUTIS-EN-CT-04-T | 0.3027 | QUTIS-EN-CT-04-T | 0.3354 | KDEG-EN-CT-05-T | 0.5326 |
| | KDEG-EN-CT-03-T | 0.2623 | KDEG-EN-CT-03-T | 0.2879 | KDEG-EN-CT-03-T | 0.4826 |
| | QUTIS-EN-CT-03-T | 0.2490 | QUTIS-EN-CT-03-T | 0.2761 | QUTIS-EN-CT-03-T | 0.4771 |
| | KDEG-EN-CT-04-T | 0.2292 | KDEG-EN-CT-04-T | 0.2539 | KDEG-EN-CT-04-T | 0.4434 |
| | QUTIS-EN-CT-02-T | 0.2041 | QUTIS-EN-CT-02-T | 0.2369 | QUTIS-EN-CT-02-T | 0.4298 |
| | QUTIS-EN-CT-01-T | 0.1834 | QUTIS-EN-CT-01-T | 0.2078 | QUTIS-EN-CT-01-T | 0.3910 |
| | CYUT-EN-CT-02-T | 0.1753 | CYUT-EN-CT-02-T | 0.1945 | CYUT-EN-CT-02-T | 0.3847 |
| | CYUT-EN-CT-01-T | 0.1555 | CYUT-EN-CT-01-T | 0.1740 | CYUT-EN-CT-01-T | 0.3547 |
| | CYUT-EN-CT-04-DN | 0.1322 | CYUT-EN-CT-04-DN | 0.1495 | CYUT-EN-CT-04-DN | 0.3348 |
| | CYUT-EN-CT-03-D | 0.1204 | CYUT-EN-CT-03-D | 0.1346 | CYUT-EN-CT-03-D | 0.3071 |
| | QUTIS-EN-CT-05-T | 0.0947 | QUTIS-EN-CT-05-T | 0.1073 | QUTIS-EN-CT-05-T | 0.2404 |

**Table 8. $\tau$ and $\tau_{ap}$ rank correlation: system rankings by different metrics. For convenience, values higher than .9 are shown in bold.**

| | | AP | Q | nDCG |
|---|---|---|---|---|
| CS (BEFORE bug fix) | AP | **1/1** | **.970/.961** | **.906**/.838 |
| | Q | **.970/.960** | **1/1** | **.922**/.864 |
| | nDCG | **.906**/.824 | **.922**/.844 | **1/1** |
| CS (AFTER bug fix) | AP | **1/1** | **.970/.957** | **.904**/.832 |
| | Q | **.970/.957** | **1/1** | **.924**/.863 |
| | nDCG | **.904**/.817 | **.924**/.844 | **1/1** |
| JA (BEFORE bug fix) | AP | **1/1** | **.956/.925** | .882/.860 |
| | Q | **.956/.927** | **1/1** | **.926/.936** |
| | nDCG | .882/.863 | **.926/.930** | **1/1** |
| JA (AFTER bug fix) | AP | **1/1** | **.985/.958** | .897/.872 |
| | Q | **.985/.958** | **1/1** | **.912/.913** |
| | nDCG | .897/.872 | **.912/.913** | **1/1** |
| CT (pool depth 50) | AP | **1/1** | **.988/.978** | **.965/.908** |
| | Q | **.988/.978** | **1/1** | **.977/.931** |
| | nDCG | **.965/.908** | **.977/.931** | **1/1** |
| CT (pool depth 100) | AP | **1/1** | **.988/.978** | **.977/.964** |
| | Q | **.988/.978** | **1/1** | **.988/.986** |
| | nDCG | **.977/.964** | **.988/.986** | **1/1** |

**Table 9. The best T-run from each team (CS, JA, CT with pool-depth-50 and pool-depth-100 assessments): "$*$" and "$**$" indicate that a run significantly outperforms (at $\alpha = 0.05$ and $\alpha = 0.01$, respectively) than one shown immediately below according to a two-sided paired bootstrap test. Note, however, that pairwise statistical significance is not transitive. (Arrows indicate rank changes compared to the results BEFORE bug fix.)**

| | run | Mean AP | run | Mean Q | run | Mean nDCG |
|---|---|---|---|---|---|---|
| CS (BEFORE bug fix) | KDEG-CS-CS-01-T | 0.4390 | KDEG-CS-CS-01-T | 0.4764 | IMU-CS-CS-01-T | 0.6761 |
| | IMU-CS-CS-01-T | 0.4266 | IMU-CS-CS-01-T | 0.4628 | KDEG-CS-CS-01-T | 0.6674 |
| | WHUCC-EN-CS-01-T | 0.4139 | WHUQA-CS-CS-01-T | 0.4528 | WHUQA-CS-CS-01-T | 0.6629 |
| | WHUQA-CS-CS-01-T | 0.4128 | WHUCC-EN-CS-01-T | 0.4499 | WHUCC-EN-CS-01-T | 0.6509 |
| | DCU-CS-CS-01-T | 0.4111$**$ | DCU-CS-CS-01-T | 0.4464$**$ | DCU-CS-CS-01-T | 0.6466$**$ |
| | KECIR-CS-CS-03-T | 0.3333 | KECIR-CS-CS-03-T | 0.3658 | KECIR-CS-CS-03-T | 0.5903 |
| | DLUT-EN-CS-03-T | 0.3312 | DLUT-EN-CS-03-T | 0.3608 | QUTIS-EN-CS-04-T | 0.5882 |
| | QUTIS-EN-CS-04-T | 0.3198 | QUTIS-EN-CS-04-T | 0.3607$*$ | DLUT-EN-CS-03-T | 0.5561$**$ |
| | WUST-CS-CS-01-T | 0.2647 | WUST-CS-CS-01-T | 0.2871 | WUST-CS-CS-01-T | 0.4819 |
| | CYUT-EN-CS-02-T | 0.1996 | CYUT-EN-CS-02-T | 0.2263 | CYUT-EN-CS-02-T | 0.4290 |
| CS (AFTER bug fix) | KDEG-CS-CS-01-T | 0.4471 | KDEG-CS-CS-01-T | 0.4865 | IMU-CS-CS-01-T | 0.6835 |
| | IMU-CS-CS-01-T | 0.4333 | IMU-CS-CS-01-T | 0.4711 | KDEG-CS-CS-01-T | 0.6759 |
| | WHUCC-EN-CS-01-T | 0.4209 | WHUQA-CS-CS-01-T | 0.4610 | WHUQA-CS-CS-01-T | 0.6706 |
| | WHUQA-CS-CS-01-T | 0.4196 | WHUCC-EN-CS-01-T | 0.4583 | WHUCC-EN-CS-01-T | 0.6588 |
| | DCU-CS-CS-01-T | 0.4187$**$ | DCU-CS-CS-01-T | 0.4557$**$ | DCU-CS-CS-01-T | 0.6545$**$ |
| | KECIR-CS-CS-03-T | 0.3411 | KECIR-CS-CS-03-T | 0.3749 | KECIR-CS-CS-03-T | 0.5981 |
| | DLUT-EN-CS-03-T | 0.3355 | QUTIS-EN-CS-04-T$\uparrow$ 1 | 0.3677 | QUTIS-EN-CS-04-T | 0.5952 |
| | QUTIS-EN-CS-04-T | 0.3255 | DLUT-EN-CS-03-T$\downarrow$ 1 | 0.3667$**$ | DLUT-EN-CS-03-T | 0.5626$**$ |
| | WUST-CS-CS-01-T | 0.2694 | WUST-CS-CS-01-T | 0.2930 | WUST-CS-CS-01-T | 0.4881 |
| | CYUT-EN-CS-02-T | 0.2036 | CYUT-EN-CS-02-T | 0.2313 | CYUT-EN-CS-02-T | 0.4347 |
| JA (BEFORE bug fix) | LTI-JA-JA-01-T | 0.3893 | LTI-JA-JA-01-T | 0.4001 | LTI-JA-JA-01-T | 0.5977 |
| | BRKLY-JA-JA-02-T | 0.3695$**$ | BRKLY-JA-JA-02-T | 0.3832$**$ | BRKLY-JA-JA-02-T | 0.5694$**$ |
| | CYUT-EN-JA-02-T | 0.1719 | CYUT-EN-JA-02-T | 0.1788 | CYUT-EN-JA-02-T | 0.3638 |
| JA (AFTER bug fix) | LTI-JA-JA-01-T | 0.4356 | LTI-JA-JA-01-T | 0.4499 | LTI-JA-JA-01-T | 0.6571 |
| | BRKLY-JA-JA-02-T | 0.4143$**$ | BRKLY-JA-JA-02-T | 0.4334$**$ | BRKLY-JA-JA-02-T | 0.6290$**$ |
| | CYUT-EN-JA-02-T | 0.1905 | CYUT-EN-JA-02-T | 0.1993 | CYUT-EN-JA-02-T | 0.3982 |
| CT (pool depth 50) | KDEG-CT-CT-05-T | 0.4844$**$ | KDEG-CT-CT-05-T | 0.5242$**$ | KDEG-CT-CT-01-T | 0.7140$**$ |
| | QUTIS-EN-CT-04-T | 0.3231$**$ | QUTIS-EN-CT-04-T | 0.3569$**$ | QUTIS-EN-CT-04-T | 0.5555$**$ |
| | CYUT-EN-CT-02-T | 0.1941 | CYUT-EN-CT-02-T | 0.2137 | CYUT-EN-CT-02-T | 0.3963 |
| CT (pool depth 100) | KDEG-CT-CT-01-T | 0.4572$**$ | KDEG-CT-CT-01-T | 0.4977$**$ | KDEG-CT-CT-01-T | 0.7056$**$ |
| | QUTIS-EN-CT-04-T | 0.3027$**$ | QUTIS-EN-CT-04-T | 0.3354$**$ | QUTIS-EN-CT-04-T | 0.5412$**$ |
| | CYUT-EN-CT-02-T | 0.1753 | CYUT-EN-CT-02-T | 0.1945 | CYUT-EN-CT-02-T | 0.3847 |

**Table 10. The best T-run from each team in terms of Mean nDCG and their system descriptions.**

| CS runs |
|---|
| IMU-CS-CS-01-T<br>(1) Word segmentation by ICTCLAS(free version)<br>(2) Combined basic keyterm based query with PRF queries<br>(3) Indri structure query with keyterm expansion based on baidu and synonyms dictionary<br>(4) Rerank documents by interpolating the results from (2) and (3) |
| KDEG-CS-CS-01-T<br>Ranking based on BB2 model |
| WHUQA-CS-CS-01-T<br>The system is based on bigram and BM25 model |
| WHUCC-EN-CS-01-T<br>index based on bi-grams and single Chinese character; BM25 retrieval model;<br>query expansion based on RSV |
| DCU-CS-CS-01-T<br>Ranked based on language model feedback retrieval using collection mixture method and Dirichlet smoothing |
| KECIR-CS-CS-03-T<br>We use okapi as ir model; and the MMD as reranking method. |
| QUTIS-EN-CS-04-T<br>Translation using four-stage Google search and Wikipedia English and Chinese page mapping;<br>and combining the translation from translate.google.com.<br>Choose the wikipedia page contain the word in the question; otherwise just ignore it.<br>Index documents using single character and NGMI-segmented words<br>Ranking using BM25 on ANT search engine.<br>Apply frequency analysis on Chinese terms collection and Google translate;<br>then form NGMI-segmented and unigram segmented query |
| DLUT-EN-CS-03-T<br>enrich resources used by query expansion |
| WUST-CS-CS-01-T VSM; query expansion |
| CYUT-EN-CS-02-T<br>Lucene indexing; PRF; OKAPI BM25; Query Expansion based on OKAPI and Wikipedia;<br>Translation from Google Translate and Wikipedia. |

| JA runs |
|---|
| LTI-JA-JA-01-T<br>Indri (language model + inference network) with character-based index.<br>Used phrase query operator for important terms.<br>Relaxed the query with combine operator over all terms in the question. |
| BRKLY-JA-JA-02-T<br>Probabilistic retrieval based on logistic regression with blind feedback on QUESTION text only. |
| CYUT-EN-JA-02-T<br>Lucene indexing; PRF; OKAPI BM25; Query Expansion based on OKAPI and Wikipedia;<br>Translation from Google Translate and Wikipedia. |

| CT runs |
|---|
| KDEG-CT-CT-01-T<br>Ranking based on BB2 model |
| QUTIS-EN-CT-04-T<br>Translation using four-stage Google search and Wikipedia English and Chinese page mapping;<br>and combining the translation from translate.google.com.<br>Choose the wikipedia page contain the word in the question; otherwise just ignore it.<br>Index documents using single character and NGMI-segmented words<br>Ranking using BM25 on ANT search engine.<br>Apply frequency analysis on Chinese terms collection and Google translate;<br>then form NGMI-segmented and unigram segmented query |
| CYUT-EN-CT-02-T<br>Lucene indexing; PRF; OKAPI BM25; Query Expansion based on OKAPI and Wikipedia;<br>Translation from Google Translate and Wikipedia. |

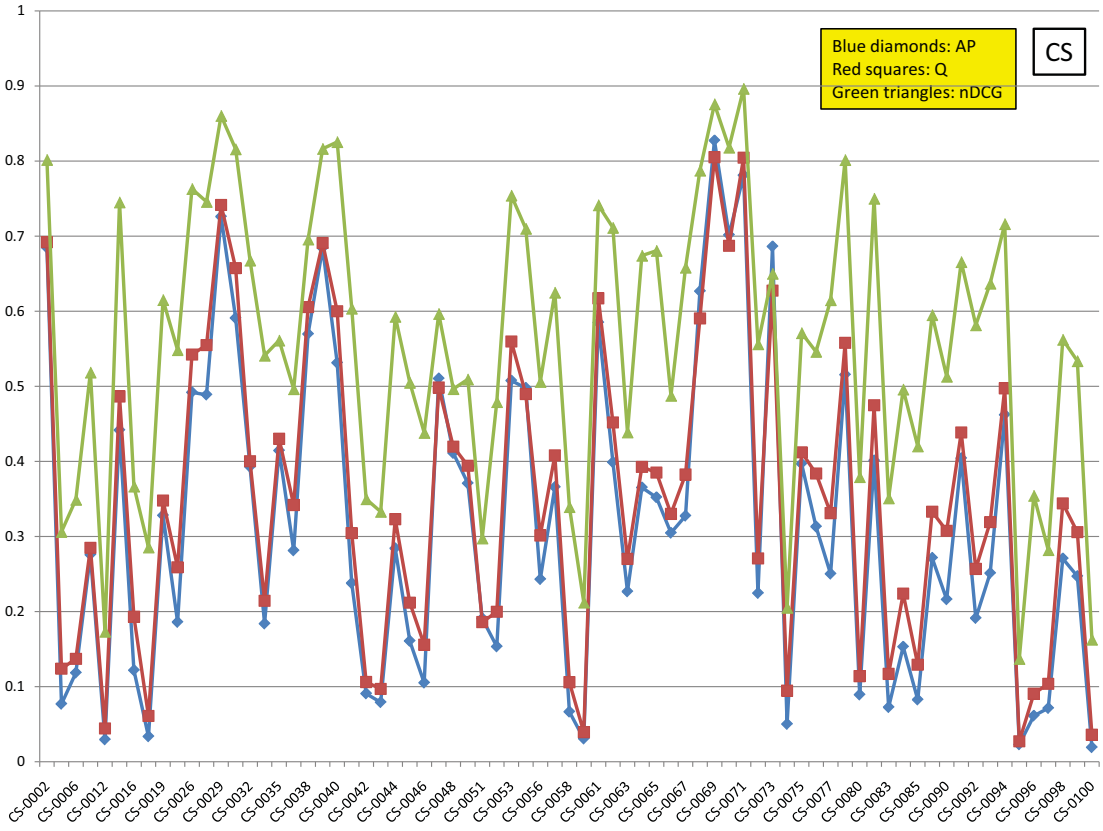


**Figure 1. Per-topic AP, Q and nDCG averaged over 48 runs for the 73 CS topics (AFTER bug fix).**

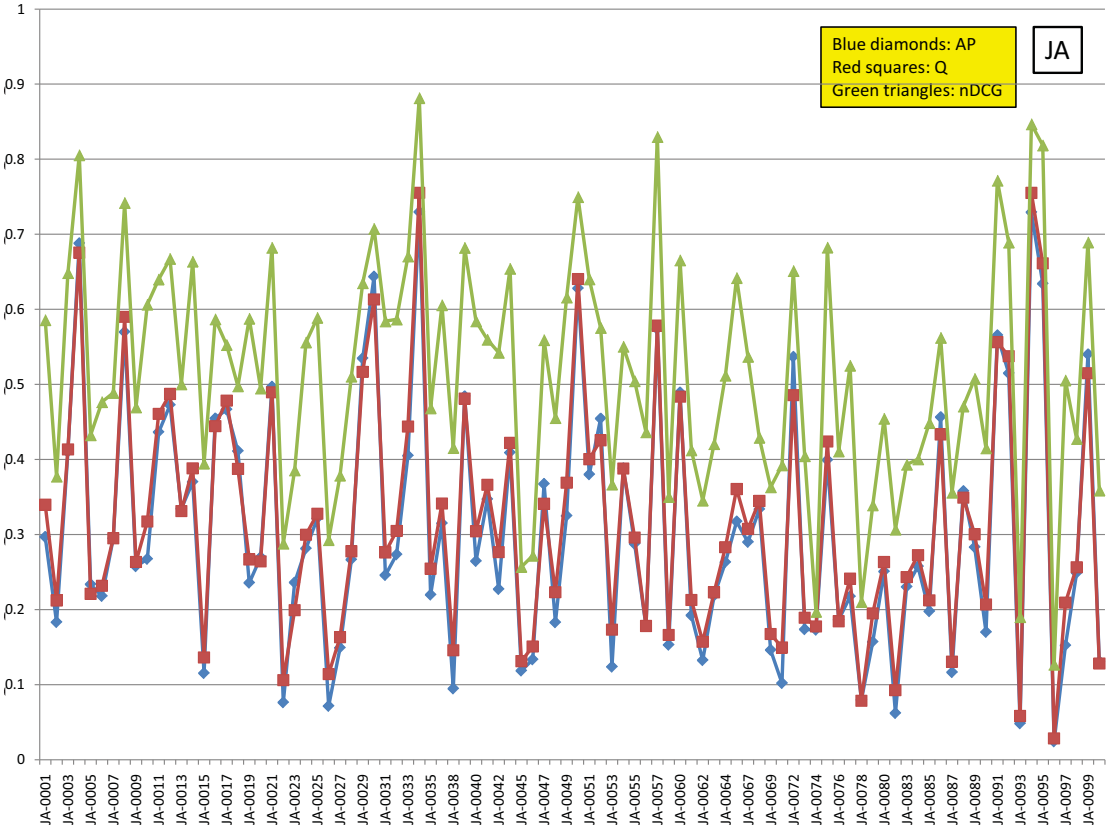


**Figure 2. Per-topic AP, Q and nDCG averaged over 17 runs for the 94 JA topics (AFTER bug fix).**

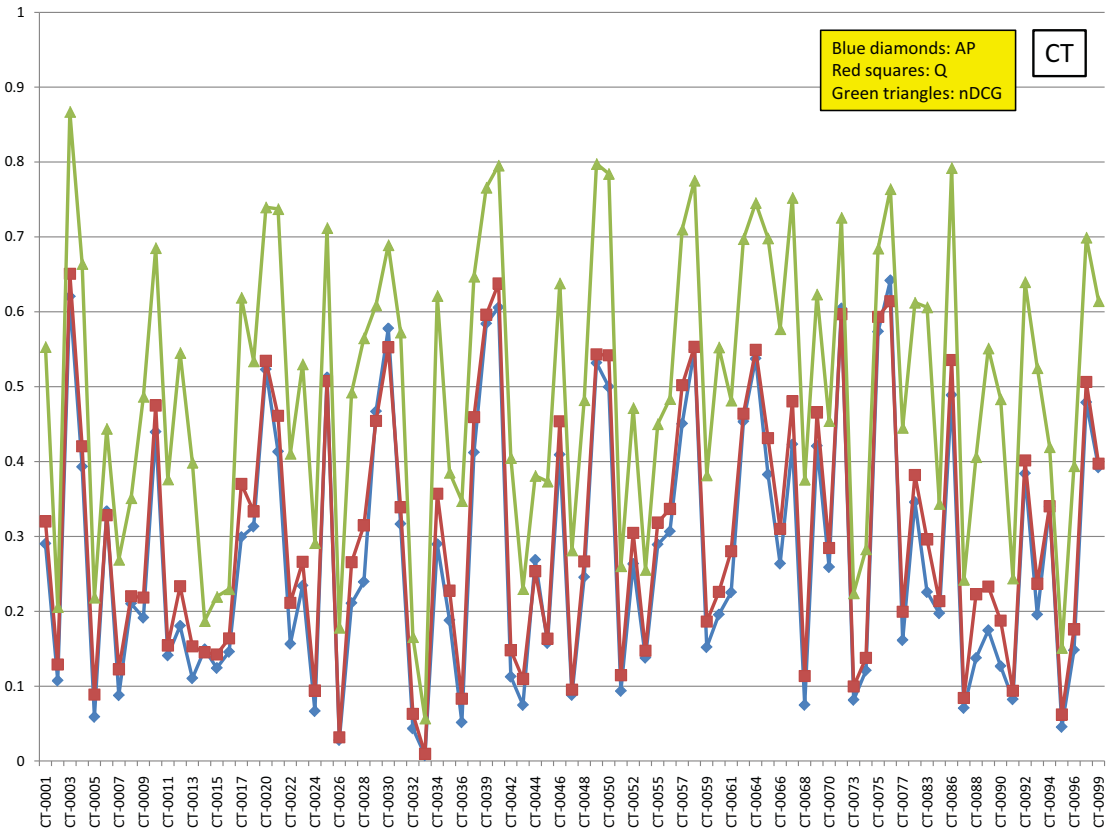


**Figure 3. Per-topic AP, Q and nDCG averaged over 19 runs for the 87 CT topics (pool depth 100).**

**Table 11. $\tau$ and $\tau_{ap}$ rank correlation: topic rankings by different metrics. For convenience, values higher than .9 are shown in bold.**

| | | AP | Q | nDCG |
|---|---|---|---|---|
| CS (BEFORE bug fix) | AP | **1/1** | **.927**/.883 | .757/.679 |
| | Q | **.927**/.888 | **1/1** | .788/.705 |
| | nDCG | .757/.703 | .788/.723 | **1/1** |
| CS (AFTER bug fix) | AP | **1/1** | **.938/.907** | .775/.685 |
| | Q | **.938/.909** | **1/1** | .795/.703 |
| | nDCG | .775/.709 | .795/.721 | **1/1** |
| JA (BEFORE bug fix) | AP | **1/1** | **.914**/.893 | .763/.714 |
| | Q | **.914**/.896 | **1/1** | .822/.772 |
| | nDCG | .763/.723 | .822/.775 | **1/1** |
| JA (AFTER bug fix) | AP | **1/1** | **.922/.902** | .744/.697 |
| | Q | **.922/.903** | **1/1** | .796/.749 |
| | nDCG | .744/.702 | .796/.753 | **1/1** |
| CT (pool depth 50) | AP | **1/1** | **.918**/.842 | .750/.622 |
| | Q | **.918**/.853 | **1/1** | .802/.692 |
| | nDCG | .750/.647 | .802/.715 | **1/1** |
| CT (pool depth 100) | AP | **1/1** | **.932**/.886 | .756/.645 |
| | Q | **.932**/.899 | **1/1** | .798/.717 |
| | nDCG | .756/.667 | .798/.716 | **1/1** |

**Table 12. Coverage of relevant documents summed across 73 CS topics (BEFORE bug fix).**

| run name | coverage. |
|---|---|
| KDEG-CS-CS-01-T | 4785 |
| KDEG-CS-CS-02-DN | 4772 |
| WHUQA-CS-CS-01-T | 4763 |
| WHUQA-CS-CS-02-T | 4750 |
| KDEG-CS-CS-03-T | 4715 |
| KDEG-EN-CS-02-DN | 4690 |
| DCU-CS-CS-01-T | 4668 |
| IMU-CS-CS-02-T | 4644 |
| KECIR-CS-CS-03-T | 4614 |
| WHUQA-EN-CS-01-T | 4613 |
| WHUQA-EN-CS-02-T | 4602 |
| KDEG-CS-CS-05-T | 4507 |
| KDEG-CS-CS-04-T | 4507 |
| IMU-CS-CS-03-T | 4468 |
| IMU-CS-CS-01-T | 4468 |
| KECIR-CS-CS-01-T | 4449 |
| WHUCC-EN-CS-01-T | 4437 |
| KECIR-CS-CS-02-T | 4428 |
| WHUCC-EN-CS-03-T | 4285 |
| QUTIS-EN-CS-04-T | 4268 |
| WHUCC-CS-CS-01-T | 4261 |
| WHUCC-EN-CS-02-T | 4258 |
| DCU-EN-CS-03-T | 4243 |
| WHUCC-CS-CS-02-T | 4216 |
| DCU-CS-CS-02-T | 4156 |
| KDEG-EN-CS-01-T | 4144 |
| DCU-EN-CS-02-T | 4098 |
| IMU-EN-CS-01-T | 4040 |
| KDEG-EN-CS-03-T | 3941 |
| KDEG-EN-CS-05-T | 3932 |
| KDEG-EN-CS-04-T | 3932 |
| QUTIS-EN-CS-03-T | 3845 |
| QUTIS-EN-CS-05-T | 3813 |
| DLUT-EN-CS-01-T | 3757 |
| DCU-EN-CS-01-T | 3751 |
| KECIR-CS-CS-05-T | 3649 |
| DLUT-EN-CS-03-T | 3647 |
| DLUT-EN-CS-02-T | 3647 |
| KECIR-CS-CS-04-T | 3641 |
| WUST-CS-CS-01-T | 3306 |
| CYUT-EN-CS-02-T | 3217 |
| CYUT-EN-CS-04-DN | 3150 |
| CYUT-EN-CS-03-D | 3042 |
| CYUT-EN-CS-01-T | 3023 |
| QUTIS-EN-CS-02-T | 2681 |
| QUTIS-EN-CS-01-T | 2315 |
| WUST-EN-CS-02-T | 2028 |
| WUST-EN-CS-01-T | 1940 |
| team name | coverage. |
| KDEG | 5177 |
| DCU | 5098 |
| KECIR | 5039 |
| WHUQA | 4988 |
| IMU | 4968 |
| WHUCC | 4900 |
| QUTIS | 4558 |
| CYUT | 3994 |
| DLUT | 3912 |
| WUST | 3760 |

**Table 13. Coverage of relevant documents summed across 73 CS topics (AFTER bug fix).**

| run name | coverage. |
|---|---|
| KDEG-CS-CS-01-T | 4792 |
| KDEG-CS-CS-02-DN | 4778 |
| WHUQA-CS-CS-01-T | 4769 |
| WHUQA-CS-CS-02-T | 4756 |
| KDEG-CS-CS-03-T | 4722 |
| KDEG-EN-CS-02-DN | 4697 |
| DCU-CS-CS-01-T | 4675 |
| IMU-CS-CS-02-T | 4651 |
| KECIR-CS-CS-03-T | 4620 |
| WHUQA-EN-CS-01-T | 4618 |
| WHUQA-EN-CS-02-T | 4607 |
| KDEG-CS-CS-05-T | 4514 |
| KDEG-CS-CS-04-T | 4514 |
| IMU-CS-CS-03-T | 4474 |
| IMU-CS-CS-01-T | 4474 |
| KECIR-CS-CS-01-T | 4455 |
| WHUCC-EN-CS-01-T | 4444 |
| KECIR-CS-CS-02-T | 4434 |
| WHUCC-EN-CS-03-T | 4293 |
| QUTIS-EN-CS-04-T | 4276 |
| WHUCC-CS-CS-01-T | 4268 |
| WHUCC-EN-CS-02-T | 4265 |
| DCU-EN-CS-03-T | 4248 |
| WHUCC-CS-CS-02-T | 4223 |
| DCU-CS-CS-02-T | 4163 |
| KDEG-EN-CS-01-T | 4151 |
| DCU-EN-CS-02-T | 4103 |
| IMU-EN-CS-01-T | 4046 |
| KDEG-EN-CS-03-T | 3945 |
| KDEG-EN-CS-05-T | 3936 |
| KDEG-EN-CS-04-T | 3936 |
| QUTIS-EN-CS-03-T | 3852 |
| QUTIS-EN-CS-05-T | 3817 |
| DLUT-EN-CS-01-T | 3762 |
| DCU-EN-CS-01-T | 3740 |
| KECIR-CS-CS-05-T | 3653 |
| DLUT-EN-CS-03-T | 3652 |
| DLUT-EN-CS-02-T | 3652 |
| KECIR-CS-CS-04-T | 3645 |
| WUST-CS-CS-01-T | 3311 |
| CYUT-EN-CS-02-T | 3219 |
| CYUT-EN-CS-04-DN | 3156 |
| CYUT-EN-CS-03-D | 3051 |
| CYUT-EN-CS-01-T | 3025 |
| QUTIS-EN-CS-02-T | 2687 |
| QUTIS-EN-CS-01-T | 2320 |
| WUST-EN-CS-02-T | 2013 |
| WUST-EN-CS-01-T | 1926 |
| team name | coverage. |
| KDEG | 5180 |
| DCU | 5093 |
| KECIR | 5044 |
| WHUQA | 4993 |
| IMU | 4972 |
| WHUCC | 4907 |
| QUTIS | 4562 |
| CYUT | 3998 |
| DLUT | 3917 |
| WUST | 3752 |

**Table 14. Coverage of relevant documents summed across 94 JA topics (BEFORE bug fix).**

| run name | coverage. |
|---|---|
| BRKLY-JA-JA-01-DN | 6515 |
| LTI-JA-JA-01-T | 6231 |
| LTI-JA-JA-02-T | 6226 |
| LTI-JA-JA-03-T | 6173 |
| BRKLY-JA-JA-02-T | 6087 |
| BRKLY-EN-JA-01-DN | 5904 |
| BRKLY-JA-JA-04-DN | 5722 |
| BRKLY-EN-JA-02-T | 5436 |
| LTI-EN-JA-01-T | 5386 |
| BRKLY-JA-JA-05-T | 5369 |
| LTI-EN-JA-02-T | 5354 |
| LTI-EN-JA-03-T | 5156 |
| CYUT-EN-JA-02-T | 4670 |
| CYUT-EN-JA-01-T | 4665 |
| CYUT-EN-JA-03-D | 3426 |
| CYUT-EN-JA-04-DN | 3213 |
| BRKLY-JA-JA-03-DN | 2964 |
| team name | coverage. |
| BRKLY | 7774 |
| LTI | 6924 |
| CYUT | 5611 |

**Table 15. Coverage of relevant documents summed across 94 JA topics (AFTER bug fix).**

| run name | coverage. |
|---|---|
| BRKLY-JA-JA-01-DN | 6505 |
| LTI-JA-JA-01-T | 6271 |
| LTI-JA-JA-02-T | 6266 |
| LTI-JA-JA-03-T | 6210 |
| BRKLY-JA-JA-02-T | 6151 |
| BRKLY-EN-JA-01-DN | 5892 |
| BRKLY-JA-JA-04-DN | 5660 |
| BRKLY-EN-JA-02-T | 5424 |
| LTI-EN-JA-01-T | 5314 |
| LTI-EN-JA-02-T | 5280 |
| BRKLY-JA-JA-05-T | 5218 |
| LTI-EN-JA-03-T | 5010 |
| CYUT-EN-JA-02-T | 4445 |
| CYUT-EN-JA-01-T | 4440 |
| CYUT-EN-JA-03-D | 3006 |
| CYUT-EN-JA-04-DN | 2787 |
| BRKLY-JA-JA-03-DN | 2611 |
| team name | coverage. |
| BRKLY | 7463 |
| LTI | 6825 |
| CYUT | 5234 |

**Table 16. Coverage of relevant documents summed across 87 CT topics (pool depth 100).**

| run name | coverage. |
|---|---|
| KDEG-CT-CT-02-DN | 4889 |
| KDEG-CT-CT-01-T | 4885 |
| KDEG-CT-CT-05-T | 4859 |
| KDEG-CT-CT-03-T | 4715 |
| KDEG-CT-CT-04-T | 4542 |
| KDEG-EN-CT-02-DN | 4095 |
| QUTIS-EN-CT-04-T | 4031 |
| KDEG-EN-CT-01-T | 3923 |
| KDEG-EN-CT-05-T | 3898 |
| QUTIS-EN-CT-03-T | 3761 |
| KDEG-EN-CT-03-T | 3608 |
| QUTIS-EN-CT-02-T | 3571 |
| KDEG-EN-CT-04-T | 3475 |
| QUTIS-EN-CT-01-T | 3321 |
| CYUT-EN-CT-02-T | 2951 |
| CYUT-EN-CT-01-T | 2769 |
| CYUT-EN-CT-04-DN | 2629 |
| CYUT-EN-CT-03-D | 2270 |
| QUTIS-EN-CT-05-T | 1901 |
| team name | coverage. |
| KDEG | 5258 |
| QUTIS | 4589 |
| CYUT | 3414 |

**Table 17. Unique relevant documents found summed across 73 CS topics (BEFORE bug fix).**

| run name | unique relevant |
|---|---|
| QUTIS-EN-CS-01-T | 23 |
| QUTIS-EN-CS-03-T | 21 |
| WHUCC-EN-CS-02-T | 20 |
| QUTIS-EN-CS-02-T | 18 |
| WHUCC-EN-CS-01-T | 17 |
| KECIR-CS-CS-02-T | 17 |
| KECIR-CS-CS-01-T | 17 |
| KECIR-CS-CS-04-T | 14 |
| QUTIS-EN-CS-04-T | 12 |
| WHUCC-EN-CS-03-T | 11 |
| WHUCC-CS-CS-02-T | 8 |
| WHUCC-CS-CS-01-T | 6 |
| QUTIS-EN-CS-05-T | 5 |
| KECIR-CS-CS-05-T | 5 |
| WUST-EN-CS-02-T | 4 |
| WUST-EN-CS-01-T | 4 |
| KDEG-EN-CS-05-T | 1 |
| KDEG-EN-CS-04-T | 1 |
| KDEG-EN-CS-03-T | 1 |
| KDEG-EN-CS-02-DN | 1 |
| KDEG-EN-CS-01-T | 1 |
| DCU-EN-CS-03-T | 1 |
| WUST-CS-CS-01-T | 0 |
| WHUQA-EN-CS-02-T | 0 |
| WHUQA-EN-CS-01-T | 0 |
| WHUQA-CS-CS-02-T | 0 |
| WHUQA-CS-CS-01-T | 0 |
| KECIR-CS-CS-03-T | 0 |
| KDEG-CS-CS-05-T | 0 |
| KDEG-CS-CS-04-T | 0 |
| KDEG-CS-CS-03-T | 0 |
| KDEG-CS-CS-02-DN | 0 |
| KDEG-CS-CS-01-T | 0 |
| IMU-EN-CS-01-T | 0 |
| IMU-CS-CS-03-T | 0 |
| IMU-CS-CS-02-T | 0 |
| IMU-CS-CS-01-T | 0 |
| DLUT-EN-CS-03-T | 0 |
| DLUT-EN-CS-02-T | 0 |
| DLUT-EN-CS-01-T | 0 |
| DCU-EN-CS-02-T | 0 |
| DCU-EN-CS-01-T | 0 |
| DCU-CS-CS-02-T | 0 |
| DCU-CS-CS-01-T | 0 |
| CYUT-EN-CS-04-DN | 0 |
| CYUT-EN-CS-03-D | 0 |
| CYUT-EN-CS-02-T | 0 |
| CYUT-EN-CS-01-T | 0 |
| team name | unique relevant |
| QUTIS | 24 |
| WHUCC | 22 |
| KECIR | 20 |
| WUST | 5 |
| KDEG | 1 |
| DCU | 1 |
| WHUQA | 0 |
| IMU | 0 |
| DLUT | 0 |
| CYUT | 0 |

**Table 18. Unique relevant documents found summed across 73 CS topics (AFTER bug fix).**

| run name | unique relevant |
|---|---|
| QUTIS-EN-CS-01-T | 23 |
| QUTIS-EN-CS-03-T | 21 |
| WHUCC-EN-CS-02-T | 20 |
| QUTIS-EN-CS-02-T | 18 |
| WHUCC-EN-CS-01-T | 17 |
| KECIR-CS-CS-02-T | 17 |
| KECIR-CS-CS-01-T | 17 |
| KECIR-CS-CS-04-T | 14 |
| QUTIS-EN-CS-04-T | 12 |
| WHUCC-EN-CS-03-T | 11 |
| WHUCC-CS-CS-02-T | 8 |
| WHUCC-CS-CS-01-T | 6 |
| QUTIS-EN-CS-05-T | 5 |
| KECIR-CS-CS-05-T | 5 |
| WUST-EN-CS-02-T | 3 |
| WUST-EN-CS-01-T | 3 |
| KDEG-EN-CS-05-T | 1 |
| KDEG-EN-CS-04-T | 1 |
| KDEG-EN-CS-03-T | 1 |
| KDEG-EN-CS-02-DN | 1 |
| KDEG-EN-CS-01-T | 1 |
| DCU-EN-CS-03-T | 1 |
| WUST-CS-CS-01-T | 0 |
| WHUQA-EN-CS-02-T | 0 |
| WHUQA-EN-CS-01-T | 0 |
| WHUQA-CS-CS-02-T | 0 |
| WHUQA-CS-CS-01-T | 0 |
| KECIR-CS-CS-03-T | 0 |
| KDEG-CS-CS-05-T | 0 |
| KDEG-CS-CS-04-T | 0 |
| KDEG-CS-CS-03-T | 0 |
| KDEG-CS-CS-02-DN | 0 |
| KDEG-CS-CS-01-T | 0 |
| IMU-EN-CS-01-T | 0 |
| IMU-CS-CS-03-T | 0 |
| IMU-CS-CS-02-T | 0 |
| IMU-CS-CS-01-T | 0 |
| DLUT-EN-CS-03-T | 0 |
| DLUT-EN-CS-02-T | 0 |
| DLUT-EN-CS-01-T | 0 |
| DCU-EN-CS-02-T | 0 |
| DCU-EN-CS-01-T | 0 |
| DCU-CS-CS-02-T | 0 |
| DCU-CS-CS-01-T | 0 |
| CYUT-EN-CS-04-DN | 0 |
| CYUT-EN-CS-03-D | 0 |
| CYUT-EN-CS-02-T | 0 |
| CYUT-EN-CS-01-T | 0 |
| team name | unique relevant |
| QUTIS | 24 |
| WHUCC | 22 |
| KECIR | 20 |
| WUST | 4 |
| KDEG | 1 |
| DCU | 1 |
| WHUQA | 0 |
| IMU | 0 |
| DLUT | 0 |
| CYUT | 0 |

**Table 19. Unique relevant documents found summed across 94 JA topics (BEFORE bug fix).**

| run name | unique relevant |
|---|---|
| BRKLY-JA-JA-01-DN | 349 |
| BRKLY-JA-JA-04-DN | 339 |
| BRKLY-JA-JA-02-T | 317 |
| BRKLY-EN-JA-01-DN | 297 |
| BRKLY-JA-JA-05-T | 296 |
| BRKLY-EN-JA-02-T | 283 |
| CYUT-EN-JA-03-D | 181 |
| CYUT-EN-JA-04-DN | 166 |
| CYUT-EN-JA-02-T | 147 |
| CYUT-EN-JA-01-T | 147 |
| LTI-JA-JA-03-T | 136 |
| LTI-JA-JA-02-T | 136 |
| LTI-JA-JA-01-T | 136 |
| BRKLY-JA-JA-03-DN | 99 |
| LTI-EN-JA-03-T | 90 |
| LTI-EN-JA-02-T | 86 |
| LTI-EN-JA-01-T | 86 |
| team name | unique relevant |
| BRKLY | 499 |
| CYUT | 207 |
| LTI | 171 |

**Table 20. Unique relevant documents found summed across 94 JA topics (AFTER bug fix).**

| run name | unique relevant |
|---|---|
| BRKLY-JA-JA-01-DN | 333 |
| BRKLY-JA-JA-02-T | 312 |
| BRKLY-JA-JA-04-DN | 310 |
| BRKLY-EN-JA-01-DN | 283 |
| BRKLY-EN-JA-02-T | 272 |
| BRKLY-JA-JA-05-T | 250 |
| LTI-JA-JA-03-T | 126 |
| LTI-JA-JA-02-T | 126 |
| LTI-JA-JA-01-T | 126 |
| CYUT-EN-JA-03-D | 118 |
| CYUT-EN-JA-02-T | 110 |
| CYUT-EN-JA-01-T | 110 |
| CYUT-EN-JA-04-DN | 103 |
| LTI-EN-JA-02-T | 80 |
| LTI-EN-JA-01-T | 80 |
| LTI-EN-JA-03-T | 79 |
| BRKLY-JA-JA-03-DN | 36 |
| team name | unique relevant |
| BRKLY | 425 |
| LTI | 151 |
| CYUT | 144 |

**Table 21. Unique relevant documents found summed across 87 CT topics (pool depth 100).**

| run name | unique relevant |
|---|---|
| KDEG-CT-CT-05-T | 385 |
| KDEG-CT-CT-01-T | 382 |
| KDEG-CT-CT-03-T | 369 |
| KDEG-CT-CT-04-T | 362 |
| KDEG-CT-CT-02-DN | 343 |
| KDEG-EN-CT-01-T | 250 |
| KDEG-EN-CT-05-T | 242 |
| KDEG-EN-CT-02-DN | 204 |
| KDEG-EN-CT-03-T | 202 |
| KDEG-EN-CT-04-T | 196 |
| CYUT-EN-CT-03-D | 76 |
| CYUT-EN-CT-04-DN | 74 |
| QUTIS-EN-CT-04-T | 58 |
| CYUT-EN-CT-01-T | 57 |
| QUTIS-EN-CT-02-T | 56 |
| CYUT-EN-CT-02-T | 56 |
| QUTIS-EN-CT-05-T | 34 |
| QUTIS-EN-CT-01-T | 32 |
| QUTIS-EN-CT-03-T | 30 |
| team name | unique relevant |
| KDEG | 420 |
| CYUT | 88 |
| QUTIS | 84 |

## 5. Run Ranking Forecasts

As mentioned earlier, we released run ranking forecasts to participants right after the run submission deadline, together with the system description of each run (See Table 10). The idea was to give participants "something to do" while they wait for the "true" rankings to arrive.

The way we created our pseudo-qrels files is very simple: We took the top 20% of each sorted depth-50 pool, and treated all of these documents as $L1$-relevant[6]. The assumption is that *documents retrieved by many runs at high ranks are relevant* [13, 16]. The value 20% was chosen because this yielded a total of 11606 "pseudo-relevant" documents across 100 CS topics, which is similar to the total number of truly relevant documents for the 97 CS topics from *NTCIR-7*, namely, 9488 documents (See [13] p.109 Table 35).

Tables 22-24 show the run ranking forecasts that were actually provided to the participants right after they submitted their runs. These tables should be compared with the "true" run rankings, shown in Tables 4, 6 and 7.

Table 25(a) shows the $\tau$ and $\tau_{ap}$ values for each pair of rankings based on the same effectiveness metric, one based on pseudo-qrels and the other based on true qrels. For computing YAR, the true ranking is taken as the ground truth. The correlation values are reasonably high for JA and CT, but it should be noted that our true qrels for these two test collections were created based on contributions from three teams only. That is, the JA and CT true qrels may be far from the "truth." We plan to "add more runs" to the JA pools to investigate this issue further in the near future.

The correlation values for CS in Table 25(a) are not as high as was expected. For example, the $\tau$ for the CS rankings with nDCG is only .571, which is much lower than the corresponding value reported for *NTCIR-7* IR4QA CS, namely, .859. We first hypothesized that this is because, while our pseudo-qrels for CS had 100 topics, the final true qrels for CS had only 73 topics after removal of topics with few relevant documents, and this had affected the system ranking. (Whereas, at NTCIR-7, we retained as many as 97 CS topics.) However, this hypothesis is not supported: Table 25(b) compares system ranking forecasts produced after removing the topics with few relevant documents for all three test collections. Thus the pseudo-qrels and the true qrels use exactly the same topic set[7]. However, it can be observed that the correlation values are similar to those in (a). That is, the fact that we lost some topics after relevance assessments is not the primary cause of many prediction errors for CS. Then, what is?

Ideally, pseudo-qrels should be useful for quickly spotting very good and very bad approaches, both within the set of runs submitted by one team (within-team) and across different teams (cross-team). However, given the limited accuracy of the run ranking forecasts, they may not be very useful for within-team comparisons where the effectiveness values of runs tend to be similar to one another. Let us therefore discuss whether pseudo-qrels are useful for predicting cross-team results.

By comparing the run ranking forecasts (Tables 22-24) with the true rankings (Tables 4, 6 and 7) at the participating *team* level, we can observe that:

- For CS, the forecast correctly identifies the top performer, namely, KDEG[8]. It also correctly identifies low performers. On the other hand, compared to Table 4, the forecast heavily overestimates DCU in Table 22: it believes that DCU is the second best CS team.
- For JA, as there are only three teams to rank, the forecast obtains the correct team ranking: LTI, BRKLY, CYUT.
- Similary for CT, as there are only three teams to rank, the forecast obtains the correct team ranking: KDEG, QUTIS, CYUT.

Why teams like DCU get overestimated in the forecast is not entirely clear, but it is probably worth mentioning that DCU covered many relevant documents (See Table 12), and only one of these documents were a unique relevant (See Table 17). That is, DCU managed to find many relevant documents which other teams were also able to find: this seems to suggest the limitation of our popularity-based pseudo-qrels approach.

Did any of the participating teams actually find the forecasts useful at all? Should we continue distributing the forecasts? If so, what level of accuracy is required?

Figures 4-6 visualise the accuracy of our run ranking forecasts in terms of Mean Q and Mean nDCG. The horizontal axis represents runs sorted by "true" effectiveness values, shown as "diamonds" in the graphs. Whereas, the squares represent the effectiveness values computed based on the pseudo-qrels.

## 6. Conclusions

This paper presented an overview of NTCIR-8 ACLIA IR4QA for retrieval of documents in Simplified Chinese, Traditional Chinese or Japanese, which involved 12 participating teams and 84 runs. Some initial findings from the organisers' point of view include:

- While 10 teams contributed CS runs, only 6 of them contributed unique relevant documents. Hence, diversifying runs is probably more important than increasing the number of participating teams in order to make our test collections less incomplete and more reusable.
- The top performing teams covered 92-96% of all known relevant documents. This suggests the possibility that our relevance assessments may be very incomplete. Incompleteness and reusability studies should be conducted with these test collections.
- The run ranking forecasts were not as accurate as was expected for CS runs. For example, a team that found many "popular" relevant documents was overestimated. On the other hand, the forecasts for CS successfully identified the top performing teams as well as the poor performers. For JA and CT, the three participating teams were correctly ranked by the pseudo-qrels.

[6] We used a similar approach at NTCIR-7, but used a fixed number of documents for every topic.

[7] Note that this is an artificial experiment just to separate out the effect of using different topic sets. Removing topics from pseudo-qrels based on the number of relevant documents is not practical.

[8] In contrast to the Q and AP rankings, the true nDCG ranking places IMU at the top in Table 3, but the difference between this run and the best KDEG run is negligible and statistically non-significant.

Finally, it appears that the questions we posed in our first IR4QA overview [13] remain largely unsolved at this point. We therefore copy and paste them here verbatim, as a reminder:

- What IR strategies work well for the purpose of QA, and for which languages? For example, does question classification help? How much?
- What are the general and language-specific challenges in crosslingual IR4QA?
- How incomplete are the IR4QA test collections? Are they reusable to some extent?
- If we conduct additional relevance assessments, how would that change the above circumstances?
- What are the best evaluation methods for IR4QA?
- How are IR4QA evaluation and the entire QA evaluation correlated?

We would like to discuss these questions with the IR4QA participants, as well as other attendees at the NTCIR-8 workshop.

## 7. Acknowledgments

We would like to thank all NTCIR-8 ACLIA participants, organisers and advisors for their efforts.

## 8. References

[1] Dongfeng, C., Shengqiao, K. and Yu, B.: Information Retrieval for Question Answer System at NTCIR-8, *Proceedings of NTCIR-8*, to appear, 2010.

[2] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM TOIS*, Vol. 20, No. 4, pp. 422-446, 2002.

[3] Larson, R.: Logistic Regression for IR4QA, *Proceedings of NTCIR-8*, to appear, 2010.

[4] Lin, M.-C., Li, M.-X., Hsu, C.-C.and Wu, S.-H.: Query Expansion from Wikipedia and Topic Web Crawler on CLIR, *Proceedings of NTCIR-8*, to appear, 2010.

[5] Liu, M., Zhou, B., Qi, L. and Zhang, Z.: Wikipedia Article Content Based Query Expansion in IR4QA System, *Proceedings of NTCIR-8*, to appear, 2010.

[6] Min, J., Jiang, J., Leveling, J., Jones, G. and Way, A.: DCU's Experiments in NTCIR-8 IR4QA Task, *Proceedings of NTCIR-8*, to appear, 2010.

[7] Mitamura, T. *et al.*: Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access, *Proceedings of NTCIR-7*, pp. 11-25, 2008. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C1/CCLQA/01-NTCIR7-OV-CCLQA-MitamuraT.pdf

[8] Mitamura, T. *et al.*: Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access, *Proceedings of NTCIR-8*, to appear, 2010.

[9] Ren, H., Ji, D. and Wan, J.: WHU Question Answering System at NTCIR-8 ACLIA Task, *Proceedings of NTCIR-8*, to appear, 2010.

[10] Sakai, T.: On the Task of Finding One Highly Relevant Document with High Precision, *Information Processing Society of Japan Transactions on Databases*, Vol.47, No.SIG 4 (TOD29), pp.13-27, 2006. Available at: http://www.jstage.jst.go.jp/article/ipsjdc/2/0/174/_pdf

[11] Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *Proceedings of the First Workshop on Evaluating Information Access (EVIA 2007)*, pp. 32-43, 2007. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/EVIA/1.pdf

[12] Sakai, T.: Evaluating Information Retrieval Metrics based on Bootstrap Hypothesis Tests, *Information Processing Society of Japan Transactions* Vol.48, No.SIG 9 (TOD35), pp. 11-28, 2007. http://www.jstage.jst.go.jp/article/ipsjdc/3/0/625/_pdf

[13] Sakai, T., Kando, N., Lin, C.-J., Mitamura, T., Shima, H., Ji, D., Chen, K.-H. and Nyberg, E.: Overview of the NTCIR-7 ACLIA IR4QA Task, *Proceedings of NTCIR-7*, pp. 77-114, 2008. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C1/IR4QA/01-NTCIR7-OV-IR4QA-SakaiT.pdf

[14] Sakai, T. and Robertson, S.: Modelling A User Population for Designing Information Retrieval Metrics, *Proceedings of the Second Workshop on Evaluating Information Access (EVIA 2008)*, pp. 30-41, 2008. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/EVIA2008/07-EVIA2008-SakaiT.pdf

[15] Sakai, T., Kando, N., Lin, C.-J., Song, R., Shima, H. and Mitamura, T.: NTCIR-7 ACLIA IR4QA Results based on Qrels Version 2, *Proceedings of NTCIR-7 online version*, 2009. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C1/IR4QA/14-NTCIR7-OV-IR4QA-SakaiT-20090626.pdf

[16] Sakai, T., Kando, N., Shima, H., Lin, C.-J. Song, R., Sugimoto, M. and Mitamura, T.: Ranking the NTCIR ACLIA IR4QA Systems without Relevance Assessments, *DBSJ Journal*, Vol. 8, No. 2, pp. 1-6, 2009. http://www.dbsj.org/Japanese/DBSJLetters/vol8/no2/dbsj-journal-08-02-001.pdf

[17] Shima, H. and Mitamura, T.: Bootstrap Pattern Learning for Open-Domain CLQA, *Proceedings of NTCIR-8*, to appear, 2010.

[18] Soboroff, I., Nicholas, C., and Cahan, P.: Ranking Retrieval Systems without Relevance Judgments, *Proceedings of ACM SIGIR 2001*, pp. 66-73, 2001.

[19] Su, X., Yan, X., Gao, G. and Wei, H.: IMU Experiment in IR4QA at NTCIR-8, *Proceedings of NTCIR-8*, to appear, 2010.

[20] Tang, L.-X., Trotman, A., Geva, S. and Xu, Y.: Wikipedia and Web document based Query Translation and Expansion for Cross-language IR, *Proceedings of NTCIR-8*, to appear, 2010.

[21] Teng, C., He, Y., Ji, D., Geng, Y., Mai, Z. and Lin, G.: Clustering and OCCC approaches in Document Re-ranking, *Proceedings of NTCIR-8*, to appear, 2010.

[22] Yilmaz, E., Aslam, J. and Robertson, S.: A New Rank Correlation Coefficient for Information Retrieval, *Proceedings of ACM SIGIR 2008*, pp. 587-594, 2008.

[23] Zezhong, L. and Degen, H.: DLUT IR4QA system in NTCIR-8, *Proceedings of NTCIR-8*, to appear, 2010.

[24] Zhou, D. and Wade, V.: The Effectiveness of Results Re-Ranking and Query Expansion in Cross-language Information Retrieval, *Proceedings of NTCIR-8*, to appear, 2010.

**Table 22. CS run results based on pseudo-qrels: mean effectiveness over 100 topics.**

| run | Mean AP | run | Mean Q | run | Mean nDCG |
|---|---|---|---|---|---|
| KDEG-CS-CS-01-T | 0.6404 | KDEG-CS-CS-01-T | 0.6639 | KDEG-CS-CS-01-T | 0.8265 |
| KDEG-CS-CS-03-T | 0.6369 | KDEG-CS-CS-03-T | 0.6591 | KDEG-CS-CS-03-T | 0.8177 |
| DCU-CS-CS-01-T | 0.6306 | DCU-CS-CS-01-T | 0.6525 | DCU-CS-CS-01-T | 0.8170 |
| KDEG-CS-CS-02-DN | 0.6173 | KDEG-CS-CS-02-DN | 0.6406 | KDEG-CS-CS-02-DN | 0.8090 |
| KDEG-EN-CS-01-T | 0.6101 | KDEG-EN-CS-01-T | 0.6334 | KDEG-EN-CS-01-T | 0.8089 |
| KDEG-EN-CS-02-DN | 0.5991 | KDEG-EN-CS-02-DN | 0.6231 | KDEG-EN-CS-02-DN | 0.8003 |
| KDEG-EN-CS-03-T | 0.5944 | KDEG-EN-CS-03-T | 0.6162 | KDEG-CS-CS-04-T | 0.7926 |
| IMU-CS-CS-01-T | 0.5901 | IMU-CS-CS-01-T | 0.6107 | KDEG-CS-CS-05-T | 0.7925 |
| DCU-EN-CS-02-T | 0.5840 | DCU-EN-CS-02-T | 0.6066 | DCU-EN-CS-02-T | 0.7898 |
| WHUQA-CS-CS-01-T | 0.5830 | WHUQA-CS-CS-01-T | 0.6042 | KDEG-EN-CS-03-T | 0.7885 |
| KDEG-CS-CS-04-T | 0.5784 | KDEG-CS-CS-04-T | 0.6031 | IMU-CS-CS-01-T | 0.7800 |
| KDEG-CS-CS-05-T | 0.5781 | KDEG-CS-CS-05-T | 0.6027 | IMU-CS-CS-03-T | 0.7676 |
| WHUQA-CS-CS-02-T | 0.5761 | WHUQA-CS-CS-02-T | 0.5983 | DCU-EN-CS-03-T | 0.7674 |
| IMU-CS-CS-03-T | 0.5569 | IMU-CS-CS-03-T | 0.5797 | KDEG-EN-CS-04-T | 0.7670 |
| DCU-EN-CS-03-T | 0.5476 | DCU-EN-CS-03-T | 0.5715 | KDEG-EN-CS-05-T | 0.7667 |
| KDEG-EN-CS-04-T | 0.5413 | KDEG-EN-CS-04-T | 0.5657 | WHUQA-CS-CS-02-T | 0.7619 |
| KDEG-EN-CS-05-T | 0.5407 | KDEG-EN-CS-05-T | 0.5651 | WHUQA-CS-CS-01-T | 0.7605 |
| QUTIS-EN-CS-04-T | 0.5305 | QUTIS-EN-CS-04-T | 0.5548 | QUTIS-EN-CS-04-T | 0.7542 |
| IMU-CS-CS-02-T | 0.5278 | IMU-CS-CS-02-T | 0.5511 | WHUCC-CS-CS-01-T | 0.7375 |
| WHUCC-CS-CS-01-T | 0.5186 | WHUCC-CS-CS-01-T | 0.5413 | IMU-CS-CS-02-T | 0.7351 |
| WHUCC-EN-CS-01-T | 0.5162 | WHUCC-EN-CS-01-T | 0.5389 | WHUCC-EN-CS-01-T | 0.7337 |
| WHUQA-EN-CS-02-T | 0.5134 | WHUQA-EN-CS-02-T | 0.5355 | WHUCC-EN-CS-02-T | 0.7215 |
| WHUQA-EN-CS-01-T | 0.5051 | WHUQA-EN-CS-01-T | 0.5266 | IMU-EN-CS-01-T | 0.7187 |
| IMU-EN-CS-01-T | 0.5004 | IMU-EN-CS-01-T | 0.5234 | WHUCC-CS-CS-02-T | 0.7153 |
| WHUCC-EN-CS-02-T | 0.4884 | WHUCC-EN-CS-02-T | 0.5114 | WHUQA-EN-CS-02-T | 0.7116 |
| WHUCC-CS-CS-02-T | 0.4834 | WHUCC-CS-CS-02-T | 0.5066 | DCU-CS-CS-02-T | 0.7093 |
| DCU-CS-CS-02-T | 0.4826 | DCU-CS-CS-02-T | 0.5059 | WHUQA-EN-CS-01-T | 0.7018 |
| WHUCC-EN-CS-03-T | 0.4478 | WHUCC-EN-CS-03-T | 0.4705 | QUTIS-EN-CS-05-T | 0.6892 |
| QUTIS-EN-CS-05-T | 0.4343 | QUTIS-EN-CS-05-T | 0.4600 | QUTIS-EN-CS-03-T | 0.6794 |
| QUTIS-EN-CS-03-T | 0.4257 | QUTIS-EN-CS-03-T | 0.4503 | WHUCC-EN-CS-03-T | 0.6778 |
| DLUT-EN-CS-03-T | 0.3882 | DLUT-EN-CS-03-T | 0.4114 | KECIR-CS-CS-01-T | 0.6533 |
| DLUT-EN-CS-02-T | 0.3882 | DLUT-EN-CS-02-T | 0.4114 | KECIR-CS-CS-02-T | 0.6470 |
| DLUT-EN-CS-01-T | 0.3817 | DLUT-EN-CS-01-T | 0.4052 | DLUT-EN-CS-03-T | 0.6461 |
| KECIR-CS-CS-01-T | 0.3406 | KECIR-CS-CS-01-T | 0.3722 | DLUT-EN-CS-02-T | 0.6461 |
| DCU-EN-CS-01-T | 0.3344 | KECIR-CS-CS-02-T | 0.3642 | DLUT-EN-CS-01-T | 0.6394 |
| KECIR-CS-CS-02-T | 0.3327 | DCU-EN-CS-01-T | 0.3581 | KECIR-CS-CS-03-T | 0.6273 |
| CYUT-EN-CS-01-T | 0.3315 | CYUT-EN-CS-01-T | 0.3517 | KECIR-CS-CS-04-T | 0.6007 |
| KECIR-CS-CS-04-T | 0.3237 | KECIR-CS-CS-03-T | 0.3511 | KECIR-CS-CS-05-T | 0.5793 |
| KECIR-CS-CS-03-T | 0.3196 | KECIR-CS-CS-04-T | 0.3481 | DCU-EN-CS-01-T | 0.5715 |
| CYUT-EN-CS-02-T | 0.3193 | CYUT-EN-CS-02-T | 0.3408 | CYUT-EN-CS-02-T | 0.5645 |
| QUTIS-EN-CS-02-T | 0.2895 | KECIR-CS-CS-05-T | 0.3088 | CYUT-EN-CS-01-T | 0.5625 |
| CYUT-EN-CS-04-DN | 0.2869 | QUTIS-EN-CS-02-T | 0.3077 | CYUT-EN-CS-04-DN | 0.5325 |
| KECIR-CS-CS-05-T | 0.2824 | CYUT-EN-CS-04-DN | 0.3070 | WUST-CS-CS-01-T | 0.5293 |
| WUST-CS-CS-01-T | 0.2758 | WUST-CS-CS-01-T | 0.2984 | QUTIS-EN-CS-02-T | 0.5024 |
| QUTIS-EN-CS-01-T | 0.2660 | QUTIS-EN-CS-01-T | 0.2817 | CYUT-EN-CS-03-D | 0.4867 |
| CYUT-EN-CS-03-D | 0.2531 | CYUT-EN-CS-03-D | 0.2722 | QUTIS-EN-CS-01-T | 0.4608 |
| WUST-EN-CS-02-T | 0.1609 | WUST-EN-CS-01-T | 0.1714 | WUST-EN-CS-01-T | 0.3671 |
| WUST-EN-CS-01-T | 0.1589 | WUST-EN-CS-02-T | 0.1704 | WUST-EN-CS-02-T | 0.3229 |

**Table 23. JA run results based on pseudo-qrels: mean effectiveness over 100 topics.**

| run | Mean AP | run | Mean Q | run | Mean nDCG |
|---|---|---|---|---|---|
| LTI-JA-JA-01-T | 0.5528 | LTI-JA-JA-01-T | 0.5724 | LTI-JA-JA-01-T | 0.7510 |
| LTI-JA-JA-02-T | 0.5521 | LTI-JA-JA-02-T | 0.5717 | LTI-JA-JA-02-T | 0.7503 |
| LTI-JA-JA-03-T | 0.5408 | LTI-JA-JA-03-T | 0.5605 | LTI-JA-JA-03-T | 0.7395 |
| LTI-EN-JA-01-T | 0.5200 | LTI-EN-JA-01-T | 0.5409 | LTI-EN-JA-01-T | 0.7273 |
| LTI-EN-JA-02-T | 0.5159 | LTI-EN-JA-02-T | 0.5364 | LTI-EN-JA-02-T | 0.7184 |
| LTI-EN-JA-03-T | 0.4972 | LTI-EN-JA-03-T | 0.5168 | BRKLY-JA-JA-01-DN | 0.7169 |
| BRKLY-JA-JA-01-DN | 0.4694 | BRKLY-JA-JA-01-DN | 0.4926 | BRKLY-JA-JA-02-T | 0.6968 |
| BRKLY-JA-JA-02-T | 0.4562 | BRKLY-JA-JA-02-T | 0.4799 | LTI-EN-JA-03-T | 0.6959 |
| BRKLY-EN-JA-01-DN | 0.4101 | BRKLY-EN-JA-01-DN | 0.4328 | BRKLY-JA-JA-04-DN | 0.6634 |
| BRKLY-JA-JA-04-DN | 0.3932 | BRKLY-JA-JA-04-DN | 0.4145 | BRKLY-EN-JA-01-DN | 0.6525 |
| BRKLY-JA-JA-05-T | 0.3783 | BRKLY-JA-JA-05-T | 0.4005 | BRKLY-JA-JA-05-T | 0.6432 |
| BRKLY-EN-JA-02-T | 0.3772 | BRKLY-EN-JA-02-T | 0.4005 | BRKLY-EN-JA-02-T | 0.6048 |
| CYUT-EN-JA-02-T | 0.2626 | CYUT-EN-JA-02-T | 0.2772 | CYUT-EN-JA-02-T | 0.5100 |
| CYUT-EN-JA-01-T | 0.2619 | CYUT-EN-JA-01-T | 0.2765 | CYUT-EN-JA-01-T | 0.5088 |
| BRKLY-JA-JA-03-DN | 0.1809 | BRKLY-JA-JA-03-DN | 0.1931 | CYUT-EN-JA-03-D | 0.3711 |
| CYUT-EN-JA-04-DN | 0.1640 | CYUT-EN-JA-03-D | 0.1711 | CYUT-EN-JA-04-DN | 0.3628 |
| CYUT-EN-JA-03-D | 0.1640 | CYUT-EN-JA-04-DN | 0.1697 | BRKLY-JA-JA-03-DN | 0.3496 |

**Table 24. CT run results based on pseudo-qrels: mean effectiveness over 100 topics.**

| run | Mean AP | run | Mean Q | run | Mean nDCG |
|---|---|---|---|---|---|
| KDEG-CT-CT-05-T | 0.6629 | KDEG-CT-CT-05-T | 0.6819 | KDEG-CT-CT-05-T | 0.8292 |
| KDEG-CT-CT-01-T | 0.6508 | KDEG-CT-CT-01-T | 0.6699 | KDEG-CT-CT-01-T | 0.8228 |
| KDEG-EN-CT-05-T | 0.6183 | KDEG-EN-CT-05-T | 0.6369 | KDEG-CT-CT-03-T | 0.7978 |
| KDEG-EN-CT-01-T | 0.6125 | KDEG-EN-CT-01-T | 0.6320 | KDEG-CT-CT-02-DN | 0.7938 |
| KDEG-CT-CT-02-DN | 0.6008 | KDEG-CT-CT-02-DN | 0.6222 | KDEG-EN-CT-01-T | 0.7936 |
| KDEG-CT-CT-03-T | 0.5924 | KDEG-CT-CT-03-T | 0.6155 | KDEG-EN-CT-05-T | 0.7930 |
| KDEG-EN-CT-02-DN | 0.5610 | KDEG-EN-CT-02-DN | 0.5824 | KDEG-CT-CT-04-T | 0.7651 |
| KDEG-EN-CT-03-T | 0.5434 | KDEG-EN-CT-03-T | 0.5654 | KDEG-EN-CT-03-T | 0.7595 |
| KDEG-CT-CT-04-T | 0.5214 | KDEG-CT-CT-04-T | 0.5476 | KDEG-EN-CT-02-DN | 0.7593 |
| QUTIS-EN-CT-04-T | 0.4750 | QUTIS-EN-CT-04-T | 0.5001 | QUTIS-EN-CT-04-T | 0.7211 |
| KDEG-EN-CT-04-T | 0.4661 | KDEG-EN-CT-04-T | 0.4903 | KDEG-EN-CT-04-T | 0.7164 |
| QUTIS-EN-CT-03-T | 0.3696 | QUTIS-EN-CT-03-T | 0.3934 | QUTIS-EN-CT-03-T | 0.6343 |
| CYUT-EN-CT-02-T | 0.3278 | CYUT-EN-CT-02-T | 0.3461 | CYUT-EN-CT-02-T | 0.5799 |
| QUTIS-EN-CT-02-T | 0.3215 | QUTIS-EN-CT-02-T | 0.3422 | QUTIS-EN-CT-02-T | 0.5648 |
| CYUT-EN-CT-01-T | 0.3137 | CYUT-EN-CT-01-T | 0.3303 | CYUT-EN-CT-01-T | 0.5535 |
| QUTIS-EN-CT-01-T | 0.2749 | QUTIS-EN-CT-01-T | 0.2940 | CYUT-EN-CT-04-DN | 0.5161 |
| CYUT-EN-CT-04-DN | 0.2616 | CYUT-EN-CT-04-DN | 0.2771 | QUTIS-EN-CT-01-T | 0.5160 |
| CYUT-EN-CT-03-D | 0.2295 | CYUT-EN-CT-03-D | 0.2437 | CYUT-EN-CT-03-D | 0.4663 |
| QUTIS-EN-CT-05-T | 0.1391 | QUTIS-EN-CT-05-T | 0.1509 | QUTIS-EN-CT-05-T | 0.2998 |

**Table 25. $\tau$ and $\tau_{ap}$ rank correlation: system rankings by pseudo-qrels vs those by true qrels, using the same effectiveness metric. (a) official forecast based on the original 100 topics; (b) "artificial" forecast using the pseudo-qrels but after removing the topics with few relevant documents.**

| | | (a) | (b) |
|---|---|---|---|
| CS | AP | .578/.458 | .583/.476 |
| (AFTER | Q | .603/.476 | .610/.491 |
| bug fix) | nDCG | .571/.443 | .599/.481 |
| JA | AP | .765/.759 | .779/.765 |
| (AFTER | Q | .765/.746 | .779/.752 |
| bug fix) | nDCG | .750/.730 | .765/.743 |
| CT | AP | .801/.683 | .801/.680 |
| (pool | Q | .789/.669 | .789/.666 |
| depth 50) | nDCG | .813/.635 | .778/.622 |
| CT | AP | .789/.572 | .789/.569 |
| (pool | Q | .778/.558 | .778/.555 |
| depth 100) | nDCG | .813/.635 | .778/.622 |

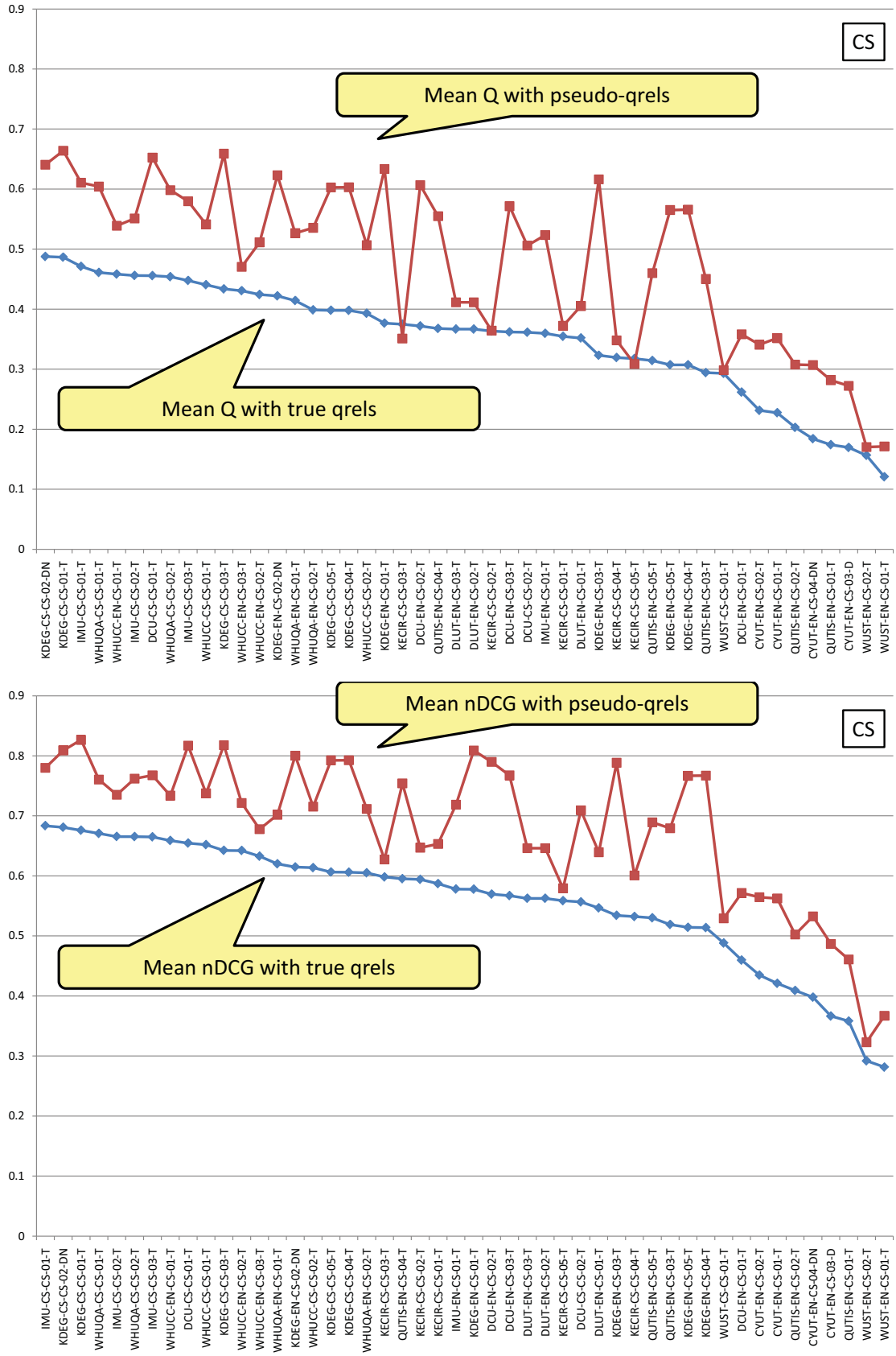


**Figure 4. Accuracy of system ranking based on pseudo-qrels with the bug-fixed qrels as the ground truth (CS runs).**

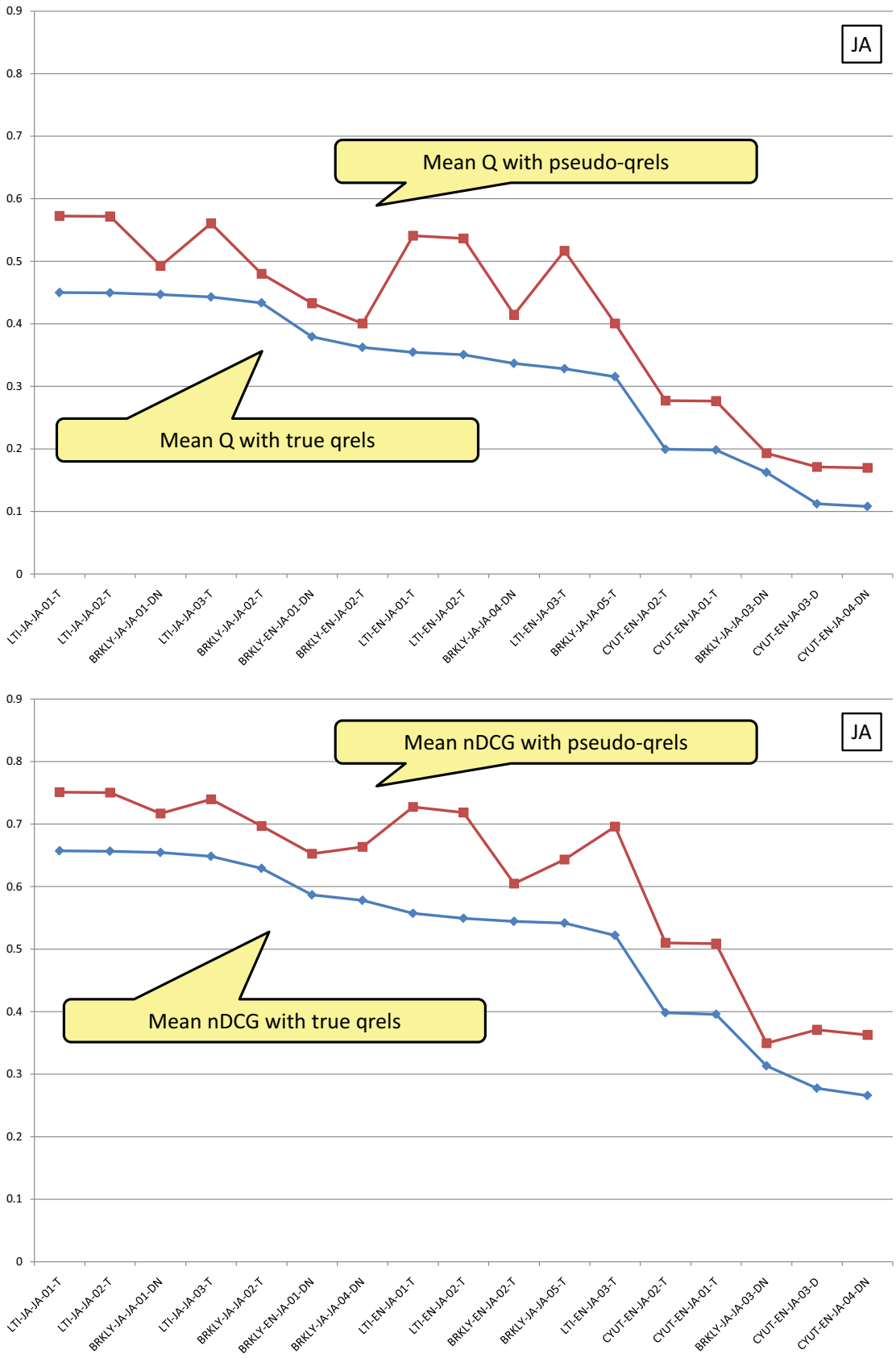


**Figure 5. Accuracy of system ranking based on pseudo-qrels with the bug-fixed qrels as the ground truth (JA runs).**


**Figure 6. Accuracy of system ranking based on pseudo-qrels with the pool-depth-100 qrels as the ground truth (CT runs).**

**Table 26. Appendix: Pool size for the CS relevance assessments. For example, "CS-0001" represents the topic ACLIA2-CS-0001.**

| topic | depth 50 | residual | depth 100 | topic | depth 50 | residual | depth 100 |
|---|---|---|---|---|---|---|---|
| CS-0001 | 302 | 265 | 567 | CS-0051 | 664 | 710 | 1374 |
| CS-0002 | 316 | 309 | 625 | CS-0052 | 185 | 166 | 351 |
| CS-0003 | 457 | 435 | 892 | CS-0053 | 439 | 545 | 984 |
| CS-0004 | 805 | 868 | 1673 | CS-0054 | 435 | 395 | 830 |
| CS-0005 | 429 | 432 | 861 | CS-0055 | 660 | 495 | 1155 |
| CS-0006 | 371 | 245 | 616 | CS-0056 | 582 | 706 | 1288 |
| CS-0007 | 808 | 707 | 1515 | CS-0057 | 604 | 582 | 1186 |
| CS-0008 | 866 | 830 | 1696 | CS-0058 | 409 | 310 | 719 |
| CS-0009 | 766 | 580 | 1346 | CS-0059 | 999 | 738 | 1737 |
| CS-0010 | 672 | 684 | 1356 | CS-0060 | 804 | 731 | 1535 |
| CS-0011 | 738 | 756 | 1494 | CS-0061 | 791 | 773 | 1564 |
| CS-0012 | 737 | 619 | 1356 | CS-0062 | 458 | 386 | 844 |
| CS-0013 | 575 | 582 | 1157 | CS-0063 | 702 | 621 | 1323 |
| CS-0014 | 457 | 451 | 908 | CS-0064 | 459 | 355 | 814 |
| CS-0015 | 388 | 310 | 698 | CS-0065 | 420 | 471 | 891 |
| CS-0016 | 335 | 415 | 750 | CS-0066 | 831 | 696 | 1527 |
| CS-0017 | 529 | 638 | 1167 | CS-0067 | 544 | 475 | 1019 |
| CS-0018 | 733 | 450 | 1183 | CS-0068 | 314 | 256 | 570 |
| CS-0019 | 294 | 198 | 492 | CS-0069 | 425 | 445 | 870 |
| CS-0020 | 620 | 412 | 1032 | CS-0070 | 538 | 597 | 1135 |
| CS-0021 | 265 | 300 | 565 | CS-0071 | 398 | 230 | 628 |
| CS-0022 | 411 | 473 | 884 | CS-0072 | 622 | 460 | 1082 |
| CS-0023 | 376 | 318 | 694 | CS-0073 | 539 | 796 | 1335 |
| CS-0024 | 396 | 402 | 798 | CS-0074 | 631 | 907 | 1538 |
| CS-0025 | 941 | 792 | 1733 | CS-0075 | 775 | 566 | 1341 |
| CS-0026 | 404 | 378 | 782 | CS-0076 | 603 | 421 | 1024 |
| CS-0027 | 548 | 614 | 1162 | CS-0077 | 486 | 278 | 764 |
| CS-0028 | 326 | 282 | 608 | CS-0078 | 343 | 288 | 631 |
| CS-0029 | 356 | 243 | 599 | CS-0079 | 1075 | 1040 | 2115 |
| CS-0030 | 281 | 320 | 601 | CS-0080 | 939 | 696 | 1635 |
| CS-0031 | 571 | 453 | 1024 | CS-0081 | 1013 | 863 | 1876 |
| CS-0032 | 331 | 283 | 614 | CS-0082 | 335 | 258 | 593 |
| CS-0033 | 560 | 423 | 983 | CS-0083 | 389 | 262 | 651 |
| CS-0034 | 502 | 429 | 931 | CS-0084 | 394 | 267 | 661 |
| CS-0035 | 892 | 734 | 1626 | CS-0085 | 679 | 442 | 1121 |
| CS-0036 | 775 | 620 | 1395 | CS-0086 | 454 | 350 | 804 |
| CS-0037 | 703 | 629 | 1332 | CS-0087 | 326 | 419 | 745 |
| CS-0038 | 1098 | 1098 | 2196 | CS-0088 | 604 | 643 | 1247 |
| CS-0039 | 478 | 594 | 1072 | CS-0089 | 1159 | 1068 | 2227 |
| CS-0040 | 220 | 133 | 353 | CS-0090 | 344 | 545 | 889 |
| CS-0041 | 425 | 338 | 763 | CS-0091 | 540 | 410 | 950 |
| CS-0042 | 817 | 667 | 1484 | CS-0092 | 563 | 414 | 977 |
| CS-0043 | 907 | 766 | 1673 | CS-0093 | 544 | 328 | 872 |
| CS-0044 | 507 | 426 | 933 | CS-0094 | 570 | 426 | 996 |
| CS-0045 | 743 | 520 | 1263 | CS-0095 | 854 | 811 | 1665 |
| CS-0046 | 684 | 494 | 1178 | CS-0096 | 797 | 591 | 1388 |
| CS-0047 | 759 | 942 | 1701 | CS-0097 | 724 | 484 | 1208 |
| CS-0048 | 1110 | 1085 | 2195 | CS-0098 | 342 | 274 | 616 |
| CS-0049 | 1022 | 997 | 2019 | CS-0099 | 428 | 317 | 745 |
| CS-0050 | 403 | 348 | 751 | CS-0100 | 500 | 377 | 877 |
| | | | | total | 58242 | 51971 | 110213 |

**Table 27. Appendix: Pool size for the JA relevance assessments. For example, "JA-0001" represents the topic ACLIA2-JA-0001.**

| topic | depth 50 | residual | depth 100 | topic | depth 50 | residual | depth 100 |
|---|---|---|---|---|---|---|---|
| JA-0001 | 202 | 126 | 328 | JA-0051 | 362 | 312 | 674 |
| JA-0002 | 368 | 393 | 761 | JA-0052 | 281 | 319 | 600 |
| JA-0003 | 297 | 220 | 517 | JA-0053 | 193 | 207 | 400 |
| JA-0004 | 192 | 129 | 321 | JA-0054 | 306 | 350 | 656 |
| JA-0005 | 360 | 338 | 698 | JA-0055 | 417 | 391 | 808 |
| JA-0006 | 345 | 343 | 688 | JA-0056 | 345 | 312 | 657 |
| JA-0007 | 337 | 298 | 635 | JA-0057 | 178 | 163 | 341 |
| JA-0008 | 233 | 263 | 496 | JA-0058 | 416 | 398 | 814 |
| JA-0009 | 337 | 313 | 650 | JA-0059 | 459 | 456 | 915 |
| JA-0010 | 239 | 190 | 429 | JA-0060 | 324 | 256 | 580 |
| JA-0011 | 353 | 328 | 681 | JA-0061 | 200 | 232 | 432 |
| JA-0012 | 212 | 169 | 381 | JA-0062 | 459 | 458 | 917 |
| JA-0013 | 342 | 355 | 697 | JA-0063 | 499 | 446 | 945 |
| JA-0014 | 229 | 239 | 468 | JA-0064 | 362 | 374 | 736 |
| JA-0015 | 307 | 283 | 590 | JA-0065 | 263 | 248 | 511 |
| JA-0016 | 333 | 268 | 601 | JA-0066 | 489 | 473 | 962 |
| JA-0017 | 354 | 329 | 683 | JA-0067 | 257 | 237 | 494 |
| JA-0018 | 310 | 280 | 590 | JA-0068 | 383 | 388 | 771 |
| JA-0019 | 296 | 281 | 577 | JA-0069 | 436 | 400 | 836 |
| JA-0020 | 372 | 341 | 713 | JA-0070 | 230 | 201 | 431 |
| JA-0021 | 262 | 181 | 443 | JA-0071 | 435 | 393 | 828 |
| JA-0022 | 198 | 114 | 312 | JA-0072 | 448 | 460 | 908 |
| JA-0023 | 402 | 355 | 757 | JA-0073 | 354 | 340 | 694 |
| JA-0024 | 316 | 279 | 595 | JA-0074 | 503 | 463 | 966 |
| JA-0025 | 293 | 263 | 556 | JA-0075 | 269 | 309 | 578 |
| JA-0026 | 174 | 234 | 408 | JA-0076 | 329 | 316 | 645 |
| JA-0027 | 377 | 346 | 723 | JA-0077 | 331 | 297 | 628 |
| JA-0028 | 269 | 275 | 544 | JA-0078 | 442 | 408 | 850 |
| JA-0029 | 206 | 237 | 443 | JA-0079 | 405 | 382 | 787 |
| JA-0030 | 326 | 318 | 644 | JA-0080 | 357 | 368 | 725 |
| JA-0031 | 188 | 196 | 384 | JA-0081 | 344 | 310 | 654 |
| JA-0032 | 238 | 252 | 490 | JA-0082 | 402 | 421 | 823 |
| JA-0033 | 346 | 296 | 642 | JA-0083 | 405 | 402 | 807 |
| JA-0034 | 205 | 138 | 343 | JA-0084 | 378 | 351 | 729 |
| JA-0035 | 263 | 317 | 580 | JA-0085 | 323 | 307 | 630 |
| JA-0036 | 237 | 239 | 476 | JA-0086 | 359 | 371 | 730 |
| JA-0037 | 344 | 357 | 701 | JA-0087 | 408 | 361 | 769 |
| JA-0038 | 233 | 162 | 395 | JA-0088 | 336 | 341 | 677 |
| JA-0039 | 245 | 343 | 588 | JA-0089 | 282 | 289 | 571 |
| JA-0040 | 262 | 208 | 470 | JA-0090 | 402 | 432 | 834 |
| JA-0041 | 269 | 245 | 514 | JA-0091 | 185 | 164 | 349 |
| JA-0042 | 288 | 238 | 526 | JA-0092 | 342 | 406 | 748 |
| JA-0043 | 264 | 355 | 619 | JA-0093 | 361 | 393 | 754 |
| JA-0044 | 171 | 132 | 303 | JA-0094 | 291 | 327 | 618 |
| JA-0045 | 462 | 418 | 880 | JA-0095 | 183 | 178 | 361 |
| JA-0046 | 453 | 422 | 875 | JA-0096 | 424 | 378 | 802 |
| JA-0047 | 321 | 320 | 641 | JA-0097 | 201 | 207 | 408 |
| JA-0048 | 263 | 208 | 471 | JA-0098 | 296 | 258 | 554 |
| JA-0049 | 183 | 202 | 385 | JA-0099 | 187 | 192 | 379 |
| JA-0050 | 146 | 139 | 285 | JA-0100 | 354 | 304 | 658 |
| | | | | total | 31417 | 30024 | 61441 |

**Table 28. Appendix: Pool size for the CT relevance assessments. For example, "CT-0001" represents the topic ACLIA2-CT-0001.**

| topic | depth 50 | residual | depth 100 | topic | depth 50 | residual | depth 100 |
|---|---|---|---|---|---|---|---|
| CT-0001 | 360 | 286 | 646 | CT-0051 | 467 | 431 | 898 |
| CT-0002 | 372 | 357 | 729 | CT-0052 | 395 | 321 | 716 |
| CT-0003 | 178 | 144 | 322 | CT-0053 | 540 | 491 | 1031 |
| CT-0004 | 250 | 265 | 515 | CT-0054 | 374 | 440 | 814 |
| CT-0005 | 466 | 353 | 819 | CT-0055 | 438 | 409 | 847 |
| CT-0006 | 390 | 446 | 836 | CT-0056 | 368 | 343 | 711 |
| CT-0007 | 521 | 430 | 951 | CT-0057 | 182 | 209 | 391 |
| CT-0008 | 496 | 452 | 948 | CT-0058 | 194 | 251 | 445 |
| CT-0009 | 355 | 323 | 678 | CT-0059 | 368 | 349 | 717 |
| CT-0010 | 257 | 283 | 540 | CT-0060 | 356 | 256 | 612 |
| CT-0011 | 387 | 366 | 753 | CT-0061 | 238 | 221 | 459 |
| CT-0012 | 294 | 259 | 553 | CT-0062 | 308 | 413 | 721 |
| CT-0013 | 333 | 285 | 618 | CT-0063 | 197 | 178 | 375 |
| CT-0014 | 469 | 442 | 911 | CT-0064 | 194 | 257 | 451 |
| CT-0015 | 576 | 489 | 1065 | CT-0065 | 189 | 156 | 345 |
| CT-0016 | 413 | 383 | 796 | CT-0066 | 331 | 321 | 652 |
| CT-0017 | 250 | 176 | 426 | CT-0067 | 174 | 168 | 342 |
| CT-0018 | 361 | 283 | 644 | CT-0068 | 256 | 194 | 450 |
| CT-0019 | 423 | 298 | 721 | CT-0069 | 305 | 355 | 660 |
| CT-0020 | 268 | 248 | 516 | CT-0070 | 472 | 392 | 864 |
| CT-0021 | 262 | 209 | 471 | CT-0071 | 179 | 151 | 330 |
| CT-0022 | 316 | 257 | 573 | CT-0072 | 226 | 196 | 422 |
| CT-0023 | 349 | 296 | 645 | CT-0073 | 463 | 416 | 879 |
| CT-0024 | 409 | 374 | 783 | CT-0074 | 329 | 327 | 656 |
| CT-0025 | 256 | 234 | 490 | CT-0075 | 356 | 421 | 777 |
| CT-0026 | 463 | 386 | 849 | CT-0076 | 161 | 302 | 463 |
| CT-0027 | 207 | 198 | 405 | CT-0077 | 206 | 249 | 455 |
| CT-0028 | 241 | 195 | 436 | CT-0078 | 277 | 302 | 579 |
| CT-0029 | 417 | 409 | 826 | CT-0079 | 367 | 372 | 739 |
| CT-0030 | 284 | 405 | 689 | CT-0080 | 297 | 278 | 575 |
| CT-0031 | 327 | 374 | 701 | CT-0081 | 420 | 398 | 818 |
| CT-0032 | 435 | 393 | 828 | CT-0082 | 373 | 331 | 704 |
| CT-0033 | 512 | 441 | 953 | CT-0083 | 201 | 176 | 377 |
| CT-0034 | 195 | 227 | 422 | CT-0084 | 438 | 389 | 827 |
| CT-0035 | 363 | 319 | 682 | CT-0085 | 316 | 353 | 669 |
| CT-0036 | 441 | 309 | 750 | CT-0086 | 187 | 233 | 420 |
| CT-0037 | 218 | 217 | 435 | CT-0087 | 364 | 285 | 649 |
| CT-0038 | 378 | 342 | 720 | CT-0088 | 168 | 206 | 374 |
| CT-0039 | 205 | 213 | 418 | CT-0089 | 291 | 261 | 552 |
| CT-0040 | 334 | 247 | 581 | CT-0090 | 282 | 191 | 473 |
| CT-0041 | 363 | 427 | 790 | CT-0091 | 455 | 419 | 874 |
| CT-0042 | 257 | 196 | 453 | CT-0092 | 193 | 232 | 425 |
| CT-0043 | 430 | 428 | 858 | CT-0093 | 314 | 255 | 569 |
| CT-0044 | 379 | 390 | 769 | CT-0094 | 360 | 290 | 650 |
| CT-0045 | 435 | 376 | 811 | CT-0095 | 346 | 337 | 683 |
| CT-0046 | 363 | 350 | 713 | CT-0096 | 401 | 383 | 784 |
| CT-0047 | 521 | 500 | 1021 | CT-0097 | 289 | 259 | 548 |
| CT-0048 | 299 | 295 | 594 | CT-0098 | 408 | 365 | 773 |
| CT-0049 | 243 | 204 | 447 | CT-0099 | 332 | 351 | 683 |
| CT-0050 | 121 | 85 | 206 | CT-0100 | 428 | 418 | 846 |
| | | | | total | 33215 | 31165 | 64380 |

**Table 29. Appendix: Number of judged nonrelevant ($L0$) and judged relevant ($L1$ and $L2$) documents for the 73 CS topics (AFTER bug fix).**

| | $L0$ | $L1$ | $L2$ | #relevant | #judged | | $L0$ | $L1$ | $L2$ | #relevant | #judged |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CS-0002 | 614 | 6 | 5 | 11 | 625 | CS-0059 | 1666 | 21 | 50 | 71 | 1737 |
| CS-0005 | 855 | 1 | 5 | 6 | 861 | CS-0061 | 1550 | 0 | 14 | 14 | 1564 |
| CS-0006 | 608 | 3 | 5 | 8 | 616 | CS-0062 | 625 | 0 | 219 | 219 | 844 |
| CS-0009 | 752 | 0 | 594 | 594 | 1346 | CS-0063 | 1317 | 1 | 5 | 6 | 1323 |
| CS-0012 | 1350 | 5 | 1 | 6 | 1356 | CS-0064 | 719 | 77 | 18 | 95 | 814 |
| CS-0015 | 483 | 2 | 213 | 215 | 698 | CS-0065 | 787 | 37 | 67 | 104 | 891 |
| CS-0016 | 745 | 0 | 5 | 5 | 750 | CS-0066 | 1448 | 0 | 79 | 79 | 1527 |
| CS-0018 | 1158 | 12 | 13 | 25 | 1183 | CS-0067 | 937 | 0 | 82 | 82 | 1019 |
| CS-0019 | 441 | 44 | 7 | 51 | 492 | CS-0068 | 471 | 39 | 60 | 99 | 570 |
| CS-0023 | 672 | 1 | 21 | 22 | 694 | CS-0069 | 851 | 5 | 14 | 19 | 870 |
| CS-0026 | 690 | 1 | 91 | 92 | 782 | CS-0070 | 1115 | 8 | 12 | 20 | 1135 |
| CS-0028 | 595 | 0 | 13 | 13 | 608 | CS-0071 | 276 | 13 | 339 | 352 | 628 |
| CS-0029 | 399 | 13 | 187 | 200 | 599 | CS-0072 | 1047 | 35 | 0 | 35 | 1082 |
| CS-0030 | 586 | 0 | 15 | 15 | 601 | CS-0073 | 1318 | 15 | 2 | 17 | 1335 |
| CS-0032 | 569 | 22 | 23 | 45 | 614 | CS-0074 | 1533 | 1 | 4 | 5 | 1538 |
| CS-0033 | 813 | 168 | 2 | 170 | 983 | CS-0075 | 874 | 0 | 467 | 467 | 1341 |
| CS-0035 | 1614 | 11 | 1 | 12 | 1626 | CS-0076 | 1018 | 0 | 6 | 6 | 1024 |
| CS-0036 | 1390 | 0 | 5 | 5 | 1395 | CS-0077 | 691 | 5 | 68 | 73 | 764 |
| CS-0038 | 2191 | 0 | 5 | 5 | 2196 | CS-0078 | 506 | 8 | 117 | 125 | 631 |
| CS-0039 | 1053 | 14 | 5 | 19 | 1072 | CS-0080 | 1498 | 15 | 122 | 137 | 1635 |
| CS-0040 | 279 | 0 | 74 | 74 | 353 | CS-0082 | 525 | 25 | 43 | 68 | 593 |
| CS-0041 | 587 | 0 | 176 | 176 | 763 | CS-0083 | 642 | 5 | 4 | 9 | 651 |
| CS-0042 | 1403 | 60 | 21 | 81 | 1484 | CS-0084 | 644 | 3 | 14 | 17 | 661 |
| CS-0043 | 1600 | 66 | 7 | 73 | 1673 | CS-0085 | 1071 | 12 | 38 | 50 | 1121 |
| CS-0044 | 905 | 17 | 11 | 28 | 933 | CS-0087 | 710 | 8 | 27 | 35 | 745 |
| CS-0045 | 1196 | 0 | 67 | 67 | 1263 | CS-0090 | 879 | 1 | 9 | 10 | 889 |
| CS-0046 | 1138 | 4 | 36 | 40 | 1178 | CS-0091 | 763 | 9 | 178 | 187 | 950 |
| CS-0047 | 1664 | 34 | 3 | 37 | 1701 | CS-0092 | 881 | 1 | 95 | 96 | 977 |
| CS-0048 | 2188 | 1 | 6 | 7 | 2195 | CS-0093 | 745 | 5 | 122 | 127 | 872 |
| CS-0049 | 2011 | 0 | 8 | 8 | 2019 | CS-0094 | 693 | 11 | 292 | 303 | 996 |
| CS-0051 | 1368 | 4 | 2 | 6 | 1374 | CS-0095 | 1571 | 66 | 28 | 94 | 1665 |
| CS-0052 | 339 | 12 | 0 | 12 | 351 | CS-0096 | 1318 | 15 | 55 | 70 | 1388 |
| CS-0053 | 969 | 7 | 8 | 15 | 984 | CS-0097 | 1191 | 0 | 17 | 17 | 1208 |
| CS-0054 | 751 | 64 | 15 | 79 | 830 | CS-0098 | 599 | 3 | 14 | 17 | 616 |
| CS-0056 | 1273 | 0 | 15 | 15 | 1288 | CS-0099 | 738 | 0 | 7 | 7 | 745 |
| CS-0057 | 1163 | 0 | 23 | 23 | 1186 | CS-0100 | 872 | 0 | 5 | 5 | 877 |
| CS-0058 | 700 | 9 | 10 | 19 | 719 | | | | | | |
| | | | | | | total | 71201 | 1025 | 4391 | 5416 | 76617 |

**Table 30. Appendix: Number of judged nonrelevant ($L0$) and judged relevant ($L1$ and $L2$) documents for the 94 JA topics (AFTER bug fix).**

| | $L0$ | $L1$ | $L2$ | #relevant | #judged | | $L0$ | $L1$ | $L2$ | #relevant | #judged |
|---|---|---|---|---|---|---|---|---|---|---|---|
| JA-0001 | 269 | 51 | 8 | 59 | 328 | JA-0050 | 240 | 44 | 1 | 45 | 285 |
| JA-0002 | 718 | 34 | 9 | 43 | 761 | JA-0051 | 360 | 76 | 238 | 314 | 674 |
| JA-0003 | 380 | 113 | 24 | 137 | 517 | JA-0052 | 508 | 57 | 35 | 92 | 600 |
| JA-0004 | 138 | 46 | 137 | 183 | 321 | JA-0053 | 391 | 2 | 7 | 9 | 400 |
| JA-0005 | 416 | 240 | 42 | 282 | 698 | JA-0054 | 625 | 27 | 4 | 31 | 656 |
| JA-0006 | 626 | 59 | 3 | 62 | 688 | JA-0055 | 775 | 27 | 6 | 33 | 808 |
| JA-0007 | 457 | 90 | 88 | 178 | 635 | JA-0056 | 633 | 19 | 5 | 24 | 657 |
| JA-0008 | 448 | 4 | 44 | 48 | 496 | JA-0057 | 218 | 91 | 32 | 123 | 341 |
| JA-0009 | 483 | 59 | 108 | 167 | 650 | JA-0058 | 736 | 6 | 72 | 78 | 814 |
| JA-0010 | 372 | 33 | 24 | 57 | 429 | JA-0060 | 418 | 133 | 29 | 162 | 580 |
| JA-0011 | 523 | 11 | 147 | 158 | 681 | JA-0061 | 412 | 5 | 15 | 20 | 432 |
| JA-0012 | 294 | 65 | 22 | 87 | 381 | JA-0062 | 901 | 5 | 11 | 16 | 917 |
| JA-0013 | 620 | 65 | 12 | 77 | 697 | JA-0063 | 511 | 84 | 350 | 434 | 945 |
| JA-0014 | 403 | 38 | 27 | 65 | 468 | JA-0064 | 668 | 52 | 16 | 68 | 736 |
| JA-0015 | 518 | 47 | 25 | 72 | 590 | JA-0065 | 399 | 21 | 91 | 112 | 511 |
| JA-0016 | 451 | 105 | 45 | 150 | 601 | JA-0067 | 483 | 6 | 5 | 11 | 494 |
| JA-0017 | 485 | 198 | 0 | 198 | 683 | JA-0068 | 765 | 6 | 0 | 6 | 771 |
| JA-0018 | 334 | 222 | 34 | 256 | 590 | JA-0069 | 748 | 38 | 50 | 88 | 836 |
| JA-0019 | 499 | 67 | 11 | 78 | 577 | JA-0070 | 416 | 6 | 9 | 15 | 431 |
| JA-0020 | 574 | 118 | 21 | 139 | 713 | JA-0072 | 899 | 7 | 2 | 9 | 908 |
| JA-0021 | 240 | 183 | 20 | 203 | 443 | JA-0073 | 654 | 37 | 3 | 40 | 694 |
| JA-0022 | 303 | 4 | 5 | 9 | 312 | JA-0074 | 961 | 0 | 5 | 5 | 966 |
| JA-0023 | 149 | 538 | 70 | 608 | 757 | JA-0075 | 532 | 43 | 3 | 46 | 578 |
| JA-0024 | 475 | 54 | 66 | 120 | 595 | JA-0076 | 549 | 79 | 17 | 96 | 645 |
| JA-0025 | 378 | 84 | 94 | 178 | 556 | JA-0077 | 599 | 28 | 1 | 29 | 628 |
| JA-0026 | 396 | 7 | 5 | 12 | 408 | JA-0078 | 787 | 53 | 10 | 63 | 850 |
| JA-0027 | 610 | 99 | 14 | 113 | 723 | JA-0079 | 781 | 1 | 5 | 6 | 787 |
| JA-0028 | 504 | 25 | 15 | 40 | 544 | JA-0080 | 699 | 13 | 13 | 26 | 725 |
| JA-0029 | 394 | 24 | 25 | 49 | 443 | JA-0081 | 639 | 5 | 10 | 15 | 654 |
| JA-0030 | 636 | 3 | 5 | 8 | 644 | JA-0083 | 787 | 3 | 17 | 20 | 807 |
| JA-0031 | 300 | 40 | 44 | 84 | 384 | JA-0084 | 605 | 23 | 101 | 124 | 729 |
| JA-0032 | 395 | 29 | 66 | 95 | 490 | JA-0085 | 442 | 33 | 155 | 188 | 630 |
| JA-0033 | 443 | 28 | 171 | 199 | 642 | JA-0086 | 725 | 3 | 2 | 5 | 730 |
| JA-0034 | 53 | 10 | 280 | 290 | 343 | JA-0087 | 739 | 21 | 9 | 30 | 769 |
| JA-0035 | 551 | 13 | 16 | 29 | 580 | JA-0088 | 668 | 6 | 3 | 9 | 677 |
| JA-0036 | 397 | 22 | 57 | 79 | 476 | JA-0089 | 516 | 50 | 5 | 55 | 571 |
| JA-0038 | 362 | 28 | 5 | 33 | 395 | JA-0090 | 804 | 0 | 30 | 30 | 834 |
| JA-0039 | 555 | 20 | 13 | 33 | 588 | JA-0091 | 150 | 104 | 95 | 199 | 349 |
| JA-0040 | 424 | 26 | 20 | 46 | 470 | JA-0092 | 735 | 4 | 9 | 13 | 748 |
| JA-0041 | 474 | 38 | 2 | 40 | 514 | JA-0093 | 742 | 12 | 0 | 12 | 754 |
| JA-0042 | 510 | 9 | 7 | 16 | 526 | JA-0094 | 601 | 1 | 16 | 17 | 618 |
| JA-0044 | 261 | 37 | 5 | 42 | 303 | JA-0095 | 217 | 13 | 131 | 144 | 361 |
| JA-0045 | 823 | 51 | 6 | 57 | 880 | JA-0096 | 777 | 25 | 0 | 25 | 802 |
| JA-0046 | 855 | 20 | 0 | 20 | 875 | JA-0097 | 384 | 11 | 13 | 24 | 408 |
| JA-0047 | 624 | 14 | 3 | 17 | 641 | JA-0098 | 478 | 48 | 28 | 76 | 554 |
| JA-0048 | 456 | 15 | 0 | 15 | 471 | JA-0099 | 297 | 68 | 11 | 79 | 376 |
| JA-0049 | 374 | 11 | 0 | 11 | 385 | JA-0100 | 530 | 58 | 70 | 128 | 658 |
| | | | | | | total | 48454 | 4551 | 3585 | 8136 | 56590 |

**Table 31. Appendix: Number of judged nonrelevant ($L0$) and judged relevant ($L1$ and $L2$) documents for the 87 CT topics.**

| | $L0$ | $L1$ | $L2$ | #relevant | #judged | | $L0$ | $L1$ | $L2$ | #relevant | #judged |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CT-0001 | 548 | 34 | 64 | 98 | 646 | CT-0048 | 553 | 21 | 20 | 41 | 594 |
| CT-0002 | 706 | 14 | 9 | 23 | 729 | CT-0049 | 205 | 202 | 40 | 242 | 447 |
| CT-0003 | 109 | 213 | 0 | 213 | 322 | CT-0050 | 162 | 43 | 1 | 44 | 206 |
| CT-0004 | 491 | 10 | 14 | 24 | 515 | CT-0051 | 870 | 27 | 1 | 28 | 898 |
| CT-0005 | 808 | 4 | 7 | 11 | 819 | CT-0052 | 635 | 0 | 81 | 81 | 716 |
| CT-0006 | 784 | 48 | 4 | 52 | 836 | CT-0053 | 988 | 21 | 22 | 43 | 1031 |
| CT-0007 | 871 | 3 | 77 | 80 | 951 | CT-0055 | 732 | 24 | 91 | 115 | 847 |
| CT-0008 | 804 | 18 | 126 | 144 | 948 | CT-0056 | 577 | 1 | 133 | 134 | 711 |
| CT-0009 | 600 | 44 | 34 | 78 | 678 | CT-0057 | 365 | 2 | 24 | 26 | 391 |
| CT-0010 | 485 | 21 | 34 | 55 | 540 | CT-0058 | 357 | 77 | 11 | 88 | 445 |
| CT-0011 | 705 | 17 | 31 | 48 | 753 | CT-0059 | 684 | 22 | 11 | 33 | 717 |
| CT-0012 | 514 | 16 | 23 | 39 | 553 | CT-0060 | 464 | 146 | 2 | 148 | 612 |
| CT-0013 | 584 | 10 | 24 | 34 | 618 | CT-0061 | 444 | 0 | 15 | 15 | 459 |
| CT-0014 | 869 | 34 | 8 | 42 | 911 | CT-0063 | 277 | 44 | 37 | 81 | 358 |
| CT-0015 | 1044 | 0 | 21 | 21 | 1065 | CT-0064 | 332 | 12 | 107 | 119 | 451 |
| CT-0016 | 777 | 1 | 18 | 19 | 796 | CT-0065 | 326 | 3 | 16 | 19 | 345 |
| CT-0017 | 400 | 8 | 18 | 26 | 426 | CT-0066 | 575 | 25 | 52 | 77 | 652 |
| CT-0019 | 406 | 31 | 284 | 315 | 721 | CT-0067 | 312 | 9 | 21 | 30 | 342 |
| CT-0020 | 361 | 60 | 95 | 155 | 516 | CT-0068 | 430 | 2 | 18 | 20 | 450 |
| CT-0021 | 287 | 13 | 171 | 184 | 471 | CT-0069 | 646 | 0 | 14 | 14 | 660 |
| CT-0022 | 563 | 0 | 10 | 10 | 573 | CT-0070 | 716 | 3 | 145 | 148 | 864 |
| CT-0023 | 476 | 8 | 161 | 169 | 645 | CT-0071 | 277 | 8 | 45 | 53 | 330 |
| CT-0024 | 766 | 5 | 12 | 17 | 783 | CT-0073 | 870 | 1 | 8 | 9 | 879 |
| CT-0025 | 386 | 78 | 26 | 104 | 490 | CT-0074 | 641 | 13 | 2 | 15 | 656 |
| CT-0026 | 835 | 12 | 2 | 14 | 849 | CT-0075 | 762 | 1 | 14 | 15 | 777 |
| CT-0027 | 395 | 7 | 3 | 10 | 405 | CT-0076 | 442 | 6 | 15 | 21 | 463 |
| CT-0028 | 418 | 0 | 18 | 18 | 436 | CT-0077 | 443 | 10 | 2 | 12 | 455 |
| CT-0029 | 816 | 4 | 6 | 10 | 826 | CT-0082 | 688 | 5 | 11 | 16 | 704 |
| CT-0030 | 670 | 10 | 9 | 19 | 689 | CT-0083 | 342 | 12 | 23 | 35 | 377 |
| CT-0031 | 648 | 24 | 29 | 53 | 701 | CT-0084 | 821 | 4 | 2 | 6 | 827 |
| CT-0032 | 822 | 3 | 3 | 6 | 828 | CT-0086 | 391 | 16 | 13 | 29 | 420 |
| CT-0033 | 940 | 10 | 3 | 13 | 953 | CT-0087 | 644 | 2 | 3 | 5 | 649 |
| CT-0034 | 405 | 0 | 17 | 17 | 422 | CT-0088 | 366 | 1 | 7 | 8 | 374 |
| CT-0035 | 676 | 4 | 2 | 6 | 682 | CT-0089 | 519 | 12 | 21 | 33 | 552 |
| CT-0036 | 706 | 8 | 36 | 44 | 750 | CT-0090 | 413 | 36 | 24 | 60 | 473 |
| CT-0037 | 404 | 14 | 17 | 31 | 435 | CT-0091 | 829 | 40 | 5 | 45 | 874 |
| CT-0039 | 385 | 16 | 17 | 33 | 418 | CT-0092 | 375 | 43 | 7 | 50 | 425 |
| CT-0040 | 238 | 6 | 337 | 343 | 581 | CT-0093 | 519 | 17 | 33 | 50 | 569 |
| CT-0042 | 429 | 22 | 2 | 24 | 453 | CT-0094 | 553 | 96 | 1 | 97 | 650 |
| CT-0043 | 845 | 3 | 10 | 13 | 858 | CT-0095 | 677 | 0 | 6 | 6 | 683 |
| CT-0044 | 762 | 5 | 2 | 7 | 769 | CT-0096 | 769 | 2 | 13 | 15 | 784 |
| CT-0045 | 358 | 3 | 450 | 453 | 811 | CT-0097 | 429 | 15 | 104 | 119 | 548 |
| CT-0046 | 700 | 1 | 12 | 13 | 713 | CT-0099 | 646 | 34 | 3 | 37 | 683 |
| CT-0047 | 901 | 116 | 4 | 120 | 1021 | | | | | | |
| | | | | | | total | 49763 | 2016 | 3474 | 5490 | 55253 |