Tackling Domain-Specific Winograd Schemas with Knowledge-Based Reasoning and Machine Learning

Suk Joon Hong[1]
School of Mathematics, University of Leeds, UK
InfoMining Co., Seoul, South Korea

Brandon Bennett
School of Computing, University of Leeds, UK

Abstract

The *Winograd Schema Challenge* (WSC) is a commonsense reasoning task that requires background knowledge. In this paper, we contribute to tackling WSC in four ways. Firstly, we suggest a keyword method to define a restricted domain where distinctive high-level semantic patterns can be found. A *thanking domain* was defined by keywords, and the data set in this domain is used in our experiments. Secondly, we develop a high-level knowledge-based reasoning method using semantic roles which is based on the method of Sharma [17]. Thirdly, we propose an ensemble method to combine knowledge-based reasoning and machine learning which shows the best performance in our experiments. As a machine learning method, we used Bidirectional Encoder Representations from Transformers (BERT) [3, 9]. Lastly, in terms of evaluation, we suggest a "robust" accuracy measurement by modifying that of Trichelair et al. [20]. As with their switching method, we evaluate a model by considering its performance on trivial variants of each sentence in the test set.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases Commonsense Reasoning, Winograd Schema Challenge, Knowledge-based Reasoning, Machine Learning, Semantics

Digital Object Identifier 10.4230/OASIcs.LDK.2021.41

Related Version *Full Version*: `https://arxiv.org/abs/2011.12081`

Supplementary Material *Software (Source Code)*: `https://github.com/hsjplus/high-level-kb-reasoning`; archived at `swh:1:dir:bf9138cbf3a41a02809ac1de2dea41d499b9198e`

1 Introduction

The *Winograd Schema Challenge* (WSC) was proposed by Levesque et al. [10] as a means to test whether a machine has human-like intelligence. It is an alternative to the well known *Turing Test* (TT) and has been designed with the motivation of reducing certain problematic aspects that affect the TT. Specifically, while the TT is subjective in nature, the WSC provides a purely objective evaluation; and, whereas passing the TT requires a machine to behave in a deceptive way, the WSC takes the form of a positive demonstration of intelligent capability.

The core problem of the WSC is to resolve the reference of pronouns occurring in natural language sentences. To reduce the possibility that the task can be accomplished by procedures based on superficial or statistical characteristics, rather than "understanding" of the sentence, it is required that the test sentences used in the WSC should be constructed in pairs, which have similar structure and differ only in some key word or phrase, and such that the correct referent of the pronoun is different in the two cases. This sentence pair, together with an

[1] corresponding author

indication of which pronoun is to be resolved and a pair of two possible candidates, is called a *Winograd Schema*. An example of a Winograd Schema from the original WSC273 data set [10] is as follows:

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

Levesque et al. [10] design Winograd schemas to require background knowledge to resolve a pronoun, which can be an evidence of *understanding*. Therefore, they aim to exclude the sentences that can be resolved by a superficial statistical association within a sentence.

In this paper, we used a keyword method to define domains in Winograd schemas. To our knowledge, this is the first work to use keywords for defining domains in WSC and explore high-level patterns in them. To use the domain-specific high-level patterns, we also develop an advanced high-level knowledge-based reasoning method by modifying the method of Sharma [17]. Furthermore, we suggest a simple ensemble method that combines knowledge-based reasoning and machine learning. By the experiments on the domain-specific data set, the ensemble method gives a better performance than each single method. Lastly, we also propose a "robust" accuracy measure that is more objective by improving the switching method of Trichelair et al. [20].

2 Related work

Knowledge-based reasoning and machine learning are the two main approaches to resolve Winograd schemas.

Knowledge-based reasoning

The paper of Levesque et al. [10] is concerned with defining a test for AI rather than proposing how the challenge should be addressed. However, in the paper's conclusion they suggest that the knowledge representation (KR) approach is the most promising. They say: *"While this approach (KR) still faces tremendous scientific hurdles, we believe it remains the most likely path to success. That is, we believe that in order to pass the WSC, a system will need to have commonsense knowledge about space, time, physical reasoning, emotions, social constructs, and a wide variety of other domains."*

KR techniques make use of explicit symbolic representations of information and inference rules. A number of researchers have taken this kind of approach. Bailey et al. [1, p. 18] propose a "correlation calculus" for representing and reasoning with background knowledge principles and use this to derive solutions to certain Winograd schemas. Sharma [17] employs automated extraction of graphical representations of a sentence structure using a semantic parser called K-Parser [18] and implements a WSC resolution procedure based on Answer Set Programming (ASP) [5].

An advantage of KR-based methods is that they provide explanations of how the answers they give are justified by logical prinicples. However, KR-based methods also face huge problems both in automating the conversion from natural language sentences to a formal representation and also in building a knowledge base that covers the *general domain* of knowledge required to address the WSC. Bailey et al. [1] do not give an automatic method to transform a natural language sentence into the form of first-order logic that they use. Though Sharma et al. [19] do use an automated method to extract background knowledge, their method is based on using a search engine, which cannot guarantee acquiring all necessary knowledge.

Table 1 Two Examples from WSC273 with each variant by negation on which Kocijan's BERT was tested.

Type	Sentence	Pred.	Answer
Ori.	Dan had to stop Bill from toying with the injured bird. **He** is very compassionate.	**Dan**	Dan
Neg.	Dan had to stop Bill from toying with the injured bird. **He** is **not** compassionate.	**Dan**	Bill
Ori.	I can't cut that tree down with that axe; **it** is too small.	**The tree**	The axe
Neg.	I can't cut that tree down with that axe; **it** is **not** small.	**The tree**	The tree

Machine learning

Contrary to the expectations expressed by the proposers of the challenge (as cited in the previous section), many researchers have applied Machine Learning (ML) methods to the WSC, and, in terms of accuracy performance, impressive results have been obtained. An early work by Rahman and Ng [13] extracts features of a WSC-like sentence by using background knowledge such as Google search counts and a large corpus, and these features are used to train the SVM ranker that gives the higher rank to the correct candidate.

More recent ML approaches mostly use a neural language model. Trinh and Le [21] introduce an approach to use a neural language model to tackle Winograd schemas. After this, Bidirectional Encoder Representations from Transformers (BERT) [3], which is a state-of-the-art language model, is also used for WSC. Kocijan et al. [9] demonstrate that the BERT fine-tuned with the data set similar to Winograd schemas gives a better performance than the BERT without fine-tuning. In addition, Sakaguchi et al. [16] give the accuracy of around 90% on the original WSC273 by fine-tuning a variant of BERT with the larger data set (WinoGrande) which is also similar to Winograd schemas.

Despite the high accuracy of BERT and other neural language model methods, some limitations have been found. Though many of the original Winograd schemas can be resolved by the language models, Trichelair et al. [20] demonstrate that they often predict wrongly on simple variants of the original sentences. Specifically, when we switch the positions of the candidates, in most cases this means that the answer should also be switched. However, the language model methods frequently give the same prediction for the switched sentence as in the original sentence. We return to this matter of switching in Section 6. Their finding implies that the real understanding of the model cannot be guaranteed by accuracy only. Furthermore, Ettinger [4] also shows that the BERT does not seem to understand negation since BERT's predictions on the masked tokens of the negated sentences are likely to be similar to its predictions on the masked tokens of the non-negated sentences.

The finding of Ettinger [4] is also supported by recent study [11] and the experiments of Kocijan's BERT on some Winograd schema sentences from WSC273 that are negated by us in Table 1. Though the answers should be changed on the negated Winograd schema sentences in this example, the BERT's predictions on them are still same as its predictions on the non-negated sentences.

Table 2 The five major high-level domain-specific reasoning patterns found in the thanking domain.

Type	Sentence
Pattern 1	Candidate1 owes candidate2, and (so) pronoun is doing good
Pattern 2	Candidate1 owes candidate2, and (so) pronoun is receiving good
Pattern 3	Candidate1 does good to candidate2 because pronoun is owing
Pattern 4	Candidate1 gives thanks to candidate2 because pronoun is being owed
Pattern 5	Candidate1 gives thanks to candidate2 because pronoun is owing

3 Semantic Domains and Keywords

Several researchers in natural language processing have suggested that semantic domains can be identified based on the occurrence of key words in text corpora [14, 6]. Assuming that keywords are related to the high-level semantic meaning of a sentence, we used a keyword method in terms of identifying a domain in Winograd Schemas. To our best knowledge, our method is the first work to use keywords regarding a domain in Winograd schemas and examine high-level patterns in a domain. Although defining a domain by keywords has weakness such as word sense ambiguity, it can be beneficial for knowledge-based reasoning which requires relevant knowledge to tackle WSC. A keyword-based domain could target narrower Winograd schema sentences that are related to smaller number of background knowledge principles since they share at least one word. In this sense, building a knowledge base for a keyword-based domain can be less costly.

For the pilot study, we chose a *thanking* domain since the thanking domain has a distinctive semantics. The thanking domain contains the sentences that have a keyword related to the normal sense of thanking. The keywords we used for the thanking domain were "thank" and "grateful". We extracted sentences that include the two keywords from WinoGrande [16] which has approximately 44K Winograd schema sentences since WSC273 contains only 273 sentences. In this extraction, we exclude the sentences including "thanks to" and "thanks in no small part to" though "thank" is within them. The reason for their exclusion is that their semantic meaning is related to causal relations, not thanking.

As a result, the number of the extracted Winograd schema sentences was 171 ($\approx 0.39\%$ of the $44,000$ Winogrande sentences). We believe that the number of them is adequate as it is compatible with the number of the original WSC273's sentences which is 273. These extracted sentences are considered to belong to the thanking domain, and we investigated the high-level reasoning patterns in the thanking domain. As shown in Table 2, the five major high-level domain-specific reasoning patterns were found. As these patterns are from the thanking domain, they are related to the relationships of "owing" and "being owed". It is common that a person who is owing would thank or do good to someone who is owed. It is interesting that around 77% (132/171) of the sentences in the thanking domain follow the only five major high-level patterns. Some of the other minor high-level patterns were also found in the thanking domain.

In addition to the high-level patterns, the Winograd schema sentences in the thanking domain have two other characteristics. The first characteristic is that more than 90% (161/171) of the sentences in the thanking domain have candidates with human names while this proportion is around 50% in WSC273. This finding can be explained by the fact that thanking is done by humans. For the second characteristic, only around 46% (80/171) of the sentences in the thanking domain can be paired while almost all the sentences can be paired in WSC273. This is due to the fact that some of the WinoGrande sentences use keywords such as "thank" for the special words or the alternative words.

4 The advanced high-level knowledge-based reasoning method

Our high-level knowledge-based reasoning method is related to the method of Sharma [17], who identifies and exploits very specific identity implications to resolve pronouns. We use a more general method of abstracting semantic relationships to identify and make use of high-level domain-specific semantic roles, based on the analysis of Winograd schemas given by Bennett [2]. According to this analysis, most Winograd sentences can be represented as having the form:

$$\phi(a, b, p) \equiv ((\alpha(a) \ \wedge \ \beta(b) \ \wedge \rho(a, b)) \ \# \ \pi(p)) \tag{1}$$

where α is the candidate a's property, β is the candidate b's property, ρ refers to a predicate that defines the candidates' relationship, # refers to the relationship between the clause of the sentence that contains candidates and the clause of the sentence that contains the pronoun, and π is the pronoun p's property. In the most common cases the relationship # is "because", but it can also be other connectives such as "and", "but", "since", or sometimes just a full stop between sentences. For instance, consider this sentence from WinoGrande:

Lawrence thanked **Craig** profusely for the assistance ... because only [**he**] helped him.

Here a and b correspond to Lawrence and Craig, and the predicates α and β refer to the roles thanker and being thanked. p corresponds to the pronoun ("he") and the predicate π refers to the role of helper. ρ can refer to (a) giving thanks to (b) and # can be "because". While this type of formula can be used for particular examples of Winograd schemas, we also used the formula to represent higher-level general principles that can potentially explain a large class of specific cases.

4.1 Building a domain-specific knowledge base

Our knowledge base is composed of two types of rules and one type of facts – rules to derive semantic roles, rules to define relationships regarding the semantic roles and high-level background knowledge principles.

Rules to derive semantic roles

We defined rules to derive semantic roles specific to the thanking domain. These semantic roles are high-level representations related to the candidates and the pronoun, and they are also grounds to derive the relationships regarding them. In the thanking domain, six major domain-specific semantic roles were found – thanker, being thanked, giver, given, helper and being helped. In the current work, we assume that each person has a role in relation to the situation being described, and we formulate rules to derive and reason about these roles. (Potentially, someone could have different roles with respect to different aspects of the situation, which would require elaboration of our framework.)

Table 3 The major rules to define the relationships between the semantic roles of the candidates.

Semantic relationship	Causal relation	Semantic role	
		X	Y
X owes Y	No	being helped	helper
X owes Y	No	given	giver
X does good to Y	Yes	helper	being helped
X does good to Y	Yes	giver	given
X gives thanks to Y	Yes	thanker	being thanked

Our rules are implemented in ASP by using K-Parser's graphical representations, and they are manually defined from the sentences in the thanking domain. For example, a simple rule for thanker can be defined as:

```
has_s(X, semantic_role, thanker) :-
        has_s(Thank,agent,X),
        has_s(Thank,instance_of,thank).
```

In order to make more generalisable rules, the following four measures were taken. The first measure is to derive the semantic role of a candidate if that of the other candidate is known (e.g. if "give" is the semantic role of a candidate, then that of the other candidate would be "given"). The second measure is for the case when no semantic roles of the candidates are known. For instance, if candidate1 is an agent of the verb to which candidate2 is a recipient, candidate1's semantic role is derived to be "giver". The third measure is to use synonyms that are manually defined in the thanking domain. The fourth measure is to use an external sentiment lexicon dictionary [8] to derive the semantic roles of "good" and "bad".

Rules to define relationships regarding the semantic roles

The domain-specific semantic roles are used to derive their relationships for the high-level representations of Winograd schema sentences. We defined the rules for the relationships using the semantic roles in the following three aspects: relationships between the semantic roles of the candidates, relations between the clause containing the candidates and the clause containing the pronoun, and property of the pronoun.

1. **Relationships between the candidates' semantic roles.**
   As the five high-level patterns in Table 2 show, the two candidates in a Winograd schema are found to have domain-specific relationships in the thanking domain. The main relationships between them are "owes", "does good to" and "gives thanks to". In order to derive the relationships between the semantic roles of the candidates, we defined the rules by using their semantic roles and the existence of causal relation. Table 3 shows the five rules to derive the relationships between the candidates.

   For instance, the second rule in Table 3 means that if the semantic role of X is "given", that of Y is "giver", and there is no causal relation then X owes Y. It is written in ASP as:

   ```
   has_s(X, owes, Y) :-
       has_s(X,semantic_role,given),
       has_s(Y,semantic_role,giver),
       not has_s(_,relation,causer).
   ```

2. **The relationship between first and second clauses of the sentence.**
   As represented in Formula (1), the structure of a Winograd schema involves a relationship between the first clause of the sentence containing the candidates and the second clause containing the pronoun ("#"). In most cases we assume that there is some kind of implication from the first clause to the second clause, corresponding to some reasoning principle. However, if the sentence is of the form "P *because* Q", then the implication will go from the second clause to the first(Q to P). In this second case, K-Parser generates a `caused_by` relationship. Hence, we have a rule that when this relation is present, the agent of the second clause (i.e. the pronoun reference) has a causal role in the first clause of the sentence (i.e. corresponds to the candidate who is the agent in the first part). This rule can be defined in ASP as follows:

```
has_s(P, relation, causer) :-
    pronoun(P),
    is_candidate(A),
    has_s(Verb1,caused_by,Verb2),
    1 {has_s(Verb1,agent,A);
    has_s(Verb1,recipient,A)},
    has_s(Verb2,agent,P).
```

3. **Property of the pronoun.**
   The semantic role of the pronoun can be the property of the pronoun ("$\pi(p)$") in Formula (1), but there can be a higher-level semantic role. For this reason, we defined rules to derive the high-level semantic role from the low-level semantic role. These rules are based on the fact that a low-level semantic role can be *a subset of* a high-level semantic role in the thanking domain. For instance, the semantic role of "helper" can be a subset of that of "doing good". We implemented these rules in ASP, and the following rule is one of them:

```
has_s(X, semantic_role, doing_good) :-
    has_s(X, semantic_role, helper).
```

High-level background knowledge principles

In our knowledge base, we also defined high-level domain-specific background knowledge principles as well as the two types of the rules above. The high-level background knowledge principles are used for the reasoning in comparison with the high-level representation of a sentence that is derived by the rules in the knowledge base. We followed the style of Sharma [17]'s background knowledge principles as a foundation, but different from Sharma [17], our background knowledge principles are based on the semantic roles' relationships derived by our knowledge base.

4.2 Transforming a Winograd schema sentence into a high-level representation

We used K-Parser to transform the Winograd schema sentences in the thanking domain into the graphical representations as Sharma [17] does. By using the rules to derive semantic roles and to derive relationships between the semantic roles, we transformed the graphical representations into high-level representations. The following is an example of the transformations from WinoGrande:

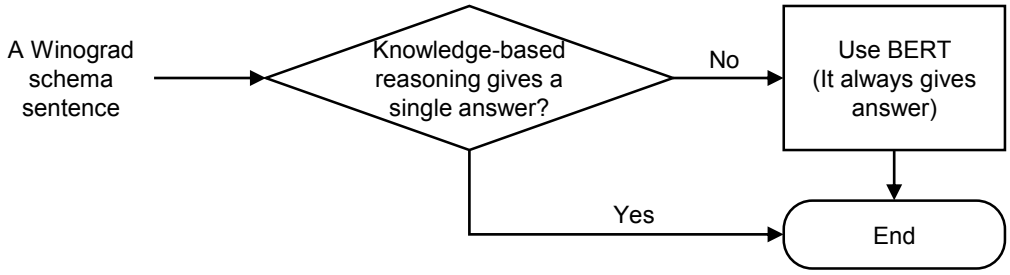

Figure 1 Our algorithmic flow of combining the knowledge-based reasoning method and the machine learning method.

Kayla cooked sticky white rice for **Jennifer**, and [**she**] was thanked for making such delicate rice.

- **The semantic roles:**
  1. Kayla: giver
  2. Jennifer: given
  3. she: being thanked
- **The relationships regarding the semantic roles:**
  1. Jennifer **owes** Kayla
  2. no causal relation
  3. she is receiving good

4.3 Reasoning to derive the answer

We used the reasoning rules of Sharma [17] with small modifications to resolve the Winograd schemas in the thanking domain. The goal of the modifications was to use the derived semantic roles for the reasoning.

In the reasoning process, each Winograd schema sentence is compared with each background knowledge principle. As a result, the answer for each sentence can be a single answer, "no answer" and multiple answers. If multiple answers have the same answers, this case is considered as a single answer.

As an example of the reasoning method, suppose a background knowledge principle is given in Sharma's form [17] as:

IF someone **owes** a person $p1$, and (consequently) a person $p2$ is receiving good **THEN** $p1$ is same as $p2$. (There is an assumption that owing occurs before receiving good.)

This background knowledge principle corresponds to the derived relationships regarding the semantic roles in the previous subsection. By applying the reasoning rules, $p1$ and $p2$ in the background knowledge principle correspond to "Kayla" and "she" in the sentence. Thus, the answer "she" = "Kayla" can be derived.

5 The simple ensemble method

We combined our advanced high-level knowledge-based reasoning method with Kocijan's BERT [9]. The aim of our ensemble method is to mitigate each method's weakness, and recent research [7] also suggests that machine learning and knowledge-based reasoning can complement each other. The weakness of the advanced high-level knowledge-based reasoning

method is that if there are no rules that can be applied in the knowledge base, no answer can be derived. With respect to weakness of language models such as BERT, their predictions are vulnerable to the small changes since it is not based on a logical relationship [20, 4].

As shown in Figure 1, we implemented a simple but effective ensemble method. If the knowledge-based reasoning method gives a single answer, the final answer will be this answer. On the other hand, if the prediction of the knowledge-based reasoning method is multiple answers or no answer, we use the BERT's prediction for the final answer. With these two conditions, the weakness of each method can be reduced.

6 "Robust" accuracy

As mentioned in Section 2, machine learning methods can give the incorrect answer on trivial variants of sentences obtained by switching the candidates [20]. This reveals an apparent weakness in these methods and a limitation in the simple evaluation of accuracy. Accuracy measurement is already quite tolerant because, since the number of the candidates are only two, the chance of predicting correctly without understanding is 50%. This is a further motivation for having a stricter form of accuracy measurement. We propose a "robust" accuracy measurement based on a generalisation of Trichelair et al. [20]. In addition to the switching, we add three more variants of each sentence by replacing the name of each candidate with the random name with the same gender if the candidates are both names. This basic method of replacing names should not affect the fundamental meaning of a sentence, and thus a model's incorrect predictions on the sentences where only the names are replaced can reveal its obvious lack of understanding. The following is an original sentence from WinoGrande in the thanking domain and its variants to measure the robust accuracy:

- **Original sentence**: **Kayla** cooked sticky white rice for **Jennifer**, and [she] was thanked for making such delicate rice.
- **The nouns switched**: **Jennifer** cooked sticky white rice for **Kayla**, and [she] was thanked for making such delicate rice.
- **The nouns replaced 1**: **Tanya** cooked sticky white rice for **Kayla**, and [she] was thanked for making such delicate rice.
- **The nouns replaced 2**: **Erin** cooked sticky white rice for **Tanya**, and [she] was thanked for making such delicate rice.
- **The nouns replaced 3**: **Lindsey** cooked sticky white rice for **Christine**, and [she] was thanked for making such delicate rice.

Only when a model predicts correctly on all of the original Winograd schema sentence and the four variants including the switched one, that prediction is considered to be "robustly" accurate. While the probability of predicting correctly on both switched and non-switched sentences out of luck is $0.5 \times 0.5 = 0.25$, the probability can go down to $(0.5)^5 \approx 0.03$ in the robust accuracy. In this sense, our robust accuracy is more objective on evaluating a model's performance. The limitation of the robust accuracy is that the candidates should be human names to make variants. In the case of no human names for the candidates, we only used the switching method to make a variant. This kind of exception is not common in the thanking domain since more than 90% of the sentences have the candidates with human names.

7 Evaluation

Our evaluation compares the performance of the following methods: GPT-2 [12], BERT-large [3], Kocijan's BERT-large [9], Kocijan's BERT-large further fine-tuned with the domain train set, our advanced high-level knowledge-based reasoning method and our ensemble method. When GPT-2 was used for resolving Winograd Schemas, partial scoring [21] was used to calculate the sentence probability of each candidate replacing the pronoun. Kocijan's BERT we used is their best performing model ("BERT_WIKI_WSCR") [9] which was fine-tuned with the WSC-like sentences[13]. We implemented Kocijan's BERT for our experiments by using the model and the code in their repository[2].

The six different methods were evaluated on the 80 paired Winograd schema sentences in the thanking domain, and the 91 non-paired sentences were used for validation. For the evaluation metrics, we used accuracy and our stricter "robust" accuracy measure.

We did two experiments with the paired sentences in the thanking domain. In the first experiment, each pair was *split* , so that one of the pair was put into the train set and the other into the test set. By its definition, 50% of the paired sentences were used for the train set, and the others were used for the test set. In the second experiment, on the other hand, each pair was put *together* either both in the train set or both in the test set in a random manner. Considering the small number of the data set and the balance with the first experiment, the second experiment also took the $50 : 50$ split between the train set and the test set.

7.1 Results

Tables 4 and 5 show the results of the two experiments respectively. Some same patterns were found in both experiments. The accuracies and the robust accuracies of our ensemble model are better than those of the other methods. Also, the models that contain a language model were found to have the lower robust accuracies than the accuracies. It demonstrates that language models, as machine learning methods, can be weak to minor changes.

Different patterns were also found between the two experiments. The accuracy of the knowledge-based reasoning method in the first experiment was higher than that in the second experiment by a large margin. It implies that the close similarity between the train set and the test set is advantageous for the knowledge-based reasoning method since the rules defined by the train set are expected to be used for the test set.

On the other hand, Kocijan's BERT-large further fine-tuned with the domain train set [9] gave the opposite results since the better accuracy was found in the second experiment, not in the first experiment. This result can be explained by the characteristics of Winograd schemas. While similar sentences have different answers in a Winograd schema, language models such as BERT are likely to give the same answer with that of the similar sentence, which leads to the wrong predictions in the first experiment. This result is compatible with the finding of Kocijan et al. [9] that training with the paired sentences shows a better performance than training with the non-paired sentences.

It is interesting that GPT-2 [12] and BERT-large [3] show the large gaps equal to or over 20% between accuracy and the "robust" accuracy in both experiments when they are not fine-tuned with WSC-like sentences. In contrast, the Kocijan's BERT-large models where fine-tuning was applied show the smaller gaps below 10% between accuracy and the "robust"

[2] `https://github.com/vid-koci/bert-commonsense`

Table 4 The results of the first experiment. These methods were tested on the same test set in the thanking domain with each pair split (between the train set and the test set).

Model	Accuracy	"Robust" accuracy
GPT-2 (no further fine-tuning) [12]	50.0% (20/40)	20.0% (8/40)
BERT-large (no further fine-tuning) [3]	57.5% (23/40)	37.5% (15/40)
Kocijan's BERT-large fine-tuned with the WSC-like data set [9]	70.0% (28/40)	62.5% (25/40)
Kocijan's BERT-large further fine-tuned with the domain train set	47.5% (19/40)	42.5% (17/40)
Our knowledge-based reasoning method	72.5% (29/40)	72.5% (29/40)
Our knowledge-based reasoning method + Kocijan's BERT-large [9] fine-tuned with the WSC-like data set[13]	**90.0**% (36/40)	**85.0**% (34/40)

accuracy in both experiments. This finding implies that the fine-tuning method applied to Kocijan's BERT-large can make language models more robust in terms of tackling Winograd schemas.

8 Conclusion

This paper demonstrates that combining both the high-level knowledge-based reasoning method and the BERT can give a better performance in the thanking domain.

In this paper, we also used the keywords method to identify a domain, and this method can be applied to specify other domains. We showed that high-level patterns were found in the domain defined by the keywords. As only one domain – the thanking domain – was tackled, future work needs to be done with more domains in Winograd schemas. Though the number of the thanking domain is 171 (around 0.39% of the number of the WinoGrande) as a pilot study, some other domains could be larger than the thanking domain. For instance, the domain that can be defined by the keywords "love" and "hate" has $1,351$ (around 3%) and 612 (around 1%) sentences respectively. If these were genuinely separate domains and the correct resolution of each schema were based on principles in the domain corresponding to the key words it contains, this would imply that tackling around 100 domains could cover almost all domains in Winograd schemas.

By modifying the method of Sharma [17] and focusing on the domain-specific semantic roles, we were able to develop a knowledge-based reasoning method that can use domain-specific high-level patterns. Though our knowledge-based method uses background knowledge principles that are built manually, we believe that our principles are more accurate than the kinds of semantic feature that could be reliably extracted from a large corpus or by using a search engine. This is because the simple statistical method used for automatically extracting

Table 5 The results of the second experiment. These methods were tested on the same test set in the thanking domain with pairs kept together (either both in the train set or both in the test set).

Model	Accuracy	"Robust" accuracy
GPT-2 (no further fine-tuning) [12]	57.5% (23/40)	15.0% (6/40)
BERT-large (no further fine-tuning)[3]	57.5% (23/40)	35.0% (14/40)
Kocijan's BERT-large [9] fine-tuned with the WSC-like data set[13]	77.5% (31/40)	70.0% (28/40)
Kocijan's BERT-large further fine-tuned with the domain train set	75.0% (30/40)	70.0% (28/40)
Our knowledge-based reasoning method	37.5% (15/40)	37.5% (15/40)
Our knowledge-based reasoning method + Kocijan's BERT-large [9] fine-tuned with the WSC-like data set[13]	**80.0**% (32/40)	**72.5**% (29/40)

knowledge is vulnerable to data bias or special usage of words in idioms (e.g. "thanks to" referring to causal relations that do not involve thanking in the normal sense of this concept). In addition, our knowledge-based method can also be used in other natural language tasks such as Choice Of Plausible Alternaties (COPA) [15]. But K-Parser used in our approach still needs to be improved as manual corrections were needed in some cases.

We also proposed the robust accuracy by improving the method of Trichelair et al. [20]. The decreased robust accuracies of language models such as BERT and GPT-2 reveal that their accuracy may not entail their real understanding.

Code repository

The code for the advanced high-level knowledge-based reasoning method (described in Section 4) can be accessed from the following repository: `https://github.com/hsjplus/high-level-kb-reasoning`

References

1 Daniel Bailey, Amelia Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. The winograd schema challenge and reasoning about correlation. In *2015 AAAI Spring Symposium Series*, USA, 2015.

2 Brandon Bennett. Logical analysis of winograd schemas. *Unpublished*, 2020.

3 Jacob Devlin, Ming W. Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv [cs.CL]*, 2018. `arXiv:1810.04805`.

4 Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020.

5 Michael Gelfond and Vladimir Lifschitz. The stable model semantics for logic programming. In *Proceedings of International Logic Programming Conference and Symposium*, pages 1070–1080, 1988.

6 Alfio Gliozzo and Carlo Strapparava. *Semantic Domains in Computational Linguistics*. Springer Berlin Heidelberg, 2009.

7 Pascal Hitzler, Federico Bianchi, Monireh Ebrahimi, and Md Kamruzzaman Sarker. Neural-symbolic integration and the semantic web. *Semantic Web*, 11(1):3–11, 2020.

8 Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2004)*, 2004.

9 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.

10 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.

11 Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. Towards generalizable neuro-symbolic systems for commonsense question answering. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 22–32. Association for Computational Linguistics, 2019.

12 Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

13 Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: The winograd schema challenge. In *EMNLP-CoNLL*, 2012.

14 Paul Rayson. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549, 2008.

15 Melissa Roemmele, Cosmin A. Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, USA, March 2011.

16 Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI-20*, 2020.

17 Arpit Sharma. Using answer set programming for commonsense reasoning in the winograd schema challenge. *arXiv [cs.AI]*, 2019. `arXiv:1907.11112`.

18 Arpit Sharma, Nguyen H. Vo, Somak Aditya, and Chitta Baral. Identifying various kinds of event mentions in k-parser output. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 82–88. Association for Computational Linguistics, 2015.

19 Arpit Sharma, Nguyen H. Vo, Somak Aditya, and Chitta Baral. Towards addressing the winograd schema challenge - building and using a semantic parser and a knowledge hunting module. In *IJCAI 2015*, pages 1319–1325, 2015.

20 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 [cs.LG]*, 2018. `arXiv:1811.01778`.

21 Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning. *arXiv [cs.AI]*, 2018. `arXiv:1806.02847`.