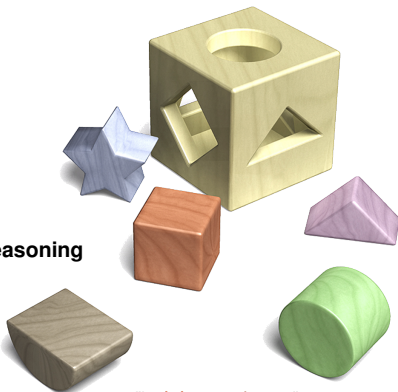


On Commonsense Domains within the Winograd Schema Challenge

Aneta Koleva

International Center for Computational Logic
Technische Universität Dresden
Germany

- ▶ Winograd Schema Challenge
- ▶ Previous Approaches
- ▶ Knowledge Types Identification and Reasoning
- ▶ Categorization of Winograd Schemas
- ▶ Conclusion



"Logic is everywhere ..."



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.



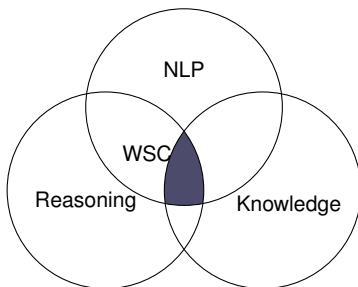
Motivation

► 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.



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

- ▶ Machine learning and deep learning techniques
- ▶ Knowledge-based system with reasoning procedures



Previous Approaches

- ▶ **Machine learning and deep learning techniques**
- ▶ **Knowledge-based system with reasoning procedures**

Name	Evaluation dataset	Results
Machine Learning Framework	Additional 943 WSs*	73%
Siamese Neural Network	* and WSC corpus	63% and 56%
Knowledge Enhanced Embeddings	PDP and WSC corpus	66.7% and 58.3%
Google's language models	PDP and WSC corpus	70% and 63.7%
OpenAI language model	WSC corpus	70.70%
Knowledge Graphs	4 WSs	100 %
Sharma's categories	71 WSs	59 %
Semantic relations categories	100 WSs	100 %
Knowledge hunting framework	WSC corpus	43.5 %



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 (Sharma and Baral, 2018)

- ▶ 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

¹kparser.org



Knowledge Types Identification and Reasoning (Sharma and Baral, 2018)

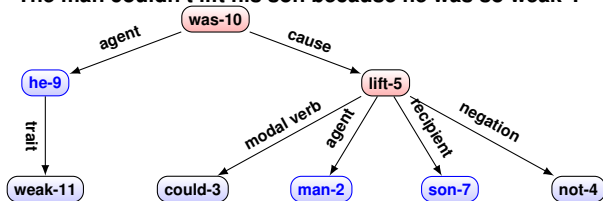
- ▶ Identified 12 **knowledge types** which cover the entire WSC dataset
- ▶ Developed a **logical reasoning algorithm**
- ▶ Evaluated on 100 problems from WSC and achieved **100%** accuracy
- ▶ Solver
 1. Semantic graph¹ of the input sentence and question
 2. Semantic graph representation of background 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

¹kparser.org



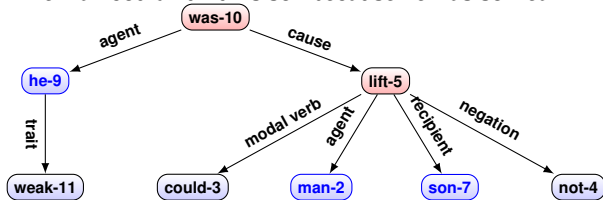
Semantic graph representation

- “The man couldn’t lift his son because he was so weak”.

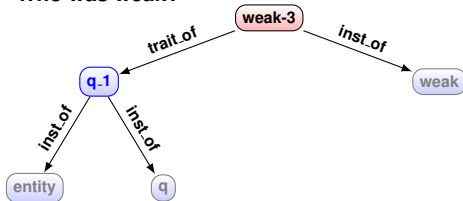


Semantic graph representation

- “The man couldn’t lift his son because he was so weak”.

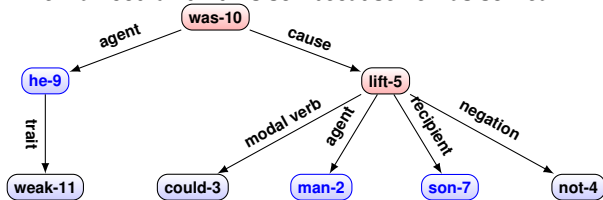


- “Who was weak?”

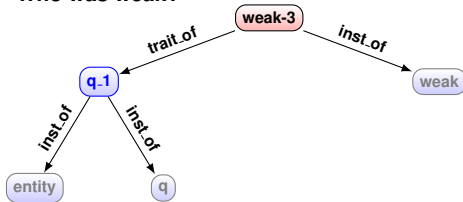


Semantic graph representation

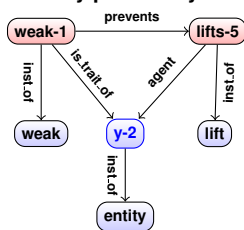
- “The man couldn’t lift his son because he was so weak”.



- “Who was weak?”



- “weak y prevents y lifts”



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. Emotional	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

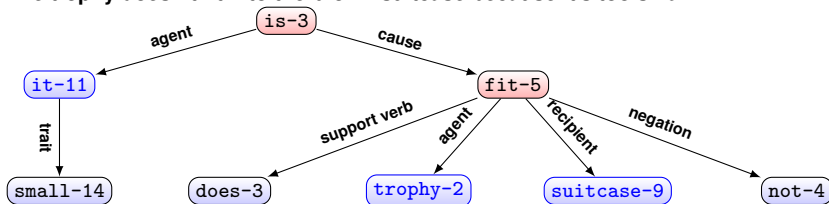
- ▶ **Strong agreement between the annotators**
Cohen's kappa score 0.66
- ▶ **Annotation Results**

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



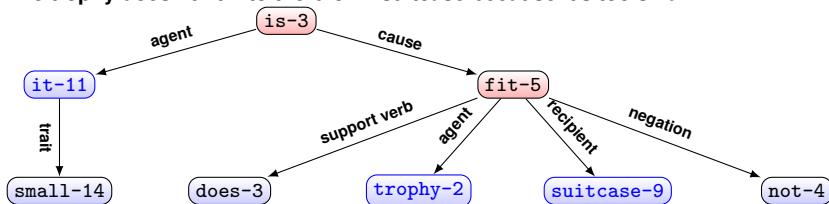
Graph Representation for Physical Category

1. The trophy doesn't fit into the brown suitcase because it's too small.

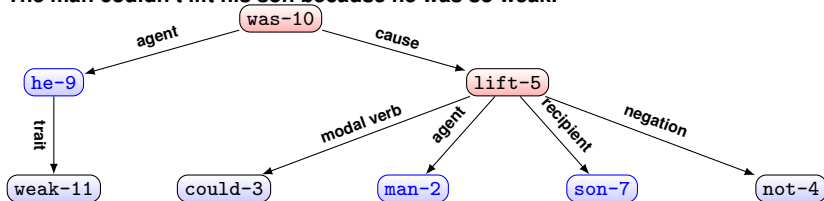


Graph Representation for Physical Category

1. The trophy doesn't fit into the brown suitcase because it's too small.



2. The man couldn't lift his son because he was so weak.



Reasoning

- ▶ Knowledge required for both examples is about **physical features**
- ▶ Similar reasoning rules for categorizing the traits
 1. `has_k(small,is_trait_of,y) :- has_k(fits,recipient,y),
not has_k(fits,modifier,could).`
 2. `has_k(weak, is_trait_of,y) :- has_k(lift,agent,y),
not has_k(lift,modifier,could).`



Reasoning

- ▶ Knowledge required for both examples is about **physical features**
- ▶ Similar reasoning rules for categorizing the traits
 1. `has_k(small,is_trait_of,y) :- has_k(fits,recipient,y),
not has_k(fits,modifier,could).`
 2. `has_k(weak, is_trait_of,y) :- has_k(lift,agent,y),
not has_k(lift,modifier,could).`
- ▶ Reasoning Algorithm
- ▶ Change of background knowledge
 - ▷ `has_k(weak,prevents,lift).`



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

- ▶ **Formalization of the characteristics for each category**
- ▶ **Knowledge-enhanced neural networks**



Thank you!

