Reasoning about Conditional Beliefs for the Winograd Schema Challenge

Denis Golovin and **Jens Claßen**
Knowledge-Based Systems Group
RWTH Aachen University
Aachen, Germany

Christoph Schwering
School of Computer Science and Engineering
The University of New South Wales
Sydney, Australia

Abstract

The Winograd Schema Challenge has been proposed as an alternative to the Turing Test for measuring a machine's intelligence by letting it solve difficult pronoun resolution problems that cannot be tackled by statistical analysis alone. While many solutions so far are based on machine learning and natural language processing, we believe that a knowledge-based approach is better suited. In particular, we propose to employ a logic for conditional beliefs that is capable of dealing with incomplete or even inconsistent information (which commonsense knowledge often is). It does so by formalising the observation that humans often reason by picturing different contingencies of what the world could be like, and then choose to believe what is thought to be most plausible. We discuss and evaluate an implementation where relevant commonsense background information is obtained from the ConceptNet semantic network, translated into our formalism, and processed by a reasoner for our logic.

1 Introduction

The *Winograd Schema Challenge* (WSC) has been proposed by Levesque, Davis and Morgenstern [2012] as an alternative to the Turing Test for measuring a machine's intelligence. While agreeing with Turing's primary concern, which is focusing on the observable behaviour exhibited by the system, the aim is to avoid some of the pitfalls that come with a free-form conversation as Turing had in mind. Most importantly, in order to pass as a person, a machine will have to resort to deception to be able to answer questions about their height, childhood, and so on. Moreover, the judgement of what passes as human-like conversation is highly subjective and may vary depending on the interrogator that is performing the test.

The central idea behind the WSC is to instead let the machine answer a number of questions of a specific form. As an example, consider the statement

> *The delivery truck zoomed by the school bus because it was going so* <u>*slow*</u>*.*

The question is "*What was going slow?*", i.e., the task is to disambiguate to which of the two parties mentioned in the statement ("the delivery truck", "the school bus") a pronoun ("it") refers. Each schema contains a *special word* (here: "slow") that, when replaced by an *alternate word* ("fast"), changes the answer. While obvious to an English-speaking human reader, such questions pose a real challenge for a machine since the given statement alone does not provide enough information to derive an answer, but rather requires having appropriate commonsense background knowledge and being able to reason about it. Winograd schemas are designed to be "Google-proof", meaning that they should not be solvable by simple statistical analysis: If instead of a "delivery truck" the above sentence spoke of a "sports car", it would not be hard to detect a correlation to words associated with fast movement.

WSC competitions that were held so far prove that the task is indeed difficult, and likely cannot be solved by classical natural language processing alone. Robust commonsense reasoning on the other hand will undoubtedly be beneficial in other areas, in particular when it comes to domestic robots and human-machine interaction in general.

Such a system will not only need access to a large corpus of background information that preferably covers all areas of the wide spectrum of everyday human life, but also be able to appropriately function in light of the fact that such information is often incomplete or even conflicting. Humans are very capable of this, and we usually do so by picturing different contingencies of what the world could be like, and then choosing to believe what we think the most plausible scenario is.

In this paper, we present an approach to the WSC that is based on a logic that formalises these concepts, the logic of conditional beliefs [Schwering *et al.*, 2017]. We designed and implemented a system that

- translates the WSC sentence into a logical formula,
- extracts relevant background information from a knowledge corpus and translates it into our logical language, and
- performs reasoning to decide what the most plausible answer is.

The rest of this paper is organised as follows. In Section 2 we discuss related work. Section 3 shows an overview of our system's architecture. We then present the formal framework of our approach in Section 4. Section 5 describes the implementation of our system, while Section 6 presents an evaluation. Finally we conclude in Section 7.

2 Related Work

Rahman and Ng [2012] view the WSC as a special form of pronoun resolution problem and propose an approach based on machine learning. Their system relies on abstract feature vectors extracted from eight different sources, among which *Google* and *Lexical Features* turned out to be most useful in their analysis. For the former, they construct a number of queries from the verb governing the pronoun in the input sentence, the words followed by it, and the two candidate parties, and then count the number of results returned by the search engine. The latter are obtained from a training set and represent the presence or absence of certain predefined structures in the sentence. They achieve an accuracy of 73.1% on their test set, where the overall dataset of 941 sentence pairs was created by 30 students and split using a 70/30 training/test ratio. It should be noted that they considered a relaxed version of the problem: while the WSC only allows for instances where resolving the pronoun requires background knowledge *not* expressed in the statement, they did not impose this condition on sentences.

Sharma et al. [2015] present a three-step method based on their semantic parser "K-parser": first, the input sentence is brought into a graph representation encoding the conceptual classes of words as well as their dependencies and semantic relations. Next, a series of Google queries is constructed where nominal entities are replaced by wild cards and verbs are substituted by their synonyms. The sentences thus found are then matched against the input by comparing their semantic graphs by a set of predefined rules. For evaluation, they selected 71 sentences from the WSC corpus, resulting in 53 answers, 49 of which were correct.

Bailey et al. [2015] introduce a new formalism for reasoning about the correlation between sentences, which can either be positive or negative. Intuitively, the correlation between two sentences is positive when after hearing the first sentence, the hearer views the second sentence as more plausible than before. Negative correlation on the other hand lowers the plausibility. For example,

The delivery truck zoomed by the school bus.

is positively correlated with

The school bus was going so slow.

but negatively with

The truck was going so slow.

The authors of the aforementioned paper present the *correlation calculus* as a deductive system to formalise this idea. It extends classical predicate calculus by a new operator $\oplus$, where $F \oplus G$ expresses that F and G are positively correlated, and $F \oplus \neg G$ denotes a negative correlation. They show how their formalism can be used to derive such correlation formulas from a set of formulas encoding background information and thus justify the answers for several WSC instances. While their work makes an important step towards solving the WSC, they admit that it so far does not address the important problem of generating the needed commonsense background axioms (these were hand coded in their examples), nor did they come up with an implementation that would allow automating the reasoning process and evaluating the approach against others. Their formalism also currently lacks the capability to handle non-monotonic aspects such as the statement

If the shelf is level, then the sculpture will not roll off.

which, when read as default rule, allows to withdraw the conclusion that the sculpture does not roll off in face of new information, e.g. that there is an earthquake.

Finally, Liu et al. [2017] propose another approach based on machine learning, which is used (a) in a knowledge acquisition step to identify cause-effect relationships between commonly used words from large text corpora and (b) to find association relationships in the form of conditional probabilities between pairs of discrete events. They manually selected a subset of 70 Winograd schemas, among which their method achieves an accuracy of 70%.

3 System Architecture Overview

The architecture of our system is visualised in Figure 1. In a preprocessing step, tuples that represent the acting parties, their actions, and their relationships are extracted from the input Winograd schema sentences. For instance, the Winograd schema from Section 1 leads to tuples *(truck, zoomed by, bus)* and *(zoomed by, goes slow)*. Then a commonsense repository, ConceptNet, is employed to derive further information about the extracted words. These facts come in the form of binary relations; for instance, it may tell us that zooming by is a form of travelling rapidly. From these facts, a conditional knowledge base is constructed in the logic $\mathcal{BO}$, and from which the system aims to infer through automated reasoning how to disambiguate the pronoun in the most plausible way. In the school bus-scenario, the system compares the plausibility of party $^{\#}1$ (the truck) versus party $^{\#}2$ (the school bus) being the slow party.

4 A Logic of Plausible Beliefs

Reasoning in our framework is based on a logic of plausible beliefs [Schwering *et al.*, 2017] called $\mathcal{BO}$. This logic features non-monotonic conditionals of the form

If *some premise* holds,
then plausibly *some consequent* is true,

which enables us to represent and reason about more and less plausible contingencies. We expect this to be crucial in the context of the WSC, where we aim to find the correct answer by comparing the plausibility of the different candidates. In the following we briefly introduce the formal machinery of this logic.

The *terms* in our language consist of infinitely many first-order variables, usually denoted as x or y, and infinitely many *(standard) names* $n \in \mathcal{N} = \{^{\#}1, ^{\#}2, \ldots\}$. *Formulas* are of the form

- $P(t_1, \ldots, t_k)$, $(t_1 = t_2)$, $(\alpha \land \beta)$, $\neg\alpha$, $\forall x \alpha$,
- $\mathbf{B}(\phi_1 \Rightarrow \psi_1)$, $\mathbf{O}\{\phi_1 \Rightarrow \psi_1, \ldots, \phi_m \Rightarrow \psi_m\}$,

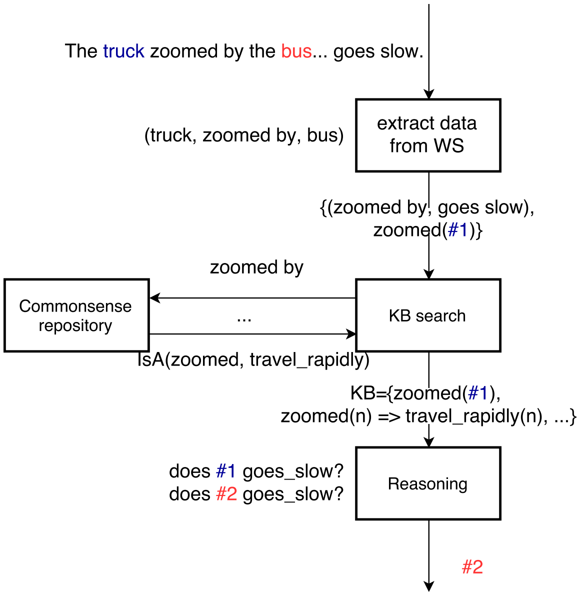
The truck zoomed by the bus... goes slow.
(truck, zoomed by, bus)
extract data from WS
{(zoomed by, goes slow), zoomed(#1)}
zoomed by
Commonsense repository
...
KB search
IsA(zoomed, travel_rapidly)
KB={zoomed(#1), zoomed(n) => travel_rapidly(n), ...}
does #1 goes_slow?
does #2 goes_slow?
Reasoning
#2

Figure 1: Components overview.

where P is a predicate symbol other than $=$, t_i are terms, α and β are formulas, and ϕ_i and ψ_i are *objective* formulas, that is, ϕ_i and ψ_i mention no $\mathbf{B}$ or $\mathbf{O}$ operators. A formula is *ground* when it contains no variables. A *sentence* is a formula without free variables. Let $\vee$, $\exists$, $\supset$, $\equiv$ be the usual abbreviations, $\top$ stand for $\forall x(x = x)$, and $\bot$ abbreviate $\neg\top$.

Standard names can be seen as special constants that represent all the individuals in the universe. They considerably simplify the semantics compared to the classical Tarskian model theory; in particular, quantification can be handled by simply substituting standard names. The formula $\mathbf{B}(\phi \Rightarrow \psi)$ intuitively expresses a conditional belief, namely the belief that if ϕ is true, then plausibly ψ is also true. The formula $\mathbf{O}\{\phi_1 \Rightarrow \psi_1, ..., \phi_m \Rightarrow \psi_m\}$ goes one step further and says that the conditional beliefs $\mathbf{B}(\phi_i \Rightarrow \psi_i)$ are *all* that is believed; everything that is not a consequence of these conditional beliefs is implicitly not believed. The concept is called *only-believing*; it generalises Levesque's only-knowing [1990] to the case of conditional beliefs and shares many properties with Pearl's System Z [1990]. Only-believing is particularly useful to capture the idea of a conditional knowledge base: the knowledge base is *all* the agent believes.

To investigate such problems, we define a possible-worlds semantics for $\mathcal{BO}$. A *world* is a set of ground non-equality atoms. An *epistemic state* $\vec{e}$ is an infinite sequence of sets of worlds $e_1, e_2, ...$ such that $e_1 \subseteq e_2 \subseteq ...$ and for some $p \in \{1, 2, ...\}$, $e_p = e_{p+1} = ...$

Intuitively, a world is a truth assignment of the ground non-equality atoms, that is, of all $P(n_1, ..., n_j)$ where the n_i are standard names. An epistemic state stratifies sets of such worlds with the subset relation. The intuition behind an epistemic state $\vec{e}$ is to model a system of spheres: e_1 contains the most-plausible worlds, e_2 adds the second-most-plausible worlds, and so on. Note that in any epistemic state $\vec{e}$ only finitely many different sets of worlds are allowed, since the definition requires that $e_p = e_{p+1} = ...$ for some p. (Just as well we could have defined an epistemic state to be a non-empty finite sequence of supersets; the present definition however is often easier to work with.)

Given an epistemic state $\vec{e}$ and a world w, we can define truth of a sentence α in $\mathcal{BO}$, written $\vec{e}, w \models \alpha$. We let α^x_n denote the result of substituting all free occurrences of x by n. The objective part of the semantics is defined inductively as follows:

1. $\vec{e}, w \models P(n_1, ..., n_j)$ iff $P(n_1, ..., n_j) \in w$;
2. $\vec{e}, w \models (n_1 = n_2)$ iff n_1 and n_2 are identical names;
3. $\vec{e}, w \models (\alpha \wedge \beta)$ iff $\vec{e}, w \models \alpha$ and $\vec{e}, w \models \beta$;
4. $\vec{e}, w \models \neg\alpha$ iff $\vec{e}, w \not\models \alpha$;
5. $\vec{e}, w \models \forall x\alpha$ iff $\vec{e}, w \models \alpha^x_n$ for all $n \in \mathcal{N}$.

Notice that Rule 5 handles quantification by substitution of standard names.

Before we proceed with the semantics of conditional belief, we define the *plausibility* $\lfloor \vec{e} \,|\, \phi \rfloor$ of an objective sentence ϕ in $\vec{e}$ as the index of the first sphere consistent with ϕ:

$$\lfloor \vec{e} \,|\, \phi \rfloor = \min\{p \mid p = \infty \text{ or } \vec{e}, w \models \phi \text{ for some } w \in e_p\},$$

where $\infty \notin \{1, 2, ...\}$ represents an "undefined" plausibility with the understanding that $p + \infty = \infty$ and $p < \infty$ for all $p \in \{1, 2, ...\}$. Then the semantics of beliefs is as follows:

6. $\vec{e}, w \models \mathbf{B}(\phi \Rightarrow \psi)$ iff for all $p \geq 1$,
   if $p \leq \lfloor \vec{e} \,|\, \phi \rfloor$ and $w' \in e_p$, then $\vec{e}, w' \models (\phi \supset \psi)$;
7. $\vec{e}, w \models \mathbf{O}\{\phi_1 \Rightarrow \psi_1, ..., \phi_m \Rightarrow \psi_m\}$ iff for all $p \geq 1$,
   $w' \in e_p$ iff $\vec{e}, w' \models \bigwedge_{i: \lfloor \vec{e} \,|\, \phi_i \rfloor \geq p} (\phi_i \supset \psi_i)$.

The intuitive meaning of Rule 6 is that $\mathbf{B}(\phi \Rightarrow \psi)$ holds iff the most-plausible ϕ worlds also satisfy ψ. Note that $\mathbf{B}(\phi \Rightarrow \psi)$ is vacuously true when there is no ϕ world. As a consequence, $\mathbf{B}(\neg\phi \Rightarrow \bot)$ can be used to express that ϕ is known.

Only-believing $\mathbf{O}\{\phi_1 \Rightarrow \psi_1, ..., \phi_m \Rightarrow \psi_m\}$ has the effect of $\mathbf{B}(\phi_i \Rightarrow \psi_i)$ plus maximising every sphere: it requires the sphere e_p to contain *all* worlds that satisfy all $(\phi_i \supset \psi_i)$ for which $\lfloor \vec{e} \,|\, \phi_i \rfloor \geq p$. It turns out that this definition gives rise to a procedure that generates the unique system of spheres $\vec{e}$ that corresponds to $\mathbf{O}\{\phi_1 \Rightarrow \psi_1, ..., \phi_m \Rightarrow \psi_m\}$ [Schwering *et al.*, 2017], which means that a conditional knowledge base has a unique semantic representation as a system of spheres.

While reasoning in the logic as presented here is undecidable given its first-order nature, a decidable (and sometimes tractable) variant based on the theory of *limited belief* has been developed and implemented [Schwering and Lakemeyer, 2016; Schwering, 2017].

5 Implementation

In order to automate the process of solving WSC through reasoning with conditional beliefs, we need to (a) translate a given WSC instance into formulas of $\mathcal{BO}$, (b) provide necessary background knowledge in the form of a KB of $\mathcal{BO}$ formulas, and (c) call the $\mathcal{BO}$ reasoner to answer the query.

5.1 Extraction of Knowledge from ConceptNet 5

Here we adopt an approach based on ConceptNet 5 [Speer and Havasi, 2012], a large semantic network that represents knowledge in a graph structure where each node is a *concept* (e.g., “zoom”, “going_fast”) and each edge corresponds to one of 28 different *relations* (e.g., “IsA”, “RelatedTo”). Apart from manually entered data, the network integrates information from various sources, including subsets of the OpenCyc and DBPedia ontologies as well as the Wiktionary and WordNet 3.0 dictionaries. The origin of each piece of information is kept in metadata associated to each edge, in particular a numerical *weight* representing the strength of the assertion. Weights have a default value of 1 but can be higher or lower, where a negative value expresses that the assertion is false or irrelevant.

As opposed to fully-fledged commonsense knowledge bases such as Cyc [Matuszek *et al.*, 2006], ConceptNet 5 does not come with a pre-defined formal semantics and/or inference engine. It can rather be viewed as an intermediate step between the intuitive, informal knowledge used by humans on the one hand and a formal, logic-based representation on the other. As developers, this gives us the freedom to attach meaning to assertions in a way that suits our application.

In our case, we identified 24 of the 28 relations to be useful and mapped them to conditional beliefs. Consider the *IsA* relation: Each corresponding tuple has the form *IsA(A,B)*, where A and B are concepts. Assuming that we represent concepts by unary predicates, we can encode the assertion as an ordinary belief (expressing that it is plausible to assume that if x is an A, it is also a B):

$$\top \Rightarrow \forall x\,(A(x) \supset B(x)) \tag{1}$$

However, thus the statement will only hold in the most plausible sphere, which essentially amounts to doing monotonic reasoning only. We hence instead use the following encoding:

$$A(x) \Rightarrow B(x) \tag{2}$$

That is, the most plausible A’s are assumed to be B’s. Similar mappings have to be defined for the remaining relations. The ones we used are summarised in Table 1, where $A \Rightarrow B$ stands for formula (2), $A \Rightarrow \neg B$ is shorthand for $A(x) \Rightarrow \neg B(x)$ etc.

Relation	**Description**	**Mapping**
IsA	A is a kind of B	$A \Rightarrow B$
NotIsA	A is not kind of B	$A \Rightarrow \neg B$
RelatedTo	A is related to B	$A \Rightarrow B$ and $B \Rightarrow A$
DerivedFrom	A is derived from B	$A \Rightarrow B$
Antonym	A is the opposite of B	$\neg(A \wedge B)$
MotivatedBy	A is motivated by B	$B \Rightarrow A$
Synonym	A is a synonym of B	$A \equiv B$
FormOf	B is the root word of A	$A \equiv B$
HasPrerequisite	in order for A to happen, B needs to happen	$A \Rightarrow B$
UsedFor	A is used for B	$A \Rightarrow B$
NotUsedFor	A is not used for B	$A \Rightarrow \neg B$
PartOf	A is part of B	$B \Rightarrow A$
HasA	B belongs to A, either as an inherent part or due to a social construct of possession	$A \Rightarrow B$
CapableOf	A is capable of B	$A \Rightarrow B$
NotCapableOf	A is not capable of B	$A \Rightarrow \neg B$
ObstructedBy	A is a goal that can be prevented by B	$B \Rightarrow \neg A$
HasProperty	A has B as a property	$A \Rightarrow B$
NotHasProperty	A has not B as a property	$B \Rightarrow \neg A$
Desires	A is a conscious entity that typically wants B	$A \Rightarrow B$
NotDesires	A is a conscious entity that typically does not want B	$A \Rightarrow \neg B$
CreatedBy	B is a process or agent that creates A	$A \Rightarrow B$
DefinedAs	A and B overlap considerably in meaning, and B is a more explanatory version of A	$A \Rightarrow B$
Entails	If A is happening, B is also happening	$A \Rightarrow B$
MannerOf	A is a specific way to do B	$A \Rightarrow B$

Table 1: ConceptNet relations and their mapping

5.2 Representing Winograd Schemas in $\mathcal{BO}$

To create the knowledge base, we start with the given WSC sentence, e.g.

The fish ate the worm. It was tasty.

Here, the word *tasty* can be replaced by *hungry* and the reference of *it* changes from *worm* to *fish*. Note that this a structure that many (but not all) Winograd schemas exhibit: They start with a statement (unambiguously) describing a situational relationship between two parties (“the fish ate the worm”), followed by an ambiguous description of an effect of it (“it was tasty”). We use the natural language processing tool ReVerb [Fader *et al.*, 2011] to obtain a structured representation for the former in the form of a triple $(\mathit{fish}, \mathit{ate}, \mathit{worm})$.

Note that it is possible to replace both parties with placeholders and still understand the sentence and resolve the pronoun. The requirement that the names of the two parties do not carry any relevant information for solving a WSC, actually intended as a main difficulty, can here be used to our advantage: Instead of “fish” we simply denote the first party by standard name #1, and similarly the “worm” as second party by #2. We will not have to worry about reasoning about the relationship between a fish and a worm, but can rather concentrate on the relation between concepts *ate* and *tasty*.

The triple $(\mathit{fish}, \mathit{ate}, \mathit{worm})$ now gives us the certain information that the *fish* (#1) is the active party in the event “ate”, which we represent by adding

$$\neg \mathit{ate}(\#1) \Rightarrow \bot \tag{3}$$

to the KB. Next, we consult ConceptNet to check if more information about the two parties is available. Specifically, we try to use the fact that the *worm* (#2) is the passive party in the event. We hence search for a concept which has similar properties as an eaten worm, that is to say something which is commonly known *to be eaten*, i.e., something which is the passive party in the event of *ate*. After determining eaten to be the passive form of ate through the common root word eat, we therefore query ConceptNet to find instances of the $\mathit{ReceivesAction}$ relation (that we only use for this specific purpose). In the example, ConceptNet suggests that the concept *apple* is commonly known to be eaten since $\mathit{ReceivesAction}(\mathit{apple}, \mathit{eaten})$ has a high weight. We hence represent that the worm is the passive party by the formula

$$\top \Rightarrow \mathit{apple}(\#2) \tag{4}$$

The new expression can be read as *it is believed that the second party is active in being an 'apple'*. The fact that the passive form of *ate* is *eaten* is further represented as follows:

$$\begin{aligned} \mathit{ate}(x) &\Rightarrow \neg \mathit{eaten}(x) \\ \neg \mathit{ate}(x) &\Rightarrow \mathit{eaten}(x) \end{aligned} \tag{5}$$

The new beliefs can be read as *it is believed that if party x is active in the event 'ate' it presumably is not active in the event 'eaten'*, and vice versa. Finally, we add the belief that if a party is active in the event of *eaten*, then it is presumably active in the event of being an *apple*:

$$\mathit{eaten}(x) \Rightarrow \mathit{apple}(x) \tag{6}$$

Note that beliefs of the form (4) to (6) are only generated in case a corresponding $\mathit{ReceivesAction}$ actually exists (which is often not the case); only the ones of the form (3) are guaranteed to be present in every WSC instance.

5.3 Reasoning

Starting with the formulas generated for the WSC sentence as demonstrated in Section 5.2, we search for a path in ConceptNet that connects the two target concepts *ate* and *tasty* identified from the WSC sentenced (using normalisation like elimination of stop words and word stemming if needed). All relations on this path are mapped to formulas as described in Section 5.1, instantiated by the standard names $\#1, \#2$ representing the two parties. We then call the $\mathcal{BO}$ reasoner on the resulting KB with the two queries

$$\begin{aligned} \mathit{tasty}(\#1) \vee \mathit{tasty}(\#2) &\Rightarrow \mathit{tasty}(\#1) \\ \mathit{tasty}(\#1) \vee \mathit{tasty}(\#2) &\Rightarrow \mathit{tasty}(\#2) \end{aligned} \tag{7}$$

A decision is then made if exactly one of them comes out as true and the other as false.

6 Evaluation

6.1 Quantitative Evaluation

We performed a preliminary experimental evaluation of our implemented system on the publicly available Winograd schema collection[1] containing 144 schemas in total. Since ConceptNet does not directly support querying paths between concepts, we search in a two-stage process. First, a subgraph of ConceptNet is generated in memory, which we do by starting with the target concepts and successively adding adjacent edges and nodes, where for each node the (at most) 30 edges with highest weight are considered. We then search for the actual path by means of the Dijkstra algorithm, alternatingly using edge costs of $1/\mathit{weight}$ or 1 (the latter because sometimes ConceptNet’s weights are not reliable).

Table 2 shows the results for increasing depth limits in terms of the number of decisions made (recall that this means that exactly one of the two queries comes out true) and the fraction of correctly answered sentences. While obviously a search depth of zero yields no results, neither does a limit of one, meaning none of the concepts mentioned in Winograd schemas are directly connected in ConceptNet, showing that these schemas are indeed hard in some sense. Once we set the limit to two or above, decisions are made with increasing accuracy (though not much better than guessing). With the current implementation we were not able to increase the depth beyond three due to reaching memory limitations.

depth	decisions	correctly answered
0	0	0
1	0	0
2	103	52 (50.5%)
3	159	85 (53.5%)

Table 2: Quantitative evaluation

After looking into some of the examples where wrong or no decisions were made, we adjusted our method slightly. First, we know that exactly one of the two parties is referred by the pronoun. Moreover, while one party (the fish) actively participates (eats) in the described scenario as encoded by (3), often the other party (the worm) is not the active one. We hence changed the antecedents of the belief conditionals in the queries (7) to formulas of the form

$$(\mathit{tasty}(\#1) \oplus \mathit{tasty}(\#2)) \wedge \neg \mathit{ate}(\#2)$$

where $\oplus$ denotes exclusive-or. Furthermore, we found that the returned paths disproportionally often contain the *RelatedTo*, *Synonym* and *IsA* relations. To discourage such overuse, we introduced a penalty on the weights for the most frequently appearing relations. These modifications, with depth limit three, resulted in 93 decisions and a success rate of 54.8%, i.e., 51 correct and 42 incorrect answers.

6.2 Qualitative Evaluation

To illustrate our system’s operation, let us a look at a schema successfully resolved by it in more depth, namely “*The delivery truck zoomed by the school bus because it is going so fast/slow*”. ReVerb maps the sentence to the triple

[1] http://www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html

$(\mathit{delivery\ truck}, \mathit{zoomed\ by}, \mathit{school\ bus})$, yielding the two target concepts $\mathit{zoomed\ by}$ and $\mathit{going\ fast}$ (or $\mathit{going\ slow}$, respectively). A search in ConceptNet returns the following path between them: $\mathit{zoomed} \xrightarrow{\mathit{RelatedTo}} \mathit{zoom} \xrightarrow{\mathit{Synonym}} \mathit{whizz} \xrightarrow{\mathit{RelatedTo}} \mathit{sound} \xrightarrow{\mathit{RelatedTo}} \mathit{wave} \xrightarrow{\mathit{RelatedTo}} \mathit{water} \xrightarrow{\mathit{RelatedTo}} \mathit{boat} \xrightarrow{\mathit{IsA}} \mathit{going\ fast}$. The KB constructed from this is then

$$\begin{aligned}
\{\neg \mathit{zoomed}({}^{\#}1) &\Rightarrow \bot, \\
\mathit{zoomed}(n) &\Rightarrow \mathit{zoom}(n), \\
\mathit{zoom}(n) &\Rightarrow \mathit{zoomed}(n), \\
\top &\Rightarrow \mathit{whizz}(n) \equiv \mathit{zoom}(n), \\
\mathit{whizz}(n) &\Rightarrow \mathit{sound}(n), \\
\mathit{sound}(n) &\Rightarrow \mathit{wave}(n), \\
\mathit{wave}(n) &\Rightarrow \mathit{sound}(n), \\
\mathit{wave}(n) &\Rightarrow \mathit{water}(n), \\
\mathit{water}(n) &\Rightarrow \mathit{wave}(n), \\
\mathit{boat}(n) &\Rightarrow \mathit{water}(n), \\
\mathit{water}(n) &\Rightarrow \mathit{boat}(n), \\
\mathit{boat}(n) &\Rightarrow \mathit{going\_fast}(n) \mid n \in \{{}^{\#}1, {}^{\#}2\}\}
\end{aligned}$$

which correctly suggests that party ${}^{\#}1$ is the correct one, i.e. the *delivery truck*. For the alternate sentence we obtain the KB

$$\begin{aligned}
\{\neg \mathit{zoomed}({}^{\#}1) &\Rightarrow \bot, \\
\mathit{zoomed}(n) &\Rightarrow \mathit{zoom}(n), \\
\mathit{zoom}(n) &\Rightarrow \mathit{zoomed}(n), \\
\mathit{zoom}(n) &\Rightarrow \mathit{travel}(n), \\
\mathit{rush}(n) &\Rightarrow \mathit{travel}(n), \\
\mathit{zoom}(n) &\Rightarrow \mathit{travel\_rapidly}(n), \\
\top &\Rightarrow \mathit{hurry}(n) \equiv \mathit{rush}(n), \\
\top &\Rightarrow \mathit{hurry}(n) \equiv \mathit{travel\_rapidly}(n), \\
\top &\Rightarrow \neg(\mathit{rush}(n) \wedge \mathit{go\_slow}(n)) \mid n \in \{{}^{\#}1, {}^{\#}2\}\}
\end{aligned}$$

which similarly relies heavily on $\mathit{RelatedTo}$, $\mathit{Synonym}$, and IsA, but additionally uses that *rush* is an antonym of *go_slow*. The system again yields the correct answer ${}^{\#}2$ (the school bus) to the query "what is going slow?".

The examples demonstrate that most generated conditionals seem reasonable and relevant for the schema, however sometimes interspersed with ones that are not as intuitive. The latter is often due to the $\mathit{RelatedTo}$ relation, whose meaning is on the one hand very vague, and which on the other hand accounts for the majority of edges in ConceptNet. In any case, even if the system yields a wrong answer, we obtain human-understandable justifications for the automatically made decisions, which we believe is a major benefit of a knowledge-based approach to the WSC.

7 Conclusion

We presented a solution to the Winograd Schema Challenge based on reasoning about conditional beliefs, where knowledge bases are generated by means of the NLP tool ReVerb and the semantic net ConceptNet 5. Unlike many other approaches that rely on searching text repositories (Google), we thus adopt a knowledge-based approach that, when it comes to the reasoning part, has the benefit that the generated $\mathcal{BO}$ KB can be regarded as a human-readable justification for the decision made by the system, which moreover comes with a clear formal semantics. We also employ a major difficulty of the WSC to our advantage, namely that the two parties in a schema do not carry any relevant information and could equally be replaced by placeholders.

While we do not achieve a very high quantitative success rate on the WSC dataset, our results are consistently better than chance, indicating that this may be a step in the right direction. Our system's performance of course heavily depends on the quality of data fed into it, both from the NLP module's and ConceptNet's side. In particular, information in ConceptNet on many topics is rather sparse and relies heavily on the *RelatedTo* relation, whose meaning is very vague and ambiguous.

There are many directions for future work. First of all, further experimentation with our system may be in order. Our mapping from ConceptNet relations to belief conditionals is not set in stone, and quite simple in the sense that many relations are represented in the same way. It would be interesting (and not much effort to implement in our existing framework) to try other translations. For example, at the moment our system neglects that there may be interactions between the different relations, e.g. we may want to deduce that X is not capable of doing Y from the facts that X is not capable of doing Z and that Z is a prerequisite of Y.

Furthermore, ConceptNet was only one possible choice for a commonsense knowledge repository, and the system may benefit from trying other alternatives such as the Never-Ending-Learning (NELL) [Mitchell *et al.*, 2015] KB. In contrast to ConceptNet, NELL employs a large number of lower level relations such as *playsInstrument* which on the one hand would reduce ambiguity, but on the other hand increase the required modelling effort, i.e., constructing a mapping.

Acknowledgments.

This work was supported by the German Research Foundation (DFG) research unit FOR 1513 on Hybrid Reasoning for Intelligent Systems, project A1.

References

[Bailey *et al.*, 2015] Dan Bailey, Amelia Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. The Winograd schema challenge and reasoning about correlation. In *Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*, 2015.

[Fader *et al.*, 2011] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics, 2011.

[Levesque *et al.*, 2012] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema Challenge. In Gerhard Brewka, Thomas Eiter, and Sheila A. McIlraith, editors, *Proceedings of the Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2012)*, pages 552–561. AAAI Press, 2012.

[Levesque, 1990] Hector J. Levesque. All I know: a study in autoepistemic logic. *Artificial Intelligence*, 42(2), 1990.

[Liu *et al.*, 2017] Quan Liu, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. Cause-effect knowledge acquisition and neural association model for solving a set of Winograd schema problems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2344–2350. AAAI Press, 2017.

[Matuszek *et al.*, 2006] Cynthia Matuszek, John Cabral, Michael Witbrock, and John Deoliveira. An introduction to the syntax and content of Cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, pages 44–49, 2006.

[Mitchell *et al.*, 2015] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, pages 2302–2310. AAAI Press, 2015.

[Pearl, 1990] Judea Pearl. System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In *Proceedings of the Third Conference on Theoretical Aspects of Reasoning about Knowledge (TARK)*, pages 121–135. Morgan Kaufmann, 1990.

[Rahman and Ng, 2012] Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: The Winograd schema challenge. In Jun'ichi Tsujii, James Henderson, and Marius Pasca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 777–789. ACL, 2012.

[Schwering and Lakemeyer, 2016] Christoph Schwering and Gerhard Lakemeyer. Decidable reasoning in a first-order logic of limited conditional belief. In *Proceedings of the Twenty-Second European Conference on Artificial Intelligence (ECAI)*, pages 1379–1387. IOS Press, 2016.

[Schwering *et al.*, 2017] Christoph Schwering, Gerhard Lakemeyer, and Maurice Pagnucco. Belief revision and projection in the epistemic situation calculus. *Artificial Intelligence*, 251:62–97, 2017.

[Schwering, 2017] Christoph Schwering. Limbo: A reasoning system for limited belief. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5246–5248. AAAI Press, 2017.

[Sharma *et al.*, 2015] Arpit Sharma, Nguyen Ha Vo, Somak Aditya, and Chitta Baral. Towards addressing the Winograd schema challenge – building and using a semantic parser and a knowledge hunting module. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1319–1325. AAAI Press, 2015.

[Speer and Havasi, 2012] Robert Speer and Catherine Havasi. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 3679–3686. European Language Resources Association (ELRA), 2012.