Complex Answer Retrieval goes into Politics
Reproducibility Study

Mahsa S. Shahshahani, Jaap Kamps, and Maarten Marx

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

Abstract. Complex Answer Retrieval (CAR) is an extension of QA to multi-document summarization approaches, in order to deal with complex questions naturally requiring answers covering multiple aspects, pieced together from multiple documents and contexts. CAR has been running at TREC since 2017, in the form of a complex passage retrieval task against Wikipedia, with solid progress and results. We apply CAR to a real world application domain: massive scale parliamentary proceedings consisting of 100s of years of recorded speeches in parliament. Even the simplest query or question on this data results in 1,000s of speeches of individual members of parliament, parties and government. This use-case presents an obvious real-world application area in need of a CAR approach. We conduct an extensive case study of Parliamentary CAR, and have the following findings: First, according to the current CAR set up and measures the approach works almost perfect. Second, deeper analysis against the underlying motivation of CAR—essentially generating a "missing" wikipedia page like summary on the topic—we see far from optimal performance. The major upshot of our reproducibility study is to both validate the utility of CAR in a major application domain, and to highlight further refinements of CAR that would be beneficial for real-world application of CAR in politics, as well as in many other domains.

Keywords: Complex Answer Retrieval · Multi-facet answer · Political domain

1 Introduction

This paper applies Complex Answer Retrieval (CAR) approaches as developed at TREC, to the domain of politics, in order to investigate their merits in an important real world application. Let us start with a motivating example.

In October 2019, just two weeks before the 43rd Canadian federal election, there was a debate in which the leaders of all of the parties showed up on stage to discuss their ideas and plans for Canada. There was a key moment in this debate when the leaders of the two main Canadian parties (Liberal and Conservative) were discussing their plans toward *climate change*, the leader of one of the other parties (New Democratic Party) jumped in with the sentence "*You do not need to choose between Mr. Delay [Trudeau, Liberals] and Mr. Deny [Scheer, Conservatives].*" This is a single sentence which summarizes all the actions taken by the two main parties toward climate change. This key

moment got lots of media attention, both appearing on several News websites and received hundreds of retweets on Twitter.[1] As this debate took place very close to election day it can impact voters' opinion and vote. But, is this a valid sentence? How should a voter or journalist validate this sentence? They need to go through all the actions taken by the previous governments and their discussions on them, and painstakingly process many years of parliamentary debates on these matters.

Complex Answer Retrieval (CAR) is a task introduced at TREC 2017 promoting researcher on questions which need multi-facet answers. This is different from complex QA in which the question is complex and needs reasoning to be understood. Here, the question is not complex, but it requires a complex and multi-facet answer. At TREC, the task is evaluated using Wikipedia as a passage pool and questions from Wikipedia headlines and the TQA dataset [5]. Our motivating example fits the CAR setting and presents a real-life application of it, allowing us to investigate how well CAR approaches work in a realistic application.

As this is paper is part of the reproducibility track, we use a freely reusable and publicly available corpus,[2] and make all our specifically analyzed data and relevance labels available.

2 Complex Answer Retrieval

The CAR task was introduced as part of TREC in 2017. In the first year of this track, the passage pool was created using one of the dumps of English Wikipedia and questions were created by joining different levels of headings in Wikipedia pages. The participants were supposed to propose models to rank passages (paragraphs from Wikipedia) for each question. In the 2018 edition, questions were both from Wikipedia headings and TQA dataset. And again in 2019, all the questions were taken from TQA dataset and the passage pool was created from paragraphs from Wikipedia. The CAR results of TREC 2019 are ongoing and have not been published yet. Hence, we will give a very brief overview of models proposed in the first two years.

Proposed models can be categorized to learning to rank models [4], neural-network based models [6], and more traditional IR models such as BM25 with or without query expansion [3,2]. In [7] the impact of different query expansion models has been studied and it showed that using query expansion can increase the performance of base model. On the TREC 2017 CAR benchmark, the best methods were based on neural network models. In contrast, the best performing models on CAR at TREC 2018 were learning to rank models. There is ongoing investigation in ways to combine both approaches, e.g., [8] used a pre-trained BERT model to re-rank top-1000 firstly ranked passages by BM25, and showed that using BERT can improve the performance by 50% in terms of MAP measure.

The main general conclusion from the CAR track over the years is that a variety of approaches is quite effective for the task, when based on the Wikipedia corpus and the passage retrieval evaluation as used at TREC.

[1] E.g., `https://www.bbc.com/news/world-us-canada-49901453`.

[2] Available from `https://www.politicalmashup.nl/`.

3 Parliamentary Complex Answer Retrieval

In this section, we will present a case study on the CAR approach applied to a novel domain of politics.

3.1 CAR in Politics

The corpus we used in this paper to examine CAR setting consists of debates in the Canadian Parliament [1] including debates from 1901 to 2014. All of the debates are documented and all of their speeches are annotated by the speaker, the political party which the speaker is associated with, and date. These debates are publicly accessible via a search system. [3]

In order to be precise, the TREC CAR task is as follows:

Definition 1. *In TREC Complex Answer Retrieval each question consists of multiple facets and the passage pool is formed by passages from a Wikipedia dump.*

In the current TREC setting, this reduces the task to a passage ranking problem, in which systems rank passages for each of the facets of the topic. The underlying motivation, or ultimate goal of CAR, would be a system that is able to create a novel Wikipedia page for each topic using its facets as headings for different sections of the created complex answer. One can view the TREC CAR passage retrieval as a necessary initial phase, extracting all passages with key nuggets of information, and further NLP processing would turn these into a coherent structured textual summary.

The task of Parliamenary CAR is as follows:

Definition 2. *In Parliamentary Complex Answer Retrieval each question consists of multiple facets and the passage pool is formed by speeches from proceedings of Parliamentary debates.*

Again, in the ideal setting, the system should be able to create a Wikipedia page for each query using its facets as headings for different sections of the created answer.

3.2 Parliamentary CAR Effectiveness

Motivated by the example in the introduction, we pick "climate change" as a case study in this paper. We want to see what actions and attitudes members of each one of the two main parties have taken towards "climate change." Speaking in CAR terminology, we define two facets for the main question ("climate change"): conservative party, liberal party. We expect the CAR system to rank speeches concerning each facet. Top speeches for each facet should contain the most important pieces of information.

For our case study, we retrieved all speeches containing "climate change" phrase from 2010 to 2014 using a simple term-match based ranking model. Then we filtered out the speeches from other political parties rather than conservative and liberal parties which are of our interest in this example. At last, we had 621 speeches to be labeled. To keep the similarity with TREC-CAR data [2] we applied almost the same labeling approach which includes the "must be mentioned," "could be mentioned," and "non-relevant" indicating how importantly the speech should be mentioned in this section

[3] `http://search.politicalmashup.nl/`

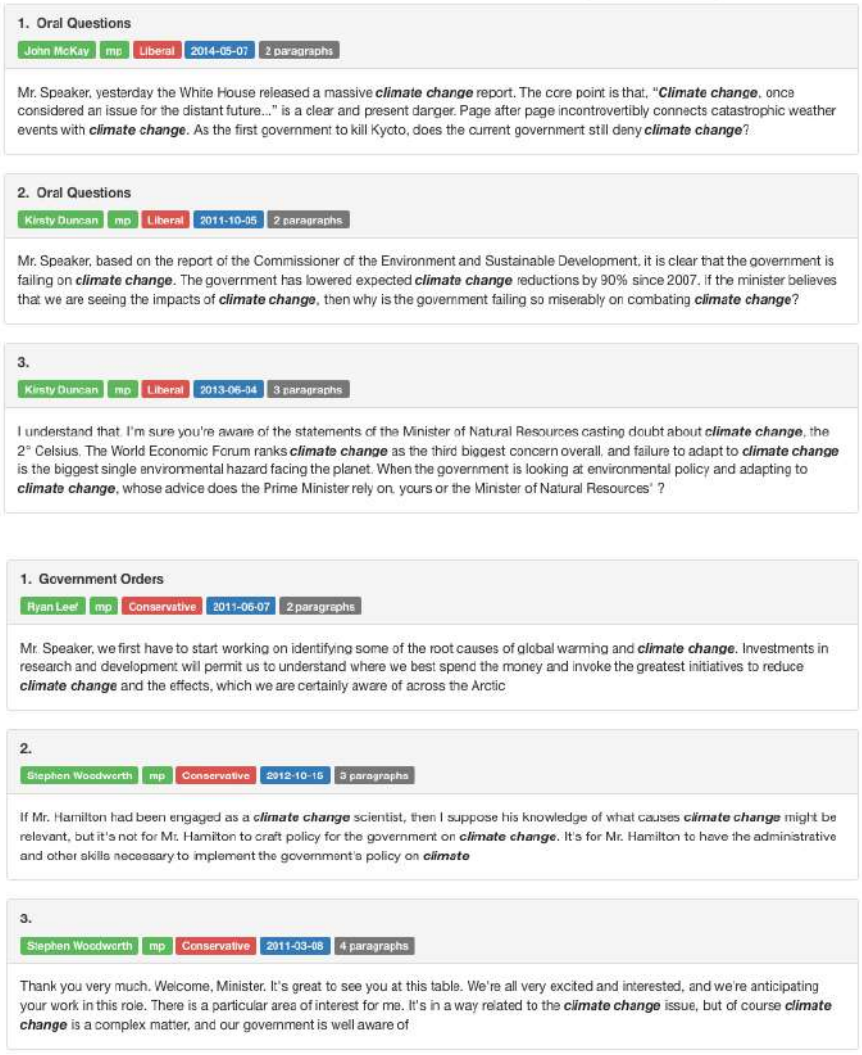

Fig. 1: Top ranked speeches for "climate change/liberals" (top) and "climate change/conservatives" (bottom).

of the created article as an answer to the query. In a next step, we labeled speeches indicated with "must be mentioned" label with "highly informative" and "informative" based on how importantly each speech should be mentioned in the ideal complex answer to the query.[4]

[4] As it turned out, the remaining set of speeches did not contain any "non-relevant" ones, as all the speeches were related to *climate change* in some sense, and the party filters are strict. This highlights one of the differences between this domain and Wikipedia.

Table 1: Parliamentary CAR results for the "climate change" case study

Question Facet	*could mention*	*must mention*	*highly informative*
`climate change`	1.000	1.000	0.6000
`climate change/Liberals`	1.000	1.000	0.1000
`climate change/Conservatives`	1.000	1.000	0.1200
Precision @ 20	1.000	1.000	0.2733

Following current TREC CAR systems, we produces a ranked-list of passages for each facet of a question. Here, we defined two facets; the top ranked speeches for "liberals" and "conservatives" facets are shown in Figure 1. Following CAR at TREC 2019, we examined the top-20 speeches for each of the facets based on our earlier labeling. All of them were among the speeches labeled with "must be mentioned." So, as shown in Table 1, the Precision is 1.0. It is not surprising as we just labeled all the speeches explicitly containing "climate change" phrase which means that normally the precision would be high while the recall can be very low. So, we see that the current CAR setting fits this data and domain very well with a very high precision.

Whilst the scope of the case study is rather small, it findings generalize to a wide range of general topics, where finding many speeches discussing this topic results in very high precision—scoring very well on current CAR measures.

4 Toward Comprehensive Answer Retrieval?

In this section, we will further analyze the CAR approach from the underlying motivation of generating a complex answer that comprehensively summarizes the topic.

In the ideal CAR setting the goal is to create a Wikipedia-like page with a section for each facet in response to the user's question. There actually exist such an ideal response for our case study in the Wikipedia page shown in Figure 2. But, are current efforts to address CAR task even close to reaching this goal? Can we reduce this task to passage ranking? How to extract the most informative and diverse and novel pieces of information among them? We also need the CAR system to identify and generate facets automatically, for which we have natural handles in the political domain in terms of the MPs, Parties, and Government vs. Opposition status, as well as further temporal facets and filters.

To address these questions, we use our refined labels on the informativeness of individual speeches. We labeled speeches with two labels by asking a simple question: *Does this speech include a line of information we expect to see a Wikipedia-like page summarizing all the actions and attitudes taken by different political parties toward climate change and its consequences?* First, we evaluated the ranked-list of speeches for the general high-level query: "climate change." As we observed, the precision in this case was 60%. Note that we still do not take redundancy of speeches, diversity and novelty factors into account. If we look into these aspects, we see that almost 90% of these "highly informative" speeches refer to the same information ("killing the Kyoto protocol"), and various information on the Wikipedia page is not covered yet.

Second, we evaluated the top-ranked speeches for each facet: liberal and conservative. The precision for these two facets was only 10% and 12% respectively. To see what

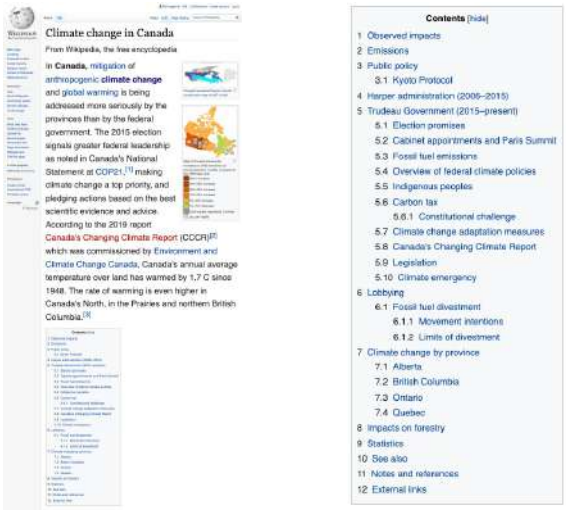

Fig. 2: An ideal Wikipedia-like complex answer on query "climate change" in political domain. Taken from: `https://en.wikipedia.org/wiki/Climate_change_in_Canada`

makes the precision for each facet very low, we examined carefully the speeches. The data that we use in this paper is conversational and consists of dialogues and questions and answers between members of Parliament. We observe that often the clearest and shortest summary of a political stance of a longer speech is not in the speeches of this particular politician. Instead, they are among the opponents' claims about them when they refer to actions they had taken before. This suggests that in order to able to extract the most informative summary about a particular politician or party, we have to look beyond the speeches of the members involved!

We opted for a case study approach as this allows us to dig deeper, and this revealed that the perfect performance (according to the TREC CAR measures) leave still much to be desired in terms of generating a comprehensive answer. This challenges the underlying assumptions of CAR reducing the task to passage ranking problem. It also suggests a path forward to refine CAR to take deeper and more complex processing of the passages into account, as concrete next steps toward the goal of generating a truly comprehensive complex answer as done by human editors on the Wikipedia page shown in Figure 2. A key element for such a next step seems to be that passages should not be considered as independent and the relationship between passages (or speeches) and information flow through them should be taken into account.

5 Conclusions

In this paper, we explored the utility of Complex Answer Retrieval systems in a new domain, speech level retrieval over massive political data. Current CAR task and TREC-CAR (2017 to 2019) datasets use Wikipedia as the passage pool to create answers. We presented a real-world use case where a CAR approach has great potential benefits, and validates the utility of CAR approaches outside the synthetic domain of Wikipedia

passages as used in TREC. We performed a representative case study to show how CAR works currently in this new domain, and showed that it is very effective based on current CAR evaluation measures. However, deeper analysis showed that this may not yet be enough to fully realize the ambitious underlying goals of CAR. Our analysis gives concrete suggestions in ways that it could move beyond passage ranking task by considering the dependencies between passages. When applied to conversational data, or other highly redundant data, this dependency is crucial and should be taken into account. In fact, this seems also beneficial for CAR systems in many other domains and applications, including Wikipedia.

Acknowledgments The data we used here is taken from a publicly available corpus, and we will make all the extracted speeches and relevance labels publicly available upon acceptance of this paper.

References

1. K. Beelen, T. A. Thijm, C. Cochrane, K. Halvemaan, G. Hirst, M. Kimmins, S. Lijbrink, M. Marx, N. Naderi, L. Rheault, R. Polyanovsky, and T. Whyte. Digitization of the Canadian parliamentary debates. *Canadian Journal of Political Science*, 50:849–864, 2017.
2. L. Dietz, B. Gamari, J. Dalton, and N. Craswell. TREC complex answer retrieval overview. In *TREC*, 2018.
3. L. Dietz, M. Verma, F. Radlinski, and N. Craswell. TREC complex answer retrieval overview. In *TREC*, 2017.
4. S. Kashyapi, S. Chatterjee, J. Ramsdell, and L. Dietz. TREMA-UNH at TREC 2018: Complex answer retrieval and news track. In *TREC*, 2018.
5. A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4999–5007, 2017.
6. S. MacAvaney, A. Yates, and K. Hui. Contextualized PACRR for complex answer retrieval. In *TREC*, 2017.
7. 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.
8. R. Nogueira and K. Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.