MRG_UWaterloo and WaterlooCormack Participation in the TREC 2017 Common Core Track

Maura R. Grossman and Gordon V. Cormack
University of Waterloo

MRG_UWaterloo Overview

The MRG_UWaterloo group from the University of Waterloo used a Continuous Active Learning ("CAL") approach [1] to identify and manually review a substantial fraction of the relevant documents for each of the 250 Common Core topics. Our primary goal was to create, with less effort, a set of relevance assessments ("qrels") comparable to the official Common Core Track qrels (*cf.* [12, 2]). To this end, we adapted for live, human-in-the-loop use, the AutoTAR CAL implementation,[1] which had demonstrated superior effectiveness as a baseline for the TREC 2015 and 2016 Total Recall Tracks [9, 5]. In total, for 250 topics, the authors spent 64.1 hours assessing 42,587 documents (on average, 15.4 mins/topic; 5.4 secs/doc), judging 30,124 of them to be relevant (70.7%).

While the principal outcome of the MRG_UWaterloo effort was a set of relevant documents for each topic, it was necessary to submit ranked lists of 10,000 documents for each topic, to be evaluated using the standard rank-based measures calculated by "trec_eval."[2] In theory, according to the probability ranking principle, the optimal strategy to maximize these measures is to construct a ranked list of the 10,000 most-likely relevant documents, with the documents ordered by their likelihood of relevance. In practice, the official qrels used for TREC evaluation are influenced by the submitted runs, confounding the theoretical optimal strategy. Participants were asked to prioritize their runs, and each participant was assured only that the ten highest-ranked documents from their highest-priority submission would be assessed for relevance, and included in the qrels. An unspecified number of additional highly-ranked documents were also to be included, depending on the results of assessing the higher-ranked documents, relative to the results of other participants' runs (*cf.* [6]).

Overall, the qrels for each topic represent a non-statistical sample of the document population, biased heavily toward documents that one or more runs deemed to have a high likelihood of relevance. To estimate the precision of our alternate qrels according to the TREC assessors, we applied a random permutation to the documents we assessed as relevant. These documents, in the order determined by the random permutation, were afforded the highest ranks in our highest-priority run ("**MRGrandrel**"), thus assuring that a random sample (*i.e.*, the first ten) would be assessed by TREC. The remaining documents were scored using the final AutoTAR model, and ranked from highest to lowest score.

Our secondary and tertiary runs ("**MRGrankel**" and "**MRGrankall**") were ordered slightly differently. The ranked lists in MRGrankrel consisted of all documents that we assessed as relevant, ordered by score; followed by the top-scoring documents that we did not assess or assessed as non-relevant, ordered by score. The ranked lists in MRGrankall consisted of the top-scoring documents, ordered by score, regardless of whether or not we had assessed them as relevant.

WaterlooCormack Overview

The WaterlooCormack submission consisted of "automatic routing" runs, as defined in TRECs 1 through 8 [7]. Document rankings were derived using logistic regression on prior relevance assessments for the same topics (but with respect to different corpora), without manual intervention. Feature engineering, learning software, and parameter settings were identical to those used in the TREC 2015 and 2016 Total Recall Tracks [9, 5], and identical to those used by the MRG_UWaterloo group.

Our highest-priority submission, "**WCrobust04**," used the TREC 2004 Robust test collection [13], which used the same 250 topics (with slightly revised narratives), for training. We formed the union of the TREC 2004 and

[1] *See* http://cormack.uwaterloo.ca/trecvm/.
[2] *See* http://trec.nist.gov/trec_eval/.

Topics:	50 NIST			33 NIST & Robust '05		
Measure:	MAP	P@10	Relret@1000	MAP	P@10	Relret@1000
MRGrandrel	0.3190	0.5660	6001	0.2752	0.5545	4425
MRGrankall	0.3538	0.6420	6010	0.3126	0.6394	4437
MRGrankrel	0.3609	0.6500	6029	0.3177	0.6455	4453
WCrobust04	0.3711	0.6460	6396	0.3462	0.6212	4779
WCrobust0405	**0.4278**	**0.7500**	**6785**	**0.4307**	**0.7788**	**5161**
WCrobust04W	0.3656	0.6580	6295	0.3405	0.6424	4687

Tab. 1: Ranked-retrieval measures based on official NIST assessments for 50 topics, and for the 33 NIST topics that were also used in the TREC 2005 Robust Track.

Topics:	50 NIST			33 NIST & Robust '05		
Measure:	MAP	P@10	Relret@1000	MAP	P@10	Relret@1000
MRGrandrel	*0.9927*	*0.9660*	*8537*	*0.9890*	*0.9667*	*6798*
MRGrankall	*0.9118*	*0.9500*	*8418*	*0.9092*	*0.9515*	*6693*
MRGrankrel	*0.9927*	*0.9660*	*8537*	*0.9890*	*0.9667*	*6798*
WCrobust04	0.2322	0.4400	4890	0.1914	0.3545	3570
WCrobust0405	**0.2570**	**0.5040**	**5258**	**0.2291**	**0.4485**	**3934**
WCrobust04W	0.2319	0.4400	4822	0.1923	0.3606	3505

Tab. 2: Ranked-retrieval measures based on MRG_UWaterloo assessments for 50 topics, and for the 33 NIST topics that were also used in the TREC 2005 Robust Track.

Common Core 2017 corpora, from which tf-idf word-based features were derived. Sofia-ML[3] was used to construct a logistic regression model from the TREC 2004 qrels, and the model was used to score the documents in the Common Core corpus. For each topic, the 10,000 highest-scoring documents were submitted, in decreasing order by score.

Our second-priority submission, "**WCrobust0405**," used the same TREC 2004 Robust Track assessments for training, augmented by assessments from the TREC 2005 Robust Track [14], which used 50 of the 250 topics, and yet another corpus. For these 50 topics, we formed the union of the three corpora from TREC 2004, TREC 2005, and Common Core 2017. We trained the model using the TREC 2004 and TREC 2005 assessments, and used the model to score the documents in the Common Core corpus. For these 50 topics, the WCrobust0405 submission consisted of the 10,000 highest-scoring documents, in decreasing order by score. For the remaining 200 Common Core topics that were not used in the TREC 2005 Robust track, the WCrobust0405 submission was identical to WCrobust04.

Our third-priority submission, "**WCrobust04W**," used the TREC 2004 Robust Track assessments for training, augmented by alternate assessments created for the TREC 6 Ad Hoc task [11], by the University of Waterloo, using Interactive Search and Judging (ISJ) [2]. TREC 6 used a subset of 50 of the 250 Common Core topics (a different subset from the 2005 Robust Track). For each of the 50 topics, we formed the union of the official TREC qrels and the Waterloo qrels; in the union, we labeled a document "relevant" if it was labeled relevant in either the TREC or Waterloo qrels, and otherwise "non-relevant." For these 50 topics, the WCrobust04W submission consisted of the 10,000 highest-scoring documents, in decreasing order by score. For the remaining 200 Common Core topics that were not used in the TREC 6 Ad Hoc task, the WCrobust04W submission was identical to WCrobust04.

Ranked-Retrieval Measures Using NIST Assessments

At the time of this writing, official relevance assessments were available only for 50 of the 250 Common Core topics (*i.e.*, the NIST subset). A dynamic "bandit" strategy [6] was employed by NIST to select documents for manual assessment by contract reviewers, starting with the ten highest-ranked documents from the highest-priority runs identified by Track participants. A total of 30,030 documents were assessed, of which 3,453 were judged "highly relevant," 5,549 were judged "relevant," and 21,028 were judged "not relevant." In the following analysis, we consider binary relevance in which the 9,002 documents judged "highly relevant" or "relevant" are considered relevant, and documents judged "not relevant," as well as unjudged documents, are considered non-relevant. In accordance with the methodology employed by the Common Core Track organizers, we consider only the highest-ranked 1,000

[3] *See* https://github.com/glycerine/sofia-ml.

		NIST				
		Rel	Nonrel	Unjudged	$\widehat{Rel}$	$\widehat{NonRel}$
Waterloo	Rel	3715	1966	3305	5524	3462
	Nonrel	213	1016	1610	-	-
	Unjudged	5074	18046	92.7M	-	-

Tab. 3: Waterloo versus NIST Agreement.

documents for each topic.

Table 1 reports the average over the 50 NIST-assessed topics of average precision (MAP), precision at rank 10 (P@10), and number of relevant documents retrieved returned at rank 1,000 (Relret@1000). Table 1 also reports the same measures, averaged over the 33 NIST topics that were also used in the TREC 2005 Robust Track. The WaterlooCormack automatic routing runs scored higher than the MRG_UWaterloo manual ad hoc runs according to the measures reported here, and substantially all the measures reported by trec_eval. Of the WaterlooCormack runs, WCrobust0405 achieved the highest score according to all measures.

Ranked-Retrieval Measures Using Waterloo Relevance Assessments

As noted in the overview above, the primary goal of the MRG_UWaterloo team was to assess substantially all of the relevant documents for each topic. To this end, the team assessed a total of 40,400 documents for relevance to the 250 Common Core topics, marking 30,122 relevant. Of the 40,400 assessments, 11,825 pertained to the 50 NIST topics, of which 8,986 were judged relevant by NIST. 9,285 of the 11,825 assessments also pertained to the Robust 2005 topics, of which 7,247 were judged relevant by NIST.

Table 2 shows the same measures as Table 1, using the MRG_UWaterloo relevance assessments. The first three rows are included for completeness, but should be discounted, as they amount to a circular assessment of the MRG_UWaterloo runs. The last three rows, on the other hand, represent the WaterlooCormack runs, which were not influenced by and did not influence the MRG_UWaterloo assessments. Although the numerical results are substantially lower, the MRG_UWaterloo-assessed scores yield the same relative result as the NIST scores: WCrobust0405 is superior by all measures.

Set-Based Measures

The aim of the MRG_UWaterloo effort, as well as the aim of the NIST assessment effort, was to identify, as effectively as possible, a complete set of documents relevant to each topic. If ground-truth relevance with respect to every topic and every document were known, we could evaluate the effectiveness of these two efforts using standard set-based measures, the most prominent of which are recall and precision. In general, however, ground truth can never be known with certainty, because the determination of relevance is based on human judgment, which is not entirely reliable. Even if the relevance determination of a competent assessor were deemed to be "reliable enough," it is feasible to render such a determination for only a small subset of the documents in the collection.

For the 50 NIST topics, the 11,825 Waterloo assessments and the 30,030 NIST assessments represent 0.01% and 0.03%, respectively, of the 92.7M assessments that would be required to render a human determination of relevance for each document in the test collection with respect to each topic. Of the 8,986 documents that the Waterloo assessors deemed relevant and the 9,002 documents that the NIST assessors deemed relevant, only 3,715 were deemed relevant by both efforts. In other words, the overlap (*i.e.*, Jaccard index) between the Waterloo and NIST assessments was 0.26.

If we deem the NIST assessments to be ground truth, Waterloo achieved (micro-averaged) recall of 0.4127 (3,715/9,002), and precision of 0.4134 (3,175/8,986). Conversely, if we deem the Waterloo assessments to be ground truth, NIST achieved (micro-averaged) recall of 0.4134 and precision of 0.4127. In either case, $F_1 = 0.4131$.

Macro-averaged recall, precision, and F_1 of Waterloo according to NIST (*i.e.*, precision, recall, and F_1 of NIST according to Waterloo) are 0.4193, 0.5091, and 0.3952, respectively.

Incomplete Assessments Versus Assessor Disagreement

Table 3 gives a 3×3 confusion matrix showing the agreement between the Waterloo and NIST assessments. It is important to note that the document selection and assessment methods are not independent. For the Waterloo

assessments, the user's relevance determinations were used to train a learning method that chose further documents to present to the user. The selection of documents for assessment by NIST was, at first, determined by the submitted runs (including those derived from the Waterloo assessments), and subsequently, by the relevance determinations of the NIST assessors. While it would be misleading to draw statistical inferences from the raw numbers, it is apparent that there are several sources of discord between the Waterloo and NIST assessment efforts.

The columns headed $\widehat{Rel}$ and $\widehat{NonRel}$ give statistical estimates of the number of documents that Waterloo judged relevant that would have been judged relevant or non-relevant by the NIST assessor, had they all been assessed. The estimate is based on the uniform sample of ten documents placed at the top of the MRGrandrel run, which was fully assessed. The difference between the Rel and $\widehat{Rel}$ is substantial, and significant ($P \approx 0.000$).

Table 4 shows the *Rel* and $\widehat{Rel}$ statistics for every topic, ordered by their difference. As can be seen from the first several rows of the table, most of the total difference between *Rel* and $\widehat{Rel}$ is accounted for by a few topics for which the number of judged documents is substantially smaller than the number of documents deemed relevant by Waterloo. Large differences are generally associated with small P-values, indicating that the large negative differences at the top of the table are unlikely attributable to chance, and reflect incompleteness in the NIST assessments.

With two notable exceptions – Topics 307 and 427 at the bottom of the table – we find that most or all of the documents we deemed relevant were judged by NIST, and that *Rel* is consistent with $\widehat{Rel}$. For both of these exceptional topics, *Rel* is substantially and significantly greater than $\widehat{Rel}$. This result is sufficiently improbable ($P \approx 0.002$ in both cases) that there must be some correlation between the sample and the NIST assessments. The sampling method (*i.e.*, drawing the first ten elements returned by the Linux utility "sort -R") is uniform and (pseudo) random, and therefore unlikely to be a causal factor.

We do know that documents in the sample, being at the top of the rankings from our highest-priority run, were assessed at the outset by the NIST assessors. Some of the other documents might also have been judged at the outset, while others were probably judged later. That is, documents in the sample were more likely to be judged at the outset than the remaining documents. For Topics 307 and 427, our results suggest that the assessors were less likely to judge the documents relevant at the outset. A possible explanation for this is that the pool of initially judged documents contained a higher proportion of relevant documents, resulting in stricter relevance assessments, as has been observed in prior work [8, 10].

Topic 307 concerned new hydroelectric projects. Three of the ten documents in our sample were coded relevant. Of the remaining seven that were marked non-relevant, five mentioned the Three Gorges hydroelectric project, though sometimes only by the name "Three Gorges Dam." We conjecture that, at the outset, the assessor judged documents mentioning "Three Gorges" to be non-relevant, but later in the process marked similar documents relevant.

Topic 427 concerned eye damage from ultraviolet exposure. Three of the ten documents in our sample were coded relevant. Of the remaining seven, five mentioned cataracts and ozone-layer depletion. We conjecture that, at the outset, the assessor coded documents about the ozone layer and cataracts as non-relevant, but later in the process marked similar documents relevant.

Discussion

We were surprised that the WaterlooCormack runs outperformed the MRG_UWaterloo runs according to ranked-retrieval measures, as we had expected that the use of corpus-specific feedback would be more effective than transfer learning from a different corpus. In accordance with our goal of achieving high recall, we spent a disproportionate amount of time assessing documents for the larger topics, which yielded no benefit – and may even have been counterproductive in terms of maximizing trec_eval scores – given the apparent incompleteness of the NIST assessments for those topics. Also, in accordance with our goal, we aimed to construe relevance broadly, but this does not fully explain the level of disagreement between the Waterloo and NIST assessments, which bears further investigation.

The problem of "concept drift" in using dynamic document-selection methods for assessment is worthy of investigation. If the order of assessment affects those assessments, how should one try to avoid any bias that might arise? The literature suggests that bias may disfavor systems that find novel relevant documents early, or that rank relevant, but perhaps not obviously relevant, documents first [3, 4].

It remains to be seen whether the disagreement between the WaterlooCormack and NIST assessments matters, as far as their use as ground truth for evaluating other IR systems. Waterloo's alternate assessments for TREC 6 [2, Fig. 1] had an overlap of 0.25 with the official NIST assessments (compared to 0.26 for the current effort), but showed comparable effectiveness for the purpose of evaluating the other TREC 6 submissions [2, 12].

A third group from Waterloo (UWaterlooMDS [15]), using a similar technique to MRG_UWaterloo, derived an

	UW Assessments		NIST Assessments			
Topic	*Rel*	*Judged*	*Rel*	$\widehat{Rel}$	$Rel - \widehat{Rel}$	P-value
325	1281	441	298	897	-599	0.0043
436	797	303	277	797	-520	0.0000
354	901	401	297	541	-244	0.1434
372	1168	371	293	467	-174	0.4508
422	518	362	246	414	-168	0.0754
439	254	128	60	178	-118	0.0040
408	255	218	111	153	-42	0.4537
626	143	138	95	129	-34	0.1862
350	113	89	53	79	-26	0.2296
426	318	260	231	254	-23	0.9080
443	69	29	18	41	-23	0.0315
399	92	80	71	92	-21	0.1281
375	330	217	112	132	-20	0.9150
330	59	59	26	41	-15	0.1440
353	81	81	38	49	-11	0.5836
416	141	136	74	85	-11	0.8734
336	43	43	25	34	-9	0.2144
394	130	128	97	104	-7	1.0000
362	114	114	85	91	-6	1.0000
400	31	31	20	25	-5	0.4041
419	29	29	12	17	-5	0.2805
423	131	130	126	131	-5	1.0000
321	67	67	56	60	-4	0.9592
367	85	85	31	34	-3	1.0000
389	36	36	29	32	-3	0.7092
435	51	39	2	5	-3	0.7137
442	83	83	55	58	-3	1.0000
620	133	129	50	53	-3	1.0000
646	76	76	59	61	-2	1.0000
677	27	27	14	16	-2	0.8037
347	16	16	4	5	-1	1.0000
355	56	54	10	11	-1	1.0000
404	22	21	3	4	-1	0.8571
614	47	47	46	47	-1	1.0000
341	54	54	16	16	-0	1.0000
344	5	5	1	1	0	1.0000
356	4	4	2	2	0	1.0000
393	131	131	79	79	0	1.0000
433	4	4	3	3	0	1.0000
445	160	110	32	32	0	1.0000
414	41	41	23	20	2	0.9299
345	38	38	30	27	3	0.6953
397	118	118	62	59	3	1.0000
378	103	90	66	62	4	1.0000
690	36	36	14	7	7	0.2888
310	164	153	60	49	11	0.9372
379	73	73	21	7	14	0.2996
363	77	77	62	46	16	0.1940
427	104	104	80	31	49	0.0024
307	177	175	140	53	87	0.0016

Tab. 4: Difference between the actual number of NIST-assessed relevant documents and the estimate from the statistical sample, ordered by difference.

alternate set qrels that had considerably lower recall, and slightly higher precision, than the MRG_UWaterloo set, but scored much higher according to the official NIST ranked-retrieval measures, including Relret@1000.

References

[1] Gordon V Cormack and Maura R Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv:1504.06868*, 2015.

[2] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In *SIGIR 1998*.

[3] Tadele T Damessie, Falk Scholer, Kalvero Järvelin, and J Shane Culpepper. The effect of document order and topic difficulty on assessor agreement. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval*, pages 73–76. ACM, 2016.

[4] Michael Eisenberg and Carol Barry. Order effects: a study of the possible influence of presentation order on user judgements of document relevance. *Journal of the American Society for Information Science*, 39(5):293, 1988.

[5] Maura R Grossman, Gordon V Cormack, and Adam Roegiest. TREC 2016 Total Recall Track Overview. In *TREC 2016*.

[6] David E Losada, Javier Parapar, and Álvaro Barreiro. Feeling lucky?: Multi-armed bandits for ordering judgements in pooling-based evaluation. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1027–1034. ACM, 2016.

[7] Stephen Robertson and Jamie Callan. Routing and filtering. In Ellen M. Voorhees and Donna K. Harman, editors, *TREC — Experiment and Evaluation in Information Retrieval*, chapter 5, pages 99–122. MIT Press, Cambridge, Massachusetts, 2005.

[8] Adam Roegiest and Gordon V. Cormack. Impact of review-set selection on human assessment for text classification. In *SIGIR 2016*.

[9] Adam Roegiest, Gordon V Cormack, Maura R Grossman, and Charles L A Clarke. TREC 2015 Total Recall Track Overview. In *TREC 2015*.

[10] Mark D Smucker and Chandra Prakash Jethani. Human performance and retrieval precision revisited. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–602. ACM, 2010.

[11] Ellen Voorhees and Donna Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). In *6th Text REtrieval Conference*, Gaithersburg, MD, 1997.

[12] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5), 2000.

[13] Ellen M. Voorhees. Overview of the TREC 2004 Robust Track. In *Proceedings of the 13th Text REtrieval Conference*, Gaithersburg, Maryland, 2004.

[14] Ellen M Voorhees. The TREC 2005 Robust Track. In *ACM SIGIR Forum*, volume 40. ACM, 2006.

[15] Haotian Zhang, Mustafa Abualsaud, Nimesh Ghelani, Angshuman Ghosh, Mark Smucker, Gordon Cormack, and Maura R Grossman. UWaterlooMDS at TREC Core Track 2017 (Notebook). In *TREC 2017*.