How Am I Doing? Evaluating Conversational Search Systems Offline

Aldo Lipani, Ben Carterette, Emine Yilmaz

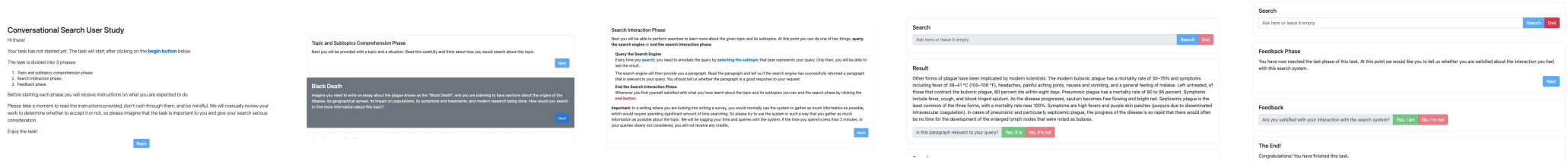
ACM Transactions on Information Systems
Volume 39 Issue 4 October 2021

SCAI - October 8th, 2021

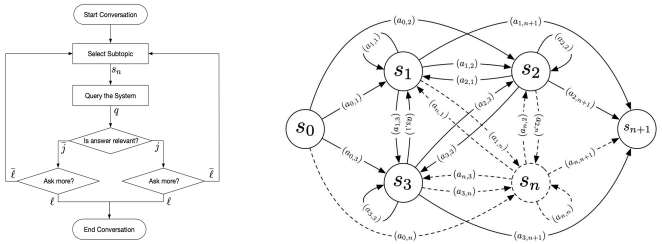
Build **test collections** for **conversational search systems**.

A Framework for Offline Evaluation

- A methodology for building test collections with relevance judgments



- An evaluation measure based on a user interaction model



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

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

	s_1	s_2	s_3	s_4	s_5	s_6
s_0	0.18	0	0.55	0	0.27	0
s_1	0	0.2	0.2	0.1	0.3	0.2
s_2	0.1	0.2	0.1	0.3	0.2	0.1
s_3	0.11	0.17	0.17	0.22	0.11	0.22
s_4	0.3	0.1	0.4	0	0.2	0
s_5	0.14	0.14	0.14	0.14	0.14	0.29

	s_1	s_2	s_3	s_4	s_5	s_6
s_0	0.18	0	0.55	0	0.27	0
s_1	0	0	0	0	0	1.0
s_2	0	0.5	0	0	0.5	0
s_3	0	0.2	0.6	0.2	0	0
s_4	0	0	0	0	0	0
s_5	0.4	0	0	0	0.4	0.2

	s_1	s_2	s_3	s_4	s_5	s_6
s_0	0	0	0	0	0	0
s_1	0	0.22	0.22	0.11	0.33	0.11
s_2	0.17	0	0.17	0.5	0	0.17
s_3	0.15	0.15	0	0.23	0.15	0.31
s_4	0.3	0.1	0.4	0	0.2	0
s_5	0	0.22	0.22	0.22	0	0.33

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.

Q: How much cold can pansies tolerate?

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

Q: Can it survive frost?

A: 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?

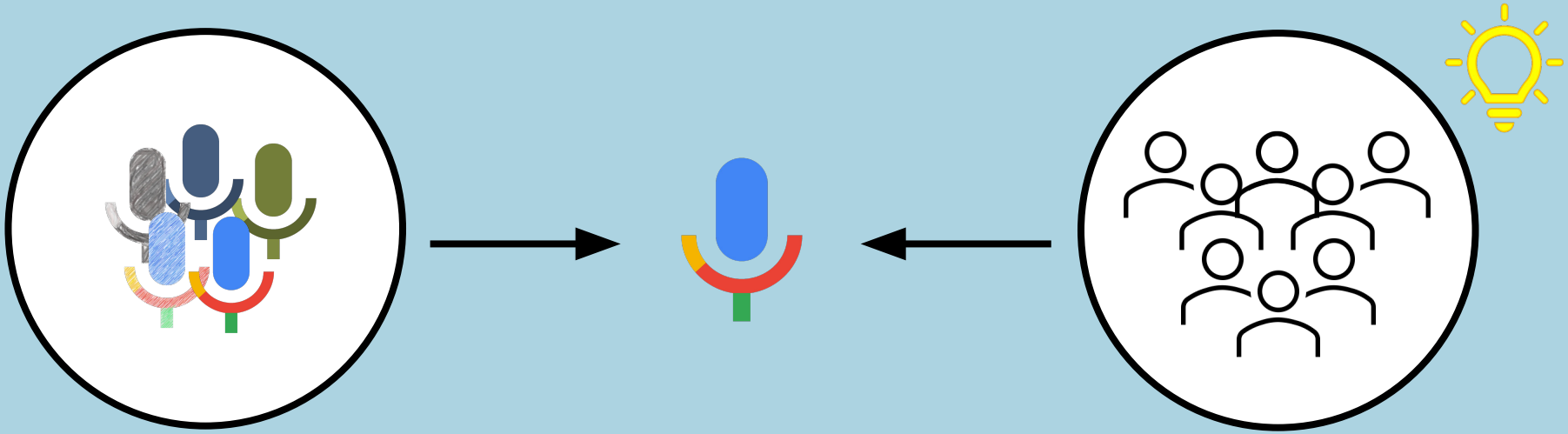
A: The herb yarrow provides frilly foliage and delicate blooms when used for cold climate gardening.

Q: How much cold can pansies tolerate?

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

Q: Can it survive frost?

A: 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

cold climate flowering plants

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.

pansy

pansy cold tolerance

Q: How much cold can pansies tolerate?

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

pansy cold tolerance

pansy frost tolerance

Q: Can it survive frost?

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

pansy frost tolerance

cold climate flowering plants

Q: What flowering plants work for cold climates?

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

yarrow

pansy cold tolerance

Q: How much cold can pansies tolerate?

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

pansy frost tolerance

pansy frost tolerance

Q: Can it survive frost?

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

pansy frost tolerance

cold climate flowering plants

Q: What flowering plants work for cold climates?

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

yarrow

yarrow cold tolerance

Q: How much cold can yarrow tolerate?

A: Achillea (yarrow) tolerates very cold temperatures, but hostas that have unfurled are subject to fre damage.

yarrow cold tolerance

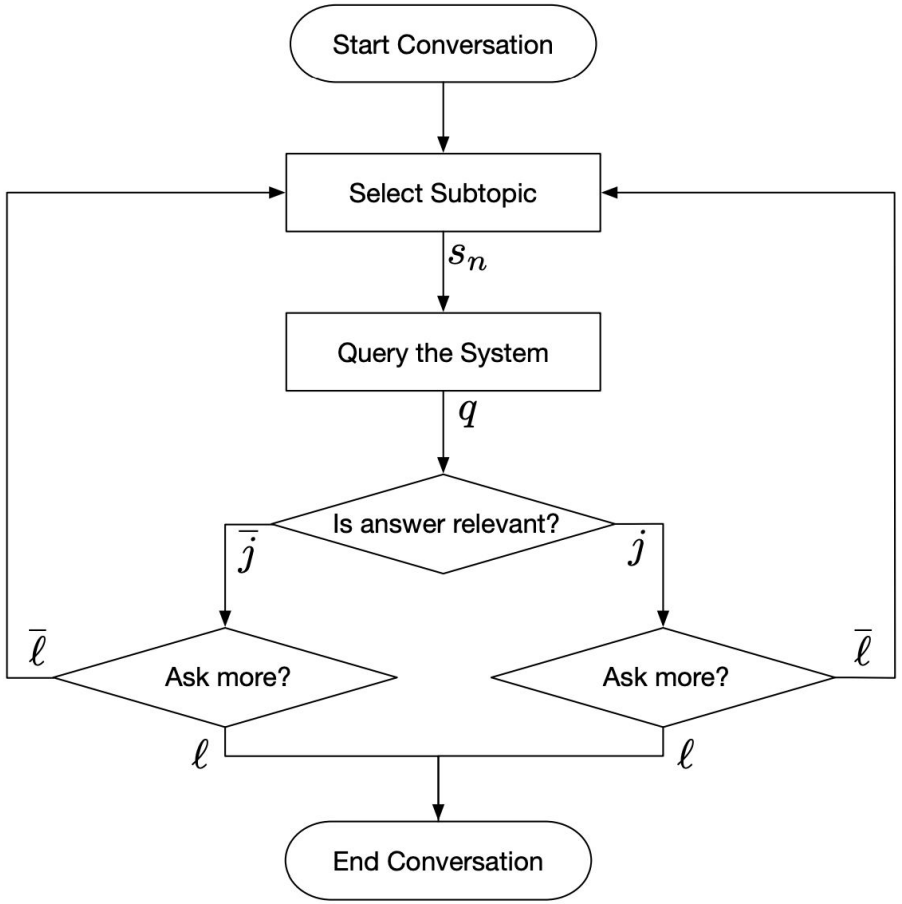
yarrow frost tolerance

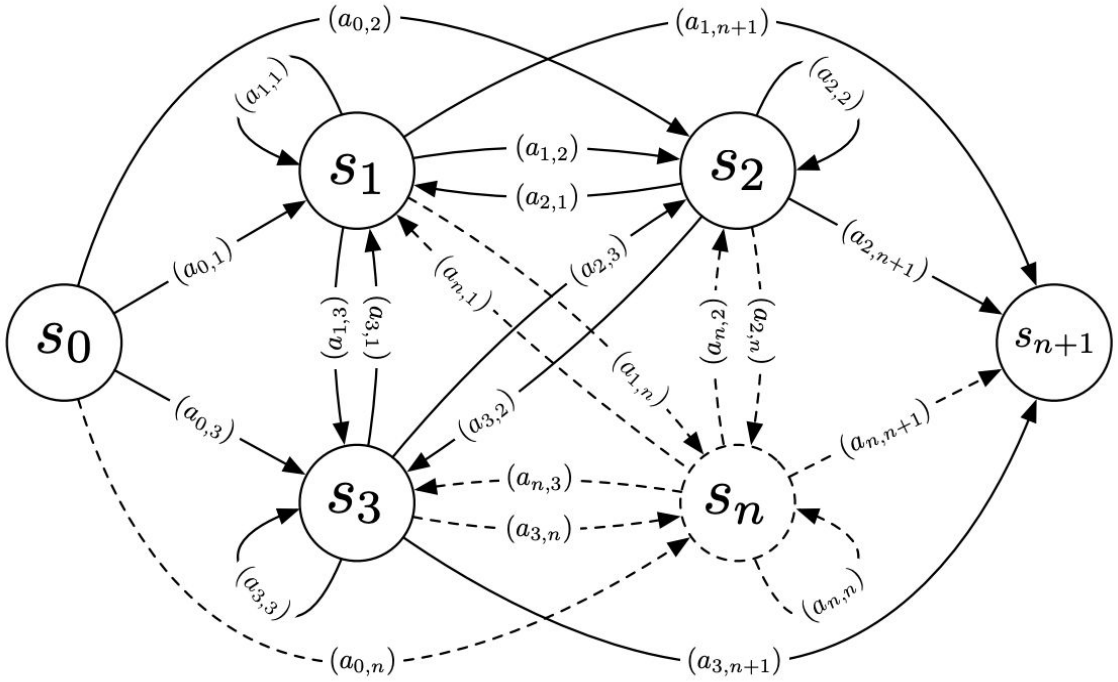
Q: Can it survive frost?

A: They can survive a light frost if they're sufficiently hardened off.

yarrow frost tolerance

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" αs

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.
- Crowdsource data for test collection queries and transition probabilities
- Evaluating conversations offline with test collections is not so hard anymore!