Incrementally Resolving References to Visually-Present Objects in a Sitauted Dialogue Setting

Casey Kennington (Bielefeld)

Objects abound and referring to visually-present objects is a very common occurrence in everyday language use. In order to produce such a referring expression, a speaker needs to be able to pick out visual features that the intended object has and determine the words that name those features such that the expression can direct a listener's attention to the referred object. The speaker can aid the listener's ability to resolve the reference by looking in the direction of the object and by providing a pointing gesture to indicate it. In order to resolve the reference, a 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 features of the object such as its colour or shape need to already be part of the listener's vocabulary and the non-linguistic gaze direction and pointing gesture of the speaker, if useful, 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 does needs to be able to do all of these things.

In my talk, I will describe a generative model and a discriminative model of reference resolution that I have been developing, each of which process the resolution of referring expressions incrementally (i.e, word for word), ground language with aspects of visual objects, and can incorporate gaze and pointing information. The generative model has a way of handling certain types of pronouns and can also take contextual (i.e., saliency) information into account. The discriminative model uses rawer features than the generative and works robustly with noisy data (from speech recognition as well as from the representation of the objects). I will describe and compare the two models, show empirically through several experiments the strengths and weaknesses each model has, and under what circumstances one might be preferred over the other. Both models are quite simple in terms of complexity and straight forward to implement.