Analysis of the INEX 2009 Ad Hoc Track Results

Jaap Kamps[1], Shlomo Geva[2], and Andrew Trotman[3]

[1] University of Amsterdam, Amsterdam, The Netherlands
kamps@uva.nl
[2] Queensland University of Technology, Brisbane, Australia
s.geva@qut.edu.au
[3] University of Otago, Dunedin, New Zealand
andrew@cs.otago.ac.nz

Abstract. This paper analyzes the results of the INEX 2009 Ad Hoc Track, focusing on a variety of topics. First, we examine in detail the relevance judgments. Second, we study the resulting system rankings, for each of the four ad hoc tasks, and determine whether differences between the best scoring participants are statistically significant. Third, we restrict our attention to particular run types: element and passage runs, keyword and phrase query runs, and systems using a reference run with a solid article ranking. Fourth, we examine the relative effectiveness of content only (CO, or Keyword) search as well as content and structure (CAS, or structured) search. Fifth, we look at the ability of focused retrieval techniques to rank articles. Sixth, we study the length of retrieved results, and look at the impact of restricting result length.

1 Introduction

This paper provides analysis of the results of the INEX 2009 Ad Hoc Track, in addition to the overview of INEX 2009 Ad Hoc track's tasks and results in [1].

We focus on a variety of topics. First, we try to understand what constitutes a "highlighted" passage, and how the new and four times larger Wikipedia collection may affect the resulting test collection. For this purpose, we examine the relevance judgments in great detail. Second, we investigate the ability of the evaluation to distinguish between different retrieval approaches. We do this by studying the resulting system rankings, for each of the four ad hoc tasks, and determine whether differences between the best scoring participants are statistically significant. Third, we dig deeper in the effectiveness of particular focused retrieval approaches, by restricting our attention to particular run types: element and passage runs, keyword and phrase query runs, and systems using a reference run with a solid article ranking. Fourth, we try to grasp the impact of structural hints using either the original XML document structure, or automatically assigned YAGO tags. We examine the relative effectiveness of content only (CO, or Keyword) search as well as content and structure (CAS, or structured) search. Fifth, we relate the focused retrieval approaches to article retrieval, by looking at the ability of focused retrieval techniques to rank articles. Sixth, we investigate the length of retrieved text per article, and the performance of focused retrieval systems under resource-limited conditions. In particular, we "cut off" the results after having retrieved the first 500 retrieved characters per article.

The rest of the paper is organized as follows. Section 2 analyzes the assessments of the INEX 2009 Ad Hoc Track. In Section 3, we report the results for the Thorough Task (Section 3.1); the Focused Task (Section 3.2); the Relevant in Context Task (Section 3.3); and the Best in Context Task (Section 3.4). Section 4 details particular types of runs (such as element versus passage, using phrases or using the reference run), and on particular subsets of the topics (such as topics with a non-trivial CAS query). Section 6 looks at the article retrieval aspects of the submissions, treating any article with highlighted text as relevant. We study the impact of result length in Section 7. Finally, in Section 8, we discuss our findings and draw some conclusions.

2 Analysis of the Relevance Judgments

In this section, we analyze the relevance assessments used in the Ad Hoc Track. The 2009 collection contains 2,666,190 Wikipedia articles (October 8, 2008 dump of the Wikipedia), which is four times larger than the earlier Wikipedia collection. What is the effect of this change in corpus size?

2.1 Topics

Topics were assessed by participants following precise instructions. The assessors used the GPXrai assessment system that assists assessors in highlighting relevant text. Topic assessors were asked to mark all, and only, relevant text in a pool of documents. After assessing an article with relevance, a separate best entry point decision was made by the assessor. The Thorough, Focused and Relevant in Context Tasks were evaluated against the text highlighted by the assessors, whereas the Best in Context Task was evaluated against the best-entry-points.

The relevance judgments were frozen on November 10, 2009. At this time 68 topics had been fully assessed. Moreover, some topics were judged by two separate assessors, each without the knowledge of the other. All results in this paper refer to the 68 topics with the judgments of the first assigned assessor, which is typically the topic author.

- The 68 assessed topics were numbered 2009*n* with *n*: 001–006, 010–015, 020, 022, 023, 026, 028, 029, 033, 035, 036, 039–043, 046, 047, 051, 053–055, 061–071, 073, 074, 076–079, 082, 085, 087–089, 091–093, 095, 096, 104, 105, 108–113, and 115

2.2 Highlighted Text

Table 1 presents statistics of the number of judged and relevant articles, and passages. In total 50,725 articles were judged. Relevant passages were found

Table 1. Statistics over judged and relevant articles per topic.

	total		# per topic				
	topics	number	min	max	median	mean	st.dev
judged articles	68	50,725	380	766	754	746.0	49.0
articles with relevance	68	4,858	5	351	52	71.4	72.5
highlighted passages	68	7,957	5	594	75.5	117.0	121.5
highlighted characters	68	18,838,137	4,453	2,776,635	97,550.5	277,031.4	442,113.9

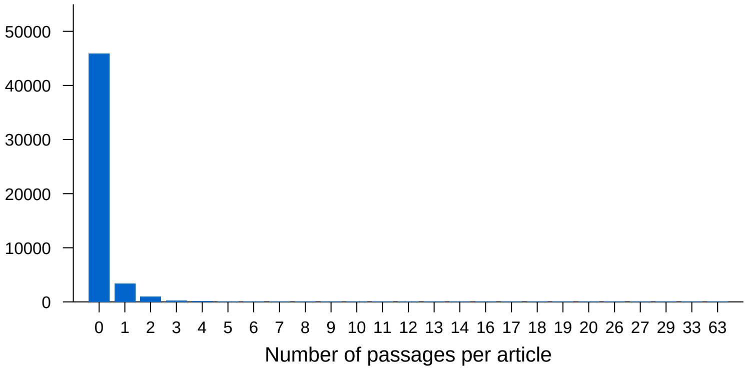

Fig. 1. Distribution of passages over articles.

in 4,858 articles. The mean number of relevant articles per topic is 71, but the distribution is skewed with a median of 52. There were 7,957 highlighted passages. The mean was 117 passages and the median was 76 passages per topic.[1]

Figure 1 presents the number of articles with the given number of passages. The vast majority of relevant articles (3,339 out of 4,858) had only a single highlighted passage, and the number of passages quickly tapers off.

2.3 Best Entry Point

Assessors where requested to provide a separate best entry point (BEP) judgment, for every article where they highlighted relevant text. Table 2 presents statistics on the best entry point offset, on the first highlighted or relevant character, and on the fraction of highlighted text in relevant articles. We first look at the BEPs. The mean BEP is well within the article with 2,493 but the distribution is very skewed with a median BEP offset of only 311. Figure 2 shows the distribution of the character offsets of the 4,858 best entry points. It is clear that the overwhelming majority of BEPs is at the beginning of the article.

The statistics of the first highlighted or relevant character (FRC) in Table 2 give very similar numbers as the BEP offsets: the mean offset of the first relevant

[1] Note that for the Focused Task the main effectiveness measures is precision at 1% recall. Given that the average topic has 117 relevant passages in 52 articles, the 1% recall roughly corresponds to a relevant passage retrieved—for many systems this will be accomplished by the first or first few results.

Table 2. Statistics over relevant articles.

	total		# per relevant article				
	topics	number	min	max	median	mean	st.dev
best entry point offset	68	4,858	2	86,545	311.5	2,493.2	6,481.8
first relevant character offset	68	4,858	2	86,545	295	2,463.0	6,375.6
length relevant documents	68	4,858	204	159,892	5,774.5	11,691.5	15,745.1
relevant characters	68	4,858	8	110,191	1,137	3,877.8	7,818.5
fraction highlighted text	68	4,858	0.00022	1.000	0.330	0.442	0.381

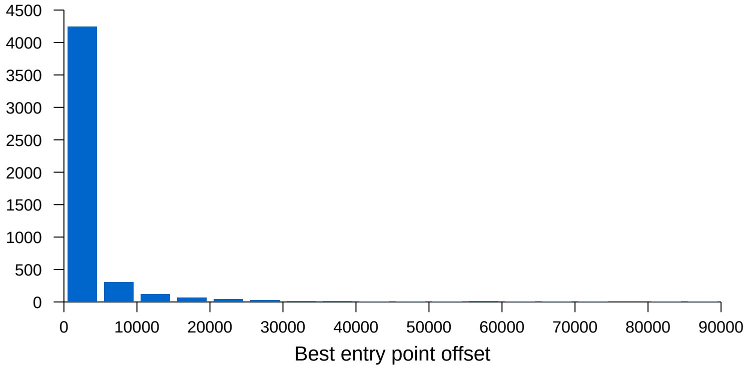

Fig. 2. Distribution of best entry point offsets.

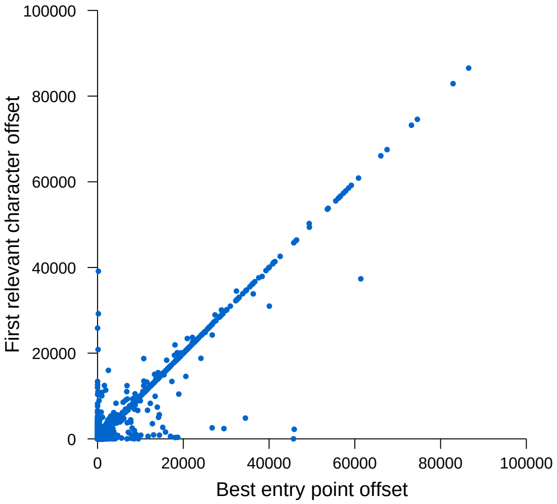

Fig. 3. Scatter plot of best entry point offsets versus the first relevant character.

character is 2,463 but the median offset is only 295. This suggests a relation between the BEP offset and the FRC offset. Figure 3 shows a scatter plot the BEP and FRC offsets. Two observations present themselves. First, there is a clear

Table 3. Top 10 Participants in the Ad Hoc Track Thorough Task: Statistical significance (t-test, one-tailed, 95%).

Participant	MAiP	1	2	3	4	5	6	7	8	9	10
p48-LIG-2009-thorough-3T	0.2855		-	-	⋆	-	⋆	-	⋆	⋆	⋆
p6-UAmsIN09article	0.2818			-	⋆	-	⋆	-	⋆	⋆	⋆
p5-BM25thorough	0.2585				⋆	-	⋆	-	⋆	⋆	⋆
p92-Lyon3LIAmanlmnt	0.2496					-	-	-	⋆	⋆	-
p60-UJM_15494	0.2435						-	-	⋆	⋆	⋆
p346-utCASartT09	0.2350							-	⋆	⋆	⋆
p10-MPII-CASThBM	0.2133								⋆	⋆	⋆
p167-09RefT	0.1390									-	-
p68-I09LIP6OWATh	0.0630										-
p25-ruc-base-coT	0.0577										

diagonal where the BEP is positioned exactly at the first highlighted character in the article. Second, there is also a vertical line at BEP offset zero, indicating a tendency to put the BEP at the start of the article even when the relevant text appears later on.

Table 2 also shows statistics on the length of relevant articles. Many articles are relatively short with a median length of 5,775 characters, the mean length is 11,691 characters. This is considerably longer than the INEX 2008 collection, where the relevant articles had a median length of 3,030 and a mean length of 6,793. The length of highlighted text in characters is on average 3,876 (mean 1,137), in comparison to an average length of 2,338 (mean 838) in 2008. Table 2 also shows that the amount of relevant text varies from almost nothing to almost everything. The mean fraction is 0.44, and the median is 0.33, indicating that typically over one-third of the article is relevant. This is considerably less than the INEX 2008 collection, where over half of the text of articles was considered relevant. The observation that the majority of relevant articles contain such a large fraction of relevant text, plausibly explains that BEPs being frequently positioned on or near the start of the article.

3 Analysis of the Ad Hoc Tasks

In this section, we discuss, for the four ad hoc tasks, the participants and their results by looking at the significance of differences between participants.

3.1 Thorough Task

We tested whether higher ranked systems were significantly better than lower ranked systems, using a t-test (one-tailed) at 95%. Table 3 shows for the best runs of the 10 best scoring groups whether a run is significantly better (indicated by "⋆") than lower ranked runs.

For the Thorough Task, we see that the performance (measured by MAiP) of the top scoring run is significantly better than the runs at rank 4, 6, 8, 9, and

Table 4. Top 10 Participants in the Ad Hoc Track Focused Task: Statistical significance (t-test, one-tailed, 95%).

Participant	iP[0.01]	1	2	3	4	5	6	7	8	9	10
p78-UWatFERBM25F	0.6333		-	-	-	-	-	-	-	-	⋆
p68-I09LIP6Okapi	0.6141			-	-	-	-	-	⋆	-	⋆
p10-MPII-COFoBM	0.6134				-	-	-	-	-	-	⋆
p60-UJM_15525	0.6060					-	-	-	-	-	⋆
p6-UamsFSsec2docbi100	0.5997						-	-	-	-	⋆
p5-BM25BOTrangeFOC	0.5992							-	-	-	⋆
p16-Spirix09R001	0.5903								-	-	⋆
p48-LIG-2009-focused-1F	0.5853									-	⋆
p22-emse2009-150	0.5844										⋆
p25-ruc-term-coF	0.4973										

10. The same holds for the second and third best run. The fourth best run is significantly better than the runs at rank 8 and 9. The fifth, sixth, and seventh ranked runs are all significantly better than the runs at rank 8, 9, and 10. Of the 45 possible pairs of runs, there are 26 (or 58%) significant differences.

3.2 Focused Task

Table 4 shows for the best runs of the 10 best scoring groups whether a run is significantly better (indicated by "⋆") than lower ranked runs. For the Focused Task, we see that the early precision (at 1% recall) is a rather unstable measure. All runs are significantly better than the run at rank 10, the second best run also is significantly better than the run at rank 8. Of the 45 possible pairs of runs, there are only 10 (or 22%) significant differences. Hence we should be careful when drawing conclusions based on the Focused Task results.

The overall MAiP measure is more stable, see the analysis of the Thorough runs before.

3.3 Relevant in Context Task

Table 5 shows for the best runs of the 10 best scoring groups whether a run is significantly better (indicated by "⋆") than lower ranked runs. For the Relevant in Context Task, we see that the top run is significantly better than ranks 2 and 4 through 10. The second best run is significantly better than ranks 5 through 10. The third, fourth, and fifth ranked systems are significantly better than ranks 6 through 10. The sixth to ninth systems are significantly better than rank 10. Of the 45 possible pairs of runs, there are 33 (or 73%) significant differences, making MAgP a very discriminative measure.

3.4 Best in Context Task

Table 6 shows for the best runs of the 10 best scoring groups whether a run is significantly better (indicated by "⋆") than lower ranked runs. For the Best in

Table 5. Top 10 Participants in the Ad Hoc Track Relevant in Context Task: Statistical significance (t-test, one-tailed, 95%).

Participant	MAgP	1	2	3	4	5	6	7	8	9	10
p5-BM25RangeRIC	0.1885		⋆	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆
p4-Reference	0.1847			-	-	⋆	⋆	⋆	⋆	⋆	⋆
p6-UamsRSCMartCMdocbi100	0.1773				-	-	⋆	⋆	⋆	⋆	⋆
p48-LIG-2009-RIC-1R	0.1760					-	⋆	⋆	⋆	⋆	⋆
p36-utampere_given30_nolinks	0.1720						⋆	⋆	⋆	⋆	⋆
p346-utCASrefR09	0.1188							-	-	-	⋆
p60-UJM_15502	0.1075								-	-	⋆
p167-09RefR	0.1045									-	⋆
p25-ruc-base-casF	0.1028										⋆
p72-umd_ric_1	0.0424										

Table 6. Top 10 Participants in the Ad Hoc Track Best in Context Task: Statistical significance (t-test, one-tailed, 95%).

Participant	MAgP	1	2	3	4	5	6	7	8	9	10
p5-BM25bepBIC	0.1711		-	-	⋆	⋆	-	⋆	⋆	⋆	⋆
p62-RMIT09titleO	0.1710			-	⋆	-	-	⋆	⋆	⋆	⋆
p10-MPII-COBIBM	0.1662				-	-	-	⋆	⋆	⋆	⋆
p48-LIG-2009-BIC-3B	0.1571					-	-	⋆	⋆	⋆	⋆
p6-UamsBAfbCMdocbi100	0.1544						-	⋆	⋆	⋆	⋆
p92-Lyon3LIAmanBEP	0.1483							-	⋆	⋆	⋆
p36-utampere_given30_nolinks	0.1207								-	-	⋆
p346-utCASrefB09	0.1056									-	-
p25-ruc-term-coB	0.1013										-
p167-09LrnRefB	0.0953										

Context Task, we see that the top run is significantly better than ranks 4 and 5, and 7 through 10. The second best run is significantly better than than ranks 4 and 7 to 10. The third, fourth, and fifth ranked runs are significantly better than than ranks 7 to 10. The seventh ranked system is better than the systems ranked 8 to 10, and the eighth ranked system better than ranks 9 to 10. Of the 45 possible pairs of runs, there are 27 (or 60%) significant differences.

4 Analysis of Run Types

In this section, we will discuss the relative effectiveness of element and passage retrieval approaches, of phase and keyword queries, and of the reference run providing solid article ranking.

4.1 Elements versus passages

We received 13 submissions using ranges of elements of FOL-passage results, from in total 4 participating groups. We will look at the relative effectiveness of element and passage runs.

Table 7. Ad Hoc Track: Runs with ranges of elements or FOL passages.

(a) Focused Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p78-UWatFERBM25F	0.6797	0.6333	0.5006	0.4095	0.1854
p5-BM25BOTrangeFOC	0.6049	0.5992	0.5619	0.5057	0.2912

(b) Relevant in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p5-BM25RangeRIC	0.3345	0.2980	0.2356	0.1786	0.1885
p36-utampere_auth_40_top30	0.2717	0.2509	0.2006	0.1583	0.1185

(c) Best in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p62-RMIT09titleO	0.3112	0.2757	0.2156	0.1673	0.1710

For three tasks there were high ranking runs using FOL passages or ranges of elements in the top 10 (discussed in Section 3). Table 7 shows the best runs using ranges of elements or FOL passages for three ad hoc tasks, there were no such submissions for the Thorough Task. As it turns out, the best focused run retrieving FOL passages was the top ranked run in Table 4; the best relevant in context retrieving ranges of elements was the top scoring run in Table 5; and the best best in context run retrieving FOL passages was the second best run in Table 6. Given the low number of submissions using passages or ranges of elements, this is an impressive result. However, looking at the runs in more detail, their character is often unlike what one would expect from a "passage" retrieval run. For Focused, *p5-BM25BOTrangeFOC* is an article retrieving run using ranges of elements, based on the CAS query. For Relevant in Context, *p5-BM25RangeRIC* is an article retrieving run using ranges of elements. For Best in Context, *p62-RMIT09titleO* is an article run using FOL passages. Hence, this is not sufficient evidence to warrant any conclusion on the effectiveness of passage level results. We hope and expect that the test collection and the passage runs will be used for further research into the relative effectiveness of element and passage retrieval approaches.

4.2 Phrase queries

We received 10 submissions based on the phrase query. Table 8 shows the best runs using the phrase query for three of the ad hoc tasks, there were no valid submissions using the phrase title for Relevant in Context. The best phrase submission for the Thorough Task did rank 5th in the overall results. The best phrase submission for the Focused Task did rank 9th in the overall results. The best phrase submission for the Best in Context Task did rank 6th in the overall results.

Although few runs were submitted, the phrase title seems competitive, but not superior to the use of the CO query. The only participant submitting both types of runs, the *Max-Planck-Institute für Informatik* for the Focused Task, had marginally better performance for the CO query run over all 68 topics, and

Table 8. Ad Hoc Track: Runs using the phrase query.

(a) Thorough Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p92-Lyon3LIAmanlmnt*	0.5196	0.4956	0.4761	0.4226	0.2496

(b) Focused Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p22-emse2009-150*	0.6671	0.5844	0.4396	0.3699	0.1470
p10-MPII-COArBPP	0.5563	0.5477	0.5283	0.4681	0.2566
p92-Lyon3LIAmanQE*	0.4955	0.4861	0.4668	0.4271	0.2522

(c) Best in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p92-Lyon3LIAmanBEP*	0.2887	0.2366	0.1815	0.1482	0.1483

marginally better performance for the combined CO and Phrase title run over the 60 topics having a proper phrase in the Phrase title field. The differences between the query types are very small. A possible explanation for this is that all CO query have been expanded to contain the same terms as the more verbose phrase query. Hence the only difference is the explicit phrase markup, which requires special handling by the search engines. The available test collection with explicit phrases marked up in 60 topics is a valuable result of INEX 2009, and it can be studied in-depth in future experiments.

4.3 Reference run

There were 19 submissions using the reference run. Table 9 shows the best runs using the reference runs for the four ad hoc tasks. For the Thorough Task, the best submission based on the reference run ranked first. For the Focused Task, the best submission based on the reference run would have ranked tenth. For the Relevant in Context Task, the best submission based on the reference run—in fact, the actual reference run itself—ranked second. For the Best in Context Task, the best submission based on the reference run ranked fourth. The results show that the reference run indeed provides competitive article ranking that forms a good basis for retrieval.

There are also considerable differences in performance of the runs based on the same reference run. This suggests that the runs do not retrieve the exact same set of articles. As explained later, in Section 6, we can look at the article rankings induced by the runs. Table 10 shows the best run of the top 10 participating groups, using the reference run. With the exception of *p36-utampere_given30_nolinks* the article rankings of the runs vary considerably.

5 Analysis of Structured Queries

In this section, we will discuss the relative effectiveness of systems using the keyword and structured queries.

Table 9. Ad Hoc Track: Runs using the reference run.

(a) Thorough Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p48-LIG-2009-thorough-3T	0.5967	0.5841	0.5444	0.5019	0.2855
p60-UJM_15494	0.5986	0.5789	0.5293	0.4813	0.2435
p346-utCASrefF09	0.4834	0.4525	0.4150	0.3550	0.1982
p167-09RefT	0.3205	0.3199	0.2779	0.2437	0.1390

(b) Focused Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p48-LIG-2009-focused-3F	0.5946	0.5822	0.5344	0.5018	0.2732
p60-UJM_15518	0.5559	0.5136	0.4003	0.3104	0.1019
p346-utCASrefF09	0.4801	0.4508	0.4139	0.3547	0.1981
p167-09LrnRefF	0.3162	0.3072	0.2512	0.2223	0.1292

(c) Relevant in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p4-Reference	0.3311	0.2936	0.2298	0.1716	0.1847
p48-LIG-2009-RIC-3R	0.3119	0.2790	0.2193	0.1629	0.1757
p36-utampere_given30_nolinks	0.3128	0.2802	0.2101	0.1592	0.1720
p346-utCASrefR09	0.2216	0.1904	0.1457	0.1095	0.1188
p167-09RefR	0.1595	0.1454	0.1358	0.1205	0.1045
p60-UJM_15503	0.1825	0.1548	0.1196	0.0953	0.1020

(d) Best in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p48-LIG-2009-BIC-3B	0.2778	0.2564	0.1969	0.1469	0.1571
p36-utampere_given30_nolinks	0.2141	0.1798	0.1462	0.1234	0.1207
p346-utCASrefB09	0.1993	0.1737	0.1248	0.0941	0.1056
p167-09LrnRefB	0.1369	0.1250	0.1181	0.1049	0.0953
p60-UJM_15508	0.1274	0.1123	0.0878	0.0735	0.0795

Table 10. Top 10 Participants in the Ad Hoc Track: Article retrieval based on the reference run.

Participant	P5	P10	1/rank	map	bpref
p4-Reference	0.6147	0.5294	0.8240	0.3477	0.3333
p36-utampere_given30_nolinks	0.6147	0.5294	0.8240	0.3477	0.3333
p48-LIG-2009-BIC-3B	0.6147	0.5294	0.8240	0.3463	0.3336
p60-UJM_15508	0.5324	0.4544	0.7020	0.2910	0.2925
p346-utCASrefB09	0.5441	0.4750	0.7494	0.2833	0.2768
p167-09RefT	0.3765	0.3603	0.5761	0.2443	0.2540

5.1 CO versus CAS

We now look at the relative effectiveness of the keyword (CO) and structured (CAS) queries. As we saw above, in Section 3, one of the best runs per group for the Relevant in Context Task, and two of the top 10 runs for the Best in Context Task used the CAS query.

All topics have a CAS query since artificial CAS queries of the form

Table 11. CAS query target elements over all 115 topics (YAGO tags slanted).

Target Element	Frequency
*	41
article	32
sec	9
group	5
p	4
music_genre	2
vehicles	1
theory	1
song	1
revolution	1
(p|sec|*person)*	1
(p|sec)	1
protest	1
(person|chemist|alchemist|scientist|physicist)	1
personality	1
museum	1
link	1
image	1
home	1
food	1
figure	1
facility	1
driver	1
dog	1
director	1
(classical_music|opera|orchestra|performer|singer)	1
bicycle	1
(article|sec|p)	1

```
//*[about(., keyword title)]
```

were added to topics without CAS title. Table 11 show the distribution of target elements, with YAGO tags in emphatic. In total 81 topics had a non-trivial CAS query.[2] These CAS topics are numbered 2009*n* with *n*: 001–009, 011–013, 015–017, 020–025, 028–032, 036, 037, 039–045, 048–053, 057, 058, 060, 061, 064–072, 074, 080, 085–096, 098, 099, 102, 105, 106, and 108–115. As it turned out, 50 of these CAS topics were assessed. The results presented here are restricted to only these 50 CAS topics.

Table 12 lists the top 10 participants measured using just the 50 CAS topics and for the Thorough Task (a and b) and the Focused Task (c and d). For the Thorough Task the best CAS run, *p5-BM25BOTthorough*, would have ranked sixth amongst the CO runs on MAiP. The two participants submitting both CO and CAS runs had better MAiP scores for the CO runs. However,

[2] Note that some of the wild-card topics (using the "*" target) in Table 11 had non-trivial about-predicates and hence have not been regarded as trivial CAS queries.

Table 12. Ad Hoc Track CAS Topics: CO runs versus CAS runs.

(a) Thorough Task: CO runs

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p48-LIG-2009-thorough-1T	0.5781	0.5706	0.5315	0.4834	0.2729
p6-UAmsIN09article	0.5900	0.5821	0.5149	0.4613	0.2629
p92-Lyon3LIAmanlmnt⋆	0.5365	0.5039	0.4794	0.4330	0.2450
p5-BM25thorough	0.6273	0.6023	0.5191	0.4620	0.2389
p60-UJM_15494	0.6034	0.5766	0.5131	0.4612	0.2280
p10-MPII-COThBM	0.6436	0.5916	0.5135	0.3783	0.1909
p167-09RefT	0.3245	0.3237	0.2682	0.2392	0.1291
p68-I09LIP6OWATh	0.4146	0.3651	0.2512	0.1963	0.0608
p25-ruc-base-coT	0.5328	0.4333	0.2538	0.1653	0.0505
p72-umd_thorough_3	0.4073	0.2893	0.1697	0.0999	0.0494

(b) Thorough Task: CAS runs

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p5-BM25BOTthorough	0.6460	0.6169	0.5359	0.4472	0.2279
p346-utCASartT09	0.5541	0.5381	0.4819	0.4136	0.2227
p10-MPII-CASThBM	0.5747	0.5308	0.4406	0.3627	0.1651

(c) Focused Task: CO runs

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p78-UWatFERBM25F	0.6742	0.6222	0.4905	0.3758	0.1737
p60-UJM_15525	0.6373	0.6127	0.5696	0.4585	0.2811
p10-MPII-COArBM	0.6201	0.6060	0.5387	0.4648	0.2684
p68-I09LIP6Okapi	0.6130	0.6005	0.5660	0.5064	0.2798
p5-ANTbigramsRangeFOC	0.6089	0.5936	0.5331	0.4531	0.2597
p48-LIG-2009-focused-3F	0.5971	0.5802	0.5205	0.4775	0.2583
p22-emse2009-150⋆	0.6453	0.5598	0.4211	0.3471	0.1371
p92-Lyon3LIAmanQE⋆	0.5185	0.5058	0.4815	0.4339	0.2472
p25-ruc-term-coF	0.6277	0.4955	0.2900	0.2065	0.0668
p167-09LrnRefF	0.3357	0.3234	0.2536	0.2211	0.1216

(d) Focused Task: CAS runs

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p6-UamsFSsec2docbi100	0.6151	0.5974	0.4851	0.4230	0.1718
p16-Spirix09R001	0.6201	0.5958	0.5386	0.4920	0.2794
p5-BM25BOTrangeFOC	0.6031	0.5954	0.5470	0.4789	0.2713
p10-MPII-CASFoBM	0.5643	0.5161	0.4454	0.3634	0.1644
p25-ruc-base-casF	0.5114	0.4775	0.4077	0.3214	0.1666
p346-utCASrefF09	0.4353	0.3955	0.3477	0.2781	0.1471
p55-doshisha09f	0.1273	0.0651	0.0307	0.0227	0.0060

the best CAS run has higher scores on early precision, iP[0.00] through iP[0.05] than any of the CO submissions. For the Focused Task the best CAS run, *p6-UamsFSsec2docbi100*, would have ranked fifth amongst the CO runs. Two participants submitting both CO and CAS runs had better iP[0.01] scores for the CO runs, one participant had a better CAS run. For Relevant in Context Task (not shown), the best CAS run, *p5-BM25BOTrangeRIC*, would have ranked third

Table 13. Top 10 Participants in the Ad Hoc Track: Article retrieval.

Participant	P5	P10	1/rank	map	bpref
p6-UamsTAbi100	0.6500	0.5397	0.8555	0.3578	0.3481
p48-LIG-2009-BIC-1B	0.6059	0.5338	0.8206	0.3573	0.3510
p62-RMIT09title	0.6029	0.5279	0.8237	0.3540	0.3488
p5-BM25ArticleRIC	0.6147	0.5294	0.8240	0.3477	0.3333
p4-Reference	0.6147	0.5294	0.8240	0.3477	0.3333
p36-utampere_given30_nolinks	0.6147	0.5294	0.8240	0.3477	0.3333
p68-I09LIP6OWA	0.6118	0.5147	0.8602	0.3420	0.3258
p10-MPII-COArBP	0.6353	0.5471	0.8272	0.3371	0.3458
p92-Lyon3LIAmanQE*	0.6265	0.5265	0.7413	0.3335	0.3416
p78-UWatFERBase	0.5765	0.5088	0.8093	0.3267	0.3205

among the CO runs. One participants submitting both CO and CAS runs had better MAgP scores for a CO run, another participant had a better CAS run. For the Best in Context Task (not shown), the best CAS run, *p5-BM25BOTbepBIC*, would rank seventh among the CO runs. All three participants submitting both CO and CAS runs had better MAgP scores for their CO runs. Overall, we see that teams submitting runs with both types of queries have higher scoring CO runs, with participant 5 as a notable exception for Focused.

6 Analysis of Article Retrieval

In this section, we will look in detail at the effectiveness of Ad Hoc Track submissions as article retrieval systems.

6.1 Article retrieval: Relevance Judgments

We will first look at the topics judged during INEX 2009, but now using the judgments to derive standard document-level relevance by regarding an article as relevant if some part of it is highlighted by the assessor. We derive an article retrieval run from every submission using a first-come, first served mapping. That is, we simply keep every first occurrence of an article (retrieved indirectly through some element contained in it) and ignore further results from the same article.

We use `trec_eval` to evaluate the mapped runs and qrels, and use mean average precision (map) as the main measure. Since all runs are now article retrieval runs, the differences between the tasks disappear. Moreover, runs violating the task requirements are now also considered, and we work with all 172 runs submitted to the Ad Hoc Track.

Table 13 shows the best run of the top 10 participating groups. The first column gives the participant, see the companion article [1, Table 3] for the full name of group. The second and third column give the precision at ranks 5 and 10, respectively. The fourth column gives the mean reciprocal rank. The fifth column gives mean average precision. The sixth column gives binary preference measures (using the top R judged non-relevant documents). No less than seven of the top 10 runs retrieve exclusively full articles: only rank two (*p48-LIG-2009-BIC-1B*), rank six (*p36-utampere_given30_nolinks*) and rank ten (*p78-UWatFERBase*) retrieve elements proper. The relative effectiveness of these article retrieval runs in terms of their article ranking is no surprise. Furthermore, we see submissions from all four ad hoc tasks. A run from the Thorough task at rank 1; runs from the Best in Context task at ranks 2 and 3; runs from the Relevant in Context task at ranks 4, 5 and 6; and runs from the Focused task at ranks 7, 8, 9 and 10.

If we break-down all runs over the original tasks, shown in Table 14, we can compare the ranking to Section 3 above. We see some runs that are familiar from the earlier tables: five Thorough runs correspond to Table 3, four Focused runs correspond to Table 4, six Relevant in Context runs correspond to Table 5, and five Best in Context runs correspond to Table 6. More formally, we looked at how the two system rankings correlate using kendall's tau.

- Over all 30 Thorough Task submissions the system rank correlation is 0.646 between MAiP and map.
- Over all 57 Focused task submissions the system rank correlation is 0.420 between iP[0.01] and map, and 0.638 between MAiP and map.
- Over all 33 Relevant in Context submissions the system rank correlation between MAgP and map is 0.598.
- Over all 37 Best in Context submissions the system rank correlation between MAgP and map is 0.517.

Overall, we see a reasonable correspondence between the rankings for the ad hoc tasks in Section 3 and the rankings for the derived article retrieval measures. The correlation between article retrieval and the "in context" tasks was much higher (0.79) for the INEX 2008 collection. This is a likely effect of the increasing length of (relevant) Wikipedia articles in the INEX 2009 collection.

7 Analysis of Result Length

In this section, we will look in detail at the impact of result length on the effectiveness of Ad Hoc Track submissions.

7.1 Impact of Result Length

Focused retrieval and XML retrieval require all, but only, relevant text to be retrieval. This could be taken to suggest that a relatively short result length is optimal. In sharp contrast, researchers found that XML-IR require careful length normalization, effectively boosting the retrieval of longer elements [2, 3].

Let us look in detail at the length of results retrieved by top scoring runs. Table 15 shows for the best Thorough runs of the 10 best scoring groups statistics on number of articles, and characters retrieved (restricted to the 68 judged topics). There is an enormous spread in the average number of characters per

Table 14. Top 10 Participants in the Ad Hoc Track: Article retrieval per task.

(a) Thorough Task

Participant	P5	P10	1/rank	map	bpref
p6-UamsTAbi100	0.6500	0.5397	0.8555	0.3578	0.3481
p48-LIG-2009-thorough-1T	0.6118	0.5191	0.8042	0.3493	0.3392
p92-Lyon3LIAmanlmnt*	0.6382	0.5279	0.7706	0.3305	0.3374
p5-BM25thorough	0.6147	0.5294	0.8240	0.3188	0.3142
p10-MPII-COThBM	0.5853	0.5206	0.8084	0.3087	0.3138
p346-utCASartT09	0.5176	0.4588	0.7138	0.2913	0.2986
p60-UJM_15486	0.5647	0.4765	0.7149	0.2797	0.2884
p68-I09LIP6OWATh	0.4735	0.4353	0.7100	0.2665	0.2745
p72-umd_thorough_3	0.5382	0.4515	0.7406	0.2486	0.2674
p167-09RefT	0.3765	0.3603	0.5761	0.2443	0.2540

(b) Focused Task

Participant	P5	P10	1/rank	map	bpref
p48-LIG-2009-focused-1F	0.6059	0.5338	0.8206	0.3569	0.3506
p5-BM25ArticleFOC	0.6147	0.5294	0.8240	0.3477	0.3333
p68-I09LIP6OWA	0.6118	0.5147	0.8602	0.3420	0.3258
p10-MPII-COArBP	0.6353	0.5471	0.8272	0.3371	0.3458
p92-Lyon3LIAmanQE*	0.6265	0.5265	0.7413	0.3335	0.3416
p78-UWatFERBase	0.5765	0.5088	0.8093	0.3267	0.3205
p60-UJM_15525	0.5824	0.4926	0.8326	0.3256	0.3169
p16-Spirix09R002	0.5206	0.4588	0.7250	0.3133	0.3149
p6-UamsFSsec2docbi100	0.5941	0.4779	0.8958	0.2985	0.2994
p346-utCASartF09	0.5176	0.4588	0.7138	0.2913	0.2986

(c) Relevant in Context Task

Participant	P5	P10	1/rank	map	bpref
p48-LIG-2009-RIC-1R	0.6059	0.5338	0.8206	0.3569	0.3506
p6-UamsRSCMartCMdocbi100	0.6324	0.5309	0.9145	0.3523	0.3374
p5-BM25ArticleRIC	0.6147	0.5294	0.8240	0.3477	0.3333
p4-Reference	0.6147	0.5294	0.8240	0.3477	0.3333
p36-utampere_given30_nolinks	0.6147	0.5294	0.8240	0.3477	0.3333
p346-utCOartR09	0.5324	0.4882	0.7448	0.3120	0.3137
p72-umd_ric_2	0.5441	0.4544	0.7807	0.2708	0.2867
p167-09RefR	0.3765	0.3603	0.5761	0.2443	0.2540
p25-ruc-base-casF	0.4441	0.4176	0.6270	0.2243	0.2523
p60-UJM_15488	0.4382	0.3853	0.6043	0.2146	0.2343

(d) Best in Context Task

Participant	P5	P10	1/rank	map	bpref
p48-LIG-2009-BIC-1B	0.6059	0.5338	0.8206	0.3573	0.3510
p62-RMIT09title	0.6029	0.5279	0.8237	0.3540	0.3488
p5-BM25AncestorBIC	0.6147	0.5294	0.8240	0.3477	0.3333
p36-utampere_given30_nolinks	0.6147	0.5294	0.8240	0.3477	0.3333
p6-UamsBAfbCMdocbi100	0.6147	0.5118	0.8531	0.3361	0.3251
p10-MPII-COBIBM	0.5824	0.5191	0.8451	0.3325	0.3315
p92-Lyon3LIAmanBEP*	0.6382	0.5279	0.7706	0.3305	0.3374
p25-ruc-term-coB	0.5206	0.4779	0.7158	0.3197	0.3251
p346-utCOartB09	0.5324	0.4882	0.7448	0.3120	0.3137
p60-UJM_15508	0.5324	0.4544	0.7020	0.2910	0.2925

Table 15. Top 10 Participants in the Thorough Task: Result length.

Participant	MAiP	# articles	# characters	# chars/art
p48-LIG-2009-thorough-3T	0.2855	588	5,621,997	9,554
p6-UAmsIN09article	0.2818	4,947	8,732,588	1,765
p5-BM25thorough	0.2585	632	5,195,586	8,213
p92-Lyon3LIAmanlmnt	0.2496	1,439	13,390,230	9,301
p60-UJM_15494	0.2435	551	1,461,857	2,648
p346-utCASartT09	0.2350	1,496	8,482,533	5,668
p10-MPII-CASThBM	0.2133	1,181	8,099,770	6,854
p167-09RefT	0.1390	1499	13,253,653	8,841
p68-I09LIP6OWATh	0.0630	976	4,400,118	4,508
p25-ruc-base-coT	0.0577	29	50,183	1,707

Table 16. Top 10 Participants in the Focused Task: Result length.

Participant	iP[0.01]	# articles	# characters	# chars/art
p78-UWFERBM25F2	0.6333	1,130	1,613,095	1,426
p68-I09LIP6Okapi	0.6141	1,485	16,868,585	11,351
p10-MPII-COFoBM	0.6134	1,319	2,137,482	1,619
p60-UJM_15525	0.6060	1,485	10,420,397	7,016
p6-UamsFSsec2docbi100	0.5997	1,213	5,745,657	4,734
p5-BM25BOTrangeFOC	0.5992	1,498	13,236,136	8,835
p16-Spirix09R001	0.5903	1,496	8,355,434	5,584
p48-LIG-2009-focused-1F	0.5853	1,357	7,570,394	5,576
p22-emse2009-150	0.5844	1,410	6,306,031	4,470
p25-ruc-term-coF	0.4973	29	55,010	1,865

article, which ranges from 1,707 to 9,554. The best run retrieves the highest number of characters per article. Recall from Section 2 that the length of a relevant article is 11,691 characters on average, and the number of relevant characters per article is 3,878 on average. Even runs that are relatively close in score seem to target radically different amounts of text per article.

Table 16 shows for the best Focused runs of the 10 best scoring groups statistics on number of articles, and characters retrieved. We see a similar spread in average number of characters per article, ranging from 1,426 to 11,351. The averages seem lower than for the Thorough Task. The best Focused run retrieves the lowest number of characters per article.

Table 17 shows for the best Relevant in Context runs of the 10 best scoring groups statistics on number of articles, and characters retrieved. We see again considerable spread in average number of characters per article, ranging from 677 to 8,841. The averages seem higher than for the Focused and Thorough Task. The second best Relevant in Context run retrieves the highest number of characters per article.

There is no analysis of result length for the Best in Context Task since for this task only a single best entry point is required.

Table 17. Top 10 Participants in the Relevant in Context Task: Result length.

Participant	MAgP	# articles	# characters	# chars/art
p5-BM25RangeRIC	0.1885	1,498	13,215,573	8,821
p4-Reference	0.1847	1,499	13,253,653	8,841
p6-UamsRSCMartCMdocbi100	0.1773	1,230	10,157,349	8,254
p48-LIG-2009-RIC-1R	0.1760	1,357	7,570,394	5,576
p36-utampere_given30_nolinks	0.1720	1,498	10,555,338	7,046
p346-utCASrefR09	0.1188	1,055	8,186,120	7,758
p60-UJM_15502	0.1075	1,102	1,446,938	1,312
p167-09RefR	0.1045	1,499	13,253,653	8,841
p25-ruc-base-casF	0.1028	660	2,814,934	4,264
p72-umd_ric_1	0.0424	464	314,342	677

Table 18. Top 10 Participants in the Thorough Task: Restricted to 500 characters per result.

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p5-BM25thorough	0.7032	0.6658	0.5511	0.4687	0.1625
p10-MPII-COThBM	0.6273	0.5009	0.3129	0.2187	0.0687
p6-UamsTSbi100	0.5908	0.4570	0.2900	0.1478	0.0445
p60-UJM_15500	0.6081	0.4896	0.2566	0.1143	0.0438
p48-LIG-2009-thorough-3T	0.6023	0.4792	0.2620	0.1283	0.0423
p25-ruc-base-coT	0.5334	0.4169	0.2387	0.1348	0.0414
p92-Lyon3LIAautolmnt	0.4651	0.3405	0.1878	0.0846	0.0280
p68-I09LIP6OkapiEl	0.3965	0.2839	0.1483	0.0692	0.0234
p72-umd_thorough_3	0.4235	0.2491	0.1103	0.0666	0.0216
p346-utCASartT09	0.5100	0.3574	0.1191	0.0288	0.0209

7.2 Limiting Result Length

In the previous section, we saw considerable spread in the numbers of characters per article retrieved. A partial explanation is the fact that making sure all relevant text is retrieved (avoiding false negatives) is easy, but making sure no non-relevant is retrieved (avoiding false positives) is very hard [4]. This leads to systems that prefer being “safe” (by retrieving whole articles or long elements) over being “sorry” (possibly missing relevant text by aiming for small elements). In many use-cases of focused retrieval there is a down-side to retrieving long excerpts or even entire documents. Think of mobile displays that can only show a certain number of characters, or think of query-biased summaries of documents that appear on the hit lists of modern search engines.

What if we limit the results to a maximum of 500 characters? For each run, we “cut off” each individual result after the first 500 retrieved characters. Table 18 shows the best Thorough run of the top 10 participating groups. This clearly hurts the overall performance, although run *p5-BM25thorough* is strikingly more effective than the other runs. Upon closer inspection, this run used a slice-and-dice approach to turn an article ranking into a list of all (overlapping) elements which contained at least one of the search terms. Recall from Table 15 that this

Table 19. Top 10 Participants in the Ad Hoc Track: Restricted to 500 characters per article.

(a)Thorough Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p10-MPII-COThBM	0.6357	0.4649	0.1893	0.0805	0.0335
p5-ANTbigramsBOTthorough	0.6337	0.4911	0.2031	0.0643	0.0326
p60-UJM_15500	0.5934	0.4305	0.1486	0.0387	0.0245
p48-LIG-2009-thorough-3T	0.5728	0.4109	0.1387	0.0295	0.0231
p6-UAmsIN09article	0.5671	0.4080	0.1452	0.0265	0.0228
p92-Lyon3LIAmanlmnt*	0.5024	0.3352	0.1253	0.0429	0.0215
p346-utCASartT09	0.5100	0.3574	0.1191	0.0288	0.0209
p25-ruc-base-coT	0.5876	0.3704	0.0965	0.0236	0.0177
p68-I09LIP6OkapiEl	0.4137	0.2355	0.0728	0.0362	0.0149
p167-09RefT	0.3131	0.2143	0.0807	0.0245	0.0137

(b) Focused Task

Participant	iP[.00]	iP[.01]	iP[.05]	iP[.10]	MAiP
p78-UWatFERBM25F	0.6776	0.5304	0.2517	0.1179	0.0414
p10-MPII-COFoBM	0.6330	0.4684	0.2025	0.0828	0.0346
p5-BM25FOC	0.6159	0.4532	0.1746	0.0504	0.0278
p60-UJM_15525	0.5931	0.4337	0.1510	0.0388	0.0251
p16-Spirix09R002	0.5626	0.4226	0.1719	0.0499	0.0256
p68-I09LIP6Okapi	0.5885	0.4138	0.1212	0.0272	0.0220
p48-LIG-2009-focused-3F	0.5665	0.4044	0.1370	0.0341	0.0230
p25-ruc-term-coF	0.6311	0.4044	0.0879	0.0413	0.0206
p22-emse2009-150*	0.6223	0.3700	0.1218	0.0352	0.0225
p6-UamsFSsec2docbi100	0.6346	0.3672	0.1212	0.0220	0.0207

(c) Relevant in Context Task

Participant	gP[5]	gP[10]	gP[25]	gP[50]	MAgP
p5-ANTbigramsRIC	0.2308	0.2069	0.1743	0.1367	0.1291
p36-utampere_given30_nolinks	0.1952	0.1909	0.1444	0.1201	0.1215
p6-UamsRSCMartCMdocbi100	0.1835	0.1565	0.1314	0.1132	0.1182
p4-Reference	0.1755	0.1671	0.1317	0.1065	0.1116
p48-LIG-2009-RIC-3R	0.1704	0.1634	0.1288	0.1039	0.1082
p60-UJM_15502	0.1471	0.1325	0.1041	0.0806	0.0857
p25-ruc-base-casF	0.1894	0.1736	0.1396	0.1248	0.0828
p346-utCASrefR09	0.1206	0.1072	0.0831	0.0670	0.0720
p167-09RefR	0.0965	0.0873	0.0800	0.0729	0.0667
p72-umd_ric_1	0.0642	0.0571	0.0444	0.0342	0.0321

run still retrieved 8,213 characters per article, so a better and more realistic filter would be to limit the number of characters retrieved per article.

We change our analysis and for each run, we "cut off" the results after having retrieved the first 500 retrieved characters per article (so any further text from the same article is ignored, and the result is removed from the run). Table 19 shows the results of restricting the official submissions to maximally 500 characters per article. This naturally leads to a much lower score on the overall measures, and a somewhat lower score on early ranks. We see some familiar runs from Tables 3–5 before, but also some new runs. The best run for the Thorough Task is a variant of the seventh ranked run in Table 3; the best run for the Focused Task was also the best run in Table 4; and the best run for the Relevant in Context Task is a variant of the best run in Table 5.

The system rank correlation (Kendall's Tau) between the official ranking and the restricted run ranking is the following.

- Over all 30 Thorough Task submissions the system rank correlation is 0.545.
- Over all 57 Focused Task submissions the system rank correlation is 0.633.
- Over all 33 Relevant in Context Task submissions the system rank correlation is 0.655.

Overall, we see a reasonable correspondence between the rankings for the ad hoc tasks in Section 3 and the rankings for the restricted runs in this section. This comes as no surprise since both task share an important aspect: finding those articles that contain relevant information.

8 Discussion and Conclusions

In this paper we analyzed the results of the INEX 2009 Ad Hoc Track. For details of the tasks, measures, and outcomes, we refer to the INEX 2009 Ad Hoc track overview paper [1]. In this paper, we focused on six different aspects of the ad hoc track evaluation, which we will discuss in turn.

First, we examined in detail the relevance judgments. The 2009 collection contained 2,666,190 Wikipedia articles (October 8, 2008 dump of the Wikipedia), which is four times larger than the earlier Wikipedia collection. What was the effect of this change in corpus size? We saw that the collection's size had little impact, but that the relevant articles were much longer (a mean length 3,030 in 2008 and 5,775 in 2009, a 52% increase), leading to a lower fraction of highlighted text per article (a mean of 58% in 2008 and 33% in 2009). This also reduced the correlation between focused retrieval and article retrieval, e.g., from 79% for the "in context" tasks in 2008 to 51–58% in 2009.

Second, we studied the resulting system rankings, for each of the four ad hoc tasks, and determined whether differences between the best scoring participants are statistically significant. The early precision measure of the focused task, interpolated precision at 1% recall, is inherently unstable, and only very few of the differences between runs are statistically significant. The overall measures, the MAiP and MAgP variants of mean average precision, are able to distinguish the majority of pairs of runs. Almost 3/4 of system pairs are significantly different with the mean average generalized precision measure of the Relevant in Context task.

Third, we restricted our attention to particular run types: element and passage runs, keyword and phrase query runs, and systems using a reference run with a solid article ranking. Thirteen submissions used ranges of elements or FOL passage results, whereas 144 submissions used element results. Still the non-element submissions were competitive with the top ranking runs for both the Focused and Relevant in Context Tasks, and the second ranking run for the Best in Context Task. Ten submissions used the explicitly annotated phrases of the phrase query. Phrase query runs were competitive with several of them in the overall top 10 results, but the impact of the phrases seemed marginal. Recall, that the exact same terms were present in the CO query, and the only difference was the phrase annotation. There were 19 submissions using the reference run providing a solid article ranking for further processing. These runs turned out to be competitive, with runs in the top 10 for all tasks. Hence the reference run was successful in helping participants to create high quality runs. However, runs based on the reference run were not directly comparable, since they had different underlying article rankings.

Fourth, we examined the relative effectiveness of content only (CO, or Keyword) search as well as content and structure (CAS, or structured) search. We found that for all tasks the best scoring runs used the CO query but some CAS runs were in the top 10 for all four tasks. Part of the explanation may be in the low number of CAS submissions (40) in comparison with the number of CO submissions (117). Only 50 of the 68 judged topics had a non-trivial CAS query, and the majority of those CAS queries made only reference to particular tags and not on their structural relations. The YAGO tags potentially expressing an information need naturally in terms of structural constraints, were popular: 36 CAS queries used them (21 of them judged). Over the 50 non-trivial CAS queries, most groups had a better performing run using the CO query. A notable exception was *QUT* who had better performance for CAS on the Focused Task.

Fifth, we looked at the ability of focused retrieval techniques to rank articles. As in earlier years, we saw that article retrieval is a reasonably effective at XML-IR: for each of the ad hoc tasks there were three article-only runs among the best runs of the top 10 groups. When looking at the article rankings inherent in all Ad Hoc Track submissions, we saw that again three of the best runs of the top 10 groups in terms of article ranking (across all three tasks) were in fact article-only runs. This also suggests that element-level or passage-level evidence is valuable for article retrieval. When comparing the system rankings in terms of article retrieval with the system rankings in terms of the ad hoc retrieval tasks, over the exact same topic set, we see a reasonable correlation. The systems with the best performance for the ad hoc tasks, also tend to have the best article rankings.

Sixth, we studied the length of retrieved results, and looked at the impact of restricting result length. We looked at the average number of characters per article that each run retrieved, and found that there is an enormous spread from less than 2,000 characters (less than the mean length of relevant text per article) to over 10,000 characters (longer than the mean length of relevant articles). Even runs scoring close on the Ad Hoc Track measures could apply radically different strategies. For many use-cases the result length is an issue, and we modified the official submission so that only the first retrieved 500 characters per article were retained. There resulting system rankings show agreement with the original scores, with a system rank correlation in the range 0.55–0.66, but also some new runs in the top 10 per task.

Acknowledgments Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grants 612.066.513, 639.072.601, and 640.001.501).

Bibliography

[1] S. Geva, J. Kamps, M. Lehtonen, R. Schenkel, J. A. Thom, and A. Trotman. Overview of the INEX 2009 ad hoc track. In S. Geva, J. Kamps, and A. Trotman, editors, *Focused Retrieval and Evaluation*, Lecture Notes in Computer Science. Springer, 2010.

[2] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. XML retrieval: What to retrieve? In C. Clarke, G. Cormack, J. Callan, D. Hawking, and A. Smeaton, editors, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 409–410. ACM Press, New York NY, 2003.

[3] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. Length normalization in XML retrieval (extended abstract). In R. Verbrugge, N. Taatgen, and L. Schomaker, editors, *BNAIC-2004: Proceedings of the 16th Belgium-Netherlands Conference on Artificial Intelligence*, pages 369–370, 2004.

[4] J. Kamps, M. Koolen, and M. Lalmas. Locating relevant text within XML documents. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, editors, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 847–849. ACM Press, New York NY, USA, 2008.

A Appendix: Full run names

Group	Run	Label	Task	Query	Results	Notes
4	617	Reference	RiC	CO	Ele	Reference run Article-only
5	744	BM25AncestorBIC	BiC	CO	Ele	Article-only
5	749	ANTbigramsRIC	RiC	CO	Ele	
5	757	BM25thorough	Tho	CO	Ele	
5	775	BM25ArticleFOC	Foc	CO	Ele	Article-only
5	776	BM25FOC	Foc	CO	Ele	
5	777	BM25RangeFOC	Foc	CO	Ran	Article-only
5	781	BM25BOTrangeFOC	Foc	CAS	Ran	Article-only
5	792	ANTbigramsRangeFOC	Foc	CO	Ran	Article-only
5	796	BM25ArticleRIC	RiC	CO	Ele	Article-only
5	797	BM25RangeRIC	RiC	CO	Ran	Article-only
5	804	BM25BOTrangeRIC	RiC	CAS	Ran	Article-only
5	807	ANTbigramsBOTthorough	Tho	CAS	Ele	
5	808	BM25BOTthorough	Tho	CAS	Ele	
5	824	BM25bepBIC	BiC	CO	Ele	Article-only
5	825	BM25BOTbepBIC	BiC	CAS	Ele	Article-only
6	634	UAmsIN09article	Tho	CO	Ele	Article-only
6	810	UamsTAbi100	Tho	CO	Ele	Article-only
6	811	UamsTSbi100	Tho	CO	Ele	
6	813	UamsFSsec2docbi100	Foc	CAS	Ele	
6	814	UamsRSCMartCMdocbi100	RiC	CO	Ele	
6	816	UamsBAfbCMdocbi100	BiC	CO	Ele	Article-only
6	817	UamsBSfbCMsec2docbi100art1	BiC	CAS	Ele	Article-only
10	618	MPII-CASFoBM	Foc	CAS	Ele	
10	619	MPII-COFoBM	Foc	CO	Ele	
10	620	MPII-CASThBM	Tho	CAS	Ele	
10	621	MPII-COThBM	Tho	CO	Ele	
10	628	MPII-COArBM	Foc	CO	Ele	Article-only
10	632	MPII-COBIBM	BiC	CO	Ele	Article-only
10	700	MPII-COArBP	Foc	CO	Ele	Article-only
10	709	MPII-COArBPP	Foc	CO	Ele	Phrases Article-only
16	872	Spirix09R001	Foc	CAS	Ele	Article-only
16	873	Spirix09R002	Foc	CAS	Ele	Article-only
22	672	emse2009-150	Foc	CO	Ele	Phrases Manual
25	727	ruc-base-coT	Tho	CO	Ele	
25	737	ruc-term-coB	BiC	CO	Ele	
25	738	ruc-term-coF	RiC	CO	Ele	
25	739	ruc-term-coF	Foc	CO	Ele	
25	898	ruc-base-casF	Foc	CAS	Ele	
25	899	ruc-base-casF	RiC	CAS	Ele	
36	688	utampere_given30_nolinks	RiC	CO	Ele	Reference run
36	701	utampere_given30_nolinks	BiC	CO	Ele	Reference run
36	708	utampere_auth_40_top30	RiC	CO	Ran	
48	682	LIG-2009-thorough-1T	Tho	CO	Ele	
48	684	LIG-2009-thorough-3T	Tho	CO	Ele	Reference run

Continued on Next Page...

Group	Run	Label	Task	Query	Results	Notes
48	685	LIG-2009-focused-1F	Foc	CO	Ele	
48	686	LIG-2009-focused-3F	Foc	CO	Ele	Reference run
48	714	LIG-2009-RIC-1R	RiC	CO	Ele	
48	716	LIG-2009-RIC-3R	RiC	CO	Ele	Reference run
48	717	LIG-2009-BIC-1B	BiC	CO	Ele	
48	719	LIG-2009-BIC-3B	BiC	CO	Ele	Reference run
55	836	doshisha09f	Foc	CAS	Ele	
60	819	UJM_15518	Foc	CO	Ele	Reference run
60	820	UJM_15486	Tho	CO	Ele	
60	821	UJM_15500	Tho	CO	Ele	
60	822	UJM_15494	Tho	CO	Ele	Reference run
60	827	UJM_15488	RiC	CO	Ele	
60	828	UJM_15502	RiC	CO	Ele	
60	829	UJM_15503	RiC	CO	Ele	Reference run
60	830	UJM_15490	BiC	CO	Ele	
60	832	UJM_15508	BiC	CO	Ele	Reference run
60	868	UJM_15525	Foc	CO	Ele	Article-only
62	895	RMIT09title	BiC	CO	Ele	Article-only
62	896	RMIT09titleO	BiC	CO	FOL	Article-only
68	679	I09LIP6Okapi	Foc	CO	Ele	Article-only
68	681	I09LIP6OWA	Foc	CO	Ele	Article-only
68	703	I09LIP6OkapiEl	Tho	CO	Ele	
68	704	I09LIP6OWATh	Tho	CO	Ele	
72	666	umd_ric_1	RiC	CO	Ele	
72	667	umd_ric_2	RiC	CO	Ele	
72	870	umd_thorough_3	Tho	CO	Ele	
78	706	UWatFERBase	Foc	CO	FOL	
78	707	UWatFERBM25F	Foc	CO	FOL	
92	694	Lyon3LIAautoBEP	BiC	CAS	Ele	Phrases
92	695	Lyon3LIAmanBEP	BiC	CO	Ele	Phrases Manual Article-only
92	697	Lyon3LIAmanQE	Foc	CO	Ele	Phrases Manual Article-only
92	698	Lyon3LIAautolmnt	Tho	CO	Ele	Phrases
92	699	Lyon3LIAmanlmnt	Tho	CO	Ele	Phrases Manual
167	651	09RefT	Tho	CO	Ele	Reference run Article-only
167	654	09LrnRefF	Foc	CO	Ele	Reference run Article-only
167	657	09RefR	RiC	CO	Ele	Reference run Article-only
167	660	09LrnRefB	BiC	CO	Ele	Reference run Article-only
346	637	utCASartT09	Tho	CAS	Ele	Article-only
346	638	utCASartF09	Foc	CAS	Ele	Article-only Invalid
346	639	utCOartR09	RiC	CO	Ele	Article-only Invalid
346	640	utCOartB09	BiC	CO	Ele	Article-only Invalid
346	645	utCASrefF09	Tho	CAS	Ele	Reference run
346	646	utCASrefF09	Foc	CAS	Ele	Reference run
346	647	utCASrefR09	RiC	CAS	Ele	Reference run
346	648	utCASrefB09	BiC	CAS	Ele	Reference run