RMIT at the 2018 TREC CORE Track

Rodger Benham
RMIT University
Melbourne, Australia

Luke Gallagher
RMIT University
Melbourne, Australia

Joel Mackenzie
RMIT University
Melbourne, Australia

Binsheng Liu
RMIT University
Melbourne, Australia

Xiaolu Lu
RMIT University
Melbourne, Australia

Falk Scholer
RMIT University
Melbourne, Australia

Alistair Moffat
The University of Melbourne
Melbourne, Australia

J. Shane Culpepper
RMIT University
Melbourne, Australia

ABSTRACT

Ad-hoc retrieval is an important problem with many practical applications. It forms the basis of web search, question-answering, and a new generation of virtual assistants being developed by several of the largest software companies in the world. In this report, we continue our exploration of the importance of multiple expressions of information needs. Our thesis is that over-reliance on a single query can lead to suboptimal performance, and that by creating multiple query representations for an information need and combining the relevance signals through fusion and relevance modeling, highly effective systems can be produced. This approach may form the basis for more complex multi-stage retrieval systems in a variety of applications.

TEAM NAME

RMIT

1 INTRODUCTION

The second TREC CORE Track[1] continues the ad-hoc evaluation campaign from 2017, where the aim is to bring the community together to solicit a diverse set of runs and to establish new methodologies for creating test collections. This year, we focused on exploring similar ideas to those we used in the previous CORE track [5], including *manual query variations* and *rank fusion*, with other ideas such as *relevance modeling* and *external resources*. In addition, we again focus on recall-oriented search to build robust runs on the new collection through combining multiple representations of an information need [8].

Inspired by the strong participation of Zhang et al. [17] and their use of relevance feedback, we solicit a shallow pool of document judgments to filter out poorly performing queries prior to rank fusion. We felt that this was a tractable and useful exercise to ensure that relevant documents are at the head of the run, while also improving the likelihood of introducing unjudged relevant documents not found by other participants. In addition, we use the resulting judgment set to select the best query per-topic, and to check if the poorly performing run RMITUQVBestM2 last year was an aberrant result due to using an external collection to decide the "best" query. Another point of inquiry is whether external query expansion combined with multiple systems on the target corpora can further improve retrieval effectiveness.

Bailey et al. [3] observed the retrieval consistency of query variations using the UQV100 collection [2]. The query variations on this collection were included in the judgment pooling process with shallow judgments, as Moffat et al. [15] showed that test collections formed without user query variability do not generalize well outside of the supplied title query. Based on these observations, we also investigate whether the CORE 2018 test collection construction methodology exhibits similar behaviour.

Research Goals. We focus on three research questions:

- **RQ1:** *Can shallow judgments from bronze-assessors be used to further improve double fusion effectiveness by filtering out non-performing queries prior to fusion?*
- **RQ2:** *Can external corpora be combined with multiple information needs in order to produce better results than the original corpora alone?*
- **RQ3:** *How robust is the new collection to multiple query variations representing the same information need?*

In the next section, we discuss how our submitted runs were formed, and in Section 3 we provide the results of our submitted runs using the fifty topics assessed by NIST, and conduct further analysis on these runs.

2 APPROACH

We now describe the various resources used to create the five submitted runs, and how these runs are generated.

Collections. The new Washington Post v2 corpus for the CORE track was parsed using the jq tool[2]. Note that Twitter was also embedded into the original unprocessed collection (json within json). All double embeddings were stripped out of the final trectext SGML formatted text produced by our scripts. Indri 5.11 and Terrier 4.2 were then used to index the resulting collection. For external query expansion, we used the Gigaword and Tipster corpora as originally described by Diaz and Metzler [11], which were also reformatted into trectext SGML format before indexing. External expansion was conducted using a patched version of Indri 5.12.[3]

[1] TREC 2018 CORE Track: http://trec-core.github.io/2018

[2] https://github.com/stedolan/jq

[3] https://github.com/diazf/indri

Details on how this collection was parsed are available in Benham et al. [5].

A peculiarity of this Washington Post corpus is that there are many duplicate documents. To improve recall, we include all duplicate documents in-place at the rank position they are retrieved (that is, duplicates are not suppressed). To determine if a document is a duplicate, we compute the MD5 hash of all document titles in the collection, and if any documents collide with a hash, those documents are considered duplicates. There are cases where this simple approach for identifying duplicates is not effective, such as when the title of the document is *"Traffic Report"*, and the document body is not the same. By selecting a threshold of only allowing in-place insertion of duplicate documents for MD5 collisions of less than 100 matches, and manually verifying that the result of this decision did not include incorrect duplicates into runs, we were satisfied with our approach.

Runs. The runs we generated for the CORE 2018 track are a logical extension of the previous ideas employed in Benham et al. [5]. We have since improved the effectiveness of our automatic runs by leveraging external relevance modeling proposed by Diaz and Metzler [11], which was recently explored by Benham et al. [6]. In their work, the Robust04 collection was treated as the target collection and the Gigaword+Tipster corpus was used as a source for relevance modeling. This time around we followed a similar methodology, but replaced the target collection with the Washington Post corpus. Empirically, we found that using Terrier's query expansion models DFree and DLH13 produced more effective runs than relevance modeling with Indri – a line of experimentation we did not explore previously. We hypothesize that combining these two approaches will be more effective, which helps to address **RQ2**. Fortunately, the second round of the CORE track offered the benefit of tuning query expansion parameters using 5-fold cross-validation on the CORE 2017 New York Times collection. This collection is more similar in composition (temporally) to the new Washington Post collection than Robust04 used in the previous year. Indri was employed to perform a parameter sweep over the New York Times collection. These parameters were also used for a number of query expansion runs with the Terrier platform.

Bailey et al. [3] proposed *double fusion*, where multiple queries for the same information need are issued to multiple systems and merged into a single, high quality SERP. Benham and Culpepper [4] showed that double fusion had the best effectiveness and risk-sensitivity trade-off space using the T_{Risk} measure [12] on the Robust04 and ClueWeb12-B corpora. Incidentally, the authors found that reciprocal rank fusion (RRF), with k fixed to 60 as proposed and recommended by Cormack et al. [7], was a marginally more effective way to perform unsupervised fusion than the rank-biased centroid (RBC) approach proposed by Bailey et al. [3], and this was used for the RMIT CORE runs in 2017. In addition, we did not perform a true double fusion in the last CORE effort – rather, we selected on a per-topic basis whether a sequential dependency model combined with query expansion should be used for the top-5 performing query variations on Robust04, or BM25 instead. Although different systems were used, they were not used in *conjunction* with each other as a source of evidence to form the topic centroid. We use a true double fusion this year to avoid tuning on a per-topic basis.

Table 2: Query variation statistics per user. Uniqueness is calculated with respect to a bag of words, as all retrieval models used are BoW. Participants marked with † were co-authors of the CORE 2017 activity that did not contribute in 2018.

Participant	Queries	Avg. Terms	Avg. Chars	% Unique
1	152	3.84	25.72	80.92%
2	120	5.09	32.68	87.50%
3†	30	3.47	24.60	73.33%
4†	59	4.90	30.56	94.92%
5	337	5.85	37.39	97.92%
6	161	5.22	33.70	94.41%
7	97	5.69	36.86	89.69%
8	322	4.90	30.32	94.72%
9	95	4.83	30.02	89.47%
10	82	6.45	37.88	92.68%
Overall	1455	5.02	31.97	89.56%

Rather than using an external collection to select the best query variation from a pool of candidates written by the authors (as we did in CORE 2017), we instead opted to form our own judgments (explained below). This is due to the unexpected finding that the best query variation from the constrained set evaluated on both collections was only the same for 12 out of 50 topics, suggesting that the *best* formulation of an information need is indeed collection dependent. We also used this judgment set to filter out queries with a zero average precision (AP) score prior to fusion to form what we hypothesize to be our most effective run; forming the run that allows us to address **RQ1**. Table 1 provides a description of each of the runs submitted, while Figure 1 shows a UML representation of how each of the submitted runs was formed.

Generating Query Variations. The approach to generating query variations was a similar process to our 2017 submission Benham et al. [5]. The authors of that paper were invited to contribute up to ten query variations per-topic. As the NIST assessed topics include 25 of the old topics which we had previously collected query variations for, we needed to gather query variations for the 25 new topics only. And since the CORE track has had a five-fold reduction in the number of topics, we were able to collect more data per-topic than in previous experiments.

Users were given spell-check suggestions for the queries they submit, using the Bing Spell Check API. All queries were case-folded and Krovetz stemmed, consistent with our participation in 2017 [5]. By the end of the collection stage, 1,455 variations were collected for the 50 topics. Table 2 shows the contributions made by each participant. The query variants are slightly shorter compared to 2017 in terms of the average number of terms (5.48 to 5.02) and the average number of characters (33.9 to 32.0). We avoid comparing the ratio of unique variations as we are only using bag of words models, rather than proximity models like last year. This data curation exercise helps to answer **RQ1** and **RQ3**.

Table 1: Description of the submitted runs to the 2018 TREC CORE track.

Run	Type	Description
RMITUQVDBFNZDM1	Manual	Authors formed a judgment pool to the top-5 of RMITUQVDBFDM3 and a title query language model run. These judgments were used to remove any query variations with a zero score prior to rank fusion. The reduced set of queries using RMITUQVDBFDM3, where documents found to have duplicates were included in-place in the ranked list.
RMITUQVDBFDM3	Manual	Query variations for the original TREC topics were generated by the authors. All query variations were run on systems with parameters shown to be effective on NYT using Indri and Terrier with query expansion, as well as external corpus query expansion using Gigaword and]Gigaword+Tipster. This was fused to make a single run using RRF k=60. Documents found with duplicates were included in-place in the ranked list.
RMITUQVBestDM2	Manual	Authors formed a judgment pool to the top-5 of RMITUQVDBFDM3 and a title query language model run. These judgments were used to select the best title-only query without fusion using the same systems as in RMITUQVDBFDM3.
RMITFDA4	Automatic	Title query runs on Indri and Terrier with query expansion, and external expansion runs from Gigaword and Tipster fused into a single run using RRF. Query expansion parameters taken from NYT judgments (collection-wide, not per-topic). A baseline for how query variations compare to titles. Duplicate documents were included in-place in the ranked list.
RMITEXTGIGADA5	Automatic	External query expansion using the Gigaword+Tipster corpus.

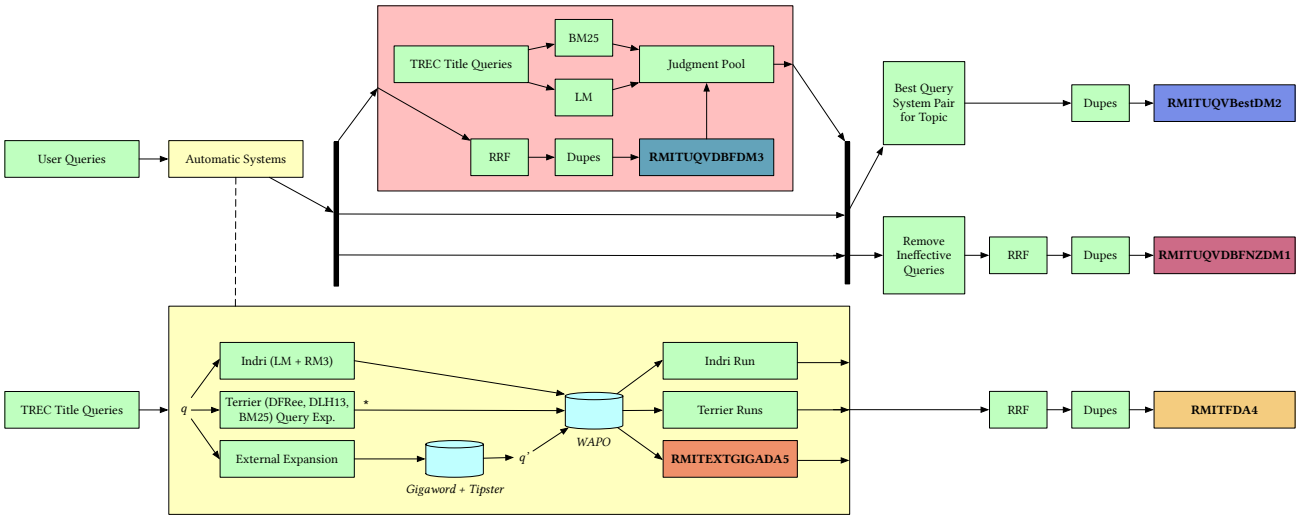

Figure 1: Flowchart representation of the runs submitted.

Judgments. In Benham et al. [5], we found that fusion of many queries yielded a more effective and robust SERP. This is in contrast to selecting the best query from a constrained set of high-quality candidates by using an external collection with relevance judgments. A preliminary investigation into the "best" query from this set shows that the ordering is highly-specific to the collection; even on corpora of similar content. This year instead of using the judgments from an external collection, we form a shallow judgment pool to a minimum depth of 5 documents per-topic, and a maximum of 15. The judgment pool was formed using the title queries supplied by NIST over BM25 and Language Modeling (LM) using Indri, as well as our submitted run RMITUQVDBFDM3 described in Table 1. On average each topic had a pool-depth of 10.40, compared to the NIST assessment average pool depth per-topic of 524.66. Figure 2 shows a screenshot of the judgment solicitation interface authors used for the assessment exercise.

Table 3 shows statistics on the judgments collected with the average document length per-assessor, and a post-hoc analysis of intra-assessor agreement with the NIST QREL set. We compute Krippendorff's α coefficient introduced by Hayes and Krippendorff

Table 3: Document judgment statistics per user, with Krippendorff α computed with respect to the NIST QREL set.

			Ratings				
Participant	Judgments	Avg. Terms / Doc	Irrelevant	Somewhat Relevant	Fully Relevant	Unique Judgments	α Agreement
5	507	348.83	189	73	245	391	0.529
9	60	474.30	31	18	11	28	0.431
8	30	369.70	19	9	2	17	0.253
10	19	454.74	14	2	3	11	−0.126
Overall	616	365.34	253	102	261	447	0.507

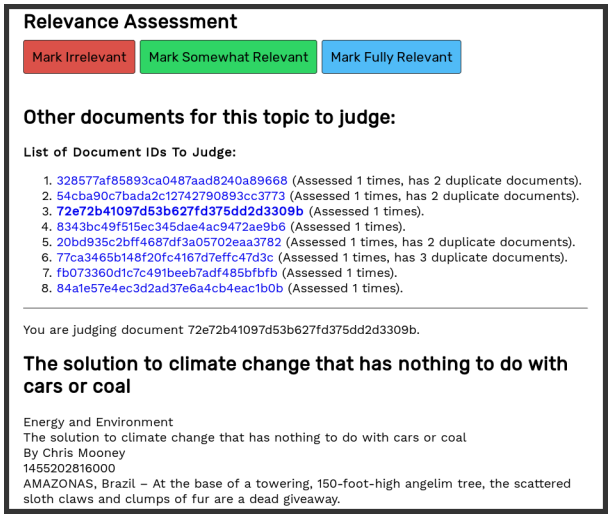

Figure 2: Screenshot of the document relevance assessment solicitation interface.

[13] to quantify this agreement with respect to the nominal dichotomous categories we define as: *Not Relevant*, *Somewhat Relevant* and *Fully Relevant*. Where multiple assessors judged the same document, the median score was taken as the true assessment, where the median could correspond to one of the three categories mentioned above. Overall we find that our relevance assessments are different to NIST with a Krippendorff's α of 0.507. If we binarize both sets of judgments to collapse the categories *Somewhat Relevant* and *Fully Relevant* to become *Relevant*, Krippendorff's alpha remains relatively unchanged with a value of 0.504, with a percentage agreement of 76.08%. As document assessment is a subjective exercise, disagreement is likely to occur. For example, on the TREC topic 336 titled *Black Bear Attacks*, two assessors for the document identifier `d6ed7028c686e5756ceb0aa0c9b62e0d` found the document to be *Not relevant* as it is about a personal account of a black bear attack, and does not discuss the frequency or possible causes for a black bear attack – however it is marked as *Fully Relevant* in the NIST QREL set. In any case, our goal for forming a judgment set is to form a general guide for inclusion of queries into a fusion pool and is not to replicate the decisions made by NIST assessors. We later show that despite the high disagreement, our use of the judgment set in the generation of RMITUQVDBFNZDM1 and RMITUQVBestDM2 runs was justified as retrieval effectiveness improved.

Given that the assessors are "bronze" judges while the NIST assessors are presumed to be "gold" assessors [1], further explorations in presentation ordering [10] or gathering multiple judgments through crowdsourcing [9] might result in higher agreement with the NIST assessors, and improve performance further.

3 RESULTS

Once again, all of our runs met the effectiveness requirements of the track organizers, meaning that all five runs from RMIT contributed to the judgment pool. We now outline a basic analysis of these systems for completeness.

Baseline Configuration. As a point of reference, we report three additional runs in the main results reported in Table 4: Title is a BM25 run on the query title, RRF is the RRF fusion of all unique variations for among each topic, and BestUQV is the top-performing single UQV from each topic. We use RRF as our basis for significance testing, as it represents a strong yet simple baseline given a set of UQVs.

Comparing Submitted Runs. Figure 3 shows the effectiveness of our submitted runs across a number of commonly used metrics including AP, NDCG@k, and RBP with persistence ϕ. In Figure 3 it is interesting to observe that of the manual runs, the best system depends on the metric and evaluation depth. While focusing on Table 4, the two automatic runs, RMITEXTGIGADA5 and RMITFDA4 were unsurprisingly outperformed by the manual runs across all metrics. In particular, the RRF baseline was statistically significantly better than RMITEXTGIGADA5 across a number of the tested metrics. On the other hand, no significance was found between RRF and any other submitted run. Interestingly, the oracle run BestUQV significantly outperformed the RRF baseline for all metrics, demonstrating the importance of how information needs are formulated. As RMITFDA4 is more effective than RMITEXTGIGADA5, and includes RMITEXTGIGADA5 in its fused run (see Figure 1) and passes a pairwise t-test over AP, we answer **RQ2** in the affirmative.

Table 5 shows the tournament matrix of wins, ties, and losses when comparing the runs head-to-head. These outcomes are consistent with the effectiveness comparison, showing at the topic level how the manual runs consistently outperform the automatic runs. RMITUQVDBFNZDM1 is the best of the five runs but it is only slightly superior to RMITUQVDBFDM3, with 32 out of the 50 topic scores within 10% of each other (the definition of "tie" used

Table 4: Comparing the submitted runs with additional runs as reference. Pairwise t-tests were conducted using a Bonferroni correction against the RRF run, with † and ‡ representing significance at $p < 0.05$ and $p < 0.01$, respectively.

System	AP	NDCG@k		RBP ϕ	
		10	20	0.80	0.95
Title	0.227‡	0.394‡	0.379‡	0.442‡	0.320‡
RRF	0.355	0.548	0.527	0.611	0.452
BestUQV	0.417‡	0.685‡	0.632‡	0.733‡	0.526‡
RMITEXTGIGADA5	0.258‡	0.424†	0.388‡	0.464‡	0.351†
RMITFDA4	0.311	0.473	0.454	0.520	0.404
RMITUQVBestDM2	0.318	0.541	0.505	0.597	0.423
RMITUQVDBFDM3	0.375	0.533	0.522	0.584	0.464
RMITUQVDBFNZDM1	0.385	0.557	0.538	0.614	0.481

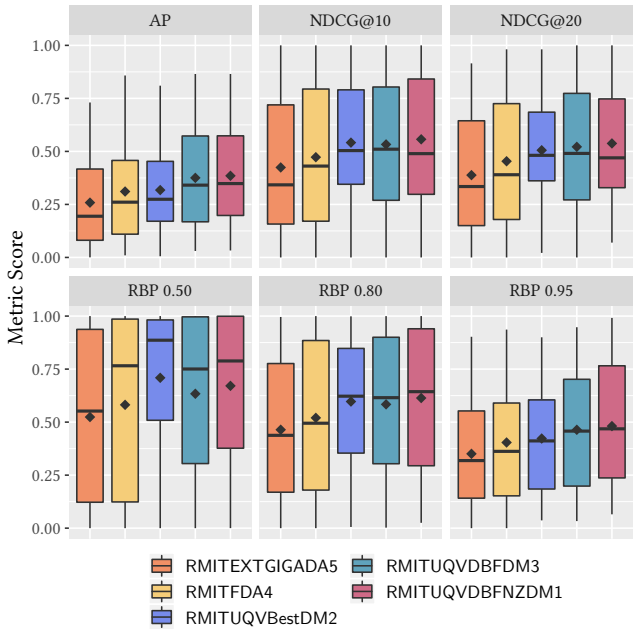

Figure 3: Comparing the submitted runs across a range of effectiveness metrics. Diamonds denote the mean effectiveness value for each system.

here). The other pairwise comparisons show larger gaps in the win and loss numbers, which is an indication of their performance difference.

All three manual runs show similar performance, but RMITUQVDBFNZDM1 is more effective than the other two. In both Table 5 and Table 4, run RMITUQVDBFNZDM1 and RMITUQVDBFDM3 are particularly of interest. The RMITUQVDBFDM3 run is a double fusion run over all query variants and several retrieval systems. As described in Table 1, RMITUQVDBFNZDM1 is also a fusion run built from RMITUQVDBFDM3, with the worst performance query variants removed, based on judgments pooled using RMITUQVDBFDM3 and an LM run at depth $d = 5$. From Figure 3, we can observe that the two runs show similar performance when evaluated using deep metrics such as AP or RBP 0.95, but a larger performance gap can be observed when evaluated with shallow metrics such as NDCG@20 and RBP 0.5. The only exception is exhibited when considering NDCG@10, where the median RMITUQVDBFDM3 score is worse than RMITUQVDBFNZDM1. Despite the greater effectiveness of RMITUQVDBFNZDM1 compared to RMITUQVDBFDM3, a pairwise t-test of their AP scores found no statistical significance, and therefore we cannot answer **RQ1** in the affirmative.

3.1 Consistency of Query Variations

Figure 4 shows, on a per-topic basis, the variance of the AP score for each submitted query variation using a simple bag-of-words ranking model (BM25). It also shows the performance of the provided title query as a diamond, using the same BM25 configuration. The left-most topics are consistently difficult, with no submitted variations (nor the title query) performing well. On the other hand, there appears to be less consistency among easy topics, as the IQR generally seems to increase with the mean AP of the topic.

Inconsistency Analysis. Looking closer, we can observe inconsistencies in the UQVs. For example, consider the three topics with the highest IQR from Figure 4:

- 804: *"women on 20s"*
- 806: *"computers and paralyzed people"*
- 811: *"car hacking"*

Topic 804 had a poor-performing title query, with an AP of 0.021 for a simple BM25 ranking. Furthermore, the highest scoring 18 query variations all contained the name *Harriet Tubman*, with 10 of these variations also including the name *Andrew Jackson*. The AP scores of these top 18 variations ranged from 0.578 to 0.834, a stark contrast to the AP score of the title query.

Topic 806 had a title query with an AP of 0.400, corresponding to the median-performing topic out of the 33 variations. For this topic, the query terms that perform well are less clear-cut, with no real trends observed in the top performing variations. Interestingly, the fourth best variation, with an AP of 0.562, did not mention *computers* at all (the query was *exoskeleton paralyzed paralysis movement*). This serves to demonstrate the high variance in query formulation, and the unpredictable behaviour that can occur for isolated query variations.

The title query for topic 811 outperformed all submitted query variations that we solicited, with an AP of 0.625. Even queries that seemingly add a slight perturbation such as *car computer hacking* or *car hacking tools* greatly reduced the performance, resulting in AP scores of 0.360 and 0.118, respectively.

While the oracle run BestUQV shows the potential upside to selecting a single yet high performing query variant on a per-topic basis, further analysis shows that UQVs exhibit high variance in their individual performance. This demonstrates why rank fusion is preferred for robustness and consistency [3, 8].

Table 5: Comparing wins, ties and losses in terms of AP score for the run in the column header, against the run listed in each row, with a 10% difference taken to be the upper threshold for a "tie".

	RMITFDA4	RMITUQVBestDM2	RMITUQVDBFDM3	RMITUQVDBFNZDM1
RMITEXTGIGADA5	25/18/7	28/11/11	36/11/3	37/10/3
RMITFDA4	–	24/10/16	22/20/8	27/17/6
RMITUQVBestDM2	–	–	25/14/11	27/17/6
RMITUQVDBFDM3	–	–	–	13/32/5

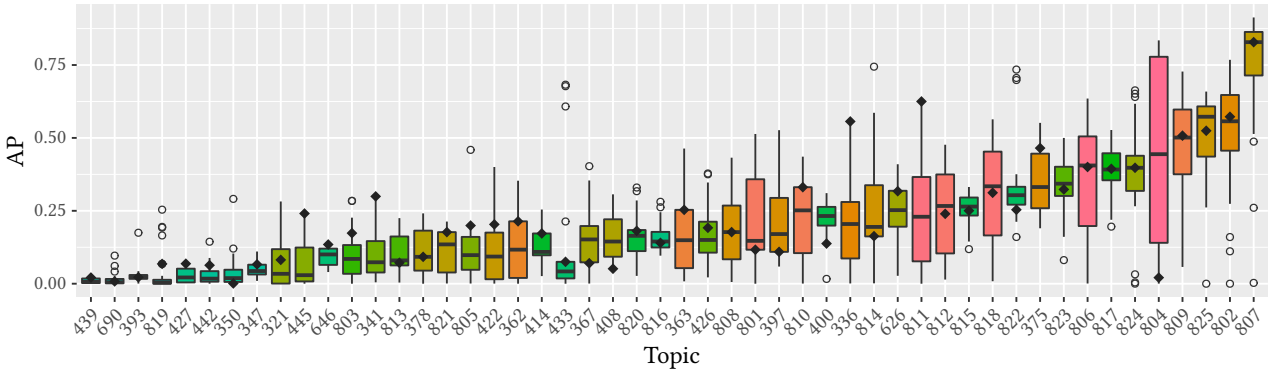

Figure 4: The per-topic BM25 based on a bag-of-words BM25 run for every query variation for each topic, sorted by *mean* AP. Clearly, some topics were consistently difficult, and others exhibit high variance depending on the query variation that is processed. Diamonds represent the TREC *title query* for each topic.

3.2 Do Query Variations Generalize?

The metric RBP [14] is by design a lower bound estimate on the retrieval effectiveness of a system. It provides a residual score that indicates the amount of unjudged documents present in an evaluation. A high residual indicates uncertainty – we simply do not know whether the unjudged documents are relevant or not. Figure 5 depicts the residuals from BM25 for RBP with $\phi = 0.95$ across all UQVs for each of the 50 topics. In contrast to Figure 4, topics that have more effective scores and less variance should overall have a lower residual score.

Consider topic 825, overall the residual and variance is low for the submitted UQVs. However, there are a number of outliers for topic 825 that can be seen in Figure 5. The background description for topic 825 is "*Does diversion of U.S. corn crops into ethanol for fuel increase food prices?*", and looking at Table 6, even the poor performing queries mention "food", "corn", "price" or "ethanol" with one exception. These appear to be important terms for the BM25 retrieval model, however, looking at the actual query variations in Table 6, there appears to be a sense of participant initiated query drift. The human element within an IR task is the strongest and weakest link – a lack of knowledge, fatigue or a momentary distraction are all viable cases for outliers. Among the outliers listed in Table 6 we see that as the RBP score improves, the residual becomes lower indicating greater confidence because there are more judged documents examined. Within the table, the MED-RBP scores (see Tan and Clarke [16]) indicate how different two ranked lists are when compared under RBP $\phi = 0.95$. This can give insight as to why things may be different for certain variants. Take, for example, variant 825-4-3 and 825-3-6. The RBP scores are essentially the same, however, the residuals differ, and 825-3-5 has more relevant information than 825-4-3. This results in an improved MED-RBP score and indicates that perhaps being more certain about what is not relevant is equally important for capturing a user's information need.

It is difficult to ascertain whether the same information need expressed in different ways is able to generalize across a collection. It certainly does work in some cases, as shown in Figure 4, with supporting evidence in Figure 5. However, there are other cases where variants that should retrieve a sensible SERP for a topic are falsely evaluated as performing poorly due to high residuals. One take on this is that query variants for the same information need may have different objectives. The variant 825-4-3 is suggestive of an open-domain style question-answer type of query, and while the information need may be similar, the level of interpretation required by the system is different. Despite this, Figure 4 and Figure 5 clearly show that there is contrasting levels of uncertainty over query variations per-topic with a fixed BM25 retrieval model. The diamond in Figure 5 shows the residual uncertainty of the title queries that contributed to the pool, where most of these residuals are below the first quartile compared query variant residuals. We find that the collection forms robust answer set to the supplied TREC title queries, however, this does not hold true for query variations, answering **RQ3**. It would be interesting in future work to explore the effect of query variants across different retrieval models.

Table 6: Topic 825. User query variation outlier analysis for RBP $\phi = 0.95$ of the variants with a relatively high residual when compared to other variants within the same topic. The difference MED-RBP $\phi = 0.95$ is computed between each variant and the original title query.

ID	Query	RBP	Residual	MED-RBP
825	*ethanol and food prices*	0.555	0.022	–
825-3-10	*corny cold war*	0.000	0.961	0.577
825-4-3	*current status of growing corn with the intention of using it for ethanol fuel impact on food price*	0.401	0.386	0.232
825-3-6	*diversion of corn to ethanol usa*	0.400	0.260	0.177
825-2-4	*impact on food prices corn into fuel*	0.470	0.197	0.112
825-5-2	*diversion united states corn crops ethanol starvation hunger in poor communities*	0.475	0.164	0.103

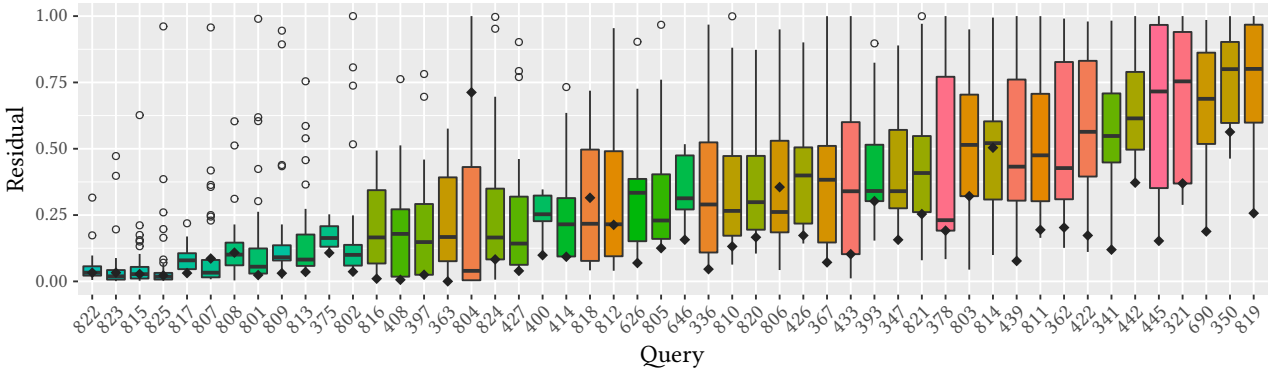

Figure 5: The per-topic residual based on a bag-of-words BM25 run across every query variation for each topic, sorted by the *mean* residual of RBP 0.95. High residuals imply that many retrieved documents were not judged. A high variance in residuals implies that some UQVs surface many unjudged documents, whereas others surface mostly judged documents. Diamonds represent the TREC *title query* for each topic.

4 CONCLUSION

RMIT (with assistance from one local friend) submitted five unique runs to the 2018 TREC CORE track of which two were automatic, and three manual. All of the submitted runs met the organizers' quality criteria for inclusion into the judgment pool. Similar to last year, we focused on user query variations and rank fusion to generate highly robust runs, with our best system RMITUQVDBFNZDM1 achieving an AP score of 0.385. RMIT placed fourth in the overall standings with respect to the number of *unique, relevant* documents found, with a total of 35. Outcomes pertaining to our research goals were mixed. Providing shallow judgments to improve double fusion effectiveness (**RQ1**) did result in an improved aggregate score, however, it did not yield statistically significant results. For **RQ2**, our findings show that improvements can be obtained by "hedging your bets" across an information need with external corpora for improved robustness. Our hypothesis for **RQ3** mirrors the conclusion drawn by Moffat et al. [15], suggesting that collections built *without* query variations are less robust than those that employ UQVs.

Acknowledgments. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP170102231), by an Australian Government Research Training Program Scholarship, and by a grant from the Mozilla Foundation.

REFERENCES

[1] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proc. SIGIR*, pages 667–674, 2008.

[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV100: A test collection with query variability. In *Proc. SIGIR*, pages 725–728, 2016.

[3] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. Retrieval consistency in the presence of query variations. In *Proc. SIGIR*, pages 395–404, 2017.

[4] R. Benham and J. S. Culpepper. Risk-reward trade-offs in rank fusion. In *Proc. Aust. Doc. Comp. Symp.*, pages 1:1–1:8, 2017.

[5] R. Benham, L. Gallagher, J. Mackenzie, T. T. Damessie, R-C. Chen, F. Scholer, A. Moffat, and J. S. Culpepper. RMIT at the 2017 TREC CORE track. In *Proc. TREC*, 2017.

[6] R. Benham, J. S. Culpepper, L. Gallagher, X. Lu, and J. Mackenzie. Towards efficient and effective query variant generation. In *Proc. DESIRES*, pages 62–67, 2018.

[7] G. V. Cormack, C. L. A. Clarke, and S. Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proc. SIGIR*, pages 758–759, 2009.

[8] J. S. Culpepper. Single query optimisation is the root of all evil. In *Proc. DESIRES*, page 100, 2018.

[9] T. T. Damessie, T. P. Nghiem, F. Scholer, and J. S. Culpepper. Gauging the quality of relevance assessments using inter-rater agreement. In *Proc. SIGIR*, pages 1089–1092, 2017.

[10] T. T. Damessie, J. S. Culpepper, K. Kim, and F. Scholer. Presentation ordering effects on assessor agreement. In *Proc. CIKM*, pages 723–732, 2018.

[11] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proc. SIGIR*, pages 154–161, 2006.

[12] B. T. Dinçer, C. Macdonald, and I. Ounis. Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proc. SIGIR*, pages 23–32, 2014.

[13] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.

[14] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Information Systems*, 27(1):2:1–2:27, 2008.

[15] A. Moffat, F. Scholer, P. Thomas, and P. Bailey. Pooled evaluation over query variations: Users are as diverse as systems. In *Proc. CIKM*, pages 1759–1762, 2015.

[16] L. Tan and C. L. A. Clarke. A family of rank similarity measures based on maximized effectiveness difference. *IEEE Trans. Know. & Data Eng.*, 27(11): 2865–2877, 2015.

[17] H. Zhang, M. Abualsaud, N. Ghelani, A. Ghosh, M. D. Smucker, G. V. Cormack, and M. R. Grossman. Uwaterloomds at the trec 2017 common core track. In *Proc. TREC*, 2017.