The Winograd Schema Challenge and Reasoning about Correlation

Daniel Bailey
Department of Computer Science
University of Nebraska at Omaha
dbailey@unomaha.edu

Amelia Harrison
Department of Computer Science
University of Texas at Austin
ameliaj@cs.utexas.edu

Yuliya Lierler
Department of Computer Science
University of Nebraska at Omaha
ylierler@unomaha.edu

Vladimir Lifschitz and **Julian Michael**
Department of Computer Science
University of Texas at Austin
{vl, julianjm}@cs.utexas.edu

Abstract

The Winograd Schema Challenge is an alternative to the Turing Test that may provide a more meaningful measure of machine intelligence. It poses a set of coreference resolution problems that cannot be solved without human-like reasoning. In this paper, we take the view that the solution to such problems lies in establishing discourse coherence. Specifically, we examine two types of rhetorical relations that can be used to establish discourse coherence: positive and negative correlation. We introduce a framework for reasoning about correlation between sentences, and show how this framework can be used to justify solutions to some Winograd Schema problems.

1 Introduction

The Winograd Schema Challenge, introduced by Levesque, Davis, and Morgenstern (2012), is an alternative to the Turing Test that may provide a more meaningful measure of machine intelligence. Nuance Communications, Inc. has announced an annual competition based on this task.

The test involves coreference resolution problems of a particular kind. The main part of a Winograd Schema is a sentence containing a pronoun, for instance:

> *The city councilmen refused the demonstrators a permit because they feared violence.*

In addition, two definite noun phrases, called "answers," are given; in the example above, the answers are *the city councilmen* and *the demonstrators*. The goal is to determine which answer provides the most natural resolution for the pronoun. For instance, the natural response to the question *Who feared violence?* is given by the first answer, *the city councilmen*.

A Winograd Schema specifies also a "special word" that occurs in the sentence and an "alternate word." Replacing the former by the latter changes the resolution of the pronoun. In the example above, the special word is *feared* and the alternate word is *advocated*. Thus every schema represents a pair of coreference resolution problems that are almost identical but have different answers. Levesque, Davis, and Morgenstern proposed to assemble a set of Winograd Schemas that are "Google-proof," in the sense that the statistical properties alone of the special word and its alternate would not justify changing the answer when the words are exchanged. To succeed in solving problems of this kind a program would have to use relevant background knowledge—it would have to "think."

In this paper we study several examples of Winograd Schemas from a list compiled by Ernest Davis.[1] In the tradition of (Hobbs 1979) and (Kehler et al. 2008), we treat coreference resolution as a by-product of a general process of establishing discourse coherence: *a resolution for a pronoun is acceptable if it makes the discourse "coherent."* From this perspective, understanding coherence is the key to coreference resolution and, in particular, to the Winograd Schema Challenge.

A study of Davis's list shows that in many cases the coherence of the solution can be explained by correlation between two clauses formed when the correct answer is substituted for the pronoun. For instance, the phrase

> *the city councilmen refused the demonstrators a permit*

is positively correlated with

> *the city councilmen feared violence*

in the sense that either one of them would cause the hearer to view the other as more plausible than before. Similarly, that phrase is correlated with

> *the demonstrators advocated violence.*

The term "correlation" usually refers to correlation between random variables as defined in probability theory. In this paper, we concentrate on the "doxastic" aspect of correlation—on its effect on a person's beliefs, as in the examples above.

As another example, consider the sentence from Schema 3 on Davis's list:

> *Joan made sure to thank Susan for all the help she had given.*

Who had given help? The coherence of the answer *Susan* can be explained by correlation between the phrases

> *Joan made sure to thank Susan*

[1] `http://www.cs.nyu.edu/faculty/davise/papers/WS.html`. The example above is Schema 1 on Davis's list.

and

Susan had given help.

These phrases are correlated in the sense that either one would cause the hearer to view the other as more plausible.

In this paper we propose a deductive system for deriving formulas expressing correlation, the "correlation calculus," and show how this system can be used to justify the answers to several examples from Davis's collection. We conjecture that the correlation calculus, in combination with appropriate axioms, can be used to justify solutions to many Winograd Schemas. Obtaining such axioms is a major problem, of course, and is an important avenue of future work (Section 5).

The probabilistic definition of correlation does not seem directly applicable to the examples above: it is not clear how to understand the probability of city councilmen refusing demonstrators a permit, or the probability of Joan thanking Susan. Nevertheless, our correlation calculus turned out to be closely related to correlation in the sense of probability theory: an event A is correlated with an event B if the conditional probability $P(A|B)$ is greater than the probability $P(A)$. We show that this idea leads to a mathematically precise semantics for correlation formulas. The correlation calculus is sound with respect to this semantics, and this fact is useful as a technical tool for proving properties of the correlation calculus.

Consider now Schema 20 with the special word *because* replaced by the alternate word *although:*

Pete envies Martin although he is very successful.

Who is very successful? The coherence of the answer *Pete* can be explained by "negative correlation" between the phrases

Pete envies Martin

and

Pete is very successful

—the former is correlated with the negation of the latter. The idea of positive correlation is often relevant when the given sentence contains the discourse connective *because*, as in the example with city councilmen. The discourse connective *although* in a sentence points to negative correlation.

We annotated the first 100 Winograd Schema sentences from Davis's list. Out of these 100 sentences, 64 exhibit positive correlation and 8 exhibit negative correlation. Thus the analysis that we develop in this paper may be applied to 72 Winograd Schema sentences out of 100 annotated examples.

Our use of positive and negative correlation is in the spirit of Asher and Lascarides (2003), who propose to establish the coherence of a discourse with respect to a rhetorical relation which connects the discourse parts. Whereas we distinguish between positive and negative correlation, they distinguish between several classes of rhetorical relation: elaboration, explanation, contrast, and others.

2 Correlation Calculus

2.1 Correlation Formulas

We begin with a signature in the sense of first-order logic. A *correlation formula* is an expression of the form $F \oplus G$, where F and G are first-order formulas. Informally speaking, if F and G are sentences, that is, have no free variables, then $F \oplus G$ expresses that (natural language texts corresponding to) F and G are correlated.[2] Free variables in a correlation formula are, informally speaking, understood as metavariables for arbitrary ground terms.

The expression $F \ominus G$ is shorthand for $F \oplus \neg G$.

2.2 Inference Rules

We now introduce the inference rules of the correlation calculus: implication, replacement, symmetry, negation, and substitution. The conclusion of each rule is a correlation formula. Both first-order sentences and correlation formulas can be used as premises.

The *implication rule* says that if a formula F implies a formula G then F and G are correlated:

$$\frac{\widetilde{\forall}(F \rightarrow G)}{F \oplus G}.$$

(The symbol $\widetilde{\forall}$ denotes universal closure.) The two *replacement rules* allow us to replace one side of a correlation formula with an equivalent formula:

$$\frac{\widetilde{\forall}(F \leftrightarrow G) \qquad F \oplus H}{G \oplus H},$$

$$\frac{\widetilde{\forall}(F \leftrightarrow G) \qquad H \oplus F}{H \oplus G}.$$

The *symmetry rule* expresses that $\oplus$ is symmetric:[3]

$$\frac{F \oplus G}{G \oplus F}.$$

According to the *negation rule*, the negations of two correlated formulas are correlated as well:

$$\frac{F \oplus G}{\neg F \oplus \neg G}.$$

The *substitution rule* allows us to substitute terms for free variables:

$$\frac{F \oplus G}{F\theta \oplus G\theta}$$

for any substitution θ of terms for variables.[4]

A *derivation* from a set Γ consisting of first-order sentences and correlation formulas is a list $C_1, \ldots, C_n$ such that

- if C_i is a first-order sentence then it is entailed by Γ, and
- if C_i is a correlation formula then it belongs to Γ or can be derived from one or two of the formulas that precede it in the list by one of the rules of the correlation calculus.

A correlation formula C is *derivable from* Γ if there exists a derivation from Γ with C as the last formula.

[2] See Section 2.4 for a more precise formulation.

[3] In the presence of the symmetry rule, any one of the replacement rules can be dropped.

[4] Here $F\theta$ stands for the result of applying θ to all free variables in F, with bound variables renamed if necessary to avoid quantifier capture.

2.3 Examples of Derivations

The formula $F \oplus F$ is derivable from the empty set of formulas:

1.	$\widetilde{\forall}(F \rightarrow F)$	logically valid
2.	$F \oplus F$	by implication rule.

The formula $P(a) \oplus \exists x P(x)$ is derivable from the empty set of formulas:

1.	$P(a) \rightarrow \exists x P(x)$	logically valid
2.	$P(a) \oplus \exists x P(x)$	by implication rule.

The formula $P(a) \oplus \forall x P(x)$ is derivable from the empty set of formulas:

1.	$\forall x P(x) \rightarrow P(a)$	logically valid
2.	$\forall x P(x) \oplus P(a)$	by implication rule
3.	$P(a) \oplus \forall x P(x)$	by symmetry rule.

The formula $G \ominus F$ is derivable from $F \ominus G$:

1.	$F \oplus \neg G$	
2.	$\neg G \oplus F$	by symmetry rule
3.	$\neg\neg G \oplus \neg F$	by negation rule
4.	$\widetilde{\forall}(\neg\neg G \leftrightarrow G)$	logically valid
5.	$G \oplus \neg F$	by replacement rule.

2.4 The Case of Complete Information

Recall that a formula $F \oplus G$ without free variables is meant to express that F and G are correlated in the sense that the message F would cause the hearer to view G as more plausible, and the message G would cause the hearer to view F as more plausible. The correlation calculus is expected to have the property that whenever $F \oplus G$ is derivable from the set of facts known to the hearer, F and G are correlated.

This is not always true, however. If the hearer knows with certainty whether F is true or false then no message can make F more plausible. Consequently, if one of the formulas F, $\neg F$ belongs to Γ then we would like $F \oplus G$ not to be derivable from Γ (and similarly if G or $\neg G$ belongs to Γ). The actual situation is exactly opposite. For any first-order formulas F and G, the correlation formula $F \oplus G$ is derivable from $\widetilde{\forall} F$:

1.	$\widetilde{\forall} F$	
2.	$\widetilde{\forall}(G \rightarrow F)$	entailed by 1
3.	$G \oplus F$	by implication rule
4.	$F \oplus G$	by symmetry rule.

It is derivable from $\widetilde{\forall}\neg F$ as well:

1.	$\widetilde{\forall}\neg F$	
2.	$\widetilde{\forall}(F \rightarrow G)$	entailed by 1
3.	$F \oplus G$	by implication rule.

To take into account these observations, we modify the interpretation of $\oplus$ to treat the case of complete information in a special way. A correlation formula $F \oplus G$ without free variables will be interpreted to mean that

F is correlated with G
or
at least one of F, G, $\neg F$, $\neg G$ is known to be true.

In Section 4 we give a mathematically precise definition of a probabilistic counterpart of this extended interpretation of $\oplus$ and prove the soundness of the correlation calculus relative to this probabilistic semantics.

3 Justifying Solutions to Winograd Schema Problems

To justify the correctness of a proposed solution to a Winograd Schema problem we encode the solution as a correlation formula in which discourse referents are represented by object constants. We then show how this correlation formula can be derived from assumptions of two kinds: ground atoms expressing the presuppositions of the discourse referents and formulas expressing relevant facts from commonsense knowledge.

3.1 The Trophy and the Suitcase

Consider the sentence from Schema 2 on Davis's list:

> *The trophy doesn't fit into the brown suitcase because it's too small.*

What is too small?

We will justify the correctness of the answer *the suitcase* as follows. The phrase *the trophy doesn't fit into the brown suitcase* can be represented by the sentence $\neg \mathit{fit\_into}(T, S)$, with the presuppositions

1. $\mathit{trophy}(T)$
2. $\mathit{suitcase}(S)$
3. $\mathit{brown}(S)$.

The phrase *the suitcase is too small* can be represented by the sentence $\mathit{small}(S)$. We will show that the correlation formula

$$\neg \mathit{fit\_into}(T, S) \oplus \mathit{small}(S) \qquad (1)$$

can be derived in the correlation calculus from the presuppositions 1–3 in combination with the following commonsense facts, which are assumed to be known to the hearer:

4. $\forall x(\mathit{suitcase}(x) \rightarrow \mathit{physical\_object}(x))$
5. $\forall x(\mathit{physical\_object}(x) \rightarrow (\mathit{small}(x) \leftrightarrow \neg \mathit{large}(x)))$
6. $\mathit{fit\_into}(x, y) \oplus \mathit{large}(y)$.

The derivation continues as follows:

7.	$\neg \mathit{large}(S) \leftrightarrow \mathit{small}(S)$	entailed by 2, 4, and 5
8.	$\mathit{fit\_into}(T, S) \oplus \mathit{large}(S)$	by substitution from 6
9.	$\neg \mathit{fit\_into}(T, S) \oplus \neg \mathit{large}(S)$	by negation rule
10.	$\neg \mathit{fit\_into}(T, S) \oplus \mathit{small}(S)$	by replacement rule.

In Axiom 6 above, x and y are, intuitively, physical objects, and moreover y is a container. Our formulation may seem too strong because it does not incorporate these assumptions about the values of x and y. Note, however, that when x and y are not objects of appropriate types, the corresponding instance of Axiom 6 holds because at least one of the conditions $\mathit{fit\_into}(x, y)$, $\mathit{large}(y)$ is known to be false (see Section 2.4).

The derivability of (1) shows that

$\neg\mathit{fit\_into}(T,S)$ is correlated with *small(S)*
or
at least one of $\mathit{fit\_into}(T,S)$, *small(S)*
is known to be true or false.

Since the formulas $\mathit{fit\_into}(T,S)$ and *small(S)* are not known to be true or false, we can conclude that the sentences $\neg\mathit{fit\_into}(T,S)$ and *small(S)* are indeed correlated.

Consider now the same example with the special word *small* replaced by the alternate word *big*:

> *The trophy doesn't fit into the brown suitcase because it's too big.*

What is too big? The correctness of the answer *the trophy* can be justified in a similar way, with Axiom 6 replaced by the axiom

$$\mathit{fit\_into}(x,y) \oplus \mathit{small}(x).$$

3.2 Lifting the Son

Consider schema 8 from Davis's list:

> *The man couldn't lift his son because he was so weak.*

Who was so weak? We will justify the answer *the man* by deriving the formula

$$\neg\mathit{can}(M,\mathit{lift}(S)) \oplus \mathit{weak}(M)$$

from the presuppositions

1. $man(M)$
2. $\mathit{son}(S)$

and the commonsense facts

3. $\forall x(\mathit{man}(x) \rightarrow \mathit{person}(x))$
4. $\forall x(\mathit{person}(x) \rightarrow (\mathit{strong}(x) \leftrightarrow \neg\mathit{weak}(x)))$
5. $\mathit{can}(x,y) \oplus \mathit{strong}(x)$.

The derivation continues:

6. $\neg\mathit{strong}(M) \leftrightarrow \mathit{weak}(M)$ entailed by 1, 3, 4
7. $\mathit{can}(M,\mathit{lift}(S)) \oplus \mathit{strong}(M)$ by substitution from 5
8. $\neg\mathit{can}(M,\mathit{lift}(S)) \oplus \neg\mathit{strong}(M)$ by negation
9. $\neg\mathit{can}(M,\mathit{lift}(S)) \oplus \mathit{weak}(M)$ by replacement.

Consider now the same example with the special word *weak* replaced by the alternate word *heavy*:

> *The man couldn't lift his son because he was so heavy.*

Who was so heavy? The answer *the son* can be justified by deriving the formula

$$\neg\mathit{can}(M,\mathit{lift}(S)) \oplus \mathit{heavy}(S)$$

from the commonsense axiom

$$\mathit{can}(x,\mathit{lift}(y)) \ominus \mathit{heavy}(y),$$

which is shorthand for

$$\mathit{can}(x,\mathit{lift}(y)) \oplus \neg\mathit{heavy}(y)$$

(see Section 2.1).

3.3 The Sculpture on the Shelf

Take now the sentence from Schema 13:

> *The sculpture rolled off the shelf because it wasn't anchored.*

What wasn't anchored? We will justify the answer *the sculpture* by showing that the correlation formula

$$\mathit{roll\_off}(Sc,Sh) \oplus \neg\mathit{anchored}(Sc)$$

can be derived from the presuppositions

1. $\mathit{shelf}(Sh)$
2. $\mathit{sculpture}(Sc)$

and the commonsense facts

3. $\forall x(\mathit{shelf}(x) \rightarrow \mathit{surface}(x))$
4. $\forall x(\mathit{sculpture}(x) \rightarrow \mathit{physical\_object}(x))$
5. $\forall xy(\mathit{physical\_object}(x) \wedge \mathit{surface}(y) \wedge \mathit{anchored}(x) \rightarrow \neg\mathit{roll\_off}(x,y))$.

The derivation continues:

6. $\mathit{roll\_off}(Sc,Sh) \rightarrow \neg\mathit{anchored}(Sc)$ entailed by 1–5
7. $\mathit{roll\_off}(Sc,Sh) \oplus \neg\mathit{anchored}(Sc)$ by implication.

Consider the same example with the special word *anchored* replaced by the alternate word *level*:

> *The sculpture rolled off the shelf because it wasn't level.*

What wasn't level? The answer *the shelf* can be justified by deriving the formula

$$\mathit{roll\_off}(Sc,Sh) \ominus \mathit{level}(Sh)$$

from axioms 1–4 and the additional commonsense fact

$$\forall xy(\mathit{physical\_object}(x) \wedge \mathit{surface}(y) \wedge \mathit{level}(y) \rightarrow \neg\mathit{roll\_off}(x,y)).$$

4 A Probabilistic Semantics of Correlation Formulas

In this section we assume that the underlying signature is finite and contains at least one object constant but no function symbols of arity greater than zero. Under these assumptions, the set $\mathbf{I}$ of Herbrand interpretations of the signature is finite.[5]

4.1 Worldviews and Satisfaction

A *worldview* is a discrete probability distribution over $\mathbf{I}$ (a function assigning a value in $[0,1]$ to each interpretation, so that the sum over $\mathbf{I}$ is 1). Every first-order sentence defines an event—the set of its models—so we may talk about its probability with respect to a worldview D:

$$P(F) = \sum_{I \in \mathbf{I}\,:\, I \models F} D(I).$$

[5] Since our formalization of the Lifting the Son example uses the function symbol *lift*, it is not covered by these semantics.

Similarly, we can talk about the probability of a set Γ of first-order sentences with respect to D:

$$P(\Gamma) = \sum_{I \in \mathbf{I}\,:\, I \models \Gamma} D(I).$$

A *correlation sentence* is a correlation formula with no free variables. *Satisfaction* is defined as follows: for any worldview D,

(i) D satisfies a first-order sentence F if

$$P(F) = 1,$$

(ii) D satisfies a correlation sentence $F \oplus G$ if the inequality

$$P(F \wedge G) > P(F)P(G)$$

holds, or D satisfies at least one of the sentences

$$F, \neg F, G, \neg G,$$

(iii) D satisfies a correlation formula $F \oplus G$ with free variables if D satisfies $F\theta \oplus G\theta$ for every substitution θ that maps the free variables to object constants,

(iv) D satisfies a set consisting of correlation formulas and first-order sentences if it satisfies all elements of the set.

Case (i) of the definition above is a generalization of the usual definition of satisfaction for first-order sentences: for any worldview D that assigns probability 1 to an interpretation I, D satisfies a formula F if and only if I satisfies F.

Furthermore, if a worldview satisfies a set Γ of first-order sentences then $P(\Gamma) = 1$. Indeed, the set of interpretations satisfying Γ is the intersection of the events $\{I \in \mathbf{I} \,:\, I \models F\}$ for all F in Γ. The probability of each of these events is 1, and there are finitely many of them.

In the case when $P(G) > 0$, the inequality in clause (ii) of the definition of satisfaction can be rewritten in terms of conditional probabilities; D satisfies $F \oplus G$ if

$$P(F|G) > P(F).$$

Let C be a correlation formula or first-order sentence. A set Γ of first-order sentences and correlation formulas *entails* C if every worldview that satisfies Γ satisfies C. In the special case when C and all elements of Γ are first-order sentences, this definition is equivalent to the usual definition in first-order logic. Indeed, it is clear that if every worldview satisfying Γ satisfies C then every interpretation satisfying Γ satisfies C. In the other direction, assume that every interpretation satisfying Γ satisfies C. Then $P(C) \geq P(\Gamma)$. If a worldview D satisfies Γ then $P(\Gamma) = 1$, so that $P(C) = 1$. Thus Γ entails C in the sense of the definition above.

4.2 Soundness of the Correlation Calculus

Consider an inference rule such that in its instances

$$\frac{C_1 \ \cdots \ C_n}{C_{n+1}} \tag{2}$$

$C_1, \ldots, C_{n+1}$ are either correlation formulas or first-order sentences. Such a rule is *sound* if for each of its instances (2) C_{n+1} is entailed by $C_1, \ldots, C_n$.

Soundness Theorem. *All rules of the correlation calculus are sound.*

Corollary. *If C is derivable from Γ in the correlation calculus then Γ entails C.*

As an example of the use of the probabilistic semantics, consider a first-order signature allowing only two distinct ground atoms p, q. We will show that the correlation formula $p \oplus q$ is not derivable from the empty set. Consider the worldview that assigns the same value $\frac{1}{4}$ to each of the 4 interpretations of this signature. This worldview does not satisfy $p \oplus q$, because

$$P(p \wedge q) = P(p) \cdot P(q).$$

Indeed, $P(p \wedge q) = \frac{1}{4}$ and $P(p) = P(q) = \frac{1}{2}$.

4.3 Proof of the Soundness Theorem

The implication rule. Consider first an instance of the implication rule where F and G are sentences:

$$\frac{F \rightarrow G}{F \oplus G}.$$

If either $P(F) = 0$ or $P(G) = 1$ then the fact that D satisfies the conclusion is immediate. Otherwise, let D be a worldview that satisfies the premise. Then for any interpretation I to which D assigns nonzero probability,

$$I \models F \rightarrow G.$$

So for any such interpretation, if $I \models F$ then $I \models G$. Consequently,

$$P(F \wedge G) = P(F).$$

On the other hand, since $P(F) > 0$ and $P(G) < 1$,

$$P(F) > P(F)P(G).$$

Consequently

$$P(F \wedge G) > P(F)P(G),$$

so that D satisfies the conclusion.

Consider now an instance of the implication rule in which F and G may contain free variables:

$$\frac{\widetilde{\forall}(F \rightarrow G)}{F \oplus G}. \tag{3}$$

Let D be a worldview that satisfies the premise. We need to show that for any substitution θ that maps the free variables to object constants, D satisfies $F\theta \oplus G\theta$. Consider the following instance of the implication rule:

$$\frac{F\theta \rightarrow G\theta}{F\theta \oplus G\theta}. \tag{4}$$

Since D satisfies the premise of (3), it satisfies the premise of (4). Since (4) is covered by the special case discussed earlier, it follows that D satisfies $F\theta \oplus G\theta$.

The replacement rule. Consider first an instance of the replacement rule where F, G, H are sentences:

$$\frac{F \leftrightarrow G \qquad F \oplus H}{G \oplus H}.$$

Let D be a worldview that satisfies both premises. Since D satisfies the first premise, it assigns probability 0 to interpretations satisfying $F \wedge \neg G$ and to interpretations satisfying $\neg F \wedge G$. Consequently,

$$\begin{aligned} P(F) &= P(F \wedge G) + P(F \wedge \neg G) \\ &= P(F \wedge G) \\ &= P(F \wedge G) + P(\neg F \wedge G) \\ &= P(G), \end{aligned}$$

so that

$$P(F) = P(G). \tag{5}$$

Similarly,

$$P(F \wedge H) = P(G \wedge H). \tag{6}$$

Since D satisfies the second premise, two cases are possible:

$$P(F \wedge H) > P(F)P(H)$$

or one of the probabilities $P(F), P(H)$ is 0 or 1. In the first case, by (5) and (6),

$$P(G \wedge H) = P(F \wedge H) > P(F)P(H) = P(G)P(H).$$

In the second case, in view of (5), one of the probabilities $P(G), P(H)$ is 0 or 1. In either case, D satisfies the conclusion.

The general case follows as in the proof for the implication rule. The proof for the other replacement rule is analogous.

The soundness of the symmetry rule is obvious.

The negation rule. Consider first an instance of the negation rule where F, G are sentences:

$$\frac{F \oplus G}{\neg F \oplus \neg G}.$$

Let D be a worldview satisfying the premise, so that either

$$P(F \wedge G) > P(F)P(G)$$

or one of the probabilities $P(F)$, $P(G)$ is 0 or 1. In the first case,

$$\begin{aligned} P(\neg F \wedge \neg G) &= P(\neg(F \vee G)) \\ &= 1 - P(F \vee G) \\ &= 1 - P(F) - P(G) + P(F \wedge G) \\ &> 1 - P(F) - P(G) + P(F)P(G) \\ &= (1 - P(F))(1 - P(G)) \\ &= P(\neg F)P(\neg G), \end{aligned}$$

so D satisfies the conclusion. In the second case, one of the probabilities $P(\neg F)$, $P(\neg G)$ is 0 or 1, and consequently D satisfies the conclusion.

The general case follows as above.

The substitution rule. Consider an instance of the substitution rule

$$\frac{F \oplus G}{F\theta \oplus G\theta},$$

and let D be a worldview satisfying the premise. For any substitution θ' that maps the free variables of $F\theta$, $G\theta$ to object constants, D satisfies

$$(F\theta)\theta' \oplus (G\theta)\theta'$$

because $(F\theta)\theta' = F(\theta\theta')$, $(G\theta)\theta' = G(\theta\theta')$.

4.4 An Unsound Inference Rule

An inference rule that may look plausible is the *transitivity rule*

$$\frac{F \oplus G \qquad G \oplus H}{F \oplus H}.$$

However, it is unsound. Indeed, consider a signature with two distinct ground atoms p, q, and the following instance of the transitivity rule:

$$\frac{p \oplus p \vee q \qquad p \vee q \oplus q}{p \oplus q}.$$

The two premises are both entailed by the empty set, as they are derivable by applying the implication rule and symmetry rule to the tautologies

$$\begin{aligned} &p \rightarrow (p \vee q), \\ &q \rightarrow (p \vee q). \end{aligned}$$

However, we have already shown that $p \oplus q$ is not entailed by the empty set, so the premises do not entail the conclusion and the transitivity rule is unsound.

5 Future Work

Our approach to reasoning about correlation relies on the availability of axioms expressing relevant commonsense knowledge. The axioms used in Section 3 are manually tailored to handle examples from Davis's list, and we would like to enerate such axioms automatically.

This difficult task can benefit from existing work on formalizing commonsense knowledge (Davis 1990; Mueller 2006). It may be possible to extract useful axioms from existing lexical and commonsense knowledge bases, such as WORDNET (Fellbaum 1998), FRAMENET (Baker, Fillmore, and Lowe 1998), VERBNET (Kipper-Schuler 2005), PROPBANK (Palmer, Gildea, and Kingsbury 2005), CONCEPTNET (Liu and Singh 2004), KNEXT (Schubert 2002), and the OPENCYC[6] project. For example, the axiom

$$\forall x(\textit{suitcase}(x) \rightarrow \textit{physical\_object}(x))$$

used in Section 3.1 is supported by the knowledge available in the WORDNET database stating that a *suitcase* is a hyponym of a *physical object*.

The automation of reasoning in the correlation calculus is another issue that will need to be studied.

Some commonsense facts expressed in Section 3 by first-order sentences can be best formalized as defaults. The formula

$$\forall xy(\textit{level}(y) \rightarrow \neg \textit{roll\_off}(x, y))$$

is an example: an earthquake can cause a sculpture to roll off a shelf even when the shelf is level. Developing a nonmonotonic approach to reasoning about correlation that will make this possible is another avenue for future work.

While we demonstrated the correlation calculus on problems in the Winograd Schema Challenge, it is possible that the approach can be expanded to other tasks relying on discourse coherence. Examples may include general coreference resolution (Kehler et al. 2008), temporal anaphora resolution (Lascarides and Asher 1993), and lexical disambiguation (Asher and Lascarides 1995).

[6] `http://www.opencyc.org/doc`

6 Related Work

The Winograd Schema Challenge is a particular restricted form of coreference resolution. There is an extensive body of work on this topic (Poesio, Ponzetto, and Versley 2010; Ng 2010). The CoNLL-2011 shared task was devoted to coreference resolution, and eighteen systems participated. The OntoNotes (Hovy et al. 2006) data was used to assess (and sometimes develop) these systems. OntoNotes consists of various genres of non-handcrafted text, as well as annotations of coreference. However, Winograd Schema problems turn out to be difficult for state-of-the-art coreference systems. For example, the top ranked Stanford system at the CoNLL-2011 shared task (Lee et al. 2011) made a mistake on 50% of the Winograd Schema problems that we experimented with. The other top-ranked systems performed similarly.

Rahman and Ng (2012) address coreference resolution problems like those in the Winograd Schema Challenge using a ranking-based machine learning approach. Schüller (2014) relates the Winograd Schema Challenge to relevance theory. Schüller's approach is closer to our own in that its focus is knowledge representation.

7 Conclusion

In this paper, we introduced and studied a calculus for deriving correlation formulas, and showed how it can be used to justify correct answers to some Winograd Schema questions. Designing the correlation calculus is only a small step towards meeting the Winograd Schema Challenge, because it sidesteps the difficult problem of generating axioms expressing the relevant commonsense and lexical information, including facts about discourse coherence. However, the correlation calculus is a stand-alone contribution to the study of discourse coherence, and may have implications for other computational linguistics tasks that rely on coherence.

Acknowledgments

Thanks to John Beavers, Ernest Davis, Selim Erdoğan, Katrin Erk, Dan Garrette, Michael Gelfond, Hector Levesque, Karl Pichotta, Peter Schüller, Ben Susman, Jesse Thomason, and the anonymous referees for useful comments. Daniel Bailey was supported by a FIRE: Fund for Investing in the Research Enterprise Grant of the University of Nebraska at Omaha. Amelia Harrison and Vladimir Lifschitz were partially supported by the National Science Foundation under Grant IIS-1422455.

References

Asher, N., and Lascarides, A. 1995. Lexical disambiguation in a discourse context. *Journal of semantics* 12(1):69–108.

Asher, N., and Lascarides, A. 2003. *Logics of Conversations*. Cambridge University Press.

Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1 (ACL/COLING)*, 86–90.

Davis, E. 1990. *Representations of Commonsense Knowledge*. Morgan Kaufmann.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Hobbs, J. 1979. Coherence and coreference. *Cognitive Science* 3:67–90.

Hovy, E.; Marcus, M.; Palmer, M.; Ramshaw, L.; and Weischedel, R. 2006. Ontonotes: The 90% solution. In *Proceedings of Human Language Technology Conference of the NAACL*.

Kehler, A.; Kertz, L.; Rohde, H.; and Elman, J. L. 2008. Coherence and coreference revisited. *Journal of Semantics*.

Kipper-Schuler, K. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. Dissertation, University of Pennsylvania.

Lascarides, A., and Asher, N. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy* 16(5):437–493.

Lee, H.; Peirsman, Y.; Chang, A.; Chambers, N.; Surdeanu, M.; and Jurafsky, D. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task (CoNLL)*.

Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The Winograd schema challenge. In *Proceedings of International Conference on Principles of Knowledge Representation and Reasoning (KR)*.

Liu, H., and Singh, P. 2004. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal* 22:211–226.

Mueller, E. 2006. *Commonsense reasoning*. Elsevier.

Ng, V. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1396–1411.

Palmer, M.; Gildea, D.; and Kingsbury, P. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–106.

Poesio, M.; Ponzetto, S. P.; and Versley, Y. 2010. Computational models of anaphora resolution: A survey.

Rahman, A., and Ng, V. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, 777–789.

Schubert, L. 2002. Can we derive general world knowledge from texts? In *Proceedings of the Second International Conference on Human Language Technology Research (HLT)*, 94–97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Schüller, P. 2014. Tackling Winograd schemas by formalizing relevance theory in knowledge graphs. In *Proceedings of International Conference on Principles of Knowledge Representation and Reasoning (KR)*.