Reasoning Requirements for Indirect Speech Act Interpretation

Vasanth Sarathy, Alexander Tsuetaki, Antonio Roque, Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University
Medford, MA 02155
{vasanth.sarathy, alexander.tsuetaki, antonio.roque, matthias.scheutz}@tufts.edu

Abstract

We perform a corpus analysis to develop a representation of the knowledge and reasoning used to interpret indirect speech acts. An indirect speech act (ISA) is an utterance whose intended meaning is different from its literal meaning. We focus on those speech acts in which slight changes in situational or contextual information can switch the dominant intended meaning of an utterance from direct to indirect or vice-versa. We computationalize how various contextual features can influence a speaker's beliefs, and how these beliefs can influence the intended meaning and choice of the surface form of an utterance. We axiomatize the domain-general patterns of reasoning involved, and implement a proof-of-concept architecture using Answer Set Programming. Our model is presented as a contribution to cognitive science and psycholinguistics, so representational decisions are justified by existing theoretical work.

1 Introduction

Understanding natural language utterances inherently involves resolving ambiguity associated with the meaning of linguistic expressions. Resolving linguistic ambiguity involves choosing one amongst many possible meanings, at the word level, at the sentence level, and at the discourse level. Word and sentence embeddings are current topics of active research and focus on selecting suitable word or sentence representations that best capture meaning. Current transformer architectures such as BERT (Devlin et al., 2018) are used to pretrain language representations so word and sentence embeddings account for nearby linguistic context. These systems are based on two key insights that we explore in this paper. The first insight is that word meanings (in the case of BERT) and span meanings (in the case of SpanBERT (Joshi et al., 2020)) cannot be understood separate from their context, which in these cases is primarily linguistic context. Using these contextualized language embeddings have led to state-of-the-art performance in several downstream tasks. However, there is still the open problem of how to incorporate extra-linguistic context. Specifically, it is unclear how to categorize and use knowledge derived from common sense knowledge about how the world works as well as situational knowledge about aspects of the physical environment in a systematic way for NLU.

The second insight is that the way context influences meaning can be entirely captured statistically from data. That is, a word encountered together with context in a particular way is likely to mean the same as the same word encountered in a similar linguistic context. The assumption is that systems with sufficient training data, or at least with powerful pretrained models are likely to interpret language correctly. However, Cohen (2019) and others have argued that pattern recognition by itself might be insufficient, and such an association might require logical reasoning beyond learning statistical similarity.

Recent work by Sarathy et al. (2019) explored both these insights and argued that for the task of coreference resolution (at least with respect to imperative discourse), to interpret a pronoun an NLU system must reason about actions and change, normative obligations and social aspects, and the intent of speakers. Knowledge from situational context combined with background domain-general knowledge is needed to perform this reasoning. Meaning is derived from choosing the interpretation that makes the "most sense" (not just in terms of statistical regularity, but in terms of coherence more broadly) given a variety of contextual factors.

We are inspired by this key idea to explore how context and reasoning can assist not only at the word level (as Sarathy et al. have done), but at the sentence level. Specifically, we address the problem of how a listener can infer the intended meaning of an utterance given contextual knowledge. Sometimes, as in **Indirect Speech Acts** (ISAs), the literal meaning of a sentence (derived from the surface form) can be different from its intended meaning (Searle, 1975; Austin, 1975; Gordon and Lakoff, 1975). For example, the utterance "can you open the door?" has a *literal meaning* of an elicitation of information from the listener regarding the listener's ability to open the door – this is because the utterance has the surface form of a question. And in fact, the speaker may in some cases be interested in whether the hearer is capable of opening the door: if the hearer is physically disabled and the speaker is a caregiver, for example. However, very frequently, the *intended meaning* of the utterance is a request that the listener perform the action of opening the door, in which case the utterance is an ISA. As this example suggests, the proper interpretation is guided by context.

Research has shown that ISA use is common in human-human communication (Levinson, 2001), as well as in human-robot task-based dialogues (Williams et al., 2018)[1]. The problem, however, is that ISAs are notoriously difficult to computationalize when their interpretation is influenced by contextual factors (Williams et al., 2018; Briggs et al., 2017). Unlike past approaches to ISA interpretation, we incorporate a broader notion of context and extended reasoning capabilities to reason about the speaker's intent and beliefs.

Our contributions include a computational formalization of (1) the preconditions, related to the speaker's intentions and beliefs, that need to be satisfied to determine if a surface form carries indirect meaning, (2) the reasoning needed to infer speaker beliefs based on domain general principles about the world (e.g., inertia), expectations of linguistic alignment, and mutual knowledge and (3) the representations needed to assimilate factual contextual evidence available to the listener into the reasoning process. To make this work more tractable, we limit our focus to speech acts that have the surface form of a question, the literal meaning of an *Ask* for information, and the intended meaning of an *Ask* for information (in which case the speech act is not an ISA) or a *Request* for action (in which case the speech act is an ISA). In the tradition of past work in computationally-aided linguistic analysis, these formalizations are derived using examples from several task-oriented dialogue corpora.

2 Approach

Our approach is based on two traditions. First, the tradition from computational linguistics of building formal representations of linguistic phenomena from corpora, guided by linguistic and cognitive scientific theories. Recent papers in this tradition were presented by Jiménez-Zafra et al. (2018) and Ilievski et al. (2018). We emphasize that the model that we build from this corpus analysis is a cognitive scientific model, which builds on theories from cognitive science and psycholinguistics as described in Section 3.

Second, the tradition of building explicit knowledge representation formalisms along with reasoners capable of deriving meaning of natural language expressions from their logical forms. Recent work in this tradition has been performed by Sarathy et al. (2019), Sharma et al. (2015), and Wilske and Kruijff (2006).

Combining these traditions, we perform a corpus analysis to extract archetypal utterances from a variety of corpora; this process includes a consensus annotation of those utterance's direct and indirect meanings, as well as of the context that influences that interpretation, as described in Section 2.1. Then we build a formal computational model of the representations and reasoning required to perform those interpretations, ensuring that the computational model performs a complete coverage of the archetypal utterances, as described in Section 2.2. The resulting model is detailed in Section 3.

[1]People not only use ISAs but repair ineffective ISAs with other ISAs, supporting the idea that humans have a strong preference for ISAs.

2.1 Corpus Analysis

We are motivated to perform a corpus analysis because of limitations in existing corpora, which include dialogue act classification and intent recognition datasets, e.g. by Li et al. (2017), Shriberg et al. (2004), and Jurafsky et al. (1998). These corpora cannot be used as-is for reasons shown by Roque et al. (2020): because ISAs are context-sensitive, **ISA Schemas** need to be developed. These are data structures that are made up of an utterance and two different context representations; in one of the contexts the utterance is an ISA, and in the other context the utterance is not an ISA. However, these ISA Schemas do not seem to be amenable to economy-of-scale authoring techniques such as crowdsourcing, so some amount of expert authoring seems to be required. For that reason, we used a corpus analysis approach suggested by Roque et al. (2020): using real-world corpora to manually author ISAs. We proceeded as follows.

First, we obtained three existing corpora, selected to provide variation (speech vs text, virtual-world vs real-world, human-human vs human-robot, navigation vs construction) while retaining a focus on task-based communication: (1) SCARE corpus (Stoia et al., 2008) in which one person helped another person navigate through a virtual environment, (2) Diorama corpus (Bennett et al., 2017) in which one person directed a tele-operated robot in arranging real-world items, and (3) Minecraft corpus (Narayan-Chen et al., 2019), in which one person directed another person in collaboratively building structures in a virtual environment.

Second, we manually searched the corpora (15,000+ utterances in 50+ hours of recordings) for pairs of human utterances with identical or very similar surface form, and in which one element of the pair was an ISA and the other pair was not. It was not necessary for the utterance to have been made by the same person, or even in the same corpus, because any system capable of interpreting ISAs should be domain-general. We found 20 such pairs in the corpus of 15,000 utterances, which is consistent with Roque et al. (2020)'s experience of obtaining a very small yield of ISA Schemas when developing them using crowsourced methods.

Third, having identified such pairs, we manually extracted the context details: scene, dialogue history, roles, constraints, and whether the intended meaning was indirect or direct. We identified the values of these parameters through consensus annotations performed by the first three authors of this paper. The utterance, and the context parameter/value slots, are what make up an ISA Schema. We consider these ISA Schema to be *archetypal* in the sense that they represent the set of ISAs in the corpus, including the ones that are unpaired. Our focus is on the paired ISAs, to investigate the ways that varying context varies interpretation for a given utterance.

2.2 Reasoner Development

Having thus identified archetypal ISA Schemas, we next developed a computational model of the logical reasoning required for their interpretation. The model is described in detail in Section 3, and was developed through the iterative development of logic programs, ensuring that the resulting programs achieved a complete coverage of the 20 extracted ISA Schemas. An important finding was the extent to which the knowledge, representations and reasoning requirements were domain-general and not tied to the particular corpora themselves. Thus, the reasoner we built, encoding these domain-general aspects, serves as a necessary starting point for others extending this work to cover other corpora or other languages.

We encoded the representations in Answer Set Prolog (ASP), also known as Answer Set Programming, a declarative logic programming paradigm that has been used for representing knowledge and reasoning non-monotonically (Gelfond and Lifschitz, 1990). ASP programs bear a superficial resemblance to Prolog or to production rule programs, but are different in how they work. In the ASP paradigm, logic programs are written to represent knowledge about the world. Variables in the program rules are then logically *grounded* in the constants representing a particular situation of interest. A *solver* is then used to compute answer sets in which the world definitions are logically satisfied (Gelfond and Kahl, 2014). We used Clingo, an ASP solver (Gebser et al., 2011), for implementing the model described in Section 3.

ASP has several characteristics that make it useful for representing knowledge in problems where context is important (Baral, 2003), such as with the ISA interpretation problems we are addressing. First, ASP allows *non-monotonic reasoning*, or adding knowledge that changes currently-existing beliefs.

Second, ASP represents two types of negation: *classical negation* is used to indicate propositions that are false, and *negation-as-failure* is used to indicate propositions that are considered false because they are not currently known to be true. Third, ASP allows for *choice rules* and *cardinality constraints*, which allow for the explicit encoding of world-related constraints to solutions.

One of the purposes of developing this model is to formalize the requirements and products of the reasoning involved. As described in Section 3, the model requires contextual evidence as input, and produces a set of candidate intents. The specifics of how this would be integrated into an embodied and situated intelligent system are beyond the scope of this paper, but while building the model we ensured that these tasks could in principle be achievable in agents built with contemporary cognitive architectures, e.g. by Ritter et al. (2019), Scheutz et al. (2019), and Laird (2012).

2.3 Related Work

Current state-of-the-art data-driven approaches to dialogue act classification employ general ML techniques like SVMs, Bayes Nets and CRFs. More recently, deep learning approaches (Liu et al., 2017; Li et al., 2019) have shown promising results on popular datasets like the Switchboard Corpus. These approaches model dependencies between neighboring linguistic expressions using topic information as additional context. An advantage of these approaches is their generality across a wide range of discourses. However, a limitation in these approaches is that because their datasets possess a limited conception of context (i.e., linguistic, or at best, the topics of conversation), they do not learn how situation-specific aspects can switch the interpretation of an utterance. These systems essentially learn *conventionalized* ISAs (i.e., those whose indirect interpretation is always dominant regardless of context). We anticipate that these approaches can be combined with our approach to help improve the performance for conventionalized ISA, when extensive reasoning may not be needed.

Rule-based approaches use context-sensitive rules to logically derive indirect meanings from surface forms (Briggs et al., 2017; Williams et al., 2015). These approaches address some of the limitations of the purely data-driven techniques by explicitly accounting for context. However, they generally do not reason over multiple aspects of context. Moreover, a single chunked rule (suggested by these approaches) is not domain-general and does not provide much in way of explanation for the intent recognition process. Asher and Lascarides (2001) provided a formal account of ISAs but were primarily focused on the semantics of conventionalized ISAs. Our approach addresses these limitations and is significantly more elaboration-tolerant in allowing for small factual variations to switch the interpretation.

Plan-based approaches (Perrault and Allen, 1980; Hinkelman and Allen, 1989; Green and Carberry, 1999) reason about ISAs using a set of plan recognition techniques to infer the goals of a speaker. None of these plan-based approaches systematically enumerate the different types of context, as our approach does. These approaches do not incorporate the influence of the speaker beliefs about plausibility, preference and normativity, which we have argued to be important and reflective of assumptions about the listener's capability and preferences. That said, these approaches could be integrated with our approach as one way to infer speaker intent.

The closest line of work is that of Wilske and Kruijff (2006), who proposed an ISA interpretation architecture for human-robot task-based interactions. They discuss the notions of "feasibility," "mode of operation," and the use of interactional history, which appear to parallel our notions of capability, contextual role, and dialogue history, respectively. However, like the other existing approaches, they do not consider how a listener models the speaker's beliefs. For example, in their work, if a listener considers an action to be feasible, it will take it as a request and immediately perform the action, independent of whether or not the speaker believed the listener to be capable.

3 Reasoning for ISA Interpretation

Brown (1980) provided an early descriptive account of ISAs, in which she elaborated on their form, what is required to choose one form over another, and what is needed for these surface forms to carry indirect meaning. We propose that these notions are useful not just descriptively to study ISAs as inherently interesting phenomena, but also computationally as a rubric to aid in understanding and interpreting

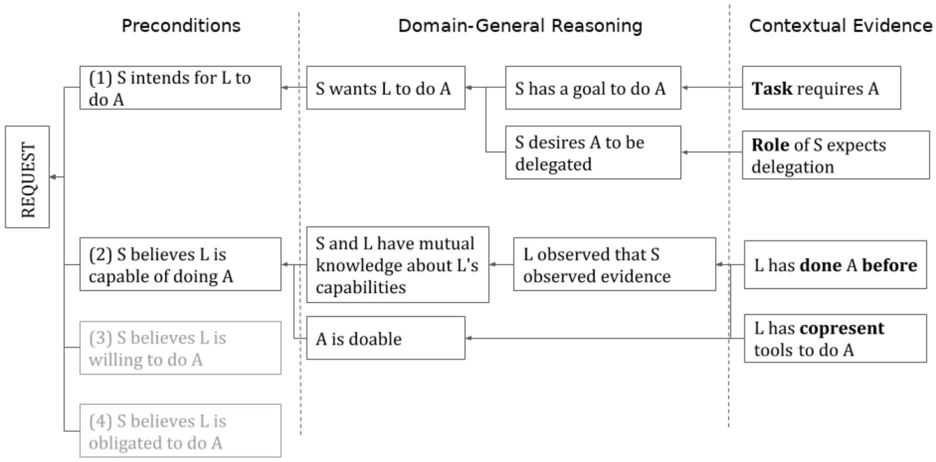

Figure 1: Types of Reasoning required to interpret the utterance "Can you do (action) A?" made from a Speaker S to a Listener L. To select the indirect meaning of a REQUEST for physical action (instead of the direct meaning of an ASK for information), L must satisfy four **Preconditions**. However, L does not have direct access to S's intentions and beliefs, and must perform **Domain-General Reasoning** (using rules about actions, mutual knowledge, social norms and obligations, and speaker preferences) based on **Contextual Evidence** (involving task, role, interactional history, and copresence). Reasoning and Evidence for Preconditions 3 and 4 follow the same pattern as for Precondition 2.

them. We computationalize the intent preconditions, and incorporate Brown's descriptive analysis of the surface form of ISAs into the architecture by allowing the surface form to dictate the necessary burden of proof required by the reasoners. Our approach, however, extends Brown's work significantly, by proposing bridge rules, reasoner-specific rules, and underlying definitions for the reasoners and their functions. Here, we systematically operationalize the principles outlined by Brown, first by formalizing the preconditions as rules in the reasoners, and then by using the surface form to influence which rules are selected for a particular utterance.

Figure 1, which shows an example of the reasoning required for interpretation, serves as an overview of our approach. The Preconditions of an interpretation are described in Section 3.1, the Domain General Reasoning is described in Section 3.2, and the Contextual Evidence required for that reasoning is described in Section 3.3.

3.1 Preconditions for Interpretation

As shown in Figure 1, each underlying communicative intent has a set of preconditions that must be met in order for a speaker to produce a surface form that conveys this intent. The theoretical justification for this comes from Brown's descriptive analysis of ISAs, which extends work by Gordon and Lakoff (1975)

As shown in Figure 1, a *Request* for a physical action (e.g., "Can you do action A?") intent must satisfy the following preconditions: (1) that the speaker S intends that the listener L **do the action** A, (2) that S believes that L plausibly **has the capability** to do A, (3) that S believes that L **is willing to** do A, (4) that S believes that L is normatively **obligated** to do A. More generally, these four preconditions also apply to other intended meanings such as *Ask* for information, where A is a speech act rather than a physical action.

Brown also suggested that the specific surface form of the ISA is influenced by the *better-knowledge principle*, which states that the direction of information flow is guided by who, S or L, has better knowledge about a particular proposition in the ISA.[2] So S may select a "can you do A?" question when S needs more evidence for satisfying precondition 2, and a "will you do A?" question when S needs more evidence for satisfying precondition 3.

[2]See also (De Ruiter, 2012) for additional discussion on "knowledge gradients".

3.2 Domain-General Reasoning

As shown in Figure 1, several sets of axioms and commonsense rules are used to test whether preconditions hold in a particular interaction between a speaker and a listener.

First we consider a set of **Domain-General Axioms** that allow an agent, regardless of situation, to reason about action and change, mutual knowledge, and linguistic alignment.

One subset of axioms involves *Reasoning about Actions and Change*. A basic component of the logical evaluation of ISAs is being able to track and infer things from the ongoing changing context. The Event Calculus (EC) is especially suited to reasoning about events and provides some axioms that are domain-independent to help guide this discussion (Kowalski and Sergot, 1989).

Formula	Meaning
$holds(F, T)$	Fluent F holds at time T
$happens(A, T)$	Ontic Action A occurs at time T
$initiated(F, T)$	Fluent F is initiated at time T
$terminated(F, T)$	Fluent F is terminated at time T

To maintain fluents once they are initiated and unless they are terminated, we have the following two rules:

```
holds(F,T+1) :- initiated(F,T),time(T).
holds(F,T+1) :- holds(F,T), not terminated(F,T), time(T).
```

Another subset of axioms involves *Reasoning about Mutual Knowledge*. One challenge in interpreting ISAs is understanding not only what the speaker knows, wants and believes, but also what the speaker thinks the agent (listener) knows, and what the speaker thinks the listener thinks the speaker thinks, and so on, ad infinitum. Herbert Clark described a heuristic-based mutual knowledge induction schema to terminate this otherwise infinite line of reasoning (Clark and Marshall, 1981). Under this schema if the agent knows that a grounds $F1$ holds, and that these grounds support a proposition $F2$, and that they know that the speaker P believes the grounds $F1$ to hold, then the agent has mutual knowledge with the speaker P. The agent can establish that if it knows that the speaker P observed the grounds (any proposition F), then they are likely to believe it holds. Moreover, if the grounds support a proposition, then, at least from the agent's perspective the supported proposition $F2$ holds as well.

Formula	Meaning
$observed(P, F, T)$	P was seen as observing F at time T
$bel(P, F)$	P believes fluent F holds
$supports(F1, F2)$	Fluent $F1$ can provide grounds for fluent $F2$ to hold
$mutual(P, F2)$	There is mutual knowledge with P about fluent F

```
holds(mutual(P,F2),T) :- holds(supports(F1,F2),T), holds(F1,T), holds(bel(P,F1),T),
    time(T).
holds(F2,T) :- holds(F1,T),holds(supports(F1,F2),T), time(T).
holds(bel(P,F),T) :- observed(P,F,T), time(T).
```

Another subset of axioms involves *Reasoning about Linguistic Alignment*. Under the theory of alignment, there is a mapping from an utterance's surface form (which could be a question, statement, command, etc.) to a direct meaning (ask, assert, request, etc.) (Asher and Lascarides, 2001). Conventionally, questions map to *Asks*, statements map to *Asserts* and commands map to *Requests*. Surface forms can also have indirect meanings and thus we need to be able to distinguish these, which we do via the following alignment axioms:

```
alignment(question,ask).
alignment(command,request).
alignment(statement,assert).
meaning(M) :- alignment(_,M).
surfaceForm(S) :- alignment(S,_).
direct(U,M) :- hasSurfaceForm(U,S), alignment(S,M), meaning(M), surfaceForm(S),
    utterance(U).
indirect(U,request) :- not direct(U,request), utterance(U).
```

Finally, a set of axioms we assume but do not explicitly discuss are *Uniqueness-of-Names* axioms which ensure that predicate relations do not clash.

Next, we consider the **Burdens of Proof** required for domain-general reasoning. As a preliminary, consider the utterance "can you do action A?" There are two types of action underlying this utterance: *ontic action* specified as A and referring to the physical act, and *epistemic action*, which in this case where the utterance is a question, is to retrieve and respond with information about A.

Formula	Meaning
$onticAction(U, A)$	Utterance U contains an ontic action A
$epistemicAction(U, M, A, Q)$	Utterance U contains an epistemic action A that holds when U has meaning M, and that pertains to proposition Q
$uttered(P, U, T)$	Interlocutor P uttered U at time T
$goal(P, A)$	Interlocutor P has the goal to do action A

As another preliminary, consider four different lines of reasoning relevant to ISA interpretation – speaker intent, plausibility, preference, normativity – each of which attempts to connect a set of facts to whether or not a particular intent holds. Each line of reasoning can be thought of as asking different questions of the facts of a given situation, as described below. These four lines of reasoning correspond directly to Brown's four preconditions for an intent to hold. But since we know from the theory of alignment that certain defaults hold we can adjust our burden of proof to make it easier to satisfy the default (or direct) interpretation than it would to satisfy the indirect interpretation. In other words, for a question "Can you install linux on your computer?" we can state that the burden to satisfy the interpretation of this utterance being an *Ask* for information should be lower than it being a *Request* for action because the surface form of it is a question and the direct meaning of questions is typically that of an *Ask* for information. One way to specify this default burden is through the use of both classical negation and negation as failure in nonmonotonic reasoning. We explain these burdens in more detail below.

First, the *Speaker Intent Burden of Proof* asks the question: does the speaker *want* the listener to have the goal of doing action (ontic or epistemic) A? If there is *no* proof that the speaker did *not* want the listener to do the action A, where A is an epistemic action, then the speaker intended for the listener to interpret the direct meaning M. For example, in the case of the utterance "Can you do action A?" if there is no proof that the speaker did not want the listener to answer the question about A, then the speaker likely intends for this to be a question and expects an answer.

```
holds(intends(P1,U,M),T) :- not -wants(P1,goal(P2,A),T), uttered(P1,U,T),
    epistemicAction(U,M,A,_), direct(U,M), holds(speaker(P1),T),
    holds(listener(P2),T), time(T), not holds(intends(P1,U,M2),T), indirect(U,M2).
```

If there is positive evidence that the speaker wanted the listener to do the ontic action A, the speaker intended for the listener to interpret the indirect meaning M. For the above example, if there is positive proof that the speaker wanted the listener to perform the physical action A, then the speaker intends for the action to be done.

```
holds(intends(P1,U,M),T) :- wants(P1,goal(P2,A),T), uttered(P1,U,T),
    onticAction(U,A), indirect(U,M), meaning(M), holds(speaker(P1),T),
    holds(listener(P2),T), time(T), not holds(intends(P1,U,M2),T), direct(U,M2).
```

In addition, we also provide some ancillary rules to help establish that once an action is performed, the intent that provoked this action is terminated. Moreover, once the intent is formed, the action to execute this intent is performed.

```
terminated(intends(P,U,M), T) :- happened(A,T), holds(bel(P,responsive(A,U,M)),T).
happened(A,T+1) :- holds(intends(P1,U,M),T), uttered(P1,U,T), action(U,A).
```

Second, the *Plausibility Burden of Proof* asks the question: does the speaker believe that the listener is *capable* of doing action A? Burden of proof rules similar to that of Speaker Intent can be formulated for these other reasoning modes as well. For example, for Plausibility reasoning, the key question is whether there is mutual knowledge about the capability of the listener.

```
holds(intends(P1,U,M),T) :- not -holds(mutual(P1,capable(P2,A)),T),
    uttered(P1,U,T), epistemicAction(U,M,A,_), direct(U,M), holds(speaker(P1),T),
```

```
    holds(listener(P2),T), time(T), not holds(intends(P1,U,M2),T), indirect(U,M2).
holds(intends(P1,U,M),T) :- holds(mutual(P1,capable(P2,A)),T), uttered(P1,U,T),
    onticAction(U,A), indirect(U,M), meaning(M), holds(speaker(P1),T),
    holds(listener(P2),T), time(T), not holds(intends(P1,U,M2),T), direct(U,M2).
```

Third, the *Preference Burden of Proof* asks the question: does the speaker believe that the listener is *willing* to do action A? This has rules analogous to the Plausibility Burden of Proof, with $willing(P2, A)$ replacing $capable(P2, A)$.

Finally, the *Normativity Burden of Proof* asks the question: does the speaker believe that the listener is *normatively obliged* to do action A? This has rules analogous to the Plausibility Burden of Proof, with $obligated(P2, A)$ replacing $capable(P2, A)$.

3.3 Contextual Evidence

As shown in Figure 1, facts and perceivable situation-specific aspects of a context connect to the burdens of proof. To represent this, we can specify some optional ancillary rules as well as commonsense knowledge that is associated with each dimension of fact. First, the **task**-related constraint aspect of context, which relates to limitations imposed by the nature of the action or of the task, as understood by the interactants. Second, the interactant **roles** aspect of context, which relates to the duties, expectations, and obligations of the interactants derived from their respective social or occupational status; whether one interactant has authority over the other, for example. Third, the **interactional history** aspect of context, which relates to the utterances and actions that have been performed in recent memory by the current interactants. Finally, the **co-presence** aspect of context, which relates to the objects and participants that are present when the utterance is made. The first three of these are motivated by dialogue context models, as for example surveyed by Jokinen and McTear (2009). The last of these is motivated by Clark's copresence heuristics (1981).

The contextual evidence is used by domain-general reasoning in the following way. To determine if a speaker wants (or does not want) the listener to do an action A as analyzed by the speaker intent reasoner, we will need to consider whether the speaker intends to have the action accomplished in the first place (i.e., it is the speaker's goal to get the action done) and whether the speaker would wish to delegate this action to the listener. We can establish this line of reasoning with the following three rules. The first rule states what is needed for the speaker to want the listener to do an action. The second and third rules state how this "want" can be negated.

```
wants(P1,goal(P2,A),T) :- holds(goal(P1,A),T), holds(delegate(P1,P2,A),T),
    holds(mutual(P1,isAvailable(P2)),T), action(_,A).
-wants(P1,goal(P2,A),T) :- -holds(goal(P1,A),T),
    holds(mutual(P1,isAvailable(P2)),T), action(_,A).
-wants(P1,goal(P2,A),T) :- -holds(delegate(P1,P2,A),T),
    holds(mutual(P1,isAvailable(P2)),T), action(_,A).
```

The following commonsense default rules specify how contextual evidence contributes to reasoning:

```
% TASK-RELATED
holds(goal(P,A),T) :- holds(requires(K,A),T), holds(currentTask(K),T),
    holds(speaker(P),T), onticAction(_,A).
-holds(goal(P,A),T) :- holds(currentTask(K),T), holds(requires(K,Q),T),
    epistemicAction(_,_,A,Q), holds(speaker(P),T).

% ROLE EXPECTATIONS
holds(delegate(P1,P2,A),T) :- holds(role(P1,leader),T), holds(role(P2,follower),T),
    onticAction(_,A).
-holds(delegate(P1,P2,A),T) :- holds(role(P1,follower),T), holds(isPresent(P2),T),
    onticAction(_,A).
:- holds(role(P,follower),T), holds(role(P,leader),T).

% INTERACTIONAL HISTORY
holds(delegate(P1,P2,A),T) :- happened(A,S), time(S), S<T, holds(speaker(P1),T),
    holds(listener(P2),T), onticAction(_,A).

% CO-PRESENCE
```

```
observed(P1,capable(P2,A),T) :- observed(P1,responsive(A,_,_),T),
    holds(isPresent(P2),T), onticAction(_,A).
holds(delegate(P1,P2,A),T) :- holds(mutual(P1,capable(P2,A)),T), onticAction(_,A).
```

The above discussion specified the evidence and reasoning with respect to Precondition 1 (speaker intent). The Preconditions 2-4 involve analogous reasoning, which is omitted for reasons of space.

4 Example Interaction

The following example shows how speaker observations update the interactional history evidence over several dialogue turns. Of particular interest is the utterance at time step 7, which has a literal meaning of an *Ask* for information but an intended meaning of a *Request* for action, making it an ISA.

```
% Initial
% p1 and p2 are present at t=0
initiated(isPresent(p1;p2),0).

% setting the current joint task
initiated(currentTask(arranging),0).

%relevant
initiated(requires(arranging,goForward; arranging,stop;arranging,grabItem;
    arranging,turn;arranging,pushItem; arranging,carryItem;arranging, moveToLoc),0).

% p1 observed that p2 is present
observed(p1,isPresent(p2),0).
initiated(supports(isPresent(p2), isAvailable(p2)),0).
observed(p1,instrumentation(p2,arms),0).
intiated(supports(instrumentation(p2,arms), capable(p2,grabItem;p2,pushItem;
    p2,carryItem)),0).
initiated(role(p1,leader),0).
initiated(role(p2,follower),0).

% Narrative
% t=1. p1 says: "You're gonna go forward"
holds(intends(p1,u1,request),1).

% t=2. p1 says: "stop"
happened(goForward,2).
observed(p1,responsive(goForward,u1, request),2).
holds(intends(p1,u2,request),2).

% t=3. p1 says: You're gonna turn right 90 degrees
happened(stop,3).
observed(p1,responsive(stop,u2,request),3).
holds(intends(p1,u3,request),3).

% t=4. p1 says: Can you grab the second box
happened(turn,4).
observed(p1,responsive(turn,u3,request),4).
holds(intends(p1,u4,request),4).

% t=5. p1 says: Can you push forward
happened(grabItem,5).
observed(p1,responsive(grabItem,u4, request),5).
holds(intends(p1,u5,request),5).

% t=6. p1 says: just go forward
happened(pushItem,6).
observed(p1,responsive(pushItem,u5, request),6).
holds(intends(p1,u6,request),6).

% t=7. p1 says: Can you grab the box?
happened(goForward,7).
observed(p1,responsive(goForward,u6, request),7).
uttered(p1,u7,7).
hasSurfaceForm(u7,question).
```

```
onticAction(u7,grabItem).
```

5 Conclusion

In this paper we present a novel cognitive scientific model of the reasoning involved in automatically interpreting Indirect Speech Acts (ISAs), empirically derived from a corpus analysis. The approach is grounded in existing cognitive science and psycholinguistic theories and provides a domain-general formal axiomatization of the reasoning needed to uncover the intended meanings of ISAs.

Per the approach, the listener reasons about the speaker's intentions and beliefs, using contextual evidence that both the speaker and listener have access to. Unlike previous approaches, we incorporate reasoning about a speaker's beliefs about the listeners capabilities, socio-normative standards, and speaker preferences. Moreover, extending past approaches, we provide a way to incorporate various aspects of extra-linguistic context into this reasoning process. Following traditions of building formal representations from corpora, we ensure that our representational choices completely cover utterances across several task-oriented dialogue corpora that feature ISAs. We focus on those utterances that have similar surface form, but different meanings due to contextual factors. In future work, we intend exploring how linguistic norms of directness, politeness and brevity, associated with ISAs (Lockshin and Williams, 2020; Wen et al., 2020) can be incorporated into the reasoning process either within or as an extension to Brown's model. We also plan to compare our approach to existing models in inferring ISAs.

We encode the formalism in a logic-based language for knowledge representation and reasoning. We use a logical representation for several reasons. First, such a representation allows for *explainable* processing; specifically, a system would be able to justify the "thought-process" behind a given interpretation. Such explainability allows the human user to detect mistakes in and debug the system, and helps build trust in the system (Reiter, 2019). Second, such a representation could be used to support online learning: if a system produced a given interpretation, and it later discovered that the interpretation was incorrect, it would have an explicit representation that it could use to reason about whether the error was in its contextual evidence, its domain-general rule application, or its preconditions. Although it may be argued from a data-driven perspective that if enough of the right type of data were embedded in the right input vector then BERT might be sufficient, the thrust of this paper is that statistical correlations do not capture the intricate reasoning that is needed to derive the meaning of ISAs. Furthermore, it is currently unknown how to perform abductive logical reasoning in a neural net. While there have been recent attempts in producing neural architectures capable of simple one-hop logical deductions, the type of reasoning we have explicated here is more sophisticated involving non-monotonic inference, default reasoning with first-order terms; we chose ASP because these features come out of the box.

Ultimately, we present this model as a standalone scientific contribution in its own right, building on the work of Brown, Clark, and the others cited in Section 3; a contribution in the fields of cognitive science and psycholinguistics. Future work will involve testing this approach on corpora of ISA Schemas (Roque et al., 2020), and considering additional types of contextual evidence and domain-general reasoning.

Acknowledgment

We are grateful to the anonymous reviewers for their helpful comments.

References

Nicholas Asher and Alex Lascarides. 2001. Indirect speech acts. *Synthese*, 128(1-2):183–228.

John Langshaw Austin. 1975. *How to do things with words.* Oxford university press.

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

Maxwell Bennett, Tom Williams, Daria Thames, and Matthias Scheutz. 2017. Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6589–6594. IEEE.

Gordon Briggs, Tom Williams, and Matthias Scheutz. 2017. Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction*, 6(1):64–94.

Gretchen P Brown. 1980. Characterizing indirect speech acts. *Computational Linguistics*, 6(3-4):150–166.

Herbert H Clark and Catherine R Marshall. 1981. Definite knowledge and mutual knowledge. *Elements of Discourse Understanding*.

Philip R Cohen. 2019. Back to the future for dialogue research: A position paper. In *The Second AAAI Workshop on Reasoning and Learning for Human-Machine Dialogues (DEEP-DIAL 2019), at the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*. Keynote Talk.

Jan P De Ruiter. 2012. *Questions: Formal, functional and interactional perspectives*, volume 12. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Martin Gebser, Benjamin Kaufmann, Roland Kaminski, Max Ostrowski, Torsten Schaub, and Marius Schneider. 2011. Potassco: The potsdam answer set solving collection. *AI Communications*, 24(2):107–124.

Michael Gelfond and Yulia Kahl. 2014. *Knowledge representation, reasoning, and the design of intelligent agents: The answer-set programming approach*. Cambridge University Press.

Michael Gelfond and Vladimir Lifschitz, 1990. *Logic programs with classical negation*, pages 579–597. MIT Press.

David Gordon and George Lakoff. 1975. Conversational postulates. *Syntax and semantics 3: Speech acts*.

Nancy Green and Sandra Carberry. 1999. Interpreting and generating indirect answers. *Computational Linguistics*, 25(3):389–435.

Elizabeth A Hinkelman and James F Allen. 1989. Two constraints on speech act ambiguity. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 212–219. Association for Computational Linguistics.

Filip Ilievski, Piek Vossen, and Stefan Schlobach. 2018. Systematic study of long tail phenomena in entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 664–674, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Salud María Jiménez-Zafra, Roser Morante, Maite Martin, and L. Alfonso Ureña-López. 2018. A review of Spanish corpora annotated with negation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 915–924, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Kristiina Jokinen and Michael McTear. 2009. *Spoken dialogue systems*. Morgan & Claypool Publishers.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1998. Johns Hopkins LVCSR Workshop-97, Switchboard discourse language modeling project, Final Report.

Robert Kowalski and Marek Sergot. 1989. A logic-based calculus of events. In *Foundations of knowledge base management*, pages 23–55. Springer.

John E Laird. 2012. *The Soar cognitive architecture*. MIT press.

Stephen C Levinson. 2001. Pragmatics. In *International Encyclopedia of Social and Behavioral Sciences: Vol. 17*, pages 11948–11954. Pergamon.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. *CoRR*, abs/1710.03957.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2019. A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392, Hong Kong, China, November. Association for Computational Linguistics.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, Copenhagen, Denmark, September. Association for Computational Linguistics.

Jane Lockshin and Tom Williams. 2020. We need to start thinking ahead: The impact of social context on linguistic norm adherence. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in minecraft. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 5405–5415.

C Raymond Perrault and James F Allen. 1980. A plan-based analysis of indirect speech acts. *Computational Linguistics*, 6(3-4):167–182.

Ehud Reiter. 2019. Natural language generation challenges for explainable AI. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7.

Frank E Ritter, Farnaz Tehranchi, and Jacob D Oury. 2019. ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(3):e1488.

Antonio Roque, Alexander Tsuetaki, Vasanth Sarathy, and Matthias Scheutz. 2020. Developing a corpus of indirect speech act schemas. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Vasanth Sarathy and Matthias Scheutz. 2019. On resolving ambiguous anaphoric expressions in imperative discourse. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*.

Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*, pages 165–193. Springer.

John R Searle. 1975. Indirect speech acts. *Syntax & Semantics, 3: Speech Act*, pages 59–82.

Arpit Sharma, Nguyen H Vo, Somak Aditya, and Chitta Baral. 2015. Towards addressing the winograd schema challenge—building and using a semantic parser and a knowledge hunting module. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. SCARE: A situated corpus with annotated referring expressions. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakesh, Morocco, May.

Ruchen Wen, Mohammed Aun Siddiqui, and Tom Williams. 2020. Dempster-Shafer theoretic learning of indirect speech act comprehension norms. In *AAAI*, pages 10410–10417.

Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. 2015. Going beyond literal command-based instructions: extending robotic natural language interaction capabilities. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Tom Williams, Daria Thames, Julia Novakoff, and Matthias Scheutz. 2018. Thank you for sharing that interesting fact!: Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 298–306. ACM.

Sabrina Wilske and Geert-Jan Kruijff. 2006. Service robots dealing with indirect speech acts. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4698–4703. IEEE.