BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski
Michael Collins, Kristina Toutanova
Motivation: Inference

- Humans can infer many things from text

“The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016.”

- Pittsburgh has a sports team called the “Penguins”
- The Sharks got second place in 2016
- The Sharks have never won the Stanley Cup
Motivation: Testing Inference is Difficult

- Crowd-sourcing interesting examples can be challenging

“The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016.”

- The Sharks advanced to the Stanley cup in 2016
- The Sharks lost to the Pittsburgh Penguins
Motivation: Testing Inference is Difficult

- Recognizing entailment is an artificial task
- Have to make a number of arbitrary decisions:
  - What things are important to infer?
  - How strictly should we define entailment?
  - What kinds of inferential abilities should be tested?
- Hard to interpret results
This Work: Natural Yes/No Questions

- Yes/No questions generated without any prompting
  - No pre-specified source text or topic
  - No knowledge of the answer
  - Not required to write yes/no questions
- Paired with passages selected by independent annotators

Does Tyrion survive in Game of Thrones?
Did the US qualify for the World Cup?
Natural Yes/No Questions

- Often require inference
- Are challenging for existing models
- Have an obvious end-task
- Real-word test of inference
Example

**Question**: Do all neurons have the same action potential?

**Passage**: In the early development, the action potential of neurons is initially carried by calcium current. The longer opening times for the calcium channels can lead to action potentials that are considerably slower than those of mature neurons.

**Answer**: ?
Example

**Question:** Do all neurons have the same action potential?

**Passage:** In the early development, the action potential of neurons is initially carried by calcium current. The longer opening times for the calcium channels can lead to action potentials that are considerably slower than those of mature neurons.

**Answer:** No
The Rest of this Talk

- Dataset Construction
- Dataset Analysis
- Transfer Learning Baselines
Dataset Construction
Collecting Questions

- Are there blue whales in the Atlantic Ocean?
- Is chess a fun game?
- Has a car ever gone the speed of sound?

Anonymized Queries

Heuristic Filtering

Manual Validation

- Are there blue whales in the Atlantic Ocean? ✅
- Is chess a fun game? ❌
- Has a car ever gone the speed of sound? ✅
Are there blue whales in the Atlantic Ocean?

Yes
The Dataset

- (Question, Paragraph, Answer) triples where the answer is either “yes” or “no”
- 9.4k train questions
- 3.2k dev/test questions
- 62% “Yes” answers
- 110 average paragraph tokens
- 90% human performance
Dataset Analysis
Question Topics

- Entertainment Media: 22.0%
- Nature/Science: 22.0%
- Sports: 11.0%
- Law/Government: 10.0%
- History: 5.0%
- Fictional Events: 4.0%
- Other: 26.0%
Paraphrasing (38.7%)

The passage explicitly asserts or refutes what is stated in the question.

**Question:** Is Tim Brown in the Hall of Fame?

**Passage:** ...Brown has also played for the Tampa Bay Buccaneers. In 2015, he was inducted into the Pro Football Hall of Fame.

**Answer:** Yes
By Example (11.8%)

The passage provides an example or counter-example to what is asserted by the question

**Question:** Has the UK been hit by a hurricane?

**Passage:** The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands...

**Answer:** Yes
Factual Reasoning (8.5%)

Answering the question requires using world-knowledge to connect what is stated in the passage to the question.

**Question:** Was Designated Survivor filmed in the White House?

**Passage:** The series is . . . filmed in Toronto, Ontario.

**Answer:** No
Implicit (8.5%)

The passage mentions or describes entities in the question in a way that would not make sense if the answer was not yes/no.

**Question:** Is static pressure the same as atmospheric pressure?

**Passage:** The aircraft designer’s objective is to ensure the pressure in the aircraft’s static pressure system is as close as possible to the atmospheric pressure...

**Answer:** No
We can conclude the answer is yes or no because, if this was not the case, it would have been mentioned in the passage.

**Question:** Did Mickey Rourke win an Oscar for the Wrestler?

**Passage:** In the 2008 film The Wrestler... Rourke received a 2009 Golden Globe award, a BAFTA award, and an Academy Award nomination...

**Answer:** No
Other Inference (25.9%)

The passage states a fact that can be used to infer whether the answer is true or false, and does not fall into any of the other categories.

---

**Question:** Is the sea snake the most venomous snake?

**Passage:** ...the venom of the inland taipan, drop by drop, is the most toxic among all snakes.

**Answer:** No
Why are Yes/No Question Interesting?

- Rarely factoid
  - Unusual to ask “Was Obama born in 1961?”
- “No” Answers usually have to be inferred
- Easy to use non-trivial kinds of reasoning when labelling them
Experiments
Simple Baselines

- Majority Guess: 62.2%
- Question-Only BERT model: 64.5%
- Passage-Only BERT model: 66.7%
- Word-Overlap Model: 62.2%
Transfer Baselines

- Supervised transfer sources:
  - Question Answering (SQuAD, QNLI, NQ)
  - Entailment (MNLI, SNLI)
  - Paraphrasing (QQP)
  - Heuristic Y/N data (MSMarco)

- Supervised tasks are used to pre-train a standard recurrent +
  co-attention model (see paper for details)

- Recent unsupervised transfer methods (BERT, OpenAI GPT,
  ELMo)
Question Answering

- Majority: 62.2
- RACE: 68.4
- None: 69.6
- SQuAD 2.0: 69.8
- QQP: 71.4
- QNLI: 72.8
- Heuristic Y/N: 71.4
- ELMO: 72.8
- NQ: 72.8
- OpenAI GPT: 72.8
- SNLI: 72.8
- MultiNLI: 72.8
- BERT: 72.8
Test Set Results

- Majority: 62.31
- Recurrent: 67.52
- Recurrent + MNLI: 74.24
- BERT: 76.7
- BERT + MNLI: 80.43
Thank You

Data: goo.gl/boolq
Will become part of the SuperGLUE benchmark (Wang et al., 2019)
  ○ super.gluebenchmark.com