Conversational Information Access II

[DAT640] Information Retrieval and Text Mining

Weronika Łajewska
University of Stavanger

01.10.2024

In this module

1. Conversational Information Access - recap
2. Conversational Question Answering
3. Conversational Search Systems
4. Mixed-initiative

Conversational Information Access - recap

Conversational voice assistants

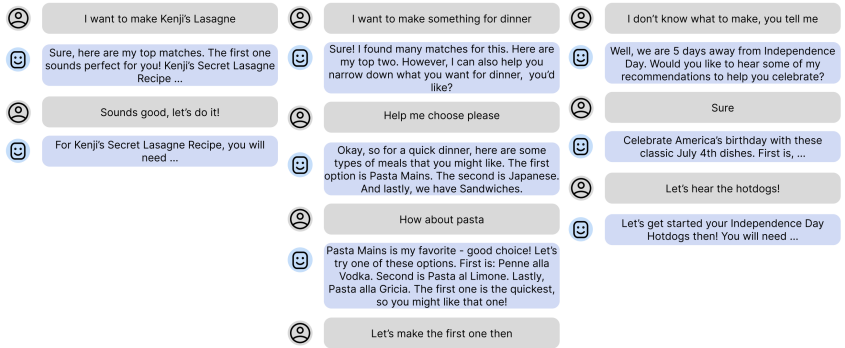


Figure: Conversational agent for solving complex real-world tasks developed at Alexa Prize Taskbot Challenge [1]; *GRILLBot-v2: Generative models for multi-modal task-oriented assistance*, University of Glasgow

[1] `https://www.amazon.science/alexa-prize/proceedings`

Conversational Information Access

- A subset of conversational AI systems that help satisfy the information needs of users via multi-turn conversations
- Specifically aims 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

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, make calls, etc.

What is a conversation?

- Oxford definition: *"Conversation is a talk, especially an informal one, between two or more people, in which news and ideas are exchanged"*

What is a conversation?

- Oxford definition: *"Conversation is a talk, especially an informal one, between two or more people, in which news and ideas are exchanged"*
- O.Henry perspective: *"Inject a few raisins of conversation into the tasteless dough of existence"*

What is a conversation?

- Oxford definition: *"Conversation is a talk, especially an informal one, between two or more people, in which news and ideas are exchanged"*
- O.Henry perspective: *"Inject a few raisins of conversation into the tasteless dough of existence"*
- Conversational Information Access definition: *"Conversation is interactive communication for exchanging information between two or more participants (i.e., humans or machines) that involves a sequence of interactions"*

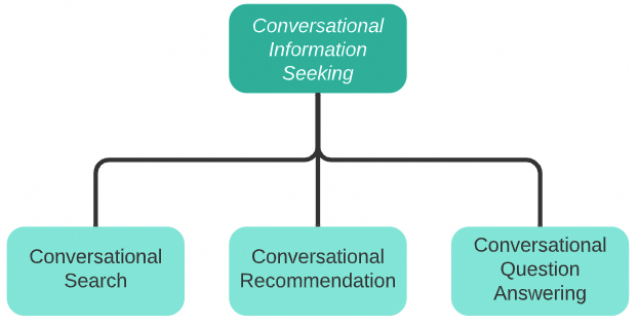
What is a conversation?

- Oxford definition: *"Conversation is a talk, especially an informal one, between two or more people, in which news and ideas are exchanged"*
- O.Henry perspective: *"Inject a few raisins of conversation into the tasteless dough of existence"*
- Conversational Information Access definition: *"Conversation is interactive communication for exchanging information between two or more participants (i.e., humans or machines) that involves a sequence of interactions"*
- *"Information seeking conversation is a conversation in which the goal of information exchange is satisfying the information needs of one or more participants"*

Conversational information seeking[2]

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



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

Conversational Question Answering

Query

- Query is an expression of user's information need
- One information need may be expressed via multiple queries - the same information need might give rise to different manifestations with different systems: for example, a few keywords are typed into the search box of a web search engine, but a fluent, well-formed natural language question is spoken to a voice assistant
- Query can take multiple forms:
  - keywords-based queries
  - natural language queries $\Leftarrow$ our focus in this lecture

Natural Language Query

- Formulated as complete sentences or questions, mimicking how humans typically speak or write
- Main categories of natural language queries:
  - Close-ended (factoid) - precise queries where users are looking for specific facts or pieces of information, e.g., *"Who won the World Cup in 2018?"*
  - Open-ended - broad and exploratory in nature queries, designed to elicit detailed information or encourage discovery, that very often are not limited to one specicifc correct answer, e.g. *"Why is machine learning important for data science?"*

Types of Open-ended Queries

- Exploratory - user wants to investigate a new or unfamiliar topic and his goal is to gather a wide range of information rather than find a specific fact, e.g. *"What are the best holidays destinations in Europe?"*
- Multifaceted - cover multiple aspects or subtopics, often requiring diverse information sources, e.g. *""What are the impacts of renewable energy adoption?""*
- Multi-perspective - require exploring a topic from several angles, considering different perspectives, arguments, or ideologies, e.g. *"What are the pros and cons of electric cars?"*

Question

How would you characterize open-ended questions?

Characteristics of Open-ended Queries

- Lack of a definitive answer
- Complexity
- Subjectivity
- Require synthesis of information
- Ambiguity
- Encourage discussion or elaboration

Approaches to QA

- Knowledge-based systems that use structured knowledge bases
- Extractive systems that return a span or snippet of text from a document
- LLM-based abstractive free-text answer generation (closed-book)
- Multi-stage retrieve-the-read approaches (open-book):
  1. Document retrieval - filtering approach to reduce the number of candidate passages
  2. Reading comprehension - extracting answer spans from retrieved documents

QA Datasets - SQuAD [3]

- Stanford Question Answering Dataset (SQuAD)
- 100,000+ questions posed by crowdworkers on a set of Wikipedia articles
- Answer for each question is selected from all possible spans in the passage
- Multiple different answer types

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

[3] *Rajpurkar, Pranav et al. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." Conference on Empirical Methods in Natural Language Processing (2016).*

QA Datasets - HotPotQA [4]

- 113k Wikipedia-based question-answer pairs
- The questions require finding and reasoning over information taken from more than one document to arrive at the answer
- The dataset provides sentence-level supporting facts required for reasoning

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

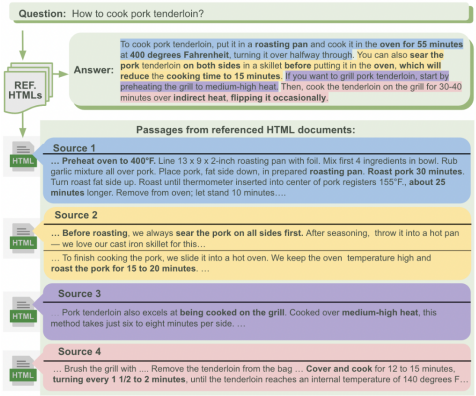
Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

[4] *Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R., & Manning, C.D. (2018). HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. Conference on Empirical Methods in Natural Language Processing.*

QA Datasets - WikiHowQA [5]

- Multi-document non-factoid QA dataset
- Questions often begin with "How to ..."
- It is based on WikiHow web-resource for INSTRUCTION questions on a wide range of topics
- It consists of 11,746 questions, each paired with a corresponding human-written answer, sourced from a diverse range of WikiHow articles



[5] *Bolotova-Baranova, V., Blinov, V., Filippova, S., Scholer, F., & Sanderson, M. (2023). WikiHowQA: A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering. Annual Meeting of the Association for Computational Linguistics.*

Conversational Search Systems

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

Question

What are the challenges in conversational search systems? (recap from last lecture)

CSS - additional challenges

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

Question

When is the conversational aspect of the search system needed?

Complex scenarios that require CSS [6]

- 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.

[6] Radlinski, F., & Craswell, N. (2017). *A Theoretical Framework for Conversational Search*. Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval.

Question

What is the difference between conversational QA and conversational search system?

Conversational QA vs Conversational Search

- QA systems provide direct answers to user questions, often summarizing multiple sources or perspectives, particularly for open-ended or complex queries
- Search systems guide users through a process of refining and focusing their search queries on the aspects they are mostly interested in, aiming to retrieve relevant documents or pieces of information; the emphasis is on navigating through a search space in addition to synthesizing a final, comprehensive response

Conversational Search Systems

- TREC CAsT
- System architecture
- Reproducibility

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

Demo

TREC workshops and resources $\Rightarrow$ TREC Browser

TREC CAsT - topic[7]

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?

[7] TREC CAsT 2020: The Conversational Assistance Track Overview

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
- In year 4, the track introduced mixed-initiative and response generation subtasks

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
- Participants are asked to prioritize the submitted runs because not all the systems that appear in the competition are pooled
- Exhaustive pools let us assume that whatever is not pooled is not relevant

Conversational Search Systems

- TREC CAsT
- System architecture
- Reproducibility

TREC 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.

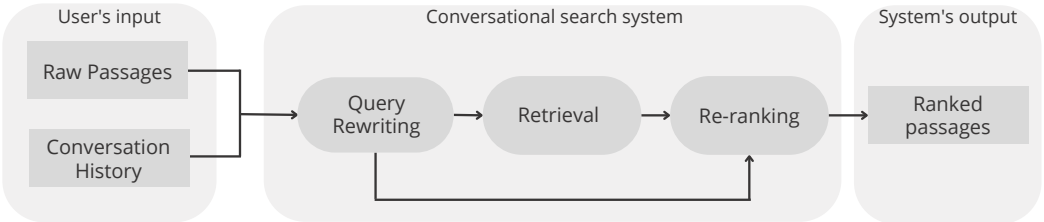


Figure: Cascade two-stage pipeline architecture with query rewriting module.

Retrieval and Re-ranking

- Inverted index with sparse retrieval (e.g., BM25)
- Learning-to-Rank implemented as a two-step retrieval with initial ranking (retrieving top-N candidate documents) and re-ranking (creating feature vectors and re-ranking top-N candidates)
  - pointwise re-ranking, e.g. monoT5, where the relevance of each passage in the ranking with respect to the query is computed independently of other passages
  - pairwise re-ranking, e.g. duoT5, where the relevance is considered in terms of pairs of passages

Retrieval and Re-ranking - cont.

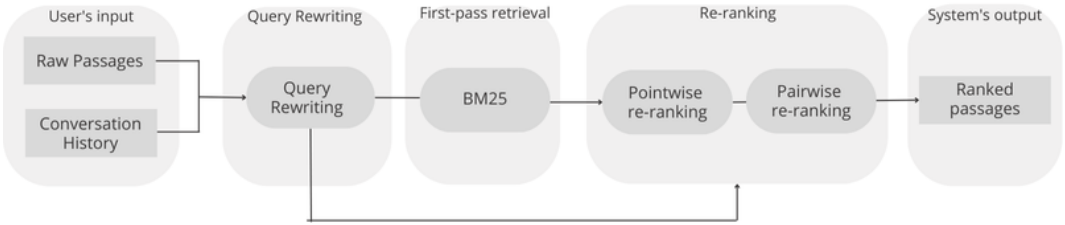


Figure: Pipeline architecture with reranking module.

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 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 - approaches

- 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 - 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

System Pipeline

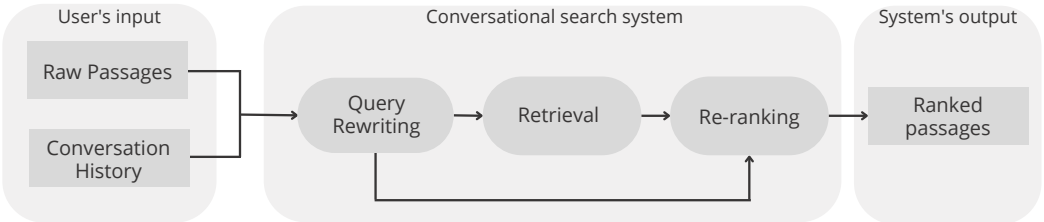


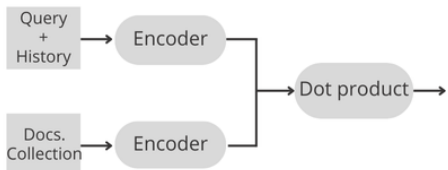
Figure: Cascade two-stage pipeline architecture with query rewriting module.

Question

Which of components will be affected if we skip QR module?

Dense Retrieval

- Dense retrieval aims to match texts in a continuous representation space learned via deep neural networks
- Dense retrieval calculates the relevance score using similarities in a learned embedding space



- The effectiveness of dense retrieval resides in learning a good representation space that maps query and relevant documents together while separating irrelevant ones.
- Representations of all documents in the collection can be pre-computed and stored in advance so during query processing, only the query representation is computed
- Dense retrieval takes into consideration the semantics of the terms in a query and is not limited to an exact match of terms

Dense Retrieval - cont.

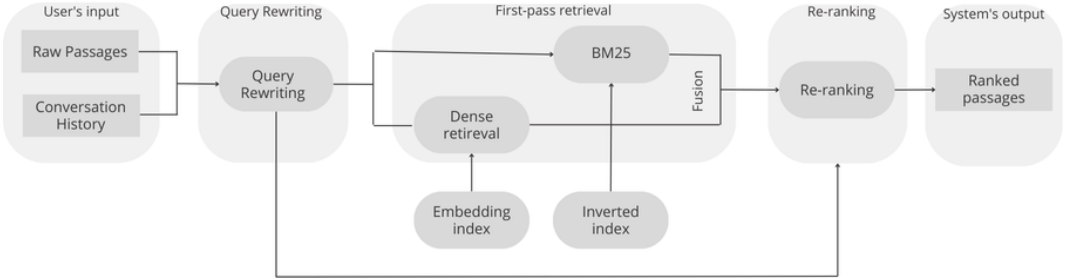


Figure: Pipeline architecture with dense retrieval.

- How to train a model to distinguish between relevant documents and irrelevant documents? $\rightarrow$ Negative sampling
- How to search through a massive index of document representations? $\rightarrow$ Embedding index

Query Expansion - recap

- Query modeling using feedback takes the results of a user's actions or previous search results to improve retrieval
- Often implemented as updates to a query, which then alters the list of documents
- Overall process is called relevance feedback, because we get feedback information about the relevance of documents
  - Explicit feedback: user provides relevance judgments on some documents
  - **Pseudo relevance feedback** (or blind feedback): we don't involve users but "blindly" assume that the top-k documents are relevant
  - Implicit feedback: infer relevance feedback from users' interactions with the search results (clickthroughs)

Query Expansion - cont.

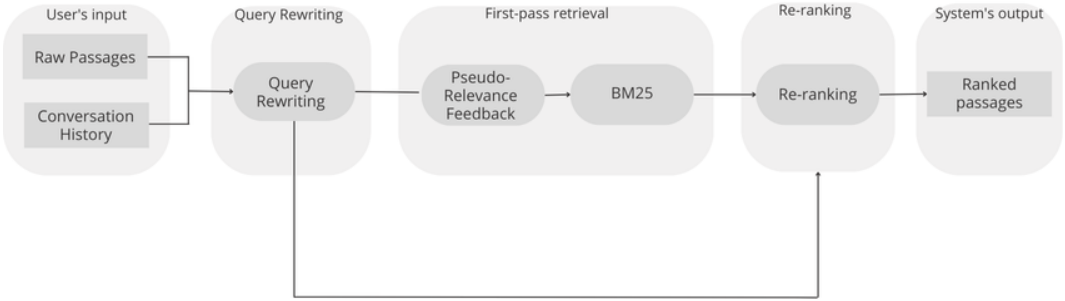


Figure: Pipeline architecture with pseudo-relevance feedback.

Document Expansion

- Document expansion techniques address the vocabulary mismatch problem: *queries can use terms semantically similar but lexically different from those used in the relevant documents*
- In document expansion learning, sequence-to-sequence models are used to modify the actual content of documents, boosting the statistics of the important terms and generating new terms to be included in a document
- Doc2Query generate new queries for which a specific document will be relevant

Document Expansion - cont.

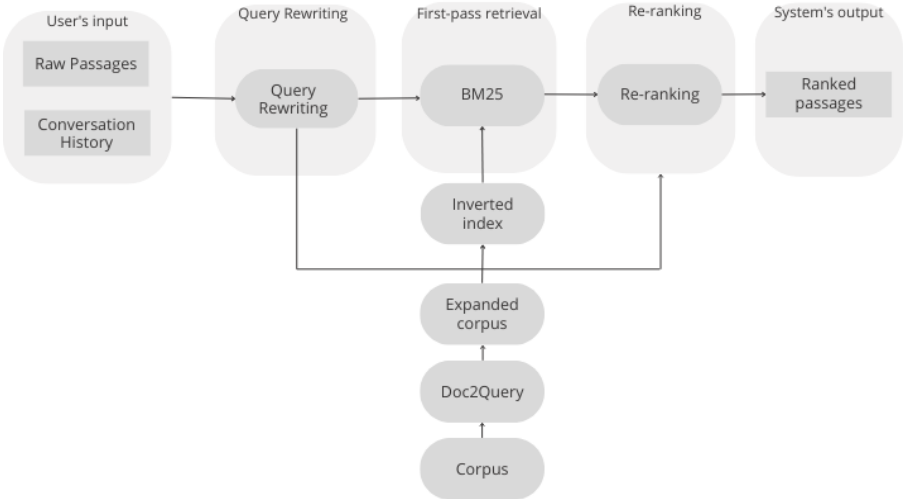


Figure: Pipeline architecture with documents expansion component.

Response Generation

- In conversational setting, identifying the top relevant passages is only an intermediate step
- Conversational response generation aims at synthesizing the information from the top retrieved passages into a single response that encapsulate the most relevant pieces of information in an easily consumable unit
- Generated response should be factual (no hallucinations), grounded in credible sources, concise/coherent, complete, transparent, and explainable

Response Generation - cont.

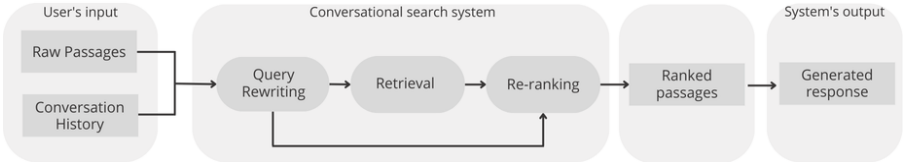


Figure: Pipeline architecture with response generation component.

Retrieval-Augmented Generation

- Retrieval-Augmented Generation (RAG) is the process of optimizing the output of a large language model, so it references an authoritative knowledge base outside of its training data sources before generating a response
- RAG involves a retriever component and a generator component
- Limitations of current RAG models:
  - Context limitation
  - Retrieval errors
  - Bias

Retrieval-Augmented Generation - cont.

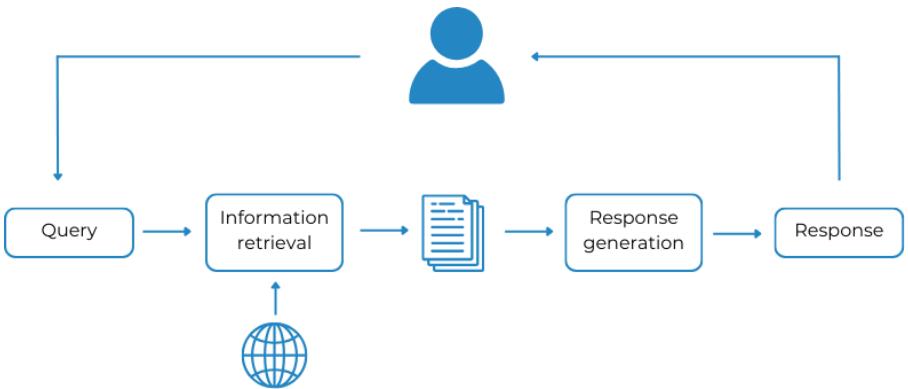


Figure: A high-level architecture of a retrieval-augmented generation system.

Demo

Grounded response generation with source attribution $\Rightarrow$ www.perplexity.ai

Conversational Search Systems

- TREC CAsT
- System architecture
- Reproducibility

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

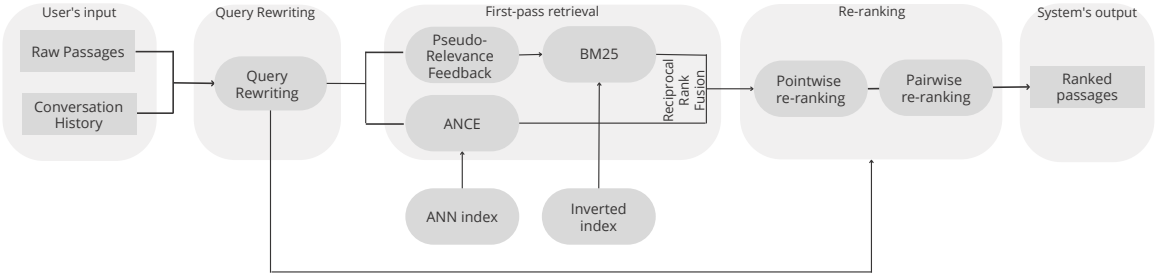


Figure: State of the art system from TREC CAsT'21

State of the art CSS - reproducibility[8]

- 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

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

Question

What challenges do you see in reproducing CSS systems?

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[9]
- 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

[9] 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

Summary

- Recap on conversational information access
- Conversational Question Answering
- Conversational Search Systems
  - TREC CAsT
  - System architecture
  - Reproducibility
- Mixed initiative

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`