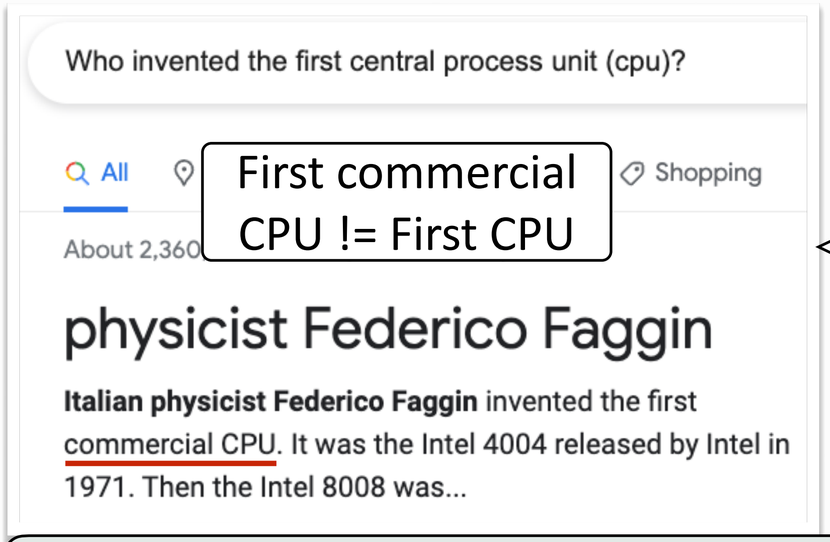
Can NLI Models Verify QA Systems' Predictions?

Jifan Chen, Eunsol Choi, Greg Durrett

TAUR LAB @UT Austin

Motivation

QA models sometimes return unreliable answers



If we can reformulate a question + proposed answer as an NLI problem, we can use off-the-shelf NLI models to check these answers and spot these mistakes.

NLI as a verifier

Question: What is the revolution period of Venus in earth days?
Answer: 243 days

Context: Venus is the second planet from the Sun… . **It** has the longest rotation period (243 days)…

Question-to-statement
Error rate: < 5%

T5

T5

Decontextualization
Error rate: < 10%

Hypothesis: The **revolution** period of Venus in earth days is 243 days.

Premise: Venus has the longest **rotation** period (243 days)…

To train the QA-NLI: gold answer and its context as positives, bad ones as negatives

Natural Language Inference

Not entailed

Model can leverage both sentence-level NLI data (e.g., MNLI) and QA datasets converted to NLI.

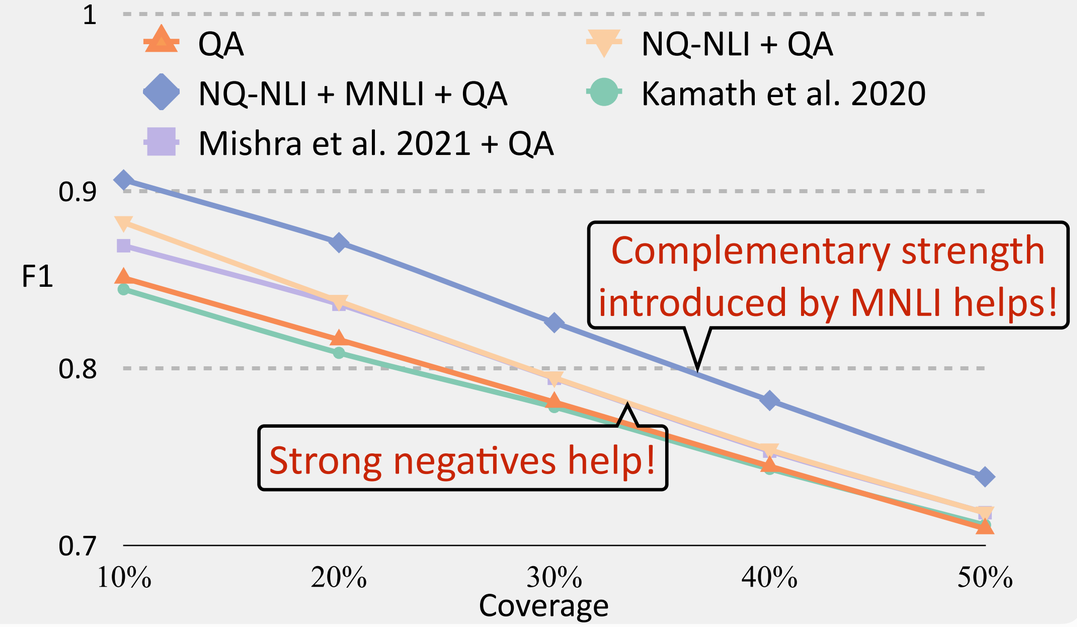
What can the model do?

Reject unanswerable questions

SQuAD2.0 → **SQuAD 1.1 QA model** → **MNLI** → Reject **78.5%** unanswerable. Accept **82.5%** answerable.

Improve prediction confidence

Selective QA: use NLI to get confidence in answer, assess performance if we answer the k% of questions with highest confidence



Detect information mismatches

Question: When was Clash Royale released in the US?
Answer: The game Clash Royale was released globally on **March 2 , 2016** Globally —> US? Need further evidence

"March 2, 2016" was labeled as the gold answer in the QA dataset, but it doesn't pass the NLI check. Are we teaching QA models to make unreliable logical leaps?