The Role of Pragmatics in Solving the Winograd Schema Challenge

Adam Richard-Bollans and **Lucía Gómez Álvarez** and **Anthony G. Cohn**
School of Computing
University of Leeds, Leeds, UK
{mm15alrb, sc14lga, a.g.cohn}@leeds.ac.uk

Abstract

Different aspects and approaches to commonsense reasoning have been investigated in order to provide solutions for the *Winograd Schema Challenge* (WSC). The vast complexities of natural language processing (parsing, assigning word sense, integrating context, pragmatics and world-knowledge, ...) give broad appeal to systems based on statistical analysis of corpora. However, solutions based purely on learning from corpora are not currently able to capture the semantics underlying the WSC – which was intended to provide problems whose solution requires knowledge and reasoning, rather than statistical analysis of superficial lexical features. In this paper we consider the WSC as a means for highlighting challenges in the field of commonsense reasoning more generally. We begin by discussing issues with current approaches to the WSC. Following this we outline some key challenges faced, in particular highlighting the importance of dealing with pragmatics. We then argue for an alternative approach which favours the use of knowledge bases where the deep semantics of the different interpretations of commonsense terms are formalised. Furthermore, we suggest using heuristic approaches based on pragmatics to determine appropriate configurations of both reasonable interpretations of terms and necessary assumptions about the world.

Introduction

The Winograd Schema Challenge (Levesque, Davis, and Morgenstern 2012) was conceived as a new benchmark in artificial intelligence, which would improve on the Turing Test (Turing 1950) by removing the need for deception and focusing more on understanding. The task is a particular type of pronoun disambiguation problem. Sentences with a pronoun and two candidate referents are given, and the task is to find the correct referent of the pronoun. As the challenge is intended to require genuine intelligence and understanding, the sentences are supposed to be constructed in such a way that syntactic constraints and semantic preference do not alone enable the disambiguation. This construction is achieved in part by finding pairs of sentences, differing only by one word but where the pronoun reference is different. For example:

> *The large ball crashed right through the table because* **it** *was made of [steel/styrofoam]. What was made of [steel/styrofoam]? Answers: The ball/the table.*[1] (1)

The pronoun 'it' refers to either the ball or the table depending on whether 'steel' or 'styrofoam' is used. In both cases the syntactic structure remains the same and, supposing that clear semantic preferences relating 'steel' and 'crashing through things' or 'styrofoam' and 'being crashed through' cannot be easily learned from mining a large corpus, it is hoped that any system which resolves the pronoun must use some sort of genuine understanding.

In the literature discussing the WSC and its motivation as a benchmark we see example reasoning processes incorporating detailed semantics of the language involved (Davis 2013; Levesque 2014; Levesque, Davis, and Morgenstern 2012; Morgenstern and Ortiz Jr 2015). This kind of approach however has not been at the forefront of proposals to the challenge. This is in large part due to the enormous complexity of dealing with natural language and constructing large enough knowledge bases to handle such varied contexts.

In order to further the symbolic approach we investigate some problems faced, mainly pragmatics. It is hoped that this sort of analysis helps to shed light on what kind of reasoning is needed where; and that heuristic methods will remove a large portion of the burden of reasoning about natural language. Along similar lines, a partial solution is provided in (Schüller 2014), using relevance theory (Sperber and Wilson 2004) to motivate selection of the best knowledge graph to describe a sentence.

In this paper we first explore what kind of reasoning capabilities we expect a system to display when solving the WSC and we analyse how some of the proposed approaches compare to this. We then consider some key challenges for solving the WSC using reasoning we consider appropriate; in particular, that pragmatics and context are very difficult to capture and semantics are hard to formalize due to vagueness. Finally, we show how pragmatic considerations can help in solving the WSC, specifically we consider how prototype theory and heuristic methods can be used to support symbolic approaches.

What kind of reasoning are we looking for?

We first consider the example above (1) given in (Levesque 2014), using the word 'styrofoam'. Humans would successfully resolve this by knowing particular properties of sty-

[1] Taken from www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection

rofoam, maybe some naive physics and even some general properties of balls and tables.

Levesque then considers what should be the outcome if we change styrofoam to XYZZY, where XYZZY is some material that we are given some facts about, one of the facts being 'It is ninety-eight percent air, making it lightweight and buoyant'. Given this fact, humans would be able to reason that the table is made of XYZZY. This is a part of intelligent behaviour that we would like to replicate, and is clearly dependent on having and being able to reason about detailed knowledge. Further, it has been suggested as a possible extension to the test to add a requirement for the solution to provide a simple explanation of its choice (Morgenstern and Ortiz Jr 2015). This need for explanation would also seem to depend on reasoning with detailed knowledge; in order to explain why the table is made of styrofoam, it seems necessary to have an understanding of the mechanics of the situation. The ability to provide an explanation is also important more generally for the field of commonsense reasoning, for example for decision support systems that need to provide justifications for decisions (Hayes-Roth, Waterman, and Lenat 1984).

Versatile solutions

The WSC was conceived as a new benchmark for artificial intelligence; as such, we hope that solutions to the WSC will provide tools for tackling a broader range of question answering tasks and commonsense challenges. In this way, solutions to the challenge should display *versatility* as well as making advances in the WSC specifically, thus representing genuine progress towards truly intelligent machines. Solutions which are over-specific to the WSC and only provide insight into this narrow set of coreference resolution problems are not likely to be 'engaging in behaviour that we would say shows thinking in people' (Levesque, Davis, and Morgenstern 2012). This is a similar but more general requirement than *elaboration tolerance* (McCarthy 1998).

The situations described in Winograd sentences (WS) are generally common/normal occurrences; however, it is desirable for AI systems to be able to reason about out-of-place objects and strange scenarios. The ability to do this displays a genuine understanding of what is going on. Levesque gives the example 'Can a crocodile run a steeplechase?' (Levesque 2014). Most humans would answer this easily using basic knowledge about crocodiles (in particular that they cannot jump) and what is necessary to be able to complete a steeplechase. Of course, as noted by Levesque, a statistical approach using the closed world assumption would be likely to get the right answer to this question too as there is little evidence of crocodiles running steeplechases. It would be less likely however to answer the question correctly if the animal was a gazelle (which presumably could run a steeplechase).

Having briefly considered the kind of solutions we are aiming for, we now look at how some existing approaches compare to this.

Existing approaches to the challenge

Since the inception of the WSC there has been some theoretical discussion on the purpose of the challenge (Davis and Marcus 2015; Levesque 2014; 2017), various methods suggested for tackling the problem (Sharma et al. 2015; Schüller 2014; Bailey et al. 2015; Rahman and Ng 2012; Peng, Khashabi, and Roth 2015), and four implementations entered into the 2016 challenge[2] (Liu et al. 2016; Isaak and Michael 2016) (two of the competitors did not release papers). The WSC is a particular type of anaphora resolution task, on which there has been much work done in the natural language processing community already (Ng 2017; Mitkov 2014; Carbonell and Brown 1988); however due to the nature of the task, necessitating the use of world knowledge, the methods employed are not wholly suitable for the challenge.

Formalizing the necessary aspects of reasoning to tackle the WSC (spatial, temporal, causal, epistemic, ...) and integrating them in one system is notoriously hard. Therefore, it is not surprising that the space of genuine proposed solutions is sparse, and that existing approaches are mostly based on statistical methods, that circumvent the need for a precise understanding of the semantics of the schemas by learning likely answers from text corpora.

In this section we analyse some of the solutions proposed along these lines. We consider both their performance and success on the challenge and also their achievements and relevance for broader commonsense reasoning, which is the ultimate aim of the WSC as a benchmark.

Machine learning approaches

Machine learning methods for anaphora resolution have been used extensively over the past two decades (Ng 2017). In this section we consider some of the best known such approaches for tackling the WSC.

The team that came first in the 2016 WSC challenge[2] used 'Commonsense Knowledge Enhanced Embeddings' (Liu et al. 2016) which works by learning word representation vectors from large text corpora while incorporating commonsense knowledge as constraints in the training process. For the competition the commonsense knowledge was obtained from CauseCom — a set of cause and effect pairs such as 'winning causes happiness' (Liu et al. 2016) — though the team has also incorporated WordNet (Miller 1995) and ConceptNet (Speer and Havasi 2012). A neural network is then trained to answer yes or no when given candidate/pronouns pairs (as vectors), and this network is then used to answer new disambiguation problems.

Though achieving a good performance on the challenge, it would be down to chance whether it correctly answers the XYZZY problem given by Levesque, whether it could be used to solve the crocodile-steeplechase problem, or in future how it could be developed to explain how it comes to the conclusion.

Rahman and Ng (2012) have worked combining multiple methods to resolve the pronoun for a large corpus of WSs. This work achieved high results on their corpus, 73.1%. However, the corpus selection has been criticized for containing redundancy (Sharma 2014). Further, the approach relies

[2] www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html

heavily on statistical methods for assessing the semantic preferences of types and events e.g. a lion is a type of predator and being the subject of a kill event makes one more likely to be the object of an arrest event. It is clear that 'lions eat zebras because they are predators' is not a 'Google-proof' WS and should be discarded. When such type distinctions are not useful, the system may rely on FrameNet (Baker, Fillmore, and Lowe 1998); in the case of 'John killed Jim so he was arrested', FrameNet gives John the role of 'killer' and Jim the role of 'victim' and the system, using statistical methods, concludes that it is more likely for a 'killer' (John) to be arrested. In this case the system resolves the pronoun successfully. However, this takes no account of the importance of the connective: changing the sentence to 'John killed Jim *after* he was arrested' should force one to re-evaluate the disambiguation.

Work by Peng et al. (2015) has been successful, achieving higher results (76.4%) than Rahman and Ng on the same corpus. The technique is similar to the FrameNet approach of (Rahman and Ng 2012) but they also take connectives into account. This approach can give crude, and clearly problematic, forms of knowledge such as '{flower has pollen} is more likely than {bee has pollen}'; to more reasonable knowledge such as 'the subject of "*be afraid of*" is more likely than the object of "*be afraid of*" to be the subject of "*get scared of*"'. Though these sorts of techniques will likely prove very useful for natural language processing, and may even manage to pass the WSC, there is a fundamental issue that these techniques are learning about the likelihood of combinations of words in corpora and there appears to be little in the way of transferable knowledge or understanding. For example, it is clear that the kind of background knowledge necessary to solve the crocodile-steeplechase problem is not present.

Rather than applying reasoning to knowledge, these techniques are geared towards mining what we may call *commonsense rules*. We discuss the nature of such rules in the following section.

Commonsense rules

It is clear that, in the WSC, it appears possible to resolve pronoun ambiguity through an appeal to normality — heavy things cannot be lifted, younger people are fitter, useless objects go in the bin while useful tools are kept in storage etc... Hence, a large part of the suggested approaches to the WSC have been about ways of finding and/or incorporating such 'commonsense rules'. We believe, however, that this is a rather crude view of commonsense reasoning and outline some problems of these approaches below.

One proposed approach is that we reduce some of the implied causation in WSs to correlation (Bailey et al. 2015). This uses 'correlation formulas' of the form $F \oplus G$, such as '*fit_into*$(x, y) \oplus$ *large*(y)' to say that 'stuff fitting into y' is correlated with 'y being large'. Some inference rules are given governing such correlation formulas and it is shown how these could be used to justify a solution to a WS. This approach is however problematic. It is analogous to a discussion in (Bunt and Black 2000) — by reducing to mere convention the reason why 'There is a howling gale in here!' is understood as a command to close the window, we are oversimplifying and missing out more important reasoning processes, including context. Similarly, if we were to find a list of commonsense correlations like '*fit_into*$(x, y) \oplus$ *large*(y)' through corpus mining, we are ripping the words out of context and may be missing out important reasoning processes.

This is not to say that conventions do not exist or form an important part of commonsense reasoning. Natural language is full of conventions that we may rely upon to communicate. For example, considering the sentence 'Sam chopped down the tree' there is a default assumption that the chopping is done with an axe. This kind of convention can be considered as part of linguistic knowledge (Pustejovsky 1991). However, reasoning based solely on conventions may be too crude, as it does not take contextual factors into consideration. Say that we know that Sam is holding a sword, then we may reject the default assumption that Sam chops down the tree with an axe. One way of dealing with the context dependency of such conventions may be to apply context frames, as in (McCarthy 1993), i.e. in the context of Sam holding a sword, the statement 'Sam chopped down the tree' suggests that Sam did the chopping with a sword rather than an axe. However, even if we can create appropriate context frames using salient aspects of context, it seems that the process of creating convention/context pairs would continue ad infinitum. We would hope that reasoning removes the necessity for a lot of these rules e.g. when someone is holding an appropriate tool, T, for performing action, A, and we are told that they performed action A, then we can assume that they have used T to do A.

The tactic for many approaches is to begin by learning commonsense knowledge from large text corpora or by integrating natural language knowledge bases. Part of the appeal of this is that knowledge can be exploited without having to translate between formal and natural language. However, the methods for extracting commonsense knowledge from the Web can be problematic. Language is used in an efficient way and commonsense knowledge is often left implicit (Schüller and Kazmi 2015).

Even if we were able to overcome some of the problems of mining commonsense, do we want to use reasoning that relies solely on these correlations and rules? Though they may be helpful for certain applications, the reasoning mechanisms need to incorporate less crude knowledge. Regarding the desire for versatility and considering some of the problems listed on the Common Sense Problem Page[3], it is clear that this approach is over-specific to the WSC. It would also clearly be hard to mine relations between crocodiles and steeplechases in this way! Moreover, any explanation of the disambiguation given by such a system would not be very enlightening. Considering schema (1) with 'steel'; explaining why 'it' refers to the ball by saying that 'steel things are more likely to crash through things than to be crashed through' is not a reasonable explanation. Even the ability to cite a salient property of steel like 'steel is hard' would be an important improvement.

The approaches outlined above at best only incorporate shallow semantic features and do not appear to exhibit the

[3] www-formal.stanford.edu/leora/commonsense/

kind of intelligent behaviour the challenge was designed to test. We believe that, in order to carry out complex inferences and really understand the world, some definitions of the natural language in terms of more refined primitives is often necessary. It is necessary to have genuine world knowledge of entities, as well as their physical, social/historical and functional attributes, as in (Bennett 2005), and be able to reason about that knowledge, e.g. crocodiles have short legs and long bodies, making them unsuitable candidates for a steeplechase, rather than superficial knowledge about relations between entities which are mined from corpora, e.g. crocodiles do not run steeplechases. A line may be drawn by the distinction between reasoning from first principles and reasoning by analogy. They can both be valid forms of reasoning, but reasoning by analogy alone is not enough to be considered intelligent.

Key challenges

This section outlines some particular problems that need resolving in order to tackle the WSC and for commonsense reasoning systems more generally.

Pragmatics

A large part of the complexity of the WSC comes from pragmatic considerations. There are varying positions on the definition of pragmatics (Carston 1999), however it is generally understood as the field concerned with extra-linguistic factors, such as context, and how they allow the understanding of a speaker's intended meaning.

Semantic considerations are clearly essential but they are generally not enough in order to reach a conclusion about the disambiguation for a WS. This is an example of *semantic underdeterminacy* — that from only considering the literal meanings of terms in a sentence and not accounting for the intended meaning, we do not obtain a truth-evaluable proposition. For example, the sentence 'Tom threw his school bag down to Ray after he reached the top of the stairs' does not contain much information if we only consider the semantics. We also need to consider the intention of the speaker and we may infer this from the decisions the speaker takes regarding the specific choice of language, what information is omitted, what is left ambiguous, the phrasing of the sentence etc... Indeed, Kempson argues that 'the articulation of semantics [does not alone] provide the full propositional content/logical form/truth conditions expressed by a sentence'(Kempson 1984).

To evidence this view, we can see that for many WSs wrongly disambiguating the pronoun does not necessarily violate world knowledge. For example, when dealing with the sentence:

The trophy does not fit into the suitcase because it$_x$ is too large[1] (2)

there are various interpretations of 'large' which give no definite disambiguation. If we imagine a trophy and suitcase to be vase-shaped, with a wide base, narrow stem and wide top, and that the trophy fits into the suitcase, it is possible that making the suitcase larger via a scale projection would make the trophy no longer fit. It is in part by making pragmatic considerations that we can assign appropriate interpretations to these terms and thus disambiguate the pronoun.

Moreover, even in the sentences where each term can be precisely and appropriately defined we can still have semantic underdeterminacy. Is it often the case that an utterance is not totally explicit and leaves the reader to fill in the gaps with available assumptions and inferences (Carston 1999). One of the ways that a hearer may fill in these gaps and infer a speaker's intention is by assuming Grice's Maxims for co-operative communication (Grice 1975); e.g. the 'Quantity Maxim', stating: 'Make your contribution as informative as is required' and 'Do not make your contribution more informative than is required'. So for instance, if a speaker goes into a lot of detail when making an utterance, we may assume that there is particular reason for this and can infer things based on this knowledge. This kind of pragmatic inference is also important for written text, and hence the WSC. Therefore, as it stands, any solution to the WSC needs some mechanisms for coping with this implicit knowledge.

In the next section we consider some particular examples of this sort of inference when addressing a WS.

Assumptions about the world

When facing any WS there are multiple commonsense principles that apply which allow us to create an accurate model of the situation. What we aim to achieve is some guidance on how to choose these principles and when they apply. To this end we examine the following WS:

Tom threw his school bag down to Ray after he$_x$ reached the [top/bottom] of the stairs. Who reached the [top/bottom] of the stairs? Answer: top: Tom. bottom: Ray.[1] (3)

We will use this example to help elucidate some of the complexities faced, including the initial position of objects and relevant objects.

The main idea of this sentence is that to throw something down to someone, that person must be below you. We then use the idea of what it means to be at the top of something, i.e. that if Ray is at the top of the stairs then he cannot be below Tom. This is however not as clear as it seems.

Initial position It is possible that Tom is on some balcony above the stairs and waits for Ray to reach the top of the stairs before throwing the bag down to Ray. So why do we like the answer 'Tom'? It appears we assume that Tom and Ray are initially in a similar location, or to be more precise, that they both have the same relation to any given landmark — in this case the stairs. Character x reaching the top of the stairs implies that x has moved upwards. Not given any information on the other character, y, we assume they have not moved and so x is likely to be above y.

Alternatively, x may have been walking along a corridor to reach the top of the stairs. In this scenario we have two locations to consider, the corridor and the stairs. We suppose that Tom and Ray are on the stairs or in the corridor. In this case it would make no sense for Ray to be at the top of the stairs, as then Tom would not be able to throw anything down to him (from the corridor or the stairs); so we suppose that it must be Tom who walks along the corridor to reach the top of the stairs and throw the bag down to Tom.

We appeal to a rule that in some narrative, unless we have reason to infer otherwise, characters are nearby/in the same place. This idea can be explained by Grice's quantity maxim i.e. there is no pertinent difference in the positions of either Tom or Ray; if there were then the quantity maxim says it should be made known.

This rule however does not always hold. Imagine we replace 'stairs' with 'swimming pool':

> *Tom threw his school bag down to Ray after he$_x$ reached the top of the swimming pool. Who reached the top of the swimming pool? Answer: Ray.*

In this scenario x reaches the top of the swimming pool, breaking the surface of the water. x is then not in a position to throw something like a school bag downwards, as it is pretty hard to throw textile objects through water. Hence, we imagine that x is not Tom, but Ray, and that Tom must be stood somewhere above the swimming pool.

Relevant objects In general in the WSC to come to a conclusion we only need to reason about entities that are explicitly mentioned. In the school bag example we reason about the two characters in the narrative, Tom and Ray, the staircase and the school bag itself. Combining knowledge of actions like 'throwing' 'reaching the top of' etc.. with knowledge of these objects. In general then, we do not need to appeal to the existence of extra entities in order to come to a conclusion. This can also be explained by the quantity maxim, the sentence should provide the necessary objects for the reader to make sense of the sentence.

However, as previously discussed, certain words or phrasings indirectly suggest the existence of certain entities, as in the 'Sam chopped down the tree' example. We can in part account for these entities by encoding into a lexicon (Pustejovsky 1991), though these are conventions that will not always hold. Therefore a defeasible reasoning process is necessary to select the most appropriate interpretation.

To conclude our discussion about assumptions about the world, we see that appropriate assumptions need to be made in order to reach the right conclusion. Further, we believe that, to varying extents, these kinds of considerations arise when analysing most WSs appearing in the collection maintained by Davis[1]. However, the assumptions are dependent on the specific situation and we need to discern somehow when the assumptions are appropriate. Deciding when to accept these assumptions should include pragmatic considerations. For example, it is lexical and semantic knowledge that *suggest* the existence of an axe in the sentence 'Sam chopped down the tree', however it is a pragmatic task to actually infer this. This motivates a heuristic process which incorporates pragmatics and gives preference to default assumptions, we will discuss this idea later.

Formalizing commonsense knowledge: level of detail and vagueness

An important issue is to recognize the level of semantics that one believes is appropriate for a solution to the WSC. Our discussion so far has motivated a detailed level of knowledge. Further, there is evidence that, even for coreference problems that would be considered easy with respect to the WSC, incorporating shallow semantic features is not enough (Durrett and Klein 2013). Yet, if we are to solve the WSC using deeper semantics, it is clear that the necessary commonsense knowledge would involve the formalization of a notoriously extensive knowledge base. How to obtain and organize such a large knowledge base is unclear.

On the one hand, due to the variety and scope necessary, mining commonsense knowledge is appealing; however, as previously discussed, the available methods and nature of text corpora pose limitations to obtaining deep knowledge, which is complex and commonly not explicit. On the other hand, hand crafted knowledge bases such as CYC (Lenat 1995), which incorporate a deeper level of knowledge, have had limited success and it is not clear how they should be exploited.

Beyond the problem of its acquisition, it is well known that commonsense knowledge is hard to formalize, particularly if the required level of detail involves the semantics of natural terms to be preserved. Vagueness and ambiguity are inherent to natural language and, for that reason, it is problematic to prescribe single strict interpretations to natural terms. To illustrate this, consider the WS (3) and imagine the case of a naive definition of a relation $\textit{at\_the\_top\_of}(x, y) \equiv x$ is on y and for any z which is part of y, x is not below z. We see that this fails for multiple reasons.

1. If Tom were one step below the very last one, it could still be considered that he is at the top of the stairs, particularly if Ray were well below him. We call it *sorites* vagueness when there is a the lack of a clear threshold of application of a term.
2. If we change 'stairs' to 'building' we might say that Tom is at the top of a building because he is on the top floor, rather than on the roof. In that case we are shifting the interpretation of the predicate to something like $\textit{at\_the\_top\_of}(x, y) \equiv z$ is the *top_part* of y and x is on z. There may also be many admissible interpretations of what it means for z to be the *top_part* of y. We call the multiplicity of conceptually distinct interpretations of natural terms *conceptual* vagueness. Further discussion on the multiple interpretations of natural language terms and their role in knowledge bases and ontologies can be found in (Bennett 2005).

Much of the work done in acquiring commonsense knowledge circumvents vagueness in different ways, such as using shallow semantics or microtheories that do not need to be consistent with one another. Various theories, however, have been proposed for dealing with vagueness. Fuzzy logic (Zadeh 1965) stands as an intuitive solution for modelling sorites vagueness by assigning *degrees* of truth. More interesting for this research, supervaluation semantics (Fine 1975) is based on the idea that vague language can be interpreted in many different precise ways, each of which can be logically conceptualised in a precisification (Bennett 2001; Gómez Álvarez and Bennett 2017), thus also offering support for modelling conceptual vagueness.

So where do all these considerations lead us? In order to reach the kind of solution we desire, we must be able to deal with semantic underdeterminacy — part of which involves deciding when to use appropriate commonsense assumptions — and also make use of a vast amount of detailed knowledge while dealing with the associated problems of vagueness.

With these issues in mind, we now consider some avenues for further work.

The role of pragmatics in solving the WSC

In the previous sections we have highlighted how current approaches, regardless of their success in solving schemas, have provided limited support for the kind of intelligent behaviour that we would like to replicate. Here, in an attempt to account for some of the key challenges, we propose an alternative approach, favouring the use of knowledge bases where the deep semantics of the different interpretations of commonsense terms are formalised. Furthermore, we suggest using heuristic approaches based on pragmatics to determine, in the context of each particular schema, appropriate configurations of both reasonable interpretations of the terms and necessary assumptions about the world.

For this purpose we first motivate the use of prototypes for categories *and* relations and then develop how heuristic methods can provide a manageable way of using pragmatic knowledge for the disambiguation of WSs.

Appealing to prototypicality

There is various work in pragmatics and cognitive science highlighting the importance of using prototypes: in utterance interpretation defaults are assigned before contextual and pragmatic considerations are taken into account (Levinson 1995; Recanati 2004) and there is also evidence for the human preference for good examples (prototypes) of some category as opposed to boundary cases and, further, that prototypes are associated with the least processing effort (Rosch 1978). In the particular scenario of a WS, we argue that the way vague terms are presented leads the reader to interpret them considering prototypical instances fitting the described scenario. For instance, when one reasons about the WS (2) involving the trophy and the suitcase, it is not necessary to worry about a precise semantic commitment for the notion of larger, but instead to evaluate the sentence considering clear cases that satisfy most of the possible interpretations.

Some of the previously discussed approaches work along similar lines, using general commonsense rules and a notion of correlation which appeal to a sense of typicality. However, we believe that this should be more nuanced and that the deep semantics of different interpretations should be preserved. Hence, we propose an approach using ideas from prototype theory (Rosch and Mervis 1975) to differentiate prototypical instances of vague terms and relations from borderline cases within a supervaluationist approach.

Much work has been done on how to pinpoint prototypical members of categories, mainly using vector analysis or conceptual spaces to find the centroid of a concept (Verheyen, Ameel, and Storms 2007; Lenci 2011). However, it is not clear how one could reason with this to resolve a WS, and further, we are not only interested in picking a prototypical example from a category, say from the class 'pet' or 'things that we eat'. Instead, we would also like to find prototypical instances of relations that can be used to compare an infinite number of objects. Although there is some work done on vector analysis for relationships between words (Mikolov, Yih, and Zweig 2013), in particular for analogy problems, it does not appear to be applicable to this sort of reasoning problem.

Suppose we have a vague term, like 'smaller'. How can we decide on prototypical instances of this relation? Adopting the supervaluation approach we would have a collection of precise interpretations of its meaning. Following motivation from (Rosch and Mervis 1975) — considering shared properties of classes — in an ideal scenario prototypical instances of 'smaller' share properties across all instances of 'smaller' i.e. a prototypical instance of smaller is considered smaller in all plausible interpretations. Consider the definitions for 'smaller' given in (Davis 2013):

1. $\mathit{Smaller}(a, b) \equiv \mathit{VolumeOf}(a) < \mathit{VolumeOf}(b)$
2. $\mathit{Smaller}(a, b) \equiv \mathit{DiameterOf}(a) < \mathit{DiameterOf}(b)$
3. $\mathit{Smaller}(a, b) \equiv a \subset b$
4. $\mathit{Smaller}(a, b) \equiv \exists s(s > 1 \wedge b = \mathit{Scale}(a, s))$

In this scenario, there are certainly pairs of objects that fall into all four categories (e.g. a sphere of radius 1 is smaller than a sphere of radius 2 in all the above senses). Hence, it would be appropriate to take the conjunction of all four definitions as a requirement for an instance to be considered a prototypical case of 'smaller'. However, in certain scenarios it may be inappropriate to take the conjunction in this way, as some definitions may be conflicting. In this case different metrics can be proposed for selecting prototypes that satisfy most of the interpretations.

Finally, our main claim in this section is twofold. On the one hand, we consider that an understanding of typicality is necessary for commonsense reasoning — that by default we should consider prototypes. On the other hand, a process which can only reason over prototypical definitions is clearly flawed in many respects as it creates over-simplification. Humans often use context to help narrow definitions, for example defining 'smaller' in a particular way makes sense when talking about 'fitting in'. Hence we believe that a good approach should reflect the diversity of possible interpretations of vague terms and that an engine based on pragmatics should guide the selection of appropriate alternatives when the prototype is not suitable.

Heuristics standing in for pragmatics

In this paper we have discussed some approaches proposed for the WSC relying on heuristic methods in different ways (Rahman and Ng 2012; Peng, Khashabi, and Roth 2015; Liu et al. 2016). Overall, we concluded that heuristics do not provide satisfactory solutions when reduced to evaluating shallow semantic notions such as correlation.

Instead, as has been argued, we believe that a good solution to the WSC should disambiguate the pronoun by considering the most plausible configuration of the scenario described, and the process of finding it should incorporate rich syntactic, semantic *and* pragmatic considerations. However, although advocating deeper semantics and symbolic based approaches that allow for *the kind of reasoning that we want* (see section above), we propose that heuristic methods have a key role in the WS resolution: that of simplifying the space of possibilities and estimating reasonably good configurations of precisifications and necessary assumptions about the world.

As we have highlighted above in order to carry out satisfactory reasoning we believe a system should give preference to both commonsense assumptions about the world as well as prototypical interpretations of the terms involved. These however should only be preferences rather than concrete rules. When to accept or reject these default assumptions requires knowledge and pragmatic understanding. The ability for this complex mix of pragmatics and world knowledge to contradict itself means that possible solutions or configurations of a described scenario are not unique. For example, when discussing the issue of throwing a school bag in a swimming pool above, the implausibility of throwing a school bag through water outweighed the assumption of Tom and Ray being in the same place. However, we may also consider that the assumption of characters being in the same place outweighs the usual interpretation of 'throw down' and 'top': supposing Tom and Ray are both stood in the swimming pool, we may interpret 'throw down' as 'throw horizontally away from the end of the swimming pool' and 'top of the swimming pool' to denote the end of the swimming pool. The result would then be to disambiguate the pronoun as 'Tom' rather than 'Ray'. This second interpretation is not wrong, however when 'throw down' and 'top' are interpreted in their usual way there is a plausible inference that Tom and Ray are not both located in the swimming pool. This would then be an example of a 'conversational implicature' (Grice 1975) and explain why the writer of the sentence did not explicitly give Tom and Ray's initial locations. Hence in the first interpretation we have a good explanation for violating the default that Tom and Ray are located in the same place and we also interpret all the terms in a usual fashion, therefore making this interpretation appear to be the valid one.

Being able to leverage these kinds of inferences is an important and difficult task in commonsense reasoning. Along these lines, one avenue (Schüller 2014) adopted in tackling the WSC has been to explore *relevance theory* (Sperber and Wilson 2004). This theory, inspired by Grice's work, is based on the idea that an utterance can have a variety of interpretations, and that it is through parsing, disambiguating terms, resolving pronouns and adding pragmatic inference as well as appropriate assumptions based on context that one can comprehend the meaning of an utterance. The principle guiding these tasks is the idea of maximizing *relevance*[4]. Schüller uses these ideas to motivate a heuristic process for reasoning over graphs, where a fitness function is employed to find relevant combinations that provide a disambiguation. Moreover, the resulting graph can be read off to get some idea of how it came to that disambiguation, potentially satisfying Morgenstern and Ortiz's requirement of a simple explanation. In spite of being preliminary research, in our view its reasonable results suggest that fruitful work can be done in further developing heuristic methods to assess the pragmatic and semantic considerations that govern reasonable disambiguations of natural language.

To conclude this section, it is our claim that this use of heuristics is much more in keeping with the nature of the WSC. That what should be simplified in order to keep the task manageable is not so much the deep semantics of natural terms, but the process of selecting and integrating relevant interpretations and background knowledge in the particular context of the resolution of each sentence.

Conclusion

In this paper we have discussed the nature of the WSC as a benchmark, highlighting the shortcomings of several current approaches and providing motivation for a more detailed level of knowledge. We have also analysed some of what we consider to be key challenges, in particular drawing attention to the need to take account of pragmatic considerations. To begin addressing these challenges, we have suggested using frameworks able to support the detailed semantics of natural terms while accounting for its vagueness. Moreover, that their complexity can be manageable with the use of prototypes, which should be identified and used by default, and, finally, that heuristic methods can be used to incorporate varying semantic interpretations as well as assumptions about the world, which maintain the pragmatic principles of cooperative communication.

In conclusion, it is our view that, while heuristic mechanisms are necessary to deal with natural language and to reduce the complexity of commonsense reasoning, they should not be used to over-simplify the semantics of natural terms. Instead, we believe that applications along the lines of theoretical studies in pragmatics can play a significant role in the selection of good interpretations of natural terms and to enrich the provided descriptions of the world with the appropriate implicit knowledge.

Acknowledgements

Thanks to Brandon Bennett for helpful discussion and to the anonymous reviewers for their useful feedback.

References

Bailey, D.; Harrison, A.; Lierler, Y.; Lifschitz, V.; and Michael, J. 2015. The Winograd Schema Challenge and Reasoning about Correlation. In *Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*.

Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The berkeley framenet project. In *Proceedings of COLING/ACL*, 86–90.

Bennett, B. 2001. What is a Forest? On the vagueness of certain geographic concepts. *Topoi* 20(2):189–201.

Bennett, B. 2005. Modes of concept definition and varieties of vagueness. *Applied Ontology* 1(1):17–26.

Bunt, H., and Black, W. 2000. The ABC of Computational Pragmatics. In Bunt, H., and Black, W., eds., *Natural Language Processing*, volume 1. Amsterdam: John Benjamins Publishing Company. 1–46.

[4] An input is said to be relevant if a worthwhile conclusion is drawn from it. An input is more relevant if it yields a greater positive cognitive effect for less processing effort.

Carbonell, J. G., and Brown, R. D. 1988. Anaphora resolution: a multi-strategy approach. In *Proceedings of the 12th Conference on Computational linguistics*, volume 1, 96–101.

Carston, R. 1999. The semantics/pragmatics distinction: A view from relevance theory. In Turner, K., ed., *The semantics/pragmatics interface from different points of view*. Oxford, UK: Elsevier. 85–125.

Davis, E., and Marcus, G. 2015. Commonsense reasoning and commonsense knowledge in Artificial Intelligence. *Communications of the ACM* 58(9):92–103.

Davis, E. 2013. Qualitative Spatial Reasoning in Interpreting Text and Narrative. *Spatial Cognition & Computation* 13(4):264–294.

Durrett, G., and Klein, D. 2013. Easy Victories and Uphill Battles in Coreference Resolution. In *EMNLP*, 1971–1982.

Fine, K. 1975. Vagueness, truth and logic. *Synthese* 30(3):265–300.

Grice, H. P. 1975. Logic and conversation. In *Syntax and Semantics, Vol. 3, Speech Acts*. New York: Academic Press. 41–58.

Gómez Álvarez, L., and Bennett, B. 2017. Classification, Individuation and Demarcation of Forests: formalising the multi-faceted semantics of geographic terms. In *13th International Conference on Spatial Information Theory*. Leibniz International Proceedings in Informatics.

Hayes-Roth, F.; Waterman, D.; and Lenat, D. 1984. *Building expert systems*. Reading, MA: Addison-Wesley.

Isaak, N., and Michael, L. 2016. Tackling the Winograd Schema Challenge Through Machine Logical Inferences. In Pearce, D., and Sofia Pinto, H., eds., *STAIRS*, volume 284 of *Frontiers in Artificial Intelligence and Applications*. IOS Press. 75–86.

Kempson, R. 1984. Pragmatics, anaphora and logical form. In Schriffin, D., ed., *Meaning, form and use in context: linguistic applications*. Washington, DC: Georgetown University Press. 1–10.

Lenat, D. B. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38(11):33–38.

Lenci, A. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proc 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 58–66. ACL.

Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The Winograd Schema Challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Levesque, H. J. 2014. On our best behaviour. *Artificial Intelligence* 212:27–35.

Levesque, H. J. 2017. *Common sense, the Turing test, and the quest for real AI*. Cambridge, MA: MIT Press.

Levinson, S. C. 1995. Three levels of meaning. In *Grammar and meaning: Essays in honour of Sir John Lyons*. Cambridge University Press. 90–115.

Liu, Q.; Jiangb, H.; Linga, Z.-H.; Zhuc, X.; Weid, S.; and Hua, Y. 2016. Commonsense Knowledge Enhanced Embeddings for Solving Pronoun Disambiguation Problems in Winograd Schema Challenge. *arXiv preprint arXiv:1611.04146*.

McCarthy, J. 1993. Notes on formalizing context. In *Proceedings of the 13th international joint conference on Artifical intelligence-Volume 1*, 555–560. Morgan Kaufmann Publishers Inc.

McCarthy, J. 1998. Elaboration tolerance. In *Common Sense*, volume 98.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *NAACL HLT*, volume 13, 746–751.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11):39–41.

Mitkov, R. 2014. *Anaphora resolution*. Routledge.

Morgenstern, L., and Ortiz Jr, C. L. 2015. The Winograd Schema Challenge: Evaluating Progress in Commonsense Reasoning. In *AAAI*, 4024–4026.

Ng, V. 2017. Machine Learning for Entity Coreference Resolution: A Retrospective Look at Two Decades of Research. In *AAAI*, 4877–4884.

Peng, H.; Khashabi, D.; and Roth, D. 2015. Solving hard coreference problems. In *Proceedings of NAACL*, 809–819.

Pustejovsky, J. 1991. The Generative Lexicon. *Computational linguistics* 17(4):409–441.

Rahman, A., and Ng, V. 2012. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*, 777–789. ACL.

Recanati, F. 2004. Pragmatics and Semantics. In *Handbook of Pragmatics*. Oxford: Blackwell. 442–462.

Rosch, E., and Mervis, C. B. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7(4):573 – 605.

Rosch, E. 1978. Principles of categorization. In Rosch, E., and Lloyd, B. B., eds., *Cognition and categorization*, volume 1. Hillsdale, NJ: Lawrence Erlbaum Associates. 27–78.

Schüller, P., and Kazmi, M. 2015. Using Semantic Web Resources for Solving Winograd Schemas: Sculptures, Shelves, Envy, and Success. In *SEMANTiCS (Posters & Demos)*, 22–25.

Schüller, P. 2014. Tackling Winograd Schemas by formalizing relevance theory in knowledge graphs. In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Sharma, A.; Vo, N. H.; Aditya, S.; and Baral, C. 2015. Towards Addressing the Winograd Schema Challenge-Building and Using a Semantic Parser and a Knowledge Hunting Module. In *IJCAI*, 1319–1325.

Sharma, A. 2014. *Solving Winograd schema challenge: Using semantic parsing, automatic knowledge acquisition and logical reasoning*. Ph.D. Dissertation, Arizona State University.

Speer, R., and Havasi, C. 2012. Representing General Relational Knowledge in ConceptNet 5. In *LREC*, 3679–3686.

Sperber, D., and Wilson, D. 2004. Relevance theory. In *Handbook of Pragmatics*. Oxford: Blackwell. 607–632.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460.

Verheyen, S.; Ameel, E.; and Storms, G. 2007. Determining the dimensionality in spatial representations of semantic concepts. *Behavior Research Methods* 39(3):427–438.

Zadeh, L. 1965. Fuzzy sets. *Information and Control* 8(3):338–353.