

# Utilizing Visual Attention for Cross-Modal Coreference Interpretation

Donna Byron, Thomas Mampilly, Vinay Sharma and Tianfang Xu

Department of Computer Science and Engineering

The Ohio State University

Columbus, Ohio, 43210, USA

{dbyron, mampilly, sharmav, xut}@cse.ohio-state.edu

## Abstract

Understanding all of the ways context modulates linguistic forms is a challenging endeavor. An important goal in computational linguistics is defining the relationship between context and referring behavior. One contextual effect on referring behavior that is observed across multiple experimental disciplines and linguistic genres is the relationship between entity salience and ambiguous referring expressions. Speakers use underspecified noun phrases, especially pronouns such as “this” and “he” but also common nouns such as “the button”, freely in discourse, relying on the addressee’s ability to understand which button or person is being referred to. This preference for certain entities, given a prior context, will be called *salience* in this paper. Salience corresponds to a prediction or expectation that a certain entity will be the topic of the upcoming discourse. Estimating the relative salience of each entity in the universe of discourse is an important task in computational models of referring behavior - both in production of felicitous noun phrases and also in interpreting connected discourse. The long-term objective of our research program is to create robust, accurate algorithms for reference interpretation in automated agents. This task is impossible without a firm understanding of

contextual effects on referring behavior.

It has been well-established in the computational linguistics literature that the discourse history can be interrogated to determine the salience of entities in a sentence one is trying to interpret. However, recent technology improvements create opportunities for human-computer conversations in which contextual factors in addition to the discourse history are in play at the same time, each impacting entity salience in different ways. The goal of our project is to create conversational software agents that can carry on a *situated* conversation with a human partner. For the purposes of this paper, situated language will be defined as language having these properties:

**Immersion** The conversation takes place within a 3D setting that is perceptually available to the conversational partners. The partners can speak to each other face to face within the setting.

**Mobility** Both conversational partners are at liberty to move about in the world, independently of each other, to gather information or change their perceptual perspective on the world.

These characteristics distinguish situated language from the bulk of interaction paradigms that have informed the development of reference processing algorithms, many of which are developed for

text or in experimental settings where speakers are not face-to-face. This experimental format was used explicitly to limit the degree to which conversational partners could exploit extra-linguistic contextual clues, such as gesture and gaze. Using new data we have generated in our lab, in this paper we examine situated language between 2 human partners. This allows us to investigate the interplay between the discourse context and the conversational setting on the interpretation of referring expressions in a visually-rich domain.

The primary focus of the present work is to develop a model of visual attention that can be used to interpret *exophors*, reference to items in the discourse setting. Similar to the way that an anaphor constitutes a repeated mention of an item introduced into the context by the linguistic history, an exophor is a repeated mention of an item already introduced into the context by the physical world, in other words, a cross-modal coreference. Our hypothesis is that the world that is visually perceptible to the conversational partners will be likely to shape the content of their discussion, especially when they are performing a task on objects in that world. Moreover, a likely source of denotations for exophors are items that the speaker's attention is directed toward simultaneously with the utterance she is producing. Given these two factors influencing the dialog, our aim is to test a method of tracking one speaker's view of the world over the course of a dialog, and use that information as input in a reference resolution algorithm to interpret the referring expressions she produces. Our eventual goal is to construct a model that fuses attentional information provided in the visual channel with that provided by the discourse history. In the present work, we perform pencil-and-paper analysis and offline simulation of our model, as a first step in developing the algorithms that will eventually be imple-

mented.

In our current design of system, we used three parameters - Recency, Uniqueness and Persistence which are obtained from the visual information of sampled frames - to calculate the salience value of each object, and chose the one with the highest salience value to be the actual referent. Experiment showed that with only the visual information, our system correctly predicts 31.3% of the referents produced by speakers while they were collaborating on a task in the virtual world. With the aid of little semantic information, which is gained by a simple string match process (e.g., "the red chair" can only refer to a chair), the accuracy becomes 52.2%, The performance is comparable with discourse-based salience calculated on the same data.