Challenges in Information-seeking QA: Unanswerable Questions and Paragraph Retrieval

Akari Asai[1], Eunsol Choi[2]

[1]University of Washington, [2]The University of Texas at Austin

Code & Data

Contact

akari@cs.washington.edu

@AkariAsai

Summary: We present comprehensive analysis on information-seeking QA datasets, namely Natural Questions[1] and TyDi QA [2].

- Our **Gold-Type & Gold-Paragraph** experiments show answerability prediction and paragraph retrieval remain hard.
- Our **Answerability Prediction** experiments shows unanswerable questions in information-seeking QA have unique characteristics, which make answerability prediction challenging even given gold context.
- Our **Unanswerability Annotation** reveal where the unanswerability arise and how we could improve answer coverage.

Background

- **Machine Reading Comprehension** questions are written by annotators who know the answers given a paragraph.

What was Maria Curie the first female recipient of?

One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who was the first female recipient of the **Nobel Prize**.

→ Nobel Prize

- **Information-seeking QA** questions are written by annotators who don't know the answers.

who is the voice of tony the tiger?

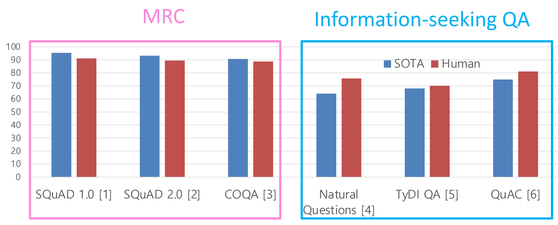
Thurl Ravenscroft - en.wikipedia

Thurl Arthur Ravenscroft was an American actor and bass singer known as the booming voice behind Kellogg's Frosted Flakes animated spokesman Tony the Tiger for more than five decades.

His voice acting career began in 1940.

→ Thurl Arthur Ravenscroft

`unanswerable`



SOTA models still struggle in information-seeking QA!!

Paragraph retrieval

Answerability Prediction

Gold Type / Gold Paragraph Experiments

Settings

- **Gold Type:** provides the answer type (-answerability prediction)
- **Gold Paragraph**: provides the gold paragraph (-paragraph retrieval)
- **Gold Type + Gold Paragraph**

Models

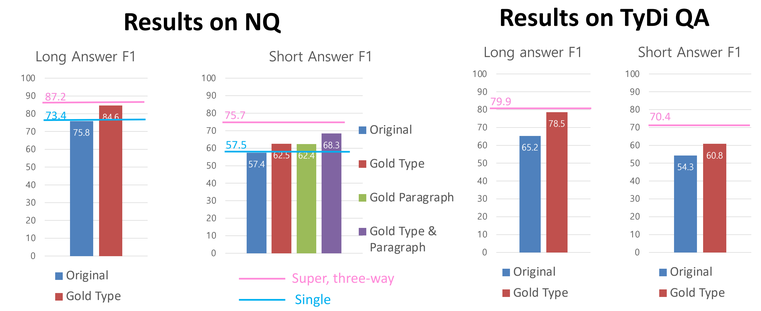
- **ETC** [7] for NQ
- **Multilingual BERT-based** model [5, 8] for TyDi QA

Human performance

- **Super** (25 way), **Single** (5 way) for NQ
- **3-way** human annotation for TyDi QA

Gold Type / Gold Paragraph Experiments

Given Gold Type and Gold Paragraph, our models outperform single human and perform on par with super / 3-way human.



Answerability Prediction Experiments

Task setup: predicting if a question is answerable or not

- **Question-only** : only use question input
- **Full-input**: use both a question and a document

Datasets

- **Natural Questions, TyDi QA**
- **SQuAD 2.0** (MRC)

Unanswerable Questions are artificially crated by annotators to confuse models

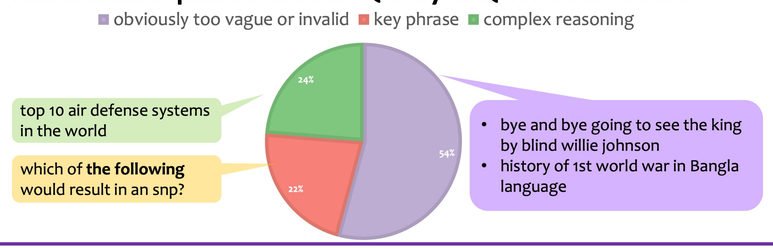
Models

- **Majority**
- **Question-only**
- **QA models**: ETC for NQ, RetreoReader [9] for SQuAD 2.0, multilingual BERT-based model for TyDi.
- **Human**

Results

Dataset	Majority	Question Only	QA model	Human
Natural questions	58.9	72.7	82.5	85.6
TyDi QA	58.2	70.2	79.4	94.0
SQuAD 2.0	50.0	63.0	94.1	--

Unanswerable questions both Q-only & QA models detect



Annotating Unanswerability – Category & Statistics

Categories of Unanswerability

[1] Retrieval Miss

Questions paired with a document that do not contain a single gold paragraph and answer.

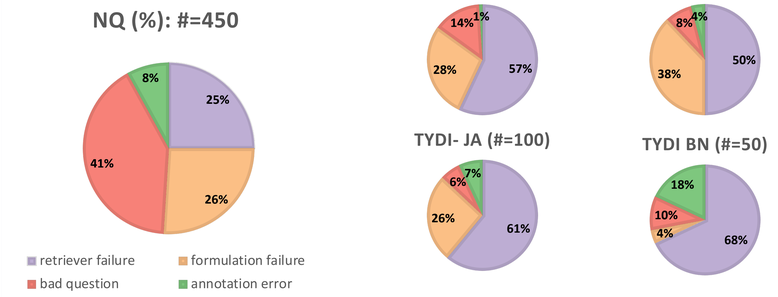
- Factoid question (retriever failure)
- Non-factoid question (formulation failure)
- Multi-evidence question (formulation failure)

[2] Invalid QA

Unanswerable questions, where annotated QA data is partially incorrect or ill-formed.

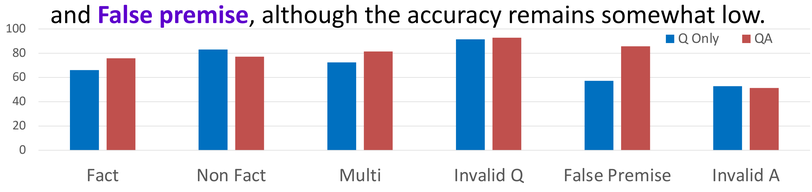
- Invalid question (bad question)
- False premise (bad question)
- Invalid answers (annotation error)

Annotation Results



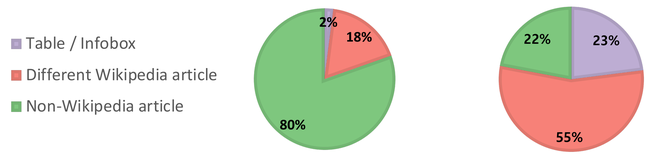
Per-category Answerability Prediction

- **Invalid question** can be easily identified even from question only.
- **Invalid answers** is hard even context information is given.
- Context input helps in **Factoid question**, **Multi-evidence question** and **False premise**, although the accuracy remains somewhat low.



Discussion: How we can improve answer coverage?

1. **New Task formulation:** extractive QA over single document may not cover all of the diverse information-seeking questions.
   → Generative answer, multi-paragraph evidence
2. **Alternative knowledge source**: improve answer coverage by using additional knowledge sources.



Reference

[1] Rajpurkar et al. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*. / [2] Rajpurkar et al. (2018). Know What You Don't Know: Unanswerable questions for SQuAD. In *ACL*. / [3] Reddy et al. (2019). CoQA: Conversational Question Answering Challenge. *TACL*.
[4] Kwiatkowski et al. (2019). Natural Questions: a Benchmark for Question Answering. *TACL*.
[5] Clark et al. (2020). TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *TACL*.
[6] Choi et al. (2018). QuAC: Question Answering in Context. In *EMNLP*. / [7] Ainslie et al. (2020). ETC: Encoding Long and Structured Inputs in Transformers. In *EMNLP*. / [8]Alberti et al. (2019). A BERT Baseline for the Natural Questions.
[9] Zhang et al. (2021). Retrospective Reader for Machine Reading Comprehension. In *AAAI*.