Using Winograd Schemas for Evaluation of Implicit Information Extraction Systems

Ivan Rygaev, Dialogue 2017
Laboratory of Computational Linguistics
Institute for Information Transmission Problems RAS, Moscow, Russia

based on
Hector Levesqe et al. 2012. Winograd Schema Challenge
and related works

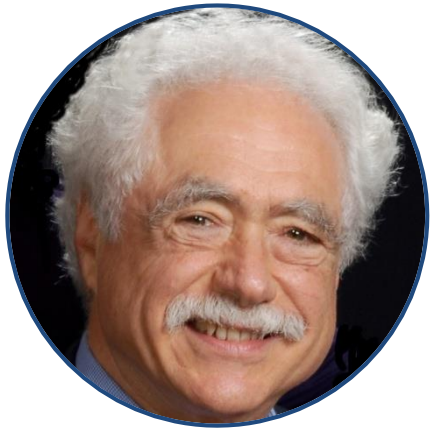
Winograd Schema Challenge

- A test for computer intelligence
- More convincing than the Turing Test that machines can think
- Based on analysis of the short text of 1-3 sentences and a question on them
- Special type of anaphora resolution problem
- Linguistic features, collocation statistics, selectional restrictions does not help
- Some kind of world knowledge is required

Key People

Hector Levesque

Ernest Davis



Terry Winograd

Leora Morgenstern

Turing Test Criticism

- Turing Test was formally passed by a chat-bot Eugene Goostman in 2014
- But does the chat-bot think?
- Is *conversation* the right way of evaluation?
  - Subjective
  - Encourage verbal acrobatics and trickery
- Turing Test requires *deception*
  - Must fool an interrogator that it is a person
  - Do we need this from an intelligent machine? For which purposes?

Winograd Schemas

- Proposed by Hector Levesque in 2011
- The trophy doesn't fit in the brown suitcase because **it**'s too *big*. What is too *big*?
  - the trophy
  - the suitcase
- Joan made sure to thank Susan for all the help **she** had *given*. Who had *given* the help?
  - Joan
  - Susan
- Terry Winograd provided the first example in 1970

Winograd Schema Structure

- Anaphora resolution problem
- There are two potential antecedents in the sentence
- Linguistic features, collocation statistics and selectional restrictions does not help much
- Changing a special word in the sentence reverts the correct answer (*big -> small*)
- The trophy doesn't fit in the brown suitcase because **it**'s too *small*. What is too *small*?
  - the trophy
  - the suitcase

Commonsense Knowledge

- People are good on Windograd Schemas
- Tests show 91-92% correct answers.
- What is required to get the right answer?
- Understanding of the verb ‘fit’
  - if A fits into B then A must be smaller than B.
- Understanding of the connective ‘because’
  - Changing it to ‘in spite of’ also reverts the answer.
- Implicit information must be extracted from the text to pass the test

WSs Preparation

- The wrong answer need not be logically inconsistent:
- Tom threw his bag down to Ray after **he** reached the *top* of the stairs. Who reached the *top* of the stairs?
  - Tom
  - Ray
- Alternate special word need not be the opposite:
- The man couldn't lift his son because **he** was so *weak/heavy*. Who was *weak/heavy*?
  - the man
  - the son

WSs Preparation

- WS must not be ‘too obvious’:
- The women stopped taking the pills because **they** were *pregnant/cancerogenic*.
  Which individuals were *pregnant/cancerogenic*?
  - the women
  - the pills
- Selectional restrictions help:
  - Only women can be pregnant, not pills
  - Only pills can be cancerogenic, not women
- The first sentence can be totally ignored

WSs Preparation

- WS must not be ambiguous for humans (both ways)
- Frank was *jealous* when Bill said that **he** was the winner of the competition. Who was the winner?
  - Frank
  - Bill
- Frank was *pleased* when Bill said that **he** was the winner of the competition. Who was the winner?
  - Frank
  - Bill
- It is not unreasonable that Bill's victory pleased Frank

Flexibility

- WSs of different difficulty allow incremental progress
- The councilmen refused to give the demonstrators a permit because **they** *feared/advocated* violence. Who *feared/advocated* violence?
  - the councilmen
  - the demonstrators
- WSs for different domains:
  - spatial vs. social relations
- WSs for specific features:
  - paraphrasing, sentiment analysis…

Approaches

- The test is agnostic to internal realization techniques:
  - Rule-based or
  - Statistical machine learning
- Both are welcome
- A deep learning solution even showed better results in the first competition in 2016
  - But it was taught on semantic resources rather than just texts

Competition

- The first competition was held in July 2016 at IJCAI conference in New York
- It was organized in two rounds:
  1. Sentences from real texts (children's' literature) rather than constructed ones. They exhibited all the properties of WS but did not have an alternative variant.
  2. Actual constructed WSs with an alternative variant
- Motivation for two rounds:
  - Not to reveal WSs to contestants who are not ready yet
  - Increase relevance of the test by using real examples

Competition

- There were 60 questions in the first round and 60 in the second one.
  - To proceed to the second round a contestant had to score at least 90% correct in the first one.
- None of the solutions achieved that score
  - The second round was not held
- The big prize was offered to the team who would achieve at least 90% in both rounds
  - Three smaller prizes were offered to the top programs achieved at least 65% in the first round

Competition Results

- Six solutions of four teams where presented:

Contestant	Number correct	Percentage correct
Patrick Dhondt	27	45%
Denis Robert	19	31.666%
Nicos Issak	29	48.33%
Quan Liu (1)	28	46.9% (48.33)*
Quan Liu (2)	29	48.33% (58.33)*
Quan Liu (3)	27	45% (58.33)*

- Random answering could yield 45%

Results Assessment

- None of the solutions got over the 65% threshold to receive even the smaller prize
- Four of the six programs showed scores around the chance level or even worse
- The best solution used deep learning algorithms. It was taught on ConceptNet, WordNet and CauseCom resources
  - CauseCom is a set of cause-effect pairs automatically collected from large text corpora
- The next test is planned for AAAI-2018 (Feb)

Conclusions

- Winograd Schema Challenge is a good test for text understanding and implicit knowledge extraction
- It allows incremental progress and can be either broad or specific to a certain domain or extracting feature
- The proposal is to organize Winograd Schema Challenge in Russian at one of the subsequent Dialogue conferences.

Thank you!