# CAsT-19: A Dataset for Conversational Information Seeking

Jeffrey Dalton
University of Glasgow
jeff.dalton@glasgow.ac.uk

Chenyan Xiong
Microsoft Research AI
chenyan.xiong@microsoft.com

Vaibhav Kumar
Carnegie Mellon University
vaibhav2@andrew.cmu.edu

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

## ABSTRACT

CAsT-19 is a new dataset that supports research on conversational information seeking. The corpus is 38,426,252 passages from the TREC Complex Answer Retrieval (CAR) and Microsoft MAchine Reading COmprehension (MARCO) datasets. Eighty information seeking dialogues (30 train, 50 test) are an average of 9 to 10 questions long. A dialogue may explore a topic broadly or drill down into subtopics. Questions contain ellipsis, implied context, mild topic shifts, and other characteristics of human conversation that may prevent them from being understood in isolation. Relevance assessments are provided for 30 training topics and 20 test topics.

CAsT-19 promotes research on conversational information seeking by defining it as a task in which effective passage selection requires understanding a question's context (the dialogue history). It focuses attention on user modeling, analysis of prior retrieval results, transformation of questions into effective queries, and other topics that have been difficult to study with existing datasets.

## KEYWORDS

Conversational information seeking; conversational search; dataset

**ACM Reference Format:**
Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. CAsT-19: A Dataset for Conversational Information Seeking. In *43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

The importance of conversation and conversational models for complex information seeking tasks is well-established within information retrieval, initially to understand user behavior during interactive search [5, 7] and later to improve search accuracy during search sessions [1]. The recent popularity of a new generation of *conversational assistants* such as Alexa, Siri, Cortana, Bixby, and Google Assistant increase the scope and importance of conversational approaches to information seeking and also introduce a broad range of new research problems [2].


*SIGIR '20, July 25–30, 2020, Virtual Event, China*

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/1122445.1122456

Research on conversational models of information seeking has been hampered by the lack of reusable datasets. Prior research studied a small number of people (e.g., 10-15) [5, 7] searching specialized corpora; or a large number of anonymous people represented in commercial search engine logs [1]. These two approaches have the advantage of authenticity because they are based on data from real information seeking sessions. However, results from small studies are often difficult to generalize, reuse, or reproduce, whereas commercial search data is difficult to obtain.

This paper presents CAsT-19 a new dataset developed for the TREC 2019 Conversational Assistance Track (CAsT) [3]. Each *topic* is a model of a person using a conversational assistant to explore or learn about a subject. A topic is a series of questions that collectively represent a complex information need that cannot be satisfied by a simple answer or single response. Topics have typical conversational artifacts, for example, ellipsis, anaphora, and implied context, and typical conversational structure, for example, drilling down into a topic, exploring a topic broadly, shifting focus, and context switching. Each response is restricted to be a short passage that could be delivered by a conversational assistant or passed to downstream language processing tasks.

CAsT-19 was designed to satisfy several goals. One goal was a low barrier to entry, thus complex information seeking is modeled as a fixed sequence of related passage ranking tasks and questions are expressed as text. The dataset focuses research on system understanding of information needs in a conversational format and on leveraging the conversational context to retrieve relevant passages. A second goal was that CAsT-19 support research on conversational information seeking for ten or more years. As a result, it was developed to include greater language complexity, including entailment and implicature, and to have topics with non-trivial discourse structure moving towards long-form dialogues. In the following sections we formalize the task, evaluation method including assessment guidelines, and related resources created.

## 2 RELATED WORK

As interest has grown in developing CIS systems, attention has turned to developing resources for evaluating CIS systems. Thomas et al. [8] introduced the Microsoft Information Seeking Conversation (MISC) dataset, which contains recorded transcripts of information seeking dialogues between people. Vtyurina et al. [9] introduced the MSDialog dataset, which contains more than 2,000 multi-turn dialogues annotated with user intent.

CIS systems are just beginning to emerge. Yang et al. [10] present a recent system that uses a neural matching network, pseudo relevance feedback, and neural question answering techniques to

**Table 1: CAsT-19 Training Topic 18.**

| **Title**: Uranus and Neptune | |
|---|---|
| **Description**: Information about Uranus and Neptune. | |
| **Turn** | **Conversation Utterances** |
| 1 | Describe Uranus. |
| 2 | What makes it so unusual? |
| 3 | Tell me about its orbit. |
| 4 | Why is it tilted? |
| 5 | How is its rotation different from other planets? |
| 6 | What is peculiar about its seasons? |
| 7 | Are there any other planets similar to it? |
| 8 | Describe the characteristics of Neptune. |
| 9 | Why is it important to our solar system? |
| 10 | How are these two planets similar to each other? |
| 11 | Can life exist on either of them? |

rank passages by their likelihood of answering questions from three CIS corpora. They found that each question improved accuracy in the early part of the conversation, but harmed accuracy later in the conversation due to the difficulty of managing lengthy contexts.

## 3 TASK DESCRIPTION

CAsT-19 defines conversational search as an information retrieval task in a conversational context. The goal of the task is to satisfy a person's information need, which is *expressed or formalized through turns in a conversation.* The response from the retrieval system is not a list of documents. Instead, it is a list of *brief text passages* (e.g., each ideally a few sentences) suitable for presentation in a voice-interface or on a mobile screen.

**Task Definition.** CAsT-19 was designed to produce a reusable dataset and be an easy entrance to research on conversational information seeking. It focuses on retrieving candidate responses for information seeking conversations. Given a natural language question $q_i$ preceded by questions $q_1, \ldots, q_{i-1}$, return a ranking of brief passages $p_{i,1}, \ldots, p_{i,n}$ that could be presented to a person or passed to downstream language processing tasks.

**Information Needs.** Exploratory information needs (*topics*) were constructed from previous TREC topics (e.g., Common Core and Session Track), MS MARCO Conversational Sessions (Section 5), and topics of general interest. The information needs were designed to be complex (require multiple rounds of elaboration), diverse (different information categories), open-domain (not requiring expert domain knowledge to access), and answerable (sufficient coverage in the collection). Topics were designed to be informational (not tasks), not require temporal or external context, not include personal or subjective decisions, avoid sensitive or controversial topics, be broad enough to support a meaningful trajectory, not be niche (i.e., be of general interest), and not be too broad.

**Conversational Sequences**. Questions were created manually for each turn $t$ in a topic. In general, they start with a general introduction of the topic and shift to exploratory information seeking trajectories. To make the corpus reusable, later turns only depend on the previous utterances, not on system responses.

When curating the conversational trajectories, multiple sources of information were used. The MS MARCO search session data was one input. Query suggestions from commercial search engines (Google and Bing) and specifically the natural language questions from the "People Also Ask" feature in Google and Bing were also used. These questions are similar to the questions released in the widely used Google Natural Language Questions dataset [6].

**Turn Guidelines.** The conversational sequences were written to mimic features of "real" dialogues, using the following guidelines.

- Topics will have topically coherent transitions between questions.
- Topics will have common natural language features including coreference, elipsis, and omissions.
- Topics will have subtopics that can be traversed breadth-first ('explore'), depth-first ('drill down'), or with other strategies.
- Some topics will have comparisons across subtopics.
- Most topic turns will require more than a short answer response (i.e., a simple fact won't suffice).
- For complexity, most turns will be contextually dependent on previous turns.

Topic questions were developed using the following guidelines.

- Questions will be locally and globally coherent.
- Questions will use well-formed natural language.
- Questions will not depend upon a prior system response.
- Questions will have several relevant passages in the top 20 of an Indri ranking, but fewer than ten.

An example topic from the CAsT-19 training set is shown in Table 1.

Topics were developed by two teams working independently. One person wrote the initial topic. At least one other person reviewed it and suggested revisions. The revision cycle repeated until everyone was satisfied. Each topic was reviewed by two or three people that were not its initial author.

**Passage Collection.** The corpus is passages from MS MARCO[1] and the TREC Complex Answer Retrieval Paragraph Collection [4].

The TREC CAR (Wikipedia) paragraphCorpus V2.0 consists of all paragraphs from Wikipedia '16. Note that this corpus has been deduplicated. It contains approximately 30 million unique paragraphs. Dietz et al. [4] provide more details about this corpus.

MS MARCO has 1M real search queries, each with 10 passages from top ranked results, yielding a corpus of approximately 8 million passages. Passage metadata includes the source URL, the MARCO queries associated with it, and relevance labels for adhoc passage retrieval. The MARCO collection contains near duplicates.

CAsT-19 was intended to include a de-duplicated, passage-based version of the Washington Post (WAPO) collection, but errors in creating deduplicated WAPO passages produced ambiguous passage ids. Prior to assessing, WAPO passage ids were filtered out of the 21 runs used to construct relevance assessing pools; these were less than 5% of results returned by the 21 systems used for pooling. Thus, CAsT-19 does not include WAPO passages.

## 4 RELEVANCE ASSESSMENTS

The following section describes the judgment criteria, labeling process, and evaluation metrics.

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

**Table 2: Judgment statistics**

| | |
|---|---|
| Topics | 20 |
| Turns | 173 |
| Assessments | 29,571 |
| Fails to meet (0) | 21,451 |
| Slightly meets (1) | 2,889 |
| Moderately meets (2) | 2,157 |
| Highly meets (3) | 1,456 |
| Fully meets (4) | 1,618 |

## 4.1 Guidelines

Passage assessment was similar to relevance assessment in other TREC settings. However, the conversational setting introduces several unique issues.

(1) *Contextualized*: The meaning of a turn and the relevance of an answer passage may depend on preceding turns in the same conversation. For example, "What is throat cancer?" followed by "Is it treatable?" Each question must be interpreted in the context established by the preceding conversation.
(2) *Coreference and omission*: As with most human conversations, many CAsT-19 turns have some form of ellipsis, for example, pronouns and implied context that omits words that can be understood from the preceding context. To aid assessment, CAsT-19 provides *resolved* versions of each turn. For example, the resolved version of "Is it treatable?" is "Is throat cancer treatable?".
(3) *Brevity and completeness*: Conversational assistants interact with people via spoken or chat interfaces that are designed for brief responses. Answer passages in the CAsT-19 corpus tend to be short. A good system will select passages that provide a complete answer in a concise response.

The relevance standard for a [turn, passage] pair is intended to represent how a person would feel if she asked the question to her favorite conversational assistant (e.g., Siri, Cortana, or Alexa) and it responded with the text in the passage. A five-point relevance scale from the Google Needs Met Rating Guideline[2] was adapted for the CAsT task with the following definitions.

(1) *Fully meets (4).* The passage is a perfect answer for the turn. It includes all of the information needed to fully answer the turn in the conversation context. It focuses only on the subject and contains little extra information.
(2) *Highly meets (3).* The passage answers the question and is focused on the turn. It would be a satisfactory answer if Google Assistant or Alexa returned this passage in response to the query. It may contain limited extraneous information.
(3) *Moderately meets (2).* The passage answers the turn, but is focused on other information that is unrelated to the question. The passage may contain the answer, but users will need extra effort to pick the correct portion. The passage may be relevant, but it may only partially answer the turn, missing a small aspect of the context.
(4) *Slightly meets (1).* The passage includes some information about the turn, but does not directly answer it. Users will find some useful information in the passage that may lead to the correct answer, perhaps after additional rounds of conversation (better than nothing).
(5) *Fails to meet (0).* The passage is not relevant to the question. The passage is unrelated to the target query.

[2]https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

## 4.2 Process

CAsT-19 relevance assessing used NIST assessors and standard NIST practices. Fifty topics, an average topic length of 9.6 questions, and 21 participating teams produced an expensive assessing task. Resource constraints limited NIST assessing to 20 randomly sampled test topics, in most cases to a depth of 8 questions.

**Pooling.** Twenty one participating teams in the TREC 2019 CAsT track each submitted up to four runs. The two highest priority rankings from each team were pooled to depth 10. The pool also includes two manual runs from the track organizers, which adds about thirty unique passages to the pool. The total pool size for all turns for the 20 assessed topics is 33,614 unique paragraphs.

**Judging.** Relevance assessing was done by NIST assessors during a three-week period. Six assessors each worked approximately 50 hours. The average labeling speed was about 100 minutes per turn, or about 35 second per passage. Assessors were provided with the raw question and also the CAsT-19 manually re-written ("resolved") question. The latter contains full information to define the passage relevance without depending on previous rounds. Assessors evaluated one topic at time in order of the turns. Thus, the conversational context was preserved in the labeling process.

173 conversational turns were judged. Table 2 shows the distribution of relevance labels. On average there are 170 unique passages per turn. Each has on average 47 passages with non-zero relevance score. Turn 75_7 had no relevant passages in the assessment pool. Topic 31_3 had all assessed passages at least slightly relevant.

Conversational information seeking poses an expensive assessing problem. Longer dialogues enable varied conversational structure, but require many more assessments. The 50 CAsT-19 test topics contain 479 questions. Evaluating all of them requires crowdworkers. Such an effort is underway, but not yet complete.

# 5 RESOURCES

In addition to the standard corpus, test topics, and assessments that make up most datasets, CAsT-19 includes several data and software resources that support use of the dataset.

**Training data.** CAsT-19 includes 30 training topics. Five of these topics have manually created relevance assessments developed by one of the authors (approximately 50 turns). Relevance assessment was performed on a three-point relevance scale.

**External data.** Building on MARCO and TREC CAR collections allows CAsT-19 to share the queries and relevance assessments developed for these datasets. These labels can be used to train models of single turn relevance.

The Conversational Search extension to the MS MARCO dataset is publicly-available information seeking session data [3]. Some of the CAsT-19 topics are based on these sessions.

[3]https://github.com/microsoft/MSMARCO-Conversational-Search

**Table 3: Frequency of four types of entity mention in CAsT-19 topics. Counts are based on manual categorization.**

| Type | Train | Test |
|---|---|---|
| Pronominal | 102 | 128 |
| Zero | 82 | 111 |
| Groups | 6 | 4 |
| Abbreviations | 29 | 15 |

**Resolved topics.** CAsT-19 dialogues include conversational phenomena such as coreference and omission. *Resolved* topics are manually-written versions that enable each question to be understood without reference to preceding conversational context.

Each question was rewritten by two CAsT-19 developers. Disagreements were resolved through discussion. On average it took 5-10 minutes to rewrite a topic (ten turns on average), and (minor) disagreements among developers was common, indicating that this is non-trivial even for those familiar with the topics.

**Entity mention annotations.** Training and test topics were manually annotated to identify and resolve entity mentions ([*mention string*, *canonical entity string*]). These annotations facilitate research on entity-based query representations. The mentions are categorized and the results shown in Table 3.

**Passage collection deduplication.** Preliminary exploration revealed that the MARCO corpus contains a significant number of near-duplicate documents. Near-duplicate detection was performed to cluster results and identify the canonical reference id per duplicate group. Only one result per duplicate group was evaluated by assessors. The recommendation is to remove duplicates from the corpus, keeping only the canonical document.

The near-duplication algorithm grouped passages based on their URLs. Within each URL group, passages were first sorted by their length in descending order. Pairwise matches between passages in the group were identified. A pairwise match is defined as the total percentage of matching words in the smaller passage with respect to the longer passage in the input pair. If this percentage match was greater than 95%, then the id of the smaller passage was added to the near-duplicate dictionary. The ids in the final duplicate dictionary were then mapped back to the MS-MARCO Ranking corpus ids (based on a prior alignment of ids between the two corpora).

**Software tools** Publicly available software tools were created to process the CAsT-19 data[4]. Scripts were provided to convert to standard TREC formats. For handling the topics, there are tools and sample code to load the conversation topics in both Python and Java. The topic files are available in multiple formats including JSON, text, and Google protocol buffers. The protocol buffer format is the canonical representation.

## 6 CONCLUSION

CAsT-19 is one of the first attempts to produce a reusable dataset to support research on conversational information seeking. It has realistic conversational structure and information seeking behavior while presenting a low barrier to entry for IR researchers just starting to study conversational information seeking. The development of CAsT-19 reveals much about the structure of conversational search, open research problems, and issues that arise when creating evaluation resources for this research area.

Existing, off-the-shelf co-reference resolution software struggled with CAsT-19 topics more than might be expected. Results using the manually resolved queries demonstrate that there is room for up to 35% improvement using manually rewritten queries over even the best automatic system [3]. These results may focus more attention on resolving co-reference problems within conversational dialogues, as opposed to the more typical narrative text setting.

We found that longer dialogues support varied conversational structure (elipsis, omission, context switch) and varied information seeking behavior (e.g., drilling down, exploring broadly, popping the stack to pursue a prior issue). Participants in the TREC 2019 CAsT track focused more on coping with conversational structure and less on tracking varied information seeking behavior, which caused accuracy for many systems to degrade later in the conversation [3]. This outcome shows the value of longer dialogues. However, longer dialogues increase assessing costs.

CAsT-19 uses static conversation sequences in which the next question is based on the seeker's interests, not the system's prior response. This choice makes the dataset reusable, but prevents study of some problems. Participants in the TREC 2019 CAsT planning workshop felt that reusability was more important than dynamic conversational structure, and recommended retaining this characteristic for the next dataset.

[4] http://treccast.ai/

## REFERENCES


[1] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. 2015. Overview of the TREC 2014 Session Track. In *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*. NIST Special Publication 500-308.

[2] J. Shane Culpepper, Fernando Diaz, and Mark Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (2018), 34–90.

[3] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2019: The Conversational Assistance Track Overview. In *The Twenty-Eighth Text Retrieval Conference Proceedings*. NIST Special Publication 1250.

[4] Laura Dietz, Ben Gamari, Jeff Dalton, and Nick Craswell. 2019. TREC Complex Answer Retrieval Overview. In *The Twenty-Seventh Text REtrieval Conference Proceedings (TREC 2018)*. NIST Special Publication 500-331.

[5] Nicholas J.Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications* 9, 3 (1995), 379–395.

[6] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics* 7 (2019), 453–466. https://tomkwiat.users.x20web.corp.google.com/papers/natural-questions/main-1455-kwiatkowski.pdf

[7] Paul Solomon. 1997. Conversation in information-seeking contexts: A test of an analytical framework. *Library & Information Science Research* 19, 3 (1997), 217–248.

[8] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *Proceedings of the International Workshop on Conversational Approaches to Information Retrieval*.

[9] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2187–2193.

[10] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *Proceeding of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 245–254.