Grounded and Transparent Response Generation for Conversational Information-Seeking Systems

Weronika Łajewska, Krisztian Balog

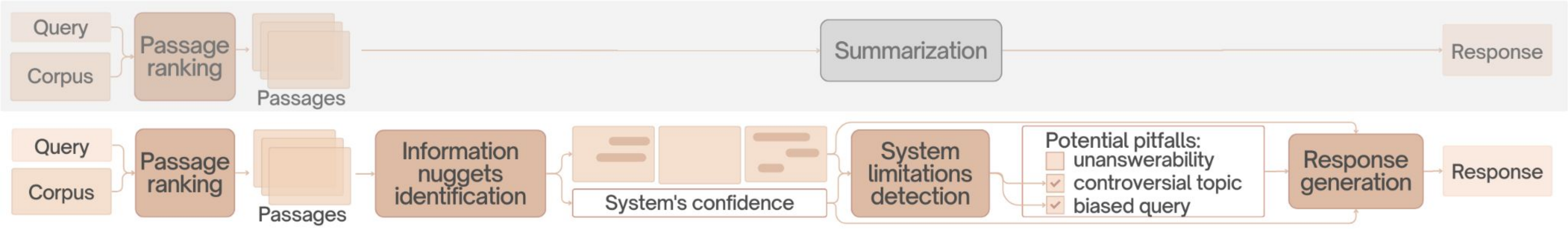
University of Stavanger, Norway

Our Motivation

- Conversational search is a less transparent setting that SERP-based interface
- Users are mostly not aware of the working mechanism of the system, its capabilities, and limitations
- Detecting hallucinations, factual errors, and/or biases in extremely difficult for users without knowledge about the topic

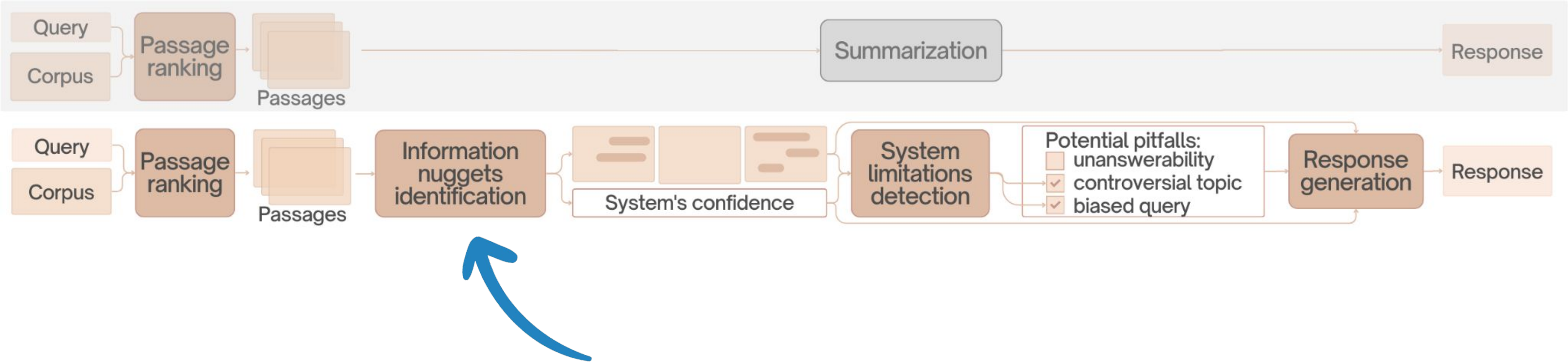
Overview of our Approach to Conversational Response Generation

"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati



Overview of our Approach to Conversational Response Generation

"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati



Towards Filling the Gap in Conversational Search: From Passage Retrieval to Conversational Response Generation

Weronika Łajewska and Krisztian Balog

University of Stavanger, Norway

CIKM'23, Birmingham

This Study

- **Problem setting:** Conversational response generation
    - It extends beyond passage retrieval + summarization
- **Goal:** snippet-level annotations of relevant passages, to enable
    1. the training of response generation models that are able to ground answers in actual statements
    2. the automatic evaluation of the generated responses in terms of completeness
- **Main contributions:**
    1. Crowdsourcing task design and protocol to collect high-quality annotations
    2. A dataset of 1.8k query-passage pairs annotated from the TREC 2020 and 2022 Conversational Assistance track

CAsT-snippets Sample

Query: I remember Glasgow hosting COP26 last year, but unfortunately I was out of the loop. What was the conference about?

Passage: HOME - UN Climate Change Conference (COP26) at the SEC – Glasgow 2021 Uniting the world to tackle climate change. The UK will host the 26th UN Climate Change Conference of the Parties (COP26) in Glasgow on 1 – 12 November 2021. The COP26 summit will bring parties together to accelerate action towards the goals of the Paris Agreement and the UN Framework Convention on Climate Change. The UK is committed to working with all countries and joining forces with civil society, companies and people on the frontline of climate change to inspire climate action ahead of COP26. COP26 @COP26 · May 25, 2021 1397069926800654339 We need to accelerate the #RaceToZero Join wef, MPPindustry, topnigel & gmunozabogabir for a series of events demonstrating the need for systemic change to accelerate the global transition to net zero. Starting May 27th Learn more #ClimateBreakthroughs | #COP26 Twitter 1397069926800654339 COP26 COP26 · May 24, 2021 1396737733649846273 #TechForOurPlanet is a new challenge programme for #CleanTech startups to pilot and showcase their solutions at #COP26! Innovators can apply to six challenges focusing around core climate issues and government priorities.

CAsT-snippets Sample

Query: I remember Glasgow hosting COP26 last year, but unfortunately I was out of the loop. What was the conference about?

Passage: HOME - UN Climate Change Conference (COP26) at the SEC – Glasgow 2021 Uniting the world to tackle climate change. The UK will host the 26th UN Climate Change Conference of the Parties (COP26) in Glasgow on 1 – 12 November 2021. The COP26 summit will bring

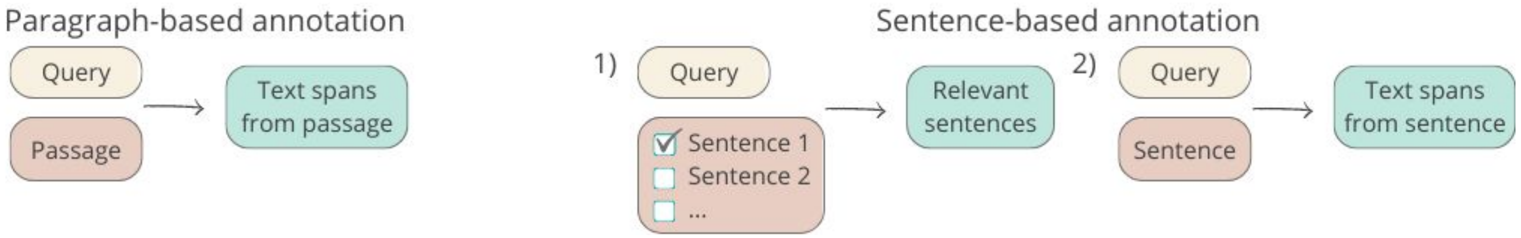
The seemingly straightforward task of highlighting relevant snippets turns out to be not that simple.

We need to accelerate the #RaceToZero Join wef, MPPindustry, topnigel & gmunozabogabir for a series of events demonstrating the need for systemic change to accelerate the global transition to net zero. Starting May 27th Learn more #ClimateBreakthroughs | #COP26 Twitter 1397069926800654339 COP26 COP26 · May 24, 2021 1396737733649846273 #TechForOurPlanet is a new challenge programme for #CleanTech startups to pilot and showcase their solutions at #COP26! Innovators can apply to six challenges focusing around core climate issues and government priorities.

Preliminary Study

A comparison of different task designs, platforms, and worker pools

- **Task designs**: paragraph-based vs. sentence-based annotation



- **Platforms and workers:**
  - Amazon MTurk (regular vs. master workers)
  - Prolific
  - Expert annotators (PhD students)

Main findings

- Relative ordering: MTurk masters > Prolific > MTurk regular
- Paragraph-level > sentence-level (w.r.t. similarity with expert annotations)

⇒ use MTurk and paragraph-based design for the large-scale data collection

Data collection

Setup

Employ a small group of trained crowd workers, selected through a qualification task, and create an extended set of guidelines with help of the annotators

Qualification task → **Discussion** → **Data collection**

Qualification task

Task consisted of: a detailed description of the problem, examples of correct annotations, a quiz, and 10 query-passage pairs to be annotated

20 workers completed/15 passed

Initial guidelines

Discussion

Feedback on qualification task

Extended guidelines

Data collection

Performed in daily batches (1 topic/batch =~46 HITs)

Individual feedback after each submitted batch

General comments/suggestions on a common Slack channel

$0.3 per HIT +$2 bonus for completing within 24h

Resulting Dataset: CAsT-snippets

371 queries, top 5 passages per query ⇒ **1855 query-passage pairs** (each annotated by 3 crowd workers)

- Data quality
  - Inter-annotator agreement exceeds even that of expert annotators
  - Similarity with expert annotations is on par with MTurk master workers
- Comparison against other datasets
  - More snippets annotated per input text; also, snippets are longer

Dataset	Input text	Avg. snippets length (tokens)	# snippets per annotation
CAsT-snippets	Paragraph	39.6	2.3
SaaC [1]	Top 10 passages	23.8	1.5
QuaC [2]	Wikipedia article	14.6	1

[1] Pengjie Ren, Zhumin Chen, Zhaochun Ren, E. Kanoulas, Christof Monz, and M. de Rijke. 2021. Conversations with Search Engines: SERP-based Conversational Response Generation. ACM Transactions on Information Systems 39, 4 (2021), 1–29
[2] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In Findings of the Association for Computational Linguistics: EMNLP 20 (EMNLP '18). 2174–2184.

Challenges Identified

Challenges pointed out by the crowd workers that need to be addressed in conversational response generation:

- Only a partial answer is present
- Temporal considerations
  - Spans may need to be excluded given the time constraints in the query
  - Assessing temporal validity can be challenging based on the paragraph alone (without larger context)
- Subjectivity of the passages originating from blogs or comments
- Indirect answers that require reasoning and background knowledge
- Determining the appropriate amount of context to include in each span
  - Balancing between being concise and being self-contained
- Determining whether the evidence or additional information is needed or an entity alone is sufficient as an answer

Summary

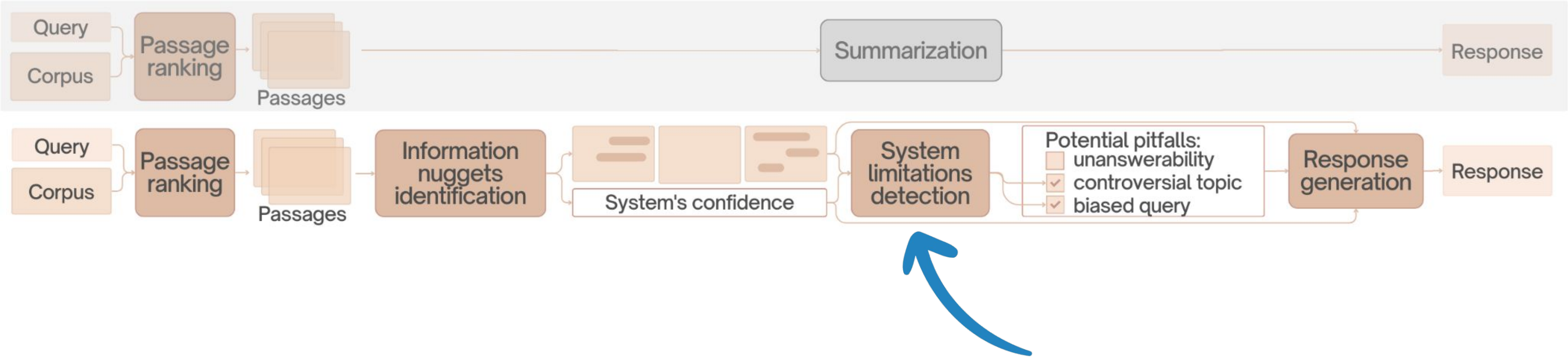
- Snippet-level annotations for conversational response generation (information-seeking queries)
- Several measures to ensure high data quality
  - Preliminary study to compare task variants and crowdsourcing platforms
  - Providing feedback and training to annotators throughout the data collection process
  - Incentive structure to engage crowd workers over a period of time and avoid worker fatigue
- Communication with workers also led to various insights regarding challenges in conversational response generation

---

Extended version on arXiv: https://arxiv.org/abs/2308.08911
Dataset: https://github.com/iai-group/CAsT-snippets

Overview of our Approach to Conversational Response Generation

"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati



Towards Reliable and Factual Response Generation: Detecting Unanswerable Questions in Information-Seeking Conversations

Weronika Łajewska, Krisztian Balog

University of Stavanger, Norway

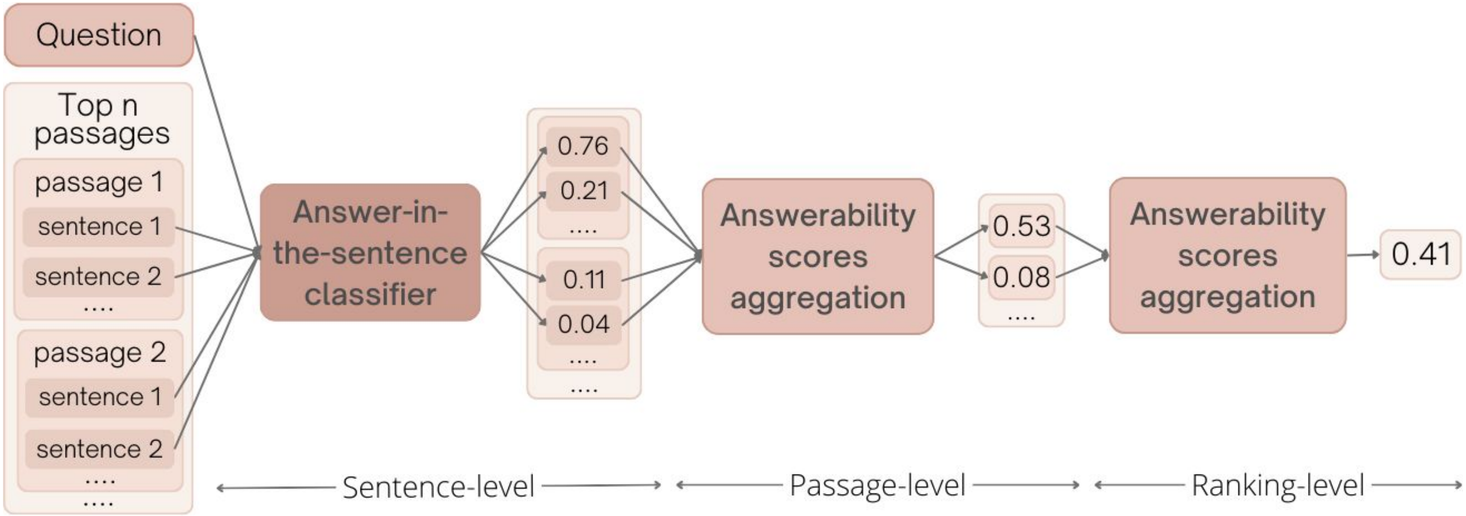
This Study

- **Problem setting:** Conversational response generation
- **Goal:** mechanism for detecting unanswerable questions for which the correct answer is not present in the corpus or could not be retrieved
- **Main contributions:**
  1. A dataset with answerability labels on three levels:
     1. sentences
     2. paragraphs
     3. rankings

	Answerable?	
	Yes	**No**
#question-sentence pairs (train+test)	6,395	19,043
#question-passage pairs (train+test)	1,778	1,932
#question-ranking pairs (test)	4,035	504

  2. A baseline approach for predicting answerability based on the top retrieved results.

Overview of our Answerability Detection Approach



Results

- Data augmentation helps answerability detection only on sentence and answer levels
- *Max* aggregation on the passage level followed by *mean* aggregation on the ranking level gives the best results
- LLMs have a limited ability to detect answerability without additional guidance.

Classifier	Sentence	Passage		Ranking	
	Acc.	Aggr.	Acc.	Aggr.	Acc.
CAsT-answerability	0.752	Max	0.634	Max	0.790
				Mean	**0.891**
		Mean	0.589	Max	0.332
				Mean	0.829
CAsT-answerability augmented with SQuAD 2.0	**0.779***	Max	0.676*	Max	0.810*
				Mean	0.848*
		Mean	0.639*	Max	0.468*
				Mean	0.672*
ChatGPT passage-level (zero-shot)			**0.787***	T=0.33	0.839*
				T=0.66	0.623*
ChatGPT ranking-level (zero-shot)					0.669*
ChatGPT ranking-level (two-shot)					0.601*

Results

Does data augmentation help answerability detection?

- Data augmentation helps answerability detection only on sentence and answer levels
- *Max* aggregation on the passage level followed by *mean* aggregation on the ranking level gives the best results
- LLMs have a limited ability to detect answerability without additional guidance.

Classifier	Sentence	Passage		Ranking	
	Acc.	Aggr.	Acc.	Aggr.	Acc.
CAsT-answerability	0.752	Max	0.634	Max	0.790
				Mean	**0.891**
		Mean	0.589	Max	0.332
				Mean	0.829
CAsT-answerability augmented with SQuAD 2.0	**0.779***	Max	0.676*	Max	0.810*
				Mean	0.848*
		Mean	0.639*	Max	0.468*
				Mean	0.672*
ChatGPT passage-level (zero-shot)			**0.787***	T=0.33	0.839*
				T=0.66	0.623*
ChatGPT ranking-level (zero-shot)					0.669*
ChatGPT ranking-level (two-shot)					0.601*

Results

Which of the two aggregation methods performs better?

- Data augmentation helps answerability detection only on sentence and answer levels
- *Max* aggregation on the passage level followed by *mean* aggregation on the ranking level gives the best results
- LLMs have a limited ability to detect answerability without additional guidance.

Classifier	**Sentence**	**Passage**		**Ranking**	
	Acc.	**Aggr.**	**Acc.**	**Aggr.**	**Acc.**
CAsT-answerability	0.752	Max	0.634	Max	0.790
				Mean	$\boxed{\mathbf{0.891}}$
		Mean	0.589	Max	0.332
				Mean	0.829
CAsT-answerability augmented with SQuAD 2.0	**0.779***	Max	0.676*	Max	0.810*
				Mean	0.848*
		Mean	0.639*	Max	0.468*
				Mean	0.672*
ChatGPT passage-level (zero-shot)			**0.787***	T=0.33	0.839*
				T=0.66	0.623*
ChatGPT ranking-level (zero-shot)					0.669*
ChatGPT ranking-level (two-shot)					0.601*

Results

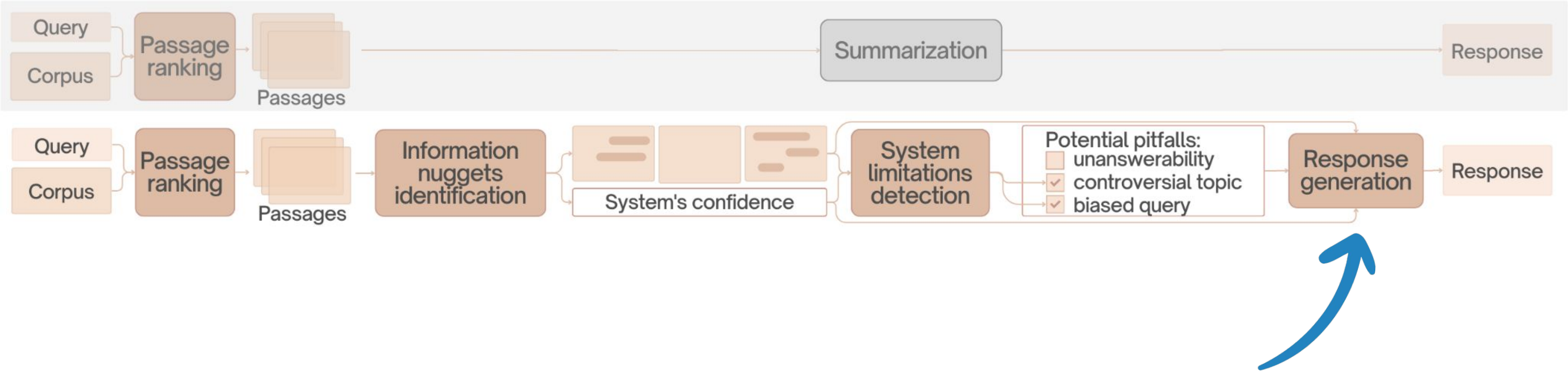
How competitive are these baselines in absolute terms?

- Data augmentation helps answerability detection only on sentence and answer levels
- *Max* aggregation on the passage level followed by *mean* aggregation on the ranking level gives the best results
- LLMs have a limited ability to detect answerability without additional guidance.

Classifier	Sentence	Passage		Ranking	
	Acc.	Aggr.	Acc.	Aggr.	Acc.
CAsT-answerability	0.752	Max	0.634	Max	0.790
				Mean	**0.891**
		Mean	0.589	Max	0.332
				Mean	0.829
CAsT-answerability augmented with SQuAD 2.0	**0.779***	Max	0.676*	Max	0.810*
				Mean	0.848*
		Mean	0.639*	Max	0.468*
				Mean	0.672*
ChatGPT passage-level (zero-shot)			**0.787***	T=0.33	0.839*
				T=0.66	0.623*
ChatGPT ranking-level (zero-shot)					0.669*
ChatGPT ranking-level (two-shot)					0.601*

Overview of our Approach to Conversational Response Generation

"A true teacher would never tell you what to do. But he would give you the knowledge with which you could decide what would be best for you to do."
— Christopher Pike, Sati



A User-Centric Analysis of Response Generation Challenges in Conversational Information-Seeking

Weronika Łajewska[1], Krisztian Balog[1], Damiano Spina[2], Johanne Trippas[2]

[1]*University of Stavanger, Norway*
[2]*RMIT University, Melbourne, Australia*

This Study

- **Problem setting:** Response generation in conversational information-seeking (CIS) scenario
- **Goal:** investigating the ability of users to recognize pitfalls in CIS responses
- **Research questions:**
    1. Can users effectively recognize the problem of query answerability and the problem of multiple viewpoints leading to response incompleteness in system responses?
    2. How do inaccurate, incomplete, and/or biased responses impact user experience?
- **Main contribution:**
    1. A novel methodology to study how users perceive query answerability and response incompleteness in CIS

Answerability Study

Query: I like hiking and Malbec wine. You mentioned some high peaks. How can I hike some high mountains and visit some wineries famous for Malbec?

Experimental Condition	Response Dimension		CIS System Response	Source
	Factual Corr.	Source		
Factually correct + valid source	✓	✓	*You can combine your love for hiking and Malbec wine by visiting Mendoza, Argentina. This picturesque city is nestled in the Andes and is renowned for its vineyards...*	https://wanderingtrader.com/argentina/top-5-argentina-tourist-attractions/
Factually correct + no source	✓	×	Same as above	–
Factually incorrect + invalid source	×	✓(invalid)	*To combine hiking and Malbec wine, plan a trip to the Penedès region in Catalonia. You can explore the Montserrat mountain range, which offers fantastic hiking opportunities, and then visit renowned wineries in the Penedès, known for its exceptional Malbec wine production. . .*	https://www.winetourism.com/wine-tasting-tours-in-penedes/ (The link is valid but the article is a website with Wine Tasting & Tours in Penedès, Spain where Malbec wine is not produced.)
Factually incorrect + no source	×	×	Same as above	–

Viewpoints Study

Query: What effects did the Watergate scandal have on President Nixon's legacy?

Experimental Condition	Response Dimension		CIS System Response
	Diversity	Balance	
Multiple viewpoints covered to the same extent	✓	✓	*The Watergate scandal initially had a severely damaging effect on President Nixon's legacy, tarnishing his image as a corrupt politician and overshadowing his accomplishments. However, in later years, Nixon's reputation underwent a degree of rehabilitation, with recognition of his roles in civil rights and diplomacy. He transitioned from being primarily associated with scandal to being acknowledged as a respected elder statesman, although his legacy remains marked by the shadow of Watergate.*
Multiple viewpoints mentioned, one covered in detail	✓	×	*The Watergate scandal had a profoundly negative impact on President Nixon's legacy, overshadowing many of his domestic achievements. It tarnished his reputation as a corrupt politician, making him a symbol of political scandal and misconduct in both American politics and popular culture. While he did enact significant legislation like creating the Environmental Protection Agency, his presidency is primarily remembered for the scandal, leading to his resignation and becoming synonymous with political corruption and disgrace.*
Single viewpoint mentioned and covered	×	×	*The Watergate scandal had a profoundly negative impact on President Nixon's legacy. It tarnished his reputation as a corrupt politician, making him a symbol of political scandal and misconduct in both American politics and popular culture. Nixon's resignation and the scandal's fallout reinforced public skepticism and criticism of the presidency, leaving a lasting impression as one of the most Shakespearean and disgraceful episodes in presidential history.*

Findings

- **Quantitative analysis →** users find it easier to identify problems with diversity and balance of viewpoints rather than factual errors and source validity in the responses
- **Analysis of a user experience →** self-reported overall satisfaction scores are not necessarily associated with the main response dimensions
- **Qualitative analysis of free-text comments →** credibility of the sources, as well as completeness, usefulness, and subjectivity of provided information impact the overall satisfaction of the users

Dependent Variable	Independent Variable(s)	F	*p*-value	Unbalanced Estimator	Effect Size
Answerability Study					
Factual Correctness	Experimental Condition	1.330	0.264	0.003	–
Confidence in Answer Accuracy		0.721	0.540	−0.002	–
Overall Satisfaction		1.190	0.313	0.002	–
Viewpoints Study					
Diversity	Experimental Condition	**31.774**	**0.000**	**0.186**	**L**
Transparency		**21.751**	**0.000**	**0.133**	**M**
Balance		**17.514**	**0.000**	**0.109**	**M**
Overall Satisfaction		**17.687**	**0.000**	**0.110**	**M**

Dependent Variable	Explanatory Variables	*p*-value
Answerability Study		
Overall Satisfaction	Familiarity	0.248
	Factual Correctness	0.069
	Confidence in Answer Accuracy	**0.012**
Viewpoints Study		
Overall Satisfaction	Familiarity	0.972
	Diversity	0.209
	Transparency	0.436
	Balance	**0.003**

⇒ simple source attribution is not enough to ensure effective interaction with the system

Communicating System Limitations in Conversational Response Generation

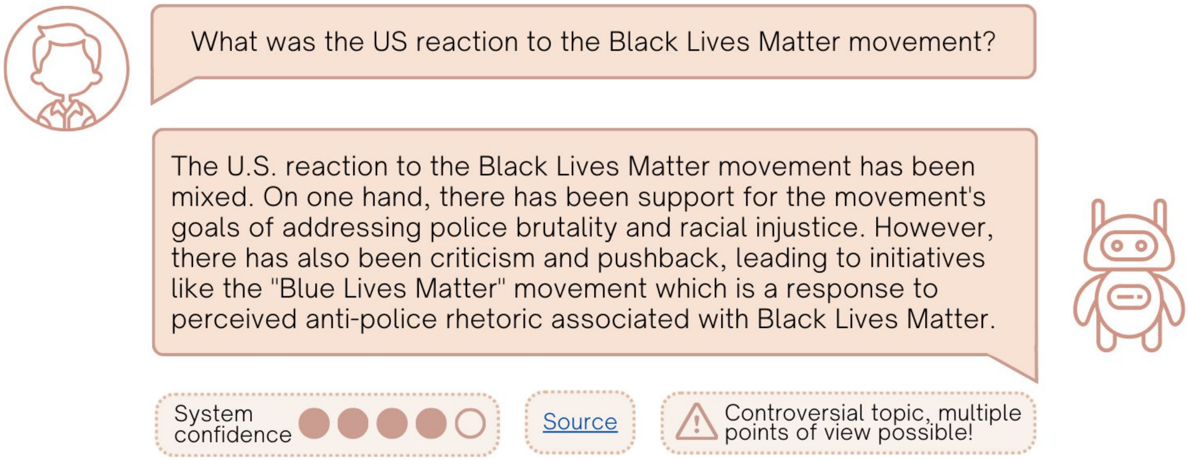
Weronika Łajewska[1], Krisztian Balog[1], Damiano Spina[2], Johanne Trippas[2]

[1]University of Stavanger, Norway
[2]RMIT University, Melbourne, Australia

This Study

Research questions:

1. How to inform user about the limitations/capabilities of the system and its confidence?
2. Given that the system reveals its capabilities and limitations, how to evaluate this enhanced response?



Open Questions

What should be included in a system response and what is the most effective way of presenting this additional information?

- Pre-use tutorial describing the system and its capabilities
- Information about potential problems/limitations of the provided response
- Information about limitations of the user query
- Information about system confidence in the provided response
- Source of the information used for generating the response
- …

Which dimensions of the response are most important?

- usefulness
- completeness
- diversity
- transparence
- …

Thank you for your attention!

Questions?