The Winograd Schema Challenge

Hector J. Levesque

AAAI, 2012

Li "Harry" Zhang (zharry@seas.upenn.edu)

04-01-2020

Penn Engineering

Problem & Motivation

- Problems of the Turing Test
  - Deception: machine needs to "assume" an identity
  - Trickery: machine can rely on evasive or canned response
- Alternative: a benchmark dataset to test understanding and reasoning
  - Binary classification (easy to evaluate)
  - Requires reasoning of some sort (can infer machine's ability beyond memorization
- A good benchmark should satisfy the following
  - involve a broad range of subjects (wide domain)
  - native English-speaking adults can pass it easily (easy for human)
  - it can be administered and graded without expert judges (clearly evaluated)
  - when people pass the test, we would say they were thinking (challenging reasoning)

Previous benchmark

- Recognizing Textual Entailment (RTE)
  - A: Time Warner is the world's largest media and internet company.
  - B: Time Warner is the world's largest company.
  - Entail? Contradict? Neutral?
- Problem of subjectivity
  - Not easily answerable by non-experts
  - E.g. confusable with "inference" in linguistics

Contributions of this work

- A new benchmark: The Winograd Schema Challenge
  - Pronoun disambiguation task
  - Binary classification
  - 100 expert-crafted examples
- Easily disambiguated by the human reader (ideally, so easily that the reader does not even notice that there is an ambiguity);
- **Not solvable by simple techniques such as selectional restrictions;**
- **No obvious statistical test over text corpora that will reliably disambiguate t**

The trophy would not fit in the brown suitcase because it was too big. What was too big?

Answer 0: the trophy
Answer 1: the suitcase

Criteria of an example

- Two parties are mentioned in a sentence by noun phrases
- A pronoun or possessive adjective is used in the sentence in reference to one of the parties, but is also of the right sort for the second party
- The question involves determining the referent of the pronoun or possessive adjective

The trophy would not fit in the brown suitcase because it was too big. What was too big?

Answer 0: the trophy
Answer 1: the suitcase

Joan made sure to thank Susan for all the help she had given. Who had given the help?

Answer 0: Joan
Answer 1: Susan

Criteria of an example

- There is a word (special word) that appears. When replaced by another word (alternate word), the answer changes.
- Guaranteed input perturbation

The trophy would not fit into the brown suitcase because it was too ⟨ ⟩. What was too ⟨ ⟩?

Answer 0:	the trophy
Answer 1:	the suitcase
special:	big
alternate:	small

Ensuring difficulty for machine

- Avoid examples that can be solved just using word-level information
- Selection restriction
  - Only the women can be pregnant
  - Only the pills can be carcinogenic
- Selection preference
  - Racecar is often associated with speed
  - School bus is often not associated with speed

The women stopped taking the pills because they were ⟨ ⟩. Which individuals were ⟨ ⟩?

Answer 0:	the women
Answer 1:	the pills
special:	pregnant
alternate:	carcinogenic

The racecar zoomed by the school bus because it was going so ⟨ ⟩. What was going so ⟨ ⟩?

Answer 0:	the racecar
Answer 1:	the school bus
special:	fast
alternate:	slow

Ensuring easiness for human

- Avoid examples that are ambiguous to human
  - Either Frank or Bill can be pleased

Frank was pleased when Bill said that he was the winner of the competition.

Answer 0: Frank
Answer 1: Bill

- Levels of difficulty based on vocabulary, corresponding to human's ability of understanding language and common sense

The large ball crashed right through the table because it was made of ⟨ ⟩. What was made of ⟨ ⟩?

Answer 0: the ball
Answer 1: the table

---

special: steel
alternate: styrofoam

Discussion and analysis

- Easy for human?
- Tests reasoning?
- Robust to statistical cues?
  - Author: "It is wildly implausible that there would be statistical or other properties of the special word or its alternate that would allow us to flip from one answer to the other in this case."
  - 13.5% examples have "word-association" (Trichelair et al., 2018)
- × Challenging for models?
  - 90.1% state-of-the-art performance by RoBERTa (Sakaguchi et al., 2019)

WSC-like tasks

- WSC (Levesque, Davis, and Morgenstern 2011)
  - The original
- PDP (Morgenstern, Davis, and Ortiz 2016)
  - 80 pronoun disambiguation problems, multiple choice
- GLUE/SuperGLUE-WSC (Wang et al. 2019)
  - aggregates the original WSC, PDP and additional examples, binary classification
- DPR (Rahman and Ng 2012), KnowRef (Emami et al. 2019), COPA (Roemmele, Bejan, and Gordon 2011), Winogender (Rudinger et al. 2018)

My takeaways

- The paper is highly avant-garde especially in 2012
  - Early idea of preventing learning from statistical cues even before word2vec
  - Criteria for building datasets still honored today
- How much do pre-trained models learn to reason based on WSC?
- How much do they rely on statistical cues?
- What's the significance of train/test split?
  - GLUE-WSC's test set does not include perturbation
  - Different examples may require different method of reasoning