Harnessing the linguistic signal in predicting within-scale variability in scalar inferences

Judith Degen

Variability in scalar inferences is a key phenomenon to be explained by pragmatic theories. Recent work in experimental pragmatics has focused on explaining inter-scale differences in the probability or strength with which a scalar inference is drawn (van Tiel et al 2016, Doran et al 2012, Benz et al 2018, Gotzner et al 2018, Sun et al 2018, Westera & Boleda 2020, Ronai & Xiang 2021). Explanations of such inter-scale variability typically take the form of correlating features of each scale under consideration with the resulting average scalar inference rate for that scale. The success of this strategy has been mixed, with effect sizes typically small and/or noisy. Moreover, items to be tested experimentally are usually hand-generated by researchers and the number of items tested per scale tends to be small. I argue that this puts us in a precarious position: the seeming regularity in inter-scale variability may be due to frequent re-use of the same set of items across experiments, the small number of items per scale, and the possible lack of representativeness of the use of scalar items real listeners encounter in the real world.

To address these issues, I present large-scale naturally occurring speech corpus data from both the <all, some> and <and, or> scales. For each scale, more than 1'200 items received inference strength judgments in web-based experiments. I show that intra-scale variability in inference strength is substantial for both scales, and is systematically predicted by features of the linguistic and extra-linguistic context. I also show that a neural network -- specifically, an LSTM-based sentence encoder using BERT word embeddings -- can successfully learn to predict a substantial amount (though not all) of the observed intra-scale variability. Attention weight analyses and other model inspection techniques indicate that the model learns to use some of the same features for interpretation that humans do.

I conclude that (i) the focus on inter-scale variability may be premature, given the large amount of intra-scale variability in inference strength; (ii) the surprisingly good performance of the neural models suggests that a lot of information about whether or not the negation of the stronger alternative was intended is contained in the linguistic signal itself; and (iii) this work opens up exciting avenues for future research investigating how much pragmatic information can be extracted from the linguistic signal itself vs from the extra-linguistic context in which that signal is produced, and how that information is integrated. Under this picture, whether "the scale" retains an explanatory role is an interesting empirical question.