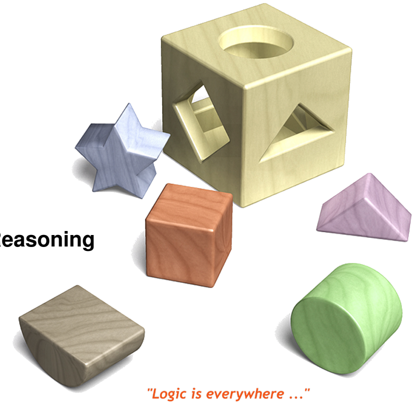
On Commonsense Domains within the Winograd Schema Challenge

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



Motivation

- **Winograd Schema Challenge (Levesque et al., 2012)**

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

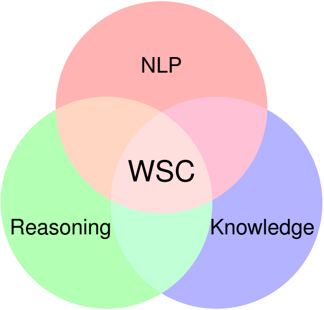
  Q: **What is too [small/large]?**

  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	trophy, suitcase
One ambiguous pronoun	it
A special word	small/ large
Question about the referent of the pronoun	What is too [small/large]
Two possible answers	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	trophy, suitcase
One ambiguous pronoun	it
A special word	small/ large
Question about the referent of the pronoun	What is too [small/large]
Two possible answers	the suitcase /the trophy

- **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

Technique	PDPs Size - Correct	WSC Size - Correct	WSC* Size - Correct	Remarks
Supervised ranking SVM model (2012)	-	-	282-30% - 205-73%	-provided additional dataset set -no evaluation on WSC dataset

Previous Approaches

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

Technique	**PDPs** Size - Correct	**WSC** Size - Correct	**WSC*** Size - Correct	**Remarks**
Supervised ranking SVM model (2012)	-	-	282-30% - 205-73%	-provided additional dataset set -no evaluation on WSC dataset
Knowledge Embeddings (2016)	60-100% - 40-66.7%	-	-	-best results in the 2016 WSC competition

Previous Approaches

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

Technique	PDPs Size - Correct	WSC Size - Correct	WSC* Size - Correct	Remarks
Supervised ranking SVM model (2012)	-	-	282-30% - 205-73%	-provided additional dataset set -no evaluation on WSC dataset
Knowledge Embeddings (2016)	60-100% - 40-66.7%	-	-	-best results in the 2016 WSC competition
Classification task with NN (2018)	-	282-100% - 157-56%	282-30% - 177-63%	-first to use substitution of the pronoun with the antecedents

Previous Approaches

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

Technique	PDPs Size - Correct	WSC Size - Correct	WSC* Size - Correct	Remarks
Supervised ranking SVM model (2012)	-	-	282-30% - 205-73%	-provided additional dataset set -no evaluation on WSC dataset
Knowledge Embeddings (2016)	60-100% - 40-66.7%	-	-	-best results in the 2016 WSC competition
Classification task with NN (2018)	-	282-100% - 157-56%	282-30% - 177-63%	-first to use substitution of the pronoun with the antecedents
Google's language models (2018)	60-100% - 42-70%	273-100% - 173-63.7%	-	-no reasoning involved in the discovery of the correct answer -state-of-the-art for PDPs

Previous Approaches

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

Technique	PDPs Size - Correct	WSC Size - Correct	WSC* Size - Correct	Remarks
Supervised ranking SVM model (2012)	-	-	282-30% - 205-73%	-provided additional dataset set -no evaluation on WSC dataset
Knowledge Embeddings (2016)	60-100% - 40-66.7%	-	-	-best results in the 2016 WSC competition
Classification task with NN (2018)	-	282-100% - 157-56%	282-30% - 177-63%	-first to use substitution of the pronoun with the antecedents
Google's language models (2018)	60-100% - 42-70%	273-100% - 173-63.7%	-	-no reasoning involved in the discovery of the correct answer -state-of-the-art for PDPs
OpenAI language models (2019)	-	273-100% - 193-70.70%	-	-current state-of-the-art for WSC -requires a lot of data for training -results are not reproducible

Previous Approaches

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

Technique	PDPs Size - Correct	WSC Size - Correct	WSC* Size - Correct	Remarks
Supervised ranking SVM model (2012)	-	-	282-30% - 205-73%	-provided additional dataset set -no evaluation on WSC dataset
Knowledge Embeddings (2016)	60-100% - 40-66.7%	-	-	-best results in the 2016 WSC competition
Classification task with NN (2018)	-	282-100% - 157-56%	282-30% - 177-63%	-first to use substitution of the pronoun with the antecedents
Google's language models (2018)	60-100% - 42-70%	273-100% - 173-63.7%	-	-no reasoning involved in the discovery of the correct answer -state-of-the-art for PDPs
OpenAI language models (2019)	-	273-100% - 193-70.70%	-	-current state-of-the-art for WSC -requires a lot of data for training -results are not reproducible
Graphs with Relevance theory (2014)	-	4-2.6% - 4-100%	-	-manual construction of graphs -first representation of WS as dependency graph

Previous Approaches

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

Technique	PDPs Size - Correct	WSC Size - Correct	WSC* Size - Correct	Remarks
Supervised ranking SVM model (2012)	-	-	282-30% - 205-73%	-provided additional dataset set -no evaluation on WSC dataset
Knowledge Embeddings (2016)	60-100% - 40-66.7%	-	-	-best results in the 2016 WSC competition
Classification task with NN (2018)	-	282-100% - 157-56%	282-30% - 177-63%	-first to use substitution of the pronoun with the antecedents
Google's language models (2018)	60-100% - 42-70%	273-100% - 173-63.7%	-	-no reasoning involved in the discovery of the correct answer -state-of-the-art for PDPs
OpenAI language models (2019)	-	273-100% - 193-70.70%	-	-current state-of-the-art for WSC -requires a lot of data for training -results are not reproducible
Graphs with Relevance theory (2014)	-	4-2.6% - 4-100%	-	-manual construction of graphs -first representation of WS as dependency graph
2 identified categories (2015)	-	71-25% - 49-69%	-	-first attempt of identifying commonsense knowledge types -developed the KParser

Previous Approaches

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

Technique	PDPs Size - Correct	WSC Size - Correct	WSC* Size - Correct	Remarks
Supervised ranking SVM model (2012)	-	-	282-30% - 205-73%	-provided additional dataset set -no evaluation on WSC dataset
Knowledge Embeddings (2016)	60-100% - 40-66.7%	-	-	-best results in the 2016 WSC competition
Classification task with NN (2018)	-	282-100% - 157-56%	282-30% - 177-63%	-first to use substitution of the pronoun with the antecedents
Google's language models (2018)	60-100% - 42-70%	273-100% - 173-63.7%	-	-no reasoning involved in the discovery of the correct answer -state-of-the-art for PDPs
OpenAI language models (2019)	-	273-100% - 193-70.70%	-	-current state-of-the-art for WSC -requires a lot of data for training -results are not reproducible
Graphs with Relevance theory (2014)	-	4-2.6% - 4-100%	-	-manual construction of graphs -first representation of WS as dependency graph
2 identified categories (2015)	-	71-25% - 49-69%	-	-first attempt of identifying commonsense knowledge types -developed the KParser
Knowledge hunting framework (2018)	-	273-100% - 119-43.5%	-	-refined query generation -developed an algorithm for scoring the retrieved sentences

Previous Approaches

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

Technique	PDPs Size - Correct	WSC Size - Correct	WSC* Size - Correct	Remarks
Supervised ranking SVM model (2012)	-	-	282-30% - 205-73%	-provided additional dataset set -no evaluation on WSC dataset
Knowledge Embeddings (2016)	60-100% - 40-66.7%	-	-	-best results in the 2016 WSC competition
Classification task with NN (2018)	-	282-100% - 157-56%	282-30% - 177-63%	-first to use substitution of the pronoun with the antecedents
Google's language models (2018)	60-100% - 42-70%	273-100% - 173-63.7%	-	-no reasoning involved in the discovery of the correct answer -state-of-the-art for PDPs
OpenAI language models (2019)	-	273-100% - 193-70.70%	-	-current state-of-the-art for WSC -requires a lot of data for training -results are not reproducible
Graphs with Relevance theory (2014)	-	4-2.6% - 4-100%	-	-manual construction of graphs -first representation of WS as dependency graph
2 identified categories (2015)	-	71-25% - 49-69%	-	-first attempt of identifying commonsense knowledge types -developed the KParser
Knowledge hunting framework (2018)	-	273-100% - 119-43.5%	-	-refined query generation -developed an algorithm for scoring the retrieved sentences
Semantic relations categories (2019)	-	100-34% - 100-100%	138-14% - 111-80%	-provided Reasoning Algorithm -identified 12 commonsense types which capture the entire WSC

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 of the 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 of the 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 scores to both sentences**

$Score_{full}$(*"the trophy"*)$=$ ***P***(The trophy doesn't fit into the brown suitcase because **the trophy** is too small)

$Score_{partial}$(*"the trophy"*)$=$ ***P***(is too big |The trophy doesn't fit into the brown suitcase because **the trophy**)

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 of the 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 scores to both sentences**

$Score_{full}$(*"the trophy"*)= ***P***(The trophy doesn't fit into the brown suitcase because **the trophy** is too small)

$Score_{partial}$(*"the trophy"*)= ***P***(is too big |The trophy doesn't fit into the brown suitcase because **the trophy**)

- **Evaluation and results**
  - **PDPs 70% accuracy**
  - **WSC 63.7% accuracy**

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

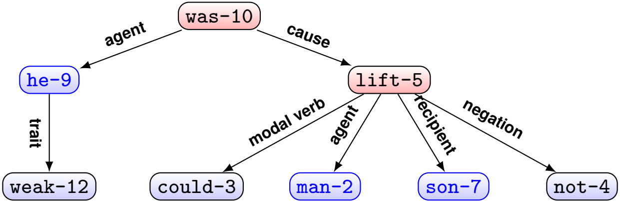
- Identified 12 knowledge types which cover the entire WSC dataset
- Categorization based on the structure of the Winograd sentence
- Developed a logical reasoning algorithm
- Evaluated on 100 problems from WSC and achieved 100% accuracy

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

- Identified 12 knowledge types which cover the entire WSC dataset
- Categorization based on the structure of the Winograd sentence
- Developed a logical reasoning algorithm
- Evaluated on 100 problems from WSC and achieved 100% accuracy
- Solver
  1. Semantic graph 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

Semantic graph representation[1]

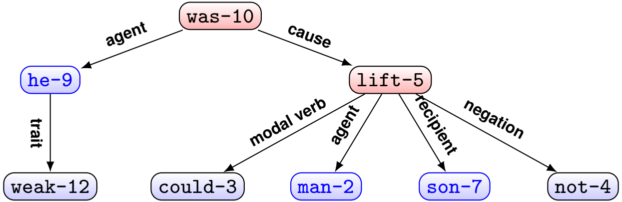
- **“The man couldn’t lift his son because he was so weak”.**



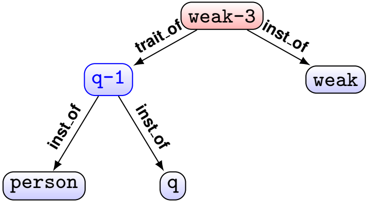
[1] kparser.org

Semantic graph representation[1]

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



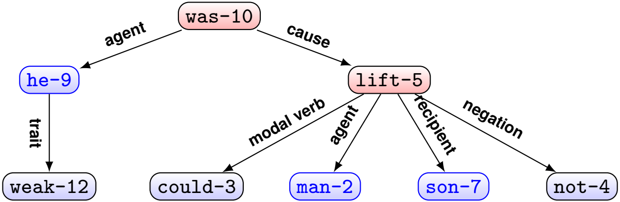
- "Who was weak?"



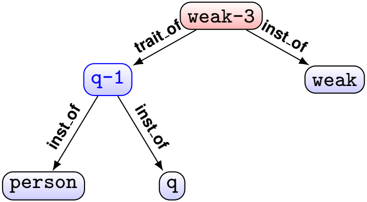
[1] kparser.org

Semantic graph representation[1]

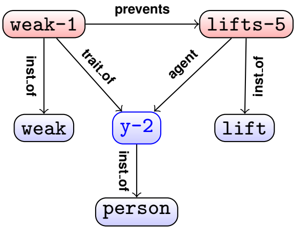
- “The man couldn’t lift his son because he was so weak”.



- “Who was weak?”

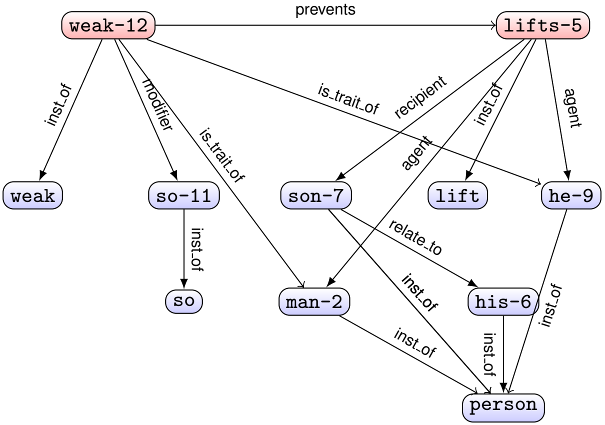


- “weak y prevents y lifts”

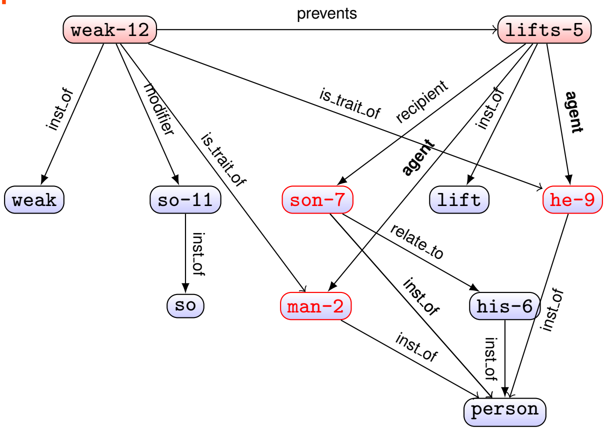


[1] kparser.org

Reasoning procedure



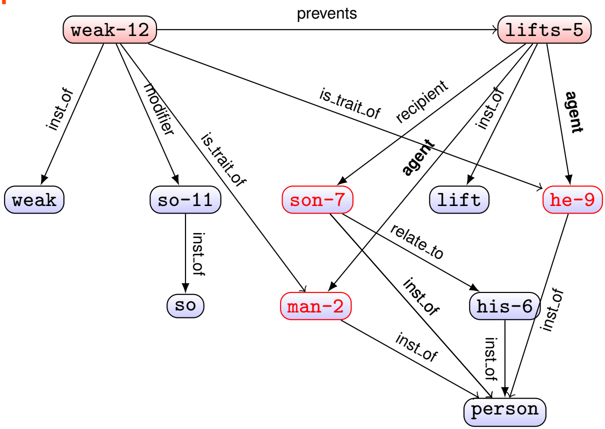
Reasoning procedure




```
has_k(weak,is_trait_of,y).
has_k(weak,prevents,lifts).
has_k(lifts,agent,y).
```

ans(q-1,he-9), ans(q-1,man-2)

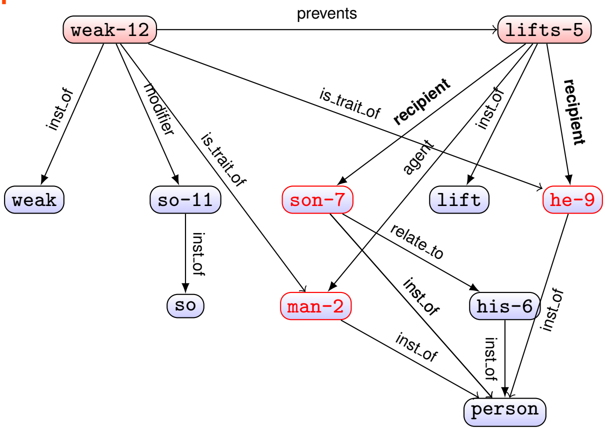
Reasoning procedure




```
has_k(weak,is_trait_of,y).
%has_k(weak,prevents,lifts).
has_k(lifts,agent,y).
```

ans(q-1,he-9), ans(q-1,man-2)

Reasoning procedure




```
has_k(weak,is_trait_of,y).
%has_k(weak,prevents,lifts).
has_k(lifts,recipient,y).
```

ans(q-1,he-9), ans(q-1,son-7)

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 of the least necessary knowledge**
    2. **Search for knowledge specific to this domain**
    3. **Apply reasoning procedure**
- **Categorization based on the content of the Winograd sentence**

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]?

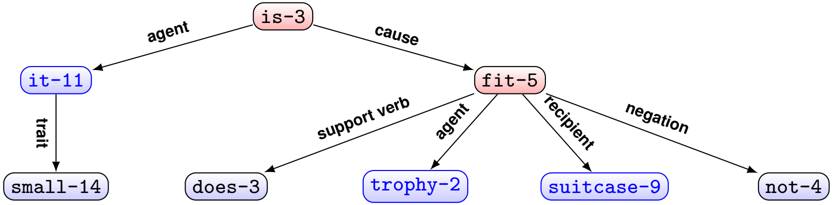
Annotation of Winograd Schemas

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

Category	Annotator 1	Annotator 2
Physical	36– 24%	39– 26%
Emotional	7– 4.6%	9– 6%
Interactions	44–29.3%	24–16%
Comparison	19–12.6%	26–17.3%
Causal	16–10.6%	18–12%
Multiple knowledge	28–18.6%	34–22.6%

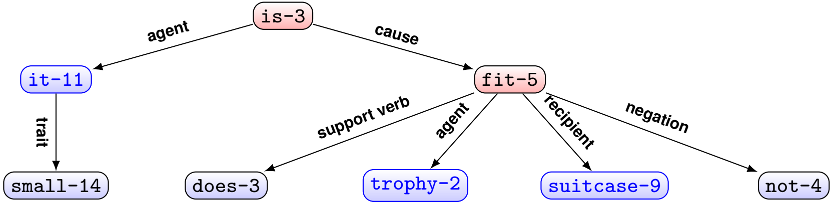
Graph Representation for Physical Category

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

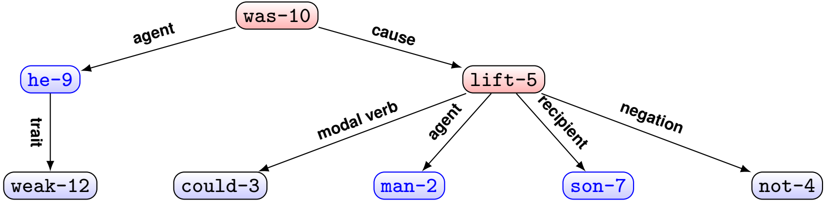


Graph Representation for Physical Category

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**

```
has_k(small,is_trait_of,y) :- has_k(fit, recipient, y),
                              not has_k(fit,modifier,could).
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**

```
has_k(small,is_trait_of,y) :- has_k(fit, recipient, y),
                              not has_k(fit,modifier,could).
has_k(weak,is_trait_of,y) :- has_k(lift, agent, y),
                             not has_k(lift,modifier,could).
```

- **Reasoning Algorithm**
- **Change of background knowledge**

- **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**

- **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

- **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!**

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.