

How Am I Doing? Evaluating Conversational Search Systems Offline

Aldo Lipani, Ben Carterette, Emine Yilmaz
ACM Transactions on Information Systems
Volume 39 Issue 4 October 2021

SCAI - October 8th, 2021



Build test collections for conversational search systems.

A Framework for Offline Evaluation

- A methodology for building test collections with relevance judgments

Conversational Search User Study

Hi there!

Your task has not started yet. The task will start after clicking on the **begin** button below.

The task is divided into 3 phases:

- Topic and subtopic comprehension phase;
- Search interaction phase;
- Feedback phase.

Before starting each phase you will receive instructions on what you are expected to do.

Please take a moment to read the instructions provided, don't rush through them, and be mindful. We will manually review your work to determine whether to accept or not, so please imagine that the task is important to you and give your search serious consideration.

Enjoy the task!

begin

Topic and Subtopic Comprehension Phase

Next you will be provided with a topic and subtopic. Read this carefully and think about how you would search about this topic.

begin

Black Death

Imagine you need to write an essay about the plague known as the "Black Death", and you are planning to have sections about the origins of the disease, its geographical spread, its impact on populations, its symptoms and treatments, and modern research being done. How would you search to find more information about this topic?

begin

Search Interaction Phase

Next you will be able to perform searches to learn more about the given topic and its subtopics. At this point you can do one of two things, **query the search engine** or **end the search interaction phase**.

Query the Search Engine

Every time you search, you need to articulate the query by **selecting the subtopic** that best represents your query. Only then, you will be able to see the result.

The search engine will then provide you a paragraph. Read the paragraph and tell us if the search engine has successfully returned a paragraph that is relevant to your query. You should also consider whether the paragraph is a good response to your request.

End the Search Interaction Phase

Whenever you find yourself satisfied with what you have learnt about the topic and its subtopics you can end the search phase by clicking the **end** button.

Important: In a setting where you are taking an online survey, you would normally use the system to gather as much information as possible, which would require spending significant amount of time searching. So please try to leave the system as soon as you have gathered as much information as possible about the topic. We will be logging your time and queries with the system. If the time you spend is less than 3 minutes, or your queries clearly not consistent, you will not receive any credits.

begin

Search

Ask here or leave it empty

Search **end**

Result

Other forms of plague have been implicated by modern scientists. The modern bubonic plague has a mortality rate of 30-70% and symptoms including fever of 38-41°C (100-106 °F), headaches, painful aching joints, nausea and vomiting, and a general feeling of malaise. Left untreated, of those that contract the bubonic plague, 85 percent die within eight days. Pneumonic plague has a mortality rate of 90 to 99 percent. Symptoms include fever, cough, and blood-tinged sputum. As the disease progresses, sputum becomes free flowing and bright red. Septicemic plague is the least common of the three forms, with a mortality rate near 100%. Symptoms are high fevers and purple skin patches (purpura due to disseminated intravascular coagulation). In cases of pneumonic and particularly septicemic plague, the progress of the disease is so rapid that there would often be no time for the development of the enlarged lymph nodes that were noted as buboes.

Is this paragraph relevant to your query?

Yes, it is **No, it's not**

Search

Ask here or leave it empty

Search **end**

Feedback Phase

You have now reached the last phase of this task. At this point we would like you to tell us whether you are satisfied about the interaction you had with this search system.

begin

Feedback

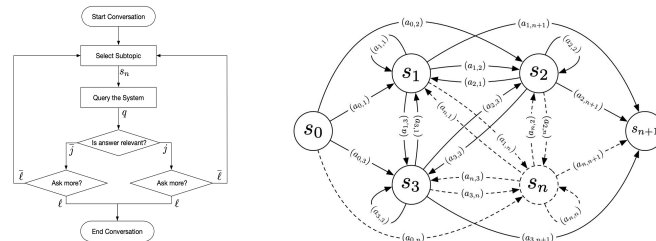
Are you satisfied with your interaction with the search system?

Yes, I am **No, I'm not**

The End!

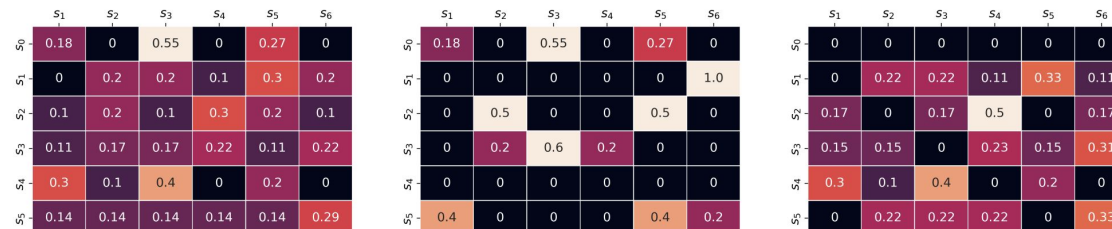
Congratulations! You have finished this task.

- An evaluation measure based on a user interaction model




$$ECS(c) = \sum_{m=1}^{|c|} j(c_m) \prod_{m'=1}^{m-1} (\alpha^+ j(c_{m'}) + \alpha^- (1 - j(c_{m'}))),$$


- An approach to collecting user interaction data to train the model






Test Collection-Based Evaluation of Conversational Search

System A


Q:  What flowering plants work for cold climates?

A: Pansies love cool weather and add tons of color to the winter landscape in frost-free regions. 


 How much cold can pansies tolerate? 


Pansies prefer temperatures during the night just a bit above freezing with 40 degrees considered ideal. 



 Can it survive frost? 


Pansies and Violas are hardy plants and will survive a frost—and even a hard freeze—for a period of time. 



System B


Q:  What flowering plants work for cold climates?

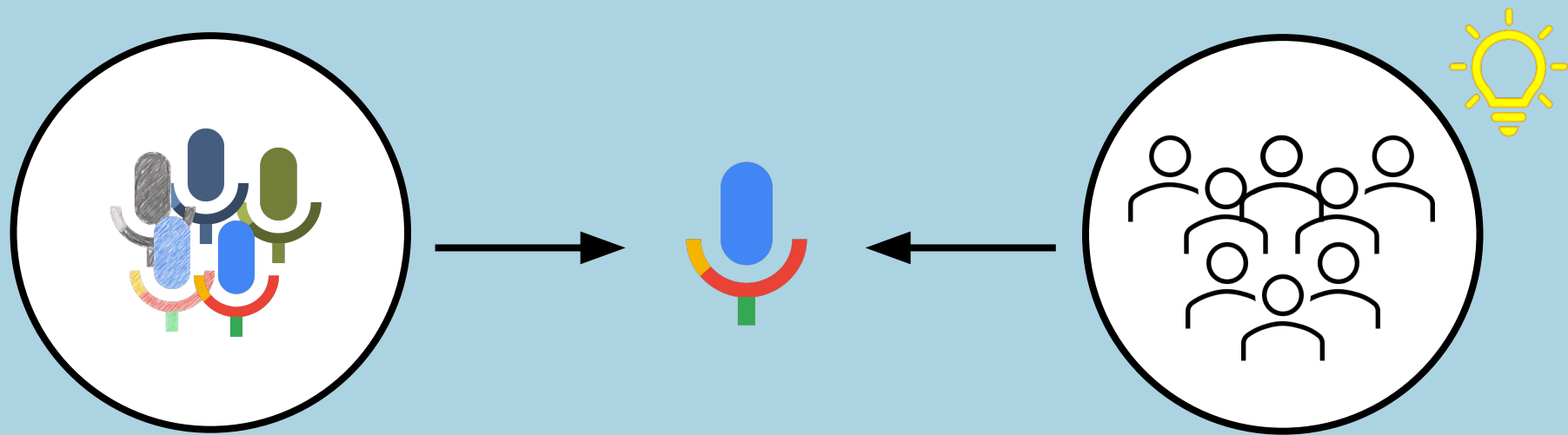
A: The herb yarrow provides frilly foliage and delicate blooms when used for cold climate gardening. 

 How much cold can pansies tolerate? 

Pansies and Violas are hardy plants and will survive a frost—and even a hard freeze—for a period of time. 

























 Can it survive frost? 

Pansies and Violas are hardy plants and will survive a frost—and even a hard freeze—for a period of time. 

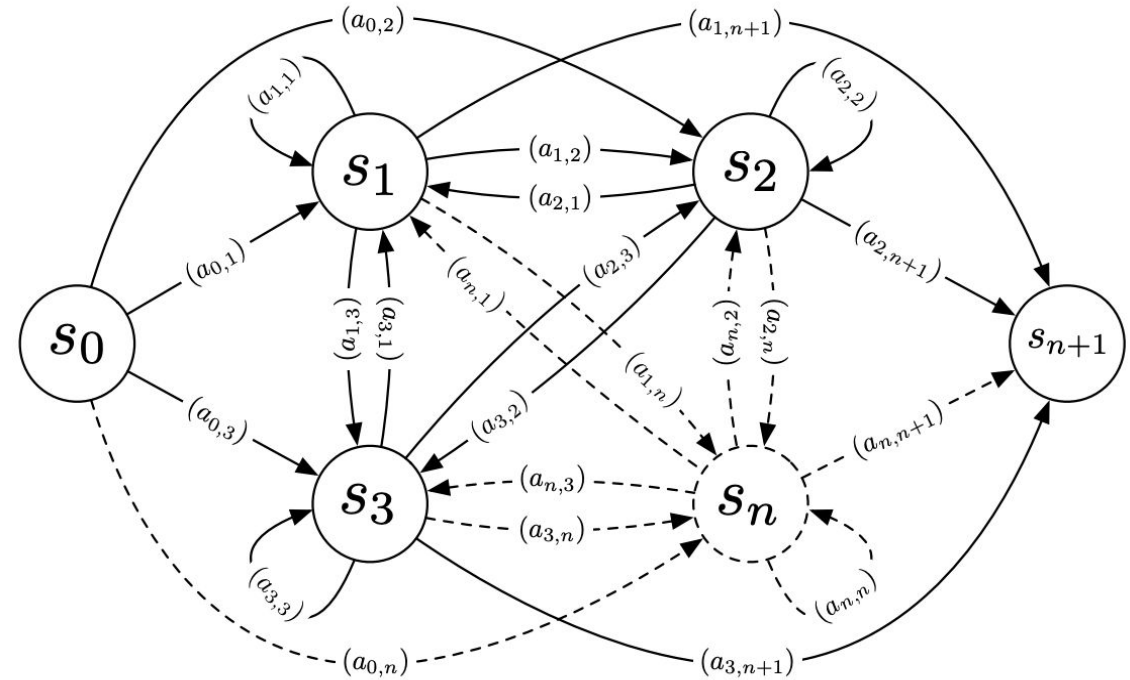
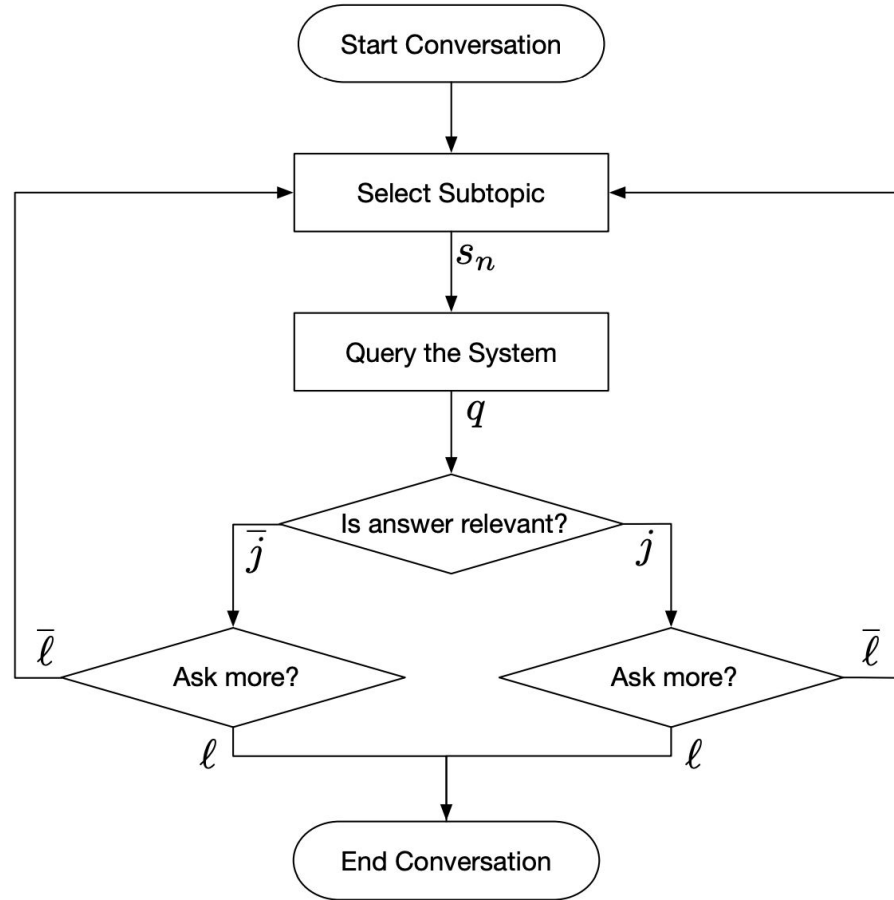


Can we simulate this?

Subtopic-Based Evaluation of Conversational Search

<p>Q:  What flowering plants work for cold climates?</p> <p>A: Pansies love cool weather and add tons of color to the winter landscape in frost-free regions. </p> <p>pansy</p>	<p>Q:  How much cold can pansies tolerate? </p> <p>A: Pansies prefer temperatures during the night just a bit above freezing with 40 degrees considered ideal. </p> <p>pansy cold tolerance</p>	<p>Q:  Can it survive frost? </p> <p>A: Pansies and Violas are hardy plants and will survive a frost—and even a hard freeze—for a period of time. </p> <p>pansy frost tolerance</p>
<p>Q:  What flowering plants work for cold climates?</p> <p>A: The herb yarrow provides frilly foliage and delicate blooms when used for cold climate gardening. </p> <p>yarrow</p>	<p>Q:  How much cold can pansies tolerate? </p> <p>A: Pansies and Violas are hardy plants and will survive a frost—and even a hard freeze—for a period of time. </p> <p>pansy frost tolerance</p>	<p>Q:  Can it survive frost? </p> <p>A: Pansies and Violas are hardy plants and will survive a frost—and even a hard freeze—for a period of time. </p> <p>pansy frost tolerance</p>
<p>Q:  What flowering plants work for cold climates?</p> <p>A: The herb yarrow provides frilly foliage and delicate blooms when used for cold climate gardening. </p> <p>yarrow</p>	<p>Q:  How much cold can yarrow tolerate? </p> <p>A: Achillea (yarrow) tolerates very cold temperatures, but hostas that have unfurled are subject to frost damage. </p> <p>yarrow cold tolerance</p>	<p>Q:  Can it survive frost? </p> <p>A: They can survive a light frost if they're sufficiently hardened off. </p> <p>yarrow frost tolerance</p>

Conversational Search Simulation Model



Components of a Simulation-Based Evaluation

Conversational search system:

- Takes a question/query, returns an answer in form of a sentence/paragraph

Test collection:

- Topics/tasks/information needs
- Subtopics/aspects/facets/subtasks/entities
- User queries that model subtopics
- Transition probabilities between subtopics
- Corpus of “answers”
- Relevance judgments of answers to subtopics

Evaluation Metric

ECS: Expected Conversation Satisfaction Loop

- Sample subtopic, then query
- Request answer for query from system
- Obtain relevance
- Increment gain, discount by length of interaction according to “persistence parameters” α s

Two variants for line 16, the subtopic

RI. sampled is independent of answer relevance

RD. sampled is conditioned on the answer relevance

Algorithm 1: Computation of ECS

Input: α^+ , α^- , \mathcal{S} , \mathcal{Q} , \mathcal{J} , system()

Output: score

```
1 score  $\leftarrow$  0
2  $p(Q = q) \leftarrow 1$ 
3 relevant  $\leftarrow$  false
4 subtopic  $\leftarrow$  start
5 subtopic  $\sim \mathcal{S}_{\text{relevant, subtopic}}$ 
6 while subtopic  $\neq$  end do
7   query  $\sim \mathcal{Q}_{\text{subtopic}}$ 
8   answer  $\leftarrow$  system(query)
9   relevant  $\leftarrow \mathcal{J}_{\text{subtopic, answer}}$ 
10  if relevant then
11    | score  $\leftarrow$  score +  $p(Q = q)$ 
12    |  $p(Q = q) \leftarrow \alpha^+ p(Q = q)$ 
13  else
14    |  $p(Q = q) \leftarrow \alpha^- p(Q = q)$ 
15  end
16  subtopic  $\sim \mathcal{S}_{\text{relevant, subtopic}}$ 
17 end
```

Comparison of ECS to Precision and RBP

	Sim.	Parameters	τ	ρ
P	RI		0.3963	0.4184
	RD		0.6606†	0.8200†
RBP	RI	$\alpha = 0.79$	0.3963	0.4184
	RD	$\alpha = 0.79$	0.6606†	0.8200†
ECS	RI	$\alpha^+ = 0.82, \alpha^- = 0.70$	0.3963	0.4184
	RD	$\alpha^+ = 0.85, \alpha^- = 0.64$	0.6972†	0.8383†

Summary

- Evaluating conversations offline with test collections is hard
- Use insights from diversity & novelty, sessions, tasks to design an evaluation framework based on simulation
- Provided results based on subtopics, and subtopics and relevance.
- Crowdsourced data for test collection queries and transition probabilities
- Evaluating conversations offline with test collections is not so hard anymore!