Do We Know What We Don't Know? Studying Unanswerable Questions beyond SQuAD 2.0

Elior Sulem, Jamaal Hay and Dan Roth



EMNLP-Findings 2021

Unanswerable Questions in Extractive QA

The Iroquois sent runners to the manor of William Johnson in upstate New York. The British Superintendent for Indian Affairs in the New York region and beyond, Johnson was known to the Iroquois as Warraghiggey, meaning "He who does great things." He spoke their languages and had become a respected honorary member of the Iroquois Confederacy in the area. In 1746, Johnson was made a colonel of the Iroquois. Later he was commissioned as a colonel of the Western New York Militia. They met at Albany, New York with Governor Clinton and officials from some of the other American colonies. Mohawk Chief Hendrick, Speaker of their tribal council, insisted that the British abide by their obligations and block French expansion. When Clinton did not respond to his satisfaction, Chief Hendrick said that the "Covenant Chain", a long-standing friendly relationship between the Iroquois Confederacy and the British Crown, was broken.

What was William Johnson's Sioux name?

I don't know

SQuAD 2.0 (Rajpurkar et al., 2018)

- Current systems trained on SQuAD 2.0 achieve good in-domain performance. For example, a system based on BERT-LARGE (Devlin et al., 2019) achieves **80.96 F1** (Has answer: 83.53 F1; No-answer: 78.40 F1) on the SQuAD 2.0 dev set.

Unanswerable Questions beyond SQuAD 2.0

- Informative evaluation requires out-of-domain test sets
  - Linzen, 2020
  - Dunietz et al., 2020
- QA applications involve out-of-domain test sets
  - Zero-shot event extraction (Lyu et al., 2021)
  - Evaluation of summarization (Deutsch et al. 2021)

New Event-Based Test Dataset

- Compiling a test event corpus for wh-questions - **ACE-whQA**, derived from ACE, focusing on time and location: 734 examples
  - **Has-answer:** Sentences that include the answer to the time or location-related question.

"She lost her seat in the **1997** election."
When was the loss?

  - **Competitive IDK:** Sentences with an entity of the same type as the expected answer.

"She travelled to **Mexico** after she lost her seat in the 1997 election"
Where was the loss?

  - **Non-Competitive IDK:** Sentences with no entity of the same type as the expected answer.

"He was arrested for his crimes"
When was the arrest?

Out-of-domain Performance

- Evaluation on ACE-whQA:
  - Low performance of a top system trained on SQuAD 2.0 (Rajpurkar et al., 2018)
  - First training on Textual Entailment (Dagan et al., 2013) that includes an IDK option ("neutral") improves the performance, in particular for non-competitive IDK questions.
  - This improvement is not replicated in the case of Binary TE (contradiction/non-contradiction).

	Baseline	Using TE	Using Binary TE
train / test	SQuAD 2.0	MNLI + SQuAD 2.0	c(MNLI) +SQuAD 2.0
Has Answer	68.75	71.68	78.13*
Compet. IDK	20.80	46.40*	26.00
Non-Compet. IDK	28.46	75.61*	47.15

F1 scores of the BERT-LARGE system evaluated on ACE-whQA.
* Significantly higher than the baseline ($p<0.05$)

Conclusion

- We provide a new test set to evaluate the ability of Extractive QA systems to identify unanswerable questions, beyond the SQuAD 2.0 domain.
- We find that SQuAD 2.0 alone is not sufficient to address IDK in these cases, even in the easy ones.
- We show that leveraging Textual Entailment can be useful, particularly for the easy cases.

Thank you for listening

The data and models can be found at
http://cogcomp.org/page/publication_view/955

eliors@seas.upenn.edu
https://www.cis.upenn.edu/~eliors/