On Resolving Ambiguous Anaphoric Expressions in Imperative Discourse

Vasanth Sarathy
Department of Computer Science
Tufts University
200 Boston Ave.
Medford, MA 02155
vasanth.sarathy@tufts.edu

Matthias Scheutz
Department of Computer Science
Tufts University
200 Boston Ave.
Medford, MA 02155
matthias.scheutz@tufts.edu

Abstract

Anaphora resolution is a central problem in natural language understanding. We study a subclass of this problem involving object pronouns when they are used in simple imperative sentences (e.g., "pick *it* up."). Specifically, we address cases where situational and contextual information is required to interpret these pronouns. Current state-of-the art statistically-driven coreference systems and knowledge-based reasoning systems are insufficient to address these cases. In this paper, we introduce, with examples, a general class of situated anaphora resolution problems, propose a proof-of-concept system for disambiguating situated pronouns, and discuss some general types of reasoning that might be needed.

Introduction

Anaphors are linguistic referring expressions whose interpretation depends on objects and entities introduced earlier in a discourse (Mitkov 2014). For example, in the sentence: "pick up the parcel and give it to me," the pronoun "it" is an anaphor that relies on its antecedent "the parcel" for its meaning; both mentions likely pointing to the same real-world parcel. Pronouns (and anaphors more generally) are used extensively in dialogue and discourse and there are often multiple antecedent candidates for an anaphor, leading to ambiguity, a problem that humans handle quite gracefully (Grosz and Sidner 1986).

Naturally, artificial agents equipped with natural language capabilities must also be able to resolve these ambiguous expressions, if they are to perform commands issued to them, whether it be a home kitchen helper robot or a voice-activated personal assistant (Tellex et al. 2013). Contemporary approaches to anaphora resolution aim at solving the sister-problem of coreference resolution, which requires linking different mentions that reference the same entity or object (Ng 2010; Mitkov 2014). The state of the art methods are statistically driven machine learners that have learned these associations from expansive datasets (Clark and Manning 2016; Wiseman et al. 2015). Unfortunately, when extra-linguistic information is needed, a situation all too common in imperative task-oriented dialog, these systems are much less accurate. Consider the simple example of the following command to a robot to move blocks shown in Figure 1:

Figure 1: Consider instructing a robot the following: *"Pick up block1. Put it on block2. Pick up block3. Put* ***it*** *on block1."* Which block does the second "it" refer to? This question is trivial for humans, but not so for many artificial systems.

Pick up $[block1]_{b1}$*. Put* $\underline{\textbf{it}}_{b1}$ *on* $[block2]_{b2}$*. Pick up* $[block3]_{b3}$*. Put* $\underline{\textbf{it}}_{b1,b2,b3}$ *on block1."* (1)

Many current coreference systems do not resolve (1) because it is not the statistical relationships that undergirds disambiguation, but instead extra-linguistic knowledge. Recognizing the importance of commonsense knowledge (Cassimatis 2008), Winograd Schemas (WS) were proposed as a gold standard evaluation dataset for these types of more difficult problems (Levesque, Davis, and Morgenstern 2012; Morgenstern and Ortiz Jr 2015). Here's an example:

"The sculpture rolled off the shelf because $[it]_{sc,sh}$ *wasn't [anchored/level]."* (2)

Resolving "it" in (2) requires at least some commonsense knowledge about the physics of cause-and-effect, gravity and friction. A number of computational approaches have been proposed that formalize such knowledge. and then use it for disambiguating pronouns. So these systems are likely to be able to resolve (1), however, in many realistic settings, what is needed is an ability to be able to reason with situation-specific knowledge that can change dynamically. Consider these three examples, all situated in a kitchen where a robot is assisting a human with a cooking task:

"Pick up the $[knife]_k$*. Cut the* $[tomato]_t$*. Put* $\underline{\textbf{it}}_{k,t}$ *down."* (3)

"Pick up the $[knife]_k$*. Cut the* $[tomato]_t$*. Put* $\underline{\textbf{it}}_{k,t,b}$ *in the* $[bowl]_b$*."* (4)

[Speaker context - A:Washing dishes, B: Cooking]
"Pick up the $[knife]_k$*. Cut the* $[tomato]_t$*. Pass* $\underline{\textbf{it}}_{k,t}$ *to me."* (5)

In each of these cases, the disambiguation of "it" requires one to consider not just the statistical relationships (like those inferred by the coreference systems) or static and timeless bits of commonsense knowledge (like those used in tackling WSs), but contextual information available to an agent that is situated and embodied in an environment. It is unclear what types of knowledge or reasoning capabilities are needed. In the most general case, the problem is very hard and been the subject of research for decades (Hobbs 1978; Lappin and Leass 1994; Winograd 1980). However, in the narrower case of imperative dialogue we can simplify the problem by focusing on the cause and effect relationships associated with *performing* actions issued by the speaker.

The goal of this paper is to unpack this problem of situated or embodied anaphora resolution. We focus on object pronouns (like "it") as used in imperative utterances within a larger discourse. We view natural language comprehension as a incremental model-building and generative process (Kamp 1981) in which the listener must either perform (or simulate) the issued actions thereby changing the surrounding world. In doing so, resolving anaphors becomes a task of associating actions with its parameters in a way that "makes sense" in this unfolding narrative. Specifically, the contributions of this paper are:

1. **(Problem Characterization)** We introduce the general class of situated anaphor resolution problems in imperative discourse. We characterize these problems by providing a set of exemplary problems and some insights into what makes them particularly special and distinct.
2. **(Proof of Concept)** We construct a proof of concept system using Answer Set Programming and Dempster-Shafer theory for solving this class of problems. The system can resolve the ambiguous anaphors in (3), (4) and (5). We present a detailed walk-through for (5).
3. **(Reasoning Characterization)** We articulate some general and domain-independent types of reasoning as well as architectural capabilities needed to solve these problems.

Solving Situated Anaphora Problems

Overview of a Proof-of-Concept System

To solve situated anaphora problems, a listener agent must reason about extra-linguistic information obtained as a result of its embodiment (i.e., sensory-motor and bodily capabilities of the agent) and its situatedness (agents interactions in context with its environment, which includes other agents). The agent's decision making is guided by the mutual knowledge shared with its interactants, which in turn is influenced by the agent's own capabilities, expectations of its interactants and general normative expectations of the society in which the agent is situated (Clark and Marshall 1981).

We propose that when tasked with disambiguating an anaphoric object pronoun as part of an imperative (e.g., "pass $\text{it}_{\text{knife,tomato}}$ to me."), the agent must bind an **action** to one of a set of two or more **object candidates**. To do so, it must reason about *three different aspects* over and above syntactic considerations, which together form the mutual knowledge, namely:

1. **Plausibility**: *Can* it perform the desired action on an object candidate?
2. **Normative**: *Should* it perform the desired action on an object candidate?
3. **Speaker Intent**: Is the speaker *intending* for it to perform the desired action on an object candidate?

We formalize these notions by suggesting that these three aspects or reasoning modes can be structured as microtheories and represented as answer set programs. Reasoning within these microtheories can happen in parallel with each reasoner returning uncertainty measures for each object candidate. We propose then combining uncertain evidence obtained from these theories using Belief-theoretic notions of evidence combination. Belief theory (a subset of which is Demspter-Shafer theory) generalizes Bayesian probability theory and provides some unique advantages over Bayesian updates to modeling epistemic and subjective uncertainty. Moreover, it has a rich history of application in sensor-fusion networks, which the proposed proof of concept system is modeled after.

In the next section, we walk through a demonstrative example in more detail. But, first, we provide some background on Answer Set Programming and Dempster-Shafer theory and provide some intuition for why they might be suitable frameworks for resolving situated anaphors.

Preliminaries

Answer Set Programming. Answer Set Programming (ASP) is a knowledge representation language useful for commonsense reasoning, especially in presence of incomplete information, defaults, exceptions and inductive definitions (Baral 2003). A logic program Π is a set of rules of the form:

$$L_0|\ldots|L_k \leftarrow L_{k+1},\ldots,L_m, \textbf{not } L_{m+1},\ldots,\textbf{not } L_n$$

Where L_is are literals in the sense of classical logic and the **not** represents *negation-as-failure*. The left and right hand sides of the rule are called the *head* and *body* of the rule, respectively. Either one of (or both) head and body can be empty. When the head is empty, i.e., $k = 0$ and the $L_0 = \bot$ the rule is called an *integrity constraint*. When the body is empty, the rule is called a *fact*. Intuitively the above rule means that if $L_{k+1},\ldots,L_m$ are true and if there is not proof that $L_{m+1},\ldots,L_n$ are true (i.e., can be safely assumed to be false), then one of $L_0|\ldots|L_k$ must be true. The semantics of ASP is based on the stable model semantics of logic programming (Gelfond and Lifschitz 1990).

ASP serves as a suitable language with which to represent knowledge in the proposed microtheories for solving situated anaphora resolution problems, for several reasons. First, ASP allows *non-monotonic reasoning*, that is adding more knowledge can change one's previous beliefs, a mode especially true of situated reasoning when the world state and context can change and evolve. Second, because ASP

allows for negation-as-failure (**not** L_i) and classical negation (-L_i), default rules can be encoded, which as we will see, allows for encoding complicated cases where, for example, certain actions are not permissible if there is no reason to think they are not forbidden. Third, ASP allows for what are known as *choice rules*. In addition to literals, the head of the rule can contain *cardinality constraints* of the form $l\{L_0, \ldots, L_k\}u$ in which l, u are integers and explicitly allow the encoding of choices. Finally, we will need to be able capture dynamic systems when reasoning about actions and ASP, through its implementation as an *incremental logic program* which allows for capturing knowledge accummulating over increasing time steps (Gebser et al. 2008). For this paper, we use ASP implementations in `clingo` and `iclingo`, which provide both grounding and solving capabilities.

Dempster-Shafer Theory. DS-Theory is a measure-theoretic mathematical framework that allows for combining pieces of uncertain evidential information to produce degrees of belief for the various events of interest (Shafer 1976) . In DS-Theory a set of elementary events of interest is called *Frame of Discernment* (FoD). The FoD is a finite set of mutually exclusive events $\Theta = \{\theta_1, ..., \theta_N\}$. The power set of Θ is denoted by $2^\Theta = \{A : A \subseteq \Theta\}$. Each set $A \subseteq \Theta$ has a certain weight, or *mass* associated with it. A *Basic Belief Assignment* (BBA) is a mapping $m_\Theta(\cdot) : 2^\Theta \rightarrow [0, 1]$ such that $\sum_{A\subseteq\Theta} m_\Theta(A) = 1$ and $m_\Theta(\emptyset) = 0$. The BBA measures the support assigned to the propositions $A \subseteq \Theta$ only. The subsets of A with non-zero mass are referred to as *focal elements* and comprise the set $\mathcal{F}_\Theta$. The triple $\mathcal{E} = \{\Theta, \mathcal{F}_\Theta, m_\Theta(\cdot)\}$ is called the *Body of Evidence* (BoE). For ease of reading, we sometimes omit $\mathcal{F}_\Theta$ when referencing the BoE. Given a BoE $\{\Theta, \mathcal{F}_\Theta, m_\Theta(\cdot)\}$, the *belief* for a set of hypotheses A is $Bel(A) = \sum_{B\subseteq A} m_\Theta(B)$. This belief function captures the total support that can be committed to A without also committing it to the complement A^c of A. The *plausibility* of A is $Pl(A) = 1 - Bel(A^c)$. Thus, $Pl(A)$ corresponds to the total belief that does not contradict A. The $uncertainty$ interval of A is $[Bel(A), Pl(A)]$, which contains the true probability $P(A)$. In the limit case with no uncertainty, we get $Pl(A) = Bel(A) = P(A)$.

DS-Theory extends Bayesian theory in several ways, allowing for some capabilities that are suitable for our purposes. First, it allows for assigning probabilistic measures to sets of these hypotheses (not just individual ones), including the set of all hypothesis. This allows DS-Theory to consider ignorant and ambiguous information, which is helpful when there is evidence that an anaphor could resolve to more than one object candidate. Second, DS-Theory does not require assuming any prior distributions over object candidates, which is useful when priors are difficult to justify. Bayesian and DS-theories do share many commonalities and DS-theory is often viewed as being a generalization.

Detailed Walk-through of an Example Situated Anaphoric Imperative Discourse

Consider the discourse D_1 from (5). In the scene, the speaker is performing a [washing dishes/cooking] task and is looking to have the robot use a knife to cut the tomato and then to pass over the [knife/tomato] once it is done. Below is the discourse comprising three utterances, as follows:

Pick up the [knife]$_k$.
Cut the [tomato]$_t$.
Pass it$_{k,t}$ to me

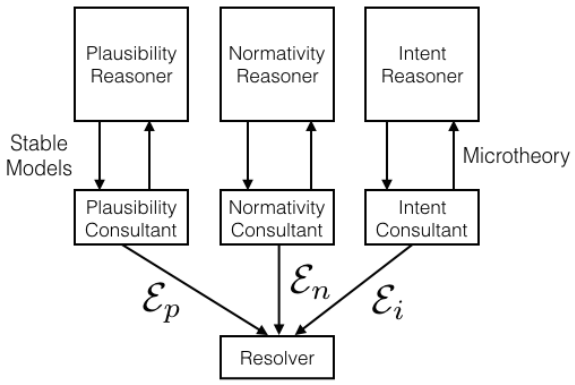

Figure 2: Approach for resolving situated anaphors. Each of the three proposed reasoning modes are encoded as microtheory templates, filled with situational information by a consultant and solved using a reasoner (e.g., answer set solver). Uncertainty measures are computed over the set of object candidates and combined to return the best guess.

We propose a knowledge representation scheme and resolution approach implemented through a resolver architecture as shown in Figure 2. Specifically, we consider three consultants, one for each reasoning mode. The consultants instantiate and run a reasoner, and later compute uncertainty metrics for object candidates. Each reasoner is an ASP solver that operates on a microtheory which is a ground logic program. The consultants maintain partial or incomplete microtheories (*microtheory template*) that are domain-independent. During resolution, the consultants "fill-in" the missing facts and rules in order to be able to complete reasoning. The microtheory templates are not solvable logic programs as they are not sufficiently ground, but merely serve as a general blueprint for a consultant to flesh out. In the next section, we will discuss properties of these templates in more detail, but for now, we will look at specific microtheories for our running example.

Once reasoning has been performed, each reasoner returns, to its corresponding consultant, answer sets (if available) that are stable models consistent with the information that the agent has used in reasoning. Thus, each reasoner answers the question of whether a set of facts relating to the discourse "makes sense" from a plausibility or normativity or intent point of view, depending on the consultant. The consultant defines a DS-theoretic mass function over these models, which in turn allows the consultant to build a BoE ($\mathcal{E}$) over the objects of interest. We will now look at each microtheory for (5) in more detail.

Plausibility Microtheory. The plausibility microtheory contains knowledge for determining if an *action* requested in an utterance *can* be performed in relation to an *object*, given the agent's action capabilities and the current situation. We use an incremental answer set programming paradigm as it is reasoning with action dynamics. First, we establish a program base that incorporates a set of facts true of the agent's (initial) current situation in its interaction with a human (named "commX" or "commander X")

```
% For incremental mode iclingo
#include <incmode>.

#program base.
% Percept types
is(obj1,object).
is(obj2,object).
is(obj3,scene).
is(obj4,loc).
is(obj5,person).
is(self,loc).

% Percept names
has(obj1,name,knife).
has(obj2,name,tomato).
has(obj3,name,kitchen).
has(obj4,name,table).
has(obj5,name,commX).

% Fluents
init(has(obj1,loc,self)).
init(has(obj2,loc,table)).

% Initial state axiom
holds(F,0):- init(F).

% Basic action definitions
action(pickup(X)) :- is(X,object).
action(putdown(X)) :- is(X,object).
action(pass(X,Y)) :- is(X,object),is(Y,person).
```

The next step is to GENERATE a set of action occurrences using ASP choice rule syntax, as follows:

```
#program step(t).
%GENERATE
{ occ(A,t) :  action(A) } = 1.
```

The syntax states that at any given time instance 't', one and only one action occurs from the set of possible actions. We can then define the effects of our actions and various action-related axioms.

```
%DEFINE
% Effect of action occuring
holds(has(X,loc,self),t) :- occ(pickup(X),t-1).
-holds(has(X,loc,table),t) :- occ(pickup(X),t-1).
holds(has(X,loc,table),t) :- occ(putdown(X),t-1).
-holds(has(X,loc,self),t) :- occ(putdown(X),t-1).
holds(has(X,loc,Y),t) :- occ(pass(X,Y),t-1).
-holds(has(X,loc,self),t) :- occ(pass(X,Y),t-1).

% Inertia axioms
holds(F,t) :- holds(F,t-1), not -holds(F,t).
-holds(F,t) :- -holds(F,t-1), not holds(F,t).

% Commonsense laws
% Objects cannot be in two locations at once.
-holds(has(X,loc,Y),t) :- holds(has(X,loc,Z),t),is(Y,loc),
Y!=Z.
```

Once, the axioms are defined we use the integrity constraint syntax of ASP to represent various action requirements and various situations that do not make plausible sense.

```
%TEST
% Cannot pick up something you are already holding
:- occ(pickup(X),t), holds(has(X,loc,self),t).

% Cannot pick up something when holding something else
:- occ(pickup(X),t), holds(has(Y,loc,self),t).

% Cannot put down something you are not holding
:- occ(putdown(X),t), -holds(has(X,loc,self),t).

% Cannot pass something you are not holding (on the table)
:- occ(pass(X,Y),t), holds(has(X,loc,table),t).

% Cannot pass if recipient already has it
:- occ(pass(X,Y),t), holds(has(X,loc,Y),t).

#program check(t).
:- query(t), t<maxlength.

#const maxlength=5.
```

The `#program check(t)` operation provides a termination of the program, which in this case is after five steps, set by the `maxlength` constant. This microtheory is essentially a planning microtheory defined over a short time horizon of five steps.

Normative Microtheory. The normative microtheory contains knowledge for answering the question of whether an action *should* be performed on an object candidate. This microtheory shares many common features with the plausibility microtheory including the types of fluents and static predicates that are used such as $is(X, Y)$ and $has(X, Z, Y)$. These predicates are intentionally quite general and are designed to be representative of a high level language that a situated agent can use (Baral, Lumpkin, and Scheutz 2017). A crucial difference in the normative microtheory are the generate choice rules and the special predicate ($has(A, permissible, X)$) used therein, as shown below

```
%GENERATE
has(obj5,is_doing,(washing_dishes;cooking)) = 1.

{ has(A,permissible,X) : is(A,action_verb),is(X,object),
is(S,person),has(S,uttered,A) } = 1.
```

The normative microtheory also has an additional choice rule associated with the task context of `washing_dishes` versus `cooking`. The task context choice rule is meant to allow the program to consider what would happen in different contexts. However, we anticipate that in real-world situations, the context is set and not necessarily choose-able

in this manner. Nevertheless, we provide this as a choice rule, so we can compare performance across two different contexts. We can then define and test "permissible" actions against what is deemed forbidden or in some instances, what is not shown to be not forbidden, depending on the normative requirements at play.

```
%DEFINE
% Forbidden to pass something that is not
% a dish to someone who is washing dishes.
has(X,function,tool) :- has(X,used_for,cutting).
has(cooking,requires,X):- has(X,used_for,eating).
has(washing_dishes,requires,X):- has(X,function,tool).
-has(A,forbidden,X) :- has(S,is_doing,T),
has(A,permissible,X), has(T,requires,X).

%TEST
:- has(A,permissible,X),not -has(A,forbidden,X).
```

Speaker Intent Microtheory. The speaker intent microtheory contains knowledge for answering the question of whether the action on an object candidate was what the speaker intended for the agent to do. Once again, this microtheory shares many of the same predicates with the normative and the plausibility microtheories. And, as with those other two theories, what is unique is the special predicate used in the generate step ($has(A, speaker\_intends, X)$).

```
%GENERATE
%Task that the speaker is doing
has(obj5,is_doing,(washing_dishes;cooking)) = 1.

{has(A,speaker_intends,X) : is(A,action_verb),is(S,person),
has(S,uttered,A),is(X,object)} = 1.
```

The potential set of action-object pairings suggested by the speaker's utterance are constrained by what actions might be relevant to the task that the speaker is performing.

```
%DEFINE
has(X,nextAction,A) :- has(X,loc,self),is(X,object),
has(A,name, putdown).
has(X,nextAction,A) :- has(X,loc,table),is(X,object),
has(A,name,pickup).

% When something is split, it is made up of multiple parts
% Speaker unlikely to use "it" when referring to
% multiple object
has(X,number_parts,multiple) :-
has(X,physical_integrity,split),is(X,object).
-has(A,speaker_intends,X):-
has(X,number_parts,multiple),
has(A,verb_pronoun_ref,W),
is(W,pronoun),has(W,name,it),is(X,object),
not has(A,speaker_intends,X).

% speaker prefers the robot to perform the next action
-has(A,speaker_intends,X):- not
has(X,nextAction,A),is(A,action_verb),is(X,object), not
has(A,speaker_intends,X).

% relevance of an object to a speaker if it helps the
% speaker
has(X,function,tool) :- has(X,used_for,cutting).
has(cooking,requires,X):- has(X,used_for,eating).
has(washing_dishes,requires,X):- has(X,function,tool).
has(X,relevant_to,S) :- has(S,is_doing,T),
has(T,requires,X), is(S,person).

%TEST
% speaker does not intend for the robot to pass
% it irrelevant things.
:- has(A,speaker_intends,X), not has(X,relevant_to,S),
has(S,uttered,A).
```

Combining Evidence with Dempster-Shafer Theory. We are reasoning about whether or not certain action-object pairings make sense. Thus, if the set of candidate objects is $O = \{o_1, \ldots, o_n\}$, we can define the DS-theoretic frame of discernment $\Theta = O$. Now, each microtheory potentially outputs a set of answer sets, $\mathcal{A} = \{A_1, \ldots, A_n\}$, where if $\mathcal{A} = \emptyset$ then the microtheory is unsatisfiable. We know that each answer set contains a set of ground predicates $A_i = \{p_i^1, \ldots, p_i^k\}$ including exactly one special generative predicate p_i^*. The generative predicate contains one of the n candidate objects O. We can define our mass function of some subset $B \subseteq \Theta, B \neq \emptyset$ as being the following:

$$m_\Theta(B) = \begin{cases} \frac{N}{|\mathcal{A}'|} & \forall b \in B, \exists i, \text{ such that } p_i^* \in A_i \text{ and} \\ & p_i^* \text{ contains } b, \text{ does not contain } b' \notin B \\ 0 & otherwise \end{cases}$$

Where $\mathcal{A}' \subseteq \mathcal{A}$ refers to those answer sets that contain the desired action verb, and N is the number of answer sets $A_i \in \mathcal{A}'$ that satisfy the specified criterion. The reasoning consultants compute these mass functions and return a BoE $\mathcal{E}$ that contains the computed mass functions for the focal elements For example, if there are two object candidates $\Theta = \{o_1, o_2\}$ and a reasoner returns three answer sets $\mathcal{A} = \{A_1, A_2, A_3\}$, with A_1 containing o_1, A_2 containing o_2 and A_3 containing both object candidates, then the BoE will contain masses $m(\{o_1\}) = 1/3, m(\{o_2\}) = 1/3, m(\{o_1, o_2\}) = 1/3$. The evidence from these sources can be combined using the Dempster's rule of combination, which aggregates evidences or confidence values from different sources, but within the same frame of discernment. The results from computing the fused uncertainties for each of discourse examples (3), (4) and (5) are shown in Table 1.

General Properties

In this section, we discuss several general, domain-independent aspects of the proposed approach.

Class of Situated Anaphora Resolution Problems

Thus far, we have presented a few examples of situated anaphora resolution problems.[1] Discourses in this class of problems share the following features:

1. Each discourse consists of a set of utterances, at least one utterance being an imperative. The imperative utterance contains at least one anaphoric referring expression.

[1] Additional examples: (1) "Walk to the green door. Enter your passcode on the panel to open it." (2) "My pen fell under the bed. Grab that broom. Use the long end to get it out." (3) "The knife is on the chair. Pull it out. Grab the knife. Push it back."

	(5) "Pass it to me." [washing dishes / cooking] $\Theta = \{knife, tomato\}$	**(4) "Put it in the bowl." [Bowl contains food]** $\Theta = \{knife, tomato, bowl\}$	**(3) "Put it down."** $\Theta = \{knife, tomato\}$
Plausibility ($\mathcal{E}_p$)	10 Stable models with "pass" $m(\{knife\}) = 0.4$ $m(\{tomato\}) = 0.3$ $m(\Theta) = 0.3$	29 Stable models with "put in" $m(\{knife\}) = 0.52$ $m(\{tomato\}) = 0.21$ $m(\{knife, tomato\}) = 0.27$	4 Stable models with "put down" $m(\{knife\}) = 0.25$ $m(\Theta) = 0.75$
Normative ($\mathcal{E}_n$)	Washing Dishes: 1 Stable Model, $m(\{knife\}) = 1.0$ | Cooking: 1 Stable Model, $m(\{tomato\}) = 1.0$	1 Stable Model $m(\{tomato\}) = 1.0$	1 Stable Model $m(\{knife\}) = 1.0$
Speaker-Intent ($\mathcal{E}_i$)	Washing Dishes: 1 Stable Model, $m(\{knife\}) = 1.0$ | Cooking: 1 Stable Model, $m(\{tomato\}) = 1.0$	1 Stable Model $m(\{tomato\}) = 1.0$	2 Stable Models $m(\{knife\}) = 0.5$ $m(\{tomato\}) = 0.5$
Combined Scores	Washing Dishes: $knife : [1.0, 1.0]$ | Cooking: $tomato : [1.0, 1.0]$	$tomato : [1.0, 1.0]$	$knife : [1.0, 1.0]$

Table 1: Computed uncertainties for each of the object candidates in three different scenarios. The bottom row contains the final uncertainties for the object candidates. Thus, for example, the agent is certain that the object pronoun must resolve to "tomato" when a speaker has the agent to "pass it" and the speaker was in the middle of a cooking task.

2. The discourse context contains, among other things, two or more candidate antecedents to the at least one anaphoric expression. The antecedents could be explicitly mentioned linguistic referring expressions or discourse entities (real-world objects or cognitive concepts) being considered by the interlocutors as part of the discourse.
3. Linking the anaphoric expression to one of the candidate antecedents requires reasoning with *extra-linguistic situational knowledge*.
4. Executing an imperative (or mentally simulating it) during the discourse can change the state of the world in a way that influences the interpretation of a subsequent anaphor. That is, the interpretation of the anaphor is not merely influenced by the semantics of a linguistic expression (as in WS), but by what happens in the world as a result of incrementally interpreting (and executing) each utterance in the discourse.

Domain-Independent Aspects of the Reasoners

Each of the microtheories share a common structure as allowed by the GENERATE-DEFINE-TEST methodology in ASP. In the English language, every simple imperative utterance with an object pronoun has no overt subject (i.e., the subject is assumed to be the speaker) and the verb is often in its bare form. This focused structure allows us to consider specifically the relationship between just the *action verb* and the *object*. In the case of pronoun use, the object is replaced by an object pronoun such as "it." From a representational standpoint we are only interested in this relationship between *action* and *object*. This means we can pre-define useful relationships between action and object for each of the reasoners and generally ask the question if an action-object pairing makes sense from the corresponding reasoner's (or microtheory's) perspective. If we let a be the action verb in the utterance O represent the variable corresponding to the object pronoun, then we have three general symbols, one for each reasoning mode, as follows:

1. Plausibility Reasoning: $\mathbf{occ}(a(O), t)$
2. Normative Reasoning: $has(a, \mathbf{permissible}, O)$
3. Speaker Intent: $has(a, \mathbf{speaker\_intends}, O)$

Moreover, many of the commonsense definitions shown in the code fragments in the previous section for each of the reasoners are in fact domain-independent such as the axioms of inertia in the plausibility reasoner, the rule associated with when an action is not forbidden (-$has(A, forbidden, O)$) in the normative reasoner and the rule relating to when an object is relevant to a speaker ($has(O, relevant\_to, S)$) in the speaker intent reasoner. In fact, we were able to model (3) and (4) in much the same way as we did with (5).

We note that not only does this generality hold within a domain (like cutting vegetables), it extends to other domains as well as in cases where it is possible to incorporate several object pronouns in sequence, each referring to different objects, as follows:

Pick up the [ladel]$_l$. Put it$_l$ in the [pot]$_p$ containing [soup]$_s$. Stir it$_{l,p,s}$. Check if it$_{l,p,s}$ is mixed. Take it$_{l,p,s}$ out and wash it$_{l,p,s}$. (6)

Related Work

Early non-statistical approaches exploited hard constraints (syntactic, semantic and morphological) and selection preferences (salience and commonsense knowledge) that humans used when disambiguating anaphors (Winograd 1980; Lappin and Leass 1994; Ge, Hale, and Charniak 1998). Object pronouns like "it" were resolved by exercising a selection preference based on salience in the discourse, which in turn was often tied to how near the antecedent was to the

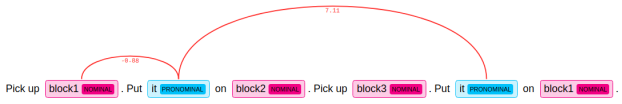

Figure 3: Incorrect resolution of a current neural coreference system (Clark and Manning 2016) on the example shown in Figure 1. (Implementation: https://huggingface.co/coref/)

anaphor, i.e., its recency (Van Deemter 2016). Although such an approach would suffice for (1), it would fail for (3), (4) and (5) in certain contexts.

Contemporary approaches aim to solve the sister-problem of coreference resolution using statistical and neural network based algorithms. Mitkov (2014) and Ng (2010) provide nice overviews of the evolution of this field. The state-of-the-art systems are trained on large corpora and they are able to recognize mention-pairs that are statistically related to one another. However, many of these systems fail when presented with situated anaphors, much like (1). Clark and Manning (2016)[2] propose one such system that does not produce correct resolution of (1) (see Figure 3), (3), (4), (5) and all the examples provided in the paper. Extracting statistical patterns are insufficient because disambiguating situated anaphors requires inference on world models that evolve through the discourse. Some have proposed specific representations to incorporate background knowledge needed to solve these more "hard coreference problems" (Peng, Khashabi, and Roth 2015). However, these systems still lack the ability to reason about plausibility, normativity and speaker intent, which we argue to be important reasoning modes in situated cases.

There have been parallel efforts in tackling the Winograd Schema Challenge and the leading approaches employ a strategy of first selecting a format to represent commonsense factual knowledge, learning vast amounts of static commonsense knowledge from online databases (ConceptNet, WordNet and CauseCom), and then performing inference or classification with this knowledge (Bailey et al. 2015; Sharma et al. 2015; Liu et al. 2016; Golovin, Claßen, and Schwering 2017). However, these approaches do not consider the three reasoning modes we discuss. Also, it is unclear how the knowledge needed for performing this situated reasoning can be acquired from these online databases.

There has been considerable work in *reference resolution*, more generally, and in applying various theories from cognitive science in order to use pragmatics (Schüller 2014; Richard-Bollans, Gomez Alvarez, and Cohn 2017; Kehler 2000; Chai, Prasov, and Qu 2006; Williams et al. 2016; Williams and Scheutz 2016; Van Deemter 2016). Unfortunately, much of this work focuses on pragmatics as they pertain to processing effort and cognitive effect on the agent, and less so on situational aspects of the agent's surroundings. The models that do unpack discourse context information have received a weaker computational treatment.

[2]https://huggingface.co/coref/ and http://corenlp.run/

Discussion and Limitations

Although we propose three specific reasoning modes, we expect that there are sub-classes of situated anaphor resolution problems that cannot be resolved with just the three reasoning modes proposed. For example, if the imperatives in the discourse represent not the intent of the speaker but of another in situations where the speaker may simply be conveying a message or an order from, for example, their superior or boss. In such cases, the agent would need the capability to reason about intent of another beyond the speaker. Our approach is by no means limited to just these three reasoning modes, and it is subject of future work to explore when and how these and other reasoning modes are triggered.

In this paper, we did not address how these microtheory templates (partial microtheories) are learned or how the agent acquires them. The question of learning is an important one and there has been extensive research efforts in acquiring and encoding knowledge from the WWW. However, situated anaphors present a unique challenge in that much of the knowledge needed to resolve them might not be explicitly available in a dataset. Instead, this knowledge may be quite implicit acquired by the agent throughout its lifetime. We are currently exploring how an embodied agent might glean these implicit rules from experience.

We have presented the first steps towards resolving ambiguous references by reasoning with situated information available to an agent when embodied in an environment. One follow-on step for this research effort is to integrate these capabilities into a cognitive robotic architecture and attempt to empirically evaluate the system and the knowledge represented therein in real human-robot interaction scenarios. One advantage of the proposed microtheories are the use of identifiers for object constants that allow for the integration of multi-modal perceptual information about the same entity to be aggregated and reasoned with and allows for the symbols to be ground in the robot's sensory-motor system.

Conclusion

Artificial agents interacting with humans will need to be able to disambiguate anaphoric expressions, which are used freely and frequently in discourse. To do so, we argue that the agent must consider what it *can*, *should* and be *expected* to do in a situation. In this paper, we propose a knowledge representation scheme to formalize domain-independent and situation-specific knowledge for each of these three considerations, and a resolution strategy for using this knowledge to disambiguate object pronouns in simple imperative sentences that are situated in real-world embodied discourse. This work advances the state of the art in anaphora resolution by reframing the disambiguation problem from being only about mention-pairs, to also being about the viability of the actions being considered as the world state evolves.

Acknowledgments

This project was supported in part by ONR MURI grant N00014-16-1-2278. We also thank Chitta Baral for his helpful suggestions and comments on this paper.

References

Bailey, D.; Harrison, A.; Lierler, Y.; Lifschitz, V.; and Michael, J. 2015. The Winograd Schema challenge and reasoning about correlation. In *Proceedings of the AAAI Spring Symposium on Symposium on Logical Formalizations of Commonsense Reasoning*.

Baral, C.; Lumpkin, B.; and Scheutz, M. 2017. A high level language for human robot interaction. *Advances in Cognitive Systems*.

Baral, C. 2003. *Knowledge representation, reasoning and declarative problem solving*. Cambridge university press.

Cassimatis, N. L. 2008. Resolving ambiguous, implicit and non-literal references by jointly reasoning over linguistic and non-linguistic knowledge. *Semantics and Pragmatics of Dialogue (LONDIAL)* 173.

Chai, J. Y.; Prasov, Z.; and Qu, S. 2006. Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research* 27:55–83.

Clark, K., and Manning, C. D. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2256–2262.

Clark, H. H., and Marshall, C. R. 1981. Definite reference and mutual knowledge. *Psycholinguistics: critical concepts in psychology* 414.

Ge, N.; Hale, J.; and Charniak, E. 1998. A statistical approach to anaphora resolution. In *Sixth Workshop on Very Large Corpora*.

Gebser, M.; Kaminski, R.; Kaufmann, B.; Ostrowski, M.; Schaub, T.; and Thiele, S. 2008. Engineering an incremental ASP solver. In *Proceedings of the International Conference on Logic Programming*, 190–205. Springer.

Gelfond, M., and Lifschitz, V. 1990. Logic programs with classical negation, logic programming.

Golovin, D.; Claßen, J.; and Schwering, C. 2017. Reasoning about conditional beliefs for the Winograd Schema Challenge. In *Proceedings of the Thirteenth International Symposium on Commonsense Reasoning (Commonsense-2017)*.

Grosz, B. J., and Sidner, C. L. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics* 12(3):175–204.

Hobbs, J. R. 1978. Resolving pronoun references. *Lingua* 44(4):311–338.

Kamp, H. 1981. A theory of truth and semantic representation. *Formal semantics-the essential readings* 189–222.

Kehler, A. 2000. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of the 17th AAAI Conference on Artificial Intelligence*, 685–690.

Lappin, S., and Leass, H. J. 1994. An algorithm for pronominal anaphora resolution. *Computational linguistics* 20(4):535–561.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, 552–561. AAAI Press.

Liu, Q.; Jiang, H.; Ling, Z.-H.; Zhu, X.; Wei, S.; and Hu, Y. 2016. Commonsense knowledge enhanced embeddings for solving pronoun disambiguation problems in Winograd Schema Challenge. *arXiv preprint arXiv:1611.04146*.

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

Morgenstern, L., and Ortiz Jr, C. L. 2015. The Winograd Schema Challenge: Evaluating progress in commonsense reasoning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 4024–4026.

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

Peng, H.; Khashabi, D.; and Roth, D. 2015. Solving hard coreference problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 809–819.

Richard-Bollans, A.; Gomez Alvarez, L.; and Cohn, A. G. 2017. The role of pragmatics in solving the Winograd Schema Challenge. In *Proceedings of 13th International Symposium on Commonsense Reasoning (Commonsense-2017)*. CEUR Workshop Proceedings.

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

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press.

Sharma, A.; Vo, N. H.; Aditya, S.; and Baral, C. 2015. Towards addressing the Winograd Schema Challenge-building and using a semantic parser and a knowledge hunting module. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 1319–1325.

Tellex, S.; Thaker, P.; Deitsl, R.; Simeonovl, D.; Kollar, T.; and Royl, N. 2013. Toward information theoretic human-robot dialog. *Robotics* 409.

Van Deemter, K. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

Williams, T., and Scheutz, M. 2016. A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

Williams, T.; Acharya, S.; Schreitter, S.; and Scheutz, M. 2016. Situated open world reference resolution for human-robot dialogue. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*.

Winograd, T. 1980. What does it mean to understand language? *Cognitive science* 4(3):209–241.

Wiseman, S. J.; Rush, A. M.; Shieber, S. M.; and Weston, J. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.