Modeling reference games with objects of unknown categories: a Bayesian decoder for zero-shot language generation

Sina Zarrieß, Bielefeld University
sina.zarriess@uni-bielefeld.de

Recent computational work in Language & Vision has extended neural language generation models with pragmatic reasoning components implemented along the lines of "rational speech acts" (RSA) [1,2,3]. These models use a Bayesian approach to decoding a neural (literal) generation component trained, e.g., on pairs of images and verbal descriptions, to produce pragmatically appropriate image captions. Another recent strand of work in Language & Vision investigates zero-shot learning [4]. Here, the task is to correctly label objects of novel categories that the model did not observe during training. In this work, we combine these lines of research and frame zero-shot learning as a challenge for pragmatic modeling. We explore zero-shot reference games as a new setting for generation models: in this game, a speaker needs to refer a novel-category object in an image, such that an addressee (who may or may not know the category) is able to identify the target object. Previous work on language generation has commonly looked at games where a referent of a familiar category needs to be discriminated among distractor referents of identical or similar categories, thereby focusing mostly on modeling attribute selection [5]. Our work on zero-shot reference games adds an additional dimension of uncertainty to this picture, namely a setting where the category of the target itself might not be known to the model and, hence, cannot be named with reasonable accuracy.

We hypothesize that Bayesian reasoning in the style of Rational Speech Acts can extend a neural generation model trained to refer to objects of known categories, towards zero-shot learning. We implement a Bayesian decoder reasoning about categorical uncertainty and show that, solely as a result of pragmatic decoding, our model produces fewer misleading object names when being uncertain about the category. Furthermore, we show that this strategy often improves reference resolution accuracies of an automatic listener. More generally, we argue that uncertain knowledge of the world that surrounds us, including novel objects, is not only a machine learning challenge: it is a common aspect of human communication, as speakers rarely have perfect representations of their environment. Thus, when referring to objects of unfamiliar or difficult-to-name categories, even human speakers might produce utterances that avoid naming the object (e.g., the blue thingy) and avoid confusing the listener. We believe that our approach is an encouraging result for scaling models in computational pragmatics to real-world conversation and its complexities.