

KARaML: Integrating Knowledge-Based and Machine Learning Approaches to Solve the Winograd Schema Challenge

Suk Joon Hong¹, Brandon Bennett², Judith Clymo² and Lucía Gómez Álvarez³

¹InfoMining Co., South Korea

²University of Leeds, UK

³TU Dresden, Germany

Abstract

The Winograd Schema Challenge (WSC) is a commonsense reasoning task introduced as an alternative to the Turing Test. While machine learning approaches using language models show high performance on the original WSC data set, their performance degrades when tested on larger data sets. Moreover, they do not provide an interpretable explanation for their answers. To address these limitations, we present KARaML, a novel asymmetric method for integrating knowledge-based and machine learning approaches to tackle the WSC.

A central idea in our work is that *semantic roles* are key for the high-level commonsense reasoning involved in the WSC. We extract semantic roles using a knowledge-based reasoning system. For this, we use relational representations of natural language sentences and define high-level patterns encoded in Answer Set Programming to identify relationships between entities based on their semantic roles. We then use the BERT language model to find the semantic role that best matches the pronoun. BERT performs better at this task than on the general WSC. We apply our ensemble method to a restricted domain of the large WSC data set, WinoGrande, and demonstrate that it achieves better performance than a state of the art pure machine learning approach.

Keywords

Winograd Schema Challenge, Knowledge Representation, Machine Learning, Semantic Roles, Natural Language Understanding, Answer Set Programming, BERT

1. Introduction

The Winograd Schema Challenge (WSC) is a commonsense reasoning test proposed in [1] to demonstrate whether a machine is “capable of producing behaviour that we would say required thought in people”. The task of the WSC is to resolve which noun a pronoun refers to in a given sentence. Winograd schema (WS) examples are typically written in pairs (which we call Winograd schema pairs). These differ in only a few words, called the *special* and the *alternate*

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), *Proceedings of the AAAI 2022 Spring Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence (AAAI-MAKE 2022)*, Stanford University, Palo Alto, California, USA, March 21–23, 2022.


✉ hsplus89@gmail.com (S. J. Hong); B.Bennett@leeds.ac.uk (B. Bennett); J.C.Clymo@leeds.ac.uk (J. Clymo); lugo476b@tu-dresden.de (L. Gómez Álvarez)

🌐 bb-ai.net (B. Bennett)

🆔 0000-0001-5020-6478 (B. Bennett)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

words. Two candidate nouns are given alongside each schema as possible referents of the target pronoun (the same candidates for each schema in the pair), and the pronoun must be resolved in opposite ways depending on which of the special or alternate words was used. The use of schema pairs is intended to ensure that syntactic clues cannot help in finding the referent of the pronoun. Instead, this must be done by using world knowledge and reasoning. The original set of WSs, known as WSC273 [1], contains only 273 instances, but more recently a dataset of around 44000 following the same style was developed through crowd-sourcing [2]. An example from WSC273 is given below. Large and small are the special and the alternate words respectively:

- The trophy doesn't fit in the brown suitcase because **it** is too *large*.
the trophy (answer) / the suitcase.
- The trophy doesn't fit in the brown suitcase because **it** is too *small*.
the trophy / **the suitcase (answer)**.

Although the instigators of the WSC had originally envisaged that formalised theories of commonsense knowledge would be required to address the challenge [1], it has been tackled by a wide variety of approaches and has highlighted some serious difficulties that arise for Knowledge Representation (KR) approaches when applied to unconstrained, general problems of natural language understanding. By contrast, language models based on Machine Learning (ML) have achieved relatively good performance on WSC test sets although they do not employ any explicit representation of the detailed knowledge that seems to be involved in resolving WSC problems. Despite this success, the language model approaches have some weaknesses. Current language model methods are brittle, in that results are sensitive to small changes in the way a problem is expressed that are irrelevant to its solution. Language model approaches to the WSC so far do not provide any justification for the answers they give. As the WSC is supposed to test 'understanding', this is a significant limitation.

Our current work explores a combined KR and ML approach to the WSC. We call our system KARaML, standing for Knowledge Assimilation based on Roles and Machine Learning, and use the semantic roles of the agents participating in the situation described to resolve the WSC problem. We use the semantic parser K-Parser [3] to extract a relational semantic representation of the schema, and ASP-based rules to determine semantic roles of the candidate nouns. We then use the language model BERT [4] to match the pronoun to one of the extracted semantic roles. This allows us to leverage the implicit knowledge in the language model and so avoid manually building or attempting to explicitly learn a large knowledge base. By using the language model in a more focused way, rather than asking it to solve the whole task, our system is able to avoid some of the fragility commonly displayed by language models, and can provide an explanation alongside its decision. We have tested our approach on a subset of the large WSC data set, WinoGrande [2] and found that it performs better than pure ML methods using BERT [4, 5].

2. Related Work

Winograd schemas have been tackled by both KR and ML approaches. A typical KR approach would aim to resolve a WS by first translating the textual form of the schema into a logical

representation, then combining this with additional axiomatised background knowledge and using rules of inference to deduce the reference of the pronoun. Early work on AI systems for natural language understanding by Hobbs [6] proposed formalised principles of *coherence* that can account for co-references in many cases. However, he noted that in some cases, establishing the reference of a pronoun also requires detailed background knowledge. Indeed, the solution of most WS examples appears to involve knowledge concerning particular physical and/or social situations and understanding of vocabulary terms as well as general principles of communication and inference.

Sophisticated formal frameworks such as Segmented Discourse Representation Theory [7] have been developed in order to explain the logic underlying coherence and co-reference. However, the complexity of such theories has been an obstacle to their implementation in practical applications. Kehler *et al.* [8] and subsequently Bennett [9] gave formal analyses that account for certain WS cases. Schüller [10] presented a general method based on relevance theory and knowledge graphs. But the level of detail required to model knowledge relevant to specific cases suggests that the extension of these kinds of approach to incorporate sufficiently comprehensive knowledge to give general coverage of WS problems would be an enormous task. Bailey *et al.* [11] proposed a ‘correlation calculus’, which uses first-order logic with a novel correlation connective, to resolve WSs. This offers the prospect of a more general form of KR-based solution in which the complex types of correlation involved in solving the WSC might be inferred from simpler assumptions but would still require large numbers of basic correlations to be represented in order to cover the huge variety of possible WS problems.

A possible way to make KR approaches more effective for particular problem types may be to focus on aspects of semantics that are especially salient for those problems. We believe that the notion of ‘semantic role’ is such an aspect, which is often decisive in establishing co-reference and hence in solving WS problems. Semantic Role Labelling (SRL) is considered to be a significant computational task for natural language understanding, and can be carried out with high accuracy by some existing systems (such as SENNA [12]), and a method of using semantic roles for co-reference resolution is described in [13]. In NLP semantic roles are primarily defined in terms of the linkage of noun phrases to verbs (e.g. as ‘subject’, ‘object’ etc.). However, in the current paper we advocate a more general idea of semantic role that is held in relation to an activity (e.g. helping, needing help) and is not strictly tied to particular verbs and grammar. This idea of semantic role is akin to that adopted in *Frame Semantics* [14].

Many systems have been developed that can translate from natural language text into some form of logical representation [15, 16]. This ‘semantic parsing’ task is extremely challenging and the results obtained are unreliable, especially for complex sentences such as those occurring in WSs. Nevertheless, the extracted representations do identify entities, properties, relationships and logical structures that can be processed by KR-based reasoning systems. Sharma [17] developed a semantic parser, K-Parser [3] to transform schemas into relational representations, and uses these to resolve WSs. This method enhanced the extracted semantic content using rules formulated in Answer Set Programming (ASP) [18]. The initial use of this method for solving WS problems also required hand crafted representation of relevant background knowledge. The method shows accuracy of around 80% on the original WSC273 set when relational representations of both schemas and background knowledge principles are manually created. To address the problem of encoding sufficient knowledge to cover a wide class of commonsense

reasoning problems, various automatic *knowledge extraction* techniques have been employed. Sharma was able to achieve a more automated solution by extracting background knowledge using Google search to obtain identity rules enabling pronoun resolution [17]. However, fewer than half of the required rules could be obtained by this automated method.

In our previous work [19] we built on Sharma’s method [20]. We used K-Parser with additional hand-coded ASP rules to extract semantic roles of the candidate nouns, similar to the pattern-based semantic relation extraction of Al-yahya *et al.* [21]. Further logical rules were then used to determine the pronoun’s referent based on its semantic role and those of the two candidates.

Regarding ML approaches, Rahman and Ng [22] obtained promising results using an SVM ranker based on a variety of linguistic features, both semantic and syntactic. More recently, approaches based on neural network language models have made significant progress on the WSC task. Using the BERT [4] language model, high accuracy for resolving WSs has been demonstrated [23, 5, 2], with up to 90% accuracy reported for WSC273 [2]. Using the BERT variant RoBERTa, which has been found to perform better on many tasks, similarly high accuracy has been obtained [24, 2].

However, it is too early to claim that machines have reached human-like ability to resolve Winograd schemas. WSC273 is a very small test set, and accuracy has been found to decrease by around 10% or more on larger WSC-like data sets. Consequently, some researchers have suggested that the strong performance on WSC273 may overstate the capability of neural language models to carry out commonsense reasoning tasks [2, 25]. Tests that focus on cases involving compositional logical structure indicate that BERT does not work well in relation to function words such as negation [26]. BERT also seems to lack robustness with respect to irrelevant small variations: simply changing proper names can cause it to give incorrect answers to some WSs which were previously answered correctly [23, 19]. This suggests that language models may work by recognising features that are, at least in some cases, only indirectly connected with genuine understanding of WS problems. This also relates to issues of transparency and explainability. Humans would expect answers to be based on general principles, whereas current methods based on language models do not provide any meaningful explanation for their answers. Whereas humans appear to employ both commonsense reasoning and intuition [27], neural language models seem to work in a way that is more similar to intuition than to logical reasoning.

In this paper we attempt to develop a new way to combine KR and ML in order to address the WSC and contribute to exploration of the general problem of natural language understanding.

3. Winograd Schema Structure and Semantic Roles

In this section we examine the syntactic and semantic structure of Winograd Schema problems in order to motivate and explain our resolution method.

3.1. Schema Structure

A WS is a sequence of tokens in which three (non-overlapping) sub-sequences are indicated: two words or phrases referring to ‘candidate’ entities and one pronoun (normally a single word). Thus, it is an expression of the form $\psi(a, b, p)$, whose meaning constrains the references the

candidate terms a and b and the pronoun p . For the expression to be considered satisfactory as a WS, any reasonable human being should either infer from it that p refers to the same thing as a , or infer from it that p refers to the same thing as b .

In nearly all WS examples, there is a clear division into two propositional components, with the first component describing a situation involving both candidates, a and b , and the second giving information involving p . Hence, a WS normally has the structure $\phi(a, b) \# \pi(p)$, where ‘#’ represents the type of connection between the two parts. In many cases the two parts are separate sentences. For these cases we can treat the connective as logical conjunction (although temporal sequence may also be implied). In other cases, the halves may be connected with words such as ‘and’, ‘because’, ‘although’, ‘since’ etc. The particular connective is relevant to pronoun resolution.¹

So the pronoun resolution problem has the following form:

$$(\phi(a, b) \# \pi(p)) \wedge p = (a|b) \rightsquigarrow (p = \kappa), \quad (1)$$

where \rightsquigarrow represents some kind of rational inference relation and κ is either a or b . The presupposition that p must be identified with exactly one of the two candidates, is represented by the notation $p = (a|b)$.

Given that we need to infer an identity between p and either a or b , there must be some aspect of the content of $\phi(a, b)$ which can be linked to the content of $\pi(p)$ in such a way that either a or b can be distinguished as the more likely co-referent of p . One way to approach this would be to tease out from ϕ what is said individually about each of the candidates and try to link that to π . Indeed, by means of semantic intuitions or by using an automated semantic parser, a given proposition $\phi(a, b)$ can typically be analysed into a combination of simpler components, $\alpha(a) \wedge \beta(b) \wedge \rho(a, b) \wedge \gamma$, where α and β represent conditions that are individually ascribed to candidates a and b respectively, ρ represents whatever information is asserted about the relationship between a and b , and γ is any additional information that does not directly involve a or b . More specifically, each of the components $\alpha, \beta, \rho, \gamma$, may correspond to a (possibly empty) set of predicates in the semantic analysis.

In ordinary natural language, there are many examples where the reference of the pronoun can be resolved just by considering the individual properties of potential candidates (α and β). Lesvesque *et. al.* [1] consider the example ‘*The women stopped taking the pills because **they** were [pregnant/carcinogenic]*’. However, although this seems to be a typical use of a pronoun, it is *not* considered to be a good schema. Lesvesque *et. al.* explicitly say that this is a poor example, since correct resolution can be determined just by considering the types of the candidates (‘women’ and ‘pills’) and the types of entity of which the attributes ‘pregnant’ and ‘carginogenic’ could be predicated. Such cases are considered too easy to demonstrate that intelligence is required to resolve them. They suggest that suitably difficult WS examples must require understanding of the situation. This would typically involve the relationship between the candidates or some property that is not merely a simple type attribute of one of the candidates.

¹For instance, in the WS273 set presented by Levesque *et al.* [1] includes the example ‘Pete envies martin [because/although] he is successful’, where swapping ‘because’ with ‘although’ changes the pronoun reference. This case was also considered by [11], which suggests that, whereas ‘because’ implies positive correlation, ‘although’ implies negative correlation.’

3.2. Semantic Role Extraction

In the majority of cases we have examined, inferences based on individual properties $\alpha(a)$ and $\beta(b)$ are not enough. In order to resolve the pronoun, one needs to extract further attributes of a and b from the *roles* they play in the relation $\rho(a, b)$. By introducing s to stand for the situation described by the relationship $\rho(a, b)$ (i.e. we *reify* the relationship), we can conceptually unpack the relation into a conjunction $\rho_1(a, s) \wedge \rho_2(b, s) \wedge \omega(s)$ representing the *semantic* roles, ρ_1 and ρ_2 of the participants in relation to s , together with any other information (ω) attributed to the situation. Furthermore, if we are concerned with distinguishing a and b in terms semantic roles that occur in some particular types of situation (e.g. situations where one person helps another), then the relevant role information can be represented by unary role properties, $\rho_1(a)$ and $\rho_2(b)$ (e.g. a gives help, and b receives help).

In fact, existing semantic parsers (such as K-Parser and SENNA) already assign role attributes to referring constituents of sentences. However, these tend to be lacking in specific semantic content and determined largely by syntactic features of their occurrence within the text. For, example a referential word or phrase might be labelled as the ‘agent’, or ‘object’ of a verb. But, like entity types, such basic role types can only be used to resolve pronouns in ‘easy’ cases. In more complex cases, pronoun resolution requires understanding the way in which entities participate in a situation; and this requires specific knowledge of the situation and the roles it involves. Thus, we suggest that pronoun resolution in WSC problems requires an additional role extraction mechanism (RE) going beyond an initial semantic parsing (SP) stage. Hence, the semantic role extraction process can be represented by the following pattern:

$$\psi(a, b, p) \equiv \phi(a, b) \# \pi(p) \xrightarrow{\text{SP}} \alpha(a) \wedge \beta(b) \wedge \rho(a, b) \wedge \gamma \xrightarrow{\text{RE}} \rho_1(a) \wedge \rho_2(b) \quad (2)$$

To illustrate our analysis, we consider the sentence “*Maria is struggling with her exams and asks for help from Rebecca, because **she** is already successful.*”. Semantic parsing will produce a formal representation similar to the following:

$$(\text{struggling_with_exams}(\text{Maria}) \wedge \text{ask_help}(\text{Maria}, \text{Rebecca}) \text{ because } \text{successful}(\text{she}))$$

from which we want to infer $p = \text{Rebecca}$.

In this example, $\alpha(a)$ corresponds to the unary property $\text{struggling_with_exams}(\text{Maria})$, $\rho(a, b)$ is the relation $\text{ask_help}(\text{Maria}, \text{Rebecca})$, the connective $\#$ is *because* and $\pi(p)$ is $\text{successful}(\text{she})$. We have not specified any individual condition β predicated of *Rebecca*, although if we identified it as a proper name of a person (e.g. by using a named-entity recognition system [28]) we could add the individual condition ‘ $\text{Person}(\text{Rebecca})$ ’. Role extraction rules, as explained later in the paper, can then be employed to infer the semantic roles of the participants.

3.3. Resolving the Pronoun

The previous subsection examined the semantic structure of WSs and motivated the extraction of semantic role attributes of the candidates from the first part of the schema. We now explain how this can be used to identify the reference of the pronoun in the second part of the schema.

Our general idea is related to the approach of Bailey *et al.* [11], who proposed an extension of first-order logic with a novel propositional connective. The statement $A \oplus B$ means that

the truth of A is positively *correlated* with the truth of B , in the sense that if a rational agent becomes aware of the truth of either of the propositions they will consider the other proposition more plausible than they would have in the absence of that information. The paper presents a proof system to capture the logic of the ‘ \oplus ’ operator and suggest that it can be used to derive complex correlations from basic correlation assumptions and beliefs. Then these derived correlations can be used for pronoun resolution. Assuming that what is said about the entity via the pronoun reference is positively correlated with what is said about it in the candidate phrase, we should be able to infer either $\phi(a, b) \oplus \pi(a)$ or $\phi(a, b) \oplus \pi(b)$ when given a schema $\phi(a, b) \# \pi(p)$.

The correlation calculus is proved to be sound with respect to statistical semantics. And, although specification of the calculus predates the successful application of language models to the WSC, it seems that it would be well suited to interfacing with a language model. Instead of requiring correlations to be determined by axioms and logical reasoning, one could potentially evaluate or compare degrees of correlation by means of language model responses.

In our setting the relevant notion of correlation is a little different. We aim to find a correlation between the role description of one of the candidates and the description involving the pronoun. Also we look for a preferential rather than an absolute correlation. Thus, we wish to determine which of the semantic roles of the candidates is more likely to apply to the pronoun, given what is said regarding the pronoun. Hence, given the extracted roles $\rho_1(a)$ and $\rho_2(b)$ and assertion $\pi(p)$ regarding the pronoun, then if p denotes a we would expect the following inequality of relative probabilities:

$$P(\rho_1(p) \mid \pi(p)) > P(\rho_2(p) \mid \pi(p)) \quad (3)$$

Note that what is said in the proposition $\pi(p)$ does not need to explicitly describe p in terms of either of the roles ρ_1 or ρ_2 ; it only needs to provide some reason to expect that one of the potential facts $\rho_1(p)$ or $\rho_2(p)$ is more likely than the other.

4. Our Approach: KARaML

In this section we introduce our system KARaML. We use the semantic roles of the agents to resolve the WSC, following the analysis in Section 3.

Figure 1 illustrates the pipeline of our method to resolve WSs. KARaML uses a combination of KR and ML methods to derive semantic roles of the candidates and pronouns by defining domain-specific background knowledge relating to these high-level semantic roles. In the figure, the element labelled ‘Semantic parsing & KR role derivation’ relates to Section 3.2 above, and the element labelled ‘LM semantic role matching’ to Section 3.3. Finally, the ‘Semantic role based reasoning’ component uses the previously derived knowledge to infer the solution. If our combined system does not have suitable rules defined for resolving a schema, we simply revert to using the language model alone. Other important features of our architecture are the asymmetric combination of KR and ML, and the selection of conceptually related sentences that follow target patterns. We address these features in detail in the coming subsections.

After addressing in detail the architecture of KARaML (in Section 5), we give results (in Section 6) which show that where our combined reasoning method is applicable, we achieve

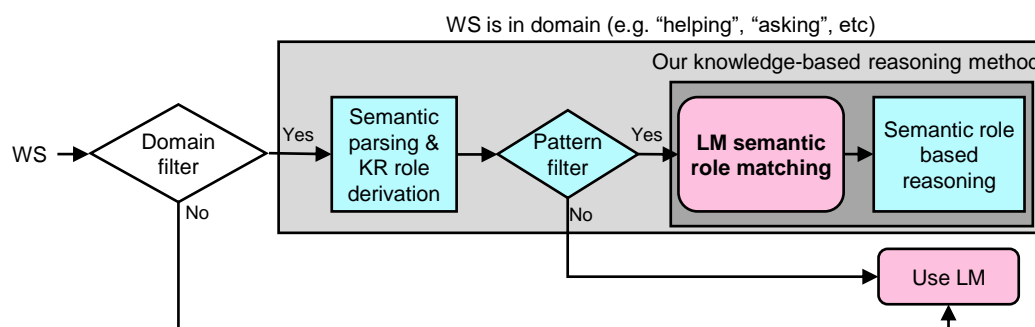


Figure 1: KARaML System Flow

better performance than using a language model alone.

A major difference from our previous work [19] is that we no longer need detailed axiomatisations of the domain’s background knowledge to infer the semantic role of the pronoun, which presented a challenge to scalability of the method. Instead, we will show that a minimal set of high-level rules for the semantic roles, coupled with the usage of a language model, is enough to obtain significant results.

4.1. An Asymmetric Combination of KR and ML

A notable feature of our approach is that we apply KR and ML methods asymmetrically with respect to different parts of a WS. Specifically the KR mode of interpretation is focused on the part of the WS that describes the candidates, whereas a neural language model is used to match a correlated semantic role for the pronoun.

A formal representation of a sentence has a rigid structure composed from specific symbolic vocabulary. This means that if we have KR representations of two related pieces of information (such as two successive sentences or clauses within a sentence) we can only draw inferences from their combined content if we have some way of aligning them. This requires both combining them in terms of a formal syntax, and also making explicit all significant semantic relationships between the vocabulary of the two parts. When dealing with representations extracted from natural language, this is a huge challenge. Not only are there an unbounded number of possible situations that might be described, but even one situation could be described in a wide variety of ways, using a wide variety of vocabulary terms. Hence, piecing together KR representations extracted from different parts of a natural language text is extremely difficult, even when connections are very clear to our intuitive understanding.

By contrast, ML techniques are more malleable, in that they do not require exact matching in order to connect one piece to another, so they can provide a mechanism for flexibly assimilating or adjoining new information to an existing KR representation. Figure 2 illustrates the potential advantage of this type of asymmetric combination.

It may still seem puzzling why we are always focusing the use of KR on the left side of the WS and reserving ML for interpreting the role of the right side. This is because our KR analysis is designed to extract roles of the candidates in WSs and, in the majority of examples, these are

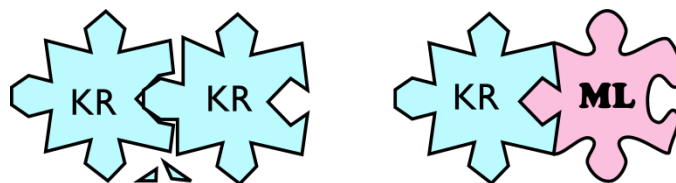


Figure 2: Knowledge Assimilation: KR+KR vs. KR+ML showing the advantage of ML’s flexibility

described primarily in the first part of the schema. In general, pronouns nearly always occur after the noun or noun-phrase with which they co-refer. In most WS examples the pronoun occurs in a following sentence or clause that does not usually make explicit the role of the pronoun referent in a way that can be directly linked to the roles of the candidates. Nevertheless, one might intuitively expect that there is a statistical correlation between the roles of the candidates in the first part and what is then said using the pronoun in the second part. Indeed, our results indicate that ML techniques can model this correlation.

4.2. Identifying Conceptually Related Sentences

In general, a WS may involve any vocabulary or domain of knowledge. This is problematic for KR approaches, which require detailed logical modelling of knowledge and semantics. We use keywords to identify restricted domains that are more manageable. Our aim is to provide a simple method for selecting related schemas for which high-level background knowledge rules can be defined.

A small number of logical rules should be sufficient to explain a significant proportion of schemas in a semantic domain. In particular we present in this paper a study of schemas obtained by identifying instances containing the keyword ‘help’ (or ‘helping’, ‘helped’, ‘helpful’ etc.). We show that the same principles extend to a larger set including also schemas that contain the keyword ‘ask’ (or ‘asking’, ‘asked’, etc.), for which only six additional rules needed to be established. This shows that domains defined in this way are flexible and able to encompass a variety of schemas. We have previously presented work on schemas containing the ‘thanking’ keyword [19]. Although our current work focuses on a few hand selected domains, it demonstrates a general approach which could be extended to cover a larger proportion of WinoGrande schemas.

In our system, WSs are first filtered for use of keywords and then compared with high-level patterns. It should be expected that there will be some overlap between domains, where a sentence references multiple concepts. If a pattern is matched, this indicates that we have suitable rules defined to understand this sentence. In the case that a sentence matches multiple patterns we propose using the correlation between candidate and pronoun roles which is identified as most significant by the language model (i.e. lowest loss).

A sentence may use knowledge from a domain without containing the relevant keyword. Provided that the sentences containing the keywords are representative of the domain and allow us to generate appropriate rules, this is not a significant limitation. We anticipate that our methods for identifying semantic roles may be extended to sentences which do not contain the relevant keyword, allowing more sentences to be resolved using the existing rules.

5. KARaML System Architecture

We now describe how we have implemented each component of KARaML from Figure 1. We first tackle the domain filter, and subsequently we introduce an ASP pattern filter based on K-Parser output. The pattern filter selects WSs that match certain semantic roles, for which domain-specific background knowledge has been encoded in ASP. Next, BERT² is used to determine which of a pair of contrasting semantic roles the pronoun has a stronger correlation to. By using the pattern filter together with the background knowledge we can infer the high-level (and not necessarily explicit) semantic roles of *a* and *b* that BERT will choose from. Finally, the derived semantic roles for the candidates and the best-matching role for the pronoun are used to infer the final answer.

In what follows we will give a detailed explanation of our system architecture and we will use a sample schema from WinoGrande as a running example:

Maria helped Elena cope with the newly diagnosed autism because **she** was *inexperienced* with the disorder.
 Maria / **Elena (answer)**

In this case, Maria (*a*) performs the semantic role of *helper* and Elena (*b*) performs the semantic role of *being_helped*. Our proposal is that the correlation between the semantic roles of the candidates and the information we have about the pronoun is a good indicator for the pronoun resolution. In the example, we note that a person being inexperienced in a situation is more likely to explain (“*because*”) needing help than giving help.

While the role **Maria** : *helper* can be derived with a relatively simple KR system, deriving that an inexperienced person is in need of help (**she** : *needing_help*) previously required further manually defined rules [19]. In our current work, this task is given to a language model, which can make use of its implicit understanding of the correlation between inexperience and need.

5.1. Domain Filter

Our pipeline begins with a domain filter that identifies the schemas that may be associated with a domain. A WS is passed into the filter, which determines whether it belongs to any of the pre-defined domains by using keywords. Our running example will be categorised using the “help” keyword. If a schema does not belong to any pre-defined domains, it is categorised as out-of-domain and will be resolved by BERT.

In our experiments, we begin by narrowing our attention to a domain centered around the keyword “help”, which contains 1356 schemas. Subsequently, we target schemas containing the keyword “ask”, amounting to 1753, which in fact respond to the same underlying patterns, thus giving rise to a more general domain. Indeed, these sets of schemas intersect, which gives further evidence that they share a common underlying semantic structure.

²Specifically, we use BERT_WIKI_WSCR from [5] throughout. This is an instance of BERT which has been additionally fine-tuned for the WSC.

5.2. Parsing and Deriving Semantic Roles

The schemas that have been assigned to domains are parsed by K-Parser, which produces a relational semantic representation of the input text containing qualitative information about the words in the text (e.g. their conceptual classes, the relationship between predications and their participants among other). Then, high-level semantic roles are derived using the output of K-Parser together with our domain-specific background knowledge rules.

Let us look at an excerpt from the parsed output of the sample schema:

```
has_s( helped_2, agent, maria_1 ).
has_s( helped_2, instance_of, help ).
has_s( she_11, trait, inexperienced_13 ).
has_s( inexperienced_13, instance_of, inexperienced ).
has_s( cope_4, caused_by, was_12 ).
```

The output from K-Parser provides us with an initial representation of a given schema, which is specified by means of predicates of the form `has_s(node1, relation, node2)`. Subsequently, the domain-specific rules are used to expand the output with relevant background knowledge for the domain, which is mostly focused on the derivation of high-level semantic roles. Two simple examples of such domain-specific rules are as follows:

```
has_s( X, semantic_role, helper ) :-
    has_s( Action, agent, X ),
    has_s( Action, instance_of, help ).

has_s( X, semantic_role, being_asked ) :-
    has_s( Ask, recipient, X ),
    has_s( Ask, instance_of, ask ).
```

Using the first rule we can straightforwardly derive `has_s(maria_1, semantic_role, helper)` as desired for our running example.

5.3. The Pattern Filter

The parsed results together with the derived semantic roles of the schema are used as inputs to the pattern filter. If a certain pattern is found by the pattern filter, that schema is to be resolved by our combined framework. If not, just BERT is used for resolving the WSs.

Our pattern filter exploits the generic structure given in formula (1) to select schemas that follow recognised patterns, using the high level semantic roles previously inferred in section 5.2. Patterns in our system are encoded in ASP and will typically fix semantic roles for one or more of the agents *a* and *b*, and possibly additional restrictions on other elements of the schema, such as forcing # to be “because” and the pronoun to be a person (rather than an inanimate object).

In this experiment we use a pattern to identify schemas where the roles of “helper” and “being helped” are likely to be relevant for pronoun resolution. The filter checks whether the semantic properties and relations extracted satisfy the conditions that: at least one of the candidate expressions has one of these roles; the pronoun refers to a person (is “he” or “she”) which is the agent of a verb in the sentence that has been identified as playing an explanatory role in the situation. The relevant pattern is defined as follows:

```

help_pattern :-
    is_candidate( C ),
    1{has_s( C, semantic_role, helper );
      has_s( C, semantic_role, being_helped )},
    pronoun( P ),
    has_s( Verb, agent, P ),
    1{has_s( P, instance_of, he );
      has_s( P, instance_of, she )},
    has_s( _, caused_by, Verb ).

```

From the 1356 schemas containing the keyword “help”, 207 satisfy this pattern, and from the 1753 schemas containing the keyword “ask”, 456 schemas match a similar pattern.

5.4. Using BERT to Identify Semantic Roles

In this phase, schemas that meet the pattern are given to BERT to extract an implicit semantic role, where the possible roles are determined by the pattern that the schema matched. Previous work (e.g. [5, 2]) uses language models such as BERT to evaluate the probabilities of each of the candidates occurring as a replacement of the pronoun. So to resolve p in $\psi(a, b, p)$ one would replace p with [MASK] and compare probabilities for [MASK] to be a or b .

$$P([MASK]=a \mid \psi(a, b, [MASK])) > P([MASK]=b \mid \psi(a, b, [MASK]))$$

In contrast, we focus on deriving which of the candidate semantic roles is most correlated to the information that is given about the pronoun. To derive this using BERT, we extract the textual fragment $\pi(p)$ and concatenate it (following a period) with a basic sentence linking the pronoun with the masked semantic role. In our experiment, we added the sentence “he/she would [MASK] help”, where the [MASK] can be either *give* or *need*. Now BERT compares the probabilities for [MASK] to be *give* or *need*, given the context $\pi(p)$ and the additional linking sentence (p would [MASK] help).

$$P([MASK]=give \mid \pi(p). p \text{ would } [MASK] \text{ help}) > P([MASK]=need \mid \pi(p). p \text{ would } [MASK] \text{ help})$$

We interpret BERT’s output as indicating the semantic roles “*giving_help*” and “*needing_help*”.

Below we can see the contrast between the input to BERT as part of our Knowledge Based strategy in contrast to a basic WSC resolution only relying on BERT:

- “She was inexperienced with the disorder. she would [MASK] help.” (Our usage of BERT to derive a semantic role.)
- “Maria helped Elena cope with the newly diagnosed autism because [MASK] was inexperienced with the disorder.” (Full WSC resolution using BERT.)

5.5. Reasoning Using Semantic Roles

This is the last phase of our system. For each schema that was matched to a pattern, the semantic roles derived in sections 5.2 and 5.4 are used as inputs. With these inputs, some background knowledge rules are needed to derive the referent of the pronoun. The background knowledge rules we use are:

- **IF** a person x *{helps / is asked by}* a person y because the pronoun p is *giving help* **THEN** p refers to x .
- **IF** a person x *{is helped by / asks}* a person y because the pronoun p is *needing help* **THEN** p refers to x .

The encoded forms of the background knowledge rules are given below as:

```
answer(X) :- is_candidate(X),
  1 {has_s( X, helps, _ ); has_s( _, asks, X )},
  pronoun(P), has_s(Verb, agent, P),
  has_s(_, caused_by, Verb),
  has_s( P, semantic_role, giving_help ).
```

```
answer(X) :- is_candidate(X),
  1 {has_s( _, helps, X ); has_s( X, asks, _ )},
  pronoun(P), has_s(Verb, agent, P),
  has_s(_, caused_by, Verb),
  has_s( P, semantic_role, needing_help ).
```

Using the domain-specific background knowledge rules and the previously derived semantic roles, we derive the answer for a WS. In our running example, statements expressing the implicit semantic role of the pronoun (“*needing_help*”) from section 5.4 and the semantic roles of the candidates from section 5.2 are added to the ASP program. Then, the condition defined in the second background knowledge rule is satisfied, and thus we can derive the answer as “Elena”.

Note that, although the implicit semantic role of the pronoun is extracted by an ML method, the reasoning used to resolve the schemas is based on interpretable rules. Hence, the rules used in resolving a schema provide an explanation of the answer.

6. Results

Table 1 shows the results of our method contrasted with two systems using BERT alone (BERT_LARGE [4] and BERT_WIKI_WSCR³ from [5]) on the 207 sentences including ‘help’ and the 457 sentences containing ‘ask’ that meet the patterns. Our method achieves accuracy of 81.64% and 75.93%, which is higher than the accuracy achieved by BERT by around 5% and 13% respectively. Moreover, for each answer from our method an explanation can be produced, in contrast to a mere quantification on the certainty of the choice as given by BERT.

In our method, we use BERT to match a pronoun with an appropriate semantic role. We checked the accuracy of BERT on this task for the sentences including ‘help’. BERT achieved 84.06%, which is higher than its accuracy in resolving Winograd Schemas directly by around 8%. By integrating KR reasoning we not only increased the overall performance of the framework, but also made better use of an existing language model’s ability.

Further strengthening our claim that BERT benefits from being given a small, focused task, we show that the accuracy for selecting the semantic role is affected by the exact prompt provided. In our main experiments we gave BERT a short prompt based only on the pronoun part of the

³This refers to BERT which has been further fine-tuned for the WSC.

	Test	Accuracy ('help')		Accuracy ('ask')	
T0.	BERT_LARGE	57.97%	(120/207)	51.86%	(237/457)
T1.	BERT_WIKI_WSCR	76.33%	(158/207)	63.02%	(288/457)
T2.	Our system	81.64%	(169/207)	75.93%	(347/457)

Table 1

Results on the sentences containing 'help' and 'ask' that meet the patterns

WS, and we later tested (for those containing 'help') the selection of semantic roles when a longer prompt containing the whole WS was given. The accuracy reduced from 84.06% to 63.29% when using the whole context rather than the pronoun part only. For example,

- “She was inexperienced with the disorder. She would [MASK] help.” (**Our usage of BERT to match a semantic role.**)
- “Maria helped Elena cope with the newly diagnosed autism because she was inexperienced with the disorder. She would [MASK] help.” (**Additional context given.**)

The finding also supports the view that the semantic role is often decisive in determining the reference of a pronoun rather than other aspects of the semantic content.

Note that the accuracy of our method as a whole (81.64%) for the schemas including 'help' is only 2.42% lower than BERT's accuracy in matching semantic roles. So, provided the pronoun role is correctly identified, our KR reasoning is very accurate. The decreased accuracy of our method compared to the semantic role prediction from BERT is mainly due to the fact that some schemas were incorrectly parsed by K-Parser.

7. Conclusion

Our new method KARaML improves on the work of [19]. In our prior work it was necessary to explicitly define a large number of rules in order to match the semantic roles for the candidates and pronouns. Here we have significantly reduced the number of rules required by instead using a language model to establish the correlation between the description of the pronoun and the semantic roles of the candidates. In addition, we improve on the performance achieved by BERT alone and we are able to generate an explanation for the chosen answer.

Our current implementation can only be applied to a subset of Winograd schemas for which domain-specific rules have been defined. For future work, if we include more domains and patterns, we will increase the coverage of our system. We would also like to apply our method to other language understanding problems such as COPA [29]. As some parsing results from K-Parser were incorrect, we intend to investigate using other parsers such as SENNA.

So far we have only made limited use of BERT to identify the likely semantic relationship of the pronoun to the candidate clause. However, the same method may be applied to identify other semantic relationships that could be exploited by a KR reasoner. Moreover, BERT could be replaced by other state-of-the-art language models such as GPT-3 [30].

More generally, our framework represents initial steps towards a progressive assimilation architecture for language understanding where we use ML to successively combine new infor-

mation into a KR representation that we have built up from prior information. This seems to provide a general way by which information expressed in natural language can be matched with predicates occurring in the formalised axioms of a KR system.

References

- [1] H. Levesque, E. Davis, L. Morgenstern, The Winograd Schema Challenge, in: The 13th Int. Conf. on Principles of Knowledge Representation and Reasoning, Italy, 2012.
- [2] K. Sakaguchi, R. L. Bras, C. Bhagavatula, Y. Choi, WinoGrande: An adversarial Winograd Schema Challenge at scale, in: AAAI-20, 2020.
- [3] A. Sharma, N. H. Vo, S. Aditya, C. Baral, Identifying various kinds of event mentions in K-Parser output, in: Procs. of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Assoc. for Comp. Linguistics, 2015, pp. 82–88.
- [4] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805[cs.CL] (2018).
- [5] V. Kocijan, A. M. Cretu, O. M. Camburu, Y. Yordanov, T. Lukasiewicz, A surprisingly robust trick for Winograd Schema Challenge, in: Procs. of the 57th Annual Meeting of the Assoc. for Comp. Linguistics, 2019, pp. 4837–4842.
- [6] J. R. Hobbs, Coherence and coreference, *Cognitive science* 3 (1979) 67–90.
- [7] N. Asher, A. Lascarides, *Logics of Conversation*, Cambridge University Press, 2003.
- [8] A. Kehler, L. Kertz, H. Rohde, J. L. Elman, Coherence and coreference revisited, *Journal of semantics* 25 (2008) 1–44.
- [9] B. Bennett, Semantic analysis of Winograd Schema no. 1, in: F. Neuhaus, B. Brodaric (Eds.), Procs. of the 12th Int. Conf. on Formal Ontology and Information Systems (FOIS 2021), *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2021.
- [10] P. Schüller, Tackling Winograd Schemas by formalizing relevance theory in knowledge graphs, in: Fourteenth Int. Conf. on the Principles of Knowledge Representation and Reasoning, 2014.
- [11] D. Bailey, A. Harrison, Y. Lierler, V. Lifschitz, J. Michael, The Winograd Schema Challenge and reasoning about correlation, in: *Logical Formalizations of Commonsense Reasoning*, AAAI Spring Symposium, Stanford University, USA, 2015.
- [12] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, *Journal of Machine Learning Research* 12 (2011) 2493–2537.
- [13] F. Kong, Y. Li, G. Zhou, Q. Zhu, P. Qian, Using semantic roles for coreference resolution, in: 2008 Int. Conf. on Advanced Language Processing and Web Information Technology, 2008, pp. 150–155.
- [14] C. J. Fillmore, Frame semantics, in: *Cognitive linguistics: Basic readings*, De Gruyter Mouton, 2008, pp. 373–400.
- [15] J. Bos, Wide-coverage semantic analysis with Boxer, in: Procs. of the 2008 Conf. on Semantics in Text Processing, STEP '08, Association for Computational Linguistics, USA, 2008, p. 277–286.

- [16] D. Das, D. Chen, A. F. Martins, N. Schneider, N. A. Smith, Frame-semantic parsing, *Computational linguistics* 40 (2014) 9–56.
- [17] A. Sharma, N. H. Vo, S. Aditya, C. Baral, Towards addressing the Winograd Schema Challenge - building and using a semantic parser and a knowledge hunting module, in: *IJCAI 2015*, 2015, pp. 1319–1325.
- [18] M. Gelfond, V. Lifschitz, The stable model semantics for logic programming, in: *Procs. of Int. Logic Programming Conf. and Symposium*, 1988, pp. 1070–1080.
- [19] S. J. Hong, B. Bennett, Tackling domain-specific Winograd Schemas with knowledge-based reasoning and machine learning, in: *3rd Conf. on Language, Data and Knowledge (LDK 2021)*, 2021.
- [20] A. Sharma, Using Answer Set Programming for commonsense reasoning in the Winograd Schema Challenge, *arXiv:1907.11112[cs.AI]* (2019).
- [21] M. Al-yahya, L. Aldhubayi, S. Al Malak, A pattern-based approach to semantic relation extraction using a seed ontology, in: *Procs. - 2014 IEEE Int. Conf. on Semantic Computing, ICSC 2014*, 2014, pp. 96–99.
- [22] A. Rahman, V. Ng, Resolving complex cases of definite pronouns: The Winograd Schema Challenge, in: *EMNLP-CoNLL*, 2012.
- [23] P. Trichelair, A. Emami, A. Trischler, K. Suleman, J. 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).
- [24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, *arXiv:1907.11692* (2019).
- [25] A. Emami, K. Suleman, A. Trischler, J. C. K. Cheung, An analysis of dataset overlap on Winograd-style tasks, in: *Procs. of the 28th Int. Conf. on Computational Linguistics, Int. Committee on Computational Linguistics, Barcelona, Spain (Online)*, 2020, pp. 5855–5865.
- [26] A. Ettinger, What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models, *Transactions of the Association for Computational Linguistics* 8 (2020) 34–48.
- [27] D. Kahneman, *Thinking, fast and slow*, Penguin, London, 2012.
- [28] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, *Lingvisticae Investigationes* 30 (2007) 3–26.
- [29] M. Roemmele, C. A. Bejan, A. S. Gordon, Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning, in: *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University, 2011.
- [30] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, *arxiv:2005.14165[cs.CL]* (2020).