University of Amsterdam at the TREC 2019 Complex Answer Retrieval Track

Mahsa S. Shahshahani Jaap Kamps Maarten Marx

University of Amsterdam
Amsterdam, the Netherlands
Email: {m.shahshahani,kamps,maartenmarx}@uva.nl

Abstract

This paper documents the University of Amsterdam's participation in the TREC 2019 Complex Answer Retrieval Track. This is the first year we actively participate in TREC CAR, attracted by the introduction to the limited "budget" of 20 passages per heading in the outline. We conducted initial exploratory experiments on making each heading contain a unique set of passages within the outline, and even do this hierarchical for each subtree and main title/article level, hence remove any redundancy between passages for different "queries" within the same title. We experimented with top-down and bottom-up filtering approaches. At the time of writing we are still in the process of analyzing the results. Some initial observations are the following. First, the restriction makes the task very challenging, as assigning any passage to the right subtree is highly non-trivial. Qualitative analysis shows that our simple heuristics often make a different decision than the editorial judges on the heading to which a passage relevant to the title's topic is assigned. Second, the fraction of judged and relevant passages per individual query or leave node is very small, making it hard to draw any definite conclusions on our experiments, and also resulting in a too small recall base to evaluate our non-pooled runs in a meaningful way. Third, when aggregating all qrels and runs to the title level, there is reasonable effectiveness of the underlying BM25 rankings, showing that the underlying passage ranking is not unreasonable, and that the hard and interesting problem is in the exact assignment of passages to the "right" headings. All our analysis in this notebook paper is preliminary, and are we will provide a more substantive analysis in the TREC 2019 proceedings.

1 Introduction

We have followed the CAR Track with great interest since it's proposal, and the introduction of the fixed "budget" of 20 passages per leave node was the reason for us to participate actively in TREC 2019 edition of the track.

TREC CAR offers a potential solution to some of the major challenges we have been working on in a line of projects using political data [5]. This political data consists of billions of speeches by individual MP's, making it notoriously hard to extract a meaning overview of the different views of MP's, political parties, or government for a general topic [2]. We have effectively applied passage retrieval and exploratory search approaches, but these still require reading many individual passages or speeches. One of the main distinctions between the Wikipedia setting and our political data, is that our semantic assignments to speaker (and based on the speaker to the party, and status in terms of member of government or opposition, etc) are strict, whereas a single Wikipedia passage at TREC CAR maybe have relevance to multiple headings. The introduction of the top 20 budget in CAR at TREC 2019, is a large step towards bridging this divide – as with the limited budget redundantly repeating the same passage at different places of the outline comes with a considerable cost, and naturally encourages a strict and non-redundant assignment. It is exactly with this question in mind, that we conducted our initial experiments in allocating a high ranked passage to the "right" node of the outline.

2 Complex Answer Retrieval

For detailed information about the CAR track's experimental setup, we refer to the overview paper (this volume) and to the track homepage [6]. We provide a high-level summary to makes this paper self-contained.

2.1 Task definition

Traditional retrieval systems could successfully satisfy current simple and entity-centric information needs. TREC-CAR track is designed to turn researchers' attention to more complex information needs which require longer and more structured answers which may need using many pieces of information on separate documents.

In the first two years of running this track (2017 and 2018) the task was focused on retrieving and ranking passages according to a complex query with different facets. In the third year (2019) the main task has been changed to order retrieved passages in order to create a fully structured document in response to a multi-facet query.

Although the task has been changed in this year, the evaluation metrics seemed to remain the same as last year means they cannot evaluate the new requirements of the task.

2.2 Dataset

Paragraph corpus The paragraph corpus which is considered as the passage-pool consists of 20 million paragraphs obtained from Wikipedia pages from a snapshot of 2016.

Outlines In this year outlines are extracted from TQA dataset. TQA queries seem to be slightly easier than hierarchical Wikipedia headlines because they contain fewer sections and the depth of queries is limited to 2 or 3.

3 TREC CAR Experiments

The main task for this year's competition is to select top K ($K = 20$) passages for each question that can make a good answer to the question. To this end we break the goal to two steps:

- First, we should rank passages in response to each question (essentially passage retrieval taking each query independently).
- Second, we should select the best passages regarding other considerations such as diversity rather than just relevancy (essentially attempting to assign each passage to the "right" part of the outline).

In order to maximize diversity, we are most interested in an extreme version of this approach where we have no redundancy within an outline, and uniquely assign each passage to the 'best' node of the outline.

As it turned out, also the editorial judges have done the same, and have not assigned the same paragraph to multiple headings of the same outline. This makes our own experiments center on the question on to what degree our simple heuristics mimicks the choice for the assignment of assessors – a very challenging task indeed!

3.1 Passage Ranking Approach

This part of the solution is like the main TREC CAR task in Y1 and Y2, essentially asking to rank passages for each heading in the outline independently.

We tested both ranking with BM25 and re-ranking with BERT (after retrieving top passages using BM25) models on Y2 data, to get an insight of which one can be more useful. Although using BERT had shown promising results on Y1 [4], it did not result in better results in comparison with BM25 on Y2 data. This is aligned with what results from TREC-CAR 2018 had showed. According to Dietz et al. [1] neural network models did not work as well as learning to rank models on Y2. Our training experiments reranking passages based on using pre-trained BERT, shown in Table 1, are confirming these observations. Qualitative analysis shows both the clear value of BERT to uncover meaningful passages not literally matching the query, but these often are unjudged and appear lower in the rankings. Qualitative analysis also shows a clear loss of precision against traditional text and word-based expansion approaches, in line with the observed higher scores of those approaches in Y2. As Y3 is very much about high precision, given the limited budget of 20 results, none of our official submissions was based on the BERT rankings.

BM25 is the traditional unsupervised ranking model which scores the relevancy of documents (here passages) in regards to queries based on the frequency

Run	MAP	MRR
BM25	0.2895	0.6223
BM25+RM3	0.3049	0.6549
BERT	0.1870	0.5118

Table 1: Training results on TREC-CAR 2018.

of common terms between the query and passage. We used the Lucene framework[1] to run the BM25 model for passage ranking, in a set-up following the infrastructure already provided by the track organizers. Rather than use our own index and tuned BM25 runs, we rerun our experiments based on the rankings provided by the organizers, in order to start from the same benchmark BM25 runs as other teams and allow for clearer comparative evaluation between teams and submissions.

Several pseudo relevance feedback models have been introduced by IR researchers to expand a query with words that occur in top-retrieved documents in response to a query (considering high ranking as an approximation of relevance). In Nanni et al. [3] the usage of feedback models has been studied for TREC CAR. We used RM3 feedback model for our submission. RM3 is based on interpolation between the language model created by the query itself and language model created by the expanded query. We use a Dirichlet smoothed language model for the feedback run. Each query is expanded with top 10 terms extracted from the top 10 feedback paragraphs.

Again, in order to allow for clearer comparative evaluation, we reran our experiments and based our submissions on the rankings provided by the organizers.

3.2 Passage Ordering Approach

The main task for this year's competition is not about passage ranking, but it is about ordering top-ranked passages in response to the multi-facet query in a way that all selected passages have:

- Highest relevance of all passages.
- Balanced coverage of all query facets as defined through headings in the outline
- Maximizing topical coherence, minimizing topic switches, i.e., first all passages about one topic, then all passages of the next topic while avoiding to interleave multiple topics.

By way of example, assume we have two queries such as "radio waves/television" and "radio waves/AM and FM radio." The general topic in these examples is

[1] https://lucene.apache.org

"radio waves" and the facets are "television" and "AM and FM radio." Here we want to rank and order passages from Paragraph Corpus in response to both queries in a way that the response include both general information about the general topic ("radio waves") and more detailed information about the facet ("television").

We applied two methods to select the passages among top-ranked documents:

Top-Down The intuition behind this approach is "safety first" – earlier years showed that finding a set of relevant passages for the title as a whole is quite effective. So let's first lock in the "best" ranked 20 passages for the top level, and then move step by step through filtering out any passage that was already select before at a higher or earlier node of the outline.

In terms of our example, we would do the most general topic of "radio waves" first (including TV, radio and other facets combined), and only then process each facet in turn (here, "television" and "AM and FM radio") and leave out any passage that already was selected before.

Bottom-Up The intuition behind this approach is assign passage to the most specific node of the outline, hence only general passages covering the whole topic should be assigned to the top level, and any passage exclusively about a particular facet should be assign to that facet. There are many reasons why this this way is preferred and it is used as a fundamental principle of information organization in library and information science (known as "Cutter's rule") for far more than a century.

In terms of our example, we would start at the leave nodes, and select the highest ranked passages for a particular facet, TV passages assigned to "television" and radio passages to "AM and FM radio", and step-by-step work our way up the tree, excluding any passage that already was selected at an earlier, more specific level of the outline.

Although more risky, if successful the bottom-up approach is intuitively the preferred approach, and all our official submissions were based on it.

4 Preliminary Results

This section presents a preliminary analysis of our results.

4.1 Relevance Judgements

Table 2 shows the number of (judged) topics and relevant passages, and their distribution over relevance levels. With 2,790 relevant passages for 303 headings, the recall base is small with less than 10 relevant passages per heading. Looking deeper into the qrels, Table 3 provides the distribution of relevant passages (at any level of relevance) over the queries or headings. We see that, indeed the

Topics	Judged topics	Rel. passages	Rel-0	Rel-1	Rel-2	Rel-3
722	303	2,790	–	693	1,577	520

Table 2: Various statistics (#) on the Y3 topics and judgments.

Level	#	Min	Max	Mean	Median	St.dev.
Heading	303	1	79	9.2	7	9.4
Title	50	12	107	50.7	47	22.8

Table 3: Various statistics on the Y3 topics and judgments per heading and aggregated per title.

recall base is small: for some topics there is only 1 relevant passage, and the distibution is skewed to the lower end with a median of 7 per query.

Looking at our non-official runs, which did not contribute to the pooling, we see very few judged and relevant results. With less than 5% of the top 20 results of a non-official run in the qrels, it is not meaningful to include results for the non-official runs at the time of writing. Note that the qrels do not contain information on judged but non-relevant passages, which would allow the use of specific measures (such as bpref or InfAP) dealing with incompleteness, although evaluating runs with so few judgments can only be done with extreme caution when interpreting the results.

Table 3 also provides the same distribution after mapping all passages to the title level, lumping together all facets into one bag of passages. We see that there are judgments for 50 titles, hence there are on average about 6 facets judged per title or outline as a whole combined, and recall base per title is less small with a minimum of 12 relevant passages, and around 50 relevant passages per title on average.

In the rest of this paper, we will restrict our analysis to our official submissions only.

4.2 TREC 2019 CAR Submissions

Although, the CAR track generously allowed up to 10 submission, the "budget" was restricted to a single one with *high* judging precedence (presumed to be assessed), and another two with *medium* precedence (possibly, but not necessarily, judged). As we did not replicate all our 2018 variations on the 2019 data, we refrained from submitting any of the remaining *low* judging precedence runs (i.e., additional submissions not meant to be assessed).

We submitted the following three submissions:

BM25+RM3 The first submission, labeled `UvABM25RM3` and *medium* judging precedence, is based on a word-based BM25 ranking, using the `section-path` queries and RM3 blind feedback. This run was essentially derived directly

Submission	MRR	Precision			NDCG			MAP
		5	10	20	5	10	20	
BM25+RM3	0.1877	0.0884	0.0779	0.0649	0.0891	0.0891	0.0891	0.0442
BottomUp-Score	0.1126	0.0607	0.0574	0.0442	0.0561	0.0561	0.0561	0.0252
BottomUp-SumScore	0.1142	0.0574	0.0597	0.0442	0.0592	0.0592	0.0592	0.0260

Table 4: Preliminary results on the TREC 2019 CAR Track's Passage Ordering Task.

by restricting the provided passage rankings to the restricted budget, just based on the retrieval status value and ignored potential overlap or dependencies between the different heading within the same outline. While our aim was to remove all redundancies and assign passages to a unique heading within the same title, this run serves as a baseline and proper comparative evaluation would profit from this baseline also contribution to the pool of passages to be judged.

BottomUp-Score The second submission, labeled `UvABottomUp2` and *medium* judging precedence, is based on BM25+RM3 and applying the *Bottom Up* approach to select paragraph to assign to each heading, allowing no redundancy. This is a straightforward heuristic where we process the source ranking, and all remaining paragraph at each heading remain ordered according to their original RSV value in the source run.

BottomUp-SumScore The third submission, labeled `UvABottomUpChangeOrder` and *high* judging precedence, is again based on the BM25+RM3 and applying again the *Bottom Up* approach to select to which node a paragraph should be assigned, allowing for no redundancy, but uses possible other occurrences of this passage in other headings to aggregate their scores, and hence rerank the passages per heading according classic RSV combination or score fusion approaches. This will change the ordering relative to the BottomUp-Score submission above.[2]

The preliminary results are shown in Table 4. As we retrieve only 20 results per topic or heading, we focus on early precision measures. A few observations present themselves. First, all scores are relatively low, clearly demonstrating how hard the task of accurate assigning passages to the right heading of the outline is, relatively to more standard document and passage retrieval tasks. Second, even the source run, BM25+RM3, scores quite low indicating that many paragraphs in that run have not been regarded as relevant by the respective assessors. Not surprisingly, the two runs de-duplicating the submissions by uniquely assigning paragraphs to unique parts of the outline score even lower

[2] Although this was our high judging precedence run, and it was validate by the eligibility script before submission, there remained an accidental duplicate passage id in the run, we report here the evaluation after removing duplicates. This may also have prevented this run from being parsed properly and contributing to the pool of passages to be assessed.

Submission	1	2	5	10	20
BM25+RM3	10.23%	10.07%	8.84%	7.79%	6.49%
BottomUp-Score	6.60%	6.77%	6.07%	5.74%	4.42%
BottomUp-SumScore	6.93%	6.27%	5.74%	5.97%	4.42%

Table 5: Percentage of results with a known relevance label (only judged and relevant) per rank.

Submission	MRR	Precision			NDCG			MAP
		5	10	20	5	10	20	
BM25+RM3	0.7064	0.4873	0.4291	0.3573	0.3830	0.3830	0.3830	0.1014
BottomUp-Score	0.4798	0.3345	0.3164	0.2436	0.2659	0.2659	0.2659	0.0610
BottomUp-SumScore	0.4878	0.3164	0.3291	0.2436	0.2706	0.2706	0.2706	0.0592

Table 6: Preliminary results on aggregated topics with a bag of relevant passages per title.

than the baseline run. This is indeed a hard and risky additional constraint, and qualitative analysis comparing difference in the runs presents cases were did retrieve the 'relevant' passage, but our heuristic approach assigned it to a different heading than the ground truth assessment did. Third, although scoring low, we can do a comparative analysis of the two variant approaches. There is no clear winner, with precision at 5 favoring the original score based approach, and the others slightly favoring the sum of scores approach.

We look a little deeper, and looked at the distribution of judged and judged relevant pages. The fraction of judged relevant over ranks are shown in Table 5 which reveal that only a small fraction of the results in our runs have a relevance label in the qrels. Recall from before that the qrels do not contain judged but not relevant labels, hence we cannot investigate directly whether this is due to these results not being assessed (hence the runs not contributing to the pooling), or whether they were all assessed up to a given rank but deemed non-relevant by the judges. Qualitative inspection of some cases didn't reveal an immediate explanation, none of the retrieved passages looked clearly off-topic. We will investigate this in further analysis.

Although we cannot exclude various system or submission format conversion errors, the runs were partly based on the independent implementation from the organizers, and those rankings looks reasonable in superficial inspection. A plausible explanation could be in the small recall base, low pooling depth, and great diversity over different submissions. To quantify this, we create synthetic versions of the qrels and our submissions, in which we assign all retrieved passages to the title as a single topic (hence retrieving more than the top 20 results). We also aggregate all the relevant pages for a particular subheading to the main title in the qrels (as previously shown in Table 3). In this way we create fewer topics with a larger recall base that can indicate the underlying ranking quality of the used approach. These additional results for analytical purposes are shown in Table 6. We make two observations. First, we see that the precision scores are in the range of what's expected for a passage retrieval or document retrieval task, and not unusually low. This is suggesting that the rankings we start out with are of reasonable quality. Second, we see again a drop in effectiveness for the additional non-redundancy processing, proving that this is the hard task even when we ignore the case where we assign a "relevant" passages to the wrong heading. Note that in this setting, we also "remove" potential assessor disagreement on to what heading to assign a given "relevant" passage to.

Our analysis present clear evidence that one of the next steps in CAR, where we enforce systems to be selective, or even require non-redundancy over all retrieved passages within the same outline, is indeed challenging. As this is a necessary step to go from passage retrieval to generating comprehensive complex answers, this is perhaps only making this task more interesting and more important to look at.

5 Conclusion

This paper documents our first participation in the TREC 2019 CAR Track.

At the time of writing we are still in the process of analysing the results. Some initial observations are the following. First, the restriction makes the task very challenging, as assigning any passage to the right subtree is highly non-trivial. Qualitative analysis shows that our simple heuristics often make a different decision than the editorial judges on the heading to which a passage relevant to the title's topic is assigned. Second, the fraction of judged and relevant passages per individual query or leave node is very small, making it hard to draw any definite conclusions on our experiments, and also resulting in a too small recall base to evaluate our non-pooled runs in a meaningful way. Third, when aggregating all qrels and runs to the title level, there is reasonable effectiveness of the underlying BM25 rankings, showing that the underlying passage ranking is not unreasonable, and that the hard and interesting problem is in the exact assignment of passages to the "right" headings.

We hope and expect that the valuable bench-marking data created at TREC CAR will be of great value to motivate, and greatly facilitate, further research into this important and hard problem.

Acknowledgments

A special word of thanks to the TREC CAR Track organizers for their amazing support during the track! This research was supported in part by the Netherlands Organization for Scientific Research (NWO, grant # CISC.CC.016, ACCESS project). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

[1] L. Dietz, B. Gamari, J. Dalton, and N. Craswell. TREC complex answer retrieval overview. In *TREC*, 2018.

[2] R. Kaptein, M. Marx, and J. Kamps. Who said what to whom? capturing the structure of debates. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 831–832. ACM Press, New York NY, USA, 2009.

[3] F. Nanni, B. Mitra, M. Magnusson, and L. Dietz. Benchmark for complex answer retrieval. In *Proceedings of the ACM SIGIR international conference on theory of information retrieval*, pages 293–296. ACM, 2017.

[4] R. Nogueira and K. Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.

[5] Political Mashup. Who said what? to whom? and why? and when? `https://www.politicalmashup.nl/`, 2019.

[6] TREC CAR. Complex answer retrieval. `http://trec-car.cs.unh.edu/`, 2019.