Testing gradient measures of relevance in discourse

Alex Warstadt (NYU) and Omar Agha (NYU)

Main Claims We experimentally test two information-theoretic measures of relevance: **Entropy reduction** (van Rooy, 2004; Rothe et al. 2018), and **KL-divergence** (Nelson et al., 2010; Hawkins et al., 2015). Results show that KL-divergence fits introspective relevance judgments better than entropy reduction. However, neither measure is adequate on its own.

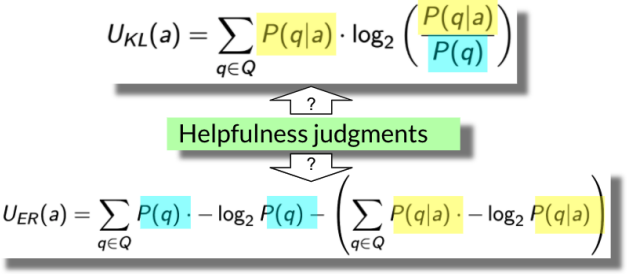
Gradient Relevance In one popular view, a response to a question counts as relevant only if it is a **partial answer** (Roberts, 2012; Groenendijk and Stokhof, 1984). However, examples like (1) show that some relevant responses are not partial answers, and that intuitively responses are *ordered* by relevance. Thus, a more inclusive and gradient definition is needed to capture these facts.

(1) Q: Is it going to rain?
It rained last week. < It's cloudy. < The forecast predicted rain.

The basic idea behind information-theoretic measures of relevance is that relevant information shifts your probability distribution over the answers to the QUD, but different measures assign high utility to different kinds of updates. We make these assumptions: A QUD is a partition over the context set. Every discourse participant's beliefs are modeled by a probability distribution over all alternatives in the QUD. Belief update is monotonic (eliminative), and we don't consider false beliefs.

Entropy reduction versus KL-divergence Entropy reduction and KL-divergence are two candidates for a gradient theory of relevance. The entropy of a probability distribution can be interpreted as a measure of uncertainty about the true answer. Therefore, upon learning some new information a, the **entropy reduction** from the prior to the posterior measures the degree to which a decreases uncertainty. Van Rooy (2004) demonstrates that, given certain assumptions, the utility of an answer is its entropy reduction. $U_{ER}(a)$ (definition below) is the utility of the answer a, under the entropy reduction theory. **KL-divergence** can be interpreted as measuring the information gained by updating from a prior distribution to a posterior. KL-divergence has been used as a measure of relevance in Bayesian models of information acquisition (Nelson et al., 2010) and the utility of questions (Hawkins et al., 2015). $U_{KL}(a)$ (definition below) is the utility of the answer a, under the KL-divergence theory. The goal of our study is to compare how these two measures of utility correlate with introspective judgments of helpfulness.

Task 1: Priors *How likely is the yes answer (given a linguistic context)?*
Task 2: Posteriors *How likely is the yes answer (given a context and a response)?*
Task 3: Helpfulness *How helpful is the response (given a context and question)?*



Methodology Using Amazon Mechanical Turk, we collected three kinds of judgments from crowdworkers: priors, posteriors and helpfulness judgments, described in the figure above. We collected each judgment from three different participants, and each rating was given on a slider from 0 to 1. We measured the correlation between both $U_{KL}(a)$ and $U_{ER}(a)$ and crowdworkers' helpfulness judgments for the answer a given the question Q. $U_{KL}(a)$ and $U_{ER}(a)$ were computed from the mean priors and posteriors, as in the formulas above.

We constructed 150 **dialogues**, where each dialogue consists of a polar question, a context, and an answer. There were 10 distinct polar questions. The **contexts** were constructed in sets of 3: negative bias, neutral bias, and positive bias. • A *negative bias* context favors a low prior

probability for the *yes* answer. • A *positive bias* context favors a high prior. • A *neutral bias* context does not favor either a high or low prior.

Each answer either favors the *yes* alternative, or favors neither alternative. The **answers** were of 5 possible types depending on the degree to which they favor *yes*: • *Exhaustive answer:* The answer rules out the *no* answer. • *Non-answer:* The answer does not directly influence credence in the possible answers. • *High certainty:* The answer makes *yes* much more likely. • *Low certainty:* The answer makes *yes* slightly more likely. • *Reductive answer:* The negation of the answer would rule out the *yes* answer (Agha & Warstadt 2020).

(2) Example prompt for posterior judgment: *positive bias* context, *high certainty* answer.

Background: It's afternoon at the office and you're ready for a snack. There was a birthday party for Lily earlier in the week, and the cake was so big that the party-goers barely made a dent in it.

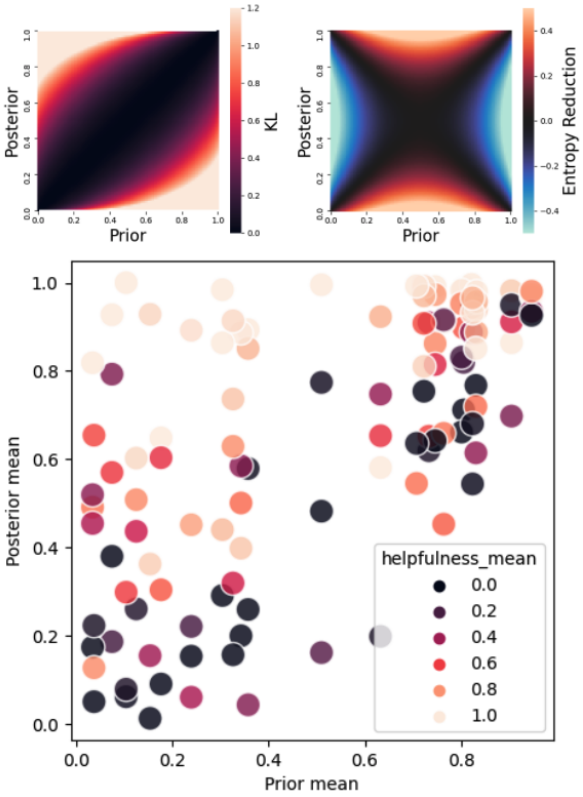
Your coworker Meaghan turns to you and says: The cake box is still on the counter in the kitchen.

How likely do you think it is that there is any cake left?

IMPOSSIBLE 50.000% CERTAIN

In our analysis, we do not treat response types or context types as categorical conditions. We use these types to elicit a diverse set of prior and posterior judgments. This allows us to test the predictions of U_{ER} and U_{KL} across a broad range of inputs.

Findings On the right, the two top heatmaps depict the KL divergence and entropy reduction as a function of the prior and the posterior probabilities of the *yes* answer. The scatterplot below shows our results, plotting helpfulness as a function of prior and posterior, with each point representing a single item. **KL-Divergence is better than entropy reduction.** When the prior is low ($<$0.33, helpfulness judgments correlate strongly with KL-divergence (Spearman's ρ=0.64) and entropy reduction does not correlate at all (ρ=0.08). However, the overall correlations are similar: ρ=0.43 for KL-divergence and ρ=0.34 for entropy reduction.



KL-divergence has systematic problems. When the prior is high ($>$0.67), the correlation of helpfulness with KL-divergence (ρ=0.40) is slightly lower than with entropy reduction (ρ=0.59). In this scenario, answers merely confirm strongly held priors. We have two hypotheses for this finding. First, this may be a reflection of the confirmation bias found across many domains of cognition. Second, confirming evidence may have the effect of reducing higher order uncertainty, i.e. uncertainty about whether one's probability distribution over alternatives is correct. We consider these worthwhile hypotheses to test in future work.

References • Agha and Warstadt 2020 (*SuB*) • Hawkins et al. 2015 (*CogSci*) • Nelson et al. 2010 (*Psychological Science*) • Roberts 2012 (*SemPrag*) • Rothe et al. 2018 (*Computational Brain and Behavior*) • van Rooy 2004 (*Journal of Philosophical Logic*)