Overview of the TREC 2020 Health Misinformation Track

Charles L. A. Clarke[1], Maria Maistro[2], Saira Rizvi[1], Mark D. Smucker[1], and Guido Zuccon[3]

[1]University of Waterloo
[2]University of Copenhagen
[3]University of Queensland

1 Introduction

TREC 2020 was the second year for the Health Misinformation track, which was named the Decision Track in 2019 [1]. Information retrieval using document collections that contain misinformation are problematic. When a search engine returns documents that contain misinformation, users may have difficulty discerning correct from incorrect information and the incorrect information can lead to incorrect decisions [5]. Decisions regarding health-related topics can be consequential, and as such we want search engines that enable users to make correct decisions. The track is designed to address the problem of health misinformation in three areas: 1) adhoc retrieval, 2) the total recall of misinformation in the collection, and 3) the evaluation of retrieval in the presence of misinformation.

The 2020 Health Misinformation track had both a recall task and an adhoc task for participants. With the onset of the coronavirus pandemic (COVID-19), the track organizers selected a document collection of news from the Common Crawl[1] that covered the first four months of 2020. The track's topics were all related to COVID-19 and posed as questions such as "Can gargling salt water prevent COVID-19?" For both tasks, NIST assessors judged documents' usefulness for answering a topic's question, and if judged to be useful, assessors then recorded if the document contained a specific answer to the question and then judged the credibility of the document. We evaluated recall runs on their ability to find all documents containing incorrect information (misinformation). For adhoc runs, we measured their ability to return useful, correct, and credible information.

2 Topics

The COVID-19 pandemic has highlighted the dangers that uncontrolled proliferation of misinformation can have on consumer health. Therefore, the topics for this track focused on the consumer health search domain relevant to COVID-19, i.e., people seeking medical advice online. Unlike other TREC tracks, the assessors did not create the topics and were instead provided with topics that included title, description, answer, narrative, and evidence fields. Figure 1 shows an example of a topic.

The final topics were shortlisted from a collection of 74 candidate topics that were collected primarily using WHO mythbusters page[2] and Harvard Medical School page on treatments for

[1]https://commoncrawl.org/2016/10/news-dataset-available/

[2]https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

COVID-19[3]. Rest of the topics were created using well known fact-checking websites (Snopes, etc). Each topic was assessed on its quality by ensuring that the corpus contains at least one example of a negative and a positive document. The final list was created by filtering out topics for which misinformation did not exist in the corpus, and the topics that were less prevalent on the internet. The topic were provided in an XML file[4].

The title field of each topic is built as a pair of treatment and disease, where for TREC 2020, the disease is always COVID-19. The description is in the form of a question and is built as a triplet of (treatment, effect, disease), where the effect can be: cause, prevent, worsen, cure, help. Only these terms were used, resulting in descriptions of form: "Can X Y COVID-19?", where X is a treatment and Y is one of the five effect terms. For example, "Can vitamin D cure COVID-19?" The answer field is either"yes" or "no". The participants were asked to consider this field as the ground truth for the task. However, since the answer corresponds to the topic writer's best understanding of medical consensus at the time of topic creation, it is not medical advice, and should not be taken as truth outside of the context of this track. For the total recall task, participants were asked to identify all documents that contradicted this field. For the adhoc task, the participants were to return the most credible and complete information supporting the answer field for each topic. Finally, the evidence field contains the URL of a page from the open Web that was used to determine this answer. This page may or may not be part of the corpus. The narrative and evidence fields were created to assist the assessors in document judgment and were not to be used by the participants to generate automatic runs. All other fields were allowed to be used.

During the the topic creation, there was cut-and-paste error resulting in a duplicate topic. The duplicate topics were 9 and 17, Convalescent plasma COVID-19. However, this error did not otherwise jeopardize the integrity of the results.

```
<topics>
<topic>
<number>0</number>
<query>ibuprofen COVID-19</query>
<description>Can ibuprofen worsen COVID-19?</description>
<alignment>no</alignment>
<narrative>Ibuprofen is an anti-inflammatory drug used to reduce fever and
treat pain or inflammation. Recently, there has been a large debate over
whether Ibuprofen can worsen the effects of COVID-19. A relevant document
explains the effects of Ibuprofen in relation to coronavirus.</narrative>
</topic>
<topic>
...
</topic>
</topics>
```

Figure 1: Example of topic for the Health Misinformation Track 2020

[3] https://www.health.harvard.edu/diseases-and-conditions/treatments-for-covid-19

[4] https://trec-health-misinfo.github.io/topics.xml

3 Document Collection

The track used the documents found in the CommonCrawl News crawl from January 1, 2020 to April 30, 2020. The crawl contains news articles from websites all over the world. The crawl is distributed in the Web ARChive (WARC) format. Each document in a WARC file contains a WARC header and the raw data from the crawl. The WARC files contain a mixture of records. The only records we used had a WARC-Type of "response". Some participants returned non-response WARC records, e.g. request records. The collection contains non-English documents, and all non-English documents were considered to be non-relevant regardless of their content. NIST provided the collection also in WET format[5]. The WET files contain only extracted content from the web pages, i.e. they parse the HTML and extract the text content. The four months of the CommonCrawl New Crawl is available via Amazon S3 at URIs:

```
s3://commoncrawl/crawl-data/CC-NEWS/2020/01
s3://commoncrawl/crawl-data/CC-NEWS/2020/02
s3://commoncrawl/crawl-data/CC-NEWS/2020/03
s3://commoncrawl/crawl-data/CC-NEWS/2020/04
```

The "docno" or document identifier used was the WARC-Record-ID field for the WARC files and the WARC-Refers-To field for the WET files. Only unique id was used, for example, from the field value of `<urn:uuid:49ecaf74-b1aa-4563-83a0-c81cece0e284>`, only the id was used: `49ecaf74-b1aa-4563-83a0-c81cece0e284`.

4 Tasks Overview

The track had two tasks: 1) Total Recall and 2) AdHoc Retrieval.

4.1 Task 1: Total Recall

For the total recall task, the goal is to identify all the documents conveying incorrect information for a specific set of topics. Documents contradicting the topic's answer are assumed to be misinformation. Participants were to identify all documents in a collection that promulgate, promote, and/or support that misinformation. For example, for the topic "Can Ibuprofen worsen COVID-19", participants were to identify all documents indicating that Ibuprofen can worsen COVID-19. Documents making this claim for the purposes of debunking it are not misinformation. Participants submitted runs that ranked documents according to the likelihood that they promulgate misinformation. Up to 10,000 documents per topic could be submitted.

4.2 Task 2: AdHoc Retrieval

For the ad-hoc retrieval task, the goal is to design a ranking model that promotes credible and correct information over incorrect information. For a given topic, participants were to return relevant, credible, and correct information that will help searchers make correct decisions. Participants were to assume that the statement included in the topic description is correct or not, based on the answer field, even if they knew current medical or other evidence suggests otherwise. Runs were allowed to contain a maximum of 1,000 documents per topic.

Note that this task is more than simply a new definition of what is relevant. There are multiple types of results: useful and correct and credible, useful and correct but not credible, etc. as well as

[5] `https://ir.nist.gov/trec-hmi/` (password protected)

incorrect and non-useful documents. It is important that search results avoid containing incorrect results. In place of notions of correctness, the credibility of the information source is useful, and useful and credible information is preferred.

5 Submitted Runs

Six groups submitted a total of 23 runs to the total recall task. Eight groups submitted 51 runs to the adhoc retrieval task. The UWaterlooMDS group submitted two baseline runs to each task on the behalf of the track organizers: `bm25_desc` and `bm25_title`. These runs used Anserini's default BM25 implementation with the WET files to retreive documents using either the topic description or the topic title as the query.

6 Evaluation

Runs were evaluated by using two pieces of software: 1) an extension of trec_eval[6] to compute multiple aspect measures [4] and 2) a script[7] to compute the compatibility measure [3, 2]. In all cases, we derive a qrels file to use with the specific measure from the original NIST qrels files.

6.1 qrels (query-relevance files)

NIST used the track's relevance assessing guidelines[8] to generate the track's qrels. The format adopted for NIST qrels file is as follows:

```
topic_id 0 doc_id usefulness-judgment answer-judgment credibility-judgment
```

where the columns are space separated. Documents were assessed by NIST assessors with respect to 3 criteria, which were recorded in the NIST qrels as follows:

- *Usefulness*: does the document contain material that the search user might find useful in answering the topic's question? Usefulness was assessed on a binary scale: 0 if the document is not useful in answering the question and 1 otherwise. This is column 4 of the qrels file.
- *Answer*: does the document answer to the question in the description field? If so, is the answer yes or no? The answer will be assessed with 3 values: 0 means that the document does not answer the question, 1 means that the answer provided by the document is "yes", and 2 means that the answer is "yes". A value of -1 means that NIST made no judgment. This is column 5 of the qrels file.
- *Credibility*: how credible is the document? Credibility was assessed on a binary scale, where 0 stands for not credible and 1 stands for credible. A value of -1 means that NIST made no judgment. This is column 6 of the qrels file.

Notes:

- When a document was judged as not useful, it was not judged for its answer nor for its credibility. In some cases, a useful document was accidentally not judged for its answer or credibility, i.e. a "skip".

[6] `https://github.com/trec-health-misinfo/Trec_eval_extension`
[7] `https://github.com/trec-health-misinfo/Compatibility`
[8] `https://trec-health-misinfo.github.io/docs/AssessingGuidelines-2020.pdf`

- Some participants submitted docids that were not WARC doc types of "response". While not explained clearly in the track guidelines, only WARC records of type "response" should have been used. In almost all cases, when an assessor was given a non-response docid to judge, it was judged "not useful". Rather than confuse matters by including non-response documents in the qrels, these qrels contain only the judgements for docs of type "response".
- There are four missing topics: 33, 35, 36, and 48. NIST ran out of time and was not able to judge all topics.
- Topics 9 and 17 are duplicate topics, but have different judgments. We accidentally duplicated the topics during the process of consolidating our selected topics and converting them to XML.

Mapping to Correctness: Before computing any evaluation measure, we map answer labels to *correctness* labels. Correctness labels are obtained from the topic answer field and the answer labels provided by the assessors: a document is correct if it contains an answer that matches the topic's given answer. A document is *incorrect* if it contains an incorrect answer. For a document to be correct or incorrect, it has to be useful.

Observe that "not correct" does not necessarily mean "incorrect". A not useful document is neither correct nor incorrect, since a not useful document is off topic. Likewise, if a document does not contain an answer or is not judged for "answer", it is neither correct nor incorrect. For example, in Figure 1 the topic's given answer is "no". If a document for that topic has an answer of "no", then it is correct; if it has an answer of "yes", then it is incorrect; and if it has no answer, then it is not correct. Thus, the final correctness qrels will be as follows:

```
topic_id 0 doc_id usefulness-judgment correctness-judgment credibility-judgment
```

where the answer label is replaced by correctness. Correctness can assume the values: -1 for not judged documents; 0 for incorrect documents; 1 for correct documents; 2 for documents that do not provide an answer.

6.2 Derived qrels

We take the NIST qrels and the topics and generate derived qrels for the various evaluation measures. For the derived qrels, we only include "relevant" documents, for not all topics contain results that meet the success criteria of a "relevant" document in the derived qrels. By excluding topics without "relevant" documents, the effectiveness measures are only computed over topics for which runs could feasible get a non-zero score.

6.2.1 Graded / preference levels

For the compatibility measure, we converted the 3 aspects judged for documents (usefulness, correctness and credibility) into a basic preference ordering or graded relevance values as reported in Table 1.

It is tempting to use the above scores to compute nDCG, but that ignores the incorrect information. A better solution is to create a set of helpful and harmful qrels. We create helpful qrels by taking all scores greater then zero. The scores define preference levels with a higher scoring document being preferred to a lower scoring documents. To create the harmful qrels, we use only the absolute value of the negative scores. Thus, the most harmful documents are those that are judged to be useful, incorrect, and credible.

Table 1: Preference ordering for documents which are mapped to graded relevance labels.

Score	Description	Usefulness	Correctness	Credibility
4	Useful, correct, credible	1	1	1
3	Useful, correct, not credible or no credibility judgment	1	1	0 or −1
2	Useful, no answer or no judgment for answer, credible	1	2 or −1	1
1	Useful, no answer or no judgment, not credible or no judgment	1	2 or −1	0 or −1
0	Not useful, ignore answer and credibility.	0	-	-
−1	Useful, incorrect, not credible or no judgment	1	0	0 or −1
−2	Useful, incorrect, credible	1	0	1

6.2.2 Binary Relevance

We created a series of qrels files in the standard qrels format for binary relevance effectiveness measures. We made the following variations to allow us to use nDCG to evaluate runs in terms of a single aspect as well as combinations:

- Usefulness. Ignores answer correctness and document credibility.
- Useful and correct. Note that a document cannot be judged correct unless it is judged useful.
- Useful and credible. Note that a document cannot be judged credible unless it is judged useful.
- Useful and correct and credible.
- Incorrect. A document is incorrect if it is useful and contains and answer that is opposite of the topic's given answer.

6.2.3 Multiple Aspect qrels

We created both two aspect and three aspect qrels. For the three aspects qrels, this is the same as the correctness qrels file except that the correctness column is mapped to 1 if the document's answer matches the topic's, and to 0 otherwise (no distinction for not judged or no answer); and the credibility column is the same except that a −1 (not judged) is mapped to 0 (not credible). The two aspect qrels are the same but only consider usefulness and one of the other two aspects.

6.3 Task 1: Total Recall

For the total recall task, we use the binary qrels where a document is "relevant" if it is incorrect. Using these qrels, we compute R-precision, which is equivalent to R-recall.

6.4 Task 2: AdHoc Retrieval

For ad-hoc retrieval, the primary method is to use compatibility of results with helpful and harmful results. For secondary analysis purposes, we computed nDCG using the binary qrels for each of the aspects and the conjunction of all three. Finally, we computed the CAM of all three aspects using mean average precision.

7 Results

Four topics were not assessed by NIST: 33, 35, 36, and 48. Therefore, they are excluded from the following analyses. Likewise, for some versions of the derived qrels, there are no "relevant" documents for some topics, and thus different measures are often computed over different number of topics.

Tables 2 and 3 report results for the adhoc task. Figure 2 plots adhoc runs' compatibility with helpful and harmful results. For two runs with the same level of compatibility with helpful results, the run with the lower compatibility with harmful results is to be preferred. Thus the h2oloo.m8 run stands out for being both very helpful and with some of the least harm. Table 4 reports results for the total recall task.

8 Acknowledgments

Thanks to Mustafa Abualsaud, Kamyar Ghajar, Linh Nhi Phan Minh, Amir Vakili Tahami, Nicole Yan, and Dake Zhang. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), the facilities of Compute Canada, the University of Waterloo, and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 893667. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

[1] M. Abualsaud, M. D. Smucker, C. Lioma, M. Maistro, and G. Zuccon. Overview of the trec 2019 decision track. In *The Twenty-Eigth Text REtrieval Conference (TREC 2019) Proceedings*, 2019. 19 pages.

[2] C. L. Clarke, A. Vtyurina, and M. D. Smucker. Offline evaluation without gain. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ICTIR '20, pages 185–192, 2020.

[3] C. L. A. Clarke, M. D. Smucker, and A. Vtyurina. Offline evaluation by maximum similarity to an ideal ranking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 225–234, 2020.

[4] C. Lioma, J. G. Simonsen, and B. Larsen. Evaluation Measures for Relevance and Credibility in Ranked Lists. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR 2017, pages 91–98, 2017. ISBN 978-1-4503-4490-6.

[5] F. A. Pogacar, A. Ghenai, M. D. Smucker, and C. L. A. Clarke. The Positive and Negative Influence of Search Results on People's Decisions About the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, pages 209–216, 2017.

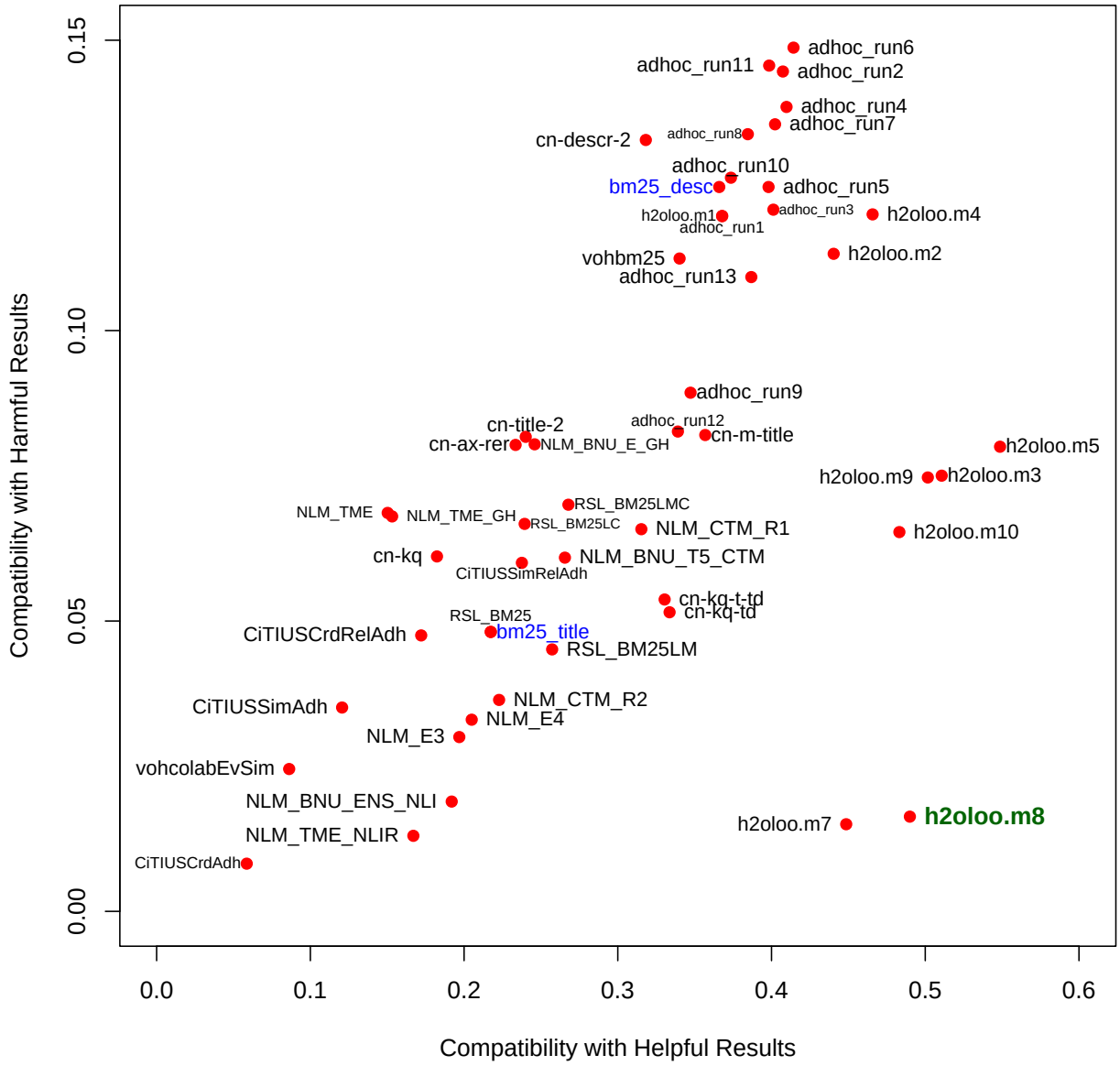

Figure 2: Adhoc results: Compatibility of runs with helpful and harmful results. A good run is helpful and not harmful. For a given level of helpfulness, a run with less harm is to be preferred.

			CAM	Compatibility		
Group	Run	Type	MAP	help	harm	help-harm
h2oloo	h2oloo.m8	auto	0.253	0.490	0.016	0.474
h2oloo	h2oloo.m5	auto	0.319	0.549	0.080	0.469
h2oloo	h2oloo.m3	auto	0.292	0.511	0.075	0.436
h2oloo	h2oloo.m7	auto	0.222	0.449	0.015	0.434
h2oloo	h2oloo.m9	auto	0.297	0.502	0.075	0.427
h2oloo	h2oloo.m10	auto	0.286	0.483	0.065	0.418
h2oloo	h2oloo.m4	auto	0.297	0.466	0.120	0.346
h2oloo	h2oloo.m2	auto	0.273	0.440	0.113	0.327
Webis	cn-kq-td	manual	0.080	0.334	0.052	0.282
KU	adhoc_run3	auto	0.250	0.401	0.121	0.280
KU	adhoc_run13	auto	0.249	0.387	0.109	0.278
Webis	cn-kq-t-td	manual	0.086	0.331	0.054	0.277
Webis	cn-m-title	manual	0.136	0.357	0.082	0.275
KU	adhoc_run5	auto	0.249	0.398	0.125	0.273
KU	adhoc_run4	auto	0.278	0.410	0.139	0.271
KU	adhoc_run7	auto	0.278	0.402	0.136	0.267
KU	adhoc_run6	auto	0.284	0.414	0.149	0.266
KU	adhoc_run2	auto	0.280	0.407	0.145	0.263
KU	adhoc_run9	auto	0.249	0.347	0.089	0.258
KU	adhoc_run12	auto	0.212	0.339	0.083	0.257
KU	adhoc_run11	auto	0.282	0.399	0.146	0.253
KU	adhoc_run8	auto	0.272	0.385	0.134	0.251
NLM	NLM_CTM_R1	auto	0.119	0.315	0.066	0.250
KU	adhoc_run1	auto	0.237	0.368	0.120	0.248
h2oloo	h2oloo.m1	auto	0.237	0.368	0.120	0.248
KU	adhoc_run10	auto	0.263	0.374	0.126	0.247
UWaterlooMDS	bm25_desc	auto	0.236	0.366	0.125	0.241
vohcolab	vohbm25	auto	0.237	0.340	0.112	0.228
RealSakaiLab	RSL_BM25LM	auto	0.138	0.257	0.045	0.212
NLM	NLM_BNU_T5_CTM	auto	0.083	0.266	0.061	0.205
RealSakaiLab	RSL_BM25LMC	auto	0.148	0.268	0.070	0.198
NLM	NLM_CTM_R2	auto	0.051	0.223	0.036	0.186
Webis	cn-descr-2	auto	0.131	0.318	0.133	0.185
CiTIUS	CiTIUSSimRelAdh	manual	0.079	0.238	0.060	0.178
NLM	NLM_BNU_ENS_NLI	auto	0.037	0.192	0.019	0.173
RealSakaiLab	RSL_BM25LC	auto	0.139	0.239	0.067	0.173
NLM	NLM_E4	auto	0.061	0.205	0.033	0.172
RealSakaiLab	RSL_BM25	auto	0.139	0.217	0.048	0.169
UWaterlooMDS	bm25_title	auto	0.139	0.217	0.048	0.169
NLM	NLM_E3	auto	0.051	0.197	0.030	0.167
NLM	NLM_BNU_E_GH	auto	0.089	0.246	0.080	0.166
Webis	cn-title-2	auto	0.114	0.240	0.082	0.158
NLM	NLM_TME_NLIR	auto	0.033	0.167	0.013	0.154
Webis	cn-ax-rer	auto	0.111	0.234	0.080	0.153
CiTIUS	CiTIUSCrdRelAdh	auto	0.036	0.172	0.048	0.125
Webis	cn-kq	manual	0.044	0.182	0.061	0.121
CiTIUS	CiTIUSSimAdh	manual	0.025	0.121	0.035	0.086
NLM	NLM_TME_GH	auto	0.086	0.153	0.068	0.085
NLM	NLM_TME	auto	0.086	0.150	0.069	0.082
vohcolab	vohcolabEvSim	manual	0.057	0.086	0.025	0.062
CiTIUS	CiTIUSCrdAdh	auto	0.004	0.059	0.008	0.050

Table 2: Adhoc run results with CAM MAP and Compatibility measures.

			nDCG on binary qrels			
Group	Run	Type	Useful	Correct	Credible	All
h2oloo	h2oloo.m5	auto	0.666	0.590	0.631	0.561
h2oloo	h2oloo.m10	auto	0.628	0.583	0.607	0.560
h2oloo	h2oloo.m9	auto	0.644	0.577	0.617	0.553
h2oloo	h2oloo.m8	auto	0.602	0.581	0.573	0.544
h2oloo	h2oloo.m3	auto	0.644	0.562	0.607	0.531
h2oloo	h2oloo.m4	auto	0.660	0.555	0.619	0.522
h2oloo	h2oloo.m7	auto	0.576	0.564	0.549	0.521
KU	adhoc_run3	auto	0.617	0.508	0.595	0.507
h2oloo	h2oloo.m2	auto	0.639	0.534	0.595	0.501
KU	adhoc_run5	auto	0.619	0.505	0.592	0.496
KU	adhoc_run13	auto	0.593	0.479	0.572	0.494
KU	adhoc_run1	auto	0.608	0.500	0.577	0.485
h2oloo	h2oloo.m1	auto	0.608	0.500	0.577	0.485
UWaterlooMDS	bm25_desc	auto	0.605	0.495	0.574	0.483
KU	adhoc_run11	auto	0.633	0.495	0.597	0.477
KU	adhoc_run2	auto	0.616	0.484	0.587	0.477
KU	adhoc_run6	auto	0.622	0.481	0.594	0.472
KU	adhoc_run10	auto	0.603	0.484	0.563	0.470
KU	adhoc_run4	auto	0.623	0.483	0.588	0.469
KU	adhoc_run7	auto	0.623	0.481	0.589	0.466
vohcolab	vohbm25	auto	0.608	0.477	0.577	0.459
KU	adhoc_run8	auto	0.619	0.475	0.584	0.457
KU	adhoc_run12	auto	0.568	0.440	0.557	0.442
KU	adhoc_run9	auto	0.586	0.451	0.576	0.441
RealSakaiLab	RSL_BM25LMC	auto	0.471	0.345	0.452	0.338
RealSakaiLab	RSL_BM25LC	auto	0.470	0.336	0.447	0.331
RealSakaiLab	RSL_BM25LM	auto	0.454	0.334	0.431	0.319
UWaterlooMDS	bm25_title	auto	0.461	0.327	0.441	0.318
RealSakaiLab	RSL_BM25	auto	0.461	0.327	0.441	0.318
Webis	cn-m-title	manual	0.443	0.337	0.420	0.318
NLM	NLM_CTM_R1	auto	0.386	0.338	0.386	0.316
vohcolab	vohcolabEvSim	manual	0.429	0.327	0.395	0.306
Webis	cn-descr-2	auto	0.385	0.316	0.360	0.305
NLM	NLM_TME_GH	auto	0.365	0.298	0.333	0.285
NLM	NLM_TME	auto	0.365	0.298	0.332	0.283
Webis	cn-title-2	auto	0.413	0.290	0.383	0.275
Webis	cn-ax-rer	auto	0.404	0.280	0.376	0.265
NLM	NLM_E4	auto	0.316	0.287	0.299	0.265
Webis	cn-kq-t-td	manual	0.341	0.280	0.332	0.264
CiTIUS	CiTIUSSimRelAdh	manual	0.332	0.243	0.305	0.235
NLM	NLM_BNU_E_GH	auto	0.293	0.236	0.279	0.228
NLM	NLM_E3	auto	0.276	0.246	0.258	0.224
Webis	cn-kq-td	manual	0.258	0.227	0.263	0.212
NLM	NLM_BNU_T5_CTM	auto	0.252	0.208	0.264	0.196
NLM	NLM_CTM_R2	auto	0.238	0.202	0.233	0.190
NLM	NLM_TME_NLIR	auto	0.192	0.175	0.172	0.165
CiTIUS	CiTIUSCrdRelAdh	auto	0.209	0.146	0.181	0.139
Webis	cn-kq	manual	0.198	0.150	0.188	0.136
NLM	NLM_BNU_ENS_NLI	auto	0.141	0.146	0.131	0.133
CiTIUS	CiTIUSSimAdh	manual	0.164	0.125	0.152	0.121
CiTIUS	CiTIUSCrdAdh	auto	0.059	0.042	0.057	0.041

Table 3: Adhoc run results with binary qrels.

Group	Run	Type	R-prec
KU	run5	auto	0.130
KU	run11	auto	0.127
KU	run7	auto	0.127
KU	run2	auto	0.124
KU	run8	auto	0.123
KU	run6	auto	0.117
KU	run10	auto	0.114
KU	run3	auto	0.112
KU	run4	auto	0.109
UWaterlooMDS	bm25-desc	auto	0.104
vohcolab	vohbm25rm3	auto	0.103
NLM	NLM_CTM_R1_C	auto	0.098
KU	run1	auto	0.094
KU	run9	auto	0.087
NLM	NLM_TME_NLIR_C	auto	0.063
UWaterlooMDS	bm25-title	auto	0.051
vohcolab	vohEvDiv_colm	manual	0.043
CiTIUS	CiTIUSCrdRelTot	auto	0.035
CiTIUS	CiTIUSSimTot	manual	0.033
vohcolab	vohEvDivTfidf	manual	0.033
NLM	NLM_BNU_E_NLI_C	auto	0.031
CiTIUS	CiTIUSCrdTot	auto	0.011
THUIR	THUIRRuleBased	manual	0.000

Table 4: Recall task results: R-precision.