VES Team at TREC Conversational Assistance Track (CAsT) 2019

Vasileios Stamatis, Leif Azzopardi, Alan Wilson

University of Strathclyde, Glasgow G1 1XQ, UK
vasileios.stamatis@strath.ac.uk

Abstract

In this work we present our submission at the TREC Conversational Assistance Track 2019. For this year's track, we have focused on developing a baseline system from which we can build upon in the future. Our system is built upon a Lucene index which serves up results (using BM25), these are then re-ranked by BERT given the conversational context.

1 Introduction

The TREC Conversational Assistance Track (CAsT) aims to advance research in conversational search systems. This is the first year that TREC CAsT is running and the focus is on candidate information ranking in context.

CAsT defines conversational search as an information retrieval task in a conversational context. Given some questions in a specific context with a specific topic, the system should be able to answer them one by one staying in that context. More specifically, in order to satisfy a user's need, the system will retrieve answers for a series of follow up questions from the user. For example a topic from the evaluation topics was the the bronze age collapse and some questions that the user can ask are:

1. Tell me about the Bronze Age collapse.
2. What is the evidence for it?
3. What are some of the possible causes?
4. Who were the Sea Peoples?
5. What was their role in it?
6. What other factors led to a breakdown of trade?

On that occasion the system should be able to answer the questions while staying in the context of the bronze age collapse. Some questions can be answered directly like the questions 1,4 but some others need to know the context from the previous turns. For this year's track we addressed this issue using coreference resolution for example replace "it" with "the Bronze Age collapse". The results after the coreference resolution for this specific example were:

1. Tell me about the Bronze Age collapse.
2. What is the evidence for the Bronze Age collapse?
3. What are some of the possible causes?
4. Who were the Sea Peoples?
5. What was the Sea Peoples role in the Bronze Age collapse?
6. What other factors led to a breakdown of trade?

In this paper we describe our run submitted for the TREC CAsT 2019. Our main contribution is a Lucene retrieval module in conjunction with a Bert Re-ranker [2].

2 Method

Collection and Materials: The data used for this task is from three datasets: MS MARCO Passage Ranking collection [5], TREC CAR paragraph collection v2.0 [3], TREC Washington Post Corpus version 2. The MS MARCO, CAR, WAPO data after the processing, cleaning etc. are 3.5, 13.5, 3.6 GB and consists of 8.8, 29.8, 9.2 million passages respectively (see Table 1).

All three datasets processed, cleaned and merged in one dataset in which each line has the format: {Document_id, Title, Document}. The final dataset has a size of 20.6 GB and consists of 47.8 millions passages (see Table 1).

The Document_id is the id of the passage in each collection with the name of the collection added at the start i.e.
{DocID: CAR_00000047dc43083f49b68399c6 deeed5c0e81c1f, Title: , Paragraph: On 28 October 1943, Fuller sailed from....}.
For WAPO as there were more than one passages with the same id, we also add another index at the end of the Document_id for each paragraph with the same id i.e.
{DocID: WAPO_ffd6b3d07764da97d7a 3b287035ff5f2-2, Title: The NSA..., Paragraph: The National Security Agency ...}.
This is happening because the WAPO dataset consists of topics with individual ids and each topic consists of many paragraphs. So the paragraphs belong to a topic have the same id.

For the title we used the topic titles for WAPO and for every passage in CAR and MARCO collections we searched for queries connected with the passage in

	CAR	MS MARCO	WAPO	Total
Size (GB)	13.5	3.5	3.6	20.6
Passages (Millions)	29.8	8.8	9.2	47.8

Table 1: Description of Different Datasets

previous years' qrel files. Where a passage had a query connected with it, we used this query as title.

During the cleaning process we removed any special characters i.e.(%,$ etc.) and urls where they existed. We also converted the text to lowercase and implemented Kstem analyzer. Using this data the index created using Lucene4IR, a toolkit for information retrieval [1].

Conversational Requests: The next step was the preprocessing of the Evaluation topics year 1 V1.0. There were 50 topics each of which has 7 to 12 questions and each questions consisting of 3-15 words. The topics were general topics i.e. medial topics, history topics, general knowledge topics etc.

For each topic the questions passed from a coreference resolution module using Stanford CoreNLP, a module for Natural Language Processing [4]. In almost every topic there were changes by the coreference resolution module. The Bronze Age Collapse example can be found in the introduction.

After that, all the questions ran against the index using BM25 algorithm within Lucene4IR and retrieved 1000 documents per query.

The final step was the re-rank using Bert. Bert is a language representation model. We used a Bert model fine tuned in the MS MARCO dataset [6] for the re-rank process. For every query we fed to bert the whole query and the whole passage to calculate their relevance.

For getting the new relevance score R we used the formula:

$$R = R_{Bert} * R_{BM25}$$

where R_{Bert}, R_{BM25} is the relevance score from Bert and BM25 respectively. We multiplied the scores in order to let the Bert model decide the final ranking because it would slightly change the score of the correct passage to a query and it dropped significantly the score of the non-relevant passage such that we got the final ranking. A flowchart of the workflow pipeline can be found in Fig. 1.

3 Results

Our system achieved NDCG = 0.3835 @ 1000 which is almost the same as the median of all systems and MAP = 0.2055 @ 1000 which is 0.03 higher than the median of all systems. The results can be found in Table 2

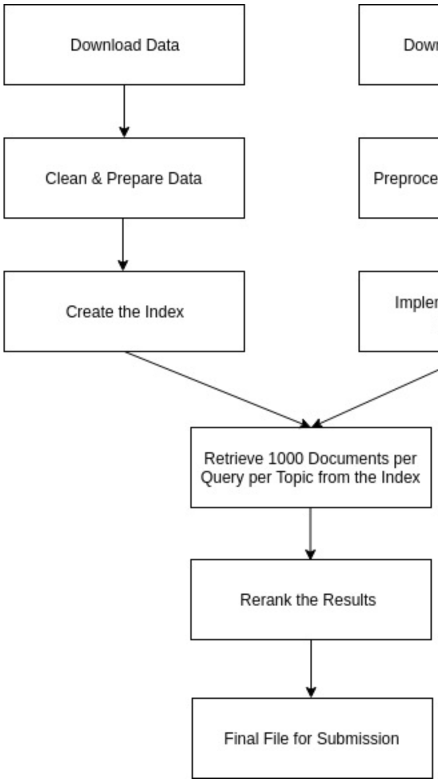
Download Data
Download Evaluation Topics
Clean & Prepare Data
Preprocess Evaluation Topics
Create the Index
Implement Coreference Resolution
Retrieve 1000 Documents per Query per Topic from the Index
Rerank the Results
Final File for Submission

Figure 1: Flow Chart Pipeline

4 Conclusion

In this work we presented our submission for CAsT 2019. We created a baseline system composed of a Lucene retrieval module and a Bert re-ranker. Further work will be focused on improving our system by embedding a topic classifier and model the history of the conversation using more advanced techniques.

	NDCG@5	MAP@5	NDCG@1000	MAP@1000
Median of All Systems	0.296	0.042	0.384	0.174
Our System (VES1000)	0.3038	0.0438	0.3835	0.2055

Table 2: NDCG & MAP Results of Systems

References

[1] Leif Azzopardi, Yashar Moshfeghi, Martin Halvey, Rami S Alkhawaldeh, Krisztian Balog, Emanuele Di Buccio, Diego Ceccarelli, Juan M Fernández-Luna, Charlie Hull, Jake Mannix, and Sauparna Palchowdhury. Lucene4IR: Developing Information Retrieval Evaluation Resources using Lucene. Technical report.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Technical report.

[3] Ben Gamari. Laura Dietz. ”TREC CAR 2.0: A Data Set for Complex Answer Retrieval”. Version 2.0, 2018. Technical report.

[4] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David Mcclosky. The Stanford CoreNLP Natural Language Processing Toolkit. Technical report.

[5] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. Technical report, 2016.

[6] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.