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

Weronika Łajewska
Krisztian Balog

Information Access and Artificial Intelligence

University of Stavanger

Paper:

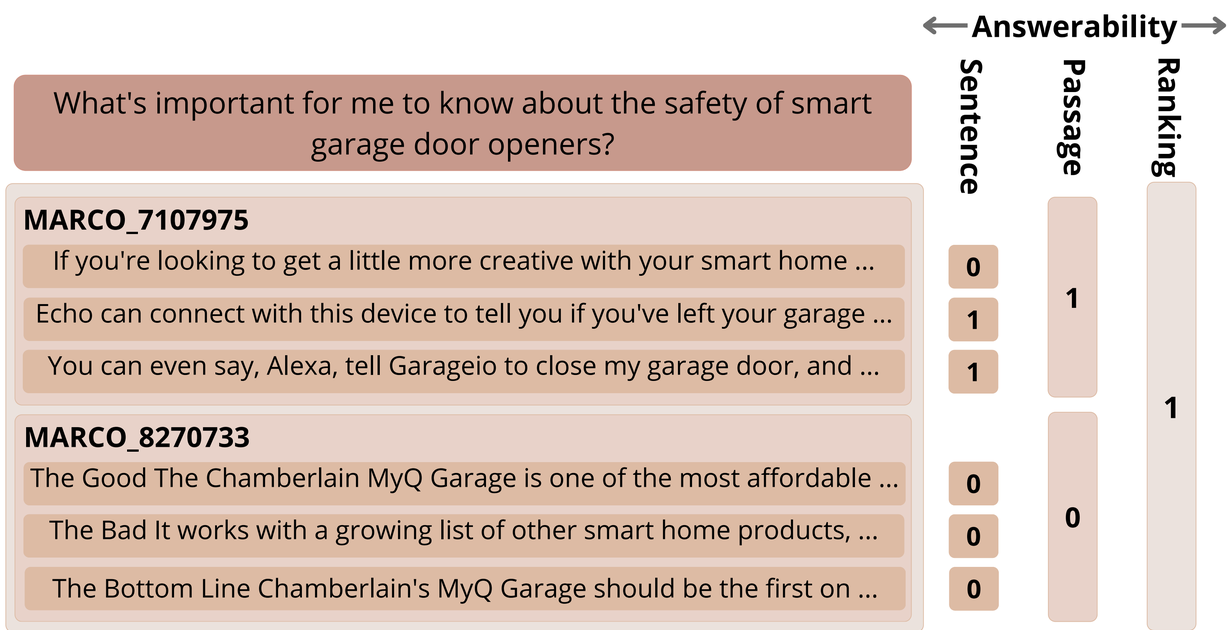
Repository:

MOTIVATION

- Conversational Information-Seeking (CIS) focuses on systems designed for dialogue-based information retrieval, where the aim is to enable interactions that closely resemble human conversation
- The answer is typically not confined to a single entity or text snippet, but rather spans across multiple sentences or even multiple passages
- Answerability extends beyond the general notion of relevance and asks for the presence of a specific answer
- The answer to the user's question may not always be contained in the top retrieved passage
- Response generated from passages not containing the answer may result in hallucinations

CONTRIBUTIONS

- We proposed a mechanism for detecting unanswerable questions for which the correct answer is not present in the corpus or could not be retrieved
- We extended the *CAsT-snippets* [1] dataset with answerability labels on the sentence, passage, and ranking levels
- We proposed a baseline approach for predicting answerability based on the top retrieved results

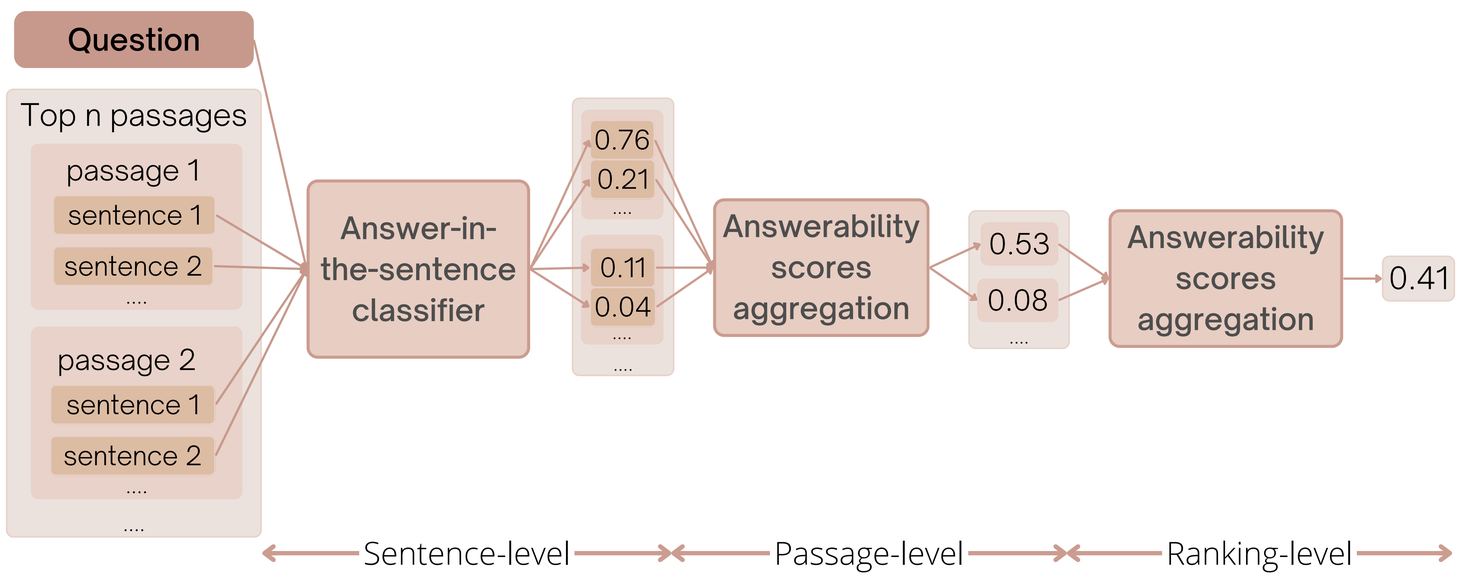


CAST-ANSWERABILITY DATASET

- 5 relevant (based on snippet-level answer annotations *CAsT-snippets* dataset [1]) and 5 non-relevant passages for each query
- Answerability labels on three levels: 1) sentence, 2) passage, and 3) ranking
- For ranking-level answerability all possible 3-element subsets of passages available for this question (both containing and not containing an answer) are considered

ANSWERABILITY DETECTION BASELINE

- Answer-in-the-sentence classifier is trained on sentence-level data
- The output of the classifier is the probability that the sentence contains (part of) the answer to the question
- The classifier is built using a BERT transformer model with a sequence classification head on top
- Sentence-level estimates are aggregated on the passage level and then further on the ranking level to determine whether the question is answerable



RESULTS

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 + SQuAD 2.0	**0.779***	Max	0.676*	Max	0.810*
				Mean	0.848*
		Mean	0.639*	Max	0.468*
				Mean	0.672*
ChatGPT (zero-shot)			**0.787***	T=0.33	0.839*
				T=0.66	0.623*
ChatGPT (zero-shot)					0.669*
ChatGPT (two-shot)					0.601*

Max aggregation on the passage level followed by *mean* aggregation on the ranking level gives the best results

Data augmentation helps answerability detection only on sentence and passage levels

LLMs have a limited ability to detect answerability without additional guidance

CAST-ANSWERABILITY IN NUMBERS

	Answerable	Unanswerable
# 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

[1] W. Łajewska and K. Balog. 2023. Towards Filling the Gap in Conversational Search: From Passage Retrieval to Conversational Response Generation. In: 32nd ACM International Conference on Information and Knowledge Management (CIKM'23)