Overview of the NTCIR-7 ACLIA IR4QA Task

Tetsuya Sakai[†] Noriko Kando[‡] Chuan-Jie Lin[⋆] Teruko Mitamura[∗]
Hideki Shima[∗] Donghong Ji[♯] Kuang-Hua Chen[♭] Eric Nyberg[∗]
†NewsWatch, Inc. ‡National Institute of Informatics ∗Carnegie Mellon University
⋆National Taiwan Ocean University ♭National Taiwan University
♯Wuhan University
tetsuyasakai@acm.org

Abstract

This paper presents an overview of the IR4QA (Information Retrieval for Question Answering) Task of the NTCIR-7 ACLIA (Advanced Cross-lingual Information Access) Task Cluster. IR4QA evaluates traditional ranked retrieval of documents using well-studied metrics such as Average Precision, but the retrieval task is embedded in the context of cross-lingual question answering. That is, document retrieval is treated as a component of the entire question answering system. This paper concentrates on how relevance assessments for the Simpified Chinese, Traditional Chinese and Japanese IR4QA test collections were obtained, and the outcome of the formal IR4QA evaluation using the three collections. For the relationship between IR4QA and the entire ACLIA task cluster, we refer the reader to the overview paper of ACLIA [17]. For details of the individual IR4QA systems, we refer the reader to the participants' reports.
Keywords: *test collections, pooling, evaluation metrics, evaluation package.*

1 Introduction

This paper presents an overview of the IR4QA (Information Retrieval for Question Answering) Task of the NTCIR-7 ACLIA (Advanced Cross-lingual Information Access) Task Cluster. IR4QA evaluates traditional ranked retrieval of documents using well-studied metrics such as Average Precision (AP), but the retrieval task is embedded in the context of cross-lingual question answering. That is, document retrieval is treated as a component of the entire question answering system. This paper concentrates on how relevance assessments for the Simpified Chinese (CS), Traditional Chinese (CT) and Japanese (JA) IR4QA test collections were obtained, and the outcome of the formal IR4QA evaluation using the three collections. For the relationship between IR4QA and the entire ACLIA task cluster, we refer the reader to the overview paper of ACLIA [17]. For details of the individual IR4QA systems, we refer the reader to the participants' reports [3, 6, 7, 14, 12, 13, 15, 16, 26, 28, 29, 31]. Table 1 provides a list of IR4QA participants.

The remainder of this paper is organised as follows. Section 2 describes our pooling method for performing relevance assessments for the CS, CT and JA document collections of ACLIA. Section 3 describes how relevance assessment data ("qrels") were obtained. Section 4 describes the IR evaluation package that we have released to the participants, and defines the three evaluation metrics we use for ranking the IR4QA systems, namely, AP, Q-measure (Q) and a version of normalised Discounted Cumulative Gain (nDCG). Section 5 describes our preliminary "pseudo-qrels" experiments, which ranks participating systems without relevance assessments. Section 6 presents the official IR4QA results using the "real" qrels, a brief overview of techniques used by the participants, and a preliminary analysis of the correlation between our pseudo-qrels and real-qrels results. Finally, Section 7 summarises our initial findings as the organisers of IR4QA.

2 Pooling

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). For example, a total of 22 CS-CS monolingual runs (i.e., runs that used Simplified Chinese topics and retrieved Simplified Chinese documents) and a total of 18 EN-CS crosslingual runs (i.e., runs that used English topics and retrieved Simplified Chinese documents) were submitted. Hence a total of 40 CS runs (i.e., runs that retrieved Simplified Chinese documents) were used in pooling for relevance assessments.

A *traditional* pooling procedure is as follows [5]. Let S be the set of systems (i.e., runs) that will con-

Table 1. IR4QA participants.

team name	organisation
BRKLY	University of California, Berkeley
CMUJAV	Language Technologies Institute, Carnegie Mellon University
CYUT	Chaoyang University of Technology
HIT	Heilongjiang Institute of Technology User Group: HIT2 NLP Joint Lab
KECIR	Shenyang Institute of Aeronautical Engineering
MITEL	Institute of Computing Technology, Chinese Academy of Sciences
NLPAI	College of Computer Science and Technology, Wuhan University of Science and Technology
NTUBROWS	CSIE, National Taiwan University
OT	Open Text Corporation
RALI	University of Montreal
TA	Toyohashi University of Technology
WHUCC	Computer Center of Wuhan University

Table 2. Number of IR4QA runs submitted.

team	CS-CS	EN-CS	CT-CT	EN-CT	JA-JA	EN-JA
BRKLY					4	
CMUJAV	2	2			5	5
CYUT		3		3		3
HIT		4				
KECIR	3					
MITEL		5	4			
NLPAI	5					
NTUBROWS			5			
OT	5		5		5	
RALI	5	4	5	4		
TA						3
WHUCC	2					
total by lang. pair	22	18∗	19	7	14	11
total by document lang.	40		26		25	

∗One team submitted seven EN-CS runs but the sixth and the seventh runs were not used for pooling and are excluded from our analyses.

tribute to the pool, and let $s \in S$. For a particular topic, let $D_X(s)$ denote the set of documents which are the top X documents of s. The depth-X pool for this topic, which we denote by P_X, is defined as:

$$P_X = \cup_{s \in S} D_X(s) \ .$$

Typically, X is a constant (e.g., $X = 100$) across topics, and all the documents in P_X are judged by relevance assessors for each topic. As a result, each document $d \in P_X$ will be assigned a relevance level: In the case of IR4QA, d can be either judged nonrelevant (which we denote as $L0$), partially relevant ($L1$) or relevant ($L2$). More details on these relevance levels will follow in Section 3.

However, due to time constraints (we only had two weeks for collecting relevance assessments of 100 topics!) and limited human resource for relevance assessments, we took a strategy similar to the one used in the NTCIR-3 PATENT task [8, 11]:

1. For each topic, create a depth-X pool, for $X = 30, 50, 70, 90, 100$.
2. For each topic, create $P'_{50} = P_{50} - P_{30}$, $P'_{70} = P_{70} - P_{50}$, $P'_{90} = P_{90} - P_{70}$ and $P'_{100} = P_{100} - P_{90}$.
3. For each topic, let the relevance assessors assess all documents in P_{30}. If the assessors complete this task for a particular topic, and if there is still enough time, let the assessors move on to P'_{50} for this topic. Similarly, if the assessors have judged all documents in P'_{50}, and if there is still enough time, let them move on to P'_{70}, and so on.

Note, for example, that judging P_{30} and then judging P'_{50} is equivalent to judging P_{50}, the depth-50 pool. However, the final pool depth can differ across topics: For some topics, the entire depth-100 pool may be judged; for others, only the depth-50 pool may be judged, depending on how efficiently the assessors can make the judgments. Tables 29-31 in the Appendix show the size of these document sets for each topic for each document collection. Among these topics, ACLIA1-CS-T{86, 331, 362}, ACLIA1-CT-T{403, 406, 412, 417, 433} and ACLIA1-JA-T{116, 127} were eventually removed from our evaluation data as no relevant documents were found for them. Hence we have 97 CS topics, 95 CT topics and 98 JA topics for formal evaluation.

3 Relevance Assessments

The documents in P_{30} and P'_X ($X \in \{50, 70, 90, 100\}$) were sorted using a simple method described below, and were presented to the relevance assessors through the *EPAN* [17] interface exactly in that order: For P_{30}, the documents were sorted first by the number of runs containing the document at or above rank 30 (the larger the better), and then by the sum of ranks of that document within

those runs (the smaller the better). Thus, if many runs contained a document within top 30, the document was presented to the assessors early. Moreover, if the ranks of this document were generally high, it was presented to the assessors even earlier. The documents in P'_X were sorted similarly, based on the number of runs containing the document at or above rank X. Note that this is in contrast to the TREC methodology which sorts pooled documents simply by document IDs. The assumptions behind our strategy are:

1. "Popular" documents (i.e., those retrieved at high ranks by many systems) are more likely to be relevant than others;
2. If there are more relevant documents near the top of the list of documents to be judged than near the bottom, then this makes it easier for the assessors to make judgments more efficiently and consistently than when relevant documents are randomly spread across the list.

Previous NTCIR tasks have used similar strategies [11]. An analysis of the IR4QA pools by Sakai and Kando supports the first of the above assumptions [24]. As for the second assumption, pools sorted by document IDs were not popular with the assessors at NTCIR-2. Although there is still room for debate regarding the above sorting technique, note that this does *not* affect which documents are judged: All documents in P_{30} (or P'_X) are judged, and no assessors are allowed to give up judgments halfway[1]. It only affects which documents are judged *first*.

The assessors manually assigned a relevance level to each pooled document shown on EPAN. The relevance levels are:

$L2$ 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.

Note that, in traditional IR evaluation, both judged nonrelevant ($L0$-relevant) documents and unjudged documents (those that never made it into the pool) are both treated as nonrelevant. We follow this strategy[2].

We released "qrels version 1" to the participants on September 1, 2008. Due to time constraints, we could not cover the depth-100 pool for many topics. The pools that the relevance assessors actually judged are indicated in bold in Tables 29-31 in the appendix. Tables 35-37 show the number of judged nonrelevant and relevant documents per topic for each language in qrels version 1[3].

[1]We regret to say that this policy was not followed strictly in practice: See Tables 29-31. We plan to fix these problems before releasing revised qrels files.

[2]A simple alternative would be to remove all unjudged documents from the original ranked list (e.g. [23]).

4 Evaluation Package and Metrics

4.1 The Package

We have developed a simple IR evaluation package for UNIX/Linux environments, available at `http://research.nii.ac.jp/ntcir/tools/ir4qa_eval-en` . It consists of a few Bourne shell scripts and a simple C program. Figure 1 outlines how the package works. For IR4QA evaluation, the following three shell scripts should be used:

IR4QA-splitqrels The input to this script is an IR4QA qrels file (i.e., ACLIA1-{CS, CT, JA}.qrels) whose format is shown in Figure 2. A qrels file contains a list of judged documents for each topic, together with the relevance level of each document[4]. The script creates a directory for each topic under current directory and stores a per-topic "rel" file under it. The rel file format is shown in Figure 3.

IR4QA-splitruns The input to this script is an IR4QA XML run file whose format is shown in Figure 4. The script just breaks the run file into per-topic ranked list of documents, called the "res" files, like the one shown in Figure 5, and stores them under the topic directories individually. For in-house experiments, researchers can create res files directly without creating a single XML run file: In this case `IR4QA-splitruns` is not required.

IR4QA-qeval This script reads the rel file and the res file and computes evaluation metrics for each topic, by calling the C program `q_eval`. The script creates a "qev" file which contains per-topic performance values. It also outputs performance values averaged across topics to standard output. As a by product, it creates a "lab" file under each topic directory, which indicates which documents in the "res" file are actually relevant.

For more details, please refer to the README file included in the package.

[3]The number of judged documents do not necessarily match the pool size shown in Tables 29-31 due to the existence of the aforementioned "pooled but unjudged" documents as well as runs that contained many illegal document IDs.

[4]Note that an IR4QA qrels file has only three fields, as opposed to previous TREC/NTCIR qrels files which had five fields. The abovementioned evaluation package includes a simple script which converts a five-field NTCIR qrels file into the new three-field format.

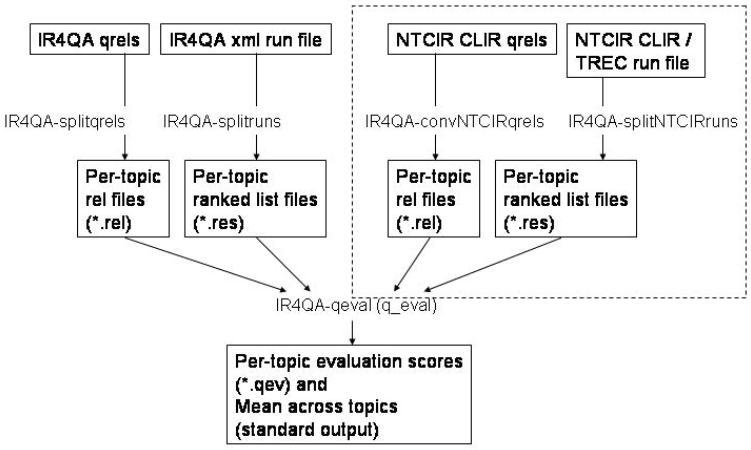

Figure 1. The IR4QA evaluation package.

4.2 The Metrics

Following the tradition at TREC, NTCIR and CLEF, we use *Average Precision* (AP) as our primary evaluation metric. However, as our relevance assessments are graded, we also use two well-studied metrics that can handle graded relevance: Q-measure (or simply Q) [21] and a version of normalised Discounted Cumulative Gain (nDCG) [9]. We define these three metrics formally below.

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, 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* at rank r is defined as $cg(r) = \sum_{i=1}^{r} g(r)$. Similarly, let $cg^*(r) = \sum_{i=1}^{r} g^*(r)$. Let β be a positive constant. Q is defined as:

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

Letting $\beta = 0$ reduces Q to AP, and using a large β makes Q more forgiving to relevant documents found near the bottom of the ranked list [21]. 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 [20].

Sakai and Robertson [25] have discussed how AP and Q can be interpreted from the viewpoint of a user population.

Let l be a document cut-off value. The version of nDCG we use is defined as:

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

The original nDCG as defined in [9] is known to be "buggy" [21]. The above version of nDCG, first used in [2] and sometimes referred to as the Microsoft version, is free from this bug[5]. Moreover, unlike the original nDCG, the choice of the logarithm base does not affect the Microsoft version. We let $l = 1000$ throughtout our analysis: That is, we use the entire document ranking to compute nDCG[6].

5 Evaluation Using Pseudo-Qrels

Prior to obtaining the real qrels, we constructed simple "pseudo-qrels" without using any manual relevance assessments and evaluated the runs using this data. The motivation was twofold:

[5] Another bug-free version is described in [10].

[6] In the aforementioned evaluation program `q_eval`, this metric is shown as "MSnDCG@1000" where MS stands for Microsoft.

```
      :
ACLIA1-JA-T1 JA-981113113 L0
ACLIA1-JA-T1 JA-981116067 L1
ACLIA1-JA-T1 JA-981119382 L0
ACLIA1-JA-T1 JA-981121145 L2
ACLIA1-JA-T1 JA-981123189 L0
      :
```

Figure 2. IR4QA qrels format: <topicID> <docID> <relevance level>.

```
      :
JA-981113113 L0
JA-981116067 L1
JA-981119382 L0
JA-981121145 L2
JA-981123189 L0
      :
```

Figure 3. Per-topic rel file format: <docID> <relevance level>.

```
<TOPIC_SET>
    <METADATA>
        <RUNID>CMUJAV-EN-JA-01-T</RUNID>
        <DESCRIPTION>Combined basic keyterm based query with PRF where
additional phrases (copula, aliase etc) are found using automatically
acquired lexico-semantic patterns</DESCRIPTION>
    </METADATA>
    <TOPIC ID="ACLIA1-JA-T1">
        <IR4QA_RESULT>
            <DOCUMENT SCORE="-6.3833" DOCID="JA-000420097" RANK="1"/>
            <DOCUMENT SCORE="-6.5087" DOCID="JA-011023090" RANK="2"/>
            <DOCUMENT SCORE="-6.5753" DOCID="JA-981116067" RANK="3"/>
                  :
            <DOCUMENT SCORE="-13.3379" DOCID="JA-980602337" RANK="1000"/>
        </IR4QA_RESULT>
    </TOPIC>
    <TOPIC ID="ACLIA1-JA-T2">
        <IR4QA_RESULT>
            <DOCUMENT SCORE="-11.2203" DOCID="JA-990601055" RANK="1"/>
            <DOCUMENT SCORE="-16.5825" DOCID="JA-000116056" RANK="2"/>
            <DOCUMENT SCORE="-16.8522" DOCID="JA-990319257" RANK="3"/>
                  :
            <DOCUMENT SCORE="-13.4261" DOCID="JA-981216060" RANK="1000"/>
        </IR4QA_RESULT>
    </TOPIC>
</TOPIC_SET>
```

Figure 4. IR4QA XML run file format.

```
JA-000420097
JA-011023090
JA-981116067
  :
JA-980602337
```

Figure 5. Per-topic res file format: <docID> (fully ordered by relevance score).

Table 3. Performances based on the *pseudo-qrels*: CS runs; 97 topics. Note that these are not the official system rankings.

run	Mean AP	run	Mean Q	run	Mean nDCG
OT-CS-CS-02-T	0.5199	OT-CS-CS-02-T	0.5673	OT-CS-CS-02-T	0.7589
CMUJAV-CS-CS-02-T	0.5189	CMUJAV-CS-CS-02-T	0.5647	CMUJAV-CS-CS-02-T	0.7519
CMUJAV-CS-CS-01-T	0.5084	CMUJAV-CS-CS-01-T	0.5549	CMUJAV-CS-CS-01-T	0.7463
MITEL-EN-CS-03-T	0.4822	MITEL-EN-CS-03-T	0.5266	OT-CS-CS-04-T	0.7256
MITEL-EN-CS-01-T	0.4728	OT-CS-CS-04-T	0.5203	MITEL-EN-CS-05-TD	0.7203
OT-CS-CS-04-T	0.4724	MITEL-EN-CS-01-T	0.5192	MITEL-EN-CS-03-T	0.7202
MITEL-EN-CS-05-TD	0.4702	MITEL-EN-CS-05-TD	0.5172	MITEL-EN-CS-01-T	0.7194
KECIR-CS-CS-02-DN	0.4701	MITEL-EN-CS-04-D	0.5105	MITEL-EN-CS-04-D	0.7129
MITEL-EN-CS-04-D	0.4643	KECIR-CS-CS-02-DN	0.5097	MITEL-EN-CS-02-T	0.7049
MITEL-EN-CS-02-T	0.4528	MITEL-EN-CS-02-T	0.4999	KECIR-CS-CS-02-DN	0.6989
KECIR-CS-CS-01-T	0.4441	KECIR-CS-CS-01-T	0.4856	OT-CS-CS-03-T	0.6857
KECIR-CS-CS-03-DN	0.4374	KECIR-CS-CS-03-DN	0.4738	KECIR-CS-CS-01-T	0.6740
CMUJAV-EN-CS-02-T	0.4302	CMUJAV-EN-CS-02-T	0.4733	OT-CS-CS-05-T	0.6737
CMUJAV-EN-CS-01-T	0.4255	CMUJAV-EN-CS-01-T	0.4712	CMUJAV-EN-CS-01-T	0.6710
OT-CS-CS-03-T	0.4130	OT-CS-CS-03-T	0.4639	CMUJAV-EN-CS-02-T	0.6697
OT-CS-CS-05-T	0.4068	OT-CS-CS-05-T	0.4596	KECIR-CS-CS-03-DN	0.6564
RALI-CS-CS-04-T	0.3976	HIT-EN-CS-01-DN	0.4417	RALI-CS-CS-04-T	0.6533
RALI-CS-CS-05-T	0.3975	RALI-CS-CS-05-T	0.4289	RALI-CS-CS-05-T	0.6532
RALI-CS-CS-02-T	0.3956	RALI-CS-CS-04-T	0.4286	RALI-CS-CS-01-T	0.6489
RALI-CS-CS-03-T	0.3950	RALI-CS-CS-02-T	0.4273	RALI-CS-CS-03-T	0.6473
HIT-EN-CS-01-DN	0.3948	RALI-CS-CS-03-T	0.4264	RALI-CS-CS-02-T	0.6454
RALI-CS-CS-01-T	0.3942	RALI-CS-CS-01-T	0.4257	HIT-EN-CS-01-DN	0.6435
WHUCC-CS-CS-02-T†	0.3862	WHUCC-CS-CS-02-T†	0.4188	HIT-EN-CS-02-T	0.6252
WHUCC-CS-CS-01-T†	0.3862	WHUCC-CS-CS-01-T†	0.4188	OT-CS-CS-01-T	0.6080
HIT-EN-CS-02-T	0.3702	HIT-EN-CS-02-T	0.4182	HIT-EN-CS-02-D	0.6037
HIT-EN-CS-02-D	0.3438	HIT-EN-CS-02-D	0.3883	WHUCC-CS-CS-02-T†	0.5963
RALI-EN-CS-04-T	0.3388	RALI-EN-CS-04-T	0.3674	WHUCC-CS-CS-01-T†	0.5963
RALI-EN-CS-05-T	0.3377	RALI-EN-CS-05-T	0.3664	HIT-EN-CS-02-DN	0.5889
NLPAI-CS-CS-02-T	0.3349	HIT-EN-CS-02-DN	0.3657	RALI-EN-CS-05-T	0.5835
RALI-EN-CS-02-T	0.3269	RALI-EN-CS-02-T	0.3565	RALI-EN-CS-04-T	0.5814
NLPAI-CS-CS-05-DN	0.3261	RALI-EN-CS-01-T	0.3547	CYUT-EN-CS-03-DN	0.5731
RALI-EN-CS-01-T	0.3257	CYUT-EN-CS-03-DN	0.3534	RALI-EN-CS-01-T	0.5698
HIT-EN-CS-02-DN	0.3210	NLPAI-CS-CS-02-T	0.3349	RALI-EN-CS-02-T	0.5663
CYUT-EN-CS-03-DN	0.3139	OT-CS-CS-01-T	0.3344	CYUT-EN-CS-02-D	0.5120
OT-CS-CS-01-T	0.3038	NLPAI-CS-CS-05-DN	0.3261	CYUT-EN-CS-01-T	0.5098
NLPAI-CS-CS-03-T	0.3010	NLPAI-CS-CS-03-T	0.3010	NLPAI-CS-CS-02-T	0.4743
NLPAI-CS-CS-01-T	0.2801	CYUT-EN-CS-02-D	0.2967	NLPAI-CS-CS-05-DN	0.4613
NLPAI-CS-CS-04-T	0.2711	CYUT-EN-CS-01-T	0.2939	NLPAI-CS-CS-03-T	0.4395
CYUT-EN-CS-02-D	0.2615	NLPAI-CS-CS-01-T	0.2801	NLPAI-CS-CS-01-T	0.4186
CYUT-EN-CS-01-T	0.2597	NLPAI-CS-CS-04-T	0.2711	NLPAI-CS-CS-04-T	0.4048

†These two runs are in fact identical: they contain the same ranked document lists for every topic.

Table 4. Performances based on the *pseudo-qrels*: CT runs; 95 topics. Note that these are not the official system rankings.

run	Mean AP	run	Mean Q	run	Mean nDCG
MITEL-CT-CT-02-T	0.5616	MITEL-CT-CT-02-T	0.6092	MITEL-CT-CT-02-T	0.7818
MITEL-CT-CT-01-T	0.5600	MITEL-CT-CT-01-T	0.6075	MITEL-CT-CT-01-T	0.7812
MITEL-CT-CT-03-D	0.5486	MITEL-CT-CT-03-D	0.5970	MITEL-CT-CT-03-D	0.7751
MITEL-CT-CT-04-T	0.5415	MITEL-CT-CT-04-T	0.5882	MITEL-CT-CT-04-T	0.7662
RALI-CT-CT-02-T	0.5017	OT-CT-CT-02-T	0.5412	RALI-CT-CT-02-T	0.7378
RALI-CT-CT-04-T	0.4999	RALI-CT-CT-02-T	0.5369	RALI-CT-CT-04-T	0.7373
OT-CT-CT-02-T	0.4945	RALI-CT-CT-04-T	0.5340	OT-CT-CT-02-T	0.7319
RALI-CT-CT-03-T	0.4837	OT-CT-CT-03-T	0.5220	RALI-CT-CT-03-T	0.7227
RALI-CT-CT-01-T	0.4762	RALI-CT-CT-03-T	0.5154	RALI-CT-CT-01-T	0.7204
OT-CT-CT-03-T	0.4734	RALI-CT-CT-01-T	0.5140	RALI-CT-CT-05-T	0.7187
RALI-CT-CT-05-T	0.4725	RALI-CT-CT-05-T	0.5095	OT-CT-CT-03-T	0.7178
OT-CT-CT-05-T	0.4546	OT-CT-CT-05-T	0.5022	OT-CT-CT-05-T	0.6990
OT-CT-CT-04-T	0.4470	OT-CT-CT-04-T	0.4922	OT-CT-CT-04-T	0.6983
NTUBROWS-CT-CT-01-T	0.4037	NTUBROWS-CT-CT-01-T	0.4366	OT-CT-CT-01-T	0.6462
OT-CT-CT-01-T	0.3647	OT-CT-CT-01-T	0.3930	NTUBROWS-CT-CT-01-T	0.6303
RALI-EN-CT-04-T	0.2854	RALI-EN-CT-02-T	0.3086	RALI-EN-CT-04-T	0.4771
RALI-EN-CT-02-T	0.2851	RALI-EN-CT-04-T	0.3085	RALI-EN-CT-02-T	0.4685
RALI-EN-CT-05-T	0.2730	RALI-EN-CT-05-T	0.2984	RALI-EN-CT-05-T	0.4635
RALI-EN-CT-01-T	0.2710	RALI-EN-CT-01-T	0.2976	RALI-EN-CT-01-T	0.4569
CYUT-EN-CT-03-DN	0.2087	CYUT-EN-CT-03-DN	0.2354	CYUT-EN-CT-03-DN	0.4181
CYUT-EN-CT-01-T	0.1988	CYUT-EN-CT-01-T	0.2253	CYUT-EN-CT-01-T	0.4107
CYUT-EN-CT-02-D	0.1907	CYUT-EN-CT-02-D	0.2162	CYUT-EN-CT-02-D	0.4030
NTUBROWS-CT-CT-03-T	0.1384	NTUBROWS-CT-CT-03-T	0.1718	NTUBROWS-CT-CT-03-T	0.3994
NTUBROWS-CT-CT-04-T	0.0925	NTUBROWS-CT-CT-04-T	0.1273	NTUBROWS-CT-CT-02-T∗	0.3676
NTUBROWS-CT-CT-02-T∗	0.0793	NTUBROWS-CT-CT-02-T∗	0.1158	NTUBROWS-CT-CT-04-T	0.3636
NTUBROWS-CT-CT-05-T∗	0.0612	NTUBROWS-CT-CT-05-T∗	0.0909	NTUBROWS-CT-CT-05-T∗	0.2933

∗The documentIDs in these two runs were all illegal: Their evaluation scores are computed here after a bug fix, even though the pools were created using the original runs.

Table 5. Performances based on the *pseudo-qrels*: JA runs; 98 topics. Note that these are not the official system rankings.

run	Mean AP	run	Mean Q	run	Mean nDCG
CMUJAV-JA-JA-04-T	0.6741	CMUJAV-JA-JA-04-T	0.7146	CMUJAV-JA-JA-04-T	0.8535
CMUJAV-JA-JA-01-T	0.6726	CMUJAV-JA-JA-01-T	0.7128	CMUJAV-JA-JA-01-T	0.8524
CMUJAV-JA-JA-05-T	0.6703	CMUJAV-JA-JA-05-T	0.7110	CMUJAV-JA-JA-05-T	0.8510
OT-JA-JA-02-T	0.6666	OT-JA-JA-02-T	0.7053	CMUJAV-JA-JA-03-T	0.8436
CMUJAV-JA-JA-03-T	0.6596	CMUJAV-JA-JA-03-T	0.7007	OT-JA-JA-02-T	0.8422
CMUJAV-JA-JA-02-T	0.6573	CMUJAV-JA-JA-02-T	0.6986	CMUJAV-JA-JA-02-T	0.8416
OT-JA-JA-04-T	0.6252	OT-JA-JA-04-T	0.6667	OT-JA-JA-04-T	0.8201
OT-JA-JA-05-T	0.5009	OT-JA-JA-05-T	0.5471	OT-JA-JA-05-T	0.7353
BRKLY-JA-JA-03-T	0.4649	BRKLY-JA-JA-03-T	0.5084	BRKLY-JA-JA-03-T	0.6987
BRKLY-JA-JA-02-DN	0.4582	BRKLY-JA-JA-02-DN	0.4997	BRKLY-JA-JA-02-DN	0.6894
BRKLY-JA-JA-01-DN	0.4447	BRKLY-JA-JA-01-DN	0.4871	BRKLY-JA-JA-01-DN	0.6836
CMUJAV-EN-JA-01-T	0.4414	BRKLY-JA-JA-02-T	0.4754	BRKLY-JA-JA-02-T	0.6798
CMUJAV-EN-JA-04-T	0.4384	CMUJAV-EN-JA-01-T	0.4738	OT-JA-JA-01-T	0.6790
CMUJAV-EN-JA-05-T	0.4371	CMUJAV-EN-JA-04-T	0.4717	OT-JA-JA-03-T	0.6391
BRKLY-JA-JA-02-T	0.4334	CMUJAV-EN-JA-05-T	0.4703	CMUJAV-EN-JA-01-T	0.6208
CMUJAV-EN-JA-03-T	0.4287	CMUJAV-EN-JA-03-T	0.4624	CMUJAV-EN-JA-04-T	0.6201
CMUJAV-EN-JA-02-T	0.4199	CMUJAV-EN-JA-02-T	0.4541	CMUJAV-EN-JA-05-T	0.6196
OT-JA-JA-01-T	0.4044	OT-JA-JA-01-T	0.4322	CMUJAV-EN-JA-03-T	0.6122
OT-JA-JA-03-T	0.3886	OT-JA-JA-03-T	0.4308	CMUJAV-EN-JA-02-T	0.6065
CYUT-EN-JA-01-T	0.1733	CYUT-EN-JA-01-T	0.1970	CYUT-EN-JA-03-DN	0.3615
CYUT-EN-JA-03-DN	0.1712	CYUT-EN-JA-03-DN	0.1927	CYUT-EN-JA-01-T	0.3610
CYUT-EN-JA-02-D	0.1528	CYUT-EN-JA-02-D	0.1723	CYUT-EN-JA-02-D	0.3326
TA-EN-JA-02-D	0.0062	TA-EN-JA-02-D	0.0073	TA-EN-JA-03-T	0.0264
TA-EN-JA-01-D	0.0060	TA-EN-JA-01-D	0.0071	TA-EN-JA-02-D	0.0205
TA-EN-JA-03-T	0.0050	TA-EN-JA-03-T	0.0067	TA-EN-JA-01-D	0.0167

1. We wanted to check that our evaluation package actually works!

2. Saving human resources for relevance assessments is always important. The extreme case is ranking system without relevant assessments [1, 27]. We wanted to investigate how such a "lazy" method correlates with traditional evaluation using real qrels.

Due to time constraints, we used an extremely simple pseudo-qrels creation method:

1. For each topic, take the depth-30 pool P_{30}, and sort the documents as described in Section 3;

2. Take the top 10 documents in the sorted list and treat all as $L1$-relevant.

Thus, according to our pseudo-qrels, $R = R(L1) = 10$ for every topic, and the relevance assessments are binary. Tables 3-5 show the Mean AP, Q and nDCG values computed based on the pseudo-qrels for the CS, CT and JA runs. Note that since the pseudo-qrels were constructed based on "majority votes", the systems are ranked by "popularity" (how closely they resemble the other systems) rather than effectiveness [1]. We shall discuss the correlation between the evaluation based on pseudo-qrels and that based on real qrels in Section 6.4.

We have also ranked the *topics* by performance averaged across runs, using the pseudo-qrels. The results are shown in Tables 32-34 in the Appendix. Following Mizzaro and Robertson [18], the AP for a particular topic averaged acrosss runs will be referred to as "Average AP" (as opposed to Mean AP), and so on.

6 Official IR4QA Results

6.1 Overview

We now present the official results based on qrels version 1. Tables 6-8 summarise the official IR4QA results by sorting the CS, CT and JA runs by Mean AP/Q/nDCG over the topic set, respectively. The tables show, for example, that:

- Some monolingual runs from OT appear to be the best among all CS runs;
- Some runs from MITEL appear to be the best among the EN-CS runs: While Mean AP and Q prefer MITEL-EN-CS-03-T, Mean nDCG prefers MITEL-EN-CS-05-TD. These runs do very well even when compared to the monolingual runs;
- Some monolingual runs from MITEL appear to be the best among all CT runs: While Mean Q and nDCG prefer MITEL-CT-CT-02-T over MITEL-CT-CT-03-D, they are equally effective according to Mean AP;
- Some runs from RALI appear to be the best among the EN-CT runs: While Mean AP and Q prefer RALI-EN-CT-05-T, Mean nDCG prefers RALI-EN-CT-04-T;
- Some monolingual runs from OT appear to be the best among all JA runs;
- Some runs from CMUJAV appear to be the best among the EN-JA runs.

However, the performance values in these tables are unnaturally high, which suggests that qrels version 1 may be very incomplete: That is, there may be many relevant documents in the document collections that we have not identified. Hence the IR4QA test collections may not be reusable at this stage: That is, it may not be suitable for evaluating systems that did not contribute to the pools. We plan to investigate this issue further.

Based on statistical significance tests, we further summarise the official results as follows: Tables 9-11 show the "best" T-run (i.e., run that used only the TITLE field of each topic as the input to the IR system) from each team according to Mean AP/Q/nDCG, sorted by performance. For each adjacent pair of runs shown in this table, we conducted a two-sided, paired bootstrap test using 1000 bootstrap samples of topics. [4, 22]. For example, the "Mean AP" column of Table 9 shows that HIT-EN-CS-02-T significantly outperforms KECIR-CS-CS-01-T at $\alpha = 0.05$, KECIR-CS-CS-01-T significantly outperforms RALI-CS-CS-05-T at $\alpha = 0.05$, RALI-CS-CS-05-T significantly outperforms WHUCC-CS-CS-01-T at $\alpha = 0.01$, and so on. Although pairwise statistical significance is not transitive, We can informally see, for example, that:

- In Table 9, OT, MITEL, CMUJAV and HIT are probably the best CS teams (Note that the MITEL run is a crosslingual one), since HIT-EN-CS-02-T significanty outperforms KECIR-CS-CS-01-T according to Mean AP and Q;
- In Table 10, MITEL and OT are probably the best CT teams since OT-CT-CT-04-T significantly outperforms RALI-CT-CT-05-T according to all three evaluation metrics;
- In Table 11, OT is probably the best JA team, since OT-JA-JA-04-T significantly outperforms BRKLY-JA-JA-01-DN according to all three evaluation metrics.

Interestingly, in Table 9, KECIR significantly outperforms RALI in terms of Mean AP, while RALI significantly outperforms KECIR in terms of Mean nDCG. Q-measure is undecided[7]. These differences arise from the different properties of the metrics: AP cannot handle graded relevance and is unforgiving for low-recall systems; Q can handle graded relevance and is unforgiving for low-recall systems; and nDCG can handle graded relevance and is relatively forgiving for low-recall systems, by definition.

We have also ranked the *topics* by performance averaged across runs, using the real qrels. This reflects the "difficulty" of topics. The results are shown in Tables 38-40 in the Appendix. For example, in Table 38, ACLIA1-CS-T80 is the "easiest" topic according to Average AP. In contrast, ACLIA1-CS-T370 is the easiest according to Average Q, while ACLIA1-CS-T340 is the easiest according to Average nDCG. The per-topic Average AP, Q and nDCG values, without the sort, are visualised in Figures 6-8.

[7]Note that two different RALI runs are involved here: Mean AP and Mean Q chose RALI-CS-CS-05-T as the best run from this team, while Mean nDCG chose RALI-CS-CS-04-T.

Table 6. Performances based on the real qrels: CS runs; 97 topics.

run	Mean AP	run	Mean Q	run	Mean nDCG
OT-CS-CS-04-T	0.6337	OT-CS-CS-04-T	0.6490	OT-CS-CS-04-T	0.8270
OT-CS-CS-02-T	0.6295	OT-CS-CS-02-T	0.6411	OT-CS-CS-02-T	0.8139
MITEL-EN-CS-03-T	0.5959	MITEL-EN-CS-03-T	0.6124	MITEL-EN-CS-05-TD	0.8003
CMUJAV-CS-CS-02-T	0.5930	MITEL-EN-CS-05-TD	0.6058	CMUJAV-CS-CS-02-T	0.7951
MITEL-EN-CS-05-TD	0.5898	CMUJAV-CS-CS-02-T	0.6055	MITEL-EN-CS-01-T	0.7949
CMUJAV-CS-CS-01-T	0.5897	CMUJAV-CS-CS-01-T	0.6028	MITEL-EN-CS-03-T	0.7947
MITEL-EN-CS-01-T	0.5849	MITEL-EN-CS-01-T	0.6005	CMUJAV-CS-CS-01-T	0.7940
MITEL-EN-CS-04-D	0.5789	MITEL-EN-CS-04-D	0.5950	MITEL-EN-CS-04-D	0.7907
MITEL-EN-CS-02-T	0.5693	OT-CS-CS-03-T	0.5859	MITEL-EN-CS-02-T	0.7847
HIT-EN-CS-01-DN	0.5690	MITEL-EN-CS-02-T	0.5858	OT-CS-CS-03-T	0.7831
OT-CS-CS-03-T	0.5659	HIT-EN-CS-01-DN	0.5840	OT-CS-CS-05-T	0.7771
OT-CS-CS-05-T	0.5645	OT-CS-CS-05-T	0.5834	HIT-EN-CS-01-DN	0.7560
HIT-EN-CS-02-T	0.5585	HIT-EN-CS-02-T	0.5745	HIT-EN-CS-02-T	0.7480
CMUJAV-EN-CS-01-T	0.5457	CMUJAV-EN-CS-01-T	0.5558	CMUJAV-EN-CS-01-T	0.7397
CMUJAV-EN-CS-02-T	0.5266	CMUJAV-EN-CS-02-T	0.5371	RALI-CS-CS-04-T	0.7276
HIT-EN-CS-02-D	0.5124	HIT-EN-CS-02-D	0.5317	CMUJAV-EN-CS-02-T	0.7254
KECIR-CS-CS-01-T	0.5013	KECIR-CS-CS-01-T	0.4842	RALI-CS-CS-03-T	0.7251
KECIR-CS-CS-02-DN	0.4864	HIT-EN-CS-02-DN	0.4827	RALI-CS-CS-05-T	0.7242
RALI-CS-CS-05-T	0.4684	RALI-CS-CS-05-T	0.4812	RALI-CS-CS-01-T	0.7192
RALI-CS-CS-01-T	0.4671	RALI-CS-CS-01-T	0.4796	HIT-EN-CS-02-D	0.7174
RALI-CS-CS-03-T	0.4657	RALI-CS-CS-03-T	0.4790	RALI-CS-CS-02-T	0.7160
HIT-EN-CS-02-DN	0.4634	RALI-CS-CS-04-T	0.4745	OT-CS-CS-01-T	0.7075
RALI-CS-CS-02-T	0.4630	RALI-CS-CS-02-T	0.4731	HIT-EN-CS-02-DN	0.6910
RALI-CS-CS-04-T	0.4622	KECIR-CS-CS-02-DN	0.4645	RALI-EN-CS-04-T	0.6701
KECIR-CS-CS-03-DN	0.4429	CYUT-EN-CS-03-DN	0.4386	RALI-EN-CS-05-T	0.6599
CYUT-EN-CS-03-DN	0.4238	KECIR-CS-CS-03-DN	0.4292	RALI-EN-CS-02-T	0.6586
RALI-EN-CS-04-T	0.4033	OT-CS-CS-01-T	0.4243	CYUT-EN-CS-03-DN	0.6578
RALI-EN-CS-02-T	0.4025	RALI-EN-CS-04-T	0.4191	KECIR-CS-CS-01-T	0.6562
RALI-EN-CS-05-T	0.4013	RALI-EN-CS-05-T	0.4181	RALI-EN-CS-01-T	0.6553
RALI-EN-CS-01-T	0.3992	RALI-EN-CS-02-T	0.4173	KECIR-CS-CS-02-DN	0.6306
WHUCC-CS-CS-02-T†	0.3806	RALI-EN-CS-01-T	0.4161	CYUT-EN-CS-01-T	0.6115
WHUCC-CS-CS-01-T†	0.3806	CYUT-EN-CS-01-T	0.3936	CYUT-EN-CS-02-D	0.6057
CYUT-EN-CS-01-T	0.3781	CYUT-EN-CS-02-D	0.3880	KECIR-CS-CS-03-DN	0.6011
CYUT-EN-CS-02-D	0.3726	WHUCC-CS-CS-02-T†	0.3626	WHUCC-CS-CS-02-T†	0.5169
OT-CS-CS-01-T	0.3702	WHUCC-CS-CS-01-T†	0.3626	WHUCC-CS-CS-01-T†	0.5169
NLPAI-CS-CS-02-T	0.1319	NLPAI-CS-CS-02-T	0.1227	NLPAI-CS-CS-02-T	0.2536
NLPAI-CS-CS-05-DN	0.1302	NLPAI-CS-CS-05-DN	0.1211	NLPAI-CS-CS-05-DN	0.2493
NLPAI-CS-CS-01-T	0.1198	NLPAI-CS-CS-01-T	0.1099	NLPAI-CS-CS-01-T	0.2383
NLPAI-CS-CS-03-T	0.1170	NLPAI-CS-CS-03-T	0.1074	NLPAI-CS-CS-03-T	0.2297
NLPAI-CS-CS-04-T	0.1117	NLPAI-CS-CS-04-T	0.1014	NLPAI-CS-CS-04-T	0.2204

†These two runs are in fact identical: they contain the same ranked document lists for every topic.

Table 7. Performances based on the real qrels: CT runs; 95 topics.

run	Mean AP	run	Mean Q	run	Mean nDCG
MITEL-CT-CT-03-D	0.5839	MITEL-CT-CT-02-T	0.6018	MITEL-CT-CT-02-T	0.7873
MITEL-CT-CT-02-T	0.5839	MITEL-CT-CT-03-D	0.6013	MITEL-CT-CT-03-D	0.7869
MITEL-CT-CT-01-T	0.5791	MITEL-CT-CT-01-T	0.5963	MITEL-CT-CT-01-T	0.7835
MITEL-CT-CT-04-T	0.5645	MITEL-CT-CT-04-T	0.5783	OT-CT-CT-04-T	0.7656
OT-CT-CT-04-T	0.5521	OT-CT-CT-04-T	0.5724	MITEL-CT-CT-04-T	0.7648
OT-CT-CT-02-T	0.5111	OT-CT-CT-02-T	0.5339	OT-CT-CT-02-T	0.7432
OT-CT-CT-03-T	0.5015	OT-CT-CT-03-T	0.5224	OT-CT-CT-03-T	0.7332
OT-CT-CT-05-T	0.4907	OT-CT-CT-05-T	0.5136	OT-CT-CT-05-T	0.7268
RALI-CT-CT-05-T	0.3952	RALI-CT-CT-05-T	0.4096	OT-CT-CT-01-T	0.6594
RALI-CT-CT-01-T	0.3921	RALI-CT-CT-01-T	0.4074	RALI-CT-CT-03-T	0.6559
RALI-CT-CT-04-T	0.3753	RALI-CT-CT-03-T	0.3922	RALI-CT-CT-04-T	0.6525
RALI-CT-CT-02-T	0.3745	RALI-CT-CT-04-T	0.3916	RALI-CT-CT-05-T	0.6516
RALI-CT-CT-03-T	0.3741	RALI-CT-CT-02-T	0.3892	RALI-CT-CT-01-T	0.6473
NTUBROWS-CT-CT-01-T	0.3587	NTUBROWS-CT-CT-01-T	0.3780	RALI-CT-CT-02-T	0.6400
OT-CT-CT-01-T	0.3228	OT-CT-CT-01-T	0.3726	NTUBROWS-CT-CT-01-T	0.5932
RALI-EN-CT-05-T	0.2723	RALI-EN-CT-05-T	0.2868	NTUBROWS-CT-CT-02-T∗	0.4993
RALI-EN-CT-01-T	0.2723	RALI-EN-CT-01-T	0.2863	NTUBROWS-CT-CT-03-T	0.4853
CYUT-EN-CT-01-T	0.2590	CYUT-EN-CT-01-T	0.2747	RALI-EN-CT-04-T	0.4845
RALI-EN-CT-04-T	0.2574	RALI-EN-CT-04-T	0.2737	RALI-EN-CT-05-T	0.4767
RALI-EN-CT-02-T	0.2572	RALI-EN-CT-02-T	0.2715	CYUT-EN-CT-01-T	0.4752
CYUT-EN-CT-03-DN	0.2516	CYUT-EN-CT-03-DN	0.2648	RALI-EN-CT-01-T	0.4750
CYUT-EN-CT-02-D	0.2458	CYUT-EN-CT-02-D	0.2620	RALI-EN-CT-02-T	0.4731
NTUBROWS-CT-CT-03-T	0.2129	NTUBROWS-CT-CT-02-T∗	0.2498	NTUBROWS-CT-CT-04-T	0.4640
NTUBROWS-CT-CT-02-T∗	0.2008	NTUBROWS-CT-CT-03-T	0.2495	CYUT-EN-CT-03-DN	0.4638
NTUBROWS-CT-CT-04-T	0.1935	NTUBROWS-CT-CT-04-T	0.2303	CYUT-EN-CT-02-D	0.4612
NTUBROWS-CT-CT-05-T∗	0.1653	NTUBROWS-CT-CT-05-T∗	0.2026	NTUBROWS-CT-CT-05-T∗	0.4041

∗The documentIDs in these two runs were all illegal: Their evaluation scores are computed here after a bug fix, even though the pools were created using the original runs.

Table 8. Performances based on the real qrels: JA runs; 98 topics.

run	Mean AP	run	Mean Q	run	Mean nDCG
OT-JA-JA-04-T	0.6979	OT-JA-JA-04-T	0.7090	OT-JA-JA-04-T	0.8650
OT-JA-JA-02-T	0.6698	OT-JA-JA-02-T	0.6808	OT-JA-JA-02-T	0.8473
BRKLY-JA-JA-01-DN	0.6278	BRKLY-JA-JA-01-DN	0.6417	BRKLY-JA-JA-01-DN	0.8168
CMUJAV-JA-JA-01-T	0.5932	CMUJAV-JA-JA-01-T	0.5996	CMUJAV-JA-JA-01-T	0.7832
CMUJAV-JA-JA-03-T	0.5885	BRKLY-JA-JA-02-T	0.5996	BRKLY-JA-JA-02-T	0.7831
CMUJAV-JA-JA-04-T	0.5845	CMUJAV-JA-JA-03-T	0.5953	OT-JA-JA-05-T	0.7818
BRKLY-JA-JA-02-T	0.5838	CMUJAV-JA-JA-04-T	0.5911	CMUJAV-JA-JA-03-T	0.7801
CMUJAV-JA-JA-02-T	0.5790	CMUJAV-JA-JA-02-T	0.5875	CMUJAV-JA-JA-04-T	0.7781
CMUJAV-JA-JA-05-T	0.5784	CMUJAV-JA-JA-05-T	0.5852	BRKLY-JA-JA-02-DN	0.7767
BRKLY-JA-JA-02-DN	0.5767	BRKLY-JA-JA-02-DN	0.5849	CMUJAV-JA-JA-02-T	0.7743
OT-JA-JA-05-T	0.5659	OT-JA-JA-05-T	0.5836	CMUJAV-JA-JA-05-T	0.7723
BRKLY-JA-JA-03-T	0.5407	BRKLY-JA-JA-03-T	0.5509	BRKLY-JA-JA-03-T	0.7475
CMUJAV-EN-JA-01-T	0.4264	OT-JA-JA-03-T	0.4481	OT-JA-JA-01-T	0.7157
OT-JA-JA-03-T	0.4254	OT-JA-JA-01-T	0.4376	OT-JA-JA-03-T	0.6666
CMUJAV-EN-JA-03-T	0.4249	CMUJAV-EN-JA-01-T	0.4344	CMUJAV-EN-JA-01-T	0.6025
CMUJAV-EN-JA-04-T	0.4229	CMUJAV-EN-JA-03-T	0.4324	CMUJAV-EN-JA-03-T	0.6010
CMUJAV-EN-JA-02-T	0.4192	CMUJAV-EN-JA-04-T	0.4306	CMUJAV-EN-JA-04-T	0.5996
CMUJAV-EN-JA-05-T	0.4187	CMUJAV-EN-JA-05-T	0.4265	CMUJAV-EN-JA-05-T	0.5971
OT-JA-JA-01-T	0.3893	CMUJAV-EN-JA-02-T	0.4265	CMUJAV-EN-JA-02-T	0.5958
CYUT-EN-JA-03-DN	0.2568	CYUT-EN-JA-03-DN	0.2545	CYUT-EN-JA-03-DN	0.4366
CYUT-EN-JA-01-T	0.2543	CYUT-EN-JA-01-T	0.2528	CYUT-EN-JA-01-T	0.4252
CYUT-EN-JA-02-D	0.2294	CYUT-EN-JA-02-D	0.2300	CYUT-EN-JA-02-D	0.4124
TA-EN-JA-02-D	0.0141	TA-EN-JA-03-T	0.0155	TA-EN-JA-03-T	0.0446
TA-EN-JA-03-T	0.0127	TA-EN-JA-02-D	0.0155	TA-EN-JA-02-D	0.0337
TA-EN-JA-01-D	0.0115	TA-EN-JA-01-D	0.0119	TA-EN-JA-01-D	0.0268

Table 9. The best T-run from each CS team: "$*$" 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.

run	Mean AP	run	Mean Q	run	Mean nDCG
OT-CS-CS-04-T	0.6337	OT-CS-CS-04-T	0.6490	OT-CS-CS-04-T	0.8270$*$
MITEL-EN-CS-03-T	0.5959	MITEL-EN-CS-03-T	0.6124	CMUJAV-CS-CS-02-T	0.7951
CMUJAV-CS-CS-02-T	0.5930	CMUJAV-CS-CS-02-T	0.6055	MITEL-EN-CS-01-T	0.7949
HIT-EN-CS-02-T	0.5585$*$	HIT-EN-CS-02-T	0.5745$**$	HIT-EN-CS-02-T	0.7480
KECIR-CS-CS-01-T	0.5013$*$	KECIR-CS-CS-01-T	0.4842	RALI-CS-CS-04-T	0.7276$**$
RALI-CS-CS-05-T	0.4684$**$	RALI-CS-CS-05-T	0.4812$**$	KECIR-CS-CS-01-T	0.6562
WHUCC-CS-CS-01-T	0.3806	CYUT-EN-CS-01-T	0.3936	CYUT-EN-CS-01-T	0.6115$**$
CYUT-EN-CS-01-T	0.3781$**$	WHUCC-CS-CS-01-T	0.3626$**$	WHUCC-CS-CS-01-T	0.5169$**$
NLPAI-CS-CS-02-T	0.1319	NLPAI-CS-CS-02-T	0.1227	NLPAI-CS-CS-02-T	0.2536

Table 10. The best T-run from each CT team: "$*$" 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.

run	Mean AP	run	Mean Q	run	Mean nDCG
MITEL-CT-CT-02-T	0.5839	MITEL-CT-CT-02-T	0.6018	MITEL-CT-CT-02-T	0.7873
OT-CT-CT-04-T	0.5521$**$	OT-CT-CT-04-T	0.5724$**$	OT-CT-CT-04-T	0.7656$**$
RALI-CT-CT-05-T	0.3952	RALI-CT-CT-05-T	0.4096	RALI-CT-CT-03-T	0.6559$**$
NTUBROWS-CT-CT-01-T	0.3587$**$	NTUBROWS-CT-CT-01-T	0.3780$**$	NTUBROWS-CT-CT-01-T	0.5932$**$
CYUT-EN-CT-01-T	0.2590	CYUT-EN-CT-01-T	0.2747	CYUT-EN-CT-01-T	0.4752

Table 11. The best T-run from each JA team: "$*$" 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.

run	Mean AP	run	Mean Q	run	Mean nDCG
OT-JA-JA-04-T	0.6979$**$	OT-JA-JA-04-T	0.7090$**$	OT-JA-JA-04-T	0.8650$**$
CMUJAV-JA-JA-01-T	0.5932	CMUJAV-JA-JA-01-T	0.5996	CMUJAV-JA-JA-01-T	0.7832
BRKLY-JA-JA-02-T	0.5838$**$	BRKLY-JA-JA-02-T	0.5996$**$	BRKLY-JA-JA-02-T	0.7831$**$
CYUT-EN-JA-01-T	0.2543$**$	CYUT-EN-JA-01-T	0.2528$**$	CYUT-EN-JA-01-T	0.4252$**$
TA-EN-JA-03-T	0.0127	TA-EN-JA-03-T	0.0155	TA-EN-JA-03-T	0.0446

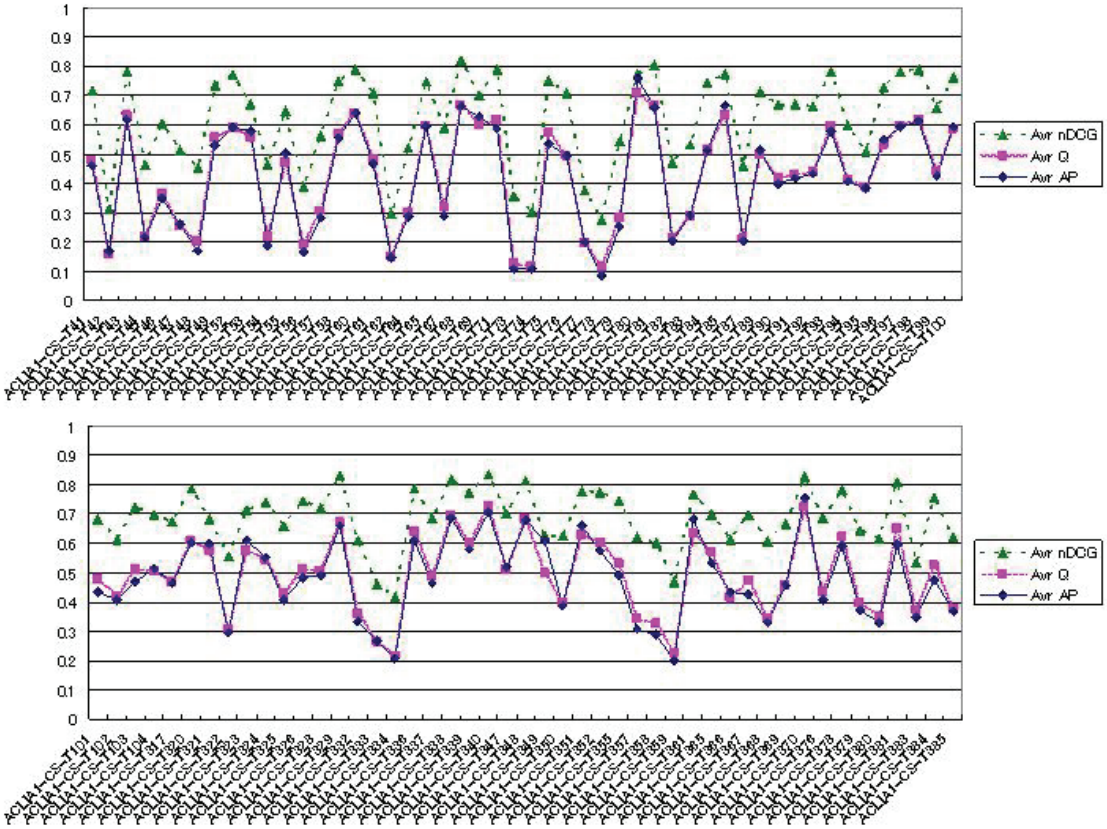

Figure 6. Per-topic Average AP, Q and nDCG values: CS topics.

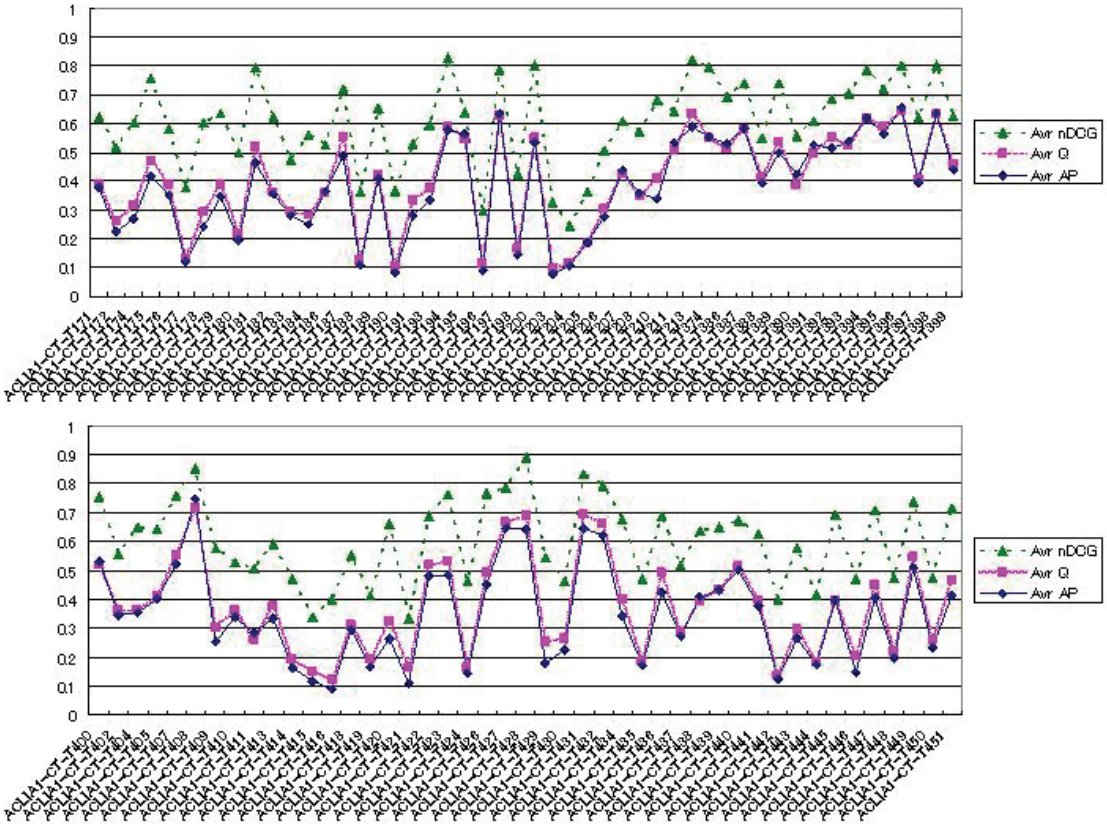

Figure 7. Per-topic Average AP, Q and nDCG values: CT topics.

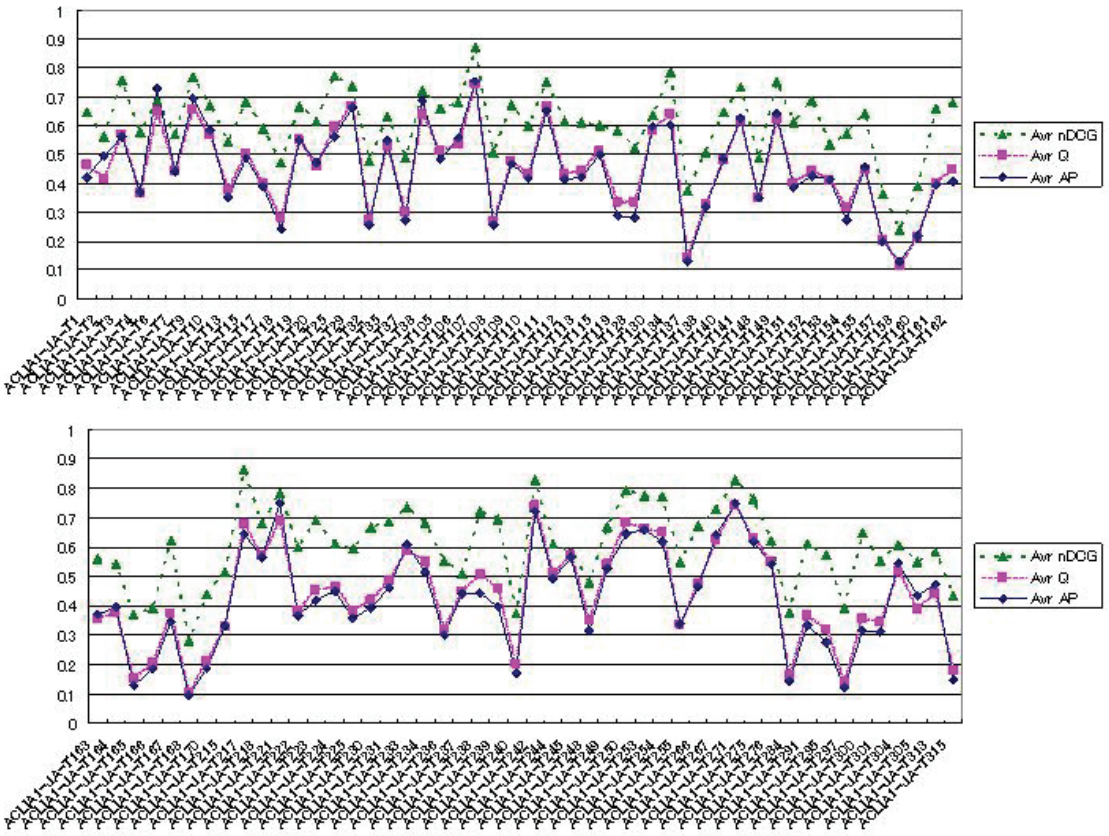

Figure 8. Per-topic Average AP, Q and nDCG values: JA topics.

Let us also look at the *coverage* of relevant documents for each run and for each 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. The coverage of relevant documents by s for this particular topic is defined as:

$$cvr(s, REL) = D(s) \cap REL \ .$$

Moreover, let $S(t)$ denote the set of runs submitted by team t. The coverage of relevant documents by t for this particular topic is defined as:

$$CVR(t, REL) = \cup_{s \in S(t)} cvr(s, REL) \ .$$

Note that this is closely related to recall.

Tables 12-14 show $cvr(s, REL)$ summed across topics for each run s. Similarly, Tables 15-17 show $CVR(t, REL)$ summed across topics for each team t. It can be observed that some runs from MITEL are good at retrieving many relevant documents for CS and CT (Tables 12 and 13), while the same goes for OT for JA (Table 14). The coverage of *teams* shows different results, because different teams submitted different number of runs with different document overlaps within each team. For example, Table 16 shows that RALI achieved high coverage as a team for CT, but this may be primarily because they submitted as many as nine runs (5 CT-CT plus 4 EN-CT): See Table 2.

We also provide system and team rankings according to the number of *unique relevant documents found*. This should reflect the "novelty" of systems/teams. For a particular topic, the number of unique relevant documents found by run $s \in S(t)$ is given by:

$$ur(s) = cvr(s, REL) - \cup_{t' = t} CVR(t', REL) \ . \quad (4)$$

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:

$$UR(t) = CVR(t, REL) - \cup_{t' = t} CVR(t', REL) \ . \quad (5)$$

Tables 18-20 show the system rankings according to $ur(s)$ summed across topics. Similarly, Tables 21-23 show the team rankings according to $UR(t)$ summed across topics. It can be observed that none of the runs is extremely "novel", since these numbers are very small compared to the total number of relevant documents across topics (See Tables 35-37). RALI may be more novel than other CS and CT teams while OT may be more novel than other JA teams: These teams are valuable for making the relevance assessments less incomplete. However, a closer look reveals that these systems/teams are "novel" for a very small

Table 12. Coverage of relevant documents summed across 97 topics: CS runs.

run name	covered docs.
MITEL-EN-CS-05-TD	8514
MITEL-EN-CS-03-T	8441
OT-CS-CS-04-T	8438
MITEL-EN-CS-01-T	8403
MITEL-EN-CS-04-D	8384
MITEL-EN-CS-02-T	8364
OT-CS-CS-02-T	8182
OT-CS-CS-03-T	8144
HIT-EN-CS-02-D	8086
OT-CS-CS-05-T	8082
OT-CS-CS-01-T	8072
CMUJAV-CS-CS-01-T	8031
HIT-EN-CS-01-DN	8027
CMUJAV-CS-CS-02-T	8024
HIT-EN-CS-02-T	7969
RALI-CS-CS-04-T	7861
RALI-CS-CS-03-T	7803
CMUJAV-EN-CS-01-T	7718
RALI-CS-CS-05-T	7700
RALI-CS-CS-02-T	7644
RALI-CS-CS-01-T	7644
HIT-EN-CS-02-DN	7491
CMUJAV-EN-CS-02-T	7426
RALI-EN-CS-04-T	7312
RALI-EN-CS-05-T	7078
RALI-EN-CS-02-T	7076
CYUT-EN-CS-03-DN	7007
RALI-EN-CS-01-T	7002
CYUT-EN-CS-01-T	6450
CYUT-EN-CS-02-D	6364
KECIR-CS-CS-01-T	5397
KECIR-CS-CS-02-DN	4378
KECIR-CS-CS-03-DN	4088
WHUCC-CS-CS-02-T	3101
WHUCC-CS-CS-01-T	3101
NLPAI-CS-CS-02-T	772
NLPAI-CS-CS-05-DN	750
NLPAI-CS-CS-01-T	739
NLPAI-CS-CS-03-T	700
NLPAI-CS-CS-04-T	677

Table 13. Coverage of relevant documents summed across topics: CT runs.

run name	covered docs.
MITEL-CT-CT-03-D	5025
MITEL-CT-CT-02-T	5014
MITEL-CT-CT-01-T	5002
OT-CT-CT-04-T	4983
OT-CT-CT-02-T	4890
MITEL-CT-CT-04-T	4882
OT-CT-CT-03-T	4833
OT-CT-CT-05-T	4819
OT-CT-CT-01-T	4817
RALI-CT-CT-04-T	4670
RALI-CT-CT-03-T	4630
RALI-CT-CT-05-T	4589
RALI-CT-CT-01-T	4552
RALI-CT-CT-02-T	4513
NTUBROWS-CT-CT-02-T$*$	4301
NTUBROWS-CT-CT-03-T	4051
NTUBROWS-CT-CT-01-T	4051
NTUBROWS-CT-CT-04-T	4025
RALI-EN-CT-04-T	3641
RALI-EN-CT-05-T	3517
RALI-EN-CT-02-T	3486
RALI-EN-CT-01-T	3467
NTUBROWS-CT-CT-05-T$*$	3338
CYUT-EN-CT-01-T	3311
CYUT-EN-CT-02-D	3164
CYUT-EN-CT-03-DN	3162

$*$The documentIDs in these two runs were all illegal: Their coverages are computed here after a bug fix, even though the pools were created using the original runs.

Table 14. Coverage of relevant documents summed across topics: JA runs.

run name	covered docs.
OT-JA-JA-04-T	8096
OT-JA-JA-02-T	8041
BRKLY-JA-JA-01-DN	7965
BRKLY-JA-JA-02-T	7817
OT-JA-JA-05-T	7674
OT-JA-JA-01-T	7668
BRKLY-JA-JA-02-DN	7553
BRKLY-JA-JA-03-T	7401
CMUJAV-JA-JA-01-T	7321
CMUJAV-JA-JA-03-T	7277
CMUJAV-JA-JA-02-T	7266
CMUJAV-JA-JA-04-T	7264
CMUJAV-JA-JA-05-T	7254
OT-JA-JA-03-T	6725
CMUJAV-EN-JA-01-T	6004
CMUJAV-EN-JA-03-T	5991
CMUJAV-EN-JA-02-T	5976
CMUJAV-EN-JA-05-T	5975
CMUJAV-EN-JA-04-T	5964
CYUT-EN-JA-01-T	4765
CYUT-EN-JA-03-DN	4721
CYUT-EN-JA-02-D	4587
TA-EN-JA-03-T	441
TA-EN-JA-02-D	369
TA-EN-JA-01-D	215

Table 15. Coverage of relevant documents summed across topics: CS-teams.

team name	covered docs.
OT	8888
MITEL	8765
RALI	8742
CMUJAV	8598
HIT	8285
CYUT	7316
KECIR	6651
WHUCC	3101
NLPAI	1222

Table 16. Coverage of relevant documents summed across topics: CT-teams.

team name	covered docs.
RALI	5132
OT	5122
MITEL	5095
NTUBROWS	4879
CYUT	3543

Table 17. Coverage of relevant documents summed across topics: JA-teams.

team name	covered docs.
OT	8241
BRKLY	8197
CMUJAV	7558
CYUT	5511
TA	494

Table 18. Unique relevant documents summed across topics: CS runs.

run name	unique relevant
RALI-EN-CS-04-T	63
RALI-CS-CS-04-T	62
RALI-CS-CS-02-T	62
RALI-EN-CS-02-T	60
RALI-CS-CS-03-T	59
RALI-EN-CS-01-T	58
RALI-EN-CS-05-T	57
RALI-CS-CS-05-T	56
RALI-CS-CS-01-T	56
OT-CS-CS-01-T	18
OT-CS-CS-05-T	17
CYUT-EN-CS-03-DN	17
CYUT-EN-CS-02-D	17
CYUT-EN-CS-01-T	16
OT-CS-CS-03-T	15
OT-CS-CS-02-T	13
OT-CS-CS-04-T	12
HIT-EN-CS-02-T	10
HIT-EN-CS-02-DN	10
HIT-EN-CS-02-D	10
HIT-EN-CS-01-DN	10
CMUJAV-EN-CS-01-T	7
KECIR-CS-CS-01-T	4
CMUJAV-EN-CS-02-T	3
CMUJAV-CS-CS-02-T	3
CMUJAV-CS-CS-01-T	3
MITEL-EN-CS-05-TD	2
MITEL-EN-CS-04-D	2
MITEL-EN-CS-03-T	2
MITEL-EN-CS-02-T	2
MITEL-EN-CS-01-T	2
WHUCC-CS-CS-02-T	1
WHUCC-CS-CS-01-T	1
KECIR-CS-CS-02-DN	1
NLPAI-CS-CS-05-DN	0
NLPAI-CS-CS-04-T	0
NLPAI-CS-CS-03-T	0
NLPAI-CS-CS-02-T	0
NLPAI-CS-CS-01-T	0
KECIR-CS-CS-03-DN	0

number of topics: RALI-EN-CS-04-T found fifty-three unique relevant documents for topic ACLIA1-CS-T42, six for ACLIA1-CS-T87, two for ACLIA1-CS-T333, and one each for ACLIA1-CS-T{322, 359}; RALI-EN-CT-05-T found sixteen for ACLIA1-CT-T442, eleven for ACLIA1-CT-T411, and one each for ACLIA1-CT-T{177, 203, 204, 416, 435}; OT-JA-JA-01-T found twelve for ACLIA1-JA-T236, eight for ACLIA1-JA-T158, six for ACLIA1-JA-T255, three each for ACLIA1-JA-T{157, 168}, and one each for ACLIA1-JA-T{160, 166, 225, 230}.

6.2 IR Techniques Used

In this section, we first provide a brief overview of IR techniques used by the IR4QA participants.

BRKLY submitted 4 JA-JA runs [13] using their Cheshire IR system. Chasen was used for creating a word-based index. As we shall see later in Table 24, pseudo-relevance feedback was moderately successful for this team, but they remark that its effect is small compared to previous NTCIRs. Could this be due to the incompleteness of the IR4QA relevance assessments? That is,

Table 19. Unique relevant documents summed across topics: CT runs.

run name	unique relevant
RALI-EN-CT-05-T	32
RALI-EN-CT-04-T	32
RALI-EN-CT-02-T	32
RALI-EN-CT-01-T	32
OT-CT-CT-05-T	5
OT-CT-CT-04-T	5
OT-CT-CT-02-T	5
OT-CT-CT-01-T	5
CYUT-EN-CT-03-DN	3
CYUT-EN-CT-02-D	3
CYUT-EN-CT-01-T	3
OT-CT-CT-03-T	2
NTUBROWS-CT-CT-05-T∗	2
MITEL-CT-CT-04-T	1
MITEL-CT-CT-03-D	1
MITEL-CT-CT-02-T	1
MITEL-CT-CT-01-T	1
RALI-CT-CT-05-T	0
RALI-CT-CT-04-T	0
RALI-CT-CT-03-T	0
RALI-CT-CT-02-T	0
RALI-CT-CT-01-T	0
NTUBROWS-CT-CT-04-T	0
NTUBROWS-CT-CT-03-T	0
NTUBROWS-CT-CT-02-T∗	0
NTUBROWS-CT-CT-01-T	0

∗The documentIDs in these two runs were all illegal: Their unique relevant documents are counted here after a bug fix, even though the pools were created using the original runs.

Table 20. Unique relevant documents summed across topics: JA runs.

run name	unique relevant
OT-JA-JA-01-T	51
OT-JA-JA-05-T	47
OT-JA-JA-03-T	34
OT-JA-JA-02-T	32
OT-JA-JA-04-T	31
BRKLY-JA-JA-01-DN	29
BRKLY-JA-JA-02-T	26
BRKLY-JA-JA-02-DN	20
CYUT-EN-JA-03-DN	19
CYUT-EN-JA-02-D	19
CYUT-EN-JA-01-T	16
BRKLY-JA-JA-03-T	11
CMUJAV-EN-JA-05-T	4
CMUJAV-EN-JA-04-T	4
CMUJAV-EN-JA-03-T	4
CMUJAV-EN-JA-02-T	4
CMUJAV-EN-JA-01-T	4
TA-EN-JA-02-D	2
TA-EN-JA-01-D	2
CMUJAV-JA-JA-05-T	1
CMUJAV-JA-JA-04-T	1
CMUJAV-JA-JA-03-T	1
CMUJAV-JA-JA-02-T	1
CMUJAV-JA-JA-01-T	1
TA-EN-JA-03-T	0

Table 21. Unique relevant documents summed across topics: CS-teams.

team name	unique relevant
RALI	66
OT	20
CYUT	18
HIT	10
CMUJAV	8
KECIR	4
MITEL	2
WHUCC	1
NLPAI	0

Table 22. Unique relevant documents summed across topics: CT-teams.

team name	covered docs.
RALI	32
OT	5
CYUT	3
NTUBROWS	2
MITEL	1

Table 23. Unique relevant documents summed across topics: JA-teams.

team name	covered docs.
OT	51
BRKLY	29
CYUT	19
CMUJAV	4
TA	2

are many of the documents captured by pseudo-relevance feedback *unjudged relevant*?

CMUJAV submitted 2 CS-CS, 2 EN-CS, 5 JA-JA and 5 EN-JA runs [12]. This team uses the IR components of their Javelin III crosslingual QA system, as well as existing natural language tools such as morphological analyzer and named entity recognizer and extensitve external translation resources. For translating questions, they combine sentence translation with key term translation, based on their observation that the former is suitable for Event and Definition questions while the latter is suitable for Biography and Relation questions. Several heuristics are used for filtering out noisy query terms. A kind of pseudo-relevance feedback that extracts new terms using lexico-semantic patterns (the "LSP-PRF" method) is proposed.

CYUT submitted 3 EN-CS, 3 EN-CT and 3 EN-JA runs [6]. For translating topics, they reused a method from NTCIR-6 that relies on Google translation and Wikipedia. In the IR phase, they used a system built around Lucene, and reused a recently-proposed query expansion method that uses Okapi BM25 with Wikipedia anchor texts as the source of expansion terms. Word segmentation tools were used for indexing Simplified and Traditional Chinese texts. As for Japanese, they

treated each character as a word. They have also done some post hoc experiments that included monolingual runs.

HIT submitted 4 EN-CS runs [31]. Google translation was used for translating English topics into Simplified Chinese and, Indri (Kullback-Leibler divergence model with Jelinek-Mercer smoothing) was used for monolingual IR. Bigrams were used for indexing. Pseudo-relevance feedback was quite successful for this team (Table 24).

KECIR submitted three CS-CS runs [3]. This team's IR system is built around Lucene, and employs word-based indexing. But for out-of-vocabulary words, single characters are indexed. Document scores are computed by a linear combination of vector-space-model and language-model scores. Three query expansion methods are compared: Rocchio relevance feedback, local context analysis and one that relies on the Baidu online encyclopedia. Query length is optimised separately for definition, biography, event and relation questions.

MITEL submitted 5 EN-CS runs and 4 CT-CT runs, and achieved high performances for both subtasks [16]. For translating the English topics, a statistical machine translation tool was used to create a phrase dictionary, and the Baidu search engine and some heuristics were used for handling out-of-vocabulary terms. In the IR phase, Lemur (Kullback-Leibler divergence model with Dirichlet smoothing) was used. Unigram, bigram and word indexes were created separately, and the corresponding three runs were merged using a linear combination of document scores.

NLPAI submitted 5 CS-CS runs, two of which used the question analysis files submitted by other teams [15]. This team employs word-based indexing, and determines how to segment a Traditional Chinese question by conducting a pilot search and examining the number of documents returned. They use two query expansion methods, one of which appears to be related to Local Context Analysis. Although their official performances are low (Table 7), this is because their runs contained only 10 documents per topic: Their post hoc experiments show that their performances range from .4241 to .4720 in Mean AP if 1000 documents are retrieved per topic, which would have been quite competitive.

NTUBROWS submitted 5 CT-CT runs [14]. They used two IR systems, Okapi and Lucene, and experimented with both word and N-gram indexes. Their techniques include query term filtering, a kind of data fusion and document reranking. Unfortunately, they included the CIRB-011 documents in their indexes, even though they are not not part of the ACLIA target collection. So their *official* performances are not high.

OT submitted 5 CS-CS, 5 CT-CT and 5 JA-JA runs, and achieved high performances for all subtasks [29]. In the JA-JA subtask, OT-JA-JA-04-T is the top performing run: it is significantly better than the second-ranked BRKLY run. OT uses Open Text Corporation's search toolkit, and successfully applies a kind of pseudo-relevance feedback which involves merging of four ranked lists (See also Table 24). This team experiments with both word and N-gram indexing, and also discusses his results in terms of an IR metric called *Generalized Success@10*, which is similar to reciprocal rank but less top-heavy.

RALI submitted 5 CS-CS, 4 EN-CS, 5 CT-CT and 4 EN-CT runs [26]. This team uses Wikipedia in several ways, including person name translation and a special treatment of biography questions. Google translation and the Google search engine are also utilised. Indri is used as the basic IR system, and a window-based passage retrieval output is converted into a ranked list of documents. Word-based indexes are created using the ICTCLAS software. Their official results (Tables 6 and 7) are not as competitive as they should be due to a bug: After the bug fix, their Mean AP performances reach .6888 for CS-CS, and .6002 for CT-CT.

TA submitted 3 EN-JA runs [7]. This team reused a crosslingual IR method proposed at NTCIR-6, which is a kind of document translation approach using statistical machine translation. Thus, Japanese documents are indexed using English terms based on translation probabilities. GETA is used as the IR engine and GIZA++ is used for building the translation model.

WHUCC submitted 2 CS-CS runs, but these runs were in fact identical [28]. This team employs Okapi BM25 for term weighting, and conducts document reranking between the initial search and the query expansion phases. The reranking process is treated as a binary classification problem, by treating the top 20 documents in the initial ranked list as positive examples and the rest as unlabelled examples. A method from the NTCIR-4 CLIR task is adopted for extracting key terms from retrieved documents. Both bigrams and single characters are used as indexing units.

One of the goals of ACLIA IR4QA was to investigate whether QA techniques such as question clas-

sification can help improve IR performance. Unfortunately, however, no team actually used such techniques for the formal runs. We encourage participants to try these approaches in their post hoc experiments.

Here, we focus on a particular IR technique, namely, Pseudo-Relevance Feedback (PRF), and discuss its effect by looking across the participating teams. Table 24 shows some selected pairs of runs, where each pair consists of a run without PRF and a corresponding run with PRF. These runs were selected based on the DESCRIPTION field of each run file. As can be seen, HIT successfully improves performance by PRF for CS; OT successfully improves performance by PRF for CT and JA; BRKLY successfully improves performance by PRF for JA. Other teams are less successful with PRF: RALI-CT-CT-04-T significantly *underperforms* RALI-CT-CT-05-T in terms of Mean AP and Q. That is, PRF can hurt performance. However, it is possible that these negative trends arise partially from the fact that our qrels are very incomplete: PRF may actually be retrieving relevant documents that are not yet listed up in the qrels.

Table 24. The effect of Pseudo-Relevance Feedback. A run that is significantly better than its counterpart is indicated by $*$ ($\alpha = 0.05$) and $**$ ($\alpha = 0.01$).

run	AP	Q	nDCG	run DESCRIPTION
HIT-EN-CS-01-DN	0.5690**	0.5840**	0.7560**	Techniques used: Google translation, 2gram, **pseudo feedback**
HIT-EN-CS-02-DN	0.4634	0.4827	0.6910	Techniques used: Google translation, 2-gram
OT-CS-CS-04-T	0.6337	0.6490	0.8270	**blind feedback** based on first 3 rows of run 02
OT-CS-CS-02-T	0.6295	0.6411	0.8139	same as 05 except that training question words also stopped
RALI-CS-CS-04-T	0.4622	0.4745	0.7276	Indri structure query, Index by overlapped passage, Word segmentation by ICTCLAS (free version), Expand person name by Wikipedia. [**Pseudo Relevance Feedback** (0.2)]
RALI-CS-CS-05-T	0.4684	0.4812	0.7242	Indri structure query, Index by overlapped passage, Word segmentation by ICTCLAS (free version), Expand person name by Wikipedia. [Baseline]
RALI-EN-CS-04-T	0.4033	0.4191	0.6701	Indri structure query, Index by overlapped passage, Word segmentation by ICTCLAS (free version), Expand person name by Wikipedia. Translation: (1) google translate (2) wikipedia entries for person names and acronyms (if exist) or google-search-api (if not a wikipedia entry). [**Pseudo Relevance Feedback** (0.2)]
RALI-EN-CS-05-T	0.4013	0.4181	0.6599	Indri structure query, Index by overlapped passage, Word segmentation by ICTCLAS (free version), Expand person name by Wikipedia. Translation: (1) google translate (2) wikipedia entries for person names and acronyms (if exist) or google-search-api (if not a wikipedia entry). [Baseline]
OT-CT-CT-04-T	0.5521**	0.5724**	0.7656**	**blind feedback** based on first 3 rows of run 02
OT-CT-CT-02-T	0.5111	0.5339	0.7432	same as 05 except that training question words also stopped
RALI-CT-CT-04-T	0.3753	0.3916	0.6525	Indri structure query, Index by overlapped passage, Word segmentation by ICTCLAS (free version), Expand person name by Wikipedia. [**Pseudo Relevance Feedback** (0.2)]
RALI-CT-CT-05-T	0.3952*	0.4096*	0.6516	Indri structure query, Index by overlapped passage, Word segmentation by ICTCLAS (free version), Expand person name by Wikipedia.
RALI-EN-CT-04-T	0.2574	0.2737	0.4845	Indri structure query, Index by overlapped passage, Word segmentation by ICTCLAS (free version), Expand person name by Wikipedia. Translation: (1) google translate (2) wikipedia entries for person names and acronyms (if exist) or google-search-api (if not a wikipedia entry). [**Pseudo Relevance Feedback** (0.2)]
RALI-EN-CT-05-T	0.2723	0.2868	0.4767	Indri structure query, Index by overlapped passage, Word segmentation by ICTCLAS (free version), Expand person name by Wikipedia. Translation: (1) google translate (2) wikipedia entries for person names and acronyms (if exist) or google-search-api (if not a wikipedia entry). [Baseline]
BRKLY-JA-JA-02-T	0.5838*	0.5996**	0.7831**	Method: Logistic Regression using the Berkeley Algorithm described in: William S. Cooper, Aitao Chen and Fredric C. Gey, "Full text retrieval based on probabilistic equations with coeffieients fitted by logistic regression", In D. K. Harman, editor, The Second Text REtrieval Conference (TREC-2), pages 57-66, March 1994. **Blind Feedback**: uses the Top 10 terms from the top 10 ranked documents of an initial run
BRKLY-JA-JA-03-T	0.5407	0.5509	0.7475	Method: Logistic Regression using the Berkeley Algorithm described in: William S. Cooper, Aitao Chen and Fredric C. Gey, "Full text retrieval based on probabilistic equations with coeffieients fitted by logistic regression", In D. K. Harman, editor, The Second Text REtrieval Conference (TREC-2), pages 57-66, March 1994. No Blind Feedback
CMUJAV-EN-JA-02-T	0.4192	0.4265	0.5958	Combined basic keyterm based query with **PRF** queries from top 5 documents
CMUJAV-EN-JA-05-T	0.4187	0.4265	0.5971	Basic keyterm based query only
CMUJAV-JA-JA-02-T	0.5790	0.5875	0.7743	Combined basic keyterm based query with **PRF** queries from top 5 documents
CMUJAV-JA-JA-05-T	0.5784	0.5852	0.7723	Basic keyterm based query only
OT-JA-JA-04-T	0.6979*	0.7090*	0.8650**	**blind feedback** based on first 3 rows of run 02
OT-JA-JA-02-T	0.6698	0.6808	0.8473	same as 05 except that training question words also stopped

Table 25. Kendall and Yilmaz/Aslam/Robertson rank correlation: System ranking by Mean AP vs Mean Q, etc.

CS runs	AP	Q	nDCG
AP	**1/1**	**.931/.930**	.823/.806
Q	**.931/.929**	**1/1**	.872/.846
nDCG	.823/.693	.872/.737	**1/1**
CT runs	AP	Q	nDCG
AP	**1/1**	**.975/.903**	.785/.745
Q	**.975/.903**	**1/1**	.809/.841
nDCG	.785/.731	.809/.830	**1/1**
JA runs	AP	Q	nDCG
AP	**1/1**	**.933/.934**	.880/.869
Q	**.933/.929**	**1/1**	**.947/.933**
nDCG	.880/.856	**.947/.922**	**1/1**

Table 26. Kendall and Yilmaz/Aslam/Robertson rank correlation: Topic ranking by Average AP vs Average Q, etc.

CS runs	AP	Q	nDCG
AP	**1/1**	**.907**/.807	.792/.625
Q	**.907**/.826	**1/1**	.867/.760
nDCG	.792/.683	.867/.789	**1/1**
CT runs	AP	Q	nDCG
AP	**1/1**	.896/.829	.761/.656
Q	.896/.836	**1/1**	.812/.714
nDCG	.761/.644	.812/.703	**1/1**
JA runs	AP	Q	nDCG
AP	**1/1**	**.909**/.841	.691/.603
Q	**.909**/.842	**1/1**	.756/.691
nDCG	.691/.600	.756/.689	**1/1**

6.3 Correlation between Two Metrics

In this section, we examine how two system rankings (or topic rankings) according to two different metrics resemble each other. For this purpose, we use Kendall's rank correlation and Yilmaz/Aslam/Robertson (YAR) rank correlation [30]. Kendall's rank correlation is a monotonic function of the probability that a *randomly chosen* pair of ranked systems 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 an equal impact. However, for the purpose of ranking retrieval systems, the ranks near the top of the list are arguably more important than those near the bottom. In light of this, the recently-proposed YAR rank correlation is a monotonic function of the probability that a randomly chosen system *and a one ranked above it* are ordered identically in the two rankings. Like Kendall's rank correlation, YAR rank correlation lies between -1 and 1, but unlike Kendall's, it is not symmetrical. When the errors (i.e., pairwise swaps with respect to the gold standard) are uniformly distributed over the ranked list being examined, YAR rank correlation is equivalent to Kendall's rank correlation.

Table 25 compares two system rankings according to two different evaluation metrics using Kendall's tau rank correlation and Yilmaz/Aslam/Roberton (YAR) rank correlation. For example, for the CS runs, the Kendall's correlation between the system ranking by Mean AP and that by Mean Q is .931; YAR correlation between Mean AP and Mean Q is .930 when the latter is taken as the ground truth, and .929 when the former is taken as the ground truth. Values higher than 0.9 are shown in bold just for convenience. Similarly, Table 26 compares two *topic* rankings according to the metrics averaged across runs. It can be observed that the rank correlations between two metrics (for ranking both systems and topics) are generally high, with Mean AP and Mean Q showing the highest correlation consistently. Recall that AP is a special case of Q, namely, Q with $\beta = 0$ (See Section 4.2).

Figures 9-11 visualise how the system rankings by Mean Q and nDCG are correlated with that by Mean AP. The systems have been sorted by Mean AP, and the Mean AP/Q/nDCG values are shown for each system. Hence a curve that goes up (from left to right) indicates inconsistency with Mean AP. For example:

- In Figure 9, KECIR-CS-CS-02-DN outperforms RALI-CS-CS-05-T according to Mean AP (.4864 vs. .4684), but Mean Q and nDCG disagree with this (.4645 vs. .4812 and .6306 vs. .7242);
- In Figure 10, NTUBROWS-CT-CT-01-T outperforms OT-CT-CT-01-T according to Mean AP and Q (.3587 vs. .3228 and .3780 vs. .3726), but Mean nDCG disagrees (.5932 vs. .6594);
- In Figure 11, CMUJAV-EN-JA-05-T outperforms OT-JA-JA-01-T according to Mean AP (.4187 vs. .3893), but Mean Q and nDCG disagree with this (.4265 vs. .4376 and .5971 vs. .7157).

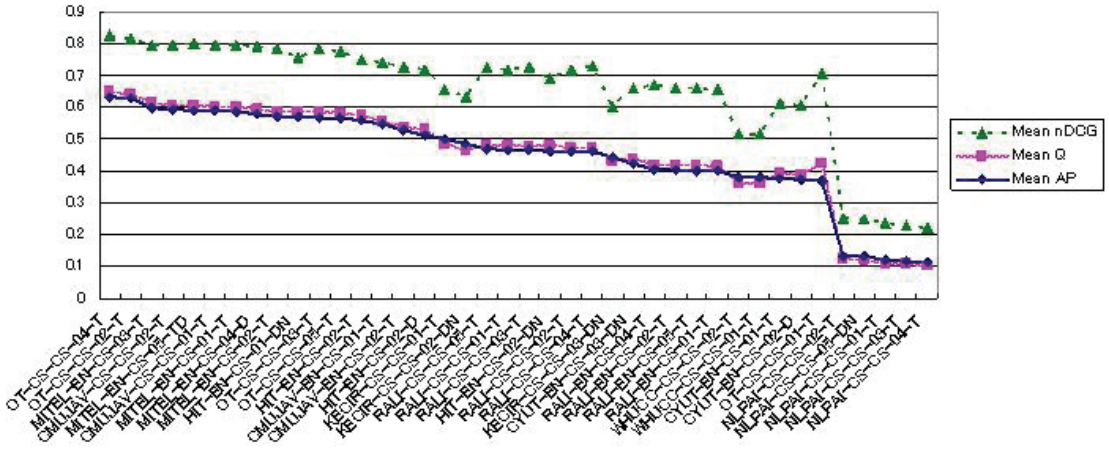

Figure 9. How system rankings by Mean Q and nDCG differ from that by Mean AP: CS runs. The systems have been sorted by Mean AP values.

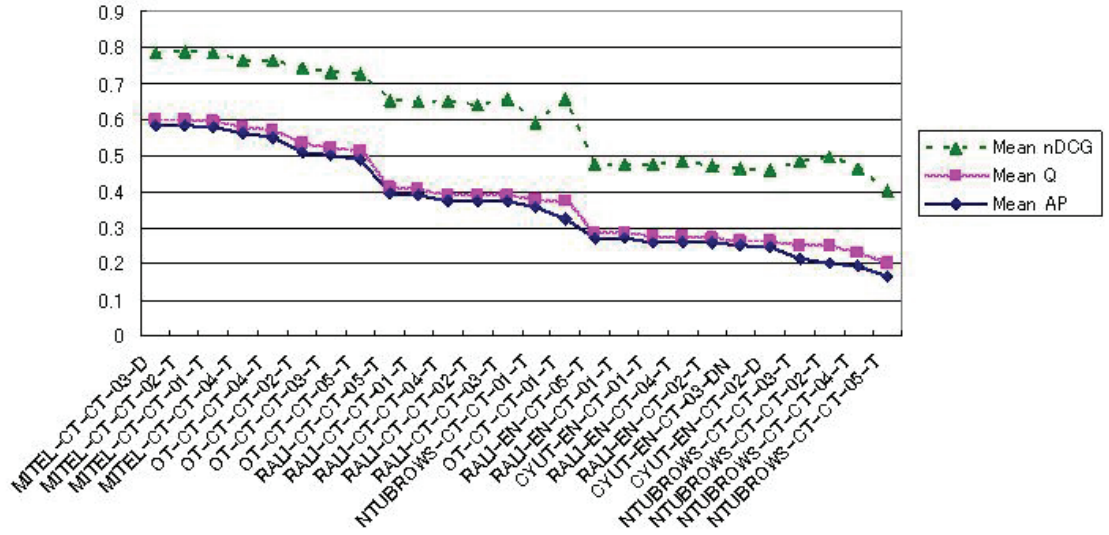

Figure 10. How system rankings by Mean Q and nDCG differ from that by Mean AP: CT runs. The systems have been sorted by Mean AP values.

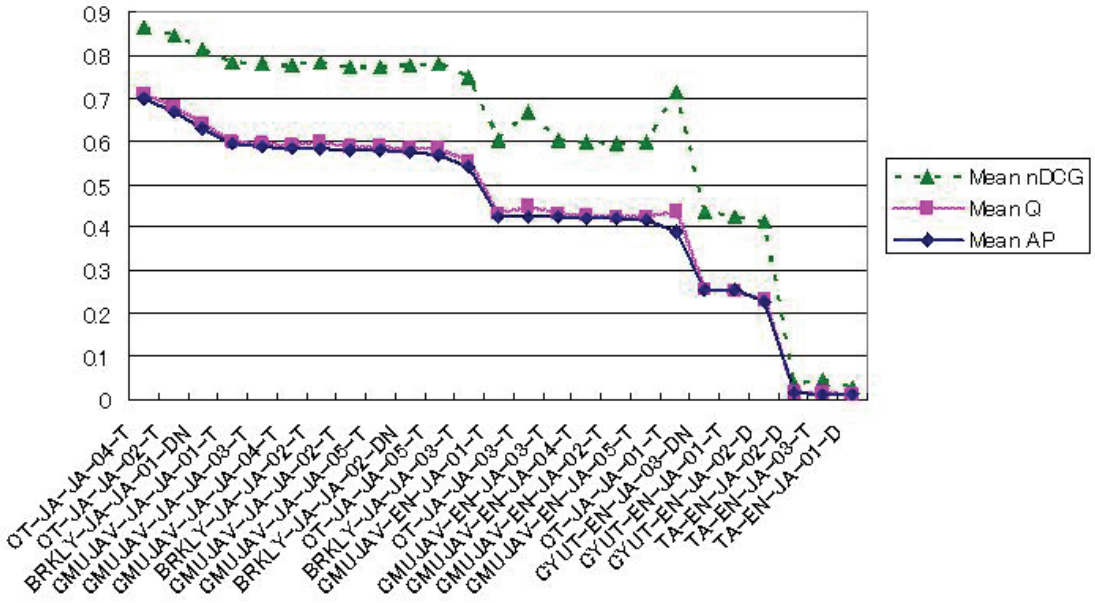

Figure 11. How system rankings by Mean Q and nDCG differ from that by Mean AP: JA runs. The systems have been sorted by Mean AP values.

Table 27. Kendall and Yilmaz/Aslam/Robertson rank correlation: System ranking by pseudo-qrels vs real qrels for each metric.

CS runs	
AP	.690/.636
Q	.700/.631
nDCG	.738/.678
CT runs	
AP	.772/.656
Q	.778/.713
nDCG	.662/.618
JA runs	
AP	.727/.512
Q	.693/.479
nDCG	.747/.473

Table 28. Kendall and Yilmaz/Aslam/Robertson rank correlation: Topic ranking by pseudo-qrels vs real qrels for each metric.

CS runs	
AP	.404/.269
Q	.432/.279
nDCG	.476/.336
CT runs	
AP	.400/.250
Q	.452/.299
nDCG	.501/.333
JA runs	
AP	.427/.433
Q	.447/.413
nDCG	.485/.456

6.4 Correlation between Pseudo-Qrels and Real Qrels

We now examine how pseudo-qrels described in Section 5 resemble real qrels. Table 27 compares a system ranking according to a metric with pseudo qrels with another one according to the same metric with real qrels. For the YAR rank correlation which is not symmetric, the ranking with real qrels is taken as the ground truth. For example, the Kendall's rank correlation with the system ranking by Mean AP with pseudo-qrels and that by Mean AP with real qrels is .690, while the corresponding YAR rank correlation is .636. It can be observed that the YAR correlation values are considerably lower than the Kendall ones for the JA-runs, suggesting that the ranking of JA-runs with pseudo-qrels has a considerable number of errors near the top of the ranked list (See also Figure 14 below).

Figures 12-14 visualise how the system rankings by Mean AP with pseudo-qrels are correlated with that by Mean AP with real qrels. The systems have been sorted by real Mean AP values, and the Mean AP values are shown for each system with both real and pseudo-qrels. As mentioned earlier, pseudo-qrels are not good at predicting the ranking of the top performers for JA (Figure 14). This possibly reflects the fact that the JA qrels are less incomplete than the CS and CT qrels: See Tables 29-31. On the other hand, the CT and JA results seem to suggest that pseudo-qrels may be good at predicting the low performers. In words, this can be summarised roughly as: *Systems that retrieve popular documents are not necessarily good; However, systems that do not retrieve popular documents are probably bad*.

Table 28 compares a *topic* ranking according to a metric with pseudo qrels with another one according to the same metric with real qrels. This time, the YAR correlations values show that ranking the CS and CT topics based on pseudo-qrels can be inaccurate. That is, pseudo-qrels is not good at predicting topic difficulty.

7 Conclusions

This paper presented an overview of the NTCIR-7 ACLIA IR4QA Task as well as some initial findings from the official results, including some positive and negative effects of PRF across participating teams. Our preliminary analysis suggests that pseudo-qrels may be useful for predicting low performers. However, it appears that they are not good at predicting top performers and predicting topic difficulty.

Together with the IR4QA participants, we would like to address the following questions in our future work:

- 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?

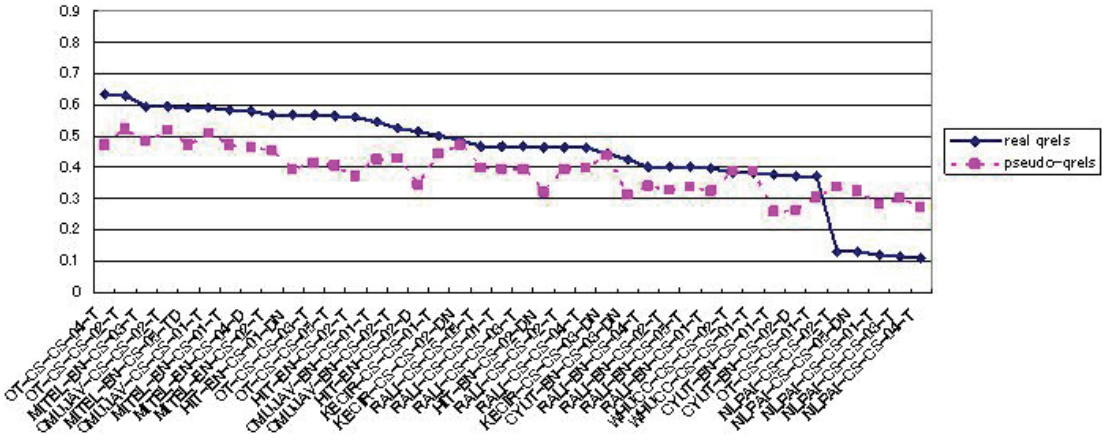

Figure 12. How system ranking by Mean AP with pseudo-qrels differs from that by Mean AP with real qrels: CS runs. The systems have been sorted by real Mean AP values.

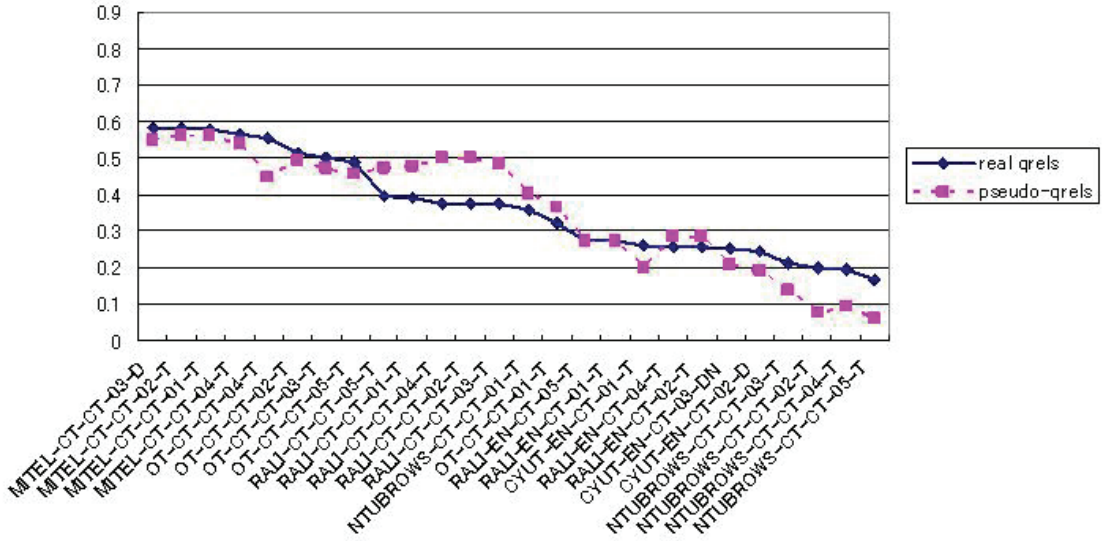

Figure 13. How system ranking by Mean AP with pseudo-qrels differs from that by Mean AP with real qrels: CT runs. The systems have been sorted by real Mean AP values.

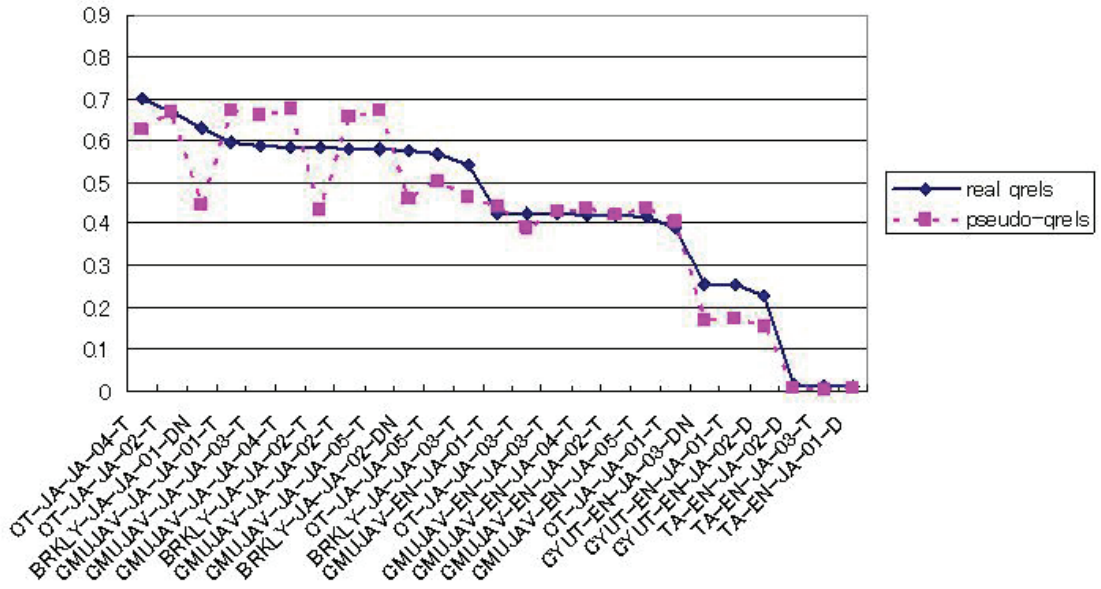

Figure 14. How system ranking by Mean AP with pseudo-qrels differs from that by Mean AP with real qrels: JA runs. The systems have been sorted by real Mean AP values.

Acknowledgments

We thank Kazuko Kuriyama (Shirayuri College), Tsuneaki Kato (University of Tokyo) and Tatsunori Mori (Yokohama National University) for their advice, and all the ACLIA participants for their hard work.

References

[1] Aslam, J. A. and Savell, R.: On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments, *Proceedings of ACM SIGIR 2003*, pp. 361-362, 2003.

[2] Burges, C. *et al.*: Learning to Rank using Gradient Descent, *Proceedings of ACM ICML 2005*, pp. 89-96, 2005.

[3] Cai, D., Li, D., Bai, Y., Zhou, B.: KECIR Information Retrieval System for NTCIR-7 IR4QA Task, *Proceedings of NTCIR-7*, to appear, 2008.

[4] Efron, B. and Tibshirani, R.: *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1993.

[5] Harman, D. K.: The TREC Test Collections, In Voorhees, E. M. and Harman, D. K. (eds.), *TREC: Experiment and Evaluation in Information Retrieval*, The MIT Press, pp. 21-52, 2005.

[6] Hsu, C.-C., Li, Y.-T., Chen, Y.-W. and Wu, S.-H.: Query Expansion via Link Analysis of Wikipedia for CLIR, *Proceedings of NTCIR-7*, to appear, 2008.

[7] Hyodo, T. and Akiba, T.: Statistical Machine Translation Based Passage Retrieval – Experiment at NTCIR-7 IR4QA, *Proceedings of NTCIR-7*, to appear, 2008.

[8] Iwayama, M., Fujii, A., Kando, N. and Takano, A.: Overview of the Patent Retrieval Task at NTCIR-3, *Proceedings of NTCIR-3*, 2003. Available at: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-OV-PATENT-IwayamaM.pdf

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

[10] Järvelin, K., Price, S. L., Delcambre, L. M. L. and Nielsen, M. L.: Discounted Cumulative Gain Based Evaluation of Multiple-Query IR Sessions, *Proceedings of ECIR 2008*, LNCS 4956, pp. 4-15, 2008.

[11] Kando, N.: Evaluation of Information Access Technologies at the NTCIR Workshop, *Proceedings of CLEF 2003*, LNCS 3237, pp. 29-43, 2004.

[12] Lao, N., Shima, H., Mitamura, T. and Nyberg, E.: Query Expansion and Machine Translation for Robust Cross-Lingual Information Retrieval, *Proceedings of NTCIR-7*, to appear, 2008.

[13] Larson, R. R. and Gey, F. C.: High Baseline Japanese Information Retrieval for Question-Answering, *Proceedings of NTCIR-7*, to appear, 2008.

[14] Liu, I.-C., Ku, L.-W., Chen, K.-H. and Chen, H.-H.: NTUBROWS System for NTCIR-7 Information Retrieval for Question Answering, *Proceedings of NTCIR-7*, to appear, 2008.

[15] Liu, M., Fang, F., Hu, Q. and Chen, J.: Question Analysis and Query Expansion in CS-CS IR4QA, *Proceedings of NTCIR-7*, to appear, 2008.

[16] Luo, W., Xia, T., Guo, J. and Liu, Q.: ICT-Crossn: The System of Cross-lingual Information Retrieval of ICT in NTCIR-7, *Proceedings of NTCIR-7*, to appear, 2008.

[17] Mitamura, T. *et al.*: Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access, *Proceedings of NTCIR-7*, to appear, 2008.

[18] Mizzaro, S. and Robertson, S.: HITS Hits TREC – Exploring IR Evaluation Results with Network Analysis, *Proceedings of ACM SIGIR 2007*, pp. 479-486, 2007.

[19] Robertson, S.: A New Interpretation of Average Precision, *Proceedings of ACM SIGIR 2008*, pp. 689-690, 2008.

[20] 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

[21] 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. Available at: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/EVIA/1.pdf

[22] 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. Available at: http://www.jstage.jst.go.jp/article/ipsjdc/3/0/625/_pdf

[23] Sakai, T. and Kando, N.: On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments, *Information Retrieval*, 2008. Available at: http://www.springerlink.com/content/k41j115214032614/fulltext.pdf

[24] Sakai, T. and Kando, N.: Are Popular Documents More Likely To Be Relevant? A Dive into the ACLIA IR4QA Pools, *Proceedings of the Second Workshop on Evaluating Information Access (EVIA 2008)*, to appear, 2008.

[25] 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)*, to appear, 2008.

[26] Shi, L., Nie, J.-Y. and Cao, G.: NTCIR-7 IR4QA Experiments at RALI, *Proceedings of NTCIR-7*, to appear, 2008.

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

[28] Teng, C. *et al.*: Information Retrieval Using PU Learning Based Re-ranking, *Proceedings of NTCIR-7*, to appear, 2008.

[29] Tomlinson, S.: Experiments in Finding Chinese and Japanese Answer Documents at NTCIR-7, *Proceedings of NTCIR-7*, to appear, 2008.

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

[31] Xiaoning, H. *et al.*: Using Google Translation in Cross-Language Information Retrieval, *Proceedings of NTCIR-7*, to appear, 2008.

Appendix

Table 29. Pool size for the CS relevance assessments. For example, "CS-T41" represents the topic ACLIA-CS-T41. Note that $P'_{50} = P_{50} - P_{30}$, and so on (See Section 2). The pools covered by qrels version 1 are indicated in bold.

topic	P_{30}	P'_{50}	P'_{70}	P'_{90}	P'_{100}	total (P_{100})	topic	P_{30}	P'_{50}	P'_{70}	P'_{90}	P'_{100}	total (P_{100})
CS-T41	**159**	**71**	73	81	35	419	CS-T100	**131**	70	63	61	25	350
CS-T42	**331**	**242**	222	160	84	1039	CS-T101	**185**	177	171	167	82	782
CS-T43	**177**	**79**	56	53	36	401	CS-T102	**127**	108	125	103	68	531
CS-T44	**286**	**158**	157	134	69	804	CS-T103	**239**	93	78	87	40	537
CS-T46	**251**	**132**	105	112	63	663	CS-T104	**140**	73	78	80	35	406
CS-T47	**269**	**126**	120	128	56	699	CS-T317	**208**	169	159	151	85	772
CS-T48	**277**	**148**	152	128	86	791	CS-T320	**155**	85	78	89	44	451
CS-T49	**177**	**75**	69	77	36	434	CS-T321	**149**	68	69	74	40	400
CS-T52	**125**	**77**	73	55	25	355	CS-T322	**327**	180	189	165	96	957
CS-T53	**252**	**162**	149	148	68	779	CS-T323	**162**	114	108	107	59	550
CS-T54	**221**	**118**	148	148	70	705	CS-T324	**147**	85	65	72	35	404
CS-T55	**228**	**92**	101	116	45	582	CS-T325	**255**	102	85	86	40	568
CS-T56	**372**	**204**	188	185	89	1038	CS-T326	**176**	111	97	117	55	556
CS-T57	**276**	**132**	117	116	59	700	CS-T328	**197**	81	69	71	41	459
CS-T58	**103**	**69**	59	59	31	321	CS-T329	**98**	87	78	93	48	404
CS-T60	**135**	**54**	64	52	25	330	CS-T331	**466**	265	230	218	110	1289
CS-T61	**188**	**89**	73	61	35	446	CS-T332	**223**	125	129	108	67	652
CS-T62	**325**	**216**	205	218	93	1057	CS-T333	**319**	**176**	153	157	83	888
CS-T64	**199**	**89**	67	67	42	464	CS-T334	**358**	220	169	165	85	997
CS-T65	**115**	**64**	51	48	27	305	CS-T336	**135**	**59**	81	86	45	406
CS-T67	**213**	**105**	112	106	57	593	CS-T337	**181**	89	69	61	44	444
CS-T68	**96**	**93**	132	182	84	587	CS-T338	**101**	58	67	95	45	366
CS-T69	**130**	**93**	156	221	109	709	CS-T339	**159**	64	58	53	35	369
CS-T71	**156**	**119**	153	165	82	675	CS-T340	**103**	57	49	50	31	290
CS-T73	**446**	**233**	201	188	85	1153	CS-T347	**151**	129	142	117	55	594
CS-T74	**438**	261	248	237	122	1306	CS-T348	**121**	45	48	39	19	272
CS-T75	**194**	**117**	150	149	72	682	CS-T349	**150**	127	138	129	66	610
CS-T76	**132**	**65**	66	57	30	350	CS-T350	**228**	152	135	159	65	739
CS-T77	**249**	**183**	143	143	55	773	CS-T351	**115**	47	56	55	31	304
CS-T78	**359**	**175**	175	135	67	911	CS-T352	**130**	64	59	73	40	366
CS-T79	**203**	**88**	101	102	44	538	CS-T355	**188**	91	82	83	48	492
CS-T80	**272**	**169**	162	171	82	856	CS-T357	**251**	108	106	99	43	607
CS-T81	**154**	**36**	37	43	17	287	CS-T358	**242**	155	135	157	69	758
CS-T82	**385**	191	169	154	82	981	CS-T359	**311**	151	146	156	65	829
CS-T83	**324**	171	127	124	74	820	CS-T361	**106**	44	50	47	29	276
CS-T84	**141**	81	84	67	34	407	CS-T362	**197**	73	105	94	47	516
CS-T85	**120**	51	47	45	34	297	CS-T365	**194**	118	119	119	68	618
CS-T86	**323**	161	172	193	109	958	CS-T366	**206**	122	102	125	57	612
CS-T87	**404**	185	168	154	64	975	CS-T367	**212**	103	86	76	41	518
CS-T89	**175**	144	101	110	45	575	CS-T368	**259**	138	106	113	52	668
CS-T90	**260**	109	87	99	45	600	CS-T369	**252**	130	112	134	56	684
CS-T91	**269**	109	85	79	37	579	CS-T370	**85**	35	67	83	42	312
CS-T92	**216**	114	87	91	49	557	CS-T376	**224**	117	102	80	44	567
CS-T93	**157**	96	113	135	64	565	CS-T378	**102**	99	125	129	62	517
CS-T94	**296**	156	166	192	87	897	CS-T379	**216**	101	113	93	49	572
CS-T95	**292**	139	125	139	69	764	CS-T380	**201**	143	107	79	31	561
CS-T96	**141**	50	70	57	24	342	CS-T381	**132**	99	104	109	63	507
CS-T97	**146**	48	53	48	29	324	CS-T383	**279**	224	190	164	90	947
CS-T98	**128**	70	62	81	43	384	CS-T384	**146**	111	101	123	65	546
CS-T99	**215**	125	131	145	69	685	CS-T385	**193**	124	139	129	65	650
							total	21132	11700	11224	11238	5638	60932

Qrels version 1 misses one depth-30-pool document for topic ACLIA1-CS-T369, as it was inadvertently left unjudged. Qrels version 1 also contains some extra judged documents for ACLIA-CS-T{334, 337, 385, 74, 81}: Some assessors went beyond the pool depth indicated in bold, but stopped halfway.

Table 30. Pool size for the CT relevance assessments. For example, "CT-T171" denotes the topic ACLIA-CT-T171. Note that $P'_{50} = P_{50} - P_{30}$, and so on (See Section 2). The pools covered by qrels version 1 are indicated in bold.

topic	P_{30}	P'_{50}	P'_{70}	P'_{90}	P'_{100}	total (P_{100})	topic	P_{30}	P'_{50}	P'_{70}	P'_{90}	P'_{100}	total (P_{100})
CT-T171	**272**	153	138	177	71	811	CT-T400	**252**	124	125	117	62	680
CT-T172	**267**	135	132	137	58	729	CT-T402	**303**	171	173	167	81	895
CT-T174	**260**	140	136	140	62	738	CT-T403	**235**	141	140	119	60	695
CT-T175	**208**	103	101	112	52	576	CT-T404	**262**	157	131	115	78	743
CT-T176	**321**	180	192	184	81	958	CT-T405	**221**	164	168	144	63	760
CT-T177	**313**	176	178	176	102	945	CT-T406	**306**	213	184	181	83	967
CT-T178	**231**	127	112	98	45	613	CT-T407	**149**	68	80	81	53	431
CT-T179	**218**	133	121	158	71	701	CT-T408	**133**	78	76	79	43	409
CT-T180	**310**	183	195	183	81	952	CT-T409	**169**	101	85	77	47	479
CT-T181	**246**	134	124	115	49	668	CT-T410	**254**	222	190	206	93	965
CT-T182	**218**	129	137	127	59	670	CT-T411	**305**	168	163	173	84	893
CT-T183	**305**	186	185	182	99	957	CT-T412	**292**	163	146	159	86	846
CT-T184	**237**	165	184	183	74	843	CT-T413	**225**	118	119	118	45	625
CT-T186	**223**	129	120	113	59	644	CT-T414	**298**	176	179	170	79	902
CT-T187	**166**	213	196	175	99	849	CT-T415	**344**	223	180	201	108	1056
CT-T188	**376**	233	226	217	108	1160	CT-T416	**340**	198	179	170	100	987
CT-T189	**211**	154	152	136	67	720	CT-T417	**254**	177	189	186	85	891
CT-T190	**339**	180	187	173	78	957	CT-T418	**207**	106	105	114	45	577
CT-T191	**198**	103	105	84	48	538	CT-T419	**324**	217	206	218	105	1070
CT-T193	**204**	135	141	153	79	712	CT-T420	**239**	138	139	125	71	712
CT-T194	**196**	95	85	97	51	524	CT-T421	**238**	147	133	157	69	744
CT-T195	**207**	98	101	98	49	553	CT-T422	**242**	155	144	152	70	763
CT-T196	**333**	212	206	210	98	1059	CT-T423	**207**	152	144	144	79	726
CT-T197	**190**	123	150	154	101	718	CT-T424	**289**	169	147	136	78	819
CT-T198	**236**	143	153	155	86	773	CT-T426	**202**	104	109	98	46	559
CT-T200	**262**	109	92	105	52	620	CT-T427	**215**	113	111	131	100	670
CT-T203	**373**	206	238	203	108	1128	CT-T428	**150**	75	77	96	42	440
CT-T204	**325**	190	193	170	107	985	CT-T429	**217**	117	112	98	43	587
CT-T205	**250**	158	158	113	55	734	CT-T430	**265**	148	165	153	80	811
CT-T206	**317**	211	170	168	74	940	CT-T431	**181**	155	153	155	80	724
CT-T207	**225**	194	236	230	111	996	CT-T432	**230**	177	168	195	86	856
CT-T208	**286**	207	189	179	75	936	CT-T433	**291**	156	155	143	77	822
CT-T210	**149**	91	107	98	47	492	CT-T434	**167**	111	111	106	41	536
CT-T211	**205**	137	147	119	54	662	CT-T435	**297**	201	159	169	80	906
CT-T213	**181**	120	100	81	48	530	CT-T436	**278**	195	164	185	102	924
CT-T374	**241**	130	100	117	41	629	CT-T437	**284**	155	179	182	77	877
CT-T386	**216**	128	107	115	54	620	CT-T438	**244**	137	140	131	65	717
CT-T387	**206**	86	79	79	37	487	CT-T439	**166**	98	100	90	53	507
CT-T388	**345**	255	254	264	121	1239	CT-T440	**182**	97	108	101	68	556
CT-T389	**247**	233	222	190	92	984	CT-T441	**237**	138	122	155	81	733
CT-T390	**227**	138	144	163	97	769	CT-T442	**281**	170	166	158	82	857
CT-T391	**177**	152	215	215	112	871	CT-T443	**278**	169	149	145	69	810
CT-T392	**260**	184	220	202	97	963	CT-T444	**263**	154	167	135	63	782
CT-T393	**209**	123	148	253	141	874	CT-T445	**223**	137	115	124	70	669
CT-T394	**167**	101	151	189	118	726	CT-T446	**314**	180	192	180	86	952
CT-T395	**393**	263	256	239	126	1277	CT-T447	**266**	153	143	154	68	784
CT-T396	**205**	90	102	85	39	521	CT-T448	**320**	204	214	191	108	1037
CT-T397	**255**	129	202	235	112	933	CT-T449	**178**	195	193	175	92	833
CT-T398	**204**	101	98	91	54	548	CT-T450	**331**	190	183	194	92	990
CT-T399	**186**	184	157	161	79	767	CT-T451	**258**	192	185	189	94	918
							total	24802	15349	15207	15143	7590	78091

Qrels version 1 misses a total of 38 depth-30-pool documents, for topics ACLIA1-CT-T{174, 175, 180, 184, 186, 190, 210, 211, 389, 391, 424, 445}, as they were inadvertently left unjudged.

Table 31. Pool size for the JA relevance assessments. For example, "JA-T1" denotes the topic ACLIA-JA-T1. Note that $P'_{50} = P_{50} - P_{30}$, and so on (See Section 2). The pools covered by qrels version 1 are indicated in bold.

topic	P_{30}	P'_{50}	P'_{70}	P'_{90}	P'_{100}	total (P_{100})	topic	P_{30}	P'_{50}	P'_{70}	P'_{90}	P'_{100}	total (P_{100})
JA-T1	**265**	**170**	**208**	**242**	**117**	1002	JA-T161	**237**	116	104	117	52	626
JA-T2	**332**	**212**	**186**	**192**	**92**	1014	JA-T162	**218**	93	115	111	54	591
JA-T3	**266**	**145**	**128**	**138**	**68**	745	JA-T163	**288**	157	148	140	71	804
JA-T4	**276**	**151**	**165**	**140**	**70**	802	JA-T164	**253**	148	143	149	63	756
JA-T6	**242**	**175**	**169**	**168**	**70**	824	JA-T165	**336**	212	200	186	93	1027
JA-T7	**244**	**147**	**162**	**161**	**91**	805	JA-T166	**272**	**177**	**174**	**169**	**89**	881
JA-T9	**257**	**116**	**103**	**102**	**46**	624	JA-T167	**244**	**149**	**133**	**135**	**60**	721
JA-T10	**283**	**264**	**240**	**211**	**101**	1099	JA-T168	**340**	**215**	**208**	**227**	**102**	1092
JA-T13	**261**	**144**	**176**	**167**	**72**	820	JA-T170	**257**	**144**	**144**	**171**	**81**	797
JA-T15	**249**	**137**	**141**	**126**	**64**	717	JA-T215	**250**	**185**	**179**	**195**	**87**	896
JA-T17	**212**	**126**	**131**	**128**	**67**	664	JA-T217	**130**	**70**	**65**	**99**	**50**	414
JA-T18	**251**	**176**	**176**	**160**	**75**	838	JA-T218	**253**	**141**	**152**	**141**	**76**	763
JA-T19	**285**	**174**	**181**	**167**	**79**	886	JA-T221	**164**	**77**	**79**	**74**	**40**	434
JA-T20	**247**	**192**	**178**	**169**	**82**	868	JA-T222	**248**	**152**	**150**	**131**	**76**	757
JA-T25	**235**	**113**	101	85	43	577	JA-T223	**236**	**141**	**130**	**124**	**64**	695
JA-T29	**217**	**241**	223	216	111	1008	JA-T224	**219**	**105**	**119**	**150**	**84**	677
JA-T32	**309**	**198**	183	179	85	954	JA-T225	**290**	**185**	**186**	**162**	**81**	904
JA-T35	**259**	143	146	244	120	912	JA-T230	**251**	**148**	**132**	**125**	**62**	718
JA-T37	**317**	222	193	205	114	1051	JA-T231	**247**	**138**	**125**	**131**	**56**	697
JA-T38	**200**	239	231	215	100	985	JA-T233	**195**	**105**	**87**	**80**	**38**	505
JA-T105	**288**	163	160	159	86	856	JA-T234	**205**	**115**	**112**	**142**	**72**	646
JA-T106	**181**	106	110	122	78	597	JA-T236	**313**	176	172	151	83	895
JA-T107	**158**	199	184	211	97	849	JA-T237	**302**	218	197	206	106	1029
JA-T108	**293**	222	198	191	86	990	JA-T238	**279**	153	168	139	75	814
JA-T109	**244**	149	155	160	71	779	JA-T239	**229**	**136**	164	152	60	741
JA-T110	**239**	**158**	**200**	**193**	**95**	885	JA-T240	**294**	244	201	191	96	1026
JA-T111	**195**	**127**	**121**	**124**	**72**	639	JA-T242	**195**	**115**	110	91	44	555
JA-T112	**184**	**113**	**97**	**127**	**53**	574	JA-T244	**238**	203	209	195	88	933
JA-T113	**187**	**103**	**108**	**108**	**63**	569	JA-T245	**213**	201	199	195	96	904
JA-T115	**178**	**129**	**136**	**142**	**73**	658	JA-T248	**230**	180	185	169	75	839
JA-T116	**245**	**153**	**132**	**148**	**76**	754	JA-T249	**224**	148	159	154	85	770
JA-T119	**232**	**145**	154	154	83	768	JA-T250	**192**	99	120	118	64	593
JA-T127	**233**	195	197	194	98	917	JA-T253	**172**	126	116	130	52	596
JA-T128	**234**	160	152	126	69	741	JA-T254	**223**	125	114	117	59	638
JA-T130	**241**	183	203	194	94	915	JA-T255	**279**	158	145	150	71	803
JA-T134	**209**	98	82	83	34	506	JA-T266	**183**	**106**	**109**	**107**	**50**	555
JA-T137	**330**	220	202	193	85	1030	JA-T267	**232**	**128**	**113**	**124**	**59**	656
JA-T138	**220**	103	105	92	40	560	JA-T271	**127**	**89**	**89**	**98**	**59**	462
JA-T140	**183**	95	112	109	54	553	JA-T275	**206**	**103**	**126**	**141**	**79**	655
JA-T141	**184**	186	191	200	104	865	JA-T276	**298**	**179**	191	190	87	945
JA-T148	**253**	**170**	**153**	**173**	**68**	817	JA-T284	**340**	203	222	218	113	1096
JA-T149	**181**	**107**	**118**	**106**	**64**	576	JA-T291	**218**	108	105	95	52	578
JA-T151	**213**	**129**	**124**	**135**	**63**	664	JA-T295	**193**	114	103	101	46	557
JA-T152	**256**	**123**	**166**	**134**	**67**	746	JA-T297	**277**	167	138	141	68	791
JA-T153	**202**	**129**	**164**	**163**	**70**	728	JA-T300	**234**	113	97	102	59	605
JA-T154	**288**	**156**	**135**	**151**	**90**	820	JA-T301	**220**	114	110	132	70	646
JA-T155	**192**	**115**	**133**	**137**	**81**	658	JA-T304	**218**	119	122	129	74	662
JA-T157	**273**	130	138	132	56	729	JA-T305	**243**	170	185	163	85	846
JA-T158	**347**	186	216	203	87	1039	JA-T313	**294**	190	189	171	84	928
JA-T160	**286**	145	163	160	84	838	JA-T315	**311**	175	166	162	80	894
							total	24266	15215	15139	15130	7478	77228

Qrels version 1 misses one depth-30-pool document for topic ACLIA1-JA-T32, as it was inadvertently left unjudged.

Table 32. Performance averaged across CS runs for each topic, using the *pseudo-qrels*. For example, "CS-T349" denotes the topic ACLIA-CS-T349. The topics are sorted by the average performance.

	AP		AP		Q		Q		nDCG		nDCG
CS-T349	0.6902	CS-T85	0.3694	CS-T68	0.7159	CS-T94	0.4126	CS-T68	0.8638	CS-T338	0.6177
CS-T317	0.6880	CS-T338	0.3670	CS-T349	0.7115	CS-T49	0.4099	CS-T84	0.8490	CS-T64	0.6172
CS-T68	0.6759	CS-T355	0.3604	CS-T317	0.7088	CS-T98	0.4024	CS-T384	0.8371	CS-T376	0.6119
CS-T84	0.6464	CS-T385	0.3588	CS-T84	0.6807	CS-T323	0.4004	CS-T349	0.8351	CS-T46	0.6110
CS-T58	0.6427	CS-T99	0.3575	CS-T58	0.6782	CS-T99	0.4003	CS-T52	0.8338	CS-T385	0.6064
CS-T329	0.6388	CS-T323	0.3553	CS-T52	0.6761	CS-T355	0.3998	CS-T58	0.8298	CS-T379	0.6047
CS-T384	0.6363	CS-T98	0.3517	CS-T329	0.6739	CS-T324	0.3955	CS-T329	0.8243	CS-T49	0.6015
CS-T52	0.6358	CS-T324	0.3511	CS-T384	0.6674	CS-T385	0.3895	CS-T317	0.8224	CS-T357	0.5994
CS-T80	0.6345	CS-T379	0.3395	CS-T381	0.6555	CS-T376	0.3778	CS-T381	0.8065	CS-T358	0.5976
CS-T381	0.6172	CS-T358	0.3363	CS-T80	0.6504	CS-T379	0.3776	CS-T75	0.7834	CS-T98	0.5957
CS-T347	0.6067	CS-T376	0.3344	CS-T75	0.6428	CS-T79	0.3676	CS-T347	0.7812	CS-T352	0.5938
CS-T75	0.6051	CS-T357	0.3306	CS-T347	0.6353	CS-T358	0.3616	CS-T378	0.7807	CS-T55	0.5835
CS-T365	0.5787	CS-T79	0.3297	CS-T365	0.6152	CS-T357	0.3616	CS-T365	0.7721	CS-T337	0.5787
CS-T71	0.5645	CS-T46	0.3171	CS-T71	0.6033	CS-T55	0.3615	CS-T71	0.7638	CS-T79	0.5773
CS-T53	0.5624	CS-T55	0.3159	CS-T378	0.6011	CS-T43	0.3519	CS-T80	0.7587	CS-T43	0.5766
CS-T383	0.5604	CS-T332	0.3140	CS-T53	0.5771	CS-T352	0.3501	CS-T104	0.7437	CS-T42	0.5763
CS-T378	0.5492	CS-T95	0.3026	CS-T383	0.5709	CS-T332	0.3465	CS-T76	0.7369	CS-T94	0.5737
CS-T102	0.5156	CS-T337	0.3025	CS-T104	0.5503	CS-T46	0.3456	CS-T326	0.7363	CS-T332	0.5737
CS-T366	0.5150	CS-T43	0.2994	CS-T102	0.5484	CS-T337	0.3365	CS-T89	0.7350	CS-T369	0.5712
CS-T326	0.5129	CS-T92	0.2992	CS-T326	0.5458	CS-T328	0.3357	CS-T60	0.7164	CS-T368	0.5662
CS-T89	0.5125	CS-T352	0.2987	CS-T89	0.5445	CS-T369	0.3326	CS-T41	0.7163	CS-T92	0.5621
CS-T104	0.5079	CS-T57	0.2967	CS-T69	0.5411	CS-T92	0.3321	CS-T93	0.7134	CS-T328	0.5608
CS-T69	0.5017	CS-T369	0.2959	CS-T366	0.5367	CS-T95	0.3317	CS-T366	0.7120	CS-T47	0.5522
CS-T101	0.4964	CS-T328	0.2878	CS-T101	0.5338	CS-T57	0.3270	CS-T102	0.7120	CS-T77	0.5408
CS-T380	0.4918	CS-T368	0.2831	CS-T93	0.5317	CS-T368	0.3215	CS-T101	0.7106	CS-T61	0.5388
CS-T93	0.4881	CS-T54	0.2801	CS-T380	0.5249	CS-T54	0.3105	CS-T53	0.7002	CS-T57	0.5278
CS-T76	0.4822	CS-T62	0.2780	CS-T76	0.5242	CS-T61	0.2965	CS-T97	0.7000	CS-T54	0.5243
CS-T41	0.4786	CS-T44	0.2753	CS-T41	0.5189	CS-T62	0.2906	CS-T320	0.6989	CS-T48	0.5152
CS-T60	0.4566	CS-T47	0.2567	CS-T60	0.5051	CS-T44	0.2886	CS-T380	0.6982	CS-T44	0.4998
CS-T340	0.4439	CS-T48	0.2521	CS-T97	0.4929	CS-T47	0.2863	CS-T361	0.6920	CS-T367	0.4901
CS-T97	0.4436	CS-T61	0.2516	CS-T340	0.4928	CS-T48	0.2822	CS-T69	0.6903	CS-T90	0.4817
CS-T320	0.4381	CS-T322	0.2250	CS-T320	0.4845	CS-T322	0.2508	CS-T340	0.6892	CS-T322	0.4813
CS-T321	0.4323	CS-T83	0.2244	CS-T100	0.4797	CS-T83	0.2494	CS-T351	0.6880	CS-T91	0.4806
CS-T100	0.4278	CS-T56	0.2100	CS-T65	0.4745	CS-T56	0.2371	CS-T348	0.6838	CS-T333	0.4704
CS-T65	0.4223	CS-T325	0.2078	CS-T361	0.4699	CS-T333	0.2307	CS-T81	0.6837	CS-T95	0.4670
CS-T67	0.4140	CS-T333	0.2054	CS-T321	0.4665	CS-T90	0.2286	CS-T336	0.6812	CS-T325	0.4633
CS-T361	0.4132	CS-T90	0.2023	CS-T351	0.4572	CS-T325	0.2282	CS-T65	0.6806	CS-T103	0.4590
CS-T81	0.4127	CS-T367	0.1968	CS-T81	0.4537	CS-T367	0.2186	CS-T100	0.6726	CS-T83	0.4543
CS-T42	0.4127	CS-T359	0.1896	CS-T348	0.4509	CS-T91	0.2147	CS-T350	0.6643	CS-T62	0.4402
CS-T351	0.4060	CS-T91	0.1877	CS-T336	0.4439	CS-T359	0.2085	CS-T85	0.6622	CS-T78	0.4280
CS-T336	0.4058	CS-T78	0.1825	CS-T370	0.4436	CS-T78	0.2051	CS-T339	0.6540	CS-T56	0.4083
CS-T348	0.4016	CS-T82	0.1674	CS-T67	0.4422	CS-T103	0.2034	CS-T383	0.6490	CS-T73	0.4023
CS-T350	0.3990	CS-T103	0.1668	CS-T339	0.4406	CS-T82	0.1862	CS-T96	0.6411	CS-T82	0.4019
CS-T77	0.3964	CS-T73	0.1652	CS-T350	0.4353	CS-T73	0.1820	CS-T355	0.6367	CS-T359	0.3905
CS-T339	0.3933	CS-T74	0.1358	CS-T96	0.4314	CS-T74	0.1570	CS-T370	0.6358	CS-T87	0.3561
CS-T96	0.3854	CS-T334	0.1288	CS-T42	0.4302	CS-T334	0.1395	CS-T67	0.6287	CS-T334	0.3277
CS-T370	0.3826	CS-T87	0.1130	CS-T338	0.4232	CS-T87	0.1291	CS-T99	0.6272	CS-T74	0.3150
CS-T49	0.3795			CS-T85	0.4179			CS-T323	0.6225		
CS-T94	0.3787			CS-T77	0.4168			CS-T324	0.6204		
CS-T64	0.3765			CS-T64	0.4164			CS-T321	0.6203		

Table 33. Performance averaged across CT runs for each topic, using the *pseudo-qrels*. For example, "CT-T210" denotes the topic ACLIA-CT-T210. The topics are sorted by the average performance.

	AP		AP		Q		Q		nDCG		nDCG
CT-T210	0.6569	CT-T181	0.3270	CT-T210	0.6956	CT-T200	0.3633	CT-T210	0.8369	CT-T441	0.5896
CT-T409	0.6362	CT-T179	0.3213	CT-T409	0.6737	CT-T189	0.3627	CT-T409	0.8147	CT-T420	0.5882
CT-T431	0.6037	CT-T414	0.3207	CT-T431	0.6350	CT-T179	0.3610	CT-T431	0.7999	CT-T179	0.5882
CT-T399	0.5935	CT-T182	0.3205	CT-T399	0.6246	CT-T175	0.3556	CT-T389	0.7759	CT-T414	0.5867
CT-T389	0.5907	CT-T200	0.3185	CT-T389	0.6221	CT-T178	0.3548	CT-T423	0.7755	CT-T390	0.5824
CT-T187	0.5551	CT-T189	0.3177	CT-T423	0.5960	CT-T182	0.3522	CT-T399	0.7734	CT-T421	0.5776
CT-T423	0.5492	CT-T178	0.3174	CT-T187	0.5960	CT-T414	0.3490	CT-T187	0.7506	CT-T206	0.5776
CT-T394	0.5464	CT-T175	0.3088	CT-T394	0.5824	CT-T429	0.3462	CT-T394	0.7457	CT-T418	0.5709
CT-T449	0.5361	CT-T444	0.3068	CT-T449	0.5773	CT-T444	0.3448	CT-T194	0.7422	CT-T400	0.5673
CT-T397	0.5327	CT-T435	0.3063	CT-T397	0.5630	CT-T424	0.3416	CT-T439	0.7405	CT-T444	0.5639
CT-T211	0.5240	CT-T429	0.2978	CT-T211	0.5583	CT-T418	0.3365	CT-T449	0.7297	CT-T388	0.5577
CT-T432	0.5136	CT-T186	0.2869	CT-T439	0.5529	CT-T435	0.3291	CT-T434	0.7231	CT-T189	0.5518
CT-T405	0.5125	CT-T400	0.2839	CT-T405	0.5510	CT-T374	0.3284	CT-T405	0.7230	CT-T447	0.5438
CT-T439	0.5097	CT-T418	0.2829	CT-T432	0.5466	CT-T400	0.3246	CT-T432	0.7196	CT-T182	0.5370
CT-T194	0.4956	CT-T411	0.2825	CT-T194	0.5383	CT-T186	0.3222	CT-T408	0.7145	CT-T438	0.5257
CT-T393	0.4914	CT-T420	0.2781	CT-T408	0.5338	CT-T411	0.3202	CT-T436	0.7078	CT-T387	0.5214
CT-T391	0.4895	CT-T374	0.2775	CT-T393	0.5322	CT-T420	0.3176	CT-T413	0.7076	CT-T443	0.5212
CT-T207	0.4888	CT-T176	0.2700	CT-T391	0.5223	CT-T387	0.3137	CT-T213	0.7040	CT-T430	0.5202
CT-T408	0.4767	CT-T174	0.2692	CT-T207	0.5147	CT-T447	0.3097	CT-T191	0.7034	CT-T186	0.5201
CT-T213	0.4743	CT-T387	0.2626	CT-T213	0.5119	CT-T174	0.2996	CT-T407	0.7019	CT-T435	0.5193
CT-T436	0.4700	CT-T447	0.2617	CT-T436	0.5073	CT-T176	0.2972	CT-T393	0.6952	CT-T172	0.5133
CT-T434	0.4590	CT-T438	0.2587	CT-T434	0.4982	CT-T438	0.2962	CT-T397	0.6928	CT-T174	0.5077
CT-T206	0.4539	CT-T430	0.2542	CT-T413	0.4825	CT-T430	0.2921	CT-T211	0.6793	CT-T404	0.4998
CT-T208	0.4463	CT-T195	0.2471	CT-T392	0.4774	CT-T195	0.2859	CT-T445	0.6752	CT-T176	0.4937
CT-T392	0.4423	CT-T442	0.2461	CT-T407	0.4754	CT-T442	0.2816	CT-T197	0.6629	CT-T411	0.4911
CT-T413	0.4408	CT-T416	0.2458	CT-T191	0.4748	CT-T416	0.2715	CT-T428	0.6613	CT-T448	0.4885
CT-T390	0.4366	CT-T450	0.2448	CT-T206	0.4694	CT-T450	0.2669	CT-T207	0.6582	CT-T446	0.4884
CT-T198	0.4357	CT-T448	0.2438	CT-T427	0.4676	CT-T443	0.2669	CT-T396	0.6560	CT-T416	0.4781
CT-T395	0.4346	CT-T446	0.2330	CT-T198	0.4671	CT-T448	0.2644	CT-T426	0.6542	CT-T195	0.4664
CT-T191	0.4331	CT-T443	0.2277	CT-T390	0.4662	CT-T446	0.2642	CT-T391	0.6525	CT-T450	0.4645
CT-T422	0.4300	CT-T415	0.2197	CT-T208	0.4654	CT-T404	0.2571	CT-T398	0.6461	CT-T171	0.4602
CT-T386	0.4247	CT-T404	0.2193	CT-T386	0.4639	CT-T171	0.2499	CT-T198	0.6362	CT-T424	0.4526
CT-T427	0.4226	CT-T171	0.2186	CT-T197	0.4617	CT-T172	0.2467	CT-T205	0.6316	CT-T437	0.4520
CT-T410	0.4212	CT-T437	0.2124	CT-T422	0.4591	CT-T415	0.2462	CT-T427	0.6309	CT-T442	0.4511
CT-T445	0.4153	CT-T172	0.2081	CT-T445	0.4561	CT-T437	0.2428	CT-T200	0.6304	CT-T415	0.4409
CT-T407	0.4139	CT-T177	0.2071	CT-T440	0.4541	CT-T177	0.2338	CT-T181	0.6259	CT-T190	0.4373
CT-T440	0.4128	CT-T419	0.2033	CT-T395	0.4500	CT-T183	0.2289	CT-T178	0.6253	CT-T183	0.4305
CT-T197	0.4091	CT-T183	0.1996	CT-T410	0.4475	CT-T419	0.2276	CT-T175	0.6234	CT-T180	0.4287
CT-T193	0.4065	CT-T402	0.1865	CT-T193	0.4443	CT-T402	0.2273	CT-T193	0.6215	CT-T419	0.4260
CT-T398	0.3926	CT-T204	0.1826	CT-T398	0.4422	CT-T204	0.1984	CT-T392	0.6209	CT-T177	0.4211
CT-T205	0.3886	CT-T203	0.1687	CT-T426	0.4337	CT-T203	0.1927	CT-T422	0.6190	CT-T402	0.4095
CT-T426	0.3856	CT-T188	0.1630	CT-T205	0.4252	CT-T188	0.1810	CT-T440	0.6178	CT-T203	0.3954
CT-T421	0.3757	CT-T196	0.1519	CT-T396	0.4179	CT-T196	0.1702	CT-T451	0.6166	CT-T188	0.3616
CT-T388	0.3674	CT-T190	0.1433	CT-T428	0.4123	CT-T180	0.1686	CT-T208	0.6146	CT-T196	0.3521
CT-T184	0.3642	CT-T180	0.1432	CT-T421	0.4034	CT-T190	0.1661	CT-T429	0.6134	CT-T204	0.3111
CT-T396	0.3614			CT-T388	0.3982			CT-T395	0.6128		
CT-T428	0.3515			CT-T184	0.3891			CT-T374	0.6116		
CT-T441	0.3435			CT-T451	0.3752			CT-T184	0.6082		
CT-T451	0.3373			CT-T441	0.3744			CT-T386	0.6047		
CT-T424	0.3280			CT-T181	0.3701			CT-T410	0.5992		

Table 34. Performance averaged across JA runs for each topic, using the *pseudo-qrels*. For example, "JA-T107" denotes the topic ACLIA-JA-T107. The topics are sorted by the average performance.

	AP		AP		Q		Q		nDCG		nDCG
JA-T107	0.7535	JA-T19	0.4106	JA-T107	0.7874	JA-T301	0.4428	JA-T107	0.8948	JA-T18	0.6030
JA-T6	0.7309	JA-T7	0.4103	JA-T6	0.7358	JA-T221	0.4428	JA-T217	0.8037	JA-T295	0.6015
JA-T29	0.7058	JA-T304	0.4067	JA-T29	0.7127	JA-T32	0.4427	JA-T271	0.8020	JA-T25	0.6000
JA-T271	0.6760	JA-T301	0.4052	JA-T271	0.7043	JA-T19	0.4423	JA-T253	0.7826	JA-T245	0.5995
JA-T253	0.6651	JA-T15	0.3912	JA-T253	0.6930	JA-T7	0.4410	JA-T6	0.7605	JA-T2	0.5981
JA-T111	0.6376	JA-T153	0.3901	JA-T109	0.6572	JA-T15	0.4334	JA-T111	0.7568	JA-T115	0.5978
JA-T109	0.6370	JA-T221	0.3887	JA-T111	0.6569	JA-T242	0.4304	JA-T20	0.7561	JA-T304	0.5972
JA-T38	0.6236	JA-T242	0.3864	JA-T38	0.6484	JA-T154	0.4251	JA-T29	0.7511	JA-T236	0.5940
JA-T20	0.6121	JA-T291	0.3847	JA-T20	0.6367	JA-T153	0.4228	JA-T109	0.7463	JA-T215	0.5896
JA-T245	0.5978	JA-T154	0.3812	JA-T217	0.6280	JA-T25	0.4222	JA-T38	0.7417	JA-T19	0.5857
JA-T130	0.5851	JA-T25	0.3792	JA-T249	0.6107	JA-T291	0.4209	JA-T249	0.7301	JA-T152	0.5786
JA-T217	0.5849	JA-T162	0.3719	JA-T276	0.6071	JA-T162	0.4152	JA-T234	0.7261	JA-T106	0.5780
JA-T276	0.5801	JA-T236	0.3630	JA-T234	0.6067	JA-T250	0.3961	JA-T275	0.7210	JA-T112	0.5776
JA-T249	0.5796	JA-T112	0.3499	JA-T10	0.6035	JA-T236	0.3956	JA-T10	0.7053	JA-T32	0.5754
JA-T10	0.5788	JA-T138	0.3491	JA-T130	0.6034	JA-T112	0.3930	JA-T233	0.7042	JA-T140	0.5696
JA-T234	0.5751	JA-T37	0.3475	JA-T245	0.5988	JA-T106	0.3873	JA-T276	0.6935	JA-T300	0.5581
JA-T239	0.5648	JA-T106	0.3457	JA-T239	0.5877	JA-T138	0.3867	JA-T149	0.6890	JA-T9	0.5565
JA-T305	0.5535	JA-T250	0.3381	JA-T305	0.5725	JA-T295	0.3840	JA-T230	0.6880	JA-T37	0.5544
JA-T1	0.5208	JA-T300	0.3309	JA-T275	0.5596	JA-T37	0.3759	JA-T239	0.6797	JA-T231	0.5515
JA-T275	0.5149	JA-T295	0.3307	JA-T1	0.5484	JA-T300	0.3660	JA-T134	0.6787	JA-T7	0.5513
JA-T35	0.5141	JA-T9	0.3221	JA-T35	0.5445	JA-T140	0.3623	JA-T1	0.6718	JA-T153	0.5485
JA-T313	0.5098	JA-T170	0.3202	JA-T230	0.5391	JA-T113	0.3584	JA-T266	0.6707	JA-T4	0.5431
JA-T230	0.5065	JA-T152	0.3184	JA-T108	0.5300	JA-T9	0.3579	JA-T254	0.6707	JA-T222	0.5421
JA-T108	0.4979	JA-T167	0.3167	JA-T266	0.5254	JA-T152	0.3545	JA-T128	0.6638	JA-T138	0.5394
JA-T266	0.4898	JA-T222	0.3151	JA-T110	0.5253	JA-T222	0.3541	JA-T161	0.6614	JA-T113	0.5375
JA-T110	0.4831	JA-T140	0.3121	JA-T313	0.5236	JA-T231	0.3536	JA-T110	0.6603	JA-T237	0.5203
JA-T267	0.4827	JA-T148	0.3117	JA-T134	0.5230	JA-T4	0.3506	JA-T305	0.6593	JA-T3	0.5179
JA-T134	0.4820	JA-T231	0.3090	JA-T149	0.5210	JA-T13	0.3500	JA-T223	0.6531	JA-T248	0.5174
JA-T244	0.4805	JA-T13	0.3087	JA-T267	0.5190	JA-T170	0.3457	JA-T141	0.6527	JA-T13	0.5166
JA-T115	0.4803	JA-T4	0.3069	JA-T233	0.5178	JA-T167	0.3452	JA-T130	0.6499	JA-T238	0.5137
JA-T17	0.4791	JA-T113	0.2961	JA-T254	0.5132	JA-T148	0.3382	JA-T108	0.6472	JA-T297	0.5086
JA-T18	0.4785	JA-T158	0.2893	JA-T115	0.5119	JA-T3	0.3310	JA-T267	0.6454	JA-T225	0.5065
JA-T149	0.4778	JA-T3	0.2871	JA-T17	0.5115	JA-T164	0.3174	JA-T154	0.6435	JA-T170	0.5064
JA-T254	0.4772	JA-T164	0.2824	JA-T244	0.5113	JA-T158	0.3025	JA-T105	0.6399	JA-T167	0.4986
JA-T218	0.4723	JA-T166	0.2629	JA-T141	0.5103	JA-T224	0.3007	JA-T301	0.6397	JA-T164	0.4700
JA-T233	0.4685	JA-T163	0.2617	JA-T218	0.5063	JA-T238	0.2974	JA-T151	0.6343	JA-T148	0.4566
JA-T128	0.4678	JA-T238	0.2597	JA-T128	0.5001	JA-T166	0.2897	JA-T242	0.6324	JA-T224	0.4553
JA-T141	0.4668	JA-T224	0.2562	JA-T18	0.4982	JA-T297	0.2862	JA-T17	0.6300	JA-T163	0.4546
JA-T240	0.4636	JA-T297	0.2535	JA-T105	0.4963	JA-T163	0.2847	JA-T221	0.6285	JA-T166	0.4485
JA-T105	0.4617	JA-T225	0.2490	JA-T240	0.4899	JA-T225	0.2826	JA-T35	0.6277	JA-T255	0.4386
JA-T248	0.4572	JA-T165	0.1943	JA-T223	0.4829	JA-T255	0.2178	JA-T119	0.6261	JA-T315	0.4302
JA-T119	0.4500	JA-T255	0.1834	JA-T119	0.4814	JA-T165	0.2153	JA-T240	0.6218	JA-T165	0.3980
JA-T2	0.4484	JA-T315	0.1741	JA-T248	0.4721	JA-T315	0.2048	JA-T155	0.6213	JA-T137	0.3759
JA-T223	0.4464	JA-T168	0.1453	JA-T151	0.4719	JA-T160	0.1671	JA-T218	0.6139	JA-T160	0.3653
JA-T237	0.4359	JA-T284	0.1446	JA-T2	0.4555	JA-T284	0.1669	JA-T244	0.6135	JA-T158	0.3425
JA-T151	0.4295	JA-T160	0.1372	JA-T215	0.4552	JA-T168	0.1556	JA-T15	0.6129	JA-T157	0.3190
JA-T215	0.4238	JA-T137	0.1338	JA-T161	0.4523	JA-T137	0.1512	JA-T313	0.6123	JA-T168	0.3079
JA-T155	0.4134	JA-T157	0.1115	JA-T304	0.4478	JA-T157	0.1296	JA-T162	0.6111	JA-T284	0.3013
JA-T32	0.4130			JA-T237	0.4474			JA-T291	0.6084		
JA-T161	0.4130			JA-T155	0.4462			JA-T250	0.6045		

Table 35. Number of judged nonrelevant ($L0$) and judged relevant ($L1$ and $L2$) documents: 97 CS topics.

	$L0$	$L1$	$L2$	#relevant	#judged		$L0$	$L1$	$L2$	#relevant	#judged
CS-T41	118	48	64	112	230	CS-T101	166	0	19	19	185
CS-T42	317	112	144	256	573	CS-T102	108	15	4	19	127
CS-T43	67	61	128	189	256	CS-T103	47	29	163	192	239
CS-T44	242	88	114	202	444	CS-T104	90	33	17	50	140
CS-T46	110	97	176	273	383	CS-T317	195	4	9	13	208
CS-T47	218	132	45	177	395	CS-T320	88	30	37	67	155
CS-T48	379	15	31	46	425	CS-T321	92	40	17	57	149
CS-T49	83	25	144	169	252	CS-T322	131	188	8	196	327
CS-T52	130	59	13	72	202	CS-T323	77	69	16	85	162
CS-T53	407	3	4	7	414	CS-T324	24	107	16	123	147
CS-T54	324	11	4	15	339	CS-T325	17	77	161	238	255
CS-T55	187	112	21	133	320	CS-T326	137	20	19	39	176
CS-T56	495	9	72	81	576	CS-T328	53	79	65	144	197
CS-T57	199	106	103	209	408	CS-T329	70	18	10	28	98
CS-T58	90	37	45	82	172	CS-T332	84	16	123	139	223
CS-T60	71	29	89	118	189	CS-T333	236	88	171	259	495
CS-T61	155	62	60	122	277	CS-T334	124	49	240	289	413
CS-T62	514	15	12	27	541	CS-T336	73	9	112	121	194
CS-T64	176	103	9	112	288	CS-T337	51	18	116	134	185
CS-T65	115	63	1	64	179	CS-T338	18	17	66	83	101
CS-T67	233	11	74	85	318	CS-T339	19	22	118	140	159
CS-T68	152	34	3	37	189	CS-T340	20	13	70	83	103
CS-T69	178	42	3	45	223	CS-T347	101	16	34	50	151
CS-T71	246	6	23	29	275	CS-T348	8	28	85	113	121
CS-T73	602	15	62	77	679	CS-T349	123	16	11	27	150
CS-T74	487	29	180	209	696	CS-T350	190	24	14	38	228
CS-T75	292	5	14	19	311	CS-T351	13	72	30	102	115
CS-T76	146	16	35	51	197	CS-T352	14	19	97	116	130
CS-T77	386	21	25	46	432	CS-T355	85	17	86	103	188
CS-T78	525	5	4	9	534	CS-T357	143	23	85	108	251
CS-T79	265	15	11	26	291	CS-T358	161	25	56	81	242
CS-T80	434	2	5	7	441	CS-T359	110	26	175	201	311
CS-T81	60	24	106	130	190	CS-T361	27	58	21	79	106
CS-T82	284	74	86	160	444	CS-T365	187	0	7	7	194
CS-T83	72	165	87	252	324	CS-T366	178	17	11	28	206
CS-T84	48	72	21	93	141	CS-T367	34	24	154	178	212
CS-T85	18	59	43	102	120	CS-T368	167	44	48	92	259
CS-T87	159	104	141	245	404	CS-T369	131	106	14	120	251
CS-T89	155	16	4	20	175	CS-T370	35	37	13	50	85
CS-T90	108	52	100	152	260	CS-T376	108	115	1	116	224
CS-T91	47	97	125	222	269	CS-T378	83	7	12	19	102
CS-T92	120	57	39	96	216	CS-T379	126	89	1	90	216
CS-T93	97	22	38	60	157	CS-T380	152	45	4	49	201
CS-T94	237	33	26	59	296	CS-T381	122	2	8	10	132
CS-T95	197	44	51	95	292	CS-T383	263	1	15	16	279
CS-T96	68	45	28	73	141	CS-T384	129	2	15	17	146
CS-T97	60	37	49	86	146	CS-T385	227	22	46	68	295
CS-T98	62	26	40	66	128						
CS-T99	182	20	13	33	215						
CS-T100	89	26	16	42	131						
						total	15243	4137	5351	9488	24731

Table 36. Number of judged nonrelevant ($L0$) and judged relevant ($L1$ and $L2$) documents: 95 CT topics.

	$L0$	$L1$	$L2$	#relevant	#judged		$L0$	$L1$	$L2$	#relevant	#judged
CT-T171	97	52	96	148	245	CT-T400	85	94	39	133	218
CT-T172	179	16	32	48	227	CT-T402	208	61	8	69	277
CT-T174	154	15	55	70	224	CT-T404	46	154	12	166	212
CT-T175	82	13	58	71	153	CT-T405	131	37	5	42	173
CT-T176	149	12	124	136	285	CT-T407	71	21	15	36	107
CT-T177	202	35	34	69	271	CT-T408	24	53	20	73	97
CT-T178	115	43	20	63	178	CT-T409	80	30	5	35	115
CT-T179	116	6	43	49	165	CT-T410	187	6	7	13	200
CT-T180	186	24	36	60	246	CT-T411	106	126	31	157	263
CT-T181	53	10	131	141	194	CT-T413	171	0	6	6	177
CT-T182	116	29	45	74	190	CT-T414	232	20	3	23	255
CT-T183	180	36	50	86	266	CT-T415	276	0	11	11	287
CT-T184	136	8	15	23	159	CT-T416	232	16	31	47	279
CT-T186	171	14	10	24	195	CT-T418	149	15	7	22	171
CT-T187	136	0	12	12	148	CT-T419	257	13	14	27	284
CT-T188	317	16	10	26	343	CT-T420	145	3	67	70	215
CT-T189	123	30	7	37	160	CT-T421	188	3	2	5	193
CT-T190	239	8	9	17	256	CT-T422	205	1	21	22	227
CT-T191	150	4	2	6	156	CT-T423	163	0	33	33	196
CT-T193	122	3	36	39	161	CT-T424	194	24	15	39	233
CT-T194	11	121	27	148	159	CT-T426	86	19	47	66	152
CT-T195	83	42	51	93	176	CT-T427	95	6	66	72	167
CT-T196	271	23	5	28	299	CT-T428	31	5	73	78	109
CT-T197	105	56	14	70	175	CT-T429	118	4	18	22	140
CT-T198	184	27	0	27	211	CT-T430	191	14	9	23	214
CT-T200	46	134	14	148	194	CT-T431	136	0	20	20	156
CT-T203	274	52	3	55	329	CT-T432	182	2	15	17	199
CT-T204	286	19	9	28	314	CT-T434	107	10	14	24	131
CT-T205	218	4	1	5	223	CT-T435	177	45	11	56	233
CT-T206	276	18	2	20	296	CT-T436	234	0	9	9	243
CT-T207	177	22	11	33	210	CT-T437	191	24	35	59	250
CT-T208	179	35	10	45	224	CT-T438	88	75	36	111	199
CT-T210	105	5	12	17	122	CT-T439	101	12	18	30	131
CT-T211	134	19	13	32	166	CT-T440	108	20	17	37	145
CT-T213	61	2	81	83	144	CT-T441	158	29	8	37	195
CT-T374	13	148	21	169	182	CT-T442	164	33	39	72	236
CT-T386	107	52	24	76	183	CT-T443	184	25	35	60	244
CT-T387	50	123	10	133	183	CT-T444	141	41	8	49	190
CT-T388	304	18	2	20	324	CT-T445	97	46	40	86	183
CT-T389	191	1	27	28	219	CT-T446	232	1	26	27	259
CT-T390	117	81	19	100	217	CT-T447	159	32	44	76	235
CT-T391	119	35	6	41	160	CT-T448	188	19	49	68	256
CT-T392	178	0	52	52	230	CT-T449	121	3	19	22	143
CT-T393	136	19	26	45	181	CT-T450	218	28	47	75	293
CT-T394	66	56	3	59	125	CT-T451	162	1	28	29	191
CT-T395	378	3	3	6	384						
CT-T396	49	106	10	116	165						
CT-T397	197	24	6	30	227						
CT-T398	79	80	15	95	174						
CT-T399	160	3	4	7	167						
						total	14396	2873	2389	5262	19658

Table 37. Number of judged nonrelevant ($L0$) and judged relevant ($L1$ and $L2$) documents: 98 JA topics.

	$L0$	$L1$	$L2$	#relevant	#judged		$L0$	$L1$	$L2$	#relevant	#judged
JA-T1	989	6	6	12	1001	JA-T163	111	147	30	177	288
JA-T2	1007	4	2	6	1013	JA-T164	147	75	31	106	253
JA-T3	461	153	129	282	743	JA-T165	306	24	6	30	336
JA-T4	661	65	74	139	800	JA-T166	759	26	95	121	880
JA-T6	811	6	4	10	821	JA-T167	602	59	59	118	720
JA-T7	739	31	34	65	804	JA-T168	1005	62	25	87	1092
JA-T9	287	181	156	337	624	JA-T170	689	32	75	107	796
JA-T10	1075	8	13	21	1096	JA-T215	847	31	15	46	893
JA-T13	771	14	21	35	806	JA-T217	340	40	33	73	413
JA-T15	608	104	5	109	717	JA-T218	660	58	44	102	762
JA-T17	582	56	24	80	662	JA-T221	286	92	55	147	433
JA-T18	791	45	2	47	838	JA-T222	668	30	54	84	752
JA-T19	649	225	9	234	883	JA-T223	584	50	61	111	695
JA-T20	847	11	10	21	868	JA-T224	578	19	75	94	672
JA-T25	196	150	1	151	347	JA-T225	540	6	357	363	903
JA-T29	432	15	9	24	456	JA-T230	364	116	236	352	716
JA-T32	419	85	2	87	506	JA-T231	544	84	69	153	697
JA-T35	201	44	14	58	259	JA-T233	324	137	44	181	505
JA-T37	293	18	6	24	317	JA-T234	630	2	14	16	646
JA-T38	176	12	10	22	198	JA-T236	82	9	222	231	313
JA-T105	237	49	2	51	288	JA-T237	292	1	8	9	301
JA-T106	108	52	20	72	180	JA-T238	136	1	141	142	278
JA-T107	138	13	7	20	158	JA-T239	201	3	152	155	356
JA-T108	223	63	7	70	293	JA-T240	244	2	48	50	294
JA-T109	221	18	5	23	244	JA-T242	155	2	151	153	308
JA-T110	878	3	2	5	883	JA-T244	220	1	16	17	237
JA-T111	630	8	0	8	638	JA-T245	190	0	23	23	213
JA-T112	517	23	34	57	574	JA-T248	147	0	83	83	230
JA-T113	533	24	11	35	568	JA-T249	198	0	26	26	224
JA-T115	641	5	9	14	655	JA-T250	126	1	64	65	191
JA-T119	340	12	25	37	377	JA-T253	129	22	20	42	171
JA-T128	227	2	5	7	234	JA-T254	126	2	95	97	223
JA-T130	219	3	19	22	241	JA-T255	107	107	65	172	279
JA-T134	150	15	44	59	209	JA-T266	329	217	9	226	555
JA-T137	292	10	28	38	330	JA-T267	483	148	11	159	642
JA-T138	171	22	26	48	219	JA-T271	344	84	34	118	462
JA-T140	136	15	31	46	182	JA-T275	516	11	114	125	641
JA-T141	169	11	4	15	184	JA-T276	484	0	10	10	494
JA-T148	680	99	37	136	816	JA-T284	289	7	44	51	340
JA-T149	500	33	43	76	576	JA-T291	167	40	11	51	218
JA-T151	612	31	20	51	663	JA-T295	158	29	6	35	193
JA-T152	563	62	120	182	745	JA-T297	241	25	11	36	277
JA-T153	696	17	12	29	725	JA-T300	144	69	21	90	234
JA-T154	772	22	26	48	820	JA-T301	189	28	3	31	220
JA-T155	581	46	27	73	654	JA-T304	190	25	3	28	218
JA-T157	544	123	61	184	728	JA-T305	212	13	12	25	237
JA-T158	478	225	43	268	746	JA-T313	280	7	5	12	292
JA-T160	272	126	33	159	431	JA-T315	281	18	12	30	311
JA-T161	140	71	26	97	237						
JA-T162	166	15	37	52	218						
						total	40473	4413	4093	8506	48979

Table 38. Performance averaged across CS runs for each topic, using the real qrels. For example, "CS-T349" denotes the topic ACLIA-CS-T349. The topics are sorted by the average performance.

	AP		AP		Q		Q		nDCG		nDCG
CS-T80	0.7582	CS-T317	0.4617	CS-T370	0.7251	CS-T41	0.4784	CS-T340	0.8344	CS-T340	0.8344
CS-T370	0.7534	CS-T41	0.4595	CS-T340	0.7233	CS-T55	0.4721	CS-T329	0.8326	CS-T329	0.8326
CS-T340	0.7061	CS-T369	0.4547	CS-T80	0.7096	CS-T367	0.4705	CS-T370	0.8285	CS-T370	0.8285
CS-T338	0.6840	CS-T101	0.4331	CS-T338	0.6939	CS-T317	0.4682	CS-T68	0.8207	CS-T68	0.8207
CS-T361	0.6827	CS-T366	0.4325	CS-T348	0.6828	CS-T369	0.4561	CS-T338	0.8173	CS-T338	0.8173
CS-T348	0.6777	CS-T92	0.4316	CS-T329	0.6698	CS-T99	0.4417	CS-T348	0.8122	CS-T348	0.8122
CS-T85	0.6657	CS-T367	0.4272	CS-T68	0.6675	CS-T92	0.4379	CS-T381	0.8081	CS-T381	0.8081
CS-T68	0.6629	CS-T99	0.4270	CS-T81	0.6642	CS-T376	0.4329	CS-T81	0.8048	CS-T81	0.8048
CS-T329	0.6604	CS-T91	0.4162	CS-T381	0.6525	CS-T91	0.4289	CS-T60	0.7899	CS-T60	0.7899
CS-T351	0.6582	CS-T325	0.4080	CS-T60	0.6410	CS-T325	0.4287	CS-T71	0.7893	CS-T71	0.7893
CS-T81	0.6580	CS-T94	0.4078	CS-T336	0.6402	CS-T102	0.4188	CS-T98	0.7889	CS-T98	0.7889
CS-T60	0.6395	CS-T102	0.4073	CS-T361	0.6333	CS-T90	0.4162	CS-T336	0.7840	CS-T336	0.7840
CS-T69	0.6271	CS-T376	0.4057	CS-T43	0.6321	CS-T94	0.4127	CS-T320	0.7837	CS-T320	0.7837
CS-T43	0.6166	CS-T90	0.3983	CS-T85	0.6312	CS-T366	0.4121	CS-T93	0.7825	CS-T93	0.7825
CS-T98	0.6143	CS-T350	0.3881	CS-T351	0.6288	CS-T379	0.3955	CS-T378	0.7821	CS-T378	0.7821
CS-T336	0.6104	CS-T95	0.3826	CS-T378	0.6213	CS-T350	0.3939	CS-T97	0.7816	CS-T97	0.7816
CS-T349	0.6100	CS-T379	0.3706	CS-T71	0.6168	CS-T95	0.3864	CS-T43	0.7815	CS-T43	0.7815
CS-T323	0.6084	CS-T385	0.3694	CS-T98	0.6156	CS-T385	0.3758	CS-T351	0.7790	CS-T351	0.7790
CS-T320	0.6032	CS-T383	0.3472	CS-T320	0.6108	CS-T383	0.3711	CS-T80	0.7757	CS-T80	0.7757
CS-T321	0.5961	CS-T46	0.3468	CS-T339	0.6030	CS-T46	0.3648	CS-T85	0.7739	CS-T85	0.7739
CS-T100	0.5946	CS-T332	0.3336	CS-T352	0.6013	CS-T332	0.3602	CS-T352	0.7738	CS-T352	0.7738
CS-T65	0.5940	CS-T380	0.3311	CS-T69	0.6006	CS-T380	0.3492	CS-T339	0.7718	CS-T339	0.7718
CS-T381	0.5933	CS-T368	0.3281	CS-T97	0.5988	CS-T368	0.3447	CS-T52	0.7701	CS-T52	0.7701
CS-T97	0.5924	CS-T357	0.3046	CS-T65	0.5982	CS-T357	0.3420	CS-T361	0.7673	CS-T361	0.7673
CS-T52	0.5890	CS-T322	0.2940	CS-T93	0.5930	CS-T358	0.3254	CS-T100	0.7620	CS-T100	0.7620
CS-T378	0.5882	CS-T83	0.2894	CS-T52	0.5902	CS-T67	0.3185	CS-T384	0.7552	CS-T384	0.7552
CS-T71	0.5859	CS-T64	0.2885	CS-T100	0.5867	CS-T322	0.3062	CS-T75	0.7527	CS-T75	0.7527
CS-T93	0.5782	CS-T358	0.2877	CS-T321	0.5734	CS-T57	0.3046	CS-T58	0.7526	CS-T58	0.7526
CS-T53	0.5770	CS-T67	0.2868	CS-T323	0.5728	CS-T64	0.2992	CS-T65	0.7483	CS-T65	0.7483
CS-T339	0.5767	CS-T57	0.2844	CS-T75	0.5720	CS-T83	0.2869	CS-T355	0.7435	CS-T355	0.7435
CS-T352	0.5740	CS-T333	0.2675	CS-T365	0.5712	CS-T79	0.2804	CS-T326	0.7434	CS-T326	0.7434
CS-T58	0.5570	CS-T47	0.2614	CS-T58	0.5661	CS-T333	0.2651	CS-T84	0.7428	CS-T84	0.7428
CS-T324	0.5522	CS-T79	0.2520	CS-T53	0.5596	CS-T47	0.2536	CS-T324	0.7408	CS-T324	0.7408
CS-T96	0.5483	CS-T44	0.2145	CS-T49	0.5570	CS-T359	0.2269	CS-T49	0.7363	CS-T49	0.7363
CS-T75	0.5378	CS-T334	0.2052	CS-T324	0.5427	CS-T54	0.2185	CS-T96	0.7278	CS-T96	0.7278
CS-T365	0.5341	CS-T82	0.2040	CS-T96	0.5315	CS-T44	0.2183	CS-T103	0.7235	CS-T103	0.7235
CS-T49	0.5297	CS-T87	0.2030	CS-T355	0.5306	CS-T82	0.2147	CS-T328	0.7210	CS-T328	0.7210
CS-T347	0.5154	CS-T359	0.1987	CS-T384	0.5247	CS-T334	0.2127	CS-T41	0.7171	CS-T41	0.7171
CS-T84	0.5146	CS-T77	0.1986	CS-T84	0.5145	CS-T87	0.2099	CS-T323	0.7131	CS-T323	0.7131
CS-T89	0.5142	CS-T54	0.1884	CS-T326	0.5106	CS-T48	0.1991	CS-T89	0.7112	CS-T89	0.7112
CS-T104	0.5115	CS-T48	0.1670	CS-T103	0.5100	CS-T77	0.1971	CS-T76	0.7106	CS-T76	0.7106
CS-T55	0.5026	CS-T42	0.1668	CS-T347	0.5085	CS-T56	0.1933	CS-T61	0.7071	CS-T61	0.7071
CS-T76	0.4927	CS-T56	0.1666	CS-T328	0.5049	CS-T42	0.1565	CS-T347	0.7001	CS-T347	0.7001
CS-T355	0.4920	CS-T62	0.1470	CS-T104	0.5044	CS-T62	0.1483	CS-T69	0.6994	CS-T69	0.6994
CS-T328	0.4913	CS-T73	0.1075	CS-T349	0.4989	CS-T73	0.1283	CS-T104	0.6969	CS-T104	0.6969
CS-T326	0.4839	CS-T74	0.1072	CS-T89	0.4987	CS-T74	0.1158	CS-T367	0.6961	CS-T367	0.6961
CS-T384	0.4733	CS-T78	0.0857	CS-T76	0.4954	CS-T78	0.1146	CS-T365	0.6958	CS-T365	0.6958
CS-T103	0.4693			CS-T337	0.4868			CS-T376	0.6859		
CS-T61	0.4672			CS-T61	0.4816			CS-T337	0.6859		
CS-T337	0.4631			CS-T101	0.4804			CS-T101	0.6815		

Table 39. Performance averaged across CT runs for each topic, using the real qrels. For example, "CT-T210" denotes the topic ACLIA-CT-T210. The topics are sorted by the average performance.

	AP		AP		Q		Q		nDCG		nDCG
CT-T408	0.7484	CT-T404	0.3544	CT-T408	0.7163	CT-T390	0.3836	CT-T428	0.8920	CT-T207	0.6052
CT-T396	0.6526	CT-T176	0.3518	CT-T431	0.6926	CT-T413	0.3757	CT-T408	0.8523	CT-T178	0.6037
CT-T431	0.6462	CT-T402	0.3478	CT-T428	0.6891	CT-T193	0.3742	CT-T431	0.8312	CT-T174	0.6032
CT-T427	0.6442	CT-T179	0.3465	CT-T427	0.6636	CT-T410	0.3633	CT-T194	0.8299	CT-T193	0.5965
CT-T428	0.6415	CT-T434	0.3435	CT-T432	0.6608	CT-T404	0.3625	CT-T213	0.8199	CT-T413	0.5908
CT-T197	0.6345	CT-T210	0.3397	CT-T396	0.6412	CT-T402	0.3608	CT-T398	0.8026	CT-T443	0.5813
CT-T398	0.6325	CT-T410	0.3396	CT-T213	0.6323	CT-T182	0.3604	CT-T200	0.8022	CT-T176	0.5809
CT-T432	0.6237	CT-T193	0.3360	CT-T398	0.6295	CT-T186	0.3577	CT-T396	0.8013	CT-T409	0.5746
CT-T394	0.6189	CT-T413	0.3351	CT-T197	0.6284	CT-T208	0.3455	CT-T432	0.7947	CT-T208	0.5729
CT-T213	0.5882	CT-T418	0.2945	CT-T394	0.6187	CT-T191	0.3334	CT-T181	0.7947	CT-T184	0.5589
CT-T387	0.5820	CT-T411	0.2847	CT-T194	0.5892	CT-T420	0.3235	CT-T374	0.7920	CT-T402	0.5582
CT-T194	0.5767	CT-T183	0.2793	CT-T395	0.5885	CT-T174	0.3141	CT-T427	0.7874	CT-T390	0.5573
CT-T195	0.5660	CT-T191	0.2788	CT-T387	0.5787	CT-T418	0.3109	CT-T394	0.7856	CT-T418	0.5533
CT-T395	0.5624	CT-T206	0.2760	CT-T187	0.5515	CT-T409	0.3022	CT-T197	0.7855	CT-T388	0.5480
CT-T374	0.5534	CT-T437	0.2715	CT-T407	0.5514	CT-T206	0.3018	CT-T426	0.7655	CT-T429	0.5461
CT-T393	0.5385	CT-T174	0.2694	CT-T392	0.5508	CT-T183	0.2958	CT-T423	0.7608	CT-T410	0.5293
CT-T200	0.5335	CT-T443	0.2643	CT-T200	0.5487	CT-T443	0.2948	CT-T407	0.7589	CT-T191	0.5288
CT-T211	0.5320	CT-T420	0.2636	CT-T374	0.5477	CT-T178	0.2925	CT-T400	0.7551	CT-T186	0.5240
CT-T400	0.5288	CT-T409	0.2514	CT-T195	0.5465	CT-T437	0.2850	CT-T175	0.7537	CT-T172	0.5189
CT-T386	0.5287	CT-T184	0.2507	CT-T449	0.5431	CT-T184	0.2799	CT-T387	0.7398	CT-T437	0.5166
CT-T391	0.5259	CT-T178	0.2403	CT-T389	0.5384	CT-T430	0.2641	CT-T449	0.7380	CT-T411	0.5058
CT-T407	0.5211	CT-T450	0.2354	CT-T423	0.5345	CT-T450	0.2626	CT-T389	0.7375	CT-T206	0.5053
CT-T392	0.5154	CT-T430	0.2265	CT-T393	0.5231	CT-T411	0.2617	CT-T395	0.7217	CT-T180	0.5000
CT-T449	0.5116	CT-T172	0.2247	CT-T400	0.5194	CT-T172	0.2592	CT-T187	0.7214	CT-T183	0.4764
CT-T440	0.5024	CT-T448	0.1995	CT-T181	0.5174	CT-T429	0.2515	CT-T451	0.7168	CT-T448	0.4756
CT-T389	0.4972	CT-T180	0.1938	CT-T422	0.5165	CT-T448	0.2213	CT-T447	0.7071	CT-T450	0.4742
CT-T187	0.4873	CT-T205	0.1850	CT-T386	0.5144	CT-T180	0.2188	CT-T393	0.7040	CT-T414	0.4723
CT-T422	0.4814	CT-T429	0.1789	CT-T440	0.5119	CT-T446	0.2017	CT-T386	0.6939	CT-T435	0.4694
CT-T423	0.4811	CT-T444	0.1758	CT-T211	0.5092	CT-T419	0.1953	CT-T445	0.6925	CT-T446	0.4690
CT-T181	0.4621	CT-T435	0.1704	CT-T391	0.4999	CT-T414	0.1914	CT-T436	0.6893	CT-T430	0.4625
CT-T426	0.4528	CT-T419	0.1657	CT-T426	0.4914	CT-T205	0.1860	CT-T422	0.6875	CT-T424	0.4623
CT-T399	0.4412	CT-T414	0.1641	CT-T436	0.4892	CT-T435	0.1853	CT-T392	0.6857	CT-T198	0.4231
CT-T207	0.4343	CT-T446	0.1475	CT-T175	0.4709	CT-T444	0.1827	CT-T210	0.6818	CT-T444	0.4165
CT-T439	0.4315	CT-T424	0.1435	CT-T451	0.4665	CT-T198	0.1682	CT-T434	0.6752	CT-T419	0.4133
CT-T436	0.4255	CT-T198	0.1422	CT-T399	0.4588	CT-T424	0.1652	CT-T440	0.6748	CT-T442	0.4015
CT-T390	0.4236	CT-T442	0.1255	CT-T447	0.4493	CT-T421	0.1624	CT-T420	0.6614	CT-T416	0.4013
CT-T175	0.4160	CT-T177	0.1192	CT-T439	0.4318	CT-T415	0.1525	CT-T189	0.6555	CT-T177	0.3792
CT-T451	0.4138	CT-T415	0.1179	CT-T207	0.4250	CT-T442	0.1364	CT-T404	0.6511	CT-T190	0.3644
CT-T189	0.4100	CT-T421	0.1088	CT-T189	0.4205	CT-T177	0.1316	CT-T439	0.6496	CT-T188	0.3626
CT-T438	0.4092	CT-T188	0.1075	CT-T388	0.4113	CT-T188	0.1255	CT-T211	0.6438	CT-T205	0.3620
CT-T447	0.4028	CT-T204	0.1040	CT-T405	0.4092	CT-T416	0.1213	CT-T405	0.6403	CT-T415	0.3403
CT-T405	0.3994	CT-T416	0.0912	CT-T210	0.4073	CT-T196	0.1141	CT-T195	0.6400	CT-T421	0.3358
CT-T445	0.3965	CT-T196	0.0886	CT-T397	0.4059	CT-T204	0.1133	CT-T438	0.6374	CT-T203	0.3276
CT-T388	0.3933	CT-T190	0.0814	CT-T434	0.4011	CT-T190	0.1062	CT-T179	0.6357	CT-T196	0.2946
CT-T397	0.3917	CT-T203	0.0769	CT-T445	0.3983	CT-T203	0.0977	CT-T399	0.6268	CT-T204	0.2468
CT-T171	0.3790			CT-T438	0.3983			CT-T441	0.6265		
CT-T441	0.3776			CT-T441	0.3965			CT-T182	0.6236		
CT-T186	0.3621			CT-T179	0.3910			CT-T171	0.6210		
CT-T208	0.3591			CT-T171	0.3905			CT-T397	0.6185		
CT-T182	0.3560			CT-T176	0.3897			CT-T391	0.6064		

Table 40. Performance averaged across JA runs for each topic, using the real qrels. For example, "JA-T107" denotes the topic ACLIA-JA-T107. The topics are sorted by the average performance.

	AP		AP		Q		Q		nDCG		nDCG
JA-T107	0.7557	JA-T1	0.4237	JA-T107	0.7467	JA-T113	0.4457	JA-T107	0.8704	JA-T151	0.6117
JA-T271	0.7466	JA-T113	0.4214	JA-T271	0.7411	JA-T7	0.4440	JA-T217	0.8626	JA-T224	0.6110
JA-T221	0.7466	JA-T110	0.4184	JA-T242	0.7407	JA-T313	0.4433	JA-T271	0.8286	JA-T291	0.6108
JA-T6	0.7309	JA-T223	0.4180	JA-T221	0.6866	JA-T110	0.4324	JA-T242	0.8275	JA-T244	0.6089
JA-T242	0.7216	JA-T153	0.4152	JA-T250	0.6839	JA-T112	0.4318	JA-T250	0.7922	JA-T304	0.6065
JA-T9	0.6950	JA-T112	0.4143	JA-T217	0.6804	JA-T230	0.4198	JA-T134	0.7840	JA-T222	0.6005
JA-T38	0.6884	JA-T162	0.4061	JA-T29	0.6699	JA-T2	0.4198	JA-T221	0.7826	JA-T115	0.5998
JA-T29	0.6643	JA-T164	0.3952	JA-T111	0.6676	JA-T153	0.4120	JA-T253	0.7756	JA-T110	0.5963
JA-T253	0.6603	JA-T239	0.3951	JA-T253	0.6641	JA-T151	0.4018	JA-T25	0.7740	JA-T225	0.5938
JA-T111	0.6509	JA-T230	0.3945	JA-T9	0.6571	JA-T161	0.4012	JA-T254	0.7709	JA-T17	0.5898
JA-T217	0.6464	JA-T161	0.3941	JA-T254	0.6498	JA-T17	0.3972	JA-T9	0.7698	JA-T245	0.5858
JA-T250	0.6454	JA-T17	0.3905	JA-T6	0.6481	JA-T305	0.3908	JA-T275	0.7628	JA-T313	0.5853
JA-T149	0.6405	JA-T151	0.3871	JA-T38	0.6411	JA-T222	0.3862	JA-T3	0.7595	JA-T119	0.5813
JA-T267	0.6400	JA-T163	0.3697	JA-T134	0.6388	JA-T225	0.3815	JA-T149	0.7557	JA-T4	0.5765
JA-T141	0.6231	JA-T4	0.3693	JA-T275	0.6314	JA-T164	0.3780	JA-T111	0.7555	JA-T295	0.5730
JA-T275	0.6173	JA-T222	0.3659	JA-T267	0.6267	JA-T13	0.3771	JA-T29	0.7387	JA-T154	0.5729
JA-T254	0.6172	JA-T225	0.3604	JA-T149	0.6215	JA-T167	0.3758	JA-T233	0.7370	JA-T7	0.5715
JA-T233	0.6071	JA-T148	0.3528	JA-T141	0.6177	JA-T291	0.3671	JA-T141	0.7340	JA-T2	0.5640
JA-T134	0.6021	JA-T13	0.3518	JA-T25	0.5947	JA-T4	0.3659	JA-T267	0.7309	JA-T163	0.5621
JA-T130	0.5959	JA-T167	0.3488	JA-T233	0.5873	JA-T163	0.3601	JA-T38	0.7243	JA-T301	0.5545
JA-T10	0.5863	JA-T255	0.3402	JA-T130	0.5855	JA-T300	0.3595	JA-T238	0.7203	JA-T236	0.5541
JA-T245	0.5661	JA-T291	0.3363	JA-T245	0.5709	JA-T148	0.3509	JA-T239	0.6948	JA-T305	0.5491
JA-T218	0.5654	JA-T215	0.3302	JA-T3	0.5694	JA-T248	0.3497	JA-T223	0.6922	JA-T255	0.5486
JA-T3	0.5637	JA-T138	0.3188	JA-T218	0.5682	JA-T301	0.3434	JA-T231	0.6879	JA-T13	0.5439
JA-T25	0.5635	JA-T300	0.3174	JA-T10	0.5646	JA-T119	0.3364	JA-T152	0.6864	JA-T164	0.5401
JA-T106	0.5581	JA-T248	0.3150	JA-T19	0.5525	JA-T128	0.3363	JA-T6	0.6858	JA-T153	0.5363
JA-T19	0.5522	JA-T301	0.3121	JA-T276	0.5504	JA-T255	0.3340	JA-T15	0.6855	JA-T128	0.5218
JA-T35	0.5500	JA-T236	0.2999	JA-T234	0.5502	JA-T215	0.3265	JA-T234	0.6824	JA-T215	0.5157
JA-T304	0.5464	JA-T119	0.2892	JA-T249	0.5424	JA-T138	0.3263	JA-T218	0.6799	JA-T237	0.5116
JA-T276	0.5424	JA-T128	0.2799	JA-T106	0.5344	JA-T295	0.3216	JA-T162	0.6794	JA-T138	0.5088
JA-T249	0.5274	JA-T295	0.2759	JA-T35	0.5295	JA-T236	0.3202	JA-T106	0.6790	JA-T108	0.5060
JA-T234	0.5152	JA-T37	0.2749	JA-T304	0.5154	JA-T154	0.3154	JA-T266	0.6718	JA-T37	0.4906
JA-T115	0.5004	JA-T154	0.2736	JA-T115	0.5126	JA-T37	0.2991	JA-T10	0.6710	JA-T148	0.4881
JA-T2	0.4974	JA-T32	0.2573	JA-T244	0.5121	JA-T18	0.2815	JA-T109	0.6702	JA-T248	0.4819
JA-T244	0.4915	JA-T108	0.2570	JA-T105	0.5114	JA-T32	0.2718	JA-T19	0.6694	JA-T32	0.4810
JA-T15	0.4867	JA-T18	0.2418	JA-T238	0.5069	JA-T108	0.2711	JA-T249	0.6689	JA-T18	0.4737
JA-T105	0.4830	JA-T160	0.2179	JA-T15	0.5052	JA-T160	0.2164	JA-T230	0.6687	JA-T170	0.4376
JA-T140	0.4827	JA-T157	0.1985	JA-T231	0.4838	JA-T170	0.2154	JA-T105	0.6601	JA-T315	0.4336
JA-T313	0.4751	JA-T166	0.1889	JA-T140	0.4804	JA-T166	0.2042	JA-T161	0.6596	JA-T297	0.3949
JA-T20	0.4719	JA-T170	0.1867	JA-T109	0.4756	JA-T157	0.2040	JA-T300	0.6492	JA-T166	0.3919
JA-T109	0.4685	JA-T240	0.1735	JA-T266	0.4730	JA-T240	0.2039	JA-T140	0.6479	JA-T160	0.3915
JA-T266	0.4645	JA-T315	0.1488	JA-T224	0.4638	JA-T315	0.1804	JA-T1	0.6473	JA-T284	0.3782
JA-T231	0.4634	JA-T284	0.1450	JA-T1	0.4637	JA-T284	0.1642	JA-T155	0.6447	JA-T240	0.3749
JA-T155	0.4582	JA-T137	0.1312	JA-T239	0.4599	JA-T165	0.1535	JA-T130	0.6359	JA-T137	0.3742
JA-T224	0.4516	JA-T158	0.1290	JA-T20	0.4566	JA-T137	0.1463	JA-T35	0.6320	JA-T165	0.3689
JA-T238	0.4440	JA-T165	0.1286	JA-T223	0.4531	JA-T297	0.1416	JA-T276	0.6240	JA-T157	0.3630
JA-T237	0.4431	JA-T297	0.1208	JA-T237	0.4516	JA-T158	0.1189	JA-T167	0.6208	JA-T168	0.2818
JA-T7	0.4405	JA-T168	0.0971	JA-T162	0.4502	JA-T168	0.1052	JA-T112	0.6184	JA-T158	0.2383
JA-T305	0.4358			JA-T155	0.4473			JA-T20	0.6160		
JA-T152	0.4274			JA-T152	0.4462			JA-T113	0.6127		