Bennett B. To appear in: Neuhaus F, Brodaric B, eds. Proceedings of the 12th Int'l Conf. (FOIS 2021), Frontiers in Artificial Intelligence and Applications by IOS Press.

Semantic Analysis of Winograd Schema No. 1

Brandon BENNETT [a,1]
[a] *School of Computing, University of Leeds, UK*

Abstract. The Winograd Schema Challenge is a general test for Artificial Intelligence, based on problems of pronoun reference resolution. I investigate the semantics and interpretation of Winograd Schemas, concentrating on the original and most famous example. This study suggests that a rich ontology, detailed commonsense knowledge as well as special purpose inference mechanisms are all required to resolve just this one example. The analysis supports the view that a key factor in the interpretation and disambiguation of natural language is the preference for *coherence*. This preference guides the resolution of co-reference in relation to both explicitly mentioned entities and also implicit entities that are required to form an interpretation of what is being described. I suggest that assumed identity of implicit entities arises from the expectation of coherence and provides a key mechanism that underpins natural language understanding. I also argue that conceptual ontologies can play a decisive role not only in directly determining pronoun references but also in identifying implicit entities and implied relationships that bind together components of a sentence.

Keywords. natural language semantics, pronoun resolution, coherence, ontology

1. Introduction

The Winograd Schema Challenge (WSC) was proposed by Levesque *et al.* [1] as an updated form of the Turing Test. It provides a method for evaluating AI systems by means of a text processing problem, whose solution seems to require both understanding of the meaning of natural language, background knowledge of physical and social situations and commonsense reasoning. Specifically, the WSC is the task of solving pronoun resolution problems having similar form to the following paradigm case (originally considered by Terry Winograd [2]):

WS1. ***The city councilmen*** *refused* ***the demonstrators*** *a permit because* **they [***feared/ advocated***]** *violence.*

Here, the pronoun to be resolved is '**they**', and its possible referents are '***the city councilmen***' and '***the demonstrators***'.

[1]Corresponding Author: Brandon Bennett, E-mail: B.Bennett@leeds.ac.uk. Contributions of my collaborators Suk Joon Hong and Judith Clymo and extremely useful suggestions from the paper's referees are gratefully acknowledged.

Each schema actually corresponds to *two*[2] problem cases, which differ in the choice of one of two alternative words or phrases, indicated here by the notation '[*A*/*B*]'. So **WS1** specifies the problems of determining the referent of the pronoun '**they**' in either of the sentences **WS1a** and **WS1b** resulting from selecting each of the two alternatives. The reason why two alternatives are given, is to guard against the possibility that the anaphora resolution can be accomplished by means of some structural analysis of the sentence that can be done without any consideration of the meaning of the sentence. Since the two alternatives are syntactically identical, apart from one word (or short phrase), but imply different resolutions of the pronoun, this prevents the resolution being determined purely by the syntactic category of the co-referring expression — or so we may hope.

Although WSC was proposed to provide a general test for AI rather than stipulating how the challenge should be addressed, the paper's conclusion advocates an approach based on knowledge representation:

> *"While this approach (KR) still faces tremendous scientific hurdles, we believe it remains the most likely path to success. That is, we believe that in order to pass the WSC, a system will need to have commonsense knowledge about space, time, physical reasoning, emotions, social constructs, and a wide variety of other domains."* [3]

In the current paper I examine **WS1** from the point of view of logic and semantics, and attempt to identify structures and principles by which WS examples can be resolved. My aim is to provide illustrations and arguments supporting the following views:

- The semantic and background knowledge and types of inference required to resolve a WS can be extremely complex.
- Coherence (and cohesion) principles are key to natural language understanding.
- Natural language interpretation is heaviliy constrained and enabled by semantic type and role relationships.
- Connections between explicitly mentioned objects and concepts are mediated by the existence of *implied entities*, as well as those explicitly mentioned.
- Ontology provides a means of specifying and identifying the relevant semantic types, roles and entities required to establish coherence preference and make inferences based on these preferences.
- Despite it being clear that enormous difficulties arise when the problem of natural language understanding by means of Ontologies and KRR techniques, this is still a good approach.

1.1. Previous Work

Although the WSC was designed primarily with KR type approaches in mind, it seems that the problem has received more attention from researchers using ML techniques. Probably the first fully automated WSC resolving system was that of [4] which used a SVM algorithm working with several linguistic features, some of which are based on semantic features relationships between words within in the WS sentences. In recent work, researchers using methods based on neural language models such as BERT [5] and RoBERTa [6] have demonstrated high statistical accuracy in resolving Winograd

[2]It is possible to devise examples with more than two cases but for simplicity of exposition we assume that we only deal with Schemas with two cases.

schemas [7, 8, 9]. An accuracy of 90% for the original WSC problem set (WSC273) is reported in [9]. However, performance goes down significantly on the larger WSC data sets [9, 10]. But despite apparent the high accuracy of BERT-based solutions, there are several reasons to suspect that BERT's understanding of sentences is superficial. It does not work well on sentences involving function words such as negation [11], and lacks of robustness to with respect to semantically insignificant variations of input sentences (e.g. cases where just proper names are changed [7, 12]). And of course, another shortcoming of the ML approaches is that they do not give any kind of explanation of their answers.

KR methods resolve Winograd schemas by creating a logical representation of a sentence and relevant background knowledge and applying inference rules. The advantage of KR is that it can give meaningful explanations for the answers. [13] define "correlation calculus" to resolve Winograd schemas by adding a novel *correlation* connective to first-order logic. However, this method requires that WSC sentences be accurately translated into first-order formulae and all relevant background knowledge needs to be manually defined in the form of correlation calculus axioms. [14] tackles the WSC by using a semantic parser (K-Parser) to extract semantic relationships from sentences and match them to identity rules that can be automatically extracted from text corpora. However, rules that successfully resolve the pronoun are found in less than half of the cases.

2. 'Coherence' and its Application to Pronoun Resolution

Informally, we may say that a text or dialogue is *coherent* if 'it fits together well'. This phrase describes language as if it were self-assembled furniture, which may be a good metaphor. But what does 'fitting together' mean in the case of language? There is no simple answer to this question. The ways in which components of language fit together are many and varied. In this paper I explore some kinds of coherence that I believe to be particularly important in natural language understanding, but, presumably because of their somewhat covert mode of operation, have not been given the attention they deserve.

As was so convincingly argued by Grice [15], and is now generally accepted, the interpretation of language is greatly dependent upon and conditioned by various principles that arise from the cooperative nature of communication. Such principles enable language to be understood in a way that is far less ambiguous than would be the case if we relied purely on the explicitly asserted content of linguistic expressions. *Coherence* is a structural property of language rather than a maxim of communication. However, it certainly plays a role in satisfying Gricean maxims, especially those of clarity and orderliness. Several researchers have suggested that coherence is a key factor in natural language understanding and have also tried to characterise more precisely what is meant by coherence and to identify [16] or even measure [17] coherence in language samples. Coherence is often considered to arise where successive clauses or sentences refer to the same things. Hence coherence is associated with *co-reference*.[3]

Hobbs [16] applied the idea of coherence to developing computational mechanisms for understanding natural language text, and argued that the principles involved arise from a relatively small number of logical principles. He also suggested that the tendency

[3]Linguists have distinguished between *coherence* and *cohesion*; the latter being applied to describe more surface level associations. In the current paper I do not use this terminology. I believe the distinction is not always clear cut.

for coherence to involve co-reference does not arise because co-reference creates coherence but rather it is the other way around: effective communication depends on certain types of progression such as elaboration and clarification, and these types of coherent progression tend to involve co-reference.

Whether or not coherence produces or is produced by co-reference does not seem to bear directly upon the current analysis. In fact, I believe that the dependency runs in both directions. But if we take the example of a particular case of pronoun resolution, such as a WS, we already know that there must be co-reference, since the pronoun must refer to something referred to elsewhere in the text. Then by appealing to the Gricean principle of non-ambiguity we can assume that there must be some principle by which the reference of the pronoun can be determined. The following example from [16] is a good example of Hobbs' approach and also a good WS example:

- John can open Bill's safe. He knows the combination.

That 'he' refers to 'John' can be explained on the basis of a principle of *elaboration*, since knowing its combination can be regarded as a more detailed explanation of being able to open a safe. This is an example general class of elaborations in which an agent's ability to do something is explained in terms of them having some kind of knowledge.

Now consider:

- John can open Bill's safe. He will have to get the combination changed soon.

Hobbs says that this is an example of a 'causal coherence relation' which depends upon knowledge of the purpose of a safe and the purpose of a combination. So it seems that Hobbs' view is that, although his relatively simple coherence relations will work in many cases, coherence may also be dependent on much more complex knowledge.

2.1. The Approach of Kehler et al.

Despite **WS1** being the original WS example and also being one for which the implied pronoun resolutions are fairly clear, explicit explanations of the principles behind its resolution are scarce in the literature. As far as I am aware, the most detailed analysis of this specific case is that of Kehler *et al.* [18]. Specifically, they say "Oversimplifying a bit, we encode the world knowledge necessary to establish explanation for **WS1** within a single axiom." The axiom they give is:

$$Fear(x,v) \wedge Advocate(y,v) \wedge Enable\_to\_cause(z,y,v) \rightsquigarrow Refuse(x,y,z) \qquad (1)$$

To clarify this they state that the implication relationship ($\rightsquigarrow$) means that the formula on the right 'plausibly follows from' those on the left.[4] They proceed to explain how (1) is sufficient to deal with both **WS1a** and **WS1b** by means of abductive inference. The idea is that when interpreting '*P because Q*', we try to match *P* with the consequence of some plausible implication an then to match *Q* with one of the antecedents of this implication. Hence, with **WS1** this would work as follows:

The first clause in both versions of **WS1** is of the form:

[4]This is stated in [18], but I have replaced the symbol $\rightarrow$ in their formula with $\rightsquigarrow$, to avoid possible confusion with a material implication.

- *Refuse*(*councillors*, *demonstrators*, *permit*),

so an explanation of this can be given by an instantiation of (1), with the variable assignment $x = councillors$, $y = demonstrators$ and $z = permit$. Using this assignment, the clauses on the left would be instantiated as:

- $Fear(councillors, violence)$,
- $Advocate(demonstrators, violence)$,
- $Enable\_to\_cause(permit, demonstrators, violence)$.

Thus, since **WS1a** contains $Fear(they, violence)$, we can match 'they' in this sentence to 'the councillors' and in **WS1b** we have $Advocate(they, violence)$, so in this case 'they' matches 'the demonstrators'.

I concur with many aspects of this analysis. (1) is a reasonable, albeit fairly coarse grained, representation of a general principle by which **WS1** may be resolved. It is indeed the case that, if an agent fears some outcome advocated by another agent, which would require some item to achieve that outcome, then that provides an explanation of why the first agent would prevent the second agent acquiring the item. I also agree that a sentence of the form '*P because Q*' can only make sense if *Q* can play a part in some possible explanation of *P*. Furthermore, if in such a sentence *Q* contains a pronoun then *P because* Q' should make sense, where Q' is formed by replacing the pronoun in *Q* with some proper noun or noun phrase occurring elsewhere in the sentence.

One could find various respects in which (1) may need refinement or elaboration. One could also complain that it is infeasible that one could formalise all the principles required to resolve all of the huge range of potential WS examples that could be devised. Neither of these seems to be a decisive argument against the use of principles of similar form to (1). However, I believe that there is another major problem facing this approach.

2.1.1. The Problem of Identifying the Appropriate Principle

The critical problem is how to identify the appropriate principle to apply to a particular WS example. There are countless reasons why one agent or group would deny something to another agent or group. Hence, any set of principles sufficient to handle a wide range of WS examples would contain many plausible inference rules with $Refuse(x, y, z)$ as the consequent. When we apply the induction rule we also make use of additional information such as $Fear(they, violence)$ to find matching explanation, and this will narrow down the choice of applicable rules. But can we expect this matching to narrow down the possibilities sufficiently to identify a single correct rule? I believe not. Or at least not with a rules that is as coarse grained as (1).

We may think that we can find the appropriate rule by means of the additional information given in the WS example. However, note that while (1) contains the predicates $Fear(x, v)$ and $Advocate(y, v)$ each of **WS1a** and **WS1b** mention only one of the corresponding relationships, so supposition that (1) explains the situation relies on inductive inference that the other also holds. But it does not seem reasonable that from knowing only $Refuse(x, y, z)$ and $Fear(x, v)$ we can deduce $Advocate(y, v)$. Consider this variation:

- The councillors refused the demonstrators a permit because they feared that pickpockets would take advantage of a crowd of unsuspecting middle class do-gooders milling around in the town square. It would be a nightmare for the local police.

In this case it is clear that the demonstrators would not be advocating the outcome that the councillors fear. And we can easily imagine scenarios involving permits and fear where a variety of different rules operate. For example:

- The councilmen gave the racketeers a permit because they feared blackmail.
- The psychologists refused the patients a certificate of mental health because they feared leaving their own house.

We have not yet established how the relationship $Enable\_to\_cause(z, y, v)$ that occurs in (1) might be used to guide selection of this rule. As we saw above by matching the relationships explicitly stated in **WS1a** in terms of the verbs 'refuse' and 'fear', to the rule (1) we get $Enable\_to\_cause(permit, demonstrators, violence)$. From the explanation given by Kehler *et al.* [18] it seems that they consider this relationship to be a further product the inductive inference although it is not necessary to resolve the pronoun 'they', which can be determined from $Fear(councillors, violence)$ alone. However, since, as I have pointed out, the presence of the relationships expressed in terms of 'refuse' and 'fear' does not seem sufficient to guarantee that the rule is appropriate, it could be argued that recognition of the relationship $Enable\_to\_cause(permit, demonstrators, violence)$ also plays a key role in identifying that the rule (1) is appropriate for this case.

But, since $Enable\_to\_cause(permit, demonstrators, violence)$ is not explicit in **WS1a**, we still face the problem of where would this come from. Nevertheless, it is plausible to argue that this kind of relationship could be part of background knowledge which must be employed in conjunction with explanatory rules such as (1) and includes information such as causal relationships that can occur involving particular types of agent and object. Indeed this seems to fit very well with Hobbs's suggestion that coherence principles based on causal relationships (such as between a safe and its combination) need to be specified as background knowledge. Hence, in addition to the rule (1) our theory would also contain:

$$Enable\_to\_cause(permit, demonstrators, violence) \tag{2}$$

This idea seems to me to be along the right lines. However, it is not without further problems. Suppose we have a knowledge base containing instances of the *Enable_to_cause* relation. It would contain cases such as *Enable_to_cause*(*knife*, *idiot*, *death*). Could such a knowledge base ever be complete or accurate? It seems highly unlikely that one could cover all possible cases of causal enablement without massive over generalisation. In the case of the particular relationship *Enable_to_cause*(*permit*, *demonstrators*, *violence*), one cannot assume that this is always relevant, even when interpreting sentences that mention the concepts 'permit', 'demonstrators' and 'violence'. For example, consider the following case:

- The demonstrators were protesting that the councillors had approved a permit for the knife throwing festival because they feared violence.

We need to recognise that the applicability of *Enable_to_cause(permit, demonstrators, violence)* to a particular situation depends on several further assumptions about the relationship between the three elements. Most obviously, it depends on the assumption that the permit relates to the demonstration being planned by the demonstrators and also that violence may arise from the demonstration. I will argue in the rest of this paper

that such assumptions are plausible and often necessary for the interpretation of natural languages. However, I shall claim that such connections do not in most cases arise from general relationships of the form of (2). Instead they arise primarily from a kind of coherence property, which, as far as I am aware, has not been emphasised in any previous work on coherence. The type of coherence to which I wish to draw attention results from a principle *identification of implied entities*. This is a principle that we apply pervasively but largely unconsciously in our interpretation of natural language. In the remainder of the paper I shall attempt to explain how this works and why it is so powerful.

3. A 'De-Coherent' Interlude

So you have probably heard the news: "The city councillors refused the demonstrators a permit because they feared violence." Let me elaborate further:

> The councillors of Bolzano had been approached by a party of Dutch clog makers, who were on their way to Lagos to hold a demonstration against import tariffs on hand-painted clogs. The prospective demonstrators were from Leiden but were requesting a permit on behalf of a group of farmers from the neighbouring town of Zoetermeer, from whom they often bought tulips, and who (because of floral oversupply in Holand) wished to relocate to Colombia to set up a tulip plantation. The opportunity to carry this out was made possible by the 'Los valientes pueden crecer' (the brave may grow) initiative, a scheme by which city councils of other countries could issue agricultural licences to farmers wishing to establish cultivation in Colombia. The reason that the Lieden clog makers had chosen to stop in Bolzano on their way to Lagos was primarily to visit some cheese makers with whom they had a trading relationship and from whom they had learned that the Bolzano council had recently issued a permit to allow some wine growers from Merano to set up a vineyard near Medellín. One of the Zoetermeer farmers had heard about this when making a tulip delivery to Lieden; and since the Bolzano council were clearly familiar with the scheme and the required paperwork, it made perfect sense for the Lieden clog makers to apply for the permit on behalf of the Zoetermeer farmers during their stay in Bolzano.
>
> However, a condition of the 'brave may grow' scheme was that such permits could only be given to prospective farmers who would not only produce nourishing food but also be capable of defending the land they would be allocated, against brutal and heavily armed drug cartels (who preferred the cultivation of cocaine to other vegetation). When the Bolzano councillors interviewed the Dutch delegation, they found them to be of very different character from the mountain toughened South Tyrolean wine growers, whose permit they had previously approved. One clog maker let slip that she and her fellow artisans were greatly worried that they might face violent aggression from the Nigerian authorities during their planned demonstration in Lagos. And, since the Dutch clog-makers seemed so afraid of potential violence from the Nigerian police, the Bolzano councillors judged that their tulip growing compatriots were likely to be of similarly meek disposition, and would be no match for Colombian drug lords. So the councillors refused to approve the permit because the demonstrators feared violence.

If you had thought it was the city councillors that feared violence, you were mistaken. Why did you think that? Probably the reason was that you applied a *coherence preference* in your interpretation of (1). You assumed that the city councillors were councillors of the *same* city in which the demonstrators planned their protest. And you assumed that the permit was a permit for these *same* demonstrators to hold a demonstra-

tion in that *same* city. But no, the situation involved several different locations, and the requested permit had no direct connection to the planned demonstration.

I am *not* arguing that **WS1a** is genuinely ambiguous. I believe that the 'correct' pronoun resolution for **WS1a**, when seen on its own is that the 'they' refers to the councillors. My counter-interpretation is very complex and artificial. My reason for constructing it was not primarily to show that the pronoun could be interpreted differently but rather to highlight the strength and pervasiveness of coherence principles that condition our interpretation of natural language. I describe my interpretation as *decoherent* because it deactivates usual coherence conventions by statements that negate identities between entities that would otherwise be assumed to be the same.

4. The Logic of 'Because'

Many of the Winograd Schemas are of the form 'ϕ because ψ'. The meaning of 'because' is somewhat difficult to define. It is generally agreed that 'ϕ because ψ' implies 'ϕ and ψ'. However, it is also clear that 'ϕ because ψ' says more than just the truth functional conjunction. Informally, we may that 'ϕ because ψ' is true whenever both ϕ and ψ are true and ψ gives an *explanation* of ϕ. Schnieder [19] presents a logical calculus for the 'because' connective using Natural Deduction style rules. The intuition underlying that system is also that 'ϕ because ψ' holds when ψ explains ϕ. However, the rules of Schnieder's calculus are limited to cases where the form of explanation is itself purely logical. For example, one rule says that from ϕ we can derive '$(\phi \vee \psi)$ *because* ϕ'.

Let us analyse 'because' in terms of what it means for one statement to provide an explanation for another. Consider a statement 'x did A because P', where A is some voluntary action performed by x. In such a case, this only makes sense if P gives some reason that explains why x would choose to do A. The explaining statement can be of many forms and can refer to a very wide range of possible factors that could motivate x to perform A. We may distinguish two broad categories of explaining statement:

- those that refer to some mental property of x (such as a belief, desire or intention),
- those that refer to some claimed fact about the world (including possible future occurrences and also the actions or possible actions of other agents in the world).

For present purposes, I shall consider only the second type of explanation. In such a case, the claimed fact P is proposed as an explanation of x's action A, without any explanation of why P would motivate this action. Thus we must fall back on an implicit, generic explanation of how a fact would motivate a action. I suggest the following:

- On the basis of P, together with other background and contextual knowledge, it is possible to reason that either:
  - ∗ doing A will have an outcome that is good for x;
  - ∗ *or*, **not** doing A may lead to a state that is bad for x;

Here, what is 'good' or 'bad' for x should be interpreted very generally. As well as material benefits or adversities, it includes conditions of status and obligation. Thus, the outcome of fulfilling an obligation or duty would be considered good and of failing to fulfil an obligation or duty would be bad.

4.1. Logical Properties of an 'Explains' Connective

I define the notation $\phi \rightsquigarrow \psi$ to means that 'ϕ can provide an explanation for ψ'.[5] From consideration of Schnieder's account of 'because' and also the specific requirements for resolving **WS1**, I propose that the $\rightsquigarrow$ connective provides the following minimal principles of inference:

Contingent Entailment: $\phi \rightsquigarrow \psi$ if ($\phi \vdash \psi$ and $\nvdash \neg\phi$ and $\nvdash \psi$)

Lexical Semantic Implication: $\phi \rightsquigarrow \psi$ if ψ can be obtained from ϕ by applying axioms expressing semantic properties and relationships among vocabulary terms.

Transitivity: If $\phi \rightsquigarrow \psi$ and $\psi \rightsquigarrow \xi$, then $\phi \rightsquigarrow \xi$.

5. A Partial 'Formalised' Solution

I now present an account of the inference patterns that underlie the resolution of **WS1a**. The presentation is 'formalised' in a weak sense. Axioms are suggested that are intended to express the logical form of valid inferences but a proof system and semantics are not given. The following notations will be used:

- $\phi \rightsquigarrow \psi$ means that 'ϕ can provide an explanation for ψ' and follows the principles stated in the previous Section.
- The variables e range over *possible events*, that is potential occurrences that may or may not actually happen.
- $\mathsf{Occurs}(e)$ means that the possible event e actually occurs.
- $\mathbf{B}_a\,\phi$ means that agent a believes that proposition ϕ is true.
- $\mathit{Good\_for}(\phi, a)$ and $\mathit{Bad\_for}(\phi, a)$ mean. respectively that ϕ being true is good for or bad for agent a.
- All un-subscripted single letter free variables (e.g. a, x, e) are taken as universally quantified with wide scope.
- Subscripted single letter free variables (e.g. a_1, e_1) are Skolem constants (i.e. existentially quantified with wide scope).

I also define the conditions where a possible event would be good (or bad) for an agent as follows:

$$\mathit{Good\_for}(e,a) \equiv_{def} (\mathsf{Occurs}(e) \rightarrow \phi) \wedge \mathit{Good\_for}(\phi,a) \quad (3)$$

$$\mathit{Bad\_for}(e,a) \equiv_{def} (\mathsf{Occurs}(e) \rightarrow \phi) \wedge \mathit{Bad\_for}(\phi,a) \quad (4)$$

Note that the formulation I use here does not include any explicit represent of time and temporal relationships. These would certainly be necessary for a more generally applicable framework, and the scenario described in **WS1** does imply certain temporal relationships. However, it seems that temporal relationships do not play an essential part in the reasoning required to justify the pronoun resolution.

[5] So the symbol has very similar, but slightly different meaning from how it was used in my earlier explanation of the formulation of Kehler *et al.* [18].

5.1. *Instantiations of* **WS1a**

The following formulae are representations of respectively the 'correct' and 'incorrect' versions of **WS1a** with for each of the possible candidates being substituted for 'they':

$$Fear(councillors, violence) \rightsquigarrow Refuse(councillors, demonstrators, permit) \quad (5)$$

$$Fear(demonstrators, violence) \rightsquigarrow Refuse(councillors, demonstrators, permit) \quad (6)$$

To demonstrate a solution for **WS1a** we need to show that from some intuitive and general principles we can derive (5) but cannot derive (6).

Note that I am ignoring any quantification that may be implicit in the noun phrases 'the demonstrators', 'the councillors', 'a permit'. Although, quantification is of course often very important in explaining reasoning, I believe that in this particular example it does not play a significant part and that trying to account for it would unnecessarily complicate the exposition.

5.2. *Explanation of a 'Preventing' Action*

If an agent believes that the occurrence of an event implies a possible state that is bad for the agent then that provides an explanation why the agent would try to prevent the event.

$$\mathbf{B}_a\, Bad\_for(e, a) \quad \rightsquigarrow \quad Try\_to\_prevent(a, e) \quad (7)$$

For an agent to fear violence means that the agent believes there is a possible event such that if it occurs it will have result that is bad for the agent. This is a semantic property of the verb 'fear':[6]

$$Fear(a, violence) \rightarrow \mathbf{B}_a\, \exists e[Violent(e) \wedge Bad\_for(e, a)] \quad (8)$$

Let us substitute *councillors* for a. Then, under the assumption that the domain of possible events includes all events that anyone might believe could exist, we can Skolemise the $\exists e$ and replace with an arbitrary event constant e_1. We then get:

$$Fear(councillors, violence) \rightarrow \mathbf{B}_{councillors}\, Bad\_for(e_1, councillors) \quad (9)$$

Since (9) expresses a purely semantic implication, it can be considered as an explanation, so entails:

$$Fear(councillors, violence) \rightsquigarrow \mathbf{B}_{councillors}\, Bad\_for(e_1, councillors) \quad (10)$$

Then, combining (10) and (7) by instantiating the variable in 7) and using the transitivity of '$\rightsquigarrow$' gives:

$$Fear(councillors, violence) \rightsquigarrow Try\_to\_prevent(councillors, e_1) \quad (11)$$

5.2.1. *Refusing a Permit is a Way of Preventing an Event*

We need establish that refusing a permit is a way of trying to prevent an event. In order to do this we need to examine how a permit is related to various agents and possible events. Some consideration will reveal that a permit is a very complex item in terms of the relationships that it involves. I suggest that, even after some simplification, a permit involves at least the following implied relationships and entities:

[6] In a more detailed representation the *Fear* relationship would be be an attitude towards a *future* event. An exploration of how one might define emotion concepts can be found in [20].

- an agent or institution with the power to issue or approve the permit,
- the person or group to which the permit confers permission,
- the activity or event that the permit permits,[7]
- the location where the permitted activity or event may take place,
- the time period during which the permitted activity may take place,
- the rules of eligibility of the permit.

Given that permits are such complex things, the formalisation of actions involving permits is quite tricky and could be done in various different ways, depending on how you want to decompose the actions and bundle together the related entities. Assuming that the list of relevant entities I have given is sufficient to uniquely determine a possible permit,[8] we can represent the permit as a functional term, with the meaning that the term denotes a permit function that is determined by its relationship to these entities. For, example Sky City council may have the power to authorise a permit for Janet Jones to use a jet-pack within designated areas of Sky City during daylight hours and subject to specified restrictions. I this case $permit(scc, jj, ujp, da, dl, r)$ would denote a potential permit involving the designated entities (abbreviated by initials) fulfilling the roles described above. In many cases, we will not know all the entities relating to the permit (e.g. we may know that Janet has a jet-pack licence but not who issued the permit or what areas or times it is valid for). In such cases we can simply replace the names of unknown entities with existentially quantified variables.

We now need to clarify and define what is meant by 'x refused y a permit'. This has considerable underlying complexity. The issuing of a permit might involve several stages, and be subject to different kinds of refusal. Also, in some cases, one might request a permit on behalf of another person (e.g. a parent on behalf of a child). For present purposes I consider a simplified but typical case, where an agent or group applies for a permit relating to that same agent or group. I will interpret the relationship $Refuse(x, y, permit)$ as a concise way of stating that an event occurs where x refuses to authorise a permit (regarding which they have authority):

$$\begin{aligned} &Refuse(x, y, permit) \leftrightarrow \\ &\quad \exists e \exists l \exists d \exists r [\ \mathsf{Occurs}(refuse(x, authorise(x, permit(x, y, e, loc, dur, rules))))] \end{aligned} \tag{12}$$

If we now consider possible explanations of why a permit might be refused, it is apparent that wanting to prevent the event that it would permit is a good general explanation for such a refusal. We can formalise this idea with the following axiom:

$$Try\_to\_prevent(x, e) \rightsquigarrow \mathsf{Occurs}(refuse(x, authorise((x,\ permit(x, c, e, l, d, r))))) \tag{13}$$

Hence, in a specific case of **WS1a**, where the councillors refuse the demonstrators a permit, this justifies the explanation:

$$Try\_to\_prevent(councillors, e_2) \rightsquigarrow Refuse(councillors, demonstrators, permit) \tag{14}$$

[7] Strictly speaking, permits will be valid for some *type* of event rather than a specific event occurrence (even in the case of a permit for a one-off event it would apply to many different ways in which a particular event could occur). While this is a significant ontological distinction, it does not appear to be critical for the **WS1** example, so I shall assume that a permit is in relation to an individual (possibly non-continuous) event entity.

[8] I am aware that we now have possible objects as well as possible events being referred to, but this seems to be necessary for interpreting the refusal of a permit.

It is important to note that e_2 refers to some particular but unspecified event. It is a Skolem constant arising from the existentially quantified implicit event for which the permit is valid.

Now from (11) and (14) together with the transitivity of '$\rightsquigarrow$' we would like to derive:

$$\mathit{Fear}(\mathit{councillors}, \mathit{violence}) \rightsquigarrow \mathit{Refuse}(\mathit{councillors}, \mathit{demonstrators}, \mathit{permit}) \qquad (15)$$

However, there is a major problem. Our analysis of **WS1a** reveals implicit references to two events: the event that the councillors fear, and the event that is permitted by the permit. We can only infer (15) if these are the same event (or at least must occur together — we could regard the violence as pertaining to an event that is only part of the whole demonstration event). I believe such an assumption is necessary for the resolution of **WS1a** and exemplifies a key mechanism for enabling natural language understanding.

5.2.2. A Default Rule for Entity Identification

One would like to have a general way of establishing identity between different entities referenced either explicitly of implicitly. In the spirit of *Ockham's razor* one can formulate a general rule of default inferences of the following form:

$$\frac{\mathscr{O}, \mathscr{K}, I \vdash \exists x \exists y[\Phi(x,y)] \quad \text{and} \quad \mathscr{O}, \mathscr{K}, I, \exists x[\Phi(x,x)] \not\vdash \bot}{\mathscr{O}, \mathscr{K}, I \vdash \exists x[\Phi(x,x)]}\ \textbf{DEI}$$

This says that, if, from ontology $\mathscr{O}$, with background knowledge $\mathscr{K}$ and some given information I (e.g. a description of some scenario), we can infer the existence of two entities satisfying some relation Φ, and it is also consistent with $\mathscr{O}$ and I that these entities may be the same, then we can (by default) infer that they are the same. For this to give reasonable inferences we would need to ensure that any semantic constraints implied by the context of x and y in Φ are enforced by $\mathscr{O}$ and $\mathscr{K}$. Even with this proviso, the rule may be too strong and it may be very difficult to determine exactly what knowledge should be incorporated within $\mathscr{K}$. Nevertheless, if applied with suitable caution **DEI** may be a useful form of inference for natural language interpretation.

5.2.3. Further Justifications of the Inference

As justification for the pronoun resolution we may seek to find a reason why the councillors would consider a violent demonstration to be bad for them:

- Every city council has responsibility for a city.
- If an agent or organisation a is responsible for some thing x, then x being in a bad condition is bad for a.
- If a violent event occurs in a location it is bad for that location.

These conditions could be represented formally as:

$$\forall x[\mathit{Councilors}(x) \rightarrow \exists y[\mathit{City}(y) \wedge \mathit{Has\_responsibility\_for}(x,y)]\,] \qquad (16)$$

$$\mathsf{Occurs}(e) \wedge \mathsf{Loc}(e, \mathit{loc}) \wedge \mathit{Violent}(e) \rightsquigarrow \mathit{Bad\_condition}(\mathit{loc}) \qquad (17)$$

$$\mathit{Has\_Responsibility\_for}(a,x) \rightarrow \mathit{Bad\_for}(\mathit{Bad\_condition}(x), a) \qquad (18)$$

Again the use of this knowledge in interpreting **WS1a** relies on identification of implied entities: we must assume that the demonstration that the demonstrators are plan-

ning will take place in the *same* city that the councillors are responsible for and also that the requested permit is a for a demonstration to take place in that *same city*.

Another reason why we might want to justify the interpretation of **WS1a** is that demonstrations are the kind of event that is likely to turn violent. However, this does not seem to be necessary for the pronoun resolution. I believe that in the following cases we would still normally interpret 'they' as referring to the councillors:

- The councillors refused the funfair organisers a permit because they feared violence.
- The councillors refused the funfair organisers a permit because they feared Joe Carson would turn up.

It is clear that fear of violence can influence agents' actions in many ways. A couple of other examples that illustrate this diversity are:

- The organisers of the demonstration decided not to apply for a permit, because they feared the event would turn violent. In fact they actually advocated violence, so they didn't want their names on a permit application form.
- The samurai offered to protect the villagers because they feared violence.

6. Conclusions

My investigation of **WS1** has only been partially successful. Despite fairly elaborate semantic analysis, the explanation of the required inferences still has several loose ends. I think it did shed light on what the problems are and how one might go about addressing them, but the methodology can certainly be called into question with respect to its generalisability. KRR approaches require huge amounts of detailed representation of both lexical semantics and world knowledge, so expanding such analysis to everything that could be described in natural language is daunting and may seem infeasible. However, in [12] I and my collaborator have investigated the coverage of a set of rules relating to the verb 'thank' and found that these rules could account for approximately 0.4% of WS examples in the large WinoGrande set [9]. Although this is a small fraction, it does lend credibility to the idea that one could incrementally build up to much greater coverage by adding knowledge domains in a modular way.

One should also consider generality in the types semantic structures and rules that have been identified. Here I believe a strong case can be made. Rules expressing general forms of plausible explanation, such as motivations for an agent carrying out an action seem (as was also observed in [12]) to be transferable to a wide range of scenarios and to many WSC examples. One can also argue that axiomatising the notion of 'reasonable explanation' may, for many purposes, be more effective than trying to give a logical theory of the philosophically problematic concept of *causality*. The need to identify implied entities is is especially salient vor **WS1**. However, my informal examination of WS examples suggestst that around 75% also depend on some kind of entity resolution (in addition to the pronoun resolution) though it is often of a less distinctive form (a very typical case is where two parts of a sentence refer to two aspects of an event).

The current analysis consists of a rather *ad hoc* combination of logical syntax with no explicit semantics. My aim was narrowed to finding a plausible path of inference to account for just one example, but ideally we would prefer a general purpose logical

language with a precise syntax and semantics. Of course, many such frameworks have already been proposed and it is apparent that they have features that address some issues raised in the current paper. For instance Minsky's *Frames* [21] and Schanks' *Scripts* [22]. group together concepts and relationships associated with a particular type of object, situation or event. Hence they would support the representation of 'permit' and its dependent entities in a way that is similar to what I proposed. The *ontologies* used in modern advanced information systems fulfil a similar role to information organisation to structures such as Frames. But, whereas Frames and Scripts typically specify conceptual structures focused around particular types of object or situation, ontology languages such as OWL are more oriented towards specifying abstract relationships between concepts, such as subsumption hierarchies. Both these forms of knowledge appear to be essential to finding coherent interpretations of natural language and especially the problem of identifying implied entities and using them to glue together the parts of a sentence. Frame type organisation of knowledge is good for identifying the auxiliary relationships and entities that surround every concept in every description, and ontologies can specify the categorial constraints on types of entity and possible relations between them that are required for establishing connections between these implied entities.

The method of investigation carried out in the current paper is likely to be significantly enhanced and generalised by incorporating insights and theories from other research that I have more recently become aware of. In particular, work in formal linguistics has developed frameworks such as Dynamic Semantics and Segmented Discourse Representation Theory [23], that provide logical representations that can capture more complex interplay of language features than is possible in the traditional, more static first-order logic. Also the notion of *bridging anaphora* [24] has long been known in the field of language processing overlaps substantially with my idea of resolving identities of implied entities, and algorithms have been developed for finding bridging links [25].

References

[1] Levesque HJ, Davis E, Morgenstern L. The Winograd Schema Challenge. In: Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference; 2012. .

[2] Winograd T. Understanding natural language. Cognitive psychology. 1972;3(1):1-191.

[3] Levesque HJ, Davis E, Morgenstern L. The Winograd Schema Challenge. In: Brewka G, Eiter T, McIlraith SA, editors. Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012. AAAI Press; 2012. .

[4] Rahman A, Ng V. Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In: EMNLP-CoNLL; 2012. .

[5] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805[csCL]. 2018.

[6] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:190711692[csCL]. 2019.

[7] Trichelair P, Emami A, Trischler A, Suleman K, Cheung JCK. How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG. arXiv:181101778[csLG]. 2018.

[8] Kocijan V, Cretu AM, Camburu OM, Yordanov Y, Lukasiewicz T. A Surprisingly Robust Trick for Winograd Schema Challenge. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 48374842.

[9] Sakaguchi K, Bras RL, Bhagavatula C, Choi Y. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. In: AAAI-20; 2020. .

[10] Emami A, Suleman K, Trischler A, Cheung JCK. An Analysis of Dataset Overlap on Winograd-Style Tasks. In: Proceedings of the 28th International Conference on Computational Linguistics; 2020. p. 5855-65.

[11] Ettinger A. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics. 2020;8:34-48.

[12] Hong SJ, Bennett B. Tackling domain-specific winograd schemas with knowledge-based reasoning and machine learning. In: Proceedings of the 3rd Conference on Language, Data and Knowledge (LDK); 2021. .

[13] Bailey D, Harrison A, Lierler Y, Lifschitz V, Michael J. The Winograd Schema Challenge and Reasoning about Correlation. In: 2015 AAAI Spring Symposium Series. USA; 2015. .

[14] Sharma A, Vo NH, Aditya S, Baral C. Towards Addressing the Winograd Schema Challenge - Building and Using a Semantic Parser and a Knowledge Hunting Module. In: IJCAI 2015; 2015. p. 1319-25.

[15] Grice HP. Logic and conversation. In: Speech acts. Brill; 1975. p. 41-58.

[16] Hobbs JR. Coherence and coreference. Cognitive science. 1979;3(1):67-90.

[17] Lapata M, Barzilay R. Automatic evaluation of text coherence: Models and representations. In: IJCAI. vol. 5; 2005. p. 1085-90.

[18] Kehler A, Kertz L, Rohde H, Elman JL. Coherence and coreference revisited. Journal of semantics. 2008;25(1):1-44.

[19] Schnieder B. A logic for 'because'. The Review of Symbolic Logic. 2011;4(3):445-65.

[20] Wierzbicka A. Defining emotion concepts. Cognitive science. 1992;16(4):539-81.

[21] Minsky M. A framework for representing knowledge. MIT-AI Laboratory; 1974. Memo 306.

[22] Schank RC, Abelson RP. Scripts, Plans, and Knowledge. In: Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1. IJCAI'75. Morgan Kaufmann; 1975. p. 151-7.

[23] Asher N, Lascarides A. Logics of Conversation. Cambridge University Press; 2003.

[24] Clark HH. Bridging. In: Theoretical issues in natural language processing; 1975. .

[25] Hou Y, Markert K, Strube M. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 2082-93.