Tackling Domain-specific Winograd Schemas with Knowledge-based Reasoning and Machine Learning

[1,2]Suk Joon Hong and [1]Brandon Bennett
[1]University of Leeds, United Kingdom, [2]InfoMining Co., South Korea

Introduction

- The **Winograd Schema Challenge (WSC)**[1] is to resolve the reference of pronouns occurring in natural language sentences.
- We tackle the WSC with knowledge-based reasoning(KR) and machine learning(ML). Here is an example from the WSC:

1. The trophy doesn't fit in the brown suitcase because **it** is too large.
   - The candidates : the trophy / the suitcase, Answer: **the trophy**
2. The trophy doesn't fit in the brown suitcase because **it** is too small.
   - The candidates : the trophy / the suitcase, Answer: **the suitcase**

Domains in WSC

- The thanking domain: the sentences that include "thank" and "grateful" were extracted from WinoGrande[2] (**171** out of 44K).
- Around **77%** of the sentences follow the five patterns.

High-level patterns in the thanking domain	
1	Candidate1 **owes** candidate2, and (so) pronoun is **doing good**
2	Candidate1 **owes** candidate2, and (so) pronoun is **receiving good**
3	Candidate1 **does good to** candidate2 because pronoun is **owing**
4	Candidate1 **gives thanks to** candidate2 because pronoun is **being owed**
5	Candidate1 **gives thanks to** candidate2 because pronoun is **owing**

Our Semantic-role Based KR Method

- Our KR method is built by modifying the method of Sharma[3].

1. Building a domain-specific knowledge base

- We define rules to derive semantic relations from K-Parser outputs

Semantic roles from K-Parser: X	Y	because relation	Semantic relationship
helper	being helped	No	Y owes X
helper	being helped	Yes	X does good to (repays) Y
giver	being given	No	Y owes X
giver	being given	Yes	X does good to (repays) Y
thanker	being thanked	Yes	X gives thanks to Y

2. Transforming a Winograd schema sentence into a high-level representation

- K-Parser and the domain-specific knowledge base are used.
- An example from WinoGrande: "**Kayla** cooked sticky white rice for **Jennifer**, and [**she**] was thanked for making such delicate rice."

Kayla	Jennifer	because relation	
giver	being given	No	Jennifer **owes** Kayla

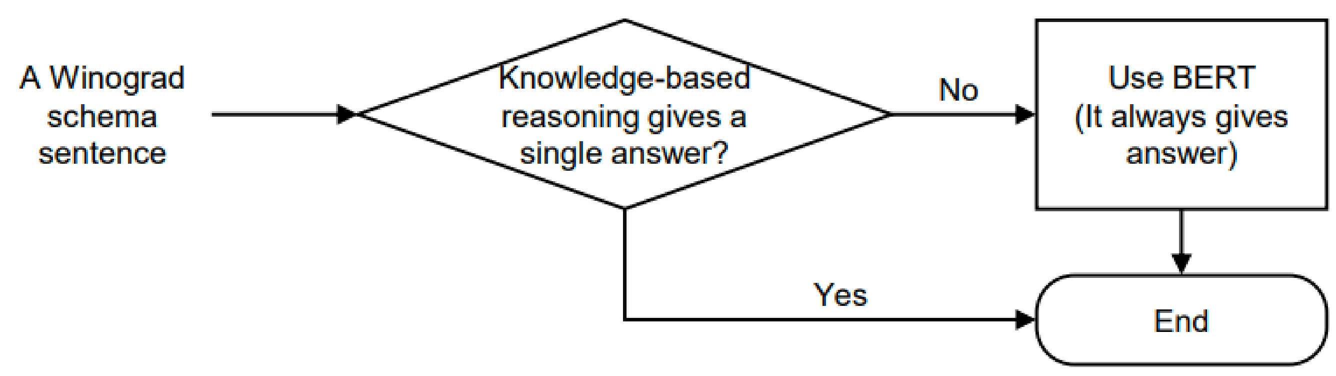
- **She** is **being thanked**, which is an instance of receiving good. Therefore, the sentence can be abstracted to **"Jennifer *owes* Kayla and she is *receiving good*."** This matches with the second high-level pattern.

3. Reasoning to derive the answer

- Answer Set Programming is used for reasoning.
- The answer can be derived by applying the background knowledge principles regarding the high-level patterns to **the abstracted sentence.**

Our Ensemble Method

- We propose a simple ensemble method by combining our semantic-role based KR method and ML (a fallback).



Robust Accuracy

- 'Robust Accuracy': A stricter form of accuracy measurement
- In addition to the switching[4], adding three more variants of each sentence by replacing the name of each candidate with the random name with the same gender
- **Predicting correctly on all the five sentences** is needed to be robustly accurate. Here is an example from WinoGrande (1: original, 2: switched, 3 ~ 5: replaced with random names):

The variants of the example sentence	
1	**Kayla** cooked sticky white rice for **Jennifer**, and [she] was thanked …
2	**Jennifer** cooked sticky white rice for **Kayla**, and [she] was thanked …
3	**Erin** cooked sticky white rice for **Tanya**, and [she] was thanked …
⋮	

Experiments

- The 80 paired Winograd schema sentences in the thanking domain were used for the experiments. In the first experiment, each pair was *split* into the train set and the test set, and in the second experiment, each pair was put together.

Results

- The accuracies and the robust accuracies of our ensemble model (KR + ML) are better than those of the other methods.
- The models that contain a language model were found to have lower robust accuracies than raw accuracies.

Model	First Experiment		Second experiment	
	Accuracy	Robust accuracy	Accuracy	Robust accuracy
GPT-2	50.0%	20.0%	57.5%	15.0%
BERT-large	57.5%	37.5%	57.5%	35.0%
Kocijan's BERT-large[5]	70.0%	62.5%	77.5%	70.0%
Kocijan's BERT-large further fine-tuned	47.5%	42.5%	75.0%	70.0%
Our KR method	72.5%	72.5%	37.5%	37.5%
Our ensemble method	**90.0%**	**85.0%**	**80.0%**	**72.5%**

Conclusion

- Our robust accuracy shows language models' predictions could be vulnerable to minor changes.
- We propose a high-level KR method based on semantic roles.
- Our keywords method is used to define the thanking domain, and it can be applied to specify other domains for future work.
- In our test set for the thanking domain, our ensemble method gives a better and more robust performance than the other approaches we tested.

References

1. Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In the 13th International Conference on Principles of Knowledge Representation and Reasoning, Italy, June 2012.
2. Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In AAAI-20, 2020.
3. Arpit Sharma. Using answer set programming for commonsense reasoning in the winograd schema challenge. arXiv:1907.11112[cs.AI], 2019.
4. Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie C. K. Cheung. How reasonable are common-sense reasoning tasks: A case-study on the winograd schema challenge and swag. arXiv:1811.01778[cs.LG], 2018.
5. Vid Kocijan, Ana M. Cretu, Oana M. Camburu, Yordan Yordanov, and Thomas Lukasiewicz. A surprisingly robust trick for winograd schema challenge. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4837—4842, 2019.