Penn COGNITIVE COMPUTATION GROUP

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

Elior Sulem, Jamaal Hay and Dan Roth

Department of Computer and Information Science
University of Pennsylvania

EMNLP Findings 2021

Unanswerable Questions in Extractive QA

- **Extractive QA**: A system must extract a correct answer to a question from a context paragraph or document.

> **Context**: John was born in New York.
> **Question:** Who was born in New York?
> **Answer:** John

- **Unanswerable Questions (IDK)**: Cases where the answer is not in the sentence.

> **Context**: John was born in New York.
> **Question:** Who was born in France?
> **Answer:** IDK

- **Existing Dataset**: SQuAD 2.0 (Rajpurkar et al., 2018)
  - Includes unanswerable questions
  - Contexts are multi-sentence paragraphs

Split	#Examples	IDK Prportion (%)
Train	130,319	33
Dev	11,873	50

Statistics for the SQuAD 2.0 dataset

ACE-whQA

New Test Dataset

We compile a test corpus for wh-questions - **ACE-whQA**, derived from ACE 2005 (Walker et al., 2006) , focusing on **time and location event arguments**. The contexts are single sentences.
It is composed of three portions:

1. **Has Answer**: The sentences include the answer to the time or location-related question.

> **Context**: She lost her seat in the 1997 election.
> **Question:** When was the loss?
> **Answer:** 1997

2. **Compet. IDK**: The sentences include an entity of the same type as the expected answer.

> **Context**: She travelled to **Mexico** after she lost her seat in the 1997 election
> **Question:** Where was the loss?
> **Answer:** IDK

3. **Non-compet. IDK:** The **se**ntences have no entity of the same type as the expected answer.

> **Context**: He was arrested for his crimes.
> **Question:** When was the arrest?
> **Answer:** IDK

Portions	#Examples	IDK Prportion (%)
Has-answer	238	0
Compet. IDK	250	100
Non-Compet. IDK	246	100

Statistics for the ACE-whQA test dataset

Evaluating on Out-of-domain Datasets

- Current systems trained on SQuAD 2.0 achieve good in-domain performance. 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.
- Informative evaluation requires **out-of-domain test sets**
  - Testing on on datasets **different from the ones they have been trained and finetuned**
  - Ask **very simple questions** whose answer is obvious to humans. (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)

Training Methods

- **BERT-based method** for training on SQuAD 2.0 (Devlin et al., 2019):
  - IDK questions are treated as questions having an answer that is a span with start and end at the [CLS] token.
    - The "no-answer" is predicted if the best non-null span is bigger than the probability of the no-answer span by a threshold θ that is selected on the dev set to maximize the F1 score.
- **Leveraging the Recognizing Textual Entailment task** (RTE; Dagan et al., 2013):
  - Finetuning BERT-LARGE on MNLI (Williams et al., 2018), removing the classification layer and then further finetuning on SQuAD 2.0.

Evaluating on ACE-whQA

	Baseline	Using RTE	Using Binary RTE
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)

- Low performance of a top system trained on SQuAD 2.0
- First training on MNLI 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 (c(MNLI); contradiction/non-contradiction).
  - Control for the size of the data
  - Control for the format similarity between TE and the test set

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 non-competitive ones.
- RTE can be useful, particularly for non-competitive IDK questions