NTCIR-7 ACLIA IR4QA Results based on Qrels Version 2

Tetsuya Sakai† Noriko Kando‡ Chuan-Jie Lin⋆
Ruihua Song† Hideki Shima∗ Teruko Mitamura∗
†Microsoft Research Asia ‡National Institute of Informatics
⋆National Taiwan Ocean University ∗Carnegie Mellon University
tetsuyasakai@acm.org

Abstract

This document is a postscript to the Overview of the NTCIR-7 ACLIA IR4QA Task [2]. At the NTCIR-7 Workshop Meeting (December 2008), participating systems of IR4QA were evaluated based on "qrels version 1," which covered the depth-30 pool for every topic and went further down the pool for a limited number of topics. Here, we report on revised results based on "qrels version 2" which covers the depth-100 pool for every topic. While the version 1 and version 2 results are generally in agreement, some differences in system rankings and significance test results suggest that the additional effort was worthwhile.

Keywords: *test collections, pooling, relevance assessments, evaluation metrics.*

1 Introduction

This document is a postscript to the Overview of the NTCIR-7 ACLIA IR4QA Task [2]. At the NTCIR-7 Workshop Meeting (December 2008), participating systems of IR4QA were evaluated based on "qrels version 1," which covered the depth-30 pool for every topic and went further down the pool for a limited number of topics. Here, we report on revised results based on "qrels version 2" which covers the depth-100 pool for every topic. To avoid confusion, we stick to the original IR4QA topics sets (97 CS, 95 CT and 98 JA) topics for evaluation (See Section 2 in the Overview), even though our additional relevance assessments did find a small number of relevant documents for a few of the "deleted" topics. As with qrels version 1, all evaluation metric values were computed using Sakai's `ir4qa_eval` [1].

Tables 1-3 show the number of judged nonrelevant and relevant documents per topic for each language in qrels version 2. These correspond to Tables 35-37 in the Overview. For CS, the average number of relevant documents per topic has increased from 9488/97=97.8 to 16475/97=169.8 (+73%); For CT, it has increased from 5262/95=55.4 to 8620/95=90.7 (+61%); For JA, it has increased from 8506/98=86.8 to 10785/98=110.1 (+79%).

Tables 4-6 show the *coverage* of version 2 relevant documents for each IR4QA run (See Section 6.1 in the Overview for the definition of coverage). These correspond to Tables 12-14 in the Overview. Note that OT-CS-CS-04-T has moved from Rank 3 (Table 12 in the Overview) to Rank 1 to receive the "best coverage award" among the CS runs (Table 4).

Similarly, Tables 7-9 show the coverage of version 2 relevant documents for each *team*. These correspond to Tables 15-17 in the Overview.

Tables 10-12 show the total number of unique relevant documents found by each run. These correspond to Tables 18-20 in the Overview.

Tables 13-15 show the total number of unique relevant documents found by each *team*. These correspond to Tables 21-23 in the Overview.

2 System Rankings

Tables 16-18 show the Mean AP, Q-measure and nDCG values for each run based on qrels version 2. These correspond to Tables 6-8 in the Overview. With qrels version 1 for CT, Q and nDCG disagreed with AP by ranking MITEL-CT-CT-02-T at the top (Table 7 in the Overview). With version 2, however, all three metrics agree that MITEL-CT-CT-03-D is the best CT run on average.

Tables 19-21 show the "best" T-runs from each team. These correspond to Tables 9-11 in the Overview. For each adjacent pair of runs shown in this table, we conducted a two-sided, paired bootstrap test using 1000 bootstrap samples of topics. Our additional relevance assessments have resulted in more system pairs with significant performance differences. For example, with qrels version 1, OT-CS-CS-04-T and MITEL-EN-CS-03-T were not significantly different from each other in terms of AP and Q (Table 9 in the Overview), but with version 2, the difference between these two runs are statistically significant at $\alpha = 0.01$ in terms of AP, Q and nDCG. This suggests

Table 1. 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	269	74	76	150	419	CS-T101	763	0	19	19	782
CS-T42	714	153	172	325	1039	CS-T102	493	32	6	38	531
CS-T43	146	103	152	255	401	CS-T103	164	35	338	373	537
CS-T44	567	102	135	237	804	CS-T104	316	48	42	90	406
CS-T46	337	129	197	326	663	CS-T317	756	6	10	16	772
CS-T47	420	196	83	279	699	CS-T320	271	45	135	180	451
CS-T48	733	21	37	58	791	CS-T321	269	56	75	131	400
CS-T49	209	37	188	225	434	CS-T322	303	195	270	465	768
CS-T52	271	71	13	84	355	CS-T323	392	70	88	158	550
CS-T53	772	3	4	7	779	CS-T324	66	114	224	338	404
CS-T54	690	11	4	15	705	CS-T325	64	103	401	504	568
CS-T55	391	131	60	191	582	CS-T326	517	20	19	39	556
CS-T56	906	20	112	132	1038	CS-T328	146	118	195	313	459
CS-T57	338	190	172	362	700	CS-T329	376	18	10	28	404
CS-T58	232	38	51	89	321	CS-T332	257	43	352	395	652
CS-T60	184	38	108	146	330	CS-T333	518	102	268	370	888
CS-T61	288	73	85	158	446	CS-T334	370	96	531	627	997
CS-T62	1029	15	13	28	1057	CS-T336	229	14	163	177	406
CS-T64	315	116	33	149	464	CS-T337	245	46	153	199	444
CS-T65	221	78	6	84	305	CS-T338	174	36	156	192	366
CS-T67	471	17	105	122	593	CS-T339	88	37	244	281	369
CS-T68	550	34	3	37	587	CS-T340	151	32	107	139	290
CS-T69	664	42	3	45	709	CS-T347	544	16	34	50	594
CS-T71	646	6	23	29	675	CS-T348	76	51	145	196	272
CS-T73	1057	19	77	96	1153	CS-T349	565	22	22	44	609
CS-T74	1020	32	254	286	1306	CS-T350	687	36	16	52	739
CS-T75	663	5	14	19	682	CS-T351	128	129	47	176	304
CS-T76	296	16	38	54	350	CS-T352	146	45	175	220	366
CS-T77	726	22	25	47	773	CS-T355	289	44	159	203	492
CS-T78	901	6	4	10	911	CS-T357	443	45	119	164	607
CS-T79	511	16	11	27	538	CS-T358	545	53	160	213	758
CS-T80	849	2	5	7	856	CS-T359	303	62	464	526	829
CS-T81	150	24	113	137	287	CS-T361	133	100	43	143	276
CS-T82	757	109	115	224	981	CS-T365	611	0	7	7	618
CS-T83	357	244	219	463	820	CS-T366	541	41	30	71	612
CS-T84	250	105	52	157	407	CS-T367	224	47	247	294	518
CS-T85	154	69	74	143	297	CS-T368	509	85	74	159	668
CS-T87	377	158	440	598	975	CS-T369	459	159	66	225	684
CS-T89	554	17	4	21	575	CS-T370	237	50	25	75	312
CS-T90	227	135	238	373	600	CS-T376	303	150	114	264	567
CS-T91	149	162	268	430	579	CS-T378	483	20	14	34	517
CS-T92	407	75	75	150	557	CS-T379	380	112	80	192	572
CS-T93	461	47	57	104	565	CS-T380	421	127	13	140	561
CS-T94	825	37	35	72	897	CS-T381	494	4	9	13	507
CS-T95	517	63	184	247	764	CS-T383	931	1	15	16	947
CS-T96	186	66	90	156	342	CS-T384	526	4	16	20	546
CS-T97	140	47	137	184	324	CS-T385	567	27	56	83	650
CS-T98	236	47	101	148	384						
CS-T99	640	27	18	45	685						
CS-T100	258	35	57	92	350						
						total	41504	5979	10496	16475	57979

that the additional relevance assessments were worthwhile.

Table 2. 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	419	178	112	290	709	CT-T400	356	140	46	186	542
CT-T172	535	23	47	70	605	CT-T402	680	93	27	120	800
CT-T174	491	18	118	136	627	CT-T404	362	224	8	232	594
CT-T175	334	28	70	98	432	CT-T405	543	69	11	80	623
CT-T176	643	26	180	206	849	CT-T407	264	39	21	60	324
CT-T177	685	74	51	125	810	CT-T408	157	99	27	126	283
CT-T178	291	94	39	133	424	CT-T409	275	45	11	56	331
CT-T179	490	14	46	60	550	CT-T410	766	6	7	13	779
CT-T180	656	41	57	98	754	CT-T411	485	279	42	321	806
CT-T181	278	24	181	205	483	CT-T413	483	0	6	6	489
CT-T182	497	35	52	87	584	CT-T414	721	24	4	28	749
CT-T183	689	75	84	159	848	CT-T415	881	8	17	25	906
CT-T184	613	25	12	37	650	CT-T416	679	36	42	78	757
CT-T186	477	20	20	40	517	CT-T418	407	14	16	30	437
CT-T187	746	0	12	12	758	CT-T419	887	24	16	40	927
CT-T188	985	21	27	48	1033	CT-T420	430	27	141	168	598
CT-T189	430	94	59	153	583	CT-T421	605	3	3	6	611
CT-T190	690	18	12	30	720	CT-T422	643	1	21	22	665
CT-T191	410	4	3	7	417	CT-T423	627	0	35	35	662
CT-T193	487	3	39	42	529	CT-T424	553	49	30	79	632
CT-T194	80	294	34	328	408	CT-T426	268	31	74	105	373
CT-T195	323	66	60	126	449	CT-T427	426	26	71	97	523
CT-T196	836	99	6	105	941	CT-T428	175	13	115	128	303
CT-T197	546	65	16	81	627	CT-T429	330	14	21	35	365
CT-T198	649	39	12	51	700	CT-T430	627	21	16	37	664
CT-T200	158	246	15	261	419	CT-T431	602	0	20	20	622
CT-T203	893	105	9	114	1007	CT-T432	729	2	15	17	746
CT-T204	910	24	9	33	943	CT-T434	358	18	24	42	400
CT-T205	619	12	4	16	635	CT-T435	602	85	14	99	701
CT-T206	869	18	2	20	889	CT-T436	801	3	9	12	813
CT-T207	901	31	11	42	943	CT-T437	641	47	71	118	759
CT-T208	684	51	13	64	748	CT-T438	339	155	64	219	558
CT-T210	371	7	12	19	390	CT-T439	336	27	25	52	388
CT-T211	504	37	20	57	561	CT-T440	374	22	18	40	414
CT-T213	271	3	133	136	407	CT-T441	554	30	10	40	594
CT-T374	90	327	26	353	443	CT-T442	562	82	76	158	720
CT-T386	393	83	26	109	502	CT-T443	536	63	90	153	689
CT-T387	214	143	40	183	397	CT-T444	460	80	21	101	561
CT-T388	1163	18	3	21	1184	CT-T445	391	61	54	115	506
CT-T389	858	1	31	32	890	CT-T446	708	6	46	52	760
CT-T390	582	102	22	124	706	CT-T447	548	55	67	122	670
CT-T391	774	39	7	46	820	CT-T448	739	25	76	101	840
CT-T392	772	3	66	69	841	CT-T449	713	4	19	23	736
CT-T393	730	26	22	48	778	CT-T450	759	44	70	114	873
CT-T394	577	59	3	62	639	CT-T451	693	4	27	31	724
CT-T395	1257	3	3	6	1263						
CT-T396	211	157	15	172	383						
CT-T397	809	24	7	31	840						
CT-T398	302	108	19	127	429						
CT-T399	682	2	4	6	688						
						total	52949	5105	3515	8620	61569

Table 3. 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	426	150	227	377	803
JA-T2	1007	4	2	6	1013	JA-T164	468	75	212	287	755
JA-T3	461	153	129	282	743	JA-T165	949	42	36	78	1027
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	285	264	75	339	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	414	161	1	162	576	JA-T225	540	6	357	363	903
JA-T29	982	15	9	24	1006	JA-T230	364	116	236	352	716
JA-T32	857	90	7	97	954	JA-T231	544	84	69	153	697
JA-T35	850	46	15	61	911	JA-T233	324	137	44	181	505
JA-T37	1021	24	6	30	1051	JA-T234	630	9	7	16	646
JA-T38	961	12	10	22	983	JA-T236	421	425	43	468	889
JA-T105	790	62	2	64	854	JA-T237	1006	1	8	9	1015
JA-T106	499	76	21	97	596	JA-T238	509	103	188	291	800
JA-T107	827	13	7	20	847	JA-T239	458	3	262	265	723
JA-T108	876	105	7	112	988	JA-T240	874	3	136	139	1013
JA-T109	755	19	5	24	779	JA-T242	329	3	210	213	542
JA-T110	878	3	2	5	883	JA-T244	902	1	16	17	919
JA-T111	630	8	0	8	638	JA-T245	881	0	23	23	904
JA-T112	517	23	34	57	574	JA-T248	486	0	353	353	839
JA-T113	533	24	11	35	568	JA-T249	744	0	26	26	770
JA-T115	641	5	9	14	655	JA-T250	493	2	97	99	592
JA-T119	727	13	28	41	768	JA-T253	526	22	47	69	595
JA-T128	732	2	5	7	739	JA-T254	467	6	162	168	635
JA-T130	893	3	19	22	915	JA-T255	594	122	86	208	802
JA-T134	420	26	60	86	506	JA-T266	329	217	9	226	555
JA-T137	958	23	49	72	1030	JA-T267	483	148	11	159	642
JA-T138	386	41	28	69	455	JA-T271	344	84	34	118	462
JA-T140	409	17	31	48	457	JA-T275	516	11	114	125	641
JA-T141	565	11	4	15	580	JA-T276	919	0	10	10	929
JA-T148	680	99	37	136	816	JA-T284	999	35	61	96	1095
JA-T149	500	33	43	76	576	JA-T291	512	52	13	65	577
JA-T151	612	31	20	51	663	JA-T295	520	31	6	37	557
JA-T152	563	62	120	182	745	JA-T297	638	35	117	152	790
JA-T153	696	17	12	29	725	JA-T300	390	139	75	214	604
JA-T154	772	22	26	48	820	JA-T301	599	38	8	46	645
JA-T155	581	46	27	73	654	JA-T304	634	25	3	28	662
JA-T157	544	123	61	184	728	JA-T305	808	14	12	26	834
JA-T158	645	239	65	304	949	JA-T313	904	7	5	12	916
JA-T160	605	194	38	232	837	JA-T315	840	23	31	54	894
JA-T161	445	125	56	181	626						
JA-T162	506	15	69	84	590						
						total	63934	5488	5297	10785	74719

Table 4. Coverage of relevant documents summed across 97 topics: CS runs.

run name	covered docs.
OT-CS-CS-04-T	14076
MITEL-EN-CS-05-TD	13966
MITEL-EN-CS-03-T	13843
MITEL-EN-CS-01-T	13823
MITEL-EN-CS-04-D	13754
MITEL-EN-CS-02-T	13741
OT-CS-CS-02-T	13683
HIT-EN-CS-02-D	13482
RALI-CS-CS-04-T	13446
HIT-EN-CS-01-DN	13430
OT-CS-CS-01-T	13419
OT-CS-CS-05-T	13407
OT-CS-CS-03-T	13368
HIT-EN-CS-02-T	13338
RALI-CS-CS-03-T	13331
CMUJAV-CS-CS-02-T	13259
CMUJAV-CS-CS-01-T	13252
RALI-CS-CS-05-T	13195
RALI-CS-CS-01-T	13080
RALI-CS-CS-02-T	13042
CMUJAV-EN-CS-01-T	12820
RALI-EN-CS-04-T	12356
CMUJAV-EN-CS-02-T	12167
HIT-EN-CS-02-DN	12151
RALI-EN-CS-02-T	11855
RALI-EN-CS-05-T	11832
RALI-EN-CS-01-T	11705
CYUT-EN-CS-03-DN	11066
CYUT-EN-CS-01-T	10089
CYUT-EN-CS-02-D	9952
KECIR-CS-CS-01-T	7466
KECIR-CS-CS-02-DN	5261
KECIR-CS-CS-03-DN	5106
WHUCC-CS-CS-02-T	3322
WHUCC-CS-CS-01-T	3322
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 5. Coverage of relevant documents summed across topics: CT runs.

run name	covered docs.
MITEL-CT-CT-03-D	7924
OT-CT-CT-04-T	7900
MITEL-CT-CT-02-T	7877
MITEL-CT-CT-01-T	7845
OT-CT-CT-02-T	7636
MITEL-CT-CT-04-T	7633
RALI-CT-CT-04-T	7527
RALI-CT-CT-03-T	7495
OT-CT-CT-05-T	7490
OT-CT-CT-01-T	7478
OT-CT-CT-03-T	7476
RALI-CT-CT-05-T	7335
RALI-CT-CT-02-T	7270
RALI-CT-CT-01-T	7254
NTUBROWS-CT-CT-02-T	6693
NTUBROWS-CT-CT-03-T	6298
NTUBROWS-CT-CT-01-T	6298
NTUBROWS-CT-CT-04-T	6249
RALI-EN-CT-04-T	5728
RALI-EN-CT-05-T	5520
RALI-EN-CT-02-T	5482
RALI-EN-CT-01-T	5435
CYUT-EN-CT-01-T	5058
NTUBROWS-CT-CT-05-T	4991
CYUT-EN-CT-03-DN	4748
CYUT-EN-CT-02-D	4740

∗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 6. Coverage of relevant documents summed across topics: JA runs.

run name	covered docs.
OT-JA-JA-04-T	10138
OT-JA-JA-02-T	10026
BRKLY-JA-JA-01-DN	9963
BRKLY-JA-JA-02-T	9723
BRKLY-JA-JA-02-DN	9404
OT-JA-JA-01-T	9315
OT-JA-JA-05-T	9309
BRKLY-JA-JA-03-T	9205
CMUJAV-JA-JA-01-T	9147
CMUJAV-JA-JA-04-T	9130
CMUJAV-JA-JA-03-T	9075
CMUJAV-JA-JA-02-T	9061
CMUJAV-JA-JA-05-T	9052
OT-JA-JA-03-T	8326
CMUJAV-EN-JA-01-T	7146
CMUJAV-EN-JA-03-T	7112
CMUJAV-EN-JA-05-T	7104
CMUJAV-EN-JA-02-T	7094
CMUJAV-EN-JA-04-T	7080
CYUT-EN-JA-01-T	5827
CYUT-EN-JA-03-DN	5765
CYUT-EN-JA-02-D	5633
TA-EN-JA-03-T	533
TA-EN-JA-02-D	502
TA-EN-JA-01-D	275

Table 7. Coverage of relevant documents summed across topics: CS-teams.

team name	covered docs.
OT	15010
RALI	14986
MITEL	14492
CMUJAV	14381
HIT	13966
CYUT	11663
KECIR	9497
WHUCC	3322
NLPAI	1222

Table 8. Coverage of relevant documents summed across topics: CT-teams.

team name	covered docs.
RALI	8359
OT	8217
MITEL	8136
NTUBROWS	7727
CYUT	5448

Table 9. Coverage of relevant documents summed across topics: JA-teams.

team name	covered docs.
OT	10411
BRKLY	10321
CMUJAV	9521
CYUT	6791
TA	632

Table 10. Unique relevant documents summed across topics: CS runs.

run name	unique relevant
RALI-EN-CS-04-T	112
RALI-EN-CS-02-T	97
RALI-CS-CS-04-T	95
RALI-CS-CS-02-T	95
RALI-CS-CS-03-T	93
RALI-EN-CS-05-T	89
RALI-CS-CS-05-T	89
RALI-EN-CS-01-T	88
RALI-CS-CS-01-T	85
HIT-EN-CS-01-DN	71
HIT-EN-CS-02-D	70
HIT-EN-CS-02-T	69
OT-CS-CS-01-T	62
HIT-EN-CS-02-DN	59
CYUT-EN-CS-02-D	55
CYUT-EN-CS-03-DN	51
CYUT-EN-CS-01-T	51
OT-CS-CS-05-T	49
OT-CS-CS-04-T	48
OT-CS-CS-02-T	48
OT-CS-CS-03-T	44
CMUJAV-EN-CS-01-T	11
MITEL-EN-CS-03-T	10
MITEL-EN-CS-05-TD	9
MITEL-EN-CS-04-D	9
MITEL-EN-CS-02-T	9
MITEL-EN-CS-01-T	9
KECIR-CS-CS-01-T	9
CMUJAV-EN-CS-02-T	6
CMUJAV-CS-CS-02-T	6
CMUJAV-CS-CS-01-T	6
KECIR-CS-CS-02-DN	3
WHUCC-CS-CS-02-T	2
WHUCC-CS-CS-01-T	2
KECIR-CS-CS-03-DN	2
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

Table 11. Unique relevant documents summed across topics: CT runs.

run name	unique relevant
RALI-EN-CT-04-T	98
RALI-EN-CT-02-T	98
RALI-EN-CT-05-T	97
RALI-EN-CT-01-T	97
CYUT-EN-CT-03-DN	14
CYUT-EN-CT-02-D	14
CYUT-EN-CT-01-T	14
RALI-CT-CT-03-T	12
OT-CT-CT-01-T	10
RALI-CT-CT-05-T	9
RALI-CT-CT-04-T	9
RALI-CT-CT-02-T	9
RALI-CT-CT-01-T	9
OT-CT-CT-05-T	6
OT-CT-CT-04-T	6
OT-CT-CT-02-T	6
OT-CT-CT-03-T	3
NTUBROWS-CT-CT-05-T$*$	3
MITEL-CT-CT-03-D	3
NTUBROWS-CT-CT-04-T	2
NTUBROWS-CT-CT-03-T	2
NTUBROWS-CT-CT-01-T	2
MITEL-CT-CT-04-T	1
MITEL-CT-CT-02-T	1
MITEL-CT-CT-01-T	1
NTUBROWS-CT-CT-02-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 12. Unique relevant documents summed across topics: JA runs.

run name	unique relevant
OT-JA-JA-01-T	71
OT-JA-JA-05-T	61
OT-JA-JA-03-T	48
BRKLY-JA-JA-01-DN	43
OT-JA-JA-02-T	41
OT-JA-JA-04-T	40
BRKLY-JA-JA-02-T	40
BRKLY-JA-JA-02-DN	31
CYUT-EN-JA-02-D	26
CYUT-EN-JA-03-DN	24
CYUT-EN-JA-01-T	22
BRKLY-JA-JA-03-T	20
CMUJAV-EN-JA-05-T	7
CMUJAV-EN-JA-04-T	7
CMUJAV-EN-JA-03-T	7
CMUJAV-EN-JA-02-T	7
CMUJAV-EN-JA-01-T	7
TA-EN-JA-02-D	4
TA-EN-JA-01-D	4
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 13. Unique relevant documents summed across topics: CS-teams.

team name	unique relevant
RALI	119
OT	78
HIT	71
CYUT	57
KECIR	12
CMUJAV	12
MITEL	10
WHUCC	2
NLPAI	0

Table 14. Unique relevant documents summed across topics: CT-teams.

team name	covered docs.
RALI	106
CYUT	14
OT	10
NTUBROWS	5
MITEL	3

Table 15. Unique relevant documents summed across topics: JA-teams.

team name	covered docs.
OT	74
BRKLY	44
CYUT	27
CMUJAV	7
TA	4

Table 16. Performances based on qrels version 2: CS runs; 97 topics.

run	Mean AP	run	Mean Q	run	Mean nDCG
OT-CS-CS-04-T	0.6184	OT-CS-CS-04-T	0.6192	OT-CS-CS-04-T	0.8086
OT-CS-CS-02-T	0.6028	OT-CS-CS-02-T	0.6010	OT-CS-CS-02-T	0.7895
CMUJAV-CS-CS-02-T	0.5733	CMUJAV-CS-CS-02-T	0.5714	CMUJAV-CS-CS-02-T	0.7680
CMUJAV-CS-CS-01-T	0.5704	MITEL-EN-CS-03-T	0.5693	CMUJAV-CS-CS-01-T	0.7673
MITEL-EN-CS-03-T	0.5670	CMUJAV-CS-CS-01-T	0.5690	MITEL-EN-CS-05-TD	0.7667
MITEL-EN-CS-05-TD	0.5606	MITEL-EN-CS-05-TD	0.5613	MITEL-EN-CS-03-T	0.7619
HIT-EN-CS-01-DN	0.5598	HIT-EN-CS-01-DN	0.5596	MITEL-EN-CS-01-T	0.7616
MITEL-EN-CS-01-T	0.5550	MITEL-EN-CS-01-T	0.5558	OT-CS-CS-03-T	0.7591
HIT-EN-CS-02-T	0.5538	OT-CS-CS-03-T	0.5538	OT-CS-CS-05-T	0.7575
MITEL-EN-CS-04-D	0.5499	OT-CS-CS-05-T	0.5535	MITEL-EN-CS-04-D	0.7571
OT-CS-CS-03-T	0.5482	HIT-EN-CS-02-T	0.5535	MITEL-EN-CS-02-T	0.7507
OT-CS-CS-05-T	0.5478	MITEL-EN-CS-04-D	0.5514	HIT-EN-CS-01-DN	0.7397
MITEL-EN-CS-02-T	0.5394	MITEL-EN-CS-02-T	0.5414	HIT-EN-CS-02-T	0.7337
CMUJAV-EN-CS-01-T	0.5233	CMUJAV-EN-CS-01-T	0.5207	RALI-CS-CS-04-T	0.7293
HIT-EN-CS-02-D	0.5073	HIT-EN-CS-02-D	0.5123	RALI-CS-CS-03-T	0.7268
CMUJAV-EN-CS-02-T	0.5044	CMUJAV-EN-CS-02-T	0.5019	RALI-CS-CS-05-T	0.7242
RALI-CS-CS-05-T	0.4852	RALI-CS-CS-05-T	0.4887	RALI-CS-CS-01-T	0.7182
RALI-CS-CS-03-T	0.4843	RALI-CS-CS-03-T	0.4876	RALI-CS-CS-02-T	0.7144
RALI-CS-CS-01-T	0.4834	RALI-CS-CS-01-T	0.4863	CMUJAV-EN-CS-01-T	0.7140
RALI-CS-CS-04-T	0.4786	RALI-CS-CS-04-T	0.4832	HIT-EN-CS-02-D	0.7016
RALI-CS-CS-02-T	0.4768	RALI-CS-CS-02-T	0.4776	OT-CS-CS-01-T	0.6999
HIT-EN-CS-02-DN	0.4477	HIT-EN-CS-02-DN	0.4542	CMUJAV-EN-CS-02-T	0.6987
KECIR-CS-CS-01-T	0.4424	RALI-EN-CS-04-T	0.4293	RALI-EN-CS-04-T	0.6713
RALI-EN-CS-04-T	0.4208	RALI-EN-CS-05-T	0.4255	HIT-EN-CS-02-DN	0.6638
RALI-EN-CS-05-T	0.4176	RALI-EN-CS-01-T	0.4236	RALI-EN-CS-05-T	0.6563
RALI-EN-CS-02-T	0.4165	RALI-EN-CS-02-T	0.4223	RALI-EN-CS-02-T	0.6551
RALI-EN-CS-01-T	0.4156	OT-CS-CS-01-T	0.4198	RALI-EN-CS-01-T	0.6508
CYUT-EN-CS-03-DN	0.4018	KECIR-CS-CS-01-T	0.4125	CYUT-EN-CS-03-DN	0.6182
OT-CS-CS-01-T	0.3830	CYUT-EN-CS-03-DN	0.4013	CYUT-EN-CS-01-T	0.5749
KECIR-CS-CS-02-DN	0.3753	CYUT-EN-CS-01-T	0.3608	KECIR-CS-CS-01-T	0.5744
CYUT-EN-CS-01-T	0.3586	CYUT-EN-CS-02-D	0.3549	CYUT-EN-CS-02-D	0.5684
KECIR-CS-CS-03-DN	0.3558	KECIR-CS-CS-02-DN	0.3498	KECIR-CS-CS-02-DN	0.5060
CYUT-EN-CS-02-D	0.3519	KECIR-CS-CS-03-DN	0.3380	KECIR-CS-CS-03-DN	0.4993
WHUCC-CS-CS-02-T†	0.2837	WHUCC-CS-CS-02-T†	0.2675	WHUCC-CS-CS-02-T†	0.4054
WHUCC-CS-CS-01-T†	0.2837	WHUCC-CS-CS-01-T†	0.2675	WHUCC-CS-CS-01-T†	0.4054
NLPAI-CS-CS-02-T	0.0990	NLPAI-CS-CS-02-T	0.0924	NLPAI-CS-CS-02-T	0.1966
NLPAI-CS-CS-05-DN	0.0979	NLPAI-CS-CS-05-DN	0.0914	NLPAI-CS-CS-05-DN	0.1934
NLPAI-CS-CS-01-T	0.0917	NLPAI-CS-CS-01-T	0.0841	NLPAI-CS-CS-01-T	0.1865
NLPAI-CS-CS-03-T	0.0882	NLPAI-CS-CS-03-T	0.0811	NLPAI-CS-CS-03-T	0.1786
NLPAI-CS-CS-04-T	0.0845	NLPAI-CS-CS-04-T	0.0768	NLPAI-CS-CS-04-T	0.1716

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

Table 17. Performances based on the qrels version 2: CT runs; 95 topics.

run	Mean AP	run	Mean Q	run	Mean nDCG
MITEL-CT-CT-03-D	0.5561	MITEL-CT-CT-03-D	0.5715	MITEL-CT-CT-03-D	0.7705
MITEL-CT-CT-02-T	0.5547	MITEL-CT-CT-02-T	0.5700	MITEL-CT-CT-02-T	0.7684
MITEL-CT-CT-01-T	0.5507	MITEL-CT-CT-01-T	0.5653	MITEL-CT-CT-01-T	0.7646
MITEL-CT-CT-04-T	0.5432	MITEL-CT-CT-04-T	0.5545	OT-CT-CT-04-T	0.7531
OT-CT-CT-04-T	0.5304	OT-CT-CT-04-T	0.5488	MITEL-CT-CT-04-T	0.7471
OT-CT-CT-03-T	0.4793	OT-CT-CT-02-T	0.4978	OT-CT-CT-02-T	0.7204
OT-CT-CT-02-T	0.4768	OT-CT-CT-03-T	0.4973	OT-CT-CT-03-T	0.7134
OT-CT-CT-05-T	0.4562	OT-CT-CT-05-T	0.4770	OT-CT-CT-05-T	0.7031
RALI-CT-CT-05-T	0.4051	RALI-CT-CT-05-T	0.4186	RALI-CT-CT-03-T	0.6673
RALI-CT-CT-01-T	0.4030	RALI-CT-CT-01-T	0.4162	RALI-CT-CT-04-T	0.6652
RALI-CT-CT-03-T	0.3874	RALI-CT-CT-03-T	0.4056	RALI-CT-CT-05-T	0.6586
RALI-CT-CT-04-T	0.3861	RALI-CT-CT-04-T	0.4034	RALI-CT-CT-01-T	0.6528
RALI-CT-CT-02-T	0.3850	RALI-CT-CT-02-T	0.3993	RALI-CT-CT-02-T	0.6496
NTUBROWS-CT-CT-01-T	0.3415	NTUBROWS-CT-CT-01-T	0.3574	OT-CT-CT-01-T	0.6424
OT-CT-CT-01-T	0.3077	OT-CT-CT-01-T	0.3533	NTUBROWS-CT-CT-01-T	0.5804
RALI-EN-CT-01-T	0.2759	RALI-EN-CT-05-T	0.2904	NTUBROWS-CT-CT-02-T	0.5115
RALI-EN-CT-05-T	0.2745	RALI-EN-CT-01-T	0.2904	NTUBROWS-CT-CT-03-T	0.4925
RALI-EN-CT-04-T	0.2628	RALI-EN-CT-04-T	0.2808	RALI-EN-CT-04-T	0.4894
RALI-EN-CT-02-T	0.2626	RALI-EN-CT-02-T	0.2769	RALI-EN-CT-05-T	0.4791
CYUT-EN-CT-01-T	0.2469	NTUBROWS-CT-CT-02-T	0.2639	RALI-EN-CT-01-T	0.4769
CYUT-EN-CT-03-DN	0.2362	CYUT-EN-CT-01-T	0.2596	RALI-EN-CT-02-T	0.4757
CYUT-EN-CT-02-D	0.2352	NTUBROWS-CT-CT-03-T	0.2577	NTUBROWS-CT-CT-04-T	0.4739
NTUBROWS-CT-CT-03-T	0.2267	CYUT-EN-CT-02-D	0.2483	CYUT-EN-CT-01-T	0.4596
NTUBROWS-CT-CT-02-T	0.2208	CYUT-EN-CT-03-DN	0.2474	CYUT-EN-CT-02-D	0.4454
NTUBROWS-CT-CT-04-T	0.2102	NTUBROWS-CT-CT-04-T	0.2411	CYUT-EN-CT-03-DN	0.4448
NTUBROWS-CT-CT-05-T	0.1780	NTUBROWS-CT-CT-05-T	0.2090	NTUBROWS-CT-CT-05-T	0.4078

∗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 18. Performances based on qrels version 2: JA runs; 98 topics.

run	Mean AP	run	Mean Q	run	Mean nDCG
OT-JA-JA-04-T	0.6999	OT-JA-JA-04-T	0.7068	OT-JA-JA-04-T	0.8632
OT-JA-JA-02-T	0.6682	OT-JA-JA-02-T	0.6748	OT-JA-JA-02-T	0.8439
BRKLY-JA-JA-01-DN	0.6376	BRKLY-JA-JA-01-DN	0.6470	BRKLY-JA-JA-01-DN	0.8192
CMUJAV-JA-JA-01-T	0.5969	BRKLY-JA-JA-02-T	0.6029	BRKLY-JA-JA-02-T	0.7854
CMUJAV-JA-JA-03-T	0.5932	CMUJAV-JA-JA-01-T	0.5987	CMUJAV-JA-JA-01-T	0.7803
CMUJAV-JA-JA-04-T	0.5925	CMUJAV-JA-JA-03-T	0.5954	CMUJAV-JA-JA-04-T	0.7786
BRKLY-JA-JA-02-T	0.5903	CMUJAV-JA-JA-04-T	0.5945	CMUJAV-JA-JA-03-T	0.7766
CMUJAV-JA-JA-02-T	0.5834	CMUJAV-JA-JA-02-T	0.5875	BRKLY-JA-JA-02-DN	0.7757
CMUJAV-JA-JA-05-T	0.5831	BRKLY-JA-JA-02-DN	0.5863	OT-JA-JA-05-T	0.7731
BRKLY-JA-JA-02-DN	0.5810	CMUJAV-JA-JA-05-T	0.5853	CMUJAV-JA-JA-02-T	0.7716
OT-JA-JA-05-T	0.5596	OT-JA-JA-05-T	0.5742	CMUJAV-JA-JA-05-T	0.7696
BRKLY-JA-JA-03-T	0.5469	BRKLY-JA-JA-03-T	0.5540	BRKLY-JA-JA-03-T	0.7484
CMUJAV-EN-JA-01-T	0.4311	OT-JA-JA-03-T	0.4417	OT-JA-JA-01-T	0.7111
CMUJAV-EN-JA-03-T	0.4283	OT-JA-JA-01-T	0.4349	OT-JA-JA-03-T	0.6571
CMUJAV-EN-JA-04-T	0.4269	CMUJAV-EN-JA-01-T	0.4342	CMUJAV-EN-JA-01-T	0.5999
OT-JA-JA-03-T	0.4234	CMUJAV-EN-JA-03-T	0.4312	CMUJAV-EN-JA-03-T	0.5979
CMUJAV-EN-JA-05-T	0.4227	CMUJAV-EN-JA-04-T	0.4297	CMUJAV-EN-JA-04-T	0.5965
CMUJAV-EN-JA-02-T	0.4227	CMUJAV-EN-JA-05-T	0.4257	CMUJAV-EN-JA-05-T	0.5943
OT-JA-JA-01-T	0.3911	CMUJAV-EN-JA-02-T	0.4255	CMUJAV-EN-JA-02-T	0.5928
CYUT-EN-JA-01-T	0.2552	CYUT-EN-JA-01-T	0.2500	CYUT-EN-JA-03-DN	0.4288
CYUT-EN-JA-03-DN	0.2543	CYUT-EN-JA-03-DN	0.2486	CYUT-EN-JA-01-T	0.4218
CYUT-EN-JA-02-D	0.2277	CYUT-EN-JA-02-D	0.2253	CYUT-EN-JA-02-D	0.4058
TA-EN-JA-02-D	0.0154	TA-EN-JA-02-D	0.0167	TA-EN-JA-03-T	0.0446
TA-EN-JA-03-T	0.0128	TA-EN-JA-03-T	0.0157	TA-EN-JA-02-D	0.0349
TA-EN-JA-01-D	0.0119	TA-EN-JA-01-D	0.0118	TA-EN-JA-01-D	0.0261

Table 19. 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	0.6184**	OT-CS-CS-04-T	0.6192**	OT-CS-CS-04-T	0.8086**
CMUJAV-CS-CS-02-T	0.5733	CMUJAV-CS-CS-02-T	0.5714	CMUJAV-CS-CS-02-T	0.7680
MITEL-EN-CS-03-T	0.5670	MITEL-EN-CS-03-T	0.5693	MITEL-EN-CS-05-TD	0.7667
HIT-EN-CS-02-T	0.5538**	HIT-EN-CS-02-T	0.5535**	HIT-EN-CS-02-T	0.7337
RALI-CS-CS-05-T	0.4852*	RALI-CS-CS-05-T	0.4887 **	RALI-CS-CS-04-T	0.7293**
KECIR-CS-CS-01-T	0.4424**	KECIR-CS-CS-01-T	0.4125	CYUT-EN-CS-01-T	0.5749
CYUT-EN-CS-01-T	0.3586*	CYUT-EN-CS-01-T	0.3608**	KECIR-CS-CS-01-T	0.5744**
WHUCC-CS-CS-01-T	0.2837**	WHUCC-CS-CS-01-T	0.2675**	WHUCC-CS-CS-01-T	0.4054**
NLPAI-CS-CS-02-T	0.0990	NLPAI-CS-CS-02-T	0.0924	NLPAI-CS-CS-02-T	0.1966

Table 20. 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	0.5547	MITEL-CT-CT-02-T	0.5700	MITEL-CT-CT-02-T	0.7684
OT-CT-CT-04-T	0.5304**	OT-CT-CT-04-T	0.5488**	OT-CT-CT-04-T	0.7531**
RALI-CT-CT-05-T	0.4051*	RALI-CT-CT-05-T	0.4186*	RALI-CT-CT-03-T	0.6673**
NTUBROWS-CT-CT-01-T	0.3415**	NTUBROWS-CT-CT-01-T	0.3574**	NTUBROWS-CT-CT-01-T	0.5804**
CYUT-EN-CT-01-T	0.2469	CYUT-EN-CT-01-T	0.2596	CYUT-EN-CT-01-T	0.4596

Table 21. 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	0.6999**	OT-JA-JA-04-T	0.7068**	OT-JA-JA-04-T	0.8632**
CMUJAV-JA-JA-01-T	0.5969	BRKLY-JA-JA-02-T	0.6029	BRKLY-JA-JA-02-T	0.7854
BRKLY-JA-JA-02-T	0.5903**	CMUJAV-JA-JA-01-T	0.5987**	CMUJAV-JA-JA-01-T	0.7803**
CYUT-EN-JA-01-T	0.2552**	CYUT-EN-JA-01-T	0.2500**	CYUT-EN-JA-01-T	0.4218**
TA-EN-JA-03-T	0.0128	TA-EN-JA-03-T	0.0157	TA-EN-JA-03-T	0.0446

3 Topic Rankings

Tables 22-24 show the *topic* rankings based performance averaged across runs, using qrels version 2. These correspond to Tables 32-34 in the Overview, and reflect the "difficulty" of topics.

4 Correlation between Metrics

Table 25 compares two system rankings according to two different evaluation metrics using Kendall's rank correlation and Yilmaz/Aslam/Robertson (YAR) rank correlation. This corresponds to Table 25 in the Overview. For example, for the CS runs, the Kendall's correlation between the system ranking by Mean AP and that by Mean nDCG is .874; YAR rank correlation between Mean AP and Mean nDCG is .845 when the latter is taken as the ground truth, and .850 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 two metrics, each averaged across runs. This corresponds to Table 26 in the Overview. In both Table 25 and Table 26, Q is consistently more highly correlated with AP than nDCG is, and Q is consistently more highly correlated with nDCG than AP is. In short, Q is like a "summary" of AP and nDCG, just as it was designed to be.

5 Correlation between Version 1 and Version 2

Table 27 compares two system rankings based on the same evaluation metric, but with qrels version 1 and version 2. The YAR correlation values are computed by treating the version 2 ranking as the ground truth. It can be observed that the correlations are high for JA, which reflects the fact that the JA relevance assessors went beyond pool depth 30 for many topics before the release of qrels version 1, as indicated in Table 31 in the Overview. That is, JA qrels version 1 was already quite reliable, relatively speaking.

Figures 1-9 visualise, for each metric, how the performance values/rankings based on qrels version 1 deviate from those based on version 2. The runs are sorted by the version 2 performance. Figures 1-3 show not only that many CS runs were "misranked" by qrels version 1, but also that they were overestimated in terms of absolute performance. In contrast, Figures 7-9 suggest that the JA qrels version 1 was a satisfactory surrogate for version 2: both the relative and absolute performances based on version 1 are quite accurate, assuming version 2 is the ground truth. Figures 4-6 show that the CT qrels version 1 also does relatively well.

6 Correlation between pseudo-qrels and Version 2

Prior to conducting relevance assessments at NTCIR-7, we generated *pseudo-qrels* using the pooled documents: For each topic, top 10 documents from the *sorted pool*, in which each document was ranked by the number of runs containing it (the larger the better) and then by the sum of ranks of the document within those runs (the smaller the better), were treated as relevant. In this paper, we refer to the pseudo-qrels data as *size-10* pseudo-qrels.

Table 28 compares two system rankings based on the same evaluation metric, but with size-10 pseudo-qrels and qrels version 2. The YAR correlation values are computed by treating the version 2 ranking as the ground truth. This corresponds to Table 27 in the Overview.

Figures 10-18 visualise, for each metric, how the performance values/rankings based on size-10 pseudo-qrels deviate from those based on version 2. In particular, Figures 10, 13 and 16, the Mean AP graphs, correspond to Figures 12, 13 and 14 in the Overview, respectively. As Figures 10-12 show, the size-10 pseudo-qrels file is not very useful for the CS runs. Whereas, Figures 13-15 show that the size-10 pseudo-qrels file mimics qrels version 2 reasonably well for the CT runs. As for the JA runs (Figures 16-18), the size-10 pseudo-qrels file predicts the ranking of the low performers relatively accurately, but is not good for predicting the ranking of the top performers.

Table 29 compares two *topic* rankings based on the same evaluation metric, but with size-10 pseudo-qrels and qrels version 2, respectively. The YAR correlation values are computed by treating the version 2 ranking as the ground truth. This corresponds to Table 28 in the Overview. The values are generally lower than those in Table 28: the size-10 pseudo-qrels files are not good at predicting topic difficulty.

The size-10 pseudo-qrels assumed that the number of relevant documents was 10 for every topic. However, according to qrels version 2, the average number of relevant ($L1$ and $L2$) documents per topic is 169.8 for CS, 90.7 for CT, and 110.1 for JA, as was mentioned earlier. Hence size-10 pseudo-qrels underestimated this number substantially, especially for CS. In another paper [3], We will report on new pseudo-qrels experiments that involve *size 100* pseudo-qrels, and show that the new pseudo-qrels files are relatively accurate at predicting the "true" system rankings.

Table 22. Performance averaged across CS runs for each topic, using qrels version 2. For example, "CS-T80" denotes the topic ACLIA-CS-T80. The topics are sorted by the average performance. (NOTE: The rightmost column of the original table (Table 38 in the Overview) is not valid: It shows the most difficult topics in terms of nDCG although it was meant to show the easiest topics. This new table corrects this mistake.)

	AP		AP		Q		Q		nDCG		nDCG
CS-T80	0.7582	CS-T378	0.4405	CS-T80	0.7096	CS-T93	0.4528	CS-T329	0.8326	CS-T320	0.6508
CS-T348	0.7249	CS-T93	0.4345	CS-T348	0.7047	CS-T55	0.4516	CS-T68	0.8207	CS-T104	0.6424
CS-T340	0.6935	CS-T90	0.4334	CS-T340	0.6885	CS-T41	0.4447	CS-T81	0.8024	CS-T369	0.6342
CS-T68	0.6629	CS-T101	0.4331	CS-T329	0.6698	CS-T321	0.4434	CS-T348	0.8006	CS-T92	0.6325
CS-T329	0.6604	CS-T41	0.4246	CS-T68	0.6675	CS-T104	0.4404	CS-T340	0.7959	CS-T90	0.6321
CS-T370	0.6603	CS-T92	0.4221	CS-T81	0.6631	CS-T90	0.4261	CS-T71	0.7893	CS-T55	0.6275
CS-T81	0.6583	CS-T384	0.4202	CS-T339	0.6341	CS-T92	0.4160	CS-T80	0.7757	CS-T317	0.6191
CS-T361	0.6404	CS-T94	0.4041	CS-T43	0.6335	CS-T94	0.4002	CS-T60	0.7744	CS-T46	0.5999
CS-T85	0.6363	CS-T349	0.3980	CS-T60	0.6269	CS-T317	0.3804	CS-T43	0.7723	CS-T99	0.5941
CS-T351	0.6360	CS-T376	0.3967	CS-T71	0.6168	CS-T383	0.3711	CS-T52	0.7603	CS-T357	0.5930
CS-T60	0.6324	CS-T317	0.3791	CS-T370	0.6114	CS-T46	0.3625	CS-T339	0.7602	CS-T380	0.5900
CS-T43	0.6319	CS-T379	0.3685	CS-T69	0.6006	CS-T376	0.3615	CS-T370	0.7546	CS-T368	0.5856
CS-T69	0.6271	CS-T46	0.3498	CS-T351	0.5979	CS-T99	0.3541	CS-T75	0.7527	CS-T94	0.5817
CS-T339	0.6226	CS-T383	0.3472	CS-T85	0.5967	CS-T380	0.3515	CS-T381	0.7496	CS-T385	0.5809
CS-T324	0.6217	CS-T95	0.3457	CS-T338	0.5963	CS-T95	0.3414	CS-T338	0.7480	CS-T321	0.5808
CS-T97	0.5890	CS-T99	0.3433	CS-T361	0.5848	CS-T368	0.3362	CS-T351	0.7479	CS-T350	0.5740
CS-T71	0.5859	CS-T380	0.3368	CS-T336	0.5770	CS-T385	0.3328	CS-T85	0.7478	CS-T376	0.5721
CS-T338	0.5837	CS-T368	0.3252	CS-T52	0.5768	CS-T379	0.3313	CS-T326	0.7434	CS-T67	0.5529
CS-T323	0.5808	CS-T385	0.3247	CS-T75	0.5720	CS-T349	0.3268	CS-T97	0.7397	CS-T358	0.5436
CS-T53	0.5770	CS-T83	0.3202	CS-T365	0.5712	CS-T357	0.3154	CS-T336	0.7363	CS-T79	0.5397
CS-T52	0.5727	CS-T350	0.3066	CS-T97	0.5708	CS-T350	0.3140	CS-T58	0.7347	CS-T383	0.5341
CS-T352	0.5511	CS-T332	0.2956	CS-T103	0.5667	CS-T332	0.3036	CS-T361	0.7342	CS-T379	0.5299
CS-T336	0.5475	CS-T322	0.2933	CS-T352	0.5646	CS-T83	0.2998	CS-T384	0.7323	CS-T349	0.5282
CS-T103	0.5472	CS-T57	0.2864	CS-T53	0.5596	CS-T57	0.2904	CS-T378	0.7319	CS-T57	0.5278
CS-T65	0.5460	CS-T357	0.2802	CS-T324	0.5584	CS-T358	0.2865	CS-T98	0.7310	CS-T102	0.5233
CS-T98	0.5455	CS-T64	0.2778	CS-T49	0.5523	CS-T67	0.2804	CS-T49	0.7271	CS-T332	0.5139
CS-T58	0.5398	CS-T333	0.2734	CS-T58	0.5460	CS-T79	0.2731	CS-T352	0.7243	CS-T83	0.5094
CS-T75	0.5378	CS-T358	0.2511	CS-T98	0.5399	CS-T64	0.2684	CS-T103	0.7191	CS-T366	0.4817
CS-T365	0.5341	CS-T67	0.2500	CS-T381	0.5220	CS-T333	0.2656	CS-T324	0.7078	CS-T47	0.4770
CS-T49	0.5283	CS-T366	0.2470	CS-T65	0.5155	CS-T322	0.2567	CS-T76	0.7050	CS-T64	0.4730
CS-T96	0.5213	CS-T102	0.2450	CS-T326	0.5106	CS-T102	0.2533	CS-T355	0.7050	CS-T54	0.4639
CS-T325	0.5183	CS-T79	0.2448	CS-T347	0.5085	CS-T366	0.2493	CS-T347	0.7001	CS-T95	0.4521
CS-T347	0.5154	CS-T47	0.2421	CS-T325	0.5077	CS-T359	0.2348	CS-T89	0.6995	CS-T44	0.4485
CS-T55	0.5054	CS-T359	0.2295	CS-T367	0.5054	CS-T47	0.2301	CS-T69	0.6994	CS-T82	0.4428
CS-T328	0.5045	CS-T334	0.2266	CS-T323	0.5053	CS-T334	0.2194	CS-T84	0.6980	CS-T322	0.4409
CS-T369	0.4980	CS-T87	0.2212	CS-T337	0.4943	CS-T54	0.2185	CS-T93	0.6972	CS-T48	0.4401
CS-T84	0.4958	CS-T44	0.2072	CS-T328	0.4906	CS-T87	0.2113	CS-T367	0.6964	CS-T333	0.4390
CS-T321	0.4950	CS-T77	0.1961	CS-T378	0.4890	CS-T44	0.2096	CS-T365	0.6958	CS-T359	0.4376
CS-T91	0.4938	CS-T54	0.1884	CS-T355	0.4879	CS-T77	0.1955	CS-T41	0.6928	CS-T87	0.4157
CS-T89	0.4908	CS-T82	0.1848	CS-T96	0.4876	CS-T82	0.1926	CS-T61	0.6901	CS-T334	0.3830
CS-T337	0.4857	CS-T42	0.1628	CS-T76	0.4855	CS-T48	0.1852	CS-T101	0.6815	CS-T77	0.3733
CS-T320	0.4852	CS-T56	0.1600	CS-T91	0.4811	CS-T56	0.1842	CS-T96	0.6796	CS-T56	0.3682
CS-T326	0.4839	CS-T48	0.1539	CS-T101	0.4804	CS-T42	0.1512	CS-T328	0.6789	CS-T73	0.3296
CS-T76	0.4809	CS-T62	0.1425	CS-T84	0.4783	CS-T62	0.1432	CS-T100	0.6737	CS-T42	0.2957
CS-T367	0.4774	CS-T74	0.0999	CS-T89	0.4759	CS-T73	0.1132	CS-T337	0.6732	CS-T62	0.2890
CS-T381	0.4726	CS-T73	0.0954	CS-T320	0.4740	CS-T74	0.1069	CS-T65	0.6725	CS-T74	0.2871
CS-T100	0.4683	CS-T78	0.0779	CS-T384	0.4726	CS-T78	0.1033	CS-T53	0.6721	CS-T78	0.2705
CS-T104	0.4611			CS-T100	0.4660			CS-T91	0.6700		
CS-T355	0.4555			CS-T61	0.4629			CS-T325	0.6654		
CS-T61	0.4514			CS-T369	0.4599			CS-T323	0.6613		

Table 23. Performance averaged across CT runs for each topic, using qrels version 2. For example, "CT-T408" denotes the topic ACLIA-CT-T408. The topics are sorted by the average performance.

	AP		AP		Q		Q		nDCG		nDCG
CT-T408	0.7621	CT-T436	0.3315	CT-T408	0.7239	CT-T193	0.3657	CT-T428	0.8990	CT-T404	0.5995
CT-T194	0.7140	CT-T193	0.3288	CT-T428	0.7133	CT-T410	0.3633	CT-T194	0.8711	CT-T391	0.5963
CT-T396	0.6982	CT-T179	0.3271	CT-T194	0.7107	CT-T405	0.3606	CT-T408	0.8582	CT-T413	0.5908
CT-T374	0.6880	CT-T434	0.3220	CT-T431	0.6926	CT-T402	0.3503	CT-T200	0.8507	CT-T193	0.5903
CT-T428	0.6838	CT-T210	0.3209	CT-T374	0.6779	CT-T182	0.3454	CT-T374	0.8437	CT-T176	0.5873
CT-T398	0.6610	CT-T186	0.3208	CT-T396	0.6761	CT-T409	0.3359	CT-T431	0.8312	CT-T189	0.5816
CT-T200	0.6582	CT-T411	0.3149	CT-T200	0.6621	CT-T186	0.3275	CT-T396	0.8158	CT-T418	0.5799
CT-T427	0.6556	CT-T184	0.3077	CT-T432	0.6608	CT-T178	0.3265	CT-T398	0.8018	CT-T207	0.5756
CT-T431	0.6462	CT-T404	0.3060	CT-T427	0.6591	CT-T404	0.3202	CT-T213	0.7972	CT-T184	0.5721
CT-T387	0.6423	CT-T183	0.3058	CT-T398	0.6453	CT-T420	0.3136	CT-T432	0.7947	CT-T443	0.5551
CT-T394	0.6260	CT-T208	0.2934	CT-T213	0.6336	CT-T174	0.3127	CT-T181	0.7947	CT-T429	0.5518
CT-T432	0.6237	CT-T409	0.2810	CT-T394	0.6254	CT-T183	0.3068	CT-T394	0.7882	CT-T390	0.5480
CT-T197	0.6121	CT-T178	0.2798	CT-T197	0.6006	CT-T206	0.3018	CT-T427	0.7809	CT-T402	0.5417
CT-T213	0.6102	CT-T206	0.2760	CT-T395	0.5885	CT-T184	0.3001	CT-T197	0.7686	CT-T208	0.5369
CT-T195	0.5965	CT-T174	0.2709	CT-T387	0.5861	CT-T418	0.2944	CT-T400	0.7611	CT-T186	0.5326
CT-T386	0.5700	CT-T450	0.2639	CT-T195	0.5660	CT-T411	0.2930	CT-T423	0.7591	CT-T437	0.5306
CT-T400	0.5635	CT-T418	0.2614	CT-T187	0.5515	CT-T450	0.2854	CT-T426	0.7547	CT-T410	0.5293
CT-T395	0.5624	CT-T437	0.2583	CT-T386	0.5485	CT-T208	0.2849	CT-T175	0.7546	CT-T172	0.5187
CT-T391	0.5143	CT-T443	0.2544	CT-T400	0.5457	CT-T437	0.2761	CT-T407	0.7527	CT-T388	0.5157
CT-T449	0.4961	CT-T420	0.2544	CT-T181	0.5329	CT-T443	0.2759	CT-T449	0.7346	CT-T411	0.5110
CT-T407	0.4907	CT-T189	0.2235	CT-T423	0.5282	CT-T429	0.2514	CT-T387	0.7272	CT-T206	0.5053
CT-T187	0.4873	CT-T430	0.2175	CT-T449	0.5280	CT-T189	0.2513	CT-T395	0.7217	CT-T450	0.4958
CT-T392	0.4867	CT-T448	0.1875	CT-T407	0.5191	CT-T430	0.2480	CT-T187	0.7214	CT-T180	0.4900
CT-T181	0.4854	CT-T172	0.1866	CT-T392	0.5183	CT-T172	0.2401	CT-T447	0.7156	CT-T424	0.4869
CT-T422	0.4814	CT-T429	0.1744	CT-T422	0.5165	CT-T448	0.2074	CT-T389	0.7148	CT-T183	0.4706
CT-T438	0.4742	CT-T180	0.1734	CT-T389	0.4986	CT-T180	0.2024	CT-T445	0.7093	CT-T446	0.4659
CT-T440	0.4738	CT-T444	0.1577	CT-T440	0.4919	CT-T446	0.1914	CT-T386	0.7058	CT-T414	0.4653
CT-T423	0.4722	CT-T419	0.1524	CT-T391	0.4848	CT-T191	0.1884	CT-T451	0.7055	CT-T435	0.4622
CT-T393	0.4599	CT-T191	0.1524	CT-T175	0.4647	CT-T419	0.1849	CT-T434	0.6933	CT-T430	0.4572
CT-T389	0.4551	CT-T414	0.1483	CT-T447	0.4623	CT-T424	0.1750	CT-T422	0.6875	CT-T448	0.4483
CT-T399	0.4377	CT-T435	0.1479	CT-T393	0.4622	CT-T414	0.1744	CT-T210	0.6772	CT-T442	0.4223
CT-T211	0.4333	CT-T424	0.1407	CT-T399	0.4598	CT-T444	0.1687	CT-T440	0.6694	CT-T416	0.4166
CT-T447	0.4273	CT-T442	0.1401	CT-T438	0.4518	CT-T435	0.1664	CT-T392	0.6673	CT-T419	0.4149
CT-T390	0.4113	CT-T446	0.1345	CT-T426	0.4494	CT-T442	0.1525	CT-T438	0.6620	CT-T190	0.3992
CT-T445	0.4094	CT-T198	0.1272	CT-T451	0.4450	CT-T421	0.1421	CT-T393	0.6610	CT-T444	0.3921
CT-T426	0.4091	CT-T177	0.1190	CT-T211	0.4379	CT-T198	0.1346	CT-T439	0.6591	CT-T191	0.3845
CT-T175	0.4071	CT-T204	0.0971	CT-T176	0.4149	CT-T177	0.1302	CT-T420	0.6563	CT-T177	0.3747
CT-T171	0.4058	CT-T421	0.0964	CT-T445	0.4133	CT-T416	0.1273	CT-T436	0.6545	CT-T203	0.3541
CT-T451	0.3947	CT-T203	0.0935	CT-T171	0.4048	CT-T203	0.1154	CT-T195	0.6475	CT-T188	0.3402
CT-T397	0.3889	CT-T416	0.0934	CT-T439	0.4018	CT-T188	0.1112	CT-T178	0.6465	CT-T205	0.3346
CT-T176	0.3851	CT-T205	0.0924	CT-T397	0.3998	CT-T190	0.1108	CT-T211	0.6345	CT-T198	0.3229
CT-T207	0.3814	CT-T188	0.0895	CT-T434	0.3914	CT-T204	0.1070	CT-T179	0.6261	CT-T421	0.3208
CT-T388	0.3785	CT-T190	0.0807	CT-T210	0.3909	CT-T205	0.1064	CT-T171	0.6211	CT-T415	0.2944
CT-T439	0.3743	CT-T196	0.0752	CT-T436	0.3871	CT-T196	0.0909	CT-T409	0.6202	CT-T196	0.2679
CT-T402	0.3592	CT-T415	0.0623	CT-T388	0.3819	CT-T415	0.0836	CT-T399	0.6165	CT-T204	0.2424
CT-T441	0.3559			CT-T207	0.3766			CT-T397	0.6130		
CT-T405	0.3509			CT-T413	0.3757			CT-T182	0.6119		
CT-T182	0.3418			CT-T390	0.3755			CT-T441	0.6100		
CT-T410	0.3396			CT-T441	0.3725			CT-T174	0.6048		
CT-T413	0.3351			CT-T179	0.3705			CT-T405	0.6000		

Table 24. Performance averaged across JA runs for each topic, using qrels version 2. For example, "JA-T242" denotes the topic ACLIA-JA-T242. The topics are sorted by the average performance.

	AP		AP		Q		Q		nDCG		nDCG
JA-T242	0.7829	JA-T152	0.4274	JA-T242	0.7866	JA-T7	0.4440	JA-T107	0.8704	JA-T113	0.6127
JA-T107	0.7557	JA-T1	0.4237	JA-T107	0.7467	JA-T313	0.4433	JA-T217	0.8626	JA-T151	0.6117
JA-T271	0.7466	JA-T305	0.4215	JA-T271	0.7411	JA-T110	0.4324	JA-T242	0.8348	JA-T224	0.6110
JA-T221	0.7466	JA-T113	0.4214	JA-T221	0.6866	JA-T112	0.4318	JA-T271	0.8286	JA-T244	0.6089
JA-T6	0.7309	JA-T110	0.4184	JA-T217	0.6804	JA-T300	0.4276	JA-T221	0.7826	JA-T304	0.6065
JA-T9	0.6942	JA-T223	0.4180	JA-T29	0.6699	JA-T230	0.4198	JA-T134	0.7811	JA-T222	0.6005
JA-T38	0.6884	JA-T153	0.4152	JA-T111	0.6676	JA-T2	0.4198	JA-T25	0.7736	JA-T115	0.5998
JA-T29	0.6643	JA-T112	0.4143	JA-T6	0.6481	JA-T163	0.4166	JA-T250	0.7680	JA-T110	0.5963
JA-T111	0.6509	JA-T300	0.4081	JA-T9	0.6450	JA-T162	0.4161	JA-T275	0.7628	JA-T225	0.5938
JA-T217	0.6464	JA-T230	0.3945	JA-T38	0.6411	JA-T153	0.4120	JA-T3	0.7595	JA-T17	0.5898
JA-T149	0.6405	JA-T17	0.3905	JA-T254	0.6396	JA-T151	0.4018	JA-T149	0.7557	JA-T245	0.5858
JA-T267	0.6400	JA-T151	0.3871	JA-T250	0.6350	JA-T17	0.3972	JA-T111	0.7555	JA-T313	0.5853
JA-T254	0.6307	JA-T161	0.3827	JA-T275	0.6314	JA-T222	0.3862	JA-T9	0.7508	JA-T295	0.5771
JA-T141	0.6231	JA-T162	0.3759	JA-T267	0.6267	JA-T225	0.3815	JA-T29	0.7387	JA-T4	0.5765
JA-T275	0.6173	JA-T164	0.3742	JA-T134	0.6261	JA-T305	0.3795	JA-T233	0.7370	JA-T154	0.5729
JA-T233	0.6071	JA-T4	0.3693	JA-T149	0.6215	JA-T13	0.3771	JA-T254	0.7368	JA-T7	0.5715
JA-T250	0.6008	JA-T222	0.3659	JA-T141	0.6177	JA-T167	0.3758	JA-T141	0.7340	JA-T163	0.5644
JA-T134	0.5989	JA-T225	0.3604	JA-T25	0.5913	JA-T161	0.3754	JA-T267	0.7309	JA-T2	0.5640
JA-T130	0.5959	JA-T236	0.3562	JA-T233	0.5873	JA-T248	0.3750	JA-T253	0.7264	JA-T119	0.5581
JA-T10	0.5863	JA-T148	0.3528	JA-T130	0.5855	JA-T4	0.3659	JA-T239	0.7264	JA-T305	0.5441
JA-T253	0.5728	JA-T13	0.3518	JA-T253	0.5725	JA-T291	0.3608	JA-T38	0.7243	JA-T13	0.5439
JA-T245	0.5661	JA-T167	0.3488	JA-T245	0.5709	JA-T148	0.3509	JA-T223	0.6922	JA-T255	0.5416
JA-T218	0.5654	JA-T255	0.3404	JA-T3	0.5694	JA-T164	0.3474	JA-T231	0.6879	JA-T153	0.5363
JA-T3	0.5637	JA-T248	0.3385	JA-T218	0.5682	JA-T128	0.3363	JA-T152	0.6864	JA-T301	0.5307
JA-T25	0.5602	JA-T215	0.3302	JA-T10	0.5646	JA-T255	0.3332	JA-T6	0.6858	JA-T128	0.5218
JA-T35	0.5523	JA-T291	0.3276	JA-T19	0.5525	JA-T215	0.3265	JA-T15	0.6855	JA-T215	0.5157
JA-T19	0.5522	JA-T138	0.2961	JA-T276	0.5504	JA-T295	0.3256	JA-T300	0.6825	JA-T236	0.5141
JA-T106	0.5498	JA-T301	0.2923	JA-T234	0.5487	JA-T138	0.3187	JA-T234	0.6824	JA-T237	0.5116
JA-T304	0.5464	JA-T128	0.2799	JA-T249	0.5424	JA-T154	0.3154	JA-T218	0.6799	JA-T138	0.5102
JA-T234	0.5443	JA-T295	0.2781	JA-T106	0.5310	JA-T236	0.3143	JA-T238	0.6778	JA-T164	0.5002
JA-T276	0.5424	JA-T32	0.2745	JA-T239	0.5303	JA-T119	0.3128	JA-T106	0.6720	JA-T108	0.4968
JA-T249	0.5274	JA-T154	0.2736	JA-T35	0.5288	JA-T301	0.3083	JA-T266	0.6718	JA-T148	0.4881
JA-T238	0.5111	JA-T119	0.2718	JA-T304	0.5154	JA-T37	0.2877	JA-T10	0.6710	JA-T37	0.4879
JA-T115	0.5004	JA-T37	0.2599	JA-T105	0.5140	JA-T18	0.2815	JA-T19	0.6694	JA-T297	0.4837
JA-T2	0.4974	JA-T108	0.2476	JA-T115	0.5126	JA-T32	0.2752	JA-T249	0.6689	JA-T248	0.4774
JA-T239	0.4935	JA-T240	0.2475	JA-T244	0.5121	JA-T108	0.2686	JA-T230	0.6687	JA-T18	0.4737
JA-T244	0.4915	JA-T18	0.2418	JA-T15	0.5052	JA-T240	0.2596	JA-T162	0.6678	JA-T32	0.4686
JA-T105	0.4889	JA-T160	0.2134	JA-T231	0.4838	JA-T170	0.2154	JA-T109	0.6638	JA-T170	0.4376
JA-T15	0.4867	JA-T157	0.1985	JA-T140	0.4752	JA-T160	0.2113	JA-T105	0.6571	JA-T315	0.4178
JA-T313	0.4751	JA-T166	0.1889	JA-T238	0.4751	JA-T297	0.2042	JA-T140	0.6476	JA-T240	0.3921
JA-T140	0.4732	JA-T170	0.1867	JA-T266	0.4730	JA-T166	0.2042	JA-T1	0.6473	JA-T166	0.3919
JA-T20	0.4719	JA-T297	0.1653	JA-T224	0.4638	JA-T157	0.2040	JA-T155	0.6447	JA-T284	0.3880
JA-T266	0.4645	JA-T158	0.1320	JA-T1	0.4637	JA-T315	0.1573	JA-T130	0.6359	JA-T160	0.3821
JA-T231	0.4634	JA-T284	0.1272	JA-T109	0.4609	JA-T165	0.1540	JA-T35	0.6318	JA-T165	0.3704
JA-T155	0.4582	JA-T315	0.1248	JA-T20	0.4566	JA-T284	0.1521	JA-T161	0.6302	JA-T157	0.3630
JA-T163	0.4574	JA-T165	0.1237	JA-T223	0.4531	JA-T158	0.1186	JA-T276	0.6240	JA-T137	0.3419
JA-T109	0.4531	JA-T137	0.1014	JA-T237	0.4516	JA-T137	0.1175	JA-T167	0.6208	JA-T168	0.2818
JA-T224	0.4516	JA-T168	0.0971	JA-T155	0.4473	JA-T168	0.1052	JA-T112	0.6184	JA-T158	0.2326
JA-T237	0.4431			JA-T152	0.4462			JA-T20	0.6160		
JA-T7	0.4405			JA-T113	0.4457			JA-T291	0.6138		

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**	**.962/.949**	.874/.845
Q	**.962/.949**	**1/1**	**.903**/.866
nDCG	.874/.850	**.903**/.873	**1/1**
CT runs	AP	Q	nDCG
AP	**1/1**	**.945/.957**	.822/.849
Q	**.945/.956**	**1/1**	.877/.888
nDCG	.822/.843	.877/.886	**1/1**
JA runs	AP	Q	nDCG
AP	**1/1**	**.920**/.909	.873/.854
Q	**.920**/.896	**1/1**	**.953/.945**
nDCG	.873/.842	**.953/.945**	**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**	.884/.817	.753/.613
Q	.884/.821	**1/1**	.835/.710
nDCG	.753/.622	.835/.738	**1/1**
CT runs	AP	Q	nDCG
AP	**1/1**	**.908**/.850	.756/.672
Q	**.908**/.845	**1/1**	.814/.746
nDCG	.756/.666	.814/.758	**1/1**
JA runs	AP	Q	nDCG
AP	**1/1**	**.911**/.873	.701/.597
Q	**.911**/.872	**1/1**	.757/.660
nDCG	.701/.597	.757/.668	**1/1**

Table 27. Kendall and Yilmaz/Aslam/Robertson rank correlation: System ranking by version 1 vs version 2 for each metric.

CS runs	
AP	.859/.830
Q	.859/.830
nDCG	**.931**/.880
CT runs	
AP	**.963/.960**
Q	**.945**/.890
nDCG	**.938**/.868
JA runs	
AP	**.973/.979**
Q	**.973/.962**
nDCG	**.967/.931**

Table 28. Kendall and Yilmaz/Aslam/Robertson rank correlation: System ranking by size-10 pseudo-qrels vs qrels version 2 for each metric.

CS runs	
AP	.621/.610
Q	.646/.608
nDCG	.700/.675
CT runs	
AP	.760/.628
Q	.748/.660
nDCG	.686/.610
JA runs	
AP	.753/.530
Q	.693/.470
nDCG	.753/.508

Table 29. Kendall and Yilmaz/Aslam/Robertson rank correlation: Topic ranking by size-10 pseudo-qrels vs qrels version 2 for each metric.

CS topics	
AP	.312/.188
Q	.366/.236
nDCG	.474/.352
CT topics	
AP	.363/.225
Q	.405/.263
nDCG	.480/.323
JA topics	
AP	.429/.423
Q	.448/.413
nDCG	.479/.439

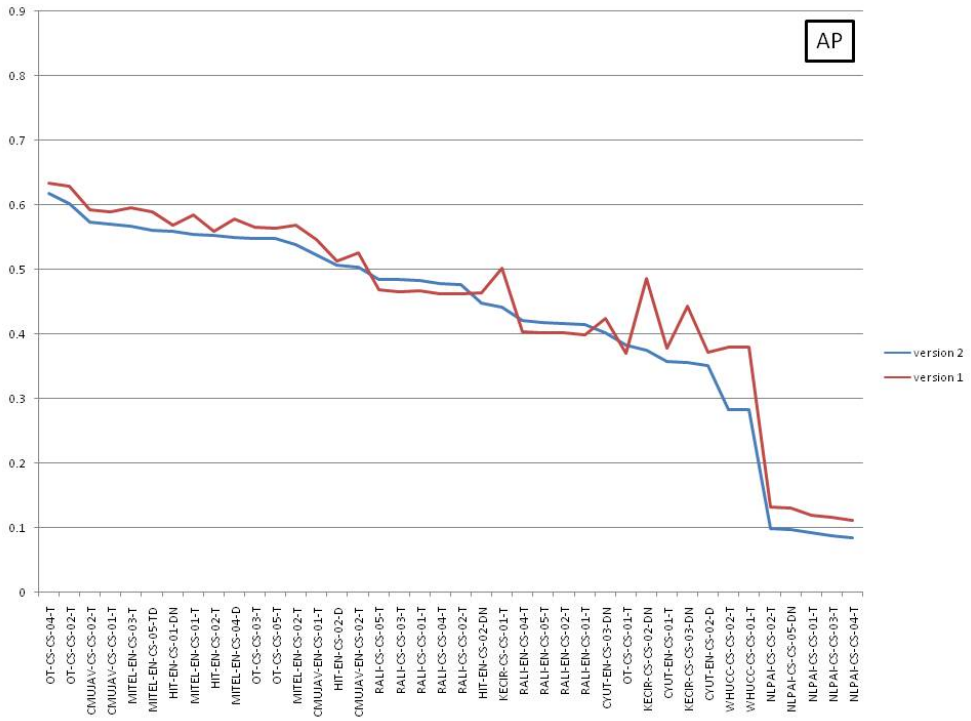

Figure 1. Mean AP values for CS runs based on qrels versions 1 and 2: Runs sorted by Mean AP based on qrels version 2.

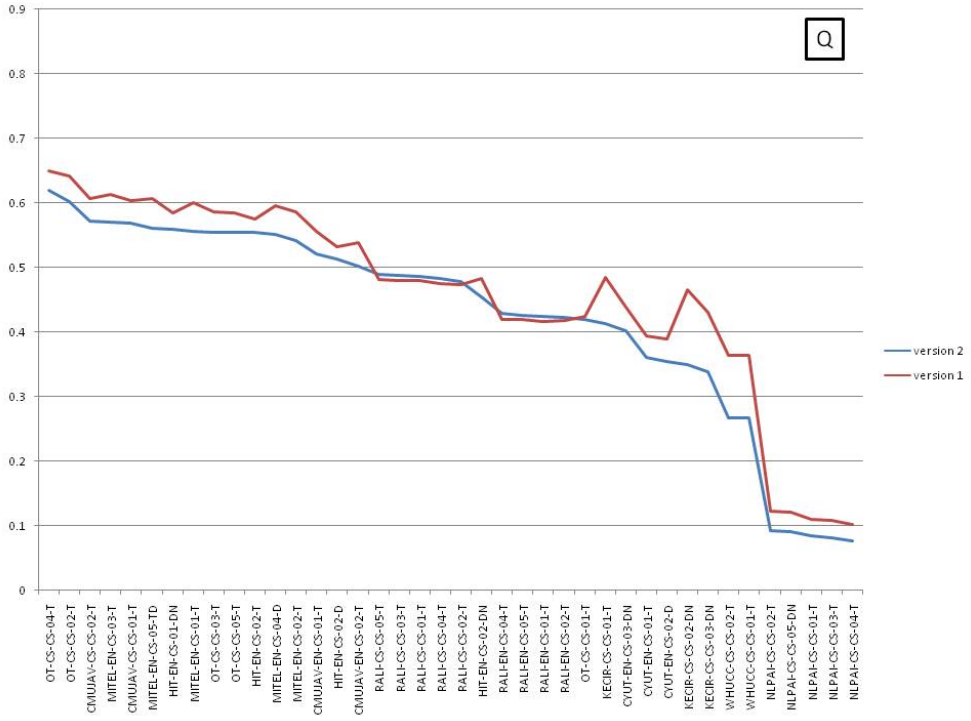

Figure 2. Mean Q values for CS runs based on qrels versions 1 and 2: Runs sorted by Mean Q based on qrels version 2.

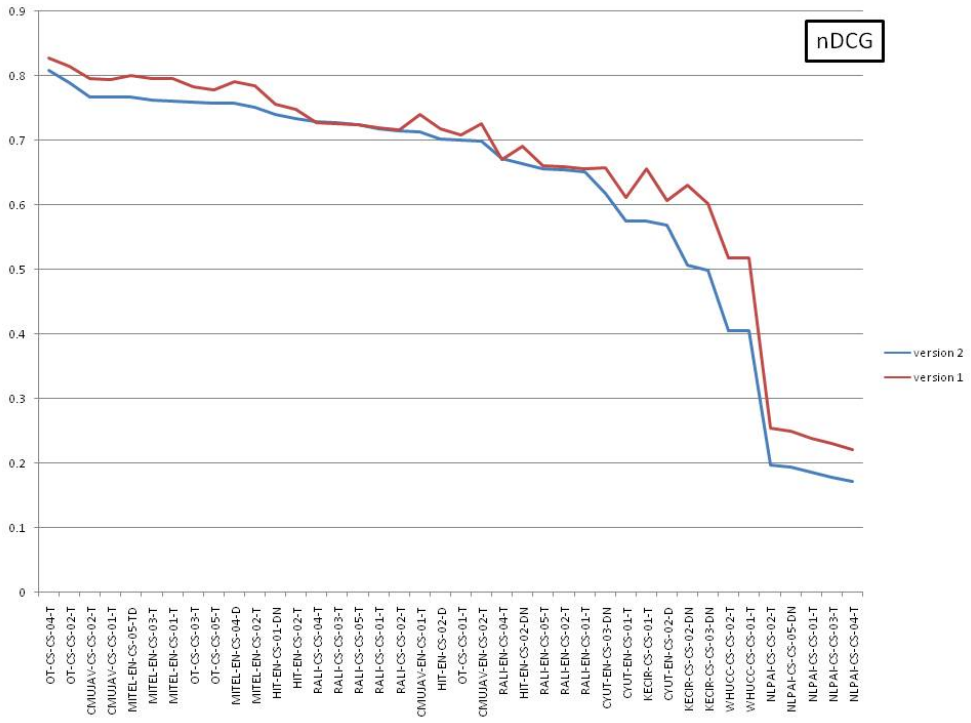

Figure 3. Mean nDCG values for CS runs based on qrels versions 1 and 2: Runs sorted by Mean nDCG based on qrels version 2.

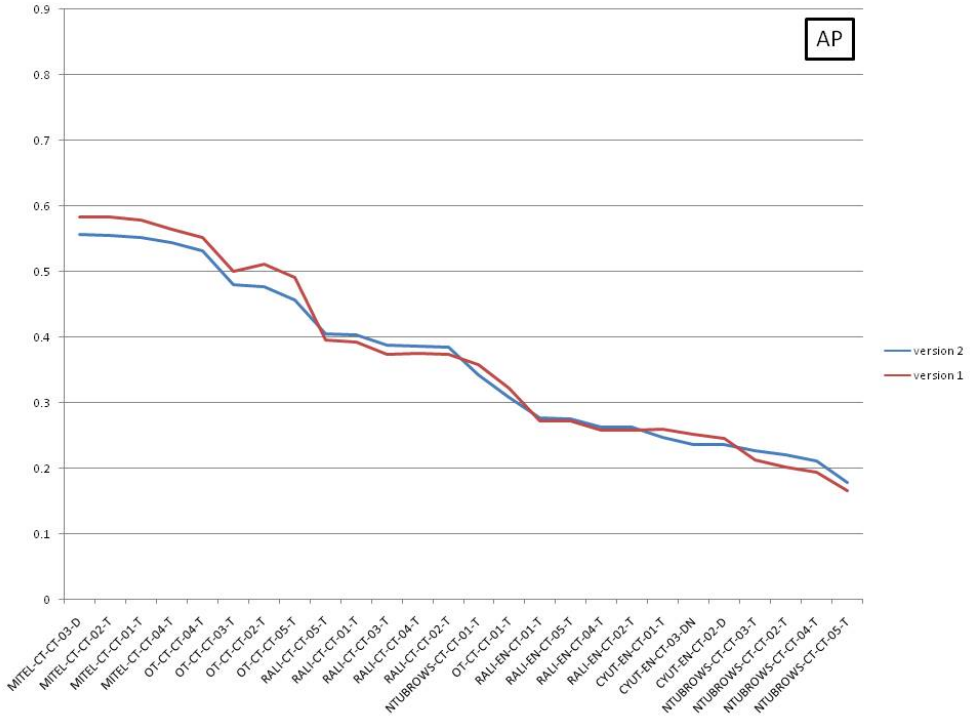

Figure 4. Mean AP values for CT runs based on qrels versions 1 and 2: Runs sorted by Mean AP based on qrels version 2.

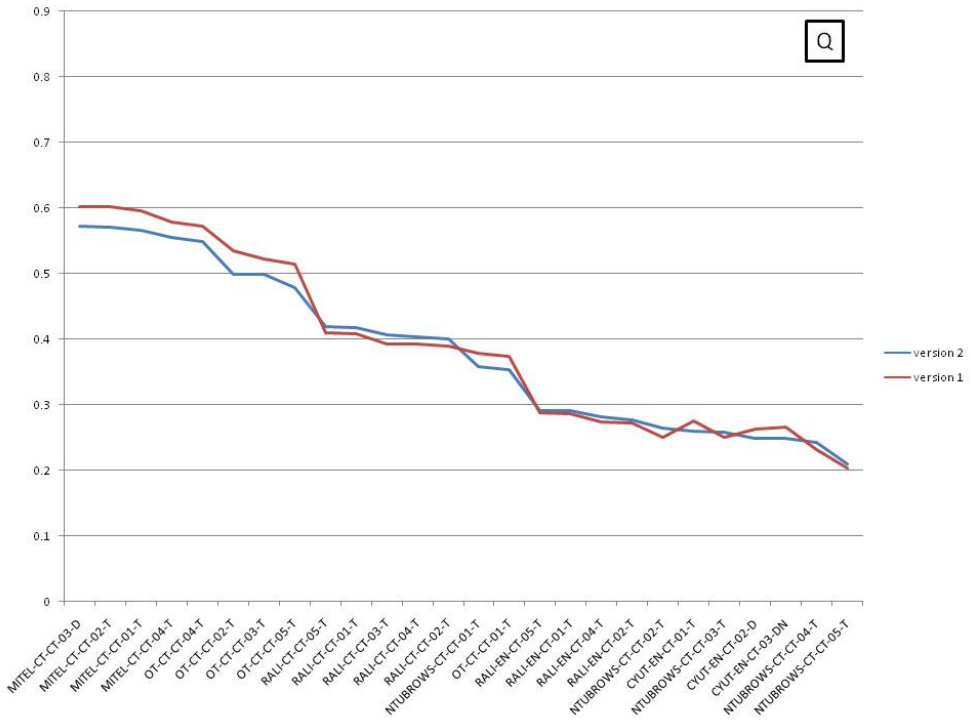

Figure 5. Mean Q values for CT runs based on qrels versions 1 and 2: Runs sorted by Mean Q based on qrels version 2.

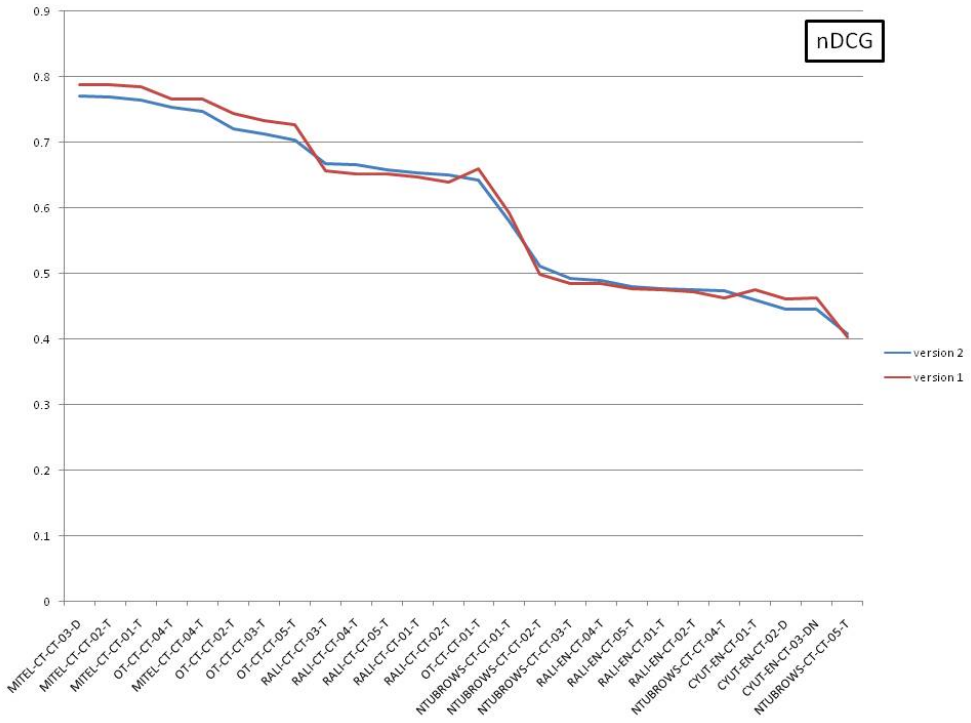

Figure 6. Mean nDCG values for CT runs based on qrels versions 1 and 2: Runs sorted by Mean nDCG based on qrels version 2.

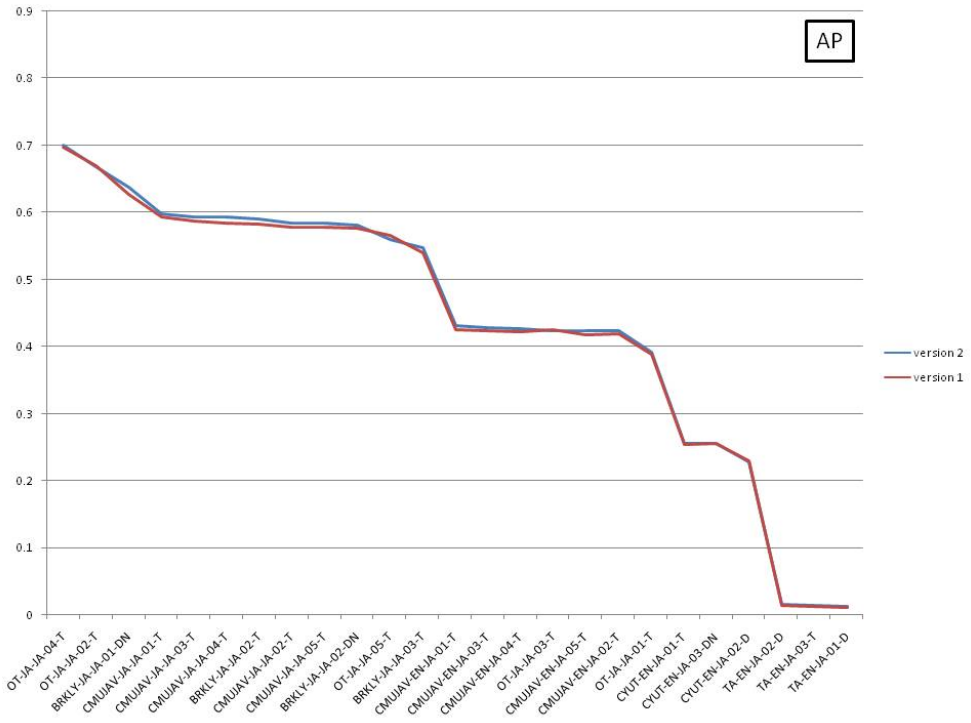

Figure 7. Mean AP values for JA runs based on qrels versions 1 and 2: Runs sorted by Mean AP based on qrels version 2.

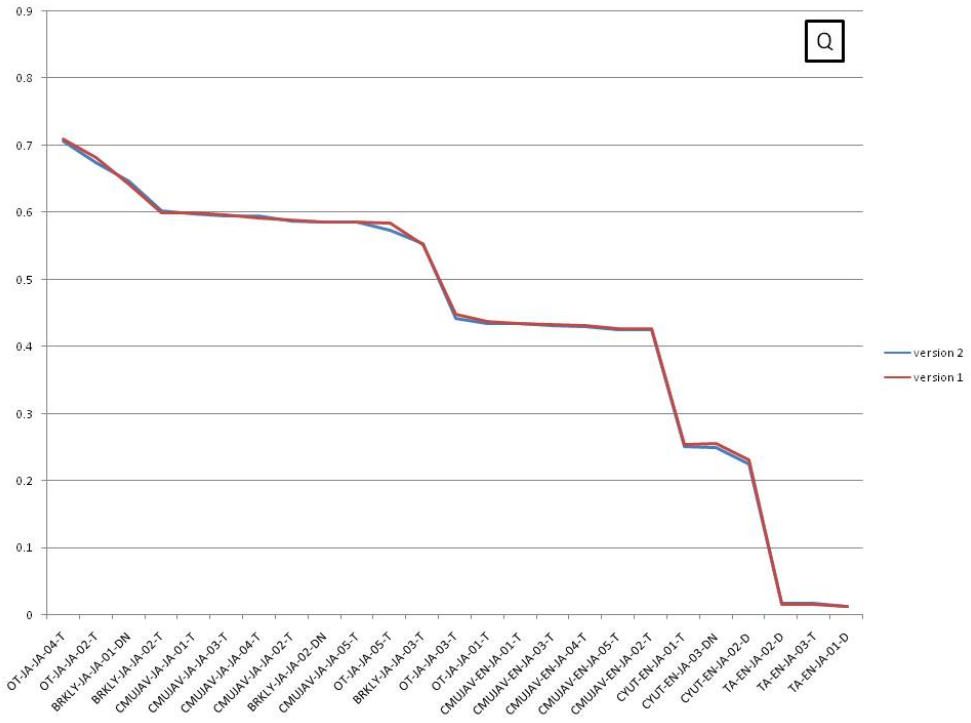

Figure 8. Mean Q values for JA runs based on qrels versions 1 and 2: Runs sorted by Mean Q based on qrels version 2.

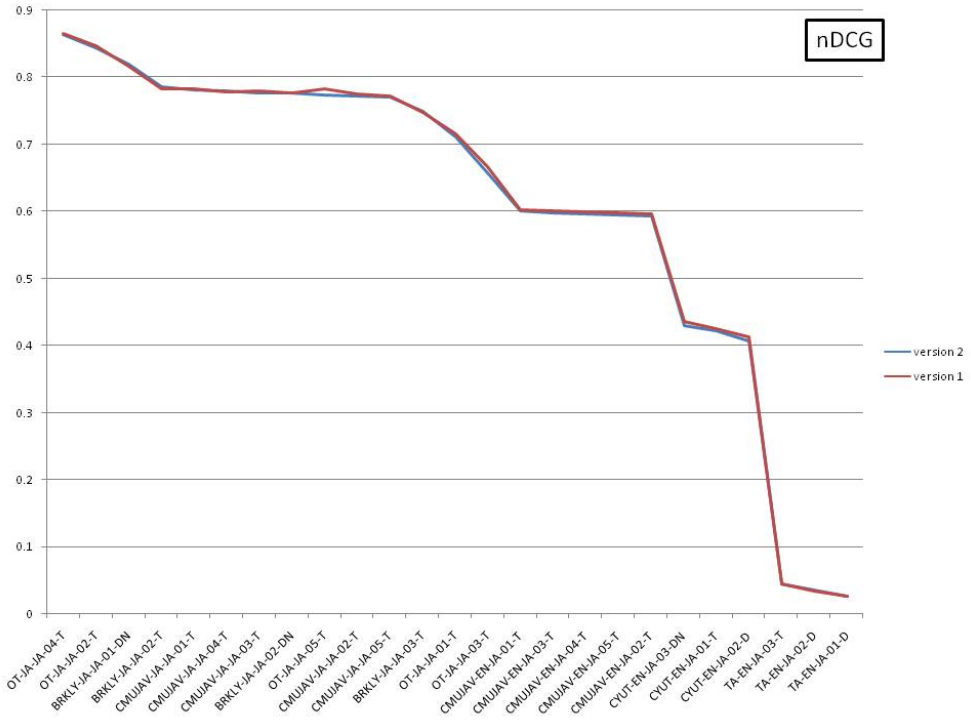

Figure 9. Mean nDCG values for JA runs based on qrels versions 1 and 2: Runs sorted by Mean nDCG based on qrels version 2.

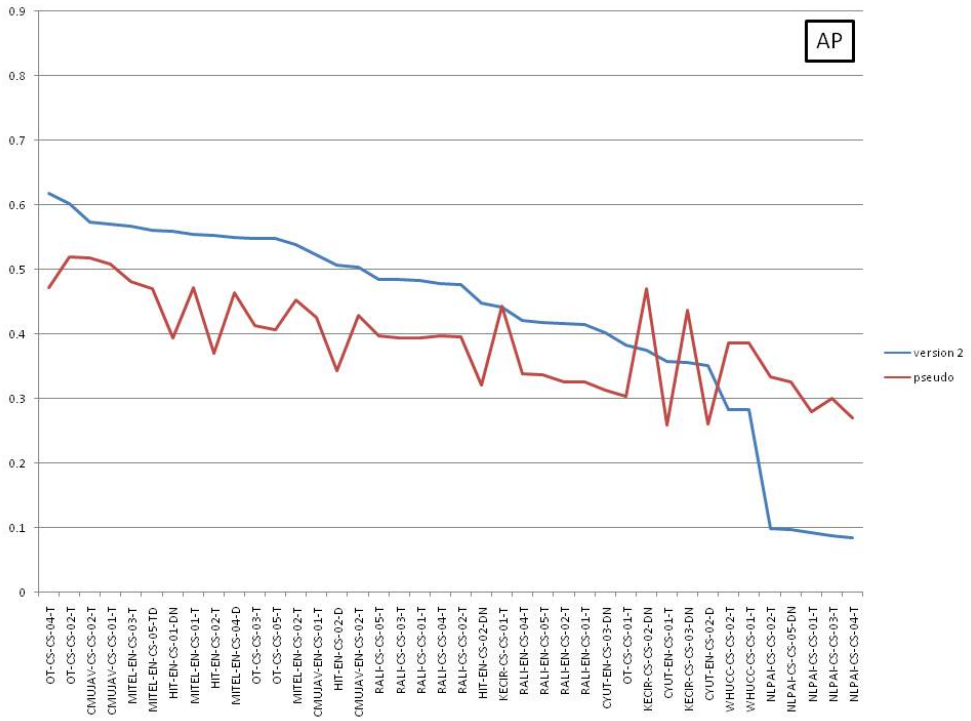

Figure 10. Mean AP values for CS runs based on size-10 pseudo-qrels and qrels versions 2: Runs sorted by Mean AP based on qrels version 2.

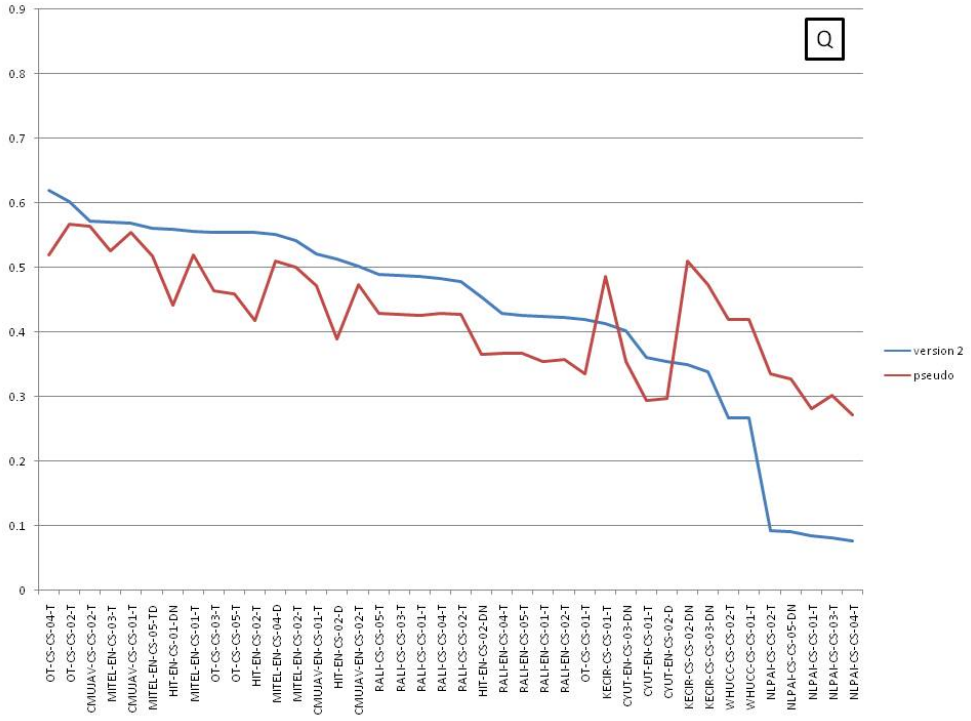

Figure 11. Mean Q values for CS runs based on size-10 pseudo-qrels and qrels versions 2: Runs sorted by Mean Q based on qrels version 2.

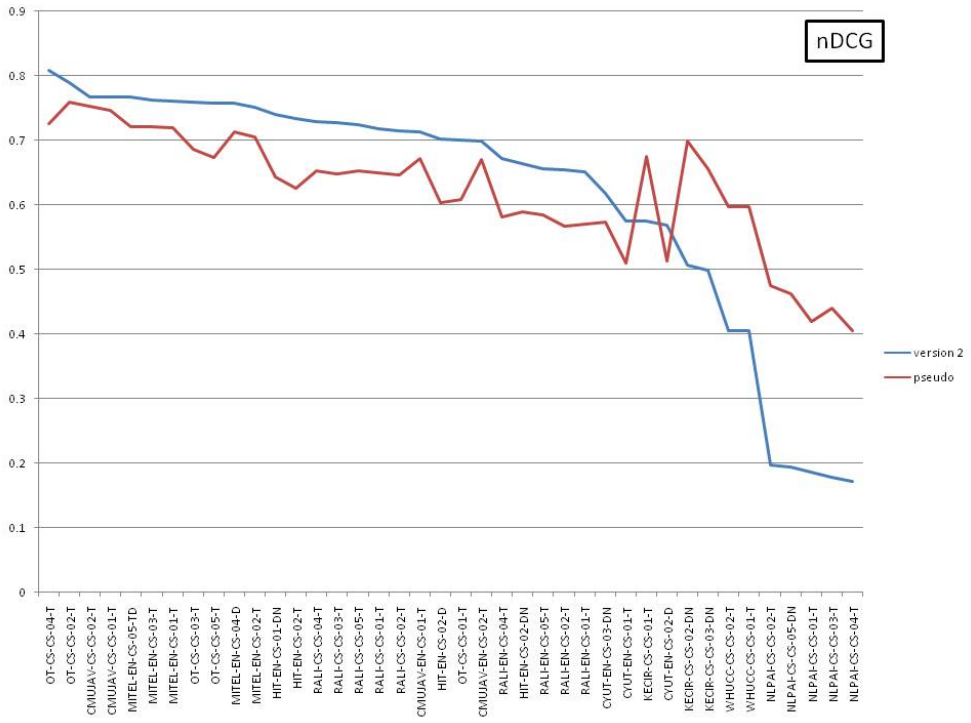

Figure 12. Mean nDCG values for CS runs based on size-10 pseudo-qrels and qrels versions 2: Runs sorted by Mean nDCG based on qrels version 2.

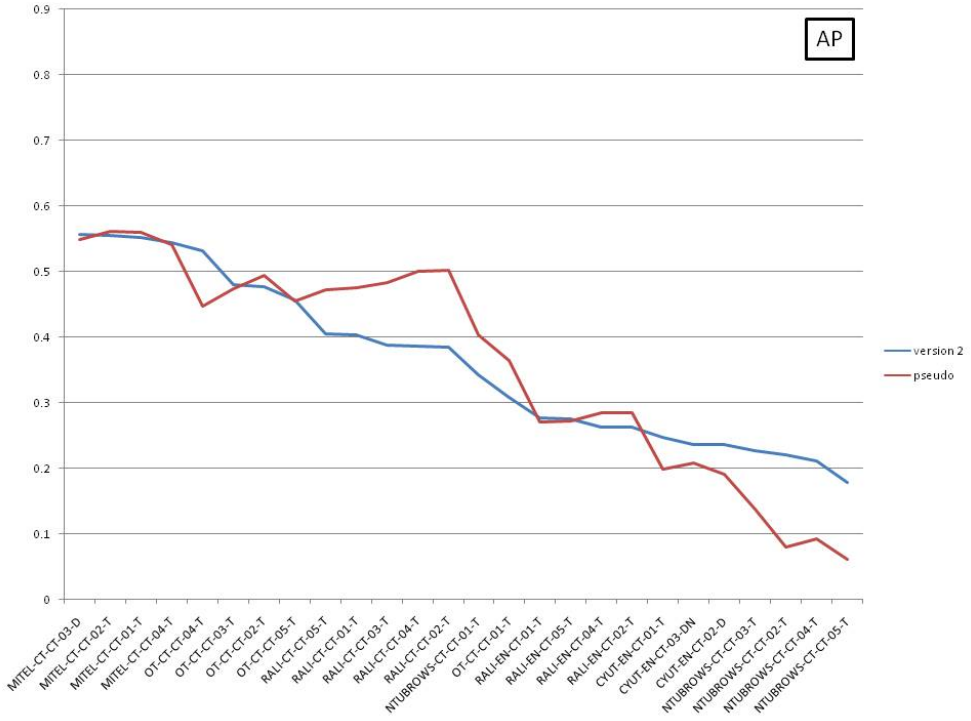

Figure 13. Mean AP values for CT runs based on size-10 pseudo-qrels and qrels versions 2: Runs sorted by Mean AP based on qrels version 2.

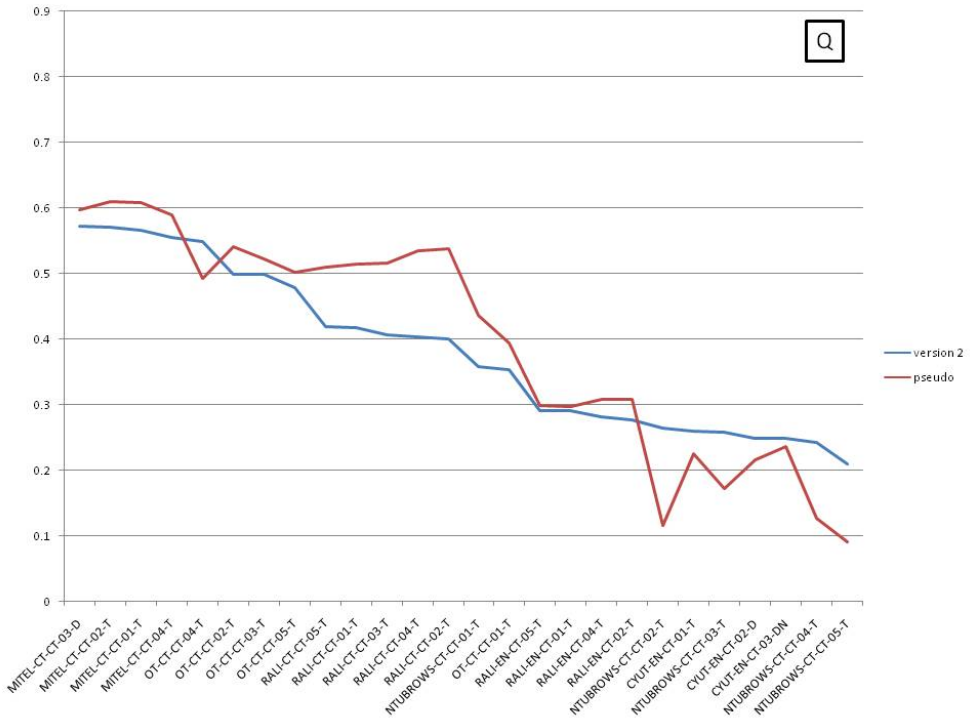

Figure 14. Mean Q values for CT runs based on size-10 pseudo-qrels and qrels versions 2: Runs sorted by Mean Q based on qrels version 2.

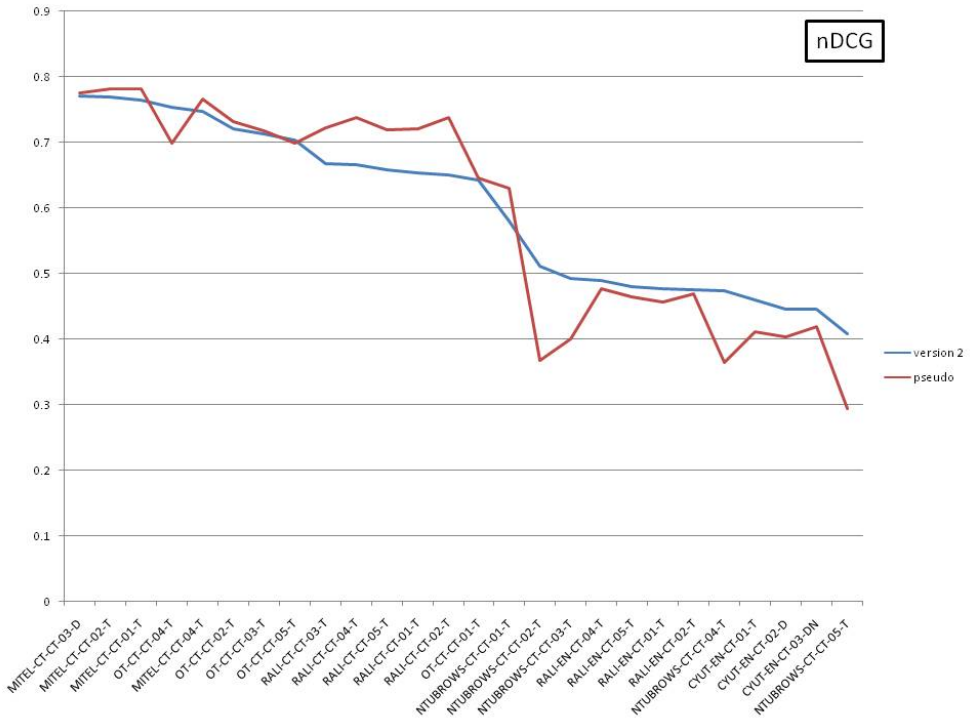

Figure 15. Mean nDCG values for CT runs based on size-10 pseudo-qrels and qrels versions 2: Runs sorted by Mean nDCG based on qrels version 2.

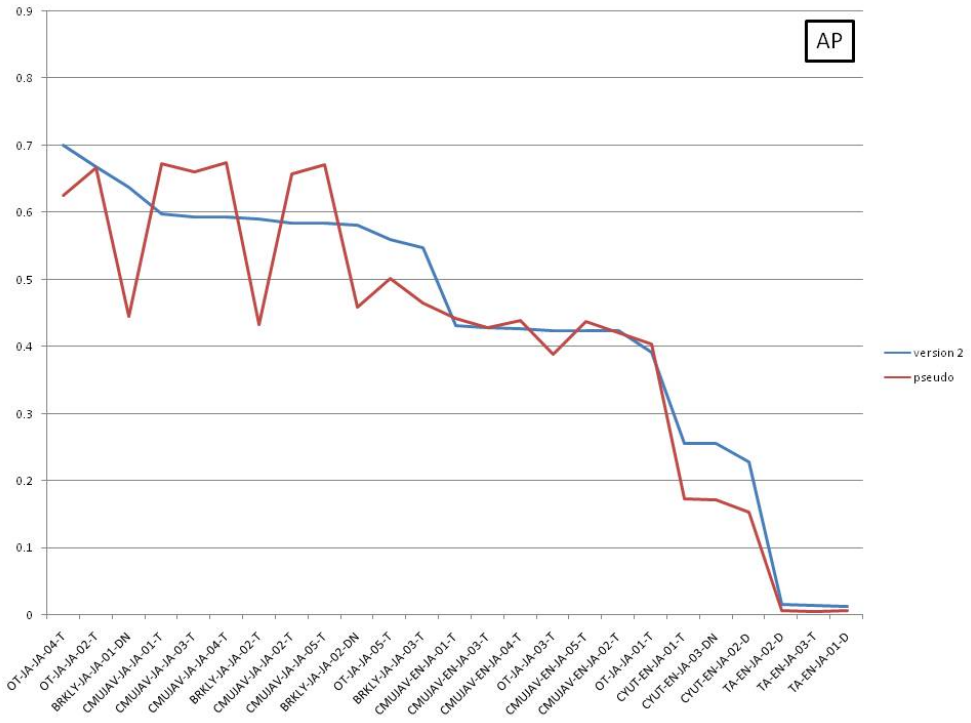

Figure 16. Mean AP values for JA runs based on size-10 pseudo-qrels and qrels versions 2: Runs sorted by Mean AP based on qrels version 2.

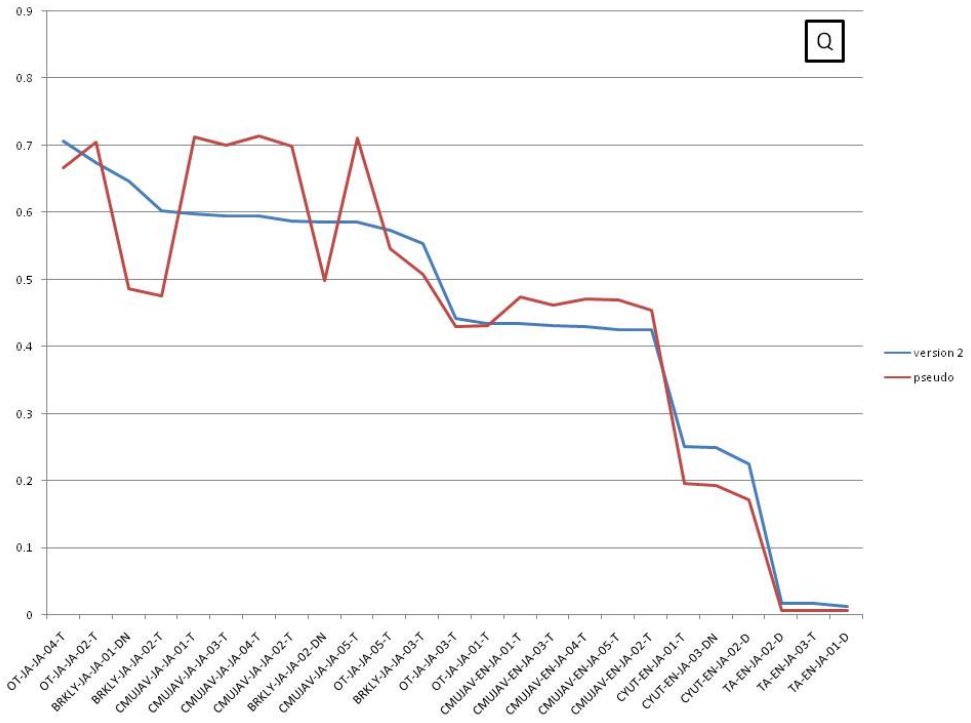

Figure 17. Mean Q values for JA runs based on size-10 pseudo-qrels and qrels versions 2: Runs sorted by Mean Q based on qrels version 2.

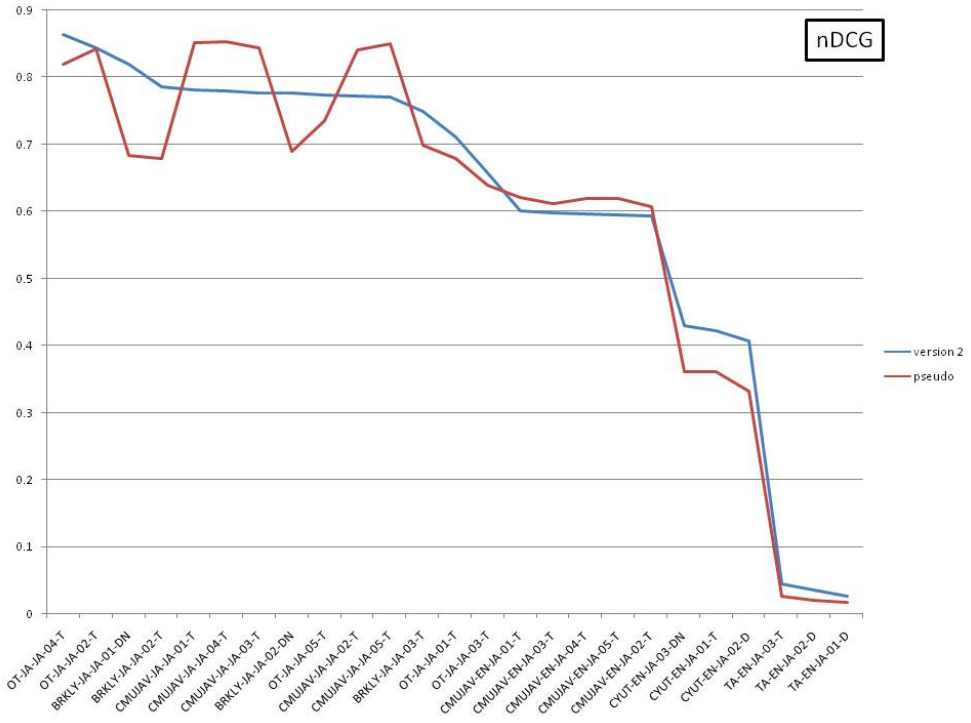

Figure 18. Mean nDCG values for JA runs based on size-10 pseudo-qrels and qrels versions 2: Runs sorted by Mean nDCG based on qrels version 2.

7 Conclusions

This document has reported on revised NTCIR-7 ACLIA IR4QA results based on "qrels version 2" which covers the depth-100 pool for every topic. While the version 1 and version 2 results are generally in agreement, some differences in system rankings and significance test results suggest that the additional effort was worthwhile.

While the size-10 pseudo-qrels files, which we created prior to the relevance assessments, do not accurately the system rankings based on qrels version 2, our new experiments using size-100 pseudo-qrels show that these new pseudo-qrels files are relatively accurate at predicting the "true" rankings. We refer the reader to [3] for more details.

References

[1] ir4qa_eval: http://research.nii.ac.jp/ntcir/tools/ir4qa_eval-en

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

[3] Sakai, T., Kando, N., Lin, C.-J., Song, R., Shima, H. and Mitamura, T.: Revisiting NTCIR ACLIA IR4QA with Additional Relevance Assessments, *IPSJ SIG Technical Reports*, 2009-FI-95 / 2009-DBS-148, to appear, 2009.