Language Understanding & Common Sense Reasoning

Joshua K. Hartshorne, Tobias Gerstenberg, & Joshua B. Tenenbaum

Language is frequently ambiguous, with the same sentence having several possible interpretations (*The children made delicious snacks*). A central challenge for the listener is to determine which of the possible intended meanings the speaker actually meant to convey. One particularly prevalent example is third-person pronouns. In principle, any of the messages in (1) could be encoded in (2):

(1) Al beat Bart at tug-of-war because Al/Bart/Carl/Darrel/Ethelridge/etc. is strong.
(2) Al beat Bart at tug-of-war because he is strong.

However, most people interpret the pronoun in (2) as referring to Al. Intuitively, common sense reasoning is implicated in this inference (cf. Winograd, 1972), though providing an account of common sense reasoning and how it is integrated into language interpretation has proven challenging.

In line with many recent computational models of pragmatics (e.g., Frank & Goodman, 2012; Goodman & Stuhlmueller, 2013) – themselves extensions of earlier theoretical approaches (e.g., Grice, 1989) – we suggest that listeners infer the speaker's intended meaning using Bayesian inference over an intuitive theory of the speaker's behavior:

$$P(\text{message} \mid \text{utterance}) \propto P(\text{utterance} \mid \text{message})\, P(\text{message})$$

As a simplification, since pronouns are rarely used to refer to previously unmentioned entities, we assume that P(utterance = (2) | message) is negligible for all messages in (1) that do not involve Al and Bart. P(message) is a function of (at least) what the speaker believes to be true and wishes to convey. What the speaker believes to be true is a function of what is true of the world and the speaker's experience with it. Note that in de-contextualized sentences like (2), we have no reason to suppose the speaker is more or less interested in conveying any of the meanings in (1), nor do we know what the speaker's range of experience is. Thus, we need only determine which of the messages in (1) is most likely to be true.

We model tug-of-war competitions as follows. We assume player strength is normally distributed and the stronger player wins (these simplifications can be relaxed but simplify calculation). We model *because* as introducing a counterfactual: *A beat B because A is strong* means that were A not strong, A would not have beaten B. We included 16 variants of (2), manipulating the adjective (*strong*, *weak*), the verb (*beat, almost beat*, *lost to, almost lost to*), and the connective (*because*, *although*). *Strong* and *weak* are modeled with prototype semantics. We model *p although q* as meaning *because of q, p was unlikely*. The model's and humans' interpretations of the pronouns correlated well (r=.92, p<.0001).

In order to ensure that our results were not specific to the tug-of-war scenario, we tested an additional 40 pairs of sentences involving explanations (3a) and 40 pairs involving results (3b):

(3) a. Al frightened Bart because he is reckless/timid.
    b. Because Al frightened Bart, he got in trouble/ran away.

In each pair, the sentences differ in terms of the most likely referent. Because we do not have a full model of the world from which to derive the probability that Al or Bart being reckless would cause Al to frighten Bart or the probability that Al frightening Bart would cause Al or Bart to get in trouble (etc.), we asked a separate set of participants to rate those probabilities, which were then used in the language model. The correlations between the model's and humans' pronoun interpretations was high for both sentences like (3a) (r=.88, p<.001) and (3b) (r=.73, p<.001).

We discuss this model and these findings in the context of recent work on the role of discourse structure, syntactic structure, and verb biases in pronoun interpretation (Hartshorne & Snedeker, 2013; Kehler & Rohde, 2013; Sagi & Rips, in press).