DCU at the TREC 2019 Conversational Assistance Track

Piyush Arora, Abhishek Kaushik and Gareth J. F. Jones

ADAPT Centre, School of Computing
Dublin City University, Dublin 9, Ireland
{piyush.arora,abhishek.kaushik,gareth.jones}@adaptcentre.ie

Abstract. We describe the DCU-ADAPT team's participation in the TREC 2019 Conversational Assistance Track (CAsT) track. The CAsT track focuses on two main aspects: i) system understanding of information needs in a conversational format, and ii) finding relevant responses using contextual information. In our participation in the CAST track, we focused on the second aspect of finding relevant information using contextual information from the queries for a conversational search system. We carried out two main investigations: i) Query formulation using syntactic analysis, and ii) Data Fusion of results for re-ranking top candidates retrieved from three different data sources used in the CAsT track. We find that using only query formulation and data fusions techniques attains average results in comparison to other submissions, which are not sufficient to answer questions in a conversational setting reliably.

Keywords: Conversational Assistance, Question Answering, Understanding Dialogue, Query Formulation, Sentence Selection

1 Introduction

The TREC CAsT track is a conversational search benchmark focusing on understanding the evolution of user information need in a conversational setting comprising of several turns in a search dialogue. The goal of this track is to identify and retrieve salient information needed for the current turn in the conversation. The overall objective of the CAsT benchmark being to improve understanding and research on conversational information seeking (CIS).

Interest in the development of conversational methods is motivated by analysis of the use of current search systems. These existing systems can be referred to as "single shot", since the user is required to enter a single query with the intention of retrieving documents sufficient to satisfy their information need in a single search operation. This can be observed to pose a number of challenges [4] including:

- The user must completely describe their information need in a single query.
- The user may not be able to adequately describe their information need.
- High cognitive load on the user in forming their query.

– A poorly specified query can make it difficult for the search engine to return relevant content.

While current search systems are able to perform information seeking and retrieval effectively with suitably defined queries, their ability to support CIS is very limited. The goal of the CAsT track is to promote and develop research activities in CIS and to create large-scale reusable test collections to enable empirical research in CIS. The main objectives of the CAsT track are: i) system understanding of information needs in a conversational format, and ii) finding relevant responses using contextual information. The CAsT track is motivated by complex search tasks requiring multiple turns (possibly across multiple sessions), e.g. as explored in the TREC Session tracks [1].

TASK Definition: In the CAsT track, participants are given a set of topics and their descriptions, and have to return potentially relevant responses for the given subsequent questions for each topic based on multiple turns as in a dialogue, as shown in Table 1. A response for each of the questions in the conversation needs to be generated by performing retrieval over a large collection of paragraphs to identify relevant information.

Topic	goat breeds
Description	Interested in buying goats that implies interest in different breeds of goats and their use (milk, meat and fur).
Question-1	What are the main breeds of goat?
Question-2	Tell me about boer goats?
Question-3	What breed is good for meat?
Question-4	Are angora goats good for it?
Question-5	What about boer goats?
Question-6	What are pygmies used for?
Question-7	What is the best for fiber production?
Question-8	How long do Angora goats live?
Question-9	Can you milk them?
Question-10	How many can you have per acre?
Question-11	Are they profitable?

Table 1. CAsT track sample topic

The main challenges of this track lie in understanding the query context and extracting salient information (e.g. named entities) from the dialogue. We examine some examples which illustrate these challenges of this track:

– **Query Formulation:** Building effective queries for conversational search systems is a complex task. Some of the key challenges associated with this are described below using examples presented in Table 1:
  1. **Pronoun Resolution:** *Are angora goats good for it?*, identifying *it* refers to meat (Question-4).

2. **Named Entity Identification:** *What are pygmies used for?*, identifying that 'pygmies' refers to 'goats' as a specific entity (Question-6).
3. **Query terms extraction:** *How long do Angora goats live?*, identifying important aspects and terms to formulate effective queries (Question-8). Just considering three terms, 'angora', 'goats', 'live', there can be different query representations:
'Angora', 'goats', 'live' – each word as a separate query
'Angora goats','live' – the named entity as a single unit "angora goats"
'Angora goats live' – all words as a single phrase

– **Data Fusion**: There are three different data collections used in this track (described in detail in Section 2). How to effectively combine results from these three different collections is a major challenge in the track which we investigate in our work.

We attempt to address the challenges described above to retrieve potential relevant responses for questions in a conversational search system. The two main questions that we investigate in this work are:

– **Query formulation**: How to effectively form queries from conversational topics spanning over multiple questions as turns and different dialogue utterances?
– **Data Fusion**: How to combine and re-rank documents from multiple data collections for the CAsT track?

The remainder of this paper is organised as follows: Section 2 introduces the dataset, tools used and the evaluation strategy of the CAsT track, Section 3 describes the approach adopted in our participation in this track, Section 4 gives results and analysis of our submissions to the track, and finally Section 5 concludes.

2 Dataset & Resources

In this section we outline the dataset, tools, resources used and the evaluation criteria adopted in this track.

2.1 Dataset

Three data collections are used in the CAsT track as follows:

– **MS MARCO**: Microsoft Machine Reading Comprehension (MS MARCO)[1] is a large scale dataset focused on machine reading comprehension, question answering, passage ranking, keyphrase extraction, and conversational search studies.

[1] http://www.msmarco.org

– **TREC CAR paragraph collection**: The TREC CAR paragraph collection is a corpus[2] of 20 million paragraphs. These are harvested from paragraphs on Wikipedia pages from a snapshot gathered in 2016 (with hyperlinks preserved). Given the large amount of duplication on Wikipedia pages, the collection was de-duplicated before the data release. These de-duplicated paragraphs are then available for passage ranking tasks.
– **TREC Washington Post Corpus (WAPO)**: The TREC Washington Post Corpus[3] contains 608,180 news articles and blog posts from January 2012 through to August 2017. The articles are stored in JSON format, and include title, byline, date of publication, kicker (a section header), article text broken into paragraphs and links to embedded images and multimedia.

2.2 Topics

Conversation evolves through turns (similar to dialogues) for a given topic. We were provided with training and validation topics with each topic consisting of about 10 questions (turns).

– **Training queries:** The training queries were divided into topic and turns. In total, there were 30 topics with each topic having around 7-11 turns. Each turn was associated with the previous turns as shown in Table 2. The queries have two sets of turns: one set associated with turns using the pronoun, while the second set was expanded (as shown in Table 2). The provided expanded queries were formed from the initial queries by the organisers by performing pronoun resolution and resolving ambiguous entity mentions.
– **Validation (test) queries:** The validation queries were also divided into topics and turns. In total, there were 50 topics with each topic having 7-11 turns. Similar to the training queries, we were provided with the raw query as well as the expanded query with the initial query modified after performing pronoun resolution and resolving entity mentions. We used the expanded queries for our investigations, since we did not have access to a suitable mechanism for resolution of pronouns and ambiguities available to us.

2.3 Tools and Data Processing

We used the spacy library[4] to perform query extraction and syntactic analysis of the queries (described later in Section 3). We used the whoosh library[5] to perform data indexing and passage retrieval over the indexed collection for the given input queries (described later in Section 3). We used the probabilistic Best Match (BM25 ranking model [7]) for our work.

[2] http://trec-car.cs.unh.edu/
[3] https://trec.nist.gov/data/wapost/
[4] https://spacy.io/usage
[5] https://whoosh.readthedocs.io/en/latest/

Topic	goat breeds	
Question (Q)	Initial question	Expanded question
Q-1	What are the main breeds of goat?	What are the main breeds of goat?
Q-2	Tell me about boer goats?	Tell me about boer goats?
Q-3	What breed is good for meat?	What **goat** breed is good for meat?
Q-4	Are angora goats good for it?	Are angora goats good for **meat**?
Q-5	What about boer goats?	Are boer goats **good for meat**?
Q-6	What are pygmies used for?	What are **pygmy goats** used for?
Q-7	What is the best for fiber production?	What **goat breed** is the best for fiber production?
Q-8	How long do Angora goats live?	How long do Angora goats live?
Q-9	Can you milk them?	Can you milk **Angora goats**?
Q-10	How many can you have per acre?	How many **Angora goats** can you have per acre?
Q-11	Are they profitable?	Are **Angora goats** profitable?

Table 2. Initial and Expanded queries

The Macro and Wapo datasets contain duplicate paragraphs. We followed the standard procedure provided by the TREC organizers to remove the duplication of records[6]. We remove special characters, stop words and numbers from the data and the input query. We perform stemming over the data collection and input query using Porter stemming [5].

2.4 Evaluation strategies

The CAsT track is similar to the task of passage retrieval and sentence selection, however the context of the query changes depending on the previous questions and the answers for these previous questions, as in a conversational setting. The top k passages retrieved for a given question are evaluated using two main evaluation mechanisms: $NDCG@k$ [3] and $MAP@k$.

3 Methodology

In this section we describe our system pipeline and the models explored in our work. The three main components of our pipeline are shown in Figure 1: *indexing*, *syntactic analysis* and *data fusion*, these components are described below.

3.1 Indexing

We indexed the dataset using the python whoosh API[7]. As outlined above, there were three datasets to be indexed with the schema (*Docid* and *content*). Due to the data collection size being quite large, we maintained a separate indexed object for each of the three datasets: MSMARCO, CAR and WAPO. Having three different indexes enables us to maintain data collection specific term and document statistics.

[6] http://www.treccast.ai/
[7] https://whoosh.readthedocs.io/en/latest/

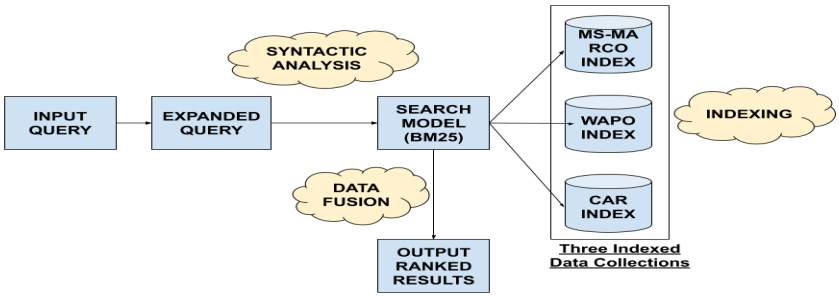

Fig. 1. System Architecture

3.2 Syntactic Analysis

As described above in Section 2, we use expanded queries provided by the track organisers in our work. We performed rich syntactic analysis of the queries. Each query was parsed and lexicon objects extracted. A combination of these lexicon objects was used to frame effective queries. The following four lexicon objects were extracted using the spacy library: *Noun*, *Noun Phrase*, *Verb Phrase* and *Adjective*. Tables 3 and 4 show input queries and corresponding lexicon objects extracted using the spacy library.

After our initial investigations with the training topics, we decided to use the eight query making models (M1−M8) listed below. These models were built using a combination of the lexicon objects. We used the "AND" boolean operator to combine the lexicon objects for a model. Based on the specificity of the query from specific to general, these eight query models were ranked from M1 to M8 based on our investigation with the training queries. While searching for a query q, we first searched the output of $M1(q)$; if there were no results, we then searched for $M2(q)$, and subsequently other model outputs $M[1-8]$ as necessary.

1. **M1:** This model was the combination of noun phrase, verb and adjective in the query
2. **M2:** This model was the combination of noun phrase and verb
3. **M3:** This model was the combination of noun phrase and adjective
4. **M4:** This model was the combination of noun, verb and adjective
5. **M5:** This model was the combination of noun and verb
6. **M6:** This model was the combination of adjective and noun
7. **M7:** This model only considered the noun phrase
8. **M8:** This model only considered the noun

ID	Text	Noun phrases
2_1	What are the main breeds of goat?	['What', 'the main breeds', 'goat']
2_2	Tell me about boer goats	['me', 'boer goats']
2_3	What goat breed is good for meat?	['What goat breed', 'meat']
2_4	Are angora goats good for meat?	['angora goats', 'meat']
2_5	Are boer goats good for meat?	['boer goats', 'meat']
2_6	What are pygmy goats used for?	['What', 'pygmy goats']
2_7	What goat breed is the best for fiber production?	['What goat breed', 'fiber production']
2_8	How long do Angora goats live?	['Angora goats']
2_9	Can you milk Angora goats?	['you', 'Angora goats']
2_10	How many Angora goats can you have per acre?	['How many Angora goats', 'you', 'acre']
2_11	Are Angora goats profitable?	['Angora goats']

Table 3. Lexicon objects-1

ID	Text	Verbs	Noun	Adjectives
2_1	What are the main breeds of goat?	['be']	['breed', 'goat']	['main']
2_2	Tell me about boer goats	['tell']	['boer', 'goat']	[]
2_3	What goat breed is good for meat?	['be']	['goat', 'breed', 'meat']	['good']
2_4	Are angora goats good for meat?	['be']	['angora', 'goat', 'meat']	['good']
2_5	Are boer goats good for meat?	['be']	['boer', 'goat', 'meat']	['good']
2_6	What are pygmy goats used for?	['be', 'use']	['pygmy', 'goat']	[]
2_7	What goat breed is the best for fiber production?	['be']	['goat', 'breed', 'fiber', 'production']	['good']
2_8	How long do Angora goats live?	['do', 'live']	['goat']	[]
2_9	Can you milk Angora goats?	['Can', 'milk']	['goat']	[]
2_10	How many Angora goats can you have per acre?	['can', 'have']	['goat', 'acre']	['many']
2_11	Are Angora goats profitable?	['be']	['goat']	['profitable']

Table 4. Lexicon objects-2

3.3 Data Fusion

We submitted four runs for the CAsT track. We used query formulation as described in Section 3.2 while performing search for potentially relevant passages. As indicated in Section 3.1, we used three separate indexes in our work. Thus for each query we retrieved three sets of ranked passages. Next, we focused on investigating different mechanisms for combining these three ranked result lists for each query. We investigated different data fusion techniques which formed our four system submissions for this track. Details of our four system submissions are as follows:

Model-1 (combination): This is the baseline run, where we performed NLP based query extraction using the spacy library to perform passage retrieval. As stated earlier, we used the officially provided expanded queries. We searched for each query using the three datasets separately using a BM25 retrieval model, and merged the retrieval results obtained for the three datasets. For merging, we performed score normalization for each ranked list and combined the ranked lists, sorting the passages in the combined list in decreasing order of relevance scores for the given topic. In cases where the scores between passages from different collections were the same, we outputted the passages in the following order of the document collection: "marco","wapo", and "car" (the collection order was chosen based on the proportion of relevant documents in the training set). Thus all the documents from a single collection such as "marco" were placed first,

then all the documents from "wapo" in cases where passages from "marco" and "wapo" had the same score.

Model-2 (datasetreorder): Similar to model-1, we performed score normalization for each ranked list and sorted the passages in a merged list in decreasing order of relevance scores for the given topic. In situations where the scores between passages from different collections were the same, we outputted the passages one at a time in the following order: "marco","wapo", and "car". Thus in this model we combined output results in a sequential manner to have mixed ranked results combined from different data collections.

Model-3 (rerankingorder): This model is similar to model-2, however we changed the document collection order to "car","wapo","marco", when combining documents from the different data collections with identical scores.

Model-4 (topicturnsort): In all of the above three models, we present the order of the document collections when combining ranked results retrieved from different collections. However, this may not be the optimum strategy. Thus in this model we performed document re-ranking using the percentage of potentially relevant documents returned. For each query we searched across all the document collections, the collection which retrieved the most results was ranked first while combining results across the three collections in cases where the scores of passages are identical. This model is different from the other three models as it is based on dynamic file ordering depending on descending order of retrieved documents.

A worked example of the four alternative data fusion strategies investigated in this work is shown in Figure 2.

4 Results and Analysis

In this section, first we discuss the nature of the *qrels* and then describe the results of our submissions to the CAsT track.

4.1 QRELs/Gold judgments

For each question within a topic, the submitted systems returned about 1000 ranked passages. A pooling technique was then adopted to generate the *qrels*. In this pooling process, the top 10 results from each different system submission were manually evaluated. Each passage was judged on a five point graded relevance scale: Fails to meet (0), Slightly meets (1), Moderately meets (2), Highly meets (3), and Fully meets (4), the information need.

Since pooling exercises are costly time wise and financially, the track organisers performed a manual evaluation exercise for 20 topics rather than the complete

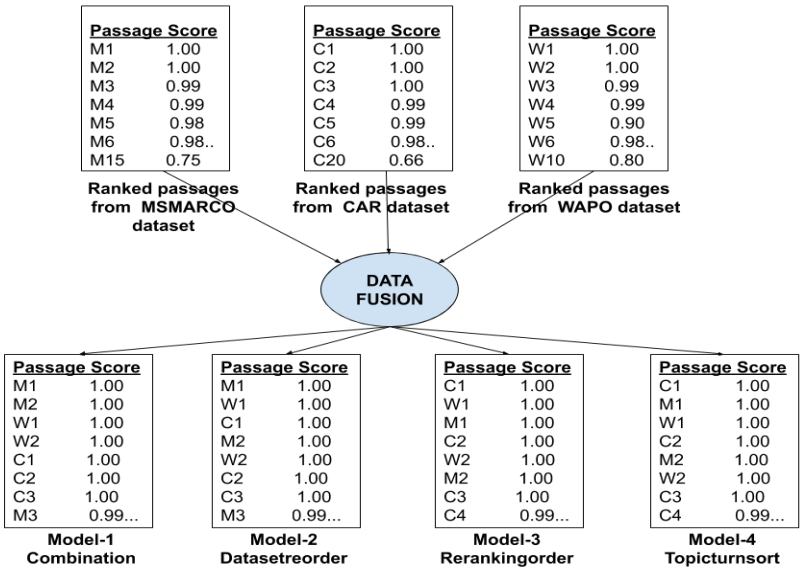

Fig. 2. Worked example of alternative Data Fusion approaches

Queries	173
Topics	20
Average turns	8.5
Number of relevant passages	8120
Total passages judged	29350

Table 5. Qrel details

50 topics provided in the validation set. Table 5 provides a basic overview of the topics for which manual evaluation was carried out.

There were issues in the de-duplication process for the WAPO dataset, hence the track organisers only considered the MS-MARCO and CAR datasets for the final evaluation of the submitted systems. The results from the WAPO dataset were omitted from the submitted system results, with the revised ranking comprising only of the Marco and CAR dataset passages for the final evaluation.

4.2 Submission results

Table 6 presents the results of our four submitted systems. For comparison, the organizers provided the results of the median system results averaged over all the turns within a topic.

All the scores of our submitted systems are below the scores of the median systems. From analysis of these results, it appears that using only syntactic information based query formulation and data fusion techniques is not that effective

	NDCG@5	MAP@5	NDCG@1000	MAP@1000
Median System	0.2960	0.0420	0.3840	0.1740
Combination	0.2481	0.0378	0.2869	0.1306
Dataset Reorder	**0.2512**	0.0360	0.2923	0.1356
Reranking Order	0.2488	0.0367	0.2937	0.1379
Topic Turn Sort	0.2427	0.0357	0.2926	0.1367

Table 6. Results for our submissions with Median System results for comparison

for retrieving potentially relevant passages for conversational search. We carried out further investigation using our system's datasetreorder run which obtained the highest NDCG@5 scores. Table 7 presents topic wise results for retrieving and ranking passages for conversational search systems. The results are quite varied across the different topics. This shows that our systems do well for some topics and not so well for others. Note that these results are averaged across all the turns.

Topic-ID	NDCG@5	MAP@5	NDCG@1000	MAP@1000
31	0.1555	0.0142	0.1149	0.0531
32	0.0827	0.0119	0.1079	0.0482
33	0.0165	0.0041	0.0481	0.0202
34	0.0627	0.0118	0.0605	0.0257
37	0.0480	0.0091	0.1010	0.0528
40	0.0588	0.0106	0.0608	0.0198
49	0.0319	0.0052	0.0506	0.0238
50	0.1059	0.0113	0.1184	0.0687
54	0.0255	0.0055	0.0466	0.0126
56	0.0814	0.0075	0.0761	0.0395
58	0.1033	0.0101	0.0903	0.0272
59	0.0761	0.0079	0.0748	0.0307
61	0.0309	0.0046	0.0249	0.0096
67	0.0764	0.0055	0.0763	0.0198
68	0.0710	0.0116	0.0799	0.0474
69	0.0894	0.0162	0.1225	0.0578
75	0.0264	0.0052	0.0698	0.0177
77	0.0876	0.0205	0.1290	0.1077
78	0.0403	0.0142	0.0692	0.0278
79	0.0768	0.0081	0.0580	0.0254
all	0.2512	0.0360	0.2923	0.1356

Table 7. Topic based results for the datasetreorder run

As the results in Table 7 are averaged across different turns, we were motivated to calculate the retrieval scores averaged across all topics but for different turns. We hypothesize that:*As we go deeper in conversation it should be more difficult to understand and hence answer the question successfully.* Table 8 presents

results averaged across all the turns. The nature of topics can vary significantly, further as the number of turns is not consistent across topics and varies considerably, linearly averaging NDCG scores across topics does not seem to be appropriate. We do not observe a clear trend to support our hypothesis that as we go deeper it becomes harder to understand and answer questions successfully. We plan to investigate these variations of results as shown in Tables 7 and 8, across different topics and turns in detail in future work.

Turns	NDCG@5	Topics
1	0.2115	20
2	0.2176	20
3	0.2648	20
4	0.3532	20
5	0.2243	20
6	0.2298	20
7	0.2620	19
8	0.1969	20
9	0.1925	7
10	0.3937	4
11	0.5299	3

Table 8. Turnwise results for datasetreorder run

5 Conclusions

In this report we describe our submissions for the CAsT track at TREC 2019. We focus on understanding of information needs in a conversational format to find relevant responses using contextual information. We investigate query formulation and data fusion techniques. Our results show that our query formulation and data fusion techniques attain average results, which are not sufficient to answer questions in a conversational setting effectively. We plan to study the variation in results across different topics and across different turns to understand the challenges to support conversational search systems better.

We anticipate a need to leverage effective semantic representation of queries, context and the passages to capture relevant information effectively. In future, we plan to explore recent work using deep learning models such as BERT [2], XLNET [8] and OPEN-GPT-2 model [6] for the task of ranking passages for conversational search systems as they have shown good results in the benchmarking tasks on question-answer ranking and machine reading comprehension [2, 6, 8].

Acknowledgement

This research is supported by Science Foundation Ireland (SFI) as a part of the ADAPT Centre at Dublin City University (Grant No: 12/CE/I2267).

References

[1] Ben Carterette et al. "Evaluating Retrieval over Sessions: The TREC Session Track 2011-2014". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. Pisa, Italy, 2016, pp. 685–688.

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[3] Kalervo Järvelin and Jaana Kekäläinen. "Cumulated gain-based evaluation of IR techniques". In: *ACM Transactions on Information Systems (TOIS)* 20.4 (2002), pp. 422–446.

[4] Abhishek Kaushik. "Dialogue-Based Information Retrieval". In: *Advances in Information Retrieval*. Ed. by Leif Azzopardi et al. Cham: Springer International Publishing, 2019, pp. 364–368. ISBN: 978-3-030-15719-7.

[5] Martin F Porter. "An algorithm for suffix stripping". In: *Program* (2006).

[6] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI Blog* 1.8 (2019).

[7] S Robertson et al. "Okapi at TREC-3". In: *NIST special publication* 500225 (1995), pp. 109–123.

[8] Zhilin Yang et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *CoRR* abs/1906.08237 (2019). arXiv: 1906.08237. URL: http://arxiv.org/abs/1906.08237.