TECHNISCHE UNIVERSITÄT DRESDEN

# On Commonsense Domains within the Winograd Schema Challenge

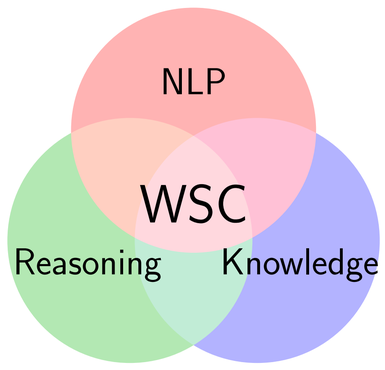
Aneta Koleva

Supervisors: Emmanuelle-Anna Dietz, Sebastian Rudolph

## Commonsense Reasoning in Computers

Hector Levesque (2011) proposes a new test for assessing computer intelligence that requires use of commonsense reasoning when predicting the correct answer.

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 (WSC)

- Structure of a 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* - statistical methods over large text corpora should not be able to resolve a WS.

## Machine-Learning vs Knowledge-Based Approaches

| Technique | PDPs Size Correct | WSC Size Correct | WSC* Size Correct | Remarks |
|---|---|---|---|---|
| Supervised ranking SVM model [6] | NA | NA | 282 - 30%<br>205 - 73% | -provided additional dataset set<br>-no evaluation on WSC dataset |
| Classification task with NN [3] | NA | 282 - 100%<br>157 - 56% | 282 - 30%<br>177 - 63% | -first to use substitution of the pronoun with the antecedents |
| Knowledge Enhanced Embeddings (KEE) [4] | 60-100%<br>40 - 66.7% | NA | NA | -best results in the 2016 WSC competition |
| Google's language models [9] | 60-100%<br>42 - 70% | 273 - 100%<br>173 - 63.7% | NA | -no reasoning involved in the discovery of the correct answer<br>-state-of-the-art for PDPs |
| OpenAI language models [5] | NA | 273 - 100%<br>193 - 70.70% | NA | -current state-of-the-art for WSC<br>-requires a lot of data for training<br>-results are not reproducible |
| Graphs with Relevance theory [7] | NA | 4 - 2.6%<br>4 - 100% | NA | -manual construction of graphs<br>-first representation of WS as dependency graph |
| 2 identified categories [8] | NA | 71 -25%<br>49 - 69% | NA | -first attempt of identifying commonsense knowledge types<br>-developed the KParser |
| Semantic relations categories [1] | NA | 100 - 34%<br>100 - 100% | 138 - 14%<br>111 - 80% | -provided Reasoning Algorithm<br>-identified 12 commonsense types which capture the entire WSC |
| Knowledge hunting framework [2] | NA | 273 - 100%<br>119 - 43.5% | NA | -refined query generation<br>-developed an algorithm for scoring the retrieved sentences |

## References

[1] C. Baral A. Sharma. Commonsense knowledge types identification and reasoning for the winograd schema challenge, 2018.

[2] A. Emami, N. De La Cruz, A. Trischler, K. Suleman, and J. Chi Kit Cheung. A knowledge hunting framework for common sense reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1949–1958, 2018.

[3] Opitz J. and Frank A. Addressing the winograd schema challenge as a sequence ranking task. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 41–52. Association for Computational Linguistics, 2018.

[4] Q. Liu, H. Jiang, Z. Ling, X. Zhu, S. Wei, and Y. Hu. Combing context and commonsense knowledge through neural networks for solving winograd schema problems. 2016.

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019.

[6] A. Rahman and V. Ng. Resolving complex cases of definite pronouns: The winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 777–789, 2012.

[7] P. Schüller. Tackling winograd schemas by formalizing relevance theory in knowledge graphs. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014, Vienna, Austria, July 20-24, 2014*, 2014.

[8] A. Sharma, Nguyen Ha Vo, Somak Aditya, and Chitta Baral. Towards addressing the winograd schema challenge - building and using a semantic parser and a knowledge hunting module. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 1319–1325, 2015.

[9] Q. V. Le T. H. Trinh. A simple method for commonsense reasoning. 2018.

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

- Identified 12 knowledge types which cover the entire WSC dataset.
- The categorization is based on the structure of the Winograd sentence.
- 10 of the knowledge types are based on different interactions between entities, actions and properties
- Developed a logical reasoning algorithm.
- Evaluated on 100 problems from WSC and achieved 100% accuracy

Extracted knowledge: "weak y prevents y lifts".
Knowledge type "Property prevents Action"
ASP encoding:

```
has_k(weak_1,is_trait_of,y_2).                1
has_k(weak_1,instance_of,weak).               2
has_k(y_2,instance_of,entity).                3
has_k(weak_1,prevents,lifts_5).               4
has_k(lifts_5,instance_of,lift).              5
has_k(lifts_5,agent,y_2).                     6
```

- The reasoning algorithm does not consider rule 4 in the reasoning process.

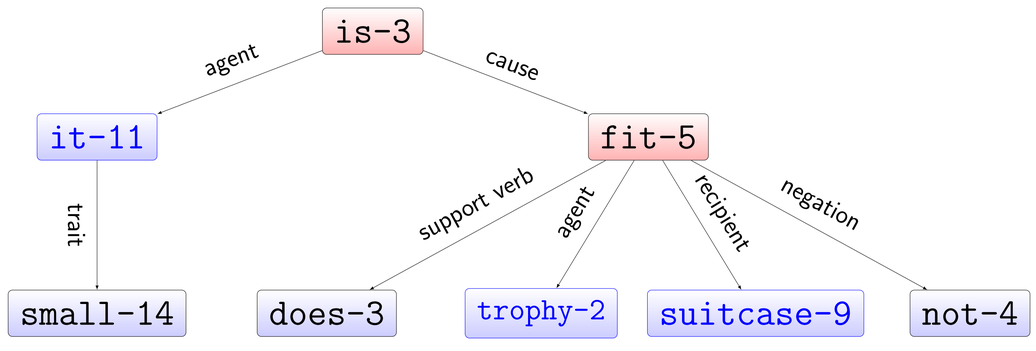
## Categorization of Winograd Schemas

- Inductively analyzed the WSC dataset and identified 6 categories.
- The categorization is based on the content of the Winograd sentence.
- Two annotators annotated the entire WSC corpus with these categories.
- Calculated Cohen's kappa - measure for inter-rater agreement $\kappa = 0.66$

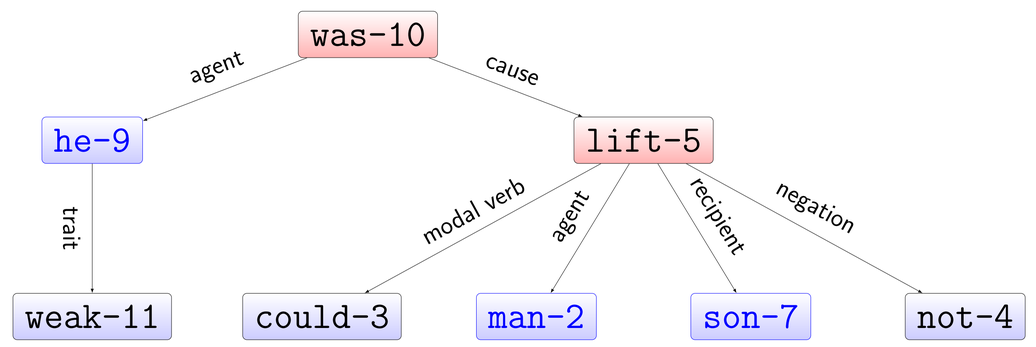
| Category | Example |
|---|---|
| 1. Physical | **S:** John couldn't see the stage with Billy in front of him because he is so [short/tall].<br>**Q:** Who is so [short/tall]? |
| 2. Emotional | **S:** Frank felt [vindicated/crushed] when his longtime rival Bill revealed that he was the winner of the competition.<br>**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].<br>**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].<br>**Q:** Who is [older/younger]? |
| 5. Causal | **S:** Pete envies Martin [because/although] he is very successful.<br>**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].<br>**Q:** Who are [snobs/fifteen]? |

## Winograd Schemas from the Physical Category

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




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




## Reasoning algorithm

Change of the formalization of the background knowledge such that it contributes to the reasoning procedure.

```
1 has_k(small_1,is_trait_of,y_2) :- has_k(fits_5, recipient, y_2),
2                                   not has_k(fits_5,modifier,could_3).
3 has_k(y_2,instance_of,entity).
4 has_k(fits_5,recipient,y_2).
```

Listing 1: Additional knowledge