Shortcomings of QA-Based Factuality Frameworks for Error Localization

TAUR LAB @UT Austin

Ryo Kamoi, Tanya Goyal, Greg Durrett *ryokamoi@utexas.edu*

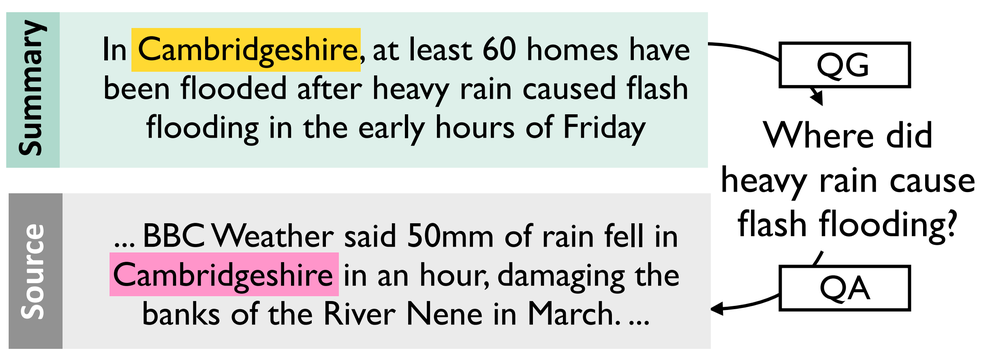
QA frameworks can detect factuality errors in summarization.
But can they localize errors? No.

The question generation (QG) step introduces fundamental problems,
and better QG will not improve span-level evaluation.

Background: QA-based Factuality Evaluation

Goal: detect factuality errors in generated summaries

- generate questions whose answers are NPs/NEs in summary
- answer these questions by referring to the source doc



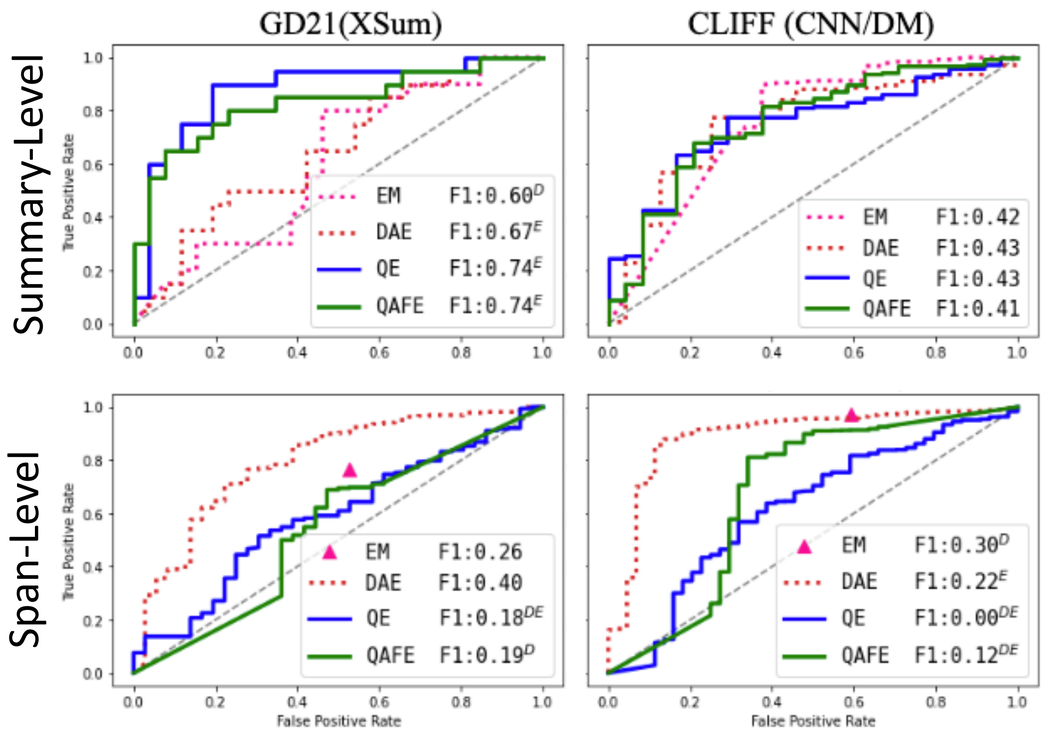
Answers from the source and summary are the same, so we expect that two have identical information.

$$\frac{1}{|S|}\sum_{\text{span}\in S}\text{score(span)} = \text{similarity(}\quad\text{,}\quad\text{)}$$

Due to this span-level decomposition, some prior work asserts that QA frameworks can localize errors.

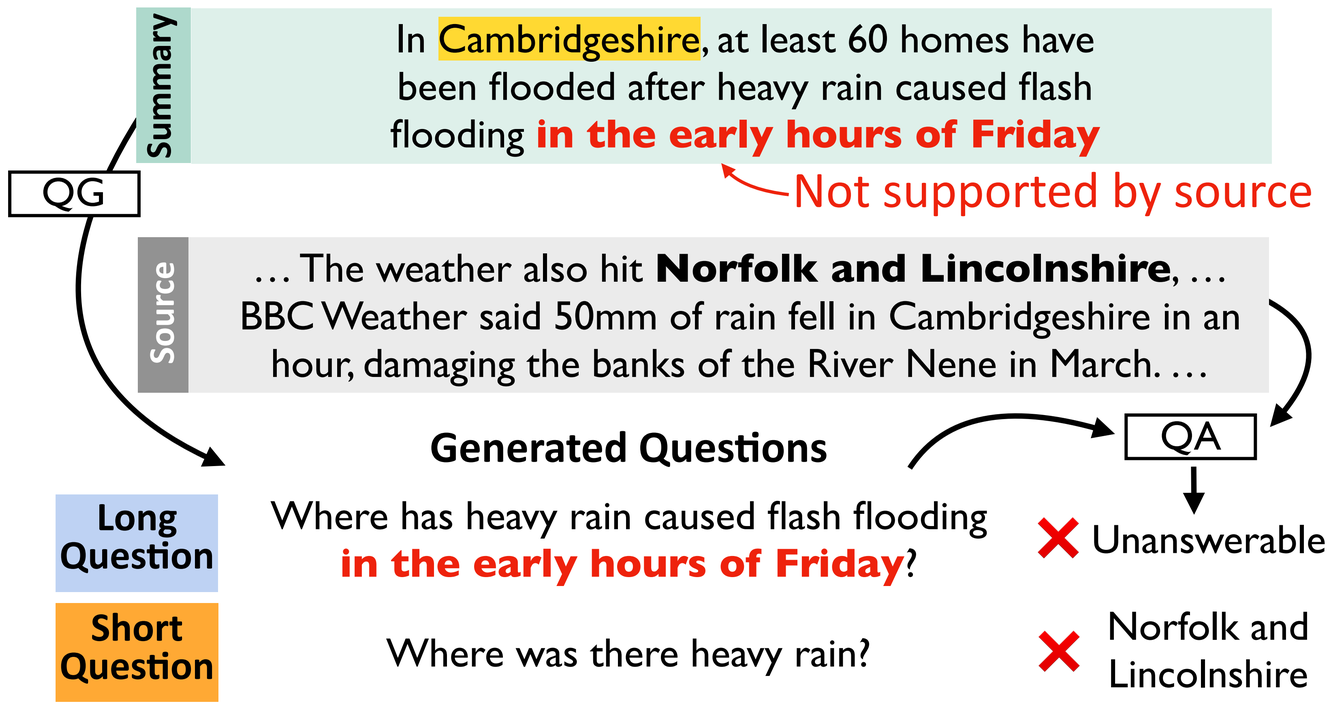
Can QA frameworks localize errors? No.

QA frameworks are good at **summary-level**. However, **span-level** evaluation of QA frameworks is worse than a naive exact match baseline.



We evaluated two QA frameworks (**QE** and **QAFE**) on three datasets (GD21, two subsets of CLIFF)

Why do QA metrics fail to localize errors?



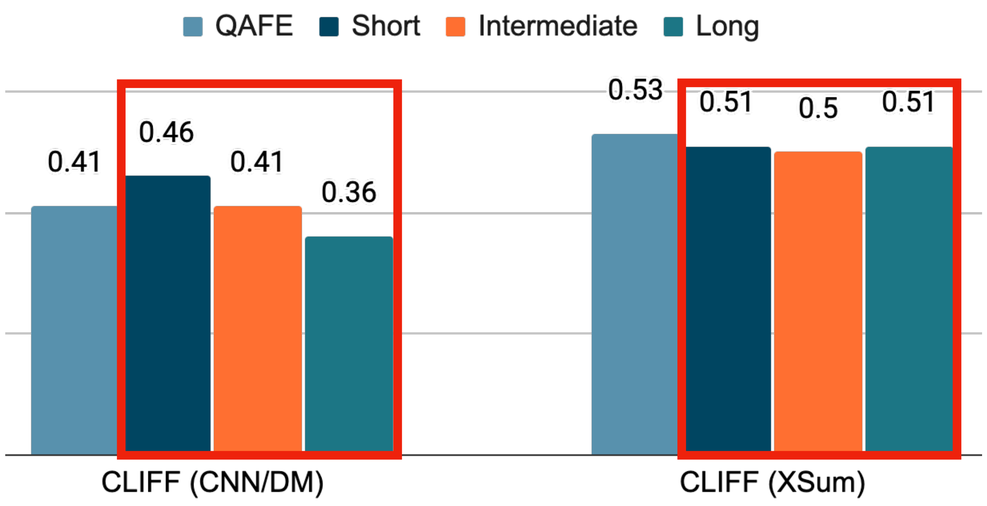
Long questions tend to **inherit errors** from erroneous summaries
Short questions can be **underspecified** when referring to source

Our analysis shows that up to 99% of questions inherit errors depending on the dataset and QA framework, and inherited errors degrade span-level performance.

Can Better QG Improve Error Localization? No.

QG cannot avoid this problem by varying the question length because QG does not know which part of summary is not factual.

We replace QG with human generated questions to test this.
Human QG does not improve span-level performance.



For each span, annotators generate multiple questions with different lengths (short, intermediate, long), but there is no single setting that improves span-level performance.

More Results and Analysis in the Full Paper

We do not see a reason to prefer QA frameworks over entailment-based metrics for summary factuality evaluation.
QA frameworks provide no localization benefit.