On Commonsense Domains within the Winograd Schema Challenge

Research Project

Aneta Koleva

Supervisors: Prof. Sebastian Rudolph Dr. Emmanuelle Dietz

28-03-2019

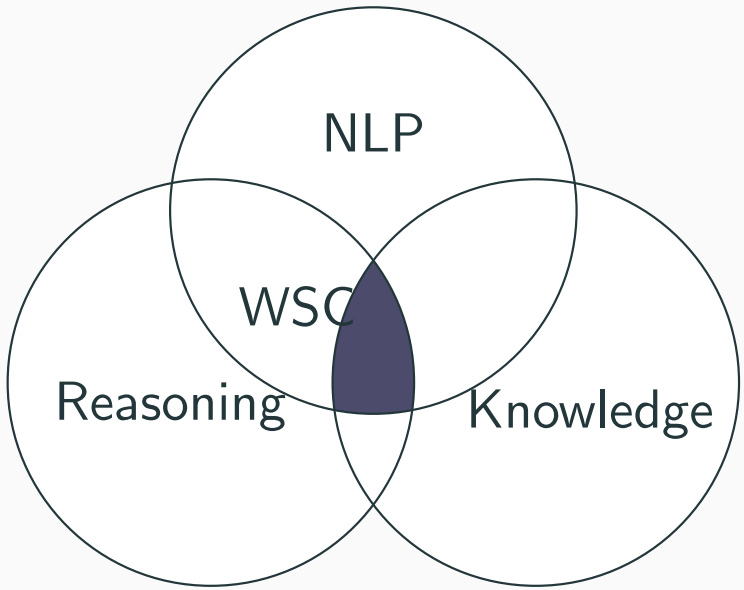
Dresden University of Technology

Motivation

- Winograd Schema Challenge (Levesque et. al, 2012)
  - S: The trophy does not fit into the brown suitcase because it is too [small/large].
  - Q: What is too [small/large]?
  - A: The suitcase/the trophy.

Motivation

- Winograd Schema Challenge (Levesque et. al, 2012)
  - S: The trophy does not fit into the brown suitcase because it is too [small/large].
  - Q: What is too [small/large]?
  - A: The suitcase/the trophy.



Outline

Description

Winograd Schema Challenge

S: The trophy does not fit into the brown suitcase because it is too [small/large].

Q: What is too [small/large]?

A: The suitcase/the trophy

Winograd Schema Challenge

S: The trophy does not fit into the brown suitcase because it is too [small/large].

Q: What is too [small/large]?

A: The suitcase/the trophy

- Winograd Schema:
  - Sentence containing two nouns, one ambiguous pronoun and a special word
  - Question asking about the referent of the pronoun
  - Two possible answers corresponding to the noun phrases in the sentence

Winograd Schema Challenge

S: The trophy does not fit into the brown suitcase because it is too [small/large].

Q: What is too [small/large]?

A: The suitcase/the trophy

- Winograd Schema:
  - Sentence containing two nouns, one ambiguous pronoun and a special word
  - Question asking about the referent of the pronoun
  - Two possible answers corresponding to the noun phrases in the sentence
- Characteristics:
  - Easy to answer for an adult English speaker
  - Always contains special word
  - Google proof

Competition

- Competition in 2016 at IJCAI-16
  - Two time-constraint rounds - 210 min. each
    - Pronoun Disambiguation Problems (PDPs) - 60
    - Parts of Winograd Schemas - 150
  - Four competitors
  - Best result: 58% correctly resolved PDPs
  - There was no second round
- Current state-of-the-art (Radford et. al, 2019) achieves 70.7% accuracy
  on the WSs dataset

Previous Approaches

Previous Approaches

- Machine learning and deep learning
  - Supervised ranking SVM
  - Supervised classification Task
  - Knowledge enhanced embeddings
  - Google's language models
  - Open AI language model
- Knowledge-based
  - Knowledge graphs with Relevance Theory
  - Semantic parsing and knowledge hunting
  - Parsing query results and assigning scores
  - Knowledge types identification

A Simple Method for Commonsense Reasoning (Trinh and Le, 2018)

- Language models trained on unlabeled data

A Simple Method for Commonsense Reasoning (Trinh and Le, 2018)

- Language models trained on unlabeled data
  - Recurrent Neural Networks
  - Trained on large datasets and on a dataset customized for WSC

A Simple Method for Commonsense Reasoning (Trinh and Le, 2018)

- Language models trained on unlabeled data
  - Recurrent Neural Networks
  - Trained on large datasets and on a dataset customized for WSC
- Substitution ambiguous pronoun
  - The trophy doesn't fit in the suitcase because the trophy is too big
  - The trophy doesn't fit in the suitcase because the suitcase is too big

A Simple Method for Commonsense Reasoning (Trinh and Le, 2018)

- Language models trained on unlabeled data
  - Recurrent Neural Networks
  - Trained on large datasets and on a dataset customized for WSC
- Substitution ambiguous pronoun
  - The trophy doesn't fit in the suitcase because the trophy is too big
  - The trophy doesn't fit in the suitcase because the suitcase is too big
- Language models assign probabilities to both sentences

A Simple Method for Commonsense Reasoning (Trinh and Le, 2018)

- Language models trained on unlabeled data
  - Recurrent Neural Networks
  - Trained on large datasets and on a dataset customized for WSC
- Substitution ambiguous pronoun
  - The trophy doesn't fit in the suitcase because the trophy is too big
  - The trophy doesn't fit in the suitcase because the suitcase is too big
- Language models assign probabilities to both sentences
- Evaluation and results
  - PDPs 70% accuracy
  - WSC 63.7% accuracy

Knowledge Types Identification and Reasoning (Anonymous Authors,2019)

- Identified 12 knowledge types which cover the entire WSC dataset
- Developed a logical reasoning algorithm
- Evaluated on 100 problems from WSC and achieved 100% accuracy

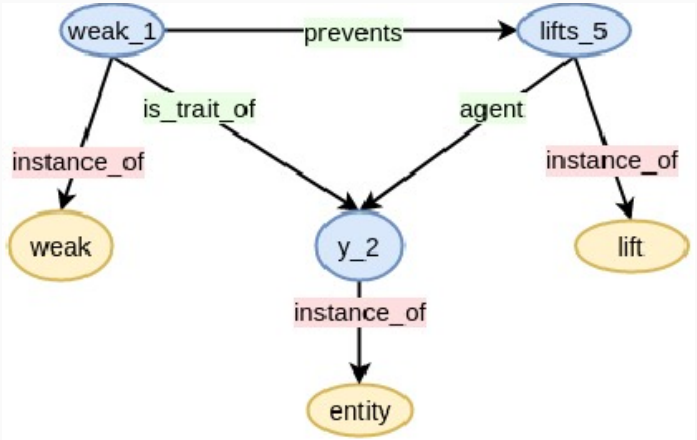
[1] kparser.org

Knowledge Types Identification and Reasoning (Anonymous Authors,2019)

- Identified 12 knowledge types which cover the entire WSC dataset
- Developed a logical reasoning algorithm
- Evaluated on 100 problems from WSC and achieved 100% accuracy
- Solver
  1. Semantic graph[1] of the input sentence and question
  2. Semantic graph representation of background knowledge
  3. Graph merging
  4. Project question graph on the merged graph
  5. Answer - the node from the merged graph which is from the same domain as the unknown node from the question graph

[1] kparser.org

- Representation of the knowledge "*weak y* **prevents** *y lifts*"



Methodology

Categorization of Winograd Schemas

- Motivation
  - Current state-of-the-art has a poor performance
  - Background knowledge is crucial for predicting the correct answer

Categorization of Winograd Schemas

- Motivation
  - Current state-of-the-art has a poor performance
  - Background knowledge is crucial for predicting the correct answer
  - Idea
    1. Analyze the input Winograd Schema and identify the domain
    2. Search for knowledge specific to this domain
    3. Apply reasoning procedure

Identified Categories

Category	Example
1. Physical	**S:** John couldn't see the stage with Billy in front of him because he is so **[short/tall]**. **Q:** Who is so [short/tall]?
2. Emotions	**S:** Frank felt **[vindicated/crushed]** when his longtime rival Bill revealed that he was the winner of the competition. **Q:** Who was the winner of the competition?
3. Interactions	**S:** Joan made sure to thank Susan for all the help she had **[given/received]**. **Q:** Who had [given/received] help?
4. Comparison	**S:** Joe's uncle can still beat him at tennis, even though he is 30 years **[older/younger]**. **Q:** Who is [older/younger]?
5. Causal	**S:** Pete envies Martin **[because/although]** he is very successful. **Q:** Who is very successful?
6. Multiple knowledge	**S:** Sam and Amy are passionately in love, but Amy's parents are unhappy about it, because they are **[snobs/fifteen]**. **Q:** Who are [snobs/fifteen]?

Annotation of Winograd Schemas

- Strong agreement between the annotators
  Cohen's kappa score 0.66
- Annotation Results

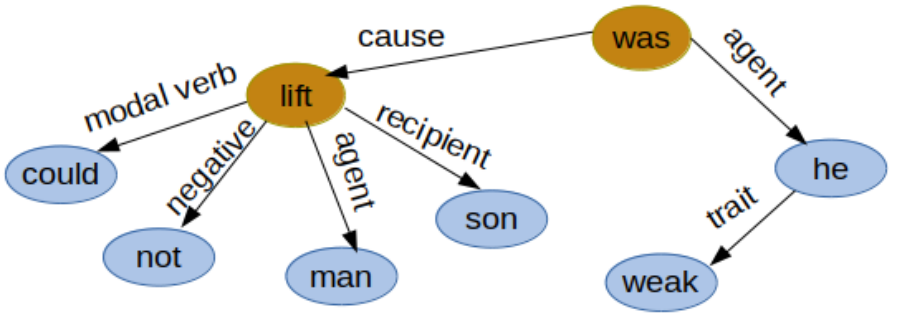
Category	**Annotator 1**	**Annotator 2**
Physical	36	39
Emotions	7	9
Interactions	44	24
Comparison	19	26
Causal	16	18
Multiple knowledge	28	34

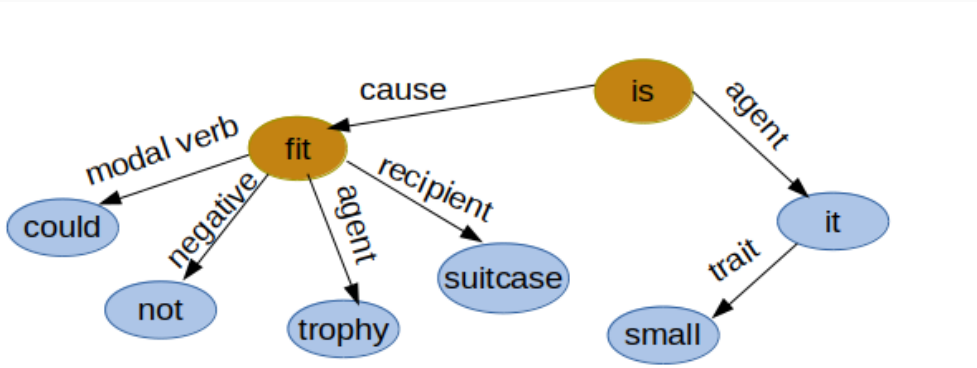
Graph Representation for Physical Category

1. The man couldn't lift his son because he was so weak.
2. The trophy doesn't fit into the brown suitcase because it's too small.

Graph Representation for Physical Category

1. The man couldn't lift his son because he was so weak.
2. The trophy doesn't fit into the brown suitcase because it's too small.





Reasoning

- Knowledge required for both examples is about physical features
- Similar reasoning rules for categorizing the traits
  1. weak(X) :- lift(X,Y), not lift(modifier, could).
  2. small(Y) :- fit(X,Y), not fit(modifier, could).

Reasoning

- Knowledge required for both examples is about physical features
- Similar reasoning rules for categorizing the traits
  1. weak(X) :- lift(X,Y), not lift(modifier, could).
  2. small(Y) :- fit(X,Y), not fit(modifier, could).
- Reasoning Algorithm
- Change of background knowledge
  - has_k(weak,prevents,lift).

Conclusion

Contributions

- Overview of different approaches towards WSC
- None achieves close to 90% accuracy
- We analyzed the entire WSC corpus and identified 6 categories
- We identified a mistake in the Reasoning Algorithm and proposed a correction

Future Work

- Better Reasoning Algorithm
- Knowledge Graphs (RDF) representation
- Knowledge-injection neural networks

Thank you!