Conversational Search Systems

[DAT640] Information Retrieval and Text Mining

Weronika Lajewska
University of Stavanger

31.10.2022

In this module

1. Conversational Search Systems
2. Mixed-initiative

Conversational Search Systems

Recap - conversational information access

- A subset of conversational AI systems that specifically aim at a task-oriented sequence of exchanges
- Supports multiple user goals, including search, recommendation and exploratory information gathering
- Requires multi-step interactions over possibly multiple modalities
- Combine elements from both task-oriented and interactive QA systems
- Consider both long-term and short-term information about the user when solving information seeking tasks

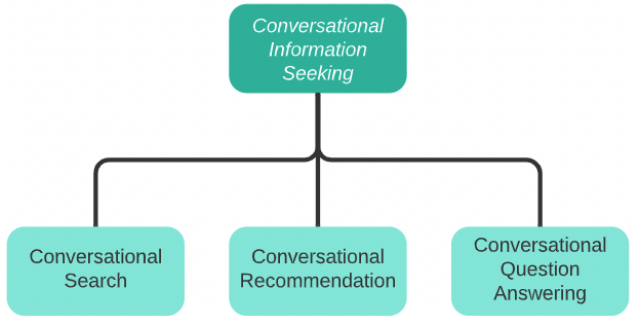
Recap - conversational information access

- Chatbot - system that mimics the unstructured conversations characteristic of informal human-human interaction
- Task-oriented dialogue system - uses conversation with users to help complete task. It can answer questions, give directions, control appliances, find restaurants, or make calls.
  - Conversational recommender system - system that can elicit the dynamic preferences of users and take actions based on their current needs through real-time multiturn interactions.

Conversational information seeking[1]

Definition

A *Conversational Information Seeking (CIS)* system is a system that satisfies the information needs of one or more users by engaging in information seeking conversations



[1] Conversational information seeking `https://arxiv.org/pdf/2201.08808.pdf`

Conversational Search System

Definition

A *conversational search system (CSS)* is a system for retrieving information that permits a mixed-initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user. The responses are expected to be concise, fluent, stateful, mixed-initiative, context-aware, and personalized.

CSS - properties

- User revealment - the system helps the user express (potentially discover) their true information need, and possibly also long-term preferences
- System revealment - the system reveals to the user its capabilities and corpus, building the user's expectations of what it can and cannot do
- Mixed initiative - the system and user both can take initiative as appropriate
- Memory - the user can reference past statements, which implicitly also remain true unless contradicted
- Set retrieval - the system can work with, manipulate and explain the sets of options/objects which are retrieved given the conversational context

CSS - additional challenges

- Filtering out superfluous, e.g, fillers, pauses, false starts
- Answer aggregation by presenting a summary of the retrieved list of results
- General knowledge about external world
- Recovering from communication breakdowns
- Understanding and reasoning about user limitations, e.g., cognitive abilities and styles, domain knowledge

Question

When is the conversational aspect of the search system needed?

Complex scenarios that require CSS

- Faceted elicitation - searching for an item with rich attributes that can be individually specified, but are much simpler to provide piecewise. As part of the search, the user is identifying aspects that can be used to describe a relevant item.
- Multi-item elicitation - searching for a single item supported by a set of nearby items, which requires estimating the relevance of the whole set of items
- Multi-item faceted elicitation - searching for a set of items directly. Not only must the system estimate the utility of each single item, it must combine the utilities of multiple items to reach an assessment of an entire set.
- Bounding choices - providing the user with precise choices rather than expecting them to come up with particular terms. It simplifies the problem of need elicitation.

Conversational Search Systems

- TREC CAsT
- Query Rewriting
- State of the art CSS

TREC CAsT - main goal

- Conversational Assistance Track (CAsT) is a part of the Text REtrieval Conference (TREC) since 2019
- It aims to advance Conversational Information Seeking research and to create a large-scale reusable test collection for CSSs
- The track addresses conversational search that is about to satisfy a user's information need expressed or formalized through multiple turns in a conversation
- The desired response is not a list of relevant documents, but rather a brief passage of a maximum of 3 sentences in length
- The topics released every year are inspired by sessions in web search and triggered to be more challenging for CSSs

TREC CAsT - differences between editions

- In TREC CAsT'19, user utterances may only refer to the information mentioned in previous user utterances
- Since 2020, utterances may refer to previous responses given by the system as well, which significantly extends the scope of contextual information that the system needs to use to understand a request
- TREC CAsT'21 is characterized by the increased dependence on previous system responses, as well as simple forms of user revealment, reformulation, topic shifts, sequence reference, and explicit feedback introduced in users' utterances
- In year 1 and year 2, the systems were using passage collections, while in year 3 the retrieval corpus from which the system response is chosen is over documents, with passages returned, which is a more realistic setting

TREC CAsT - topic[2]

Title: Steroid use in US sports **Description**: The history of steroid use in US sports.	
Turn	**Conversation Utterances**
1	What's the history of steroid use in sports in the US?
2	What were Ziegler's improvements?
3	Why are they banned?
4	Are there visible signs?
5	That sounds easy to spot. How do they get away with it?
6	What is the NFL policy?
7	Isn't that speed?
8	What is the difference between the two policies?
9	I heard it even affects card players. Didn't bridge also have a problem?
10	I know what bridge is. I heard there was a drug scandal recently.
11	Does the article have more about it?

[2]TREC CAsT 2021: The Conversational Assistance Track Overview

TREC CAsT - evaluation

- The evaluation of TREC CAsT tasks takes into consideration two dimensions of the ranking evaluation:
  - the ranking depth focused on the earlier positions (3,5) for the conversational scenario
  - the turn depth focused on deeper rounds to capture the ability of the system to understand the context of the whole conversation
- The main evaluation metric is the mean NDCG@3 with all conversation rounds averaged using uniform weights
- Additionally, the Recall@1000, MAP and MRR are calculated to evaluate each system

TREC CAsT - pooling

- The assessment pools are formed using the top ten passages from up to four runs per group
- The systems taken into consideration in the pooling need to be intrinsically different in order to achieve higher diversity
- Exhaustive pools let us assume that whatever is not pooled is not relevant
- Participants are asked to prioritize the submitted runs because not all the systems that appear in the competition are pooled

CAsT systems architecture

- There is an established two-step passage ranking architecture
- The first-pass passage retrieval is usually performed using standard unsupervised IR techniques, which is followed by re-ranking using a neural model
- Additionally, most of the systems are using a query rewriting component, where the original query is de-contextualized using a neural model to be independent of the previous turns

TREC CAsT - system architecture - cont.

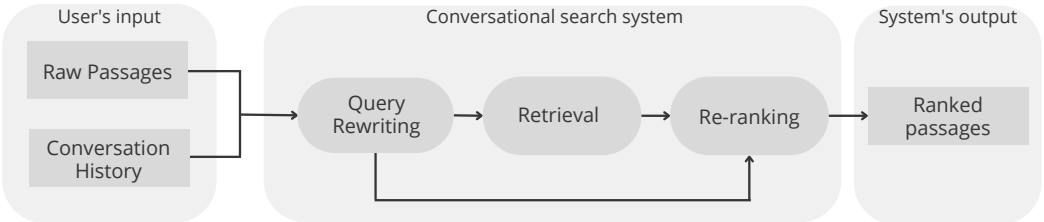


Figure: Cascade two-stage pipeline architecture with query rewriting module.

Conversational Search Systems

- TREC CAsT
- Query Rewriting
- State of the art CSS

Query Rewriting

Definition

The goal of *Conversational Query Rewriting (CQR)* is to produce an informative stand-alone, de-contextualized query for each raw query. Specifically, given a conversational history $H = [q\_1, r\_1, q\_2, r\_2, ..., q\_i - 1, r\_i - 1]$, where r_k is a response provided by the system for kth query q_k and q_i is a current raw query, the task of CQR consists of the filtering out unnecessary information in H and generating the de-contextualized query $\hat{q}\_i$.

Query Rewriting - example

Raw query	Rewrite
I remember there was a global coffee shortage some time ago. Can you tell me more about that?	Can you tell me more about the global coffee shortage?
What caused the drought?	What caused the drought in Brazil last month?
I also heard that other Latin American countries had coffee production issues. Was the disruption widespread?	Was the disruption of coffee production in other Latin American countries widespread?

Table: Examples of query rewrites generated for utterances from TREC CAsT 2021 dataset.

Query Rewriting - feature-based methods

- Unsupervised feature-based methods - expanding an original query by words chosen from conversation history or additional metadata provided by the organizers (e.g., the title of the session)
- Supervised feature-based methods use linguistic features based on dependency parsing, coreference resolution, named entity resolution, or part-of-speech tagging

Query Rewriting - feature-based method example[3]

- Historical query expansion - term importance estimation through frequency-based signals
- Intuition:
  - Observation #1: A session is centered around a main topic, and the turns in the session dive deeper into several subtopics, each of which only lasts a few turns
  - Observation #2: The degree of ambiguity divides utterances into three categories: The first category includes utterances with clear intents, which can be directly treated as ad-hoc queries. The second category contains those starting a subtopic, and the last category is composed of ambiguous utterances that continue a subtopic

[3] Multi-Stage Conversational Passage Retrieval:An Approach to Fusing Term Importance Estimation and Neural Query Rewriting

Query Rewriting - feature-based method example[4] - cont.

Main steps:

- extracting the main topic and subtopic keywords from the utterance
- measuring the ambiguity of the utterance
- expanding queries for the ambiguous utterances with the main topic and subtopic keywords extracted from previous turns

Title: Career choice for Nursing and Physician's Assistant	
Turn (i)	Conversation utterances (u_i)
1	What is a physician's assistant?
2	What are the educational requirements required to become one?
3	What does it cost?
4	What's the average starting salary in the UK?
5	What about in the US?
6	What school subjects are needed to become a registered nurse?
7	What is the PA average salary vs an RN?
8	What the difference between a PA and a nurse practitioner?
9	Do NPs or PAs make more?
10	Is a PA above a NP?
11	What is the fastest way to become a NP?
12	How much longer does it take to become a doctor after being an NP?

[4] Multi-Stage Conversational Passage Retrieval:An Approach to Fusing Term Importance Estimation and Neural Query Rewriting

Query Rewriting - supervised methods

- Supervised neural query rewriting approaches are characterized by utilizing large pre-trained language models
- In particular, generative models such as GPT-2 model, or the T5 model are used
- They are mostly fine-tuned on the CANARD dataset or QReCC
- In some cases, generated query reformulation is further expanded with terms chosen from conversation history, with its own paraphrase, or the related topic sentences from the semantically-associated documents

Conversational Search Systems

- TREC CAsT
- Query Rewriting
- State of the art CSS

State of the art CSS - main components

- Inverted index and ANN index
- Query rewriting:
  - T5 model fine-tuned on the QReCC dataset
- First-pass retrieval:
  - Sparse retrieval with BM25 on queries extended with pseudo-relevance feedback
  - Dense retrieval with ANCE dense retriever
  - Final ranking is a reciprocal rank fusion of rankings returned by sparse and dense retrievers
- Re-ranking:
  - pointwise re-ranking with monoT5
  - pairwise re-ranking with duoT5

State of the art CSS - schema of architecture

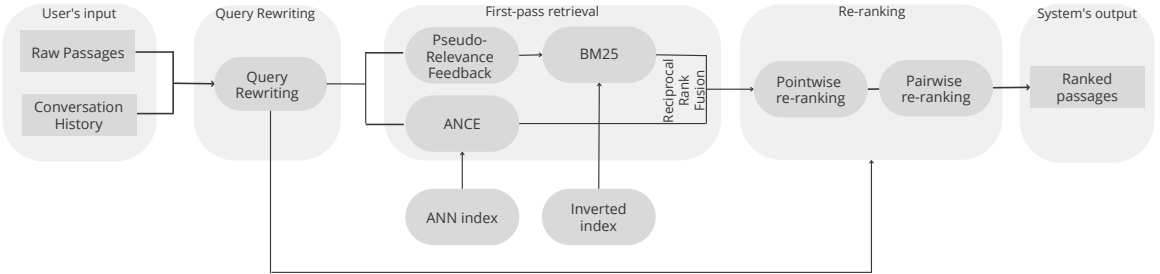


Figure: State of the art system from TREC CAsT'21

State of the art CSS - reproducibility[5]

- The reproducibility study aims to check whether the measurement can be obtained with:
  - stated precision by a different team using the same measurement procedure
  - the same measuring system
  - under the same operating conditions
  - in the same or a different location on multiple trials
- For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts

[5] https://www.acm.org/publications/policies/artifact-review-badging

State of the art CSS - key challenges in reproducibility

- TREC systems are commonly regarded as reference points for effectiveness comparison
- TREC CAsT is a competition, so the top-performing teams are not willing to share their code with the community
- Top performing systems are very complex, they use many parameters and usually several stages
- The end-to-end performance of the system may not explicitly indicate which component is causing a drop in effectiveness
- Intermediate files, such as file rewrites and first-pass retrieval rankings are usually not shared, which makes component-based analysis impossible

Mixed-initiative

Mixed-initiative

- Mixed-initiative interaction - a flexible interaction strategy in which each agent (human or computer) contributes what it is best suited at the most appropriate time[6]
- Mixed-initiative systems can take control of the communication either at the dialogue level (e.g., by asking for clarification or requesting elaboration) or at the task level (e.g., by suggesting alternative courses of action)
- Mixed-initiative is more human-like and flexible because both actors can independently contribute to the conversation
- Main goal: improving the passage ranking effectiveness by allowing the systems to take the initiative at any point in a conversation

[6] Principles of Mixed-Initiative User Interfaces, https://dl.acm.org/doi/pdf/10.1145/302979.303030

Mixed-initiative at TREC CAsT

- In the mixed-initiative sub-task at TREC CAst, the system can pose questions to the user to gain additional context for a turn
- The supported questions types for the task are: (1) elicit the task, (2) ask for feedback, or (3) clarify ambiguity

Example

User: What are some cool things to do in California?
MI-System: California is very large, what area would you like to visit?
MI-User: I'd like to explore Northern California.

System revealment

- System revealment - the system allowing the user to learn about the system's abilities, building the user's expectations of what it can and cannot do
- System revealment in terms of its confidence in the provided response:
  - the information need is not clear and the system is not able to find the answer $\rightarrow$asking clarifying questions
  - the information need is clear to the system but the question is unanwserable in the collection used $\rightarrow$handling ananswerable question
- Main goals: CSS's explainability and transparency

Clarifying questions

- Search queries are often short, and the underlying user intent may be ambiguous
- This makes it challenging for search engines to predict possible intents, only one of which may pertain to the current user
- To address this issue, search engines often diversify the result list and present documents relevant to multiple intents of the query
- Clarifying questions can be used by the system to resolve ambiguity, to prevent potential errors, and in general to clarify user's requests and response
- Main challenge: trade-off between the efficiency of the conversation and the accuracy of the information need as the system has to decide between how important it is to clarify and how risky it is to infer or impute the underspecified or missing details

Demo

E13-1 - Clarifying questions

Unanswerable questions

- Open challenges in the conversational search:
  - Detecting questions for which no good answer exists in a corpus and informing the user that the answer has not been found
  - Detecting questions for which the answer is partially present and identifying the missing part
- Datasets containing unaswerable questions:
  - SQuAD 2.0 dataset created for extractive reading comprehension tasks
  - QuAC dataset containing information-seeking dialogs over sections from Wikipedia articles with unanswerable and open-ended questions

Unanswerable questions - master thesis project

- Project is based on a conversational search system (used in the previous editions of TREC CAsT) built with three components: query rewriting, first-pass retrieval, and re-ranking
- The goal is to build a classifier for the answerability of the given query that uses the final ranking provided by the system
- The student is expected to build a baseline classifier and one advanced method
- The experiments should focus on the performance of the classifier and on different ways of handling its output in the conversational search system
- The student is expected to propose two different methods of handling the unanswerable questions in the system's response
- An alternative task can be creating a new dataset for handling unanswerable questions in conversational search systems

Summary

- Conversational Search Systems
  - Recap on conversational information access
  - TREC CAsT
  - Query rewriting
  - State of the art
- Mixed initiative
  - Clarifying questions
  - System revealment

Reading

- *A Theoretical Framework for Conversational Search*, Filip Radlinski, Nick Craswell `https://dl.acm.org/doi/pdf/10.1145/302979.303030`
- *Conversational search (Dagstuhl Seminar 19461)*, Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein `https://drops.dagstuhl.de/opus/volltexte/2020/11983/pdf/dagrep_v009_i011_p034_19461.pdf`
- *Conversational information seeking, Chapter 2*, Hamed Zamani, Johanne R. Trippas, Jeff Dalton, Filip Radlinski, `https://arxiv.org/abs/2201.08808`