Advances in Cognitive Systems X (2013) X-XX Submitted XX/XXXX; published XX/XXXX

Towards an Indexical Model of Situated Language Comprehension for Real-World Cognitive Agents

Shiwali Mohan SHIWALI@UMICH.EDU
Aaron H. Mininger MININGER@UMICH.EDU
John E. Laird LAIRD@UMICH.EDU
Computer Science and Engineering, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

We propose a computational model of situated language comprehension based on the Indexical Hypothesis that generates meaning representations by translating amodal linguistic symbols to modal representations of beliefs, knowledge, and experience external to the linguistic system. The model incorporates multiple information sources including perceptions, domain knowledge, and short and long-term experiences during comprehension. We show that exploiting diverse information sources can alleviate ambiguities that arise from contextual use of under-specific referring expressions and unexpressed argument alternations of verbs. The model is being used for supporting linguistic interactions in Rosie, an agent implemented in Soar that learns from instruction.

1. Introduction

As computational agents become pervasive in human society as collaborators, the challenge of supporting flexible human-agent interaction is becoming increasingly important. Natural language has emerged as a strong contender for the modality of human-agent interaction as it is the primary means of communication in human societies. Recent research in the design of interactive, intelligent agents has shown that linguistic interaction is not only useful in collaboration for human-agent tasks (Fong, Nourbakhsh, & Dautenhahn, 2003; Kollar et al., 2010), but it also facilitates novel concept acquisition in interactive agents (Cantrell, Schermerhorn, & Scheutz, 2011; Tellex et al., 2011; Mohan et al., 2012). This has motivated research on comprehensive models of natural language for collaborative task execution and learning in human-agent teams.

In a joint activity, the speaker and the hearer try to achieve diverse communicative goals in order to make progress on a task. Various types of utterances are employed for expressing the communicative goal. Imperative sentences such as *Take out the trash* convey that the speaker intends the hearer to complete a task. The joint communicative goal is for the hearer to identify the intended task, relevant objects, and correctly instantiate the task goals. Other utterances such as assertions (*Rice is in the pantry*) may be used to establish shared beliefs about the state for joint activity. Questions (*Where is the milk?*) may be used to supplement perceptual information by relying on the collaborative partner's knowledge.

Communication between collaborators who are simultaneously embedded in a shared task is *situated*. The speaker's linguistic utterances refer to objects, spatial configurations, and actions in

the shared environment. To respond and react to utterances, the hearer should be able to associate the amodal linguistic symbols (*words*) and constructions (*phrases*) to modal representations of perceptions, state, domain knowledge, goals, and policies that are required for reasoning about and manipulating the environment.

Being situated provides a common ground of shared perceptions, goals, and domain knowledge which can be exploited during linguistic communication. Information that is apparent from the current state of the environment or is a component of shared beliefs can be left out of the linguistic utterance by the speaker. This results in more efficient (fewer words) but ambiguous utterances. Humans frequently use referring expressions such as *it* or *that cylinder* that do not by themselves provide enough discriminative information for unambiguous resolution. The speaker assumes that the hearer can exploit extra-linguistic information, such as the context of the ongoing discourse, for unambiguous comprehension. Certain imperative sentences such as *take out the trash* incompletely specify the action by omitting information such as the location where the *trash* should be moved to. Such ambiguities make situated comprehension a significant challenge for interactive agents.

1.1 Contribution and Claims

In this paper, we study the utility of the Indexical Hypothesis (Glenberg & Robertson, 1999) in developing comprehension models for collaborative agents. These agents are embedded in real-world tasks which require the use of complex representations for probabilistic perceptual processing, continuous spatial reasoning, and goal-oriented task execution. To support situated communication, comprehension models must not only perform syntactical analyses, but also synthesize meaning representations by associating linguistic information with representations in other cognitive modules.

The Indexical Hypothesis explains how sentences become meaningful through grounding their interpretation in situated action. The hypothesis asserts that comprehending a sentence requires three processes: first, *indexing* words and phrases to referents that establishes the contents of the linguistic input; second, *deriving* affordances from these referents; and third, *meshing* these affordances under the guidance of physical constraints along with the constraints provided by the syntax of the sentence. According to the hypothesis, the linguistic information specifies the situation by identifying which components (objects, relationships, etc.) are relevant, and the semantic and experiential knowledge associated with these components augments the linguistic input with details that are required for reasoning and taking action. In this formulation of language comprehension, linguistic symbols (words) and constructions (grammatical units) are cues to the hearer to search their perceptions, domain knowledge, and long and short-term experiences to identify the referents intended by the speaker and to compose them together.

Earlier work on indexical comprehension identifies the processes that humans use for comprehension (Glenberg & Robertson, 1999) and provides supporting empirical data from human studies (Kaschak & Glenberg, 2000). It does not describe the knowledge representations and computational processes required for implementation of such models on intelligent agents. The contribution of our work is a computational model of indexical comprehension that precisely defines the representations and processes described in the Indexical Hypothesis. Our claims are as follows.

1. The Indexical Model for comprehension can be used effectively by agents that act and learn in physical environments. This claim is established by demonstration: we describe an implementa-

tion of the model for Rosie (RObotic Situated Instructable Entity), an agent (Mohan et al., 2012) that learns about various aspects of its environment through instruction.

2. The Indexical Model exploits diverse knowledge and experience of the domain to address ambiguity in semantic interpretation of linguistic input. This claim is evaluated by demonstrating the utility of incorporating knowledge and experience in language comprehension on two ubiquitous problems - referring expression resolution and unexpressed argument alternation of verbs.

The rest of the paper is organized as follows. Section 2 provides a description of our robotic domain and a brief overview of Rosie. Section 3 describes the indexical model. In section 4, we describe how the model addresses complexities that arise from ambiguity in its linguistic input. Section 5 discusses the related work on designing comprehension models for agents. We conclude with summarizing the paper and identifying our future directions in Section 6.

2. System Overview

Rosie is a instructable agent implemented in the Soar cognitive architecture (Laird, 2012). It is embodied as a robotic arm that can manipulate small foam blocks on a table-top. The workspace also contains four named locations: *pantry*, *garbage*, *table*, and *stove*. These locations have associated simulated functionalities. For example, a *stove* can be turned on and off, and the *pantry* can be opened and closed. These functionalities change the state of the world. For example, when the *stove* is turned on, it changes the simulated state of an object on it to *cooked*.

2.1 Perception, Actuation, and Interaction

Rosie senses the environment through a Kinect camera sensor. The perception system segments the scene into objects and extracts features for three perceptual properties: color, shape, and size. These properties along with the position and bounding volume of the objects in the world are provided to Rosie and are used for perceptual and spatial reasoning. For locations and objects, the simulated state (such as *open*, *on*, *cooked*) is also included in its description.

To act in the world, Rosie sends discrete primitive commands to the controller. The commands include object manipulation: `pointTo(obj)`, `pickUp(obj)`, and `putDown(x,y)`; and simulated location operation: `open(loc)`, `close(loc)`, `turnOn(stove)`, and `turnOff(stove)`. The robot controller converts these discrete commands to continuous closed loop policies, which change the state of the environment. Human instructors can interact with Rosie through a simple chat interface. Instructor's utterances are pre-processed using the Link-Grammar parser to extract parts of speech and syntactical structure. Rosie responds using semantic structures that are translated to language using templates. The instructor can point to an object by clicking on the object in the camera feed.

2.2 Learning with Instruction

Rosie begins with procedural knowledge for parsing language, maintaining interactions, and learning from instruction. It also knows how to perform primitive actions in the world. Through situated interactive instruction, Rosie can learn novel words along with the concepts they are grounded in. For nouns and adjectives (such as *red*, *large*, or *cylinder*), the agent learns new classifications of perceptual features (color, size, and shape) from interactive training. For prepositions (such as *right*

of), the agent learns combinations of primitive spatial predicates. For verbs (such as *move*), the agent learns novel task knowledge.

Learning in Rosie is active. Whenever it encounters a new word that it cannot comprehend by associating it with known concepts, it initiates interactions to learn the concept and the grounding of the word. The interactions are situated in the environment. Through instruction, the mentor provides specific examples of the concepts in the environment. When the instruction is complete, Rosie can comprehend the word and use the associated concept for classification, spatial reasoning, and action. As the human-agent interaction is linguistic, ambiguities may arise during instruction. A common form of ambiguity arises due to the use of under-specific referring expressions such *it* or *that object*. Other ambiguities arise from imprecise description of actions such as *move the red cylinder to the table* that does not identify what relationship should be established between the *red cylinder* and the *table*. Rosie's comprehension model must be able to alleviate such ambiguities by incorporating information from its state perceptions, domain knowledge, and experience.

We now give a brief description of the representations used in Rosie. Detailed explanations can be found in our earlier work (Mohan et al., 2012). Rosie's beliefs about the current state are held in its working memory. These beliefs are *object-oriented* and are derived from its perceptions of the world, from its experiential knowledge of the world encoded in its long-term memory, and interactions with the collaborator human.

Rosie's visual knowledge is encoded in its perceptual memory. It accumulates training examples that are used to classify objects in terms of visual attributes: color, size, and shape. Each visual attribute has a kNN classifier associated with it. Each class within the kNN is referred to using a perceptual symbol. For example, the domain of the *color* attribute may contain perceptual symbols `C22`, `C53`, `C49`, each of which correspond to colors known to Rosie. All perceptible objects are represented in working memory along with the known value assignment to its visual attributes.

Rosie's spatial knowledge is distributed between its semantic memory and spatial-visual system (SVS). Rosie learns and represents spatial prepositions such as *on* and *near* as compositions of known primitive spatial literals in SVS that encode alignment along axes and distance between objects. It generates symbolic representations of spatial relationships between perceptible objects using this knowledge. This representation is useful for reasoning about existing spatial relationships on the workspace (such as *the red cylinder is on the stove*) and executing actions that establish specific spatial relationships between arguments (such as *put the red cylinder on the stove*).

Rosie can learn goal-oriented tasks such as *cook a steak*, that require it to achieve a composition of spatial and state predicates by executing a policy defined hierarchically over primitive actions. Its task knowledge is distributed across its semantic and procedural memories. While the semantic memory stores a task-concept network that includes the goal definition of the task and constraints over how the goal should be instantiated, the procedural memory encodes the task's availability conditions, policy, and termination conditions represented as rules and implemented through operator proposal, selection, and application.

3. Indexical Comprehension

Consider the imperative sentence *move the large red cylinder to right of the blue triangle*. We assume that through this sentence, the collaborator intends for the hearer to execute the requested action. The goal indexical comprehension is to identify the referents of the linguistic input and

compose them to generate an action instantiation that is grounded in the modal symbols that Rosie uses to reason about and manipulate its environment. Following the Indexical Hypothesis, comprehension is carried out as follows.

3.1 Indexing

After preliminary lexical processing, it is established that the linguistic input contains two referring expressions (REs: *a red cylinder* and *the blue triangle*), a spatial preposition (*to the right of*), and a verb (*move*). The goal of the *indexing* step is to identify the referents for these linguistic units. The model uses a simple referential grammar: nouns and adjectives refer to visual properties; referring expressions refer to objects; prepositions refer to spatial relationships; and verbs refer to tasks. Figure 1 shows the objects (O12, O32) and semantic networks A, B, and C that form the referent set of REs, prepositions, nouns/adjectives, and verbs. We introduce the term *indexical maps* for structures in semantic memory that encode how linguistic symbols (nouns/adjectives, spatial prepositions, and action verbs) are associated with perceptual symbols, spatial compositions, and action-concept networks. We describe how indexical maps are used during comprehension below.

3.1.1 Indexing Referring Expressions

To index REs (*the red cylinder*), the model must first index the descriptive words (*red* and *cylinder*). For each of these words, the model queries the semantic memory for a node that was previously learned to be associated with the lexical string. For the string *red*, the memory returns node L1 (refer to Figure 1). Node L1 *maps* the lexical string *red* to the corresponding perceptual symbol C22 which is a class in the color classifier. Once the model has retrieved perceptual symbols for all words, it searches working memory for objects that have the required perceptual symbols. These objects are the intended referents of the RE. In cases where the RE is under-specific (e.g. *this block*), there may be multiple objects that match, resulting in ambiguity. The model can use other kinds of information to resolve such ambiguities (details in Section 4.1). For the sake of simplicity now, we assume that only one object (O12) matches the cue. This object is included in the referent set ($R^o_{r,c} = \texttt{O12}$) for the RE *the red cylinder*. Similarly, $R^o_{b,t} = \{\texttt{O32}\}$ for the RE *the blue triangle*. If these sets are empty, it is an indication that Rosie lacks knowledge to generate groundings of the RE in which case it would prompt the human instructor to provide examples to learn from.

3.1.2 Indexing Prepositions

Prepositions are indexed in a similar fashion. For a preposition string (*right*), the model queries the semantic memory for an indexical node that had previously been learned and is associated with it. On the retrieval of the requested node (*P3*), the model creates the referent set $R^s_{right} = \{\texttt{P3}\}$. If the set is empty, Rosie asks the instructor to provide an example of *right-of* in the environment.

3.1.3 Indexing Verbs

To index the verb *move*, the model queries the semantic memory for a node that is connected to the string *move*. The memory returns the node L2. Then the model retrieves the mapping node M2 that associates the verb to domain knowledge of the task - the goal definition G2 and procedural operator node P2. The referent set for the verb consists of the task-concept network, $R^a_{move} = \{\texttt{(M2,P2,G2)}\}$.

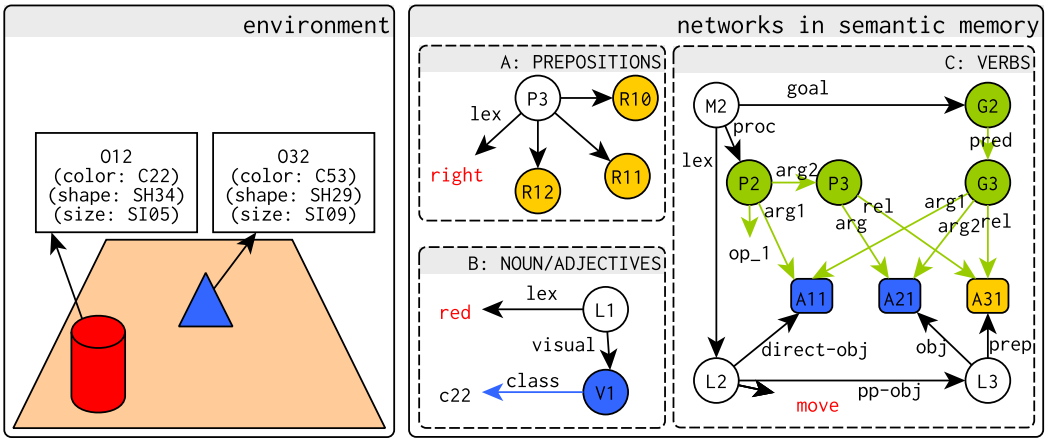

Figure 1: Environment state and the knowledge encoded in Rosie's semantic memory. The white nodes (circles) represent indexical maps between amodal linguistic symbols (in red) and modal domain knowledge. Yellow nodes represent spatial symbols and slots (round rectangles), blue nodes represent visual symbol and slots, and green nodes represent procedural symbols.

3.2 Instantiate Domain Knowledge

Once the referents have been identified, the next step is to retrieve the domain knowledge associated with them and instantiate it under the syntactical constraints of the sentence. The model begins by retrieving the previously learned syntactical nodes associated with the verb *move*. The sentence *move the large red cylinder to right of the blue triangle* has a direct-object (RE, *the large red cylinder*) and a prepositional-object (RE, *the blue triangle*) connected to the verb through the preposition *right*. Following this syntactical structure, the model retrieves the direct-object node `A11` and the pp-object node `L3`. `L3` is further expanded to retrieve nodes `A21` and `A31`. `A11` and `A21` are slot nodes that can receive sets of objects in the environment. `A11` is filled by $R^o_{r,c}$ as *the red cylinder* is the direct-object of the verb *put* and `A21` is filled by $R^o_{b,t}$ as *the blue cylinder* is the RE in the prepositional phrase of the verb. `A31` is a spatial slot which is filled by R^s_{right}, the referent set for *right*.

Then, the model expands the domain knowledge nodes `P2` and `G2`. The subgraph (`P2, P3, A11, A21, A31`) governs how the policy operator `op_1` is instantiated. The subgraph (`G2, G3, A11, A21, A31`) governs the instantiation of goal of the task. The values of the slot nodes (`A11, A21, A31`) determine the contents of the goal and the policy operator `op_1`. Instantiation of domain knowledge results in the interpretation set I_s which contains elements that encode: `execute` policy `op_1` over arguments drawn from sets $R^o_{r,c}$, $R^o_{b,t}$, R^s_{right} until the goal defined over them is achieved.

This step was described as *deriving the affordances* in the original formulation of the Indexical Hypothesis. However, the term *instantiating domain knowledge* better describes our formulation.

3.3 Meshing

The interpretation set I_s is the set of different groundings of the imperative sentence. I_s can have several elements arising from under-constraining cues in the linguistic input. However, only a subset

of these groundings can be executed in the environment given physical constraints and spatial relationship between objects. Let A be a set of tasks that can to be executed in the current state based on their availability conditions. Intersection of the sets, $I_s \cap A$ is the set of tasks that the instructor intends Rosie to execute. If this set contains a single element, that task operator is selected and executed. If this set contains multiple elements, further interaction or internal reasoning is necessary for resolution. The cardinality of the referent sets (R) is used to determine the source of the ambiguity. Rosie asks questions to gather information that will reduce the cardinality of the ambiguous set. If the $I_s \cap A = \phi$, Rosie does not have enough knowledge to generate the correct groundings for the required task. This is an opportunity to learn the task. Rosie then begins a learning interaction by prompting the human collaborator to present an example execution of the task.

4. Dealing with Complexities

Various issues can arise while attempting to generate an unambiguous and complete interpretation of an utterance. Ambiguities arise when the linguistic cues under-specify their referents resulting in multiple elements in their referent sets and consequently, multiple interpretation. One such ambiguity arises due to the contextual use of ambiguous referring expressions. We describe how this ambiguity is addressed in the Indexical Model in Section 4.1. Other issues arise when the information required to instantiate a policy is not completely specified in the linguistic input. An example of this is unexpressed argument alternation of verbs. This is addressed in Section 4.2.

4.1 Reference Resolution

Humans use a variety of surface forms to refer to the same entity. A few of these forms, such as definite noun phrases (*the large red cylinder on the table*), may uniquely identify the intended referent from the current shared perceptions. However, a majority of REs encountered in conversations, such as noun phrases with indefinite determiners (*a cylinder*), demonstrative/diectic pronouns (*this*, *that*), and personal pronouns (*it*) are ambiguous.

For generation and comprehension of REs, the communicative goal is the identification of the intended object by the hearer. The form of REs and other linguistic (word order) and phonetic (intonation) aspects are influenced by the co-operative speaker's assumptions about the relative salience of referents to the hearer. An object might gain more salience than others because it is useful in performing a task, it is being pointed at, it changes appearance, or it is unexpected. The ongoing discourse also makes objects more salient. Speakers make assumptions about which objects are more salient to the hearers and use these assumptions to choose an appropriate RE. More salient objects can be referred to by less informative REs as the hearer can exploit saliency for disambiguation.

Gundel et al. (1993) express the notion of the current and historical salience of an object to the hearer as its *cognitive status*. They propose the Givenness Hierarchy (GH) that relates the cognitive status of objects with different appropriate RE surface forms. The GH identifies six cognitive statuses, however, only four are relevant to this paper. Those four statuses (and the appropriate REs) are: *in-focus* (personal pronouns) $>$ *activated* (demonstrative pronouns, demonstrative noun phrases) $>$ *uniquely-identifiable* (definite noun phrase) $>$ *type-identifiable* (indefinite noun phrase). Each status in the GH is the necessary and sufficient condition for use of the corresponding RE and entails all the lower statuses. The choice of a RE form by the speaker is indicative of what cognitive

status is useful for resolution. Given that cognitive status of an object and the hearer's knowledge about the environment, the information in the RE uniquely identifies the intended referent.

4.1.1 Non-linguistic Contexts

Knoeferle and Crocker (2006) identify two dimensions of the interaction between the linguistic and situated context: *informational* and *temporal*. The informational dimension indicates that along with language, hearers use perceptual information and domain knowledge (discussed earlier in Section 3) for comprehension. The temporal dimension indicates that cognitive attentional processes are closely associated with utterance generation and comprehension. While REs such as noun phrases (lower in the GH) exploit the informational dimension of language-context interaction, ambiguous REs such as pronouns (higher in the GH) exploit the temporal dimension. To process the complete range of RE forms in the GH, the model exploits both the informational (described earlier) and the temporal dimensions (described below).

Interaction: When conversation participants communicate, they focus their attention on only a small portion of what each of them perceives, knows, and believes. Some entities (objects, relationships, actions) are central to information transfer at a certain point in dialog and hence, are more salient than others. This is exploited by both the speaker and the hearer. It allows the speaker to refer to focused entities with minimal information and allows the hearer to heuristically constrain the set of possible referents, reducing cognitive load on both.

Rosie has a model of instructional interaction (Mohan et al., 2012) that is based on the computational theory of task-oriented discourse by Grosz and Sidner (1986) that organizes the discourse structure according to the goals of the task. The current state of the human-agent interaction is represented by three elements. *Events* cause change, either in the environment (`actions` such as `pick-up(O32)`), the dialog (`utterances` such as *Where is the red cylinder?*), or Rosie's knowledge (`learning events` such explanation-based learning). A *segment* is a contiguous sequence of events that serves a specific *purpose* and organizes the dialog into purpose-oriented, hierarchical blocks in accordance with Rosie's goals. The *purpose* of a segment is determined based on pre-encoded heuristics about instructional interactions. Finally, the focus of the interaction is captured in a *stack* of active segments. When a new segment is created, it is pushed onto the stack. The top segment of the stack influences the agent's processing by suggesting a purpose that Rosie should act to achieve. When the purpose of the top segment is achieved, it is popped from the stack.

The *stack* maintains a set of all referents (objects, spatial predicates, actions) that are related to the events in the active segments of the stack. The set of objects (O^{stack}) is most pertinent to this paper because this set identifies all objects that have been referred to in the current discourse, making them more salient than other perceivable objects.

Attention: Object referents that have been brought to attention, either through linguistic or extra-linguistic means, but are not in the focus of the ongoing communication are usually referred to by demonstrative pronouns or demonstrative noun phrases (*this*, *that cylinder*) (Gundel et al., 1993). The extra-linguistic means may include unexpected behavior and pointing by the speaker. To resolve such REs, Rosie must maintain the history of references to objects its perceptions.

To capture attention to various objects in the environment, we use the architectural recency-based activation in Soar's semantic memory. The recency-based activation biases retrieval from semantic memory towards the most recently accessed memory. An object is *accessed* only if it

was pointed at or was a component of an action or learning. Anytime an object is accessed Rosie stores its representation in its semantic memory. Each store boosts the activation of the object in accordance with recency computation. A completely ordered subset O^{active} of the highest activated n objects is retrieved in the Rosie's working memory. These are combined with objects in focus to give a set of objects Rosie is *attending* to ($O^{attend} = O^{stack} \cup O^{active}$). This formulation of attention combines linguistic and extra-linguistic salience.

4.1.2 Resolving References in the Indexical Model

In section 3.1 we described indexing of referring expressions in simple cases where the words in the RE and their corresponding perceptual symbols by themselves uniquely identified the referent object. The following steps give the details on how an ambiguous RE is indexed to objects by incrementally adding diverse types of information.

0. *Maintain cognitive status*. Following the Givenness Hierarchy, the model maintains different cognitive statuses for objects.
   - Objects in the interaction stack (O^{stack}) have the *in-focus* status.
   - Objects that are being attended to (O^{active}) have the *activated* status.
   - Objects in perceptions ($O^{percept}$) have the *identifiable* status.
1. *Assign resolution type*. For any RE r, the model determines its resolution type based on its surface form. If the RE is -
   - a definite noun phrase (*the red cylinder*), demonstrative pronoun (*this*), or personal (*it*) pronouns, the speaker has a specific intended referent and comprehension should unambiguously determine it (*unique* resolution).
   - an indefinite noun phrase (*a red cylinder*), it indicates that there is no specific intended referent and any object that fits the noun phrase can be used for resolution (*any* resolution).
2. *Determine the candidate referent set*. The model exploits the heuristic that surface forms of REs are indicative of which set contains the intended referent. The candidate referent set is -
   - $R^o_r = O^{stack}$ for pronouns (*it*),
   - $R^o_r = O^{attend}$ for demonstrative pronouns (*this*, *that*) and noun phrases (*this cylinder*),
   - $R^o_r = O^{percept}$ for definite (*the cylinder*) and indefinite (*a cylinder*) noun phrases.
3. *Apply the visual filter*. The visual filter exploits Rosie's knowledge about perceptual symbols and how they relate to words to identify the referents for REs with descriptive words (*the red cylinder*). The model indexes each descriptive word (*red*, *cylinder*) in a noun phrase, and then the model looks up its corresponding perceptual symbol. These perceptual symbols are collected into a set as a cue. All the objects in the candidate set (R^o_r) whose working memory representations do not contain this cue are deleted from this set.
4. *Apply the spatial filter*. If the RE uses spatial reference (*the cylinder on the right of the pantry*), referent sets of both noun phrases ($R^o_{cylinder}$, R^o_{pantry}) are obtained. The model indexes the preposition *right* to retrieve the corresponding spatial relationship predicate `P3`. Items in $R^o_{cylinder}$ that do not satisfy the relationship `P3` with any item in R^o_{pantry} are deleted. This is a meshing step that combines linguistic information with the domain knowledge and the perceptual state.

5. *Apply the task filter*. If the REs are used with verbs, such as in an action command (*put the cylinder in the pantry*), the model can use the knowledge of task restrictions to constrain the interpretation of REs. To access this knowledge, the model indexes the verb to retrieve a task-operator and its corresponding goal. During meshing, it looks at all the task-operator instantiations that are applicable in the current environmental state under the physical constraints and Rosie's knowledge of object affordances. Any object that does not occur in the arguments of currently applicable task instantiations is removed from R_r^o of the RE.
6. *Obtain partial ordering*. The elements of the referent set ($r \in R_r^o$) are partially ordered based on their cognitive status and resolution type. If resolution is *unique* (from step 1) then $r_i \in O^{stack} > r_j \in O^{active} > r_k \in O^{percept}$. If resolution is *any*, then all objects have equal preference.
7. *Resolve*. After applying all available filters, if R_r^o contains only a single object, that object is selected as the intended referent. If it contains multiple objects, the model uses the partial ordering obtained earlier to select the object highest in the order, as the intended referent. If the partial ordering is not informative for resolution, the model initiates a sub-dialog to obtain more information from the instructor. If the resolution is *any*, all objects have equal preference and one is chosen at random.

4.1.3 Evaluation and Analysis

Experiments: We generated a corpus of 25 instructor utterances that addresses different capabilities of Rosie. This corpus contains instruction sequences that teach and query Rosie about objects and their attributes, present and verify grounded examples of spatial prepositions, and teach verbs. This corpus contains references to three objects in the scene. These objects are referred to using varying forms of referring expressions including 12 instances of personal pronouns (such as *it*), 4 instances of demonstrative pronouns (such as *this*), 3 instances of demonstrative phrases (such as *that cylinder*), and 14 varying length noun phrases with different descriptive words (such as *the red cylinder*). We evaluated various models of comprehension that exploit informational (I) and temporal (T) of non-linguistic context. The baseline model *p* uses the context derived from perceptual semantics only. Model *p+t* exploits the restrictions derived from task knowledge along with perceptual semantics. Model *p+t+a* exploits the temporal dimension by encoding the attentional state. Model *p+t+a+d* encodes both the attentional and dialog states. Each of the comprehension models was evaluated using the instruction corpus on different scenarios of increasing perceptual ambiguity in the environment obtained by adding distractor objects as shown in Figure 2.

Evaluation metric: Rosie is an interactive agent that engages the human instructor in a sub-dialog if it fails at any stage in its processing. On failing to resolve ambiguous referring expressions in sentences, Rosie asks questions to obtain more information that will constrain its resolution. The instructor can, then, incrementally provide more identifying information. These question-answer pairs (*object identification* queries) are informative of how ambiguous an RE is to the model given the ambiguity in the current scenario and the contexts. Note that the instructor could have provided all the identifying information in a single response (*Which object?*, *the blue cylinder in the pantry*). However, letting Rosie take the initiative in resolution ensures that it accumulates the minimum information required for unique identification in the current situation. This number of *object identification* queries in such a set up is directly correlated with the number of objects in R_r^o after all filters have been applied.

Model	Description	Dimension
p	perceptual semantics	I
p+t	plus task knowledge	I
p+t+a	plus attentional context	I + T
p+t+a+d	plus dialog context	I + T

Scenario Ambiguity	
Ambiguity 1	only intended referents
Ambiguity 2	perceptually distinct distractors
Ambiguity 3	distractors with different shapes, same colors
Ambiguity 4	distractors with same colors & shapes

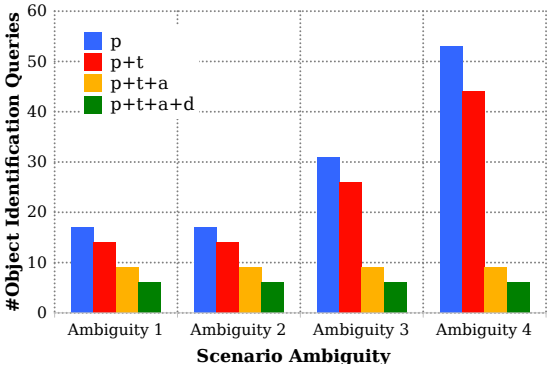

Figure 2: (on left) Different models and scenarios used for evaluation of referring expression resolution. (on right) Number of object queries asked by Rosie for referring expression resolution.

Results: The graph (in Figure 2) shows the number of *object identification* queries asked by Rosie while using different comprehension models in scenarios with varying perceptual ambiguity. The design allows the model to obtain more information through dialog if the RE is ambiguous. The model reliably integrates information provided incrementally over several interactions for resolution. Consequently, all REs were eventually correctly resolved in all models in all scenarios. The model can exploit the informational and temporal dimensions effectively for resolution. In comparison, co-reference resolution in Stanford CoreNLP (Lee et al., 2012) that exploits only the linguistic information, failed to correctly resolve 10 (28.6%) references. These results suggest that grounded contexts are essential for robust comprehension in an embodied agent.

The baseline model *p* that only exploits the contexts derived from perceptual semantics generates the most number of queries for all levels of ambiguity. The model *p+t* is able to use its knowledge about the task to constrain resolution and therefore requires fewer queries for achieving the same resolution results. The models that exploit both the temporal and informational dimensions require even fewer queries to achieve similar performance across all scenarios. Conversing with agents that only encode the informational dimension of non-linguistic context usually requires wordy REs such as *the red cylinder in the pantry* that have to be repeated in all interactions related to that object. The use of informational dimension of the context for comprehension allows the use of shorter referring expressions (*it*, *this cylinder*) resulting in *efficient* communication.

As perceptual ambiguity in the environment increases, models that exploit only the informational dimension (*p*, *p+t*) require more perceptual information for resolving REs. Models that exploit the temporal dimension (*p+t+a*, *p+t+a+d*) ask the same number of queries across all scenarios, demonstrating that use of co-reference is an efficient way of communicating about objects in human-agent dialogs. It allows the instructor to communicate the intended referent without incorporating large amounts of information in utterances in perceptually ambiguous scenarios.

4.2 Unexpressed Argument Alternations of Verbs

In Rosie, the goal of comprehension of an imperative sentence is to correctly instantiate a task that can be executed in the environment. The verb of the sentence identifies the task and its objects (and the corresponding syntactical structure) instantiate the action arguments such that a policy can be

executed in the environment to achieve the goals of the task. The syntactical structure of English verbs is flexible and often omits objects. Consider for an example, an imperative *take the trash out to the curb* that informs the hearer that the direct-object *trash* has to be placed on the location *curb*. An alternative imperative sentence that is used to convey the same meaning is *take the trash out*. The location where the *trash* should be place is left unexpressed. These variations pose a significant challenge to a system that seeks to generate a precise action interpretation of the sentence that can be executed in the environment.

Humans generate and comprehend such sentences by relying on the shared knowledge of the domain. In the example, both the speaker and the hearer know that the *trash* is usually put on the *curb*. This allows the speaker to omit the location in the sentence (*take the trash out*) for the sake of communicative efficiency. The choice of this syntax by the speaker indicates that they assume that the hearer can fill the missing location from their knowledge of the domain. Upon hearing the utterance, the hearer must exploit their domain knowledge and generate an appropriate and complete representation of the action.

4.2.1 Exploiting the Instructional Experience

To deal with imperative sentences with unexpressed information about the action, the model relies on Rosie's prior experiences of interacting with the instructor and acting in the domain. Consider the verb *move* and the variations of the imperatives that can be constructed from it - (a) *move the green object to the right of the table* and (b) *move the green object to the table*. In (a), the direct-object *the green object*, the location *the table*, and the spatial relationship between them (*right of*) are completely specified. In (b), the spatial relationship is omitted with an understanding that there is a default configuration (*on*) between the object and the location that can be used for action.

The default configuration can be extracted from the experience of learning how to perform the *move* task. When Rosie is asked to execute a task for the first time, it leads the instructor through a series of interactions to learn the structure of the task. Let's assume that Rosie does not know how to perform *move*. On getting the imperative sentence (a), Rosie asks a question about the goal (*what is the goal of the task?*). The human instructor replies, *the goal is the green object is to the right of the table*. By analyzing the imperative sentence and the goal description, Rosie extracts a general schema that relates the linguistic structure of the utterance to the goal of the task. It uses a simple heuristic that information (object, location, and relationship) that is specified in the imperative sentence can be generalized away in the goal definition. It is assumed that future instances of the verb *move* will completely specify the goal. At a later stage, Rosie receives the sentence (b). Using its knowledge of the goal definition, Rosie attempts to generate an instantiation. This fails because no relationship is specified. Rosie asks the instructor to describe the goal in this situation. The instructor may reply with *the goal is the green object is on the table*. By comparing the current situation (for sentence (b)) and its experience with sentence (a), Rosie deduces that the verb *move* may be used in two alternations. The representation of *move* is augmented to reflect that if the relationship is not specified, Rosie should attempt to establish the *on* relationship between the object and the location. This augmentation to the task-concept network of the verb *move* is shown in Figure 3.1 as dotted edges and nodes. Note that Figure 3.1 is an augmented version of the network C in Figure 1

To comprehend *move* in later instances, when the comprehension model indexes into this representation of the action, the model can use the default values to complete the argumentation of the

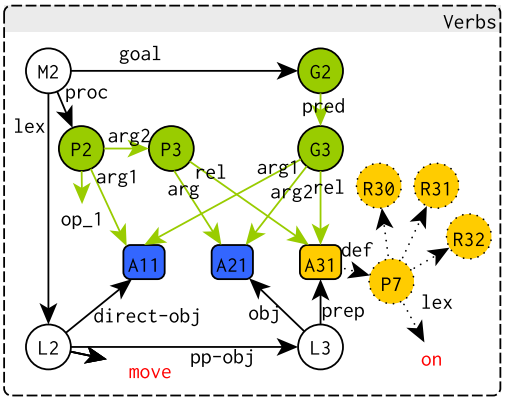

Figure 3.1. Declarative knowledge for *move* after the training episode.

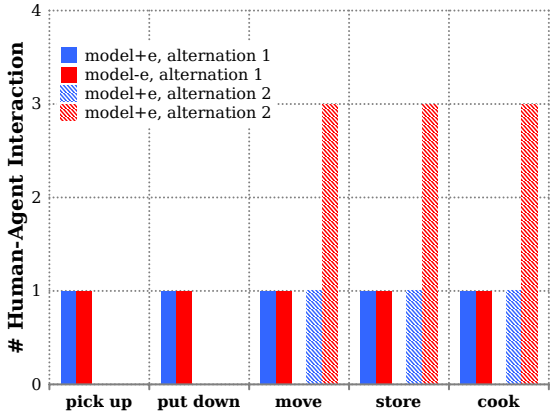

Figure 3.2. Number of interactions required for comprehending verbs with different alternations.

action if those values are not specified in the linguistic input itself. This allows the model to use Rosie's instructional experience to fill in information that is not specified in the linguistic input but is essential for action.

4.2.2 Evaluation and Analysis

Experiment: In an environment with four objects, Rosie was instructed to perform eight instances of five tasks using an equal distribution over the alternations of the relevant verb. The verbs used in the experiment have the following characteristics.

- The verb *pick* takes a direct-object and does not have any missing argument alternation. Example: *pick up the red cylinder*.
- The verb *put* takes a direct-object and a prepositional-object and does not have any alternation. Example *put down the red cylinder on the table*.
- The verb *move* has two alternations. The first alternation specifies the object, the prepositional object, and the spatial relationship (as in *move the red cylinder to the right of the table*). The second alternation does not specify the spatial relationship between the direct and prepositional object (as in *move the red object to the table*).
- The verb *store* has two alternations. The first alternation contains the object, the prepositional object, and the intended spatial relationship between them (in *store the red cylinder in the pantry*). The second alternation leaves the prepositional object unexpressed (as in *store the red cylinder*).
- The verb *cook* has two alternations. The first one specifies the instrument used for cooking along with the object to be cooked (in *cook the steak on the stove*). The second one leaves the instrument unexpressed (in *cook the steak*).

The first two verbs are primitives that have been pre-encoded in Rosie. The last three verbs are acquired through human-agent linguistic interaction. For training, Rosie was taught the task with the first alternation of the verb. After it successfully learns the task, it was asked to perform the task using the second alternation. Any questions asked by Rosie during this training episode were

appropriately answered. Two variations of the comprehension model were evaluated. Model+e uses Rosie's instructional experience to augment the linguistic input that is missing information required for task execution. Model-e is a lesioned version of model+e and does not exploit the instructional experience. It relies on asking the instructor a question for the missing information. Both models were given the same instructional experience (12 interactions for *move* and 16 interactions for *cook*).

Results: The graph (in Figure 3.2) shows the number of interactions that occurred during the comprehension of task commands in model+e (in blue) and model-e (in red). The patterned bars correspond to the first alternation and the plain bars correspond to the second alternation (if applicable) of the verb. For verbs without any alternations (*pick*, and *put*), both models take equal number of interactions to execute the task (one per task instance). For verbs that may have unexpressed argument alternations, the models behave differently for different alternations. For the first alternation in which all information is specified, both models take one interaction per task. However, for the second alternation that leaves some argumentation unexpressed, model+e takes only one interaction per task for performance because it is able to use the knowledge extracted from its learning experience to fill in the missing information. Model-e must ask questions to gather the missing information in unexpressed verb argumentation resulting in more human-agent interactions (3 per task). Both models comprehend both alternations of verbs and correctly execute the task.

5. Related Work

Research in the NLP community has recently begun to look at grounding meanings of sentences in observed world states various domains including navigation tasks (Chen & Mooney, 2011), RoboCup sportscasting (Liang, Jordan, & Klein, 2009), and question answering (Liang, Jordan, & Klein, 2013). The focus of this work has been on acquisition of grounded lexicon and semantic parsers and assumes a fairly simplistic agent with propositional state and action representations. The complexity of language comprehension is completely encoded in the parser, which is learned from aligned corpora of agent behavior and the text that describes it. Such batch-learning, statistical models of comprehension tend to be comprehensive and robust to errors in linguistic input. It is unclear how these models can be extended to collaborative agents that are engaged in situated communication. The simplistic representation of the world state and dynamics poses problems in adapting the comprehension model to agents embedded in physical environments that require complex, relational state representations for reasoning and action.

In the robotics community, grounded comprehension has been studied in the context of describing a visual scene (Roy, 2002), understanding descriptions of scene (Gorniak & Roy, 2004), understanding and spatial directions (Kollar et al., 2010), understanding natural language commands for navigation (Tellex et al., 2011). These comprehension models work with complex state and action representations required for reasoning about physical worlds. Their primary focus has been on acquisition of grounding models through batch-learning from descriptions of robot's perceptions or behavior. They do not address the challenges for comprehension that arise from ambiguities in natural language for interactive agents.

Other research in the robotics community has focused explicitly on problems that arise in situated comprehension. Scheutz et al. (2004) present a visually grounded, filter-based model for reference resolution that is implemented on a robot with audio and video inputs. Ambiguities are resolved by accounting for attentional context arising from fixations in the work area. In a related

work, Kruijff et al. (2007) demonstrated incremental parsing at multiple levels that includes contexts derived from the dialog context and declarative pre-encoded selectional restrictions along with visual semantics. Although these models only address specific issues such as contextual reference resolution, our work can be viewed as an extension of these efforts to develop situated comprehension models for intelligent agents. These works have several desirable qualities including spoken dialog processing and online, incremental comprehension which will inform our future work.

6. Concluding Remarks and Future Work

In comparison to standard approaches to semantics and meaning representations prevalent in NLP community, the Indexical approach to language comprehension affords several advantages. The representation of semantics or meaning can be diverse and modality specific, allowing the use of established representations in the AI community. For example, in order to represent actions, we use pre-conditions, policy, and goals and to reason about environmental dynamics. This is in contrast to previous approaches that either encode semantics as amodal symbols that are not grounded in real-world experiences or that use simple propositional representations that cannot scale to complex environments. An additional advantage to using standard representations is that standard learning algorithms can be exploited to expand the agent's knowledge and its situated comprehension capabilities.

In the formulation of comprehension as a search over short and long-term experiential knowledge, non-linguistic context has a natural role. It provides constraints over the hypotheses space and guides search. Non-linguistic context can be derived from various sources including the ongoing discourse, reasoning, task knowledge, and attentional mechanisms. We show that exploiting diverse contexts in our model is useful in reducing ambiguity in referring expression resolution. Experiential knowledge augments the linguistic input by incorporating knowledge from prior experiences with the environment. This is useful in situations where the linguistic input such as in *take the trash out* is under-specific and does not encode enough information for reasoning and action.

The focus of our future work will be on studying other linguistic ambiguities that may arise in instructional interactions and how they can be addressed by incorporating various situated information sources. An concern is that one verb word may indicate different task goals and policies. For example, the sentences *store the rice* and *store the milk* indicate different goal locations (*pantry* for *rice* and *refrigerator* for *milk*). The comprehension model should be able to use the semantic categorization of the arguments to instantiate the goal with appropriate locations. Other ambiguities arise in determining the site of preposition phrase attachment. In the sentence *store the red cylinder on the green block in the pantry*, it is ambiguous if the phrase *in the pantry* attaches to the verb *store* directly, or to the phrase *on the green block*. This can be resolved by the incorporating the current state of the environment. A different dimension of future research is incremental comprehension that would lead to robust performance on incomplete and ungrammatical linguistic input.

References

Cantrell, R., Schermerhorn, P., & Scheutz, M. (2011). Learning Actions from Human-Robot Dialogues. *Proceedings of the IEEE Symposium on Robot and Human Interactive Communication* (pp. 125–130).

Chen, D. L., & Mooney, R. J. (2011). Learning to interpret natural language navigation instructions from observations. *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 859–865).

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems*, *42*, 143–166.

Glenberg, A. M., & Robertson, D. A. (1999). Indexical understanding of instructions. *Discourse Processes*, *28*, 1–26.

Gorniak, P., & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470.

Grosz, B., & Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*, 175–204.

Gundel, J. K., Hedberg, N., Zacharski, R., & Fraser, S. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*, 274–307.

Kaschak, M. P., & Glenberg, A. M. (2000). Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of Memory and Language*, *43*, 508–529.

Knoeferle, P., & Crocker, M. W. (2006). The Coordinated Interplay of Scene, Utterance, and World Knowledge: Evidence from Eye Tracking. *Cognitive Science*, *30*, 481–529.

Kollar, T., Tellex, S., Roy, D., & Roy, N. (2010). Toward understanding natural language directions. *Proceeding of the 5th ACM/IEEE International Conference on Human-robot Interaction - HRI '10* (pp. 259–267).

Kruijff, G.-J., Lison, P., Benjamin, T., Jacobsson, H., & Hawes, N. (2007). Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. *Proceedings from the Symposium on Language and Robots* (pp. 55–64).

Laird, J. E. (2012). *The Soar cognitive architecture*. MIT Press.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2012). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*.

Liang, P., Jordan, M. I., & Klein, D. (2009). Learning semantic correspondences with less supervision. *Proceedings of the 47th Annual Meeting of the Association of Computational Linguistics* (pp. 91–99).

Liang, P., Jordan, M. I., & Klein, D. (2013). Learning dependency-based compositional semantics. *Computational Linguistics*, *39*, 389–446.

Mohan, S., Mininger, A., Kirk, J., & Laird, J. (2012). Acquiring grounded representation of words with situtated interactive instruction. *Advances in Cognitive Systems*, *2*.

Roy, D. (2002). Learning visually grounded words and syntax for a scene description task. *Computer Speech & Language*.

Scheutz, M., Eberhard, K., & Andronache, V. (2004). A real-time robotic model of human reference resolution using visual constraints. *Connection Science*, *16*, 145–167.

Tellex, S., Kollar, T., Dickerson, S., & Walter, M. (2011). Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. *Proceedings of the Association for Advancement of Artificial Intelligence*.