Advanced Evaluation

[DAT640] Information Retrieval and Text Mining

Krisztian Balog & Nolwenn Bernard
University of Stavanger

October 24, 2022

In this module

1. Evaluating Conversational Search and Recommender Systems

2. Data collection exercise

Recap

q001	doc00320	0
q001	doc01321	4
q001	doc17754	2
q001	doc44510	3
q001	doc90974	0
...	...	
q211	doc55204	2
q211	doc09077	0
q211	doc13201	1
q211	doc89210	1
q211	doc67828	4
...	...	

`f2015.qrels`

Question

What are the assumptions made in offline test collections for ranking evaluation?

Assumptions in offline retrieval evaluation

- Queries are a representative sample of actual user needs
  - Head vs. torso vs. tail queries
- User judgments are represenative of an average user of the actual user population
- There is a finite set of items to choose from
  - Items can only be retrieved, not generated
- All relevant answers have been identified
  - This is achieved by pooling

From relevance assessments to evaluation measures

- Relevance assessments are for (query, item) pairs
- Evaluation measures express how well the system ranks items for a given query w.r.t. the ground truth
- Evaluation measures have an (implicit or explicit) underlying user model
  - The user model describes how users interact with the ranked list

Question

Assume that your boss tells you that the search engine of the site you're in charge of should be replaced. They show you a few queries where the search doesn't work as expected and suggest to switch to a different search service provider who handle those queries much better. How would you react to this?

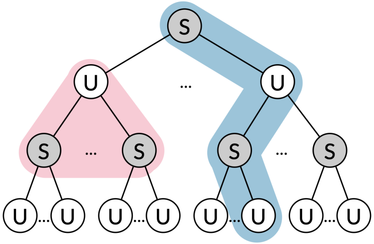
Evaluating Conversational Search and Recommender Systems

Types of evaluation

- Component-level vs. end-to-end
- Automatic vs. human vs. live evaluation

Challenges

Turn-based (offline) evaluation
✅ Possible to create a reusable test collection for a specific subtask
❌ Limited to a single turn, does not measure overall user satisfaction



End-to-end (human) evaluation
✅ Possible to annotate entire conversations
❌ Expensive, time-consuming, does not scale, does not yield a reusable test collection

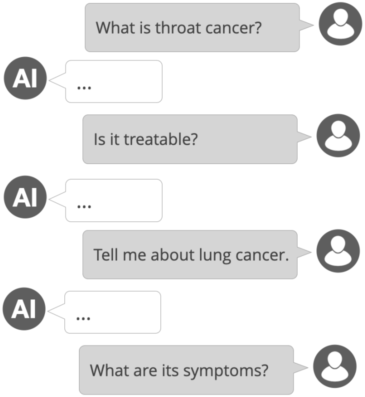
Evaluating Conversational Search and Recommender Systems

- Offline Evaluation
- Evaluation Using User Simulation

Conversational search

Task: given a natural language query, return a passage from a collection as an answer.

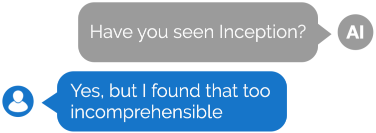
- TREC Conversational Assistance track
- Evaluation is performed with respect to answer relevance



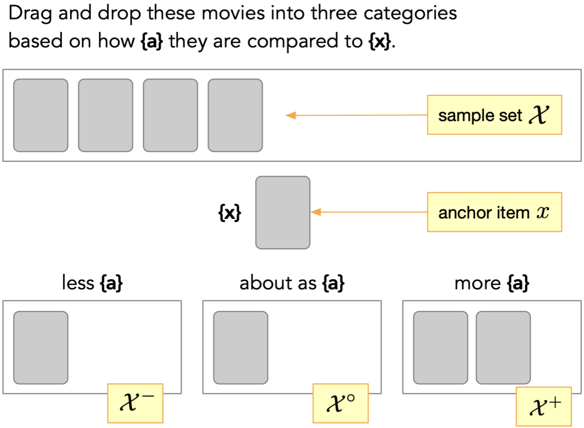
Critiquing in conversational recommendations

Task: given an item recommendation, incorporate natural language feedback (critiques) on items, which comes in the form of *soft attributes*

- A *soft attribute* is a property of an item that is not a verifiable fact that can be universally agreed upon, and where it is meaningful to compare two items and say that one item has more of the attribute than another
- Example soft attributes for movies: artsy, light-hearted, intense, predictable, violent ...



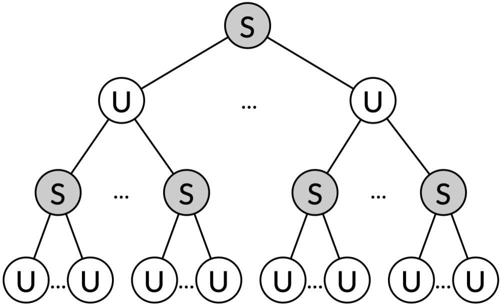
Purpose-built annotation interface



Evaluating Conversational Search and Recommender Systems

- Offline Evaluation
- Evaluation Using User Simulation

Challenges



Objectives

The user simulator should produce responses that a real user would give in a certain dialog situation.

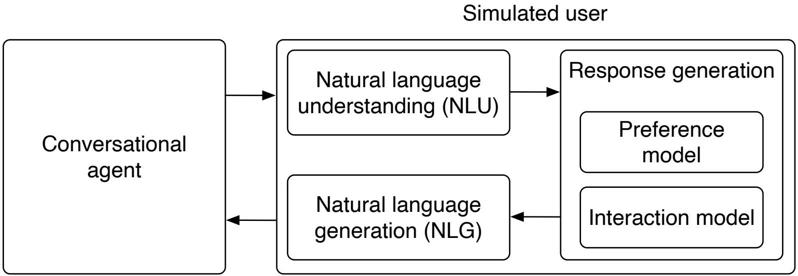
- Enable automatic assessment of conversational agents
- Make no assumptions about the inner workings of conversational agents

Formally

- For a given system S and user population U, the goal of user simulation U^* is to predict the performance of S when used by U, denoted as $M(S, U)$
- For two systems S_1 and S_2, U^* should be such that

$$M(S_1, U) < M(S_2, U) \Rightarrow M(S_1, U^*) < M(S_2, U^*)$$

User simulator archictecture

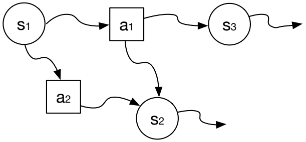


- **NLU**: Translate an agent utterance into a structured format
- **Response generation**: Determine the next user action based on the understanding of the agent's utterance
- **NLG**: Turn a structured response representation into natural language

Modeling simulated users

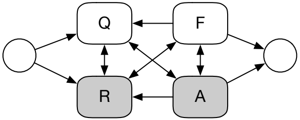
- Model dialogue as a Markov Decision Process (MDP)
- Every MDP is formally described by a finite state space $\mathcal{S}$, a finite action set $\mathcal{A}$, and transition probabilities P
- At each time step (dialogue turn) t, the dialogue manager is in a particular state $s_t \in \mathcal{S}$
- By executing action $a_t \in \mathcal{A}$, it transitions into the next state s_{t+1} according to the transition probability $P(s_{t+1}|s_t, a_t)$
- The Markov property ensures that the state at time $t+1$ depends only on the state and action at time t:

$$P(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0) = P(s_{t+1}|s_t, a_t) \ .$$

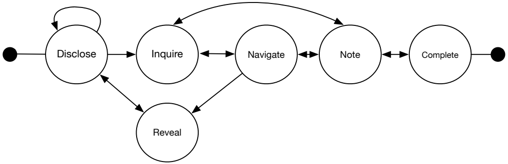


Interaction model

- Defines the space of agent and user actions and the possible transitions between them
- Examples



Simplest model with two user actions (query, feedback) and agent actions (request, answer)



User actions specific to the task of conversational item recommendation

Instantiating simulators

- Specifying the space of user and agent actions
- Training the NLU for the particular agents
- Setting up NLG (template-based/machine-learned) for the particular scenario
- It has to be grounded in actual user data!

Data collection exercise

Motivations

- Conversations mix *conversational goals*: question answering, search, and recommendation
- Lack of datasets with conversations mixing goals
- Provide resources to support development of new conversational agent handling multiple goals

Virtual shopping assistant

- Objective: collect human-human conversations mixing conversational goals in the retail domain
- Roles
  - Client: you are looking to buy an item and need the help of an assistant
  - Retail assistant: provide help to a client in order to satisfy their needs

Today's session

- Two groups: (1) clients and (2) retail assistants
- Topics of discussion available
  - Books
  - Sports and outdoors
  - CDs and vinyl
  - Grocery and gourmet food
  - Office products
  - Cell phones and accessories
  - Musical instruments
- Link to the chat platform
- Aim: complete 2 conversations

Data collection session

- We plan to perform a large scale data collection session
- Link to the registration form

Summary

- Thinking carefully about assumptions behind evaluation
- Evaluating conversational search and recommender systems
  - Types of evaluation (turn-based vs. end-to-end, automatic vs. human)
  - Advantages and limitations of offline evaluation
  - Examples of automatic turn-based evaluation
  - User simulation for automatic end-to-end evaluation

Reading

- Zhang and Balog. **Evaluating Conversational Recommender Systems via User Simulation**. In: *26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20)*
  - https://arxiv.org/pdf/2006.08732
- Balog et al. **On Interpretation and Measurement of Soft Attributes for Recommendation**. In: *44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*
  - https://arxiv.org/pdf/2105.09179