A Robotic Model of Human Reference Resolution

Matthias Scheutz(*) and **Virgil Andronache**(*) and **Kathleen Eberhard**(**)
(*) Department of Computer Science and Engineering
(**) Department of Psychology
University of Notre Dame, Notre Dame, IN 46556, USA
mscheutz,vandrona,keberhar@nd.edu

Abstract

Evidence from psychology suggests that humans process definite descriptions that refer to objects present in a visual scene incrementally upon hearing them, rather than constructing explicit parse trees after the whole sentence was said, which are then used to determine the referents. In this paper, we describe a real-time distributed robotic architectures for human reference resolution that demonstrates various interactions of auditory, visual, and semantic processing components hypothesized to underlie human processes.

Introduction

Processing natural language sentences in typical AI algorithms involves a syntactic analysis of the sentence, followed by the construction of one or more parse trees, for each of which the assignment of meaning can subsequently be attempted. This approach, however, can lead to a large number of possible unnecessary parse trees, especially, for simple instructions like "Put the red block on the blue block on the green block on the yellow block..." (e.g., the red block could be on the blue, and should be put green block on top of a yellow block), where referents can be incrementally restricted based on the context in which the sentences is encountered. Evidence from psychology suggests that humans, rather than constructing explicit parse trees, use an incremental method for determining the referents of such expressions based on context that is available to them.

In this paper, we use experimental results from studies with humans as a basis for the development of a robotic architecture that verifies the method of reference resolutions hypothesized for humans. Moreover, we demonstrate a parallel implementation of the proposed architecture, which shows the real-time interactions of auditory, visual, and semantic processing components in the architecture framework APOC (Scheutz forthcoming).

A Model of Human Reference Resolution

The domain under consideration is a simple *Blocks World Domain*, in which blocks of different color can be stacked (e.g., see the left part of Figure 1). Blocks can exhibit one of the following relationships: they can be *on*, *under*, or *next-to*

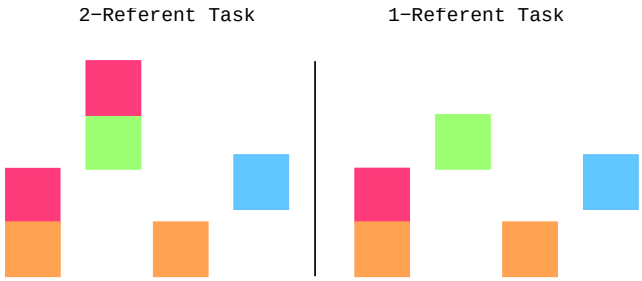

Figure 1: The overall setup of the experiment for the resolution of referents in instructions like "Put the red block on the orange block on the blue block."

other blocks, or they can be *single* (i.e., without any immediately adjacent block). In this domain, it is easily possible to arrange situations, where referential expressions are ambiguous. For example, "the red block" fails to pick out a referent if there are two or more red blocks in a scene. However, if "the red block" is followed by "on the blue block", and if there is only one red block on a blue block, the unique reference can be established.

In our model we consider sentences like the following: "Put the red block on the orange block on the blue block". Syntactically, this sentence has two parse trees, which can be expressed logically as follows: PUT(REDBLOCK,ON(ORANGEBLOCK,BLUEBLOCK)) or PUT(ON(REDBLOCK,ORANGEBLOCK),BLUEBLOCK). Suppose we are given the environment on the left in Figure 1. Starting with the first parse tree, the resolution of the reference of ON(ORANGEBLOCK,BLUEBLOCK)) fails as there is no orange block on a blue block. Consequently, in a typical computational approach, this parse tree is discarded and the tree will be considered (if it exists). In the above case, the reference to ON(REDBLOCK,ORANGEBLOCK) succeeds.

Interestingly, this method based on all possible parse trees is not the way humans resolve reference when the referents are visually co-present. Rather, humans seem to resolve reference incrementally–without building explicit parse trees and using backtracking to build an alternative tree if the previous one fails: upon hearing "red" in the above sentence and scenario, they build a set of possible

referents $PR(red)$, which at this point contains two red blocks. Since the set is not a singleton, the subsequent words are taken to specify constraints on PR and thus to belong to the same referential phrase. Hence, upon hearing "on" a constraint will be instantiated that limits the referents in $PR(red)$ based on the set $PR(orange)$, which will be formed upon hearing "orange". Since the newly formed set $PR(on(red, orange))$ is a singleton, the referent, which is the argument for "put" has been established to be the single element in $PR(on(red, orange))$. Hence, subsequent words are not taken to be part of the definite description anymore.

Now consider a situation depicted on the right in Figure 1, where the referent of "the red block" is uniquely specified. Taking again the above sentence "Put the red block on the orange block on the blue block", a classical parser would produce the unique parse PUT(ON(REDBLOCK,ORANGEBLOCK),BLUE). Hence, there is no ambiguity as to what the sentences seems to ask for. Yet, for human subjects, this sentences causes problems, because according to the above described algorithm reference to a unique object is established upon hearing "red", at which point "on" is taken to belong to "put". Consequently, the second argument of "put", i.e., the location at which to place the red block, is taken to be the referent of "orange". Although this is a set with two possible referents, the precondition of "put"–that no block can be on top of a target of the *put action*–eliminates the orange block underneath the red block, and thus reduces the set of possible referents for "orange" to a singleton (namely, the orange block on the right). Hence, at this point, all the information is present that allows subjects to carry out the put action. Yet, the sentence has not finished, but rather the additional, but seemingly superfluous phrase "on the blue block" is heard. At this point, human subjects attempt to make sense of that phrase. While some subjects interpret it as another put action (after having put the red block from the orange block on the left onto the orange block on the right), others simply ignore it.

Model and Architecture

Several psychological experiments have established the above effect (e.g., (Tanenhaus et al. 1995)) and suggest the above algorithm described for the human behavior. In particular, the assumption is that resolving reference is an incremental process that does not depend on prior established parse trees, which, if a conflict is encountered, are discarded to build alternative trees (if possible). The architecture that implements the above algorithm is depicted in Figure 2.

Each rectangle indicates a component type, whose function is indicated by its label. While some types only have one instance in the virtual machine (e.g., speech recognition), others can have multiple ones (e.g., blob tracking). For space reasons, we can only briefly sketch the functionality of the architecture.

Lexical items are processed in real-time using *Sphinx 2* (see HTTP://WWW.SPEECH.CS.CMU.EDU/SPHINX/). The recognized words are passed to the word analysis node, which ensures that only one word at a time is processed (this node effectively "wraps" the Sphinx system within APOC).

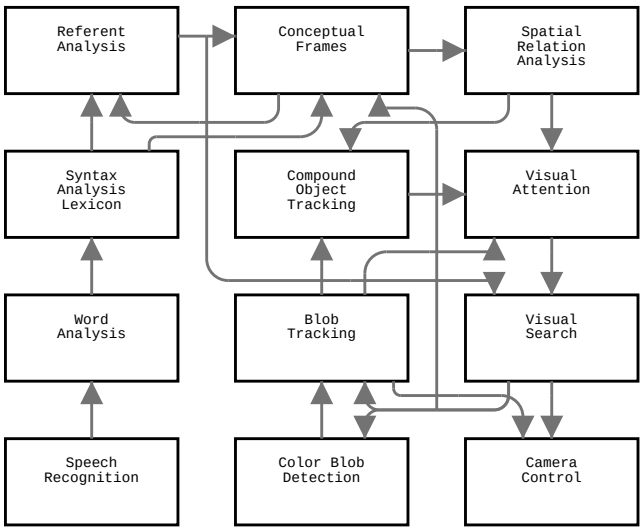

Figure 2: A high-level view of the components employed in the architecture used in the reference resolution task.

The syntactic analysis looks up the word in a lexicon and either initiates a referential analysis (if a referential expression is expected), or instantiates the associated conceptual frame, if it exists (e.g., as in the case of "put"), which will pass control to the referential analysis to obtain its arguments. The referential analysis influences the visual search, which will look for items that match the meaning of the expression built so far. This process involves the instantiation of trackers that lock onto recognized objects in the visual scene, from which more complex trackers can be built based on relational, spatial properties obtained from spatial analysis (e.g., "on"). The visual attention node focuses on one of possibly many trackers (in the cases considered here this is always the one closest to the last focus of attention). Complex trackers, in this architecture, are the "meaning" of relational terms that denote relationship among primitive objects (such as color blocks) as represented by instantiated trackers. When a referent has been uniquely determined, the referential analysis passes control back to the instantiated conceptual frame, which then either expects more arguments, or, if complete, triggers an action (not depicted here). During the whole process, all components are concurrently active, and the camera is moving, if necessary, in such a way as to keep all trackers within the camera's view. While all components could run on separate computers, in the experimental tests the whole architecture was run autonomously on a ActivMedia Pioneer Peoplebot.

References

Scheutz, M. (forthcoming) "APOC - An Architecture Framework for Complex Agents." In Darryl Davis. *Visions of Mind*. Idea Group Inc.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. 1995 Integration of visual and linguistic information in spoken language comprehension *Science*, 268, 1632-1634.