Playing a Real-world Reference Game using the Words-as-Classifiers Model of Reference Resolution

Casey Kennington
CITEC, Bielefeld University
ckennington@cit-ec.uni-bielefeld.de

Soledad Lopez Gambino
CITEC, Bielefeld University
m.lopez_gambino@uni-bielefeld.de

David Schlangen
CITEC, Bielefeld University
david.schlangen@uni-bielefeld.de

Abstract

When referring to visually-present objects, an elementary site of language use, sometimes there isn't enough information to resolve the speaker's intended object. When this happens, more information needs to be elicited from the speaker. In this demo, we will show a simple system that uses the word-as-classifiers model to resolve referring expressions to objects, as well as a simple interaction manager that determines if there is enough information to fully resolve the reference–if not, more information is elicited from the speaker. The modules are implemented and distributed with InproTK.

1 Introduction

Reference to visually-present objects is a foundational language game. Among children's earliest communicative attempts are acts indicating objects for other people; for example, pointing to or displaying an object (Wittek and Tomasello, 2005) where the words of those references are *grounded* in the features of the objects being referred (Harnad, 1990). This setting of language use is situated dialogue where interlocutors can perceive each other, the objects in their shared space, and they can perceive each other's unfolding referring expressions (REs), often resolving the referred object before the RE is complete.

In this demo, we present a system that plays a similar language game: using the words-as-classifiers model of reference resolution (WAC$_{rr}$; explained below), we have a system that can resolve referring expressions (REs) incrementally to real-world objects with an additional component: an *interaction manager* (IM), that determines if more information should be elicited from the speaker.

In the following section we will describe the WAC$_{rr}$ model and how it fits into the system. That will be followed by a description of the interaction manager and the system implementation.

2 The Words-as-Classifiers Model of Reference Resolution

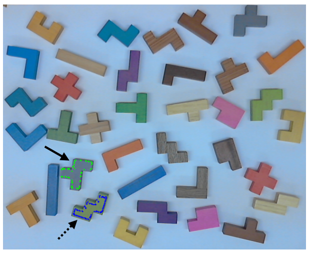

Figure 1: Example episode for where a referred *target* object is outlined in green, a *landmark* object (used to aid reference to the target) in blue; arrows added for presentation. An example RE for this would be: *the gray object on the bottom left above the green w*.

The basis of WAC$_{rr}$ is a model of *word meaning* as a function from visual features of an object to a judgment of how well that object "fits" a particular word.[1] The model can learn word meanings for picking out properties of single objects REs; e.g., *green* in *the green book* (Kennington et al., 2015) and picking out relations between two objects; e.g., *next to* (Kennington and Schlangen, 2015). These word meanings are learned from instances of language use.

These are then applied in the context of an actual reference. This application gives the desired result of a probability distribution over candidate objects, where the probability expresses the strength of belief in the object falling in the *extension* of the expression. We model two different types of composition, of what we call *simple ref-*

[1] This idea follows in spirit from Larsson (2013)

erences and *relational references*. These applications are compositional in the sense that the meanings of the more complex constructions are a function of those of their parts.

The meanings are represented as logistic regression classifiers. We train these classifiers using a corpus of REs coupled with representations of the scenes in which they were used (example in Figure 1) and an annotation of the referent of that RE. Meanings of relational words are trained in a similar fashion, except that they are presented a vector of features of a *pair* of objects, such as their euclidean distance.

During application, to get a distribution from a single word, we apply the word classifier to all candidate objects and normalise. To compose the evidence from individual words into a prediction for a 'simple' RE, we average the contributions of its constituent words. Relational REs are composed by combining two 'simple' REs via a learned classifier for a relation word. More details can be found in the two papers cited above in this section; they further show that the model is robust in reference resolution tasks despite noisy representations of scenes and speech (i.e., ASR).

3 The Interaction Manager

In a dialogue setting, making use of the distribution over objects requires an additional interaction manager which addresses certain cases in which the continuity of the game might be in jeopardy. This could happen if the user has not yet referred to any object (for example, when taking a long time to plan the RE) or if the speaker has already referred to an object but the information provided is not enough for the system to make a decision. Specifically, for this demo, the IM decides whether to select an object (i.e., the argmax of the distribution from WAC$_{rr}$) or if more information is needed.

4 Implementation

Figure 2 shows a schematic the overall system. The ASR (here, Kaldi[2]), WAC$_{rr}$ and the IM have been implemented as modules in INPROTK (Baumann and Schlangen, 2012).[3] For the logistic regression classifiers in WAC$_{rr}$, we use the Apache Mahout Java library trained on a corpus of REs to objects in a scene.[4] We also have a module that can process a video feed of pentomino objects from a standard webcam (example in Figure 1) in real-time and provide the low-level features (e.g., RGB/HSV values, x,y coordinates, number of edges, etc.) of the scene to the WAC$_{rr}$ module. The IM operates by reacting to lack of speech input from the ASR module. After a certain amount of time has elapsed and no voice activity has been detected, a timeout signal prompts the user to speak. Or, if the system has received an RE from the user but the information it contains is not enough to resolve the reference, after a certain amount of silence a simple clarification request is produced as a means of prompting the user to add further information.

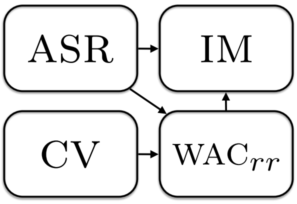

Figure 2: Scheme of the system: ASR and CV modules inform the WAC$_{rr}$ module, which produces a distribution over objects; the IM module determines selection or elicitation.

Acknowledgements This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

Timo Baumann and David Schlangen. 2012. The InproTK 2012 Release. In *Proceedings of NAACL-HLT*.

Stevan Harnad. 1990. The Symbol Grounding Problem. *Physica D*, 42:335–346.

Casey Kennington and David Schlangen. 2015. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of ACL*, Beijing, China. Association for Computational Linguistics.

Casey Kennington, Livia Dia, and David Schlangen. 2015. A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution. In *Proceedings of IWCS*. Association for Computational Linguistics.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*.

Angelika Wittek and Michael Tomasello. 2005. Young children's sensitivity to listener knowledge and perceptual context in choosing referring expressions. *Applied Psycholinguistics*, 26(04):541–558.

[2] http://kaldi.sourceforge.net

[3] https://bitbucket.org/inpro/inprotk

[4] http://mahout.apache.org/