Modeling the pragmatics of explanation

Erin Bennett (erindb@stanford.edu), **Noah D. Goodman** (ngoodman@stanford.edu)
Department of Psychology, Stanford University.

Abstract

Everyday explanations are speech acts. As such, explanations offered and their interpretations likely vary with contextual factors. Drawing from work on language understanding these factors might include: possible alternative explanations, the knowledge states of explainer and explainee, the goals of each, and the topic of conversation. We explore a formal model of explanation choice and interpretation derived from recent progress on Bayesian modeling of counterfactuals (Lucas & Kemp, 2015) and of pragmatic language use (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). Our model realizes a semantics grounded in counterfactuals over causal models (Lewis, 1973), but modulated by communicative pressures and social reasoning (Grice, 1970; Lewis, 1969). Our model is able to predict a variety of intuitions about explanation in simple contexts. For example, suppose that Alice and Bob are playing a game of tug-of-war where the stronger player usually wins. If we learn that Alice won, we would infer that she is probably strong. However if someone chose to tell us, "Alice won because Bob was weak," we have the intuition that Alice is likely weak. (On the other hand, simply knowing that Alice won and that Bob is weak gives very little new information about Alice's strength.)

We first explore how different causal models yield different types of explanation: causal explanations are the natural case from simple causal models, while teleological explanations follow from a causal model that includes a (notional) decision-making agent. We then turn to the predicted effects of pragmatics factors on explanation. We investigate how changes to the common background knowledge, to the epistemic states of each participant, and to the question under discussion affect judgments of how good an explanation is in context. For example, suppose Bob either generally follows in his older sister Alices footsteps or deliberately does the opposite. If an explainer were asked, "Why is Bob playing soccer?" she might reply with one of two explanations: "Because Alice is playing soccer," or "Because Bob usually does whatever Alice does." If the explainer knows that the questioner doesn't know one of these facts, then the explanation with new information becomes more likely. If the explainer knows that the questioner has some goal in mind (e.g. if the questioner is interested in buying Bob sporting equipment, or more generally in predicting Bobs preferences or future actions) then this goal will also guide her choice of explanation.

In a series of behavioral experiments, we test whether the interpretations of our pragmatic explainer model and its preferred explanations under different communicative contexts match people's interpretations of and preferences for different explanations. We manipulate domain structure, knowledge state of the questioner, and topic/goals of the conversation. We use common-sense scenarios similar to the examples above, which, although simple, allow for rich and complex inferences based on different explanations. We conclude with a discussion of how our model and results relate to theories of explanation, emphasizing problems for our model and future prospects.

Keywords: explanations; counterfactuals; pragmatics

Background

Causal Selection

Hesslow (1988) discusses the issue of *causal selection* as a seperate task from *causal attribution*. The latter determines whether a condition can truthfully be considered a contributing cause. The former determines, of possible contributing causes, which is the best *explanation*.

Hesslow claims that an explanation must always explain why an effect occurred in the actual situation but did not occur (or would not have occurred) in some contextually-specified comparison class of situations. He uses this account to unify many previously proposed criteria for causal selection. The relevant comparison class must depend on the interlocuters' knowledge of what alternative situations are possible, and on the topic of their conversation. Under some topics, the utility, moral acceptability, temporal proximity, typicality or stability of the different alternatives might matter to the generation of the comparison class.

Hilton (1996) desribes an informal model of Gricean pragmatics in explanations. He explains between casual selction processes of causal "discounting" and causal "backgrounding" in terms of different Gricean maxims. Causal discounting, where one cause becomes a less good explanation as a result of an alternative cause gaining salience, seems to be an effect of changes in the underlying generative model (*quality*, or truthfulness) or in the question under discussion (*relevance*). Causal backgrounding, where a cause becomes a less good explanation as a result of becoming especially predictable, seems to happen as a result of changes what common knowledge is assumed between the interlocuters (*informativity*).

Reuter, Kirfel, van Riel, and Barlassina (2014) describe factors that seem to guide causal selection that they consider to be independent of pragmatic effects or the underlying causal structure: temporal proximity and morality. They show through a series of experiments that the most recent potential causes A and B is regarded as the best explanation of an effect E when the causal structure is "A and B implies E", and that this effect is overridden by a tendency to hold as responsible any cause that violates a norm (e.g. if A broke a rule, and B did not, then A will be held responsible for the effect jointly caused by both A and B).

Woodward (2011), responding to some arguments that people's notion of "cause" is purely subjective and contextual, argues that some factors that may seem pragmatic and context-dependent are actually systematic and invariably matter to determining whether something is considered a "cause" or just an "enabling condition". The factors that Woodward cites as probably systematic and independent of context are "stability" and "specificity". By "stability", he means the sufficiency of the cause to bring about the effect in other counterfactual scenarios. By "specificity", he means the extent to which minor changes in the manner in which the cause occurred would have brought about corresponding changes in

the way the effect occurred.

Counterfactuals

Lucas and Kemp (2015) ...

Jimnez-Leal (2013) show through a series of experiments that reasoning about cause and reasonging about counterfactuals might be the same process. Previous results had shown that people respond differently to questions about cause than to questions about counterfactuals, but ... show that this is probably due to task demands in the language used to elicit causal versus counterfactual judgments.

Rational Speech Act Models

RSA models (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) and related models (Franke, 2011; Russell, 2012).

References

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, *4*, 1–82.

Goodman, N., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*.

Grice, H. P. (1970). *Logic and conversation*.

Hesslow, G. (1988). *The problem of causal selection.* Harvester Press, Brighton.

Hilton, D. J. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, *2*(4), 273–308.

Jimnez-Leal, W. (2013). Causal selection and counterfactual reasoning. *Revista Colombiana de Psicologia*, *22*, 179–197.

Lewis, D. (1969). Convention: A philosophical study.

Lewis, D. (1973). Causation. *The journal of philosophy*, 556–567.

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*.

Reuter, K., Kirfel, L., van Riel, R., & Barlassina, L. (2014). The good, the bad, and the timely: how temporal order and moral judgment influence causal selection. *Fronteirs in Psychology*.

Russell, B. (2012). *Probabilistic reasoning and the computation of scalar implicatures*. Unpublished doctoral dissertation, Brown University.

Woodward, J. (2011). *Causes, conditions, and the pragmatics of causal explanation.*