Bounded-rational theory of mind for conversational implicature

Oleg Kiselyov (FNMOC) and Chung-chieh Shan (Rutgers University)

Game-theoretic and relevance-theoretic accounts of pragmatic reasoning often implement Grice's maxim of Manner by penalizing utterances for their syntactic complexity or processing effort. In ongoing work, we propose to explain this penalty as the *risk of misinterpretation* in a bounded-rational theory of mind. The core of this explanation is an executable representation of probabilistic inference in which probabilistic models and inference algorithms are expressed as interacting programs in the same language. The payoff is that the same machinery for interpreting utterances also predicts how potential utterances would be interpreted, or misinterpreted, by others.

The hearer. We make the standard assumption that the grammar specifies which possible worlds w are compatible with the (literal) meaning of which (declarative) utterances u. We notate this relation by $w \vDash u$ as usual. To interpret a trusted assertion u, the hearer starts with some prior probability distribution $\Pr(W = w)$ over worlds and updates it to the posterior distribution $\Pr(W = w \mid w \vDash u)$ by removing worlds incompatible with u. This interpretation task, like the larger task of making decisions under uncertainty, is thus an inference problem on a probabilistic model.

We express a probabilistic model as a program that invokes primitive operations for random choice and evidence observation. For example, we express a model for interpreting utterances as a program that first chooses a world w randomly then observes that the utterance u is grammatical and that $w \vDash u$. We then express a variety of inference algorithms as programs that interact with the model and its continuations, much as an operating system interacts with a user application and its requests or a debugger interacts with a debuggee and its breakpoints. In case the inference algorithm is approximate, misinterpretation may result because the posterior distribution computed and the decision made depend on chance as well as how the grammar and the prior distribution are represented.

The speaker. As in a signaling game, we model a speaker who uses an utterance to coordinate with a hearer. For example, suppose that the speaker knows who are coming to the party and needs to tell the hearer a headcount so that the hearer prepares an appropriate amount of food—having too much food is bad, but having too little food is much worse. Because we express probabilistic models and inference algorithms in the same language, using the same primitive for random choice, it is straightforward to represent how the speaker's probabilistic model includes the hearer's decision making, in particular inference algorithms the hearer might use.

In the party-food example, the typical model quantifies how a precise headcount report, being intuitively more complex, is easier to misinterpret than an approximate one. Therefore, even a perfectly rational speaker, not just a boundedly rational one, may give an approximate headcount (*a dozen*, *some*) to accommodate a boundedly rational hearer. This result may not be specific to the parsing algorithms that humans happen to use, but a consequence of the fundamental computational complexity of interpretation.