Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computer Speech and Language ■■ (2016) ■■–■■

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A simple generative model of incremental reference resolution for situated dialogue ☆

Casey Kennington [1],*, David Schlangen

Universitätsstraße 25, 33615 Bielefeld, Germany

Received 25 September 2015; accepted 22 April 2016
Available online

Abstract

Referring to visually perceivable objects is a very common occurrence in everyday language use. In order to produce expressions that refer, the speaker needs to be able to pick out visual properties that the referred object has and determine the words that name those properties, such that the expression can direct a listener's attention to the intended object. The speaker can aid the listener by looking in the direction of the object and by providing a pointing gesture to indicate it. In order to resolve the reference, the listener has a difficult job to do: simultaneously use all of the linguistic and non-linguistic information; the words of the referring expression that denote properties of the object, such as its colour or shape, need to already be known, and the non-linguistic gaze direction and pointing gesture of the speaker need to be incorporated. Crucially, the listener does not wait until the end of the referring expression before she begins to resolve it; rather, she is interpreting it as it unfolds. A model that resolves referring expressions as the listener must be able to do all of these things. In this paper, we present such a generative model of reference resolution. We explain our model and show empirically through a series of experiments that the model can work *incrementally* (i.e., word for word) as referring expressions unfold, can incorporate multimodal information such as gaze and pointing gestures in two ways, can learn a grounded meaning of words in the referring expression, can incorporate contextual (i.e., saliency) information, and is robust to noisy input such as automatic speech recognition transcriptions, as well as uncertainty in the representation of the candidate objects.
Published by Elsevier Ltd.

1. Introduction

From eating to working, people interact with visually perceivable objects almost continually. These objects have colour, shape, and other visible properties that distinguish them from each other. Furthermore, people don't just look at and manipulate these objects, they also *talk about them*; i.e., **refer** to them by uttering descriptive noun phrases (which are very common in day-to-day speech (Poesio and Vieira, 1997). In this paper, we focus on referring expressions that are made to direct an interlocutor's attention to a visually present object. Such a setting with visually present objects presupposes that a speaker (who utters the referring expression) and the listener are co-located in a situated dialogue setting. Such a setting of co-located, face-to-face dialogue is a basic and fundamental setting of language use (Fillmore, 1975, p. 152), providing a foundational setting for the task. However, there are also complications: when two people are in a situated dialogue setting and they refer to objects around them, they can make use of more

☆ This paper follows from and extends our previous work in Kennington et al. (2013, 2014, 2015a, 2015b).

* Corresponding author at: 1910 University Dr, Boise, ID 83725, USA.
E-mail address: casey.kennington@cs.boisestate.edu (C. Kennington).

[1] The work presented here was done while the first author was affiliated with Bielefeld University.

http://dx.doi.org/10.1016/j.csl.2016.04.002
0885-2308/Published by Elsevier Ltd.

than just speech to refer to objects; gaze and gestures are also communicative. Another very crucial aspect of referring to objects is the fact that the listener does not wait until the end of the referring expression to being resolving the referred object (Spivey et al., 2002; Tanenhaus, 1995); the listener comprehends as much as possible, as early as possible as the utterance unfolds.

A computational model that resolves referring expressions in a situated dialogue setting would need to take all of these constraints into account: it would need to *ground* language with object properties (i.e., learn, on its own, that certain words are uttered when certain properties are visually present; e.g., the word "red" is used to identify red objects), it would need to process the referring expressions *incrementally* (i.e., word for word), handle uncertainty about the environment, and it would need to be able to make use of information from gaze and pointing gestures. In this paper, we present such a model and show empirically through several experiments that an implemented component based on this model does in fact work incrementally, incorporates gaze and pointing gestures, and can work despite noise from speech and object representations. We also show that the model can resolve a specific type of pronoun (i.e., exophoric), that it can make use of contextual information (i.e., salience information before the referring expression even begins), and that the model has been tested and works robustly in 3 different natural languages: German, Japanese, and English. We refer to this model as the *simple incremental update model* (SIUM).

In the following section, we describe the task of reference resolution. That is followed by a description of our generative model, with some toy examples. We then identify some open questions about the model which are addressed in the experiments below. Following the experiments, we give related work, provide an overall discussion of the model, and conclude.

2. General formalisation of the reference resolution task

To show a more concrete example of the kind of reference resolution (RR) we are interested in, suppose that two friends find themselves among some interesting-looking objects, as depicted in Fig. 1.

The speaker (indicated by *S* in the figure) finds one object particularly interesting (indicated by *I* in the figure) and wants to talk about it with his friend, the listener (indicated in the figure by *L*). In order to begin talking about the object, *S* must first draw *L*'s attention to that object with some kind of referential utterance (indicated by *U* in the figure). The overall progression of what must happen in order for *S* to draw *L*'s attention to *I* is outlined in the following:

1. *S* perceives object *I*
2. *S* forms the intention of talking about *I* with *L*
3. *S* initially indicates *I* to *L* in one of the following ways (via an indication event *U*):
   (a) a descriptive phrase (e.g., *the red cross*)
   (b) a demonstrative phrase (e.g., while pointing, *that*)
   (c) a combination of descriptive and demonstrative phrases (e.g., while pointing, *that red cross*)

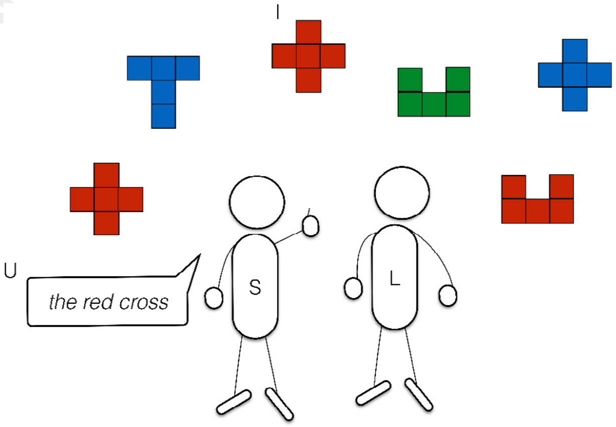

Fig. 1.

Please cite this article in press as: Casey Kennington, David Schlangen, A simple generative model of incremental reference resolution for situated dialogue , Computer Speech & Language (2016), doi: 10.1016/j.csl.2016.04.002