Language meaning at different levels of abstraction

Erin Bennett

June 20, 2018

1 Introduction

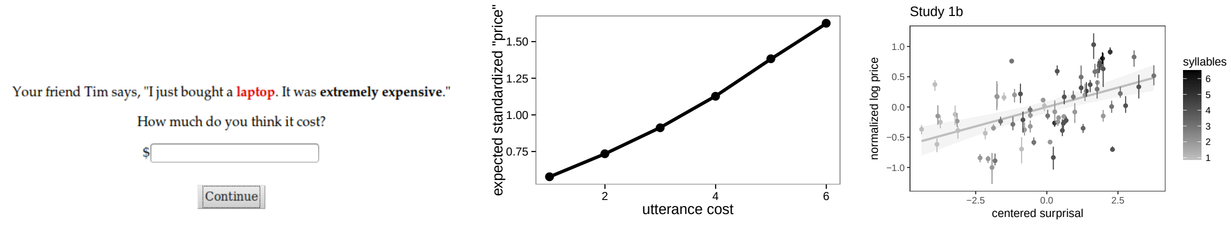
I have two main sections of work that address language meaning. In the first section, I discuss several extensions to a pragmatic model of scalar adjectives. In the second, I discuss sentence vectors and discourse: how to learn sentence representations from discourse relations, and how to describe discourse dynamics in terms of sentence vectors.

2 Adjectives

2.1 Intensifiers

In our paper (?, ?), we extend ? (?)'s model (Appendix A) to cover adjective phrases with intensifying adverbs (e.g. *extremely*, *very*). We assume that each intensified adjective phrase i has its own threshold θ_i and include the bare adjective, the intensified adjective, and *no utterance* as possible alternatives. These alternative utterances vary in their cost c_i, which the pragmatic listener can use to infer the thresholds.

In a series of MTurk studies, we demonstrate that, as predicted by this model, longer and less frequent intensified adjective phrases correspond to higher values (Figure 1). The effect of length holds for even novel intensifiers, inducating that this can be an active, on-line inference.


Your friend Tim says, "I just bought a laptop. It was **extremely expensive**."

How much do you think it cost?

$

Figure 1: Screenshot, model predictions, and results of Intensifiers study 1b. As surprisal increases, participants' scaled responses increase.

2.1.1 Next steps for Intensifiers

This section could be complete as is, or could be extended in the following ways:

- The pragmatic model currently makes a qualitative prediction. But we could specify more details and quantitatively predict responses. This would require both a specific way to calculate utterance cost (which the linear regression models in the paper would basically provide) and the prior distributions over prices (which we also need for the sorites section).

- The model predicts that as intensifiers become more frequent, they will lose their strength. We could try to test this prediction, e.g. by manipulating the frequencies of adverbs within an experiment or by somehow estimating the strength of intensifiers from historical corpora.

2.2 Sorites

Also based on ? (?)'s model of scalar adjectives, we (Dan Lassiter, Noah, and I) explored the "paradox of the heap", or *sorites* paradox. The observation is that the following two premises seem true:

A million grains of sand is a heap of sand. *(Concrete premise)*
A heap of sand with one grain of sand taken away is still a heap. *(Inductive premise)*

However, if these premises actually are both completely true, then we should be able apply the inducitive premise thousands of times and conclude that a single grain of sand is a heap.

We can resolve this paradox with ? (?)'s model if the threshold θ for a scalar adjective is inferred separately in each of its different contexts. We use the adjective *expensive* and explore the following premises for different values of V and ε.

A sweater that costs $\$V$ is expensive. *(Concrete premise)*
A sweater that costs $\$\varepsilon$ less than an expensive sweater is expensive. *(Inductive premise)*

We collected ratings for the concrete and inductive premises for the adjective *expensive* for different objects and dollar amounts and compared these ratings to model predictions. Qualitatively, we show that as the difference amount ε increases, people are less and less willing to endorse the inductive premise.

The meaning of *expensive* for a sweater depends on the distribution over prices for sweaters. We elicited price distributions for five different objects using a binned histogram method (?, ?).

We initially modeled the concrete premise and inductive premises by first using the pragmatic listener's inference of the price d and threshold θ for "The sweater is expensive." For the concrete premise, we then determined whether V was above or below θ. For the inductive premise, we determined whether $d - \varepsilon$ was above θ.

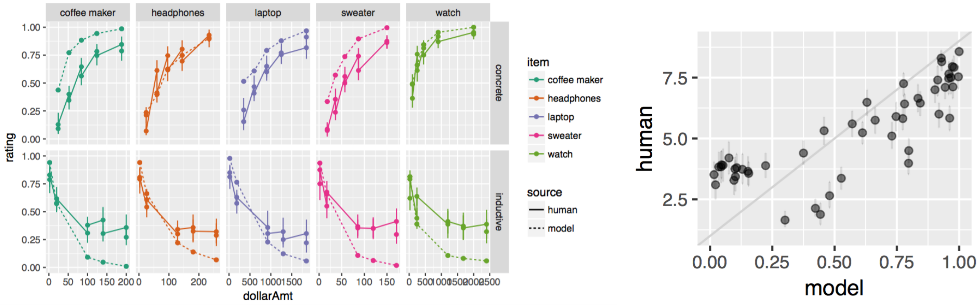


Figure 2: Best model fit for sorites studies.

Using these semantics and a log-normal fit to our binned histogram data, we found a reasonable model fit to human endorsements (Figure 2). However, our number of participants was small, model fit is very sensitive to the prior elicitation method, and the semantics we used was pretty clunky. We attempted to run a BDA with these semantics on all of our pilot studies together and got a very poor fit ($R^2 = 0.12$).

2.2.1 Next steps for Sorites

I'd like to run a replication with the following changes:

- Choose number of participants with a power analysis.
- Clean up concrete prompts to set up for an S_1 model:

  Noah bought a sweater for \$V.
  How much do you agree with the following statement:
  The sweater was expensive.

- Clean up inductive prompts to set up for an S_1 model:

  Noah bought a sweater. Noah's sweater was expensive.
  Mike bought a different sweater for \$$\varepsilon$ less.
  How much do you agree with the following statement:
  Mike's sweater was expensive.

- Use a pragmatic speaker S_1 as an endorsement model for the premises (?, ?, ?).

$$[[\text{The sweater was expensive.}]]^{V,\theta} = \begin{cases} 1, & \text{if } V > \theta \\ 0, & \text{otherwise} \end{cases}$$

$$\begin{aligned} \text{Concrete Endorsement} =& S_1(\text{“The sweater was expensive”} | d = V) \\ \text{Inductive Endorsement} =& S_1(\text{“The sweater was expensive”} | d_{Mike} = (d_{Noah} - \varepsilon)) \\ & \text{where } d_{Noah} \sim L_1(\text{“The sweater was expensive”}) \end{aligned}$$

- Assume lognormal priors, since this was empirically a good fit, and it lets us get quality estimates for the tails of distributiions.
- Use BDA for prior and endorsement data jointly

3 Sentence vectors

3.1 DisSent

In our manuscript (?, ?), we use marked discourse relations (e.g. *but*, *because*, *so*) to learn sentence vectors. This is a self-supervised method, where we extract pairs of sentences using dependency parsing (Figure 3). We then use a model developed for supervised sentence relation pairs (?, ?) to jointly train the individual sentence vectors and a classifier that uses those vectors to predict the relation. On a variety of generalization tasks that depend on sentence meaning, we use sentence vectors as features for a simple logistic regression. DisSent vectors perform well relative to state-of-the-art sentence vectors (Figure 3).

3.1.1 Next steps for DisSent

This section could be complete as is, but it would be cool if we could make better sentence vectors. These sentence vectors are useful, but they obviously don't capture sentence meaning perfectly.

One possible direction to improve embeddings would be to use other kinds of self-supervision (e.g. punctuation).

Another direction is to develop a model of discourse markers that better captures how people use them in natural language. With a model that more accurately captures how people use discourse markers, we might

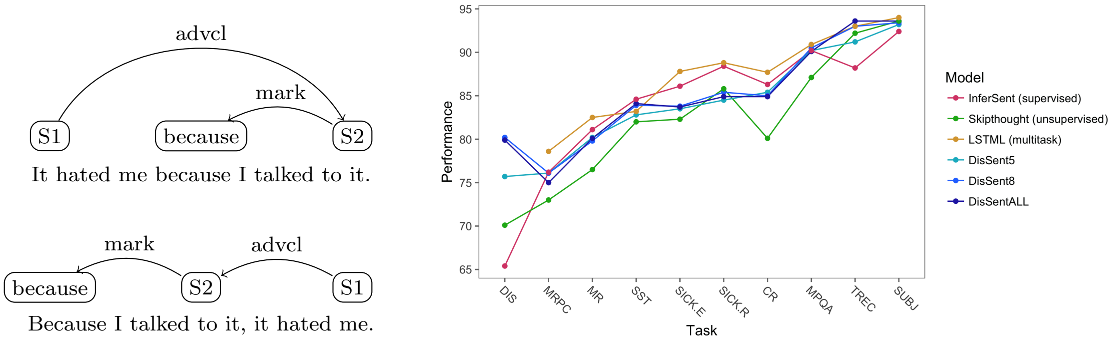


Figure 3: Dependency relations used for extraction and generalization performance of DisSent sentence vectors relative to Skipthought and InferSent.

be able to use them more effectively to learn sentence vectors. We've spent some time brainstorming models where different discourse markers are modeled as transoformations, e.g. a function from the meaning of S1 to the meaning of S2, or from the previous topic of the discourse to a new topic. It's possible that the next section, in which we describe the trajectories of documents within a given sentence embedding space, could inspire more interesting models for training better sentence vectors.

3.2 Describing discourse dynamics

Given a way of computing sentence vectors, how can we describe the trajectories of different kinds of documents through this high-dimensional space?

I've collected some visualizations for Skipthought vectors of Wikipedia sentences. We see greater similarity between nearby sentences than further away sentences in a document, indicating a kind of graded coherence. This coherence exists within paragraph boundaries but across paragraph boundaries we tend to get especially high displacements (Figure 4).

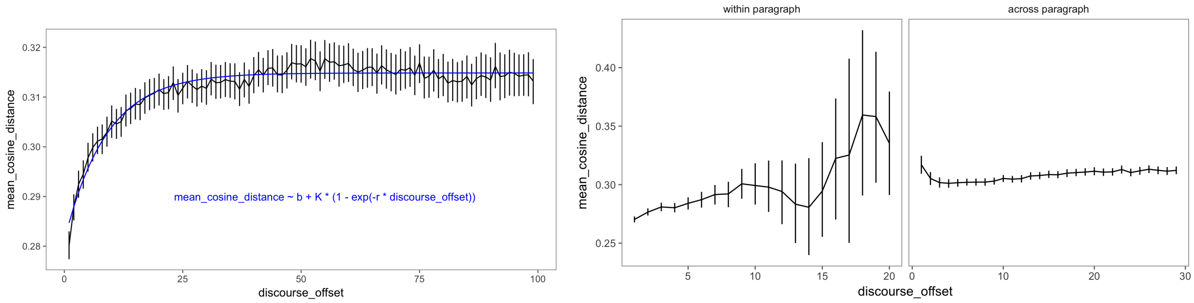


Figure 4: The distance between a random sentence in a document and the sentence N steps away increases exponentially in general. But the distances between sentences separated by only a paragraph boundary are even higher than between two random, distant sentences in a document.

3.2.1 Next steps for Discourse Dynamics

This is the most open-ended project. Some of the questions I'd like to address are:

- What kinds of displacements occur across paragraphs, different kinds of punctuation, different discourse markers, etc.. Can we predict the relation from the displacement? Do some seperators have higher or lower displacement magnitudes?

- What do different genres, modalities, writers, speakers, etc. share in their discourse dynamics? How do they differ?
- Can we differentiate the kinds of sentence vectors that will correspond to sensible meanings vs. unusual, nonsensical, or impossible meanings?
- Do different sentence embedding models lead to different descriptions of discourse dynamics? How so?
- How might the descriptions of the resulting discourse spaces and trajectories inform sentence embedding models?

References

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Degen, J., & Goodman, N. D. (2014). Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the thirty-sixth annual conference of the Cognitive Science Society.*

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Franke, M., Dablander, F., Schöller, A., Bennett, E. D., Degen, J., Tessler, M. H., ... Goodman, N. D. (2016). What does the crowd believe? a hierarchical approach to estimating subjective beliefs from empirical data. In *Proceedings of the 38th annual conference of the cognitive science society.*

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818-829.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and Linguistic Theory (SALT) 23* (pp. 587–610).

Leffel, T., Cremers, A., Gotzner, N., & Romoli, J. (2016). Vagueness and the derivation of structural implicatures. *Unpublished manuscript*.

Bennett, E. D., & Goodman, N. D. (2018). Extremely costly intensifiers are stronger than quite costly ones. *Cognition*, *178*, 147-161.

Nie, A., Bennett, E. D., & Goodman, N. D. (Manuscript). Dissent: Sentence representation learning from explicit discourse relations. *arXiv:1710.04334*.

Tessler, M. H., & Franke, M. (n.d.). Not unreasonable: Carving vague dimensions with contraries and contradictions.

Appendix A Adjective model

The intensifiers and sorites sections are based on ? (?)'s RSA model of scalar adjectives. In this model, the meaning of the adjective depends on a threshold parameter θ. For the adjctive *expensive*, the adjective is true of an object X when the price of X is greater than θ.

$$\begin{aligned} d &= price(X) \\ [[\textit{no utterance}]](d,\theta) &= 1 \\ [[\text{X is expensive}]](d,\theta) &= \begin{cases} 1, & \text{if } d > \theta \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

The price and threshold are jointly inferred by a pragmatic listener according to RSA (?, ?).

$$
\begin{aligned}
L_0(d|u,\theta) &\propto \delta_{[[u]](d,\theta)} \cdot P(d)\\
S(u|d,\theta) &\propto e^{\alpha(\log[L_0(d|u,\theta)]-c)}\\
L_1(d,\theta|u) &\propto S(u|w,\theta) \cdot P(w) \cdot P(\theta)
\end{aligned}
$$

Appendix B Other open questions

Multilingual DisSent We worked on generalizing our results for English to other languages, but got poor results, likely due to the inferier quality of the word embeddings and a need to tune the extraction method to these other languages.

Intensifiers and negation Intensifiers behave in interesting ways with negation (*not very tall*). ? (?) showed that different adjective scales behave differently with *not very* (e.g. *not very late* necessarily means late, whereas *not very fast* doesn't necessarily mean fast). ? (?) extended our intensifiers model to account for double-negated scalar adjectives (e.g. *not unhappy*) that essentially included ambiguity over the negation scope, and this could be a nice starting point to modeling negated intensifiers.

Alternative or lesioned models for sorites It might be nice to have alternative models to compare to.