OPEN MIND

Discoveries in Cognitive Science

an open access journal

Citation:

DOI:
http://dx.doi.org/10.1098/rsif.2013.0969

Supplemental Materials:
http://dx.doi.org/10.1098/rsif.2013.0969

Received:
Accepted:
Published: 26 January 2014

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:
Author Name
Corresponding author email address

The MIT Press

Interpreting Metaphors: Distributional Semantics for Bayesian Pragmatics

1

2

Keywords: metaphor, informativity, Bayesian pragmatics, distributional semantics

ABSTRACT

Humans interpret metaphors, like *Life is a journey* or *My lawyer is a shark*, with relative ease, incorporating contextual knowledge to determine which aspects of the predicate (*journey, shark*) are true of the subject (*life, my lawyer*).

One theory of the process underlying metaphor interpretation is Gricean: a listener reasons about a cooperative and informative speaker (who in turn reasons about the listener) to update their beliefs about the subject and the dimensional of meaning that the speaker is trying to convey.

This reasoning process can be modeled in the *Rational Speech Acts framework* (Frank & Goodman, 2012) as a nested inference, where listeners and speakers are Bayesian agents, who make inferences about the state of the world, and the optimal utterance, respectively. However, previous instantiations of these models have required a hand-specified semantics, restricting the generality of the model and the scope of empirical investigation into the effectiveness of pragmatic reasoning for metaphor interpretation.

We present a method to combine empirically learned word embeddings with a Rational Speech Acts model of metaphor. This allows us to interpret arbitrary predicative metaphors without manually stipulating the denotations of the words they contain. We compare our model to a word embedding model without explicit pragmatic reasoning, on human judgments, and find a significant preference for our model.

INTRODUCTION

Metaphor presents a compelling theoretical challenge for the understanding of meaning in natural language. For instance, on hearing (1) in a context where the subject, Jane, is known to be a consultant, a listener might infer that Jane is not literally a soldier, but rather that she shares certain attributes with soldiers (perhaps determination, endurance, or ruthlessness).

(1) Jane is a soldier

The Gricean view on metaphor takes the meaning conveyed by (1) in a given context to be the result of a process of *pragmatic reasoning*, about a speaker who is trying to communicate truthfully, informatively, and relevantly. That is, the listener attempts to jointly deduce what Jane must be like and what aspect of Jane is plausibly relevant, such that the speaker would have chosen the predicate *soldier* over other alternatives.

Modeling metaphor Obtaining metaphorical interpretations for utterances like (1) is within the scope of a probabilistic formalization of Gricean reasoning, the *Rational Speech Acts* framework (Goodman & Stuhlmüller, 2013). This framework models pragmatic interpretation and production of language via probabilistic models of speakers and listeners, who reason about each other in a nested fashion, with the assumption of cooperativity and a shared semantics.

Kao, Bergen, and Goodman (2014) extend the framework by introducing a new model, L_1^Q, which is able to interpret metaphors. It does so by the mechanism of *projection functions*[1] which dictate the dimension of the world that the speaker cares about communicating. This, in turn, allows for a model of a listener which jointly determines the state of the world (e.g. what Jane is like) and the aspect of the world the speaker cares to communicate (e.g. Jane's determination). Crucially, this listener assumes an informative speaker: one whose choice of utterance maximizes the probability of communicating the world w up to the projection q to the speaker's model of the listener.

This model provides an account of predicative metaphors (those of the form *A is a B* like (1)) and adjectival metaphors (like *fiery temper*). However, in order to generate predictions from the model, it is necessary to manually construct a semantics for the words involved. This makes empirical validation of the model difficult, because of the bias introduced by the hand-chosen semantics.

Our contribution We adapt the model proposed by (Kao et al., 2014) to a distributional semantics, where the denotations of words correspond to points in an abstract, and empirically determined, vector space. Such *word embedding* spaces are a core NLP tool, and have been shown to represent various notions of semantic similarity via the geometry of the vector space.

Adapting the L_1^Q model to allow for denotations of this kind requires us to introduce a new semantics, to generalize the notion of a projection to a vector space, where it amounts to a linear projection, and to develop an approximate inference algorithm to calculate the now continuous posterior distribution of the listener model.

By doing so, we obtain a model capable of interpreting arbitrary predicative and adjectival metaphors, without the need for a hand-specified semantics.

This allows us to assess to efficacy of the Gricean view of metaphor (as formalized in L_1^Q), by evaluating our model on human judgments. In particular, we show that our model significantly outperforms a baseline which uses a distributional semantics without explicit pragmatic reasoning.

OVERVIEW OF METAPHOR

Metaphor exists in many syntactic forms, and generally eludes easy definition. For present purposes, we focus on copular predicates (of the form: *A is a B*, like *Jane is a soldier*) and AN noun phrases (of the form [adjective noun] like *fiery temper*). We refer to the predicated or modified noun (*Jane*, *temper*) as the *target* of the metaphor and the predicate or adjective

[1] This is often referred to in RSA literature as a *Question under Discussion*.

(*soldier*, *fiery*) as the *source* (see (Lakoff & Johnson, 1980) for the more general sense of these terms).

For a given metaphor, only certain properties of the source are described by the target, and which these are depend on the metaphor and the context. For instance, (2) likely conveys that the bread is like a rock with respect to hardness, while (3) likely conveys that the sleeping dog is like a rock with respect to its responsiveness. However, we could also imagine a context in which the dog is very heavy, so that a more natural inference is that it is like a rock with respect to its weight.

(2) The bread is a rock.

(3) The sleeping dog is a rock.

While certain metaphors are conventional - comparing someone to a lion tends to connote bravery - examples like (3) suggest that the interpretation of a metaphor is contextually dependent. The benefit of the Gricean view of metaphor is the ability to explain this dependence on context, in a way which takes into account an underlying semantics (e.g. the conventional meanings of *bread*, *dog* and *rock*).

A BAYESIAN MODEL OF METAPHOR INTERPRETATION

The Rational Speech Acts framework (RSA) provides an elegant and practical way of formalizing Gricean pragmatics as nested Bayesian inference (Frank and Goodman (2012)).[2]

In this framework, listeners and speakers are represented as conditional probability distributions. Speakers are distributions $P(U|W)$ over possible utterances given worlds, and listeners distributions $P(W|U)$ over possible worlds given utterances, where W is the set of possible states, and U is the set of utterances available to a speaker. The most basic version of RSA (see, for example, TODO CITATION) is incapable of interpreting metaphors, due to the strict adherence to truth observed by the speaker. To address this, (Kao et al., 2014) proposes a model L_1^Q, defined in (6).

(4) $L_0(w|u) = \frac{[\![u]\!](w)P_L(w)}{\sum_{w' \in W} [\![u]\!](w')P_L(w')}$

(5) $S_1(u|w,q) \propto \sum_{w'} \delta_{q(w)=q(w')} * L_0(w'|u)$

(6) $L_1^Q(w,q|u) \propto S_1(u|q,w) * P_L(w) * P_{L_Q}(q)$

We can view L_1^Q as a model of AN metaphor interpretation, the focus of our experiment in section , in the following way.[3] Suppose that the goal of a speaker is to choose an adjective u, out of some set U, which communicates some modification of the noun (e.g. *temper*). The goal of the listener is then to infer the desired modification of *temper*, on hearing an adjective, like *fiery*. but note that predicative metaphors can be treated in a similar way. The posterior distribution of the literal listener L_0, which reasons about the semantics, represents the literal meaning of the AN metaphor, while L_1^Q has the capability of deriving a metaphorical interpretation.

[2] For an interactive introduction, we recommend http://gscontras.github.io/ESSLLI-2016/chapters/1-introduction.html.

[3] Note that an account for predicative metaphors, with a more conventional truth-conditional semantics can be provided in a similar fashion.

To make this precise, and derive predictions from L_1^Q, five things must be provided: a set W of states, a set U of utterances, a set Q of projections, a prior P_L representing the listener's uncertainty over W, and a semantics $[\![\cdot]\!]$. Jointly, we say that these determine an *interpretation* of L_1^Q.

One possible interpretation treats points in the state space W as sets of properties $w \subset P$. We refer to this as a *set theoretic* interpretation of L_1^Q, and give a concrete example below:

- $W = \{(\text{intense}), ()\}$
- $P_L = \{\}$
- $U = \{\text{fiery}, \text{warm}\}$
- $[\![\cdot]\!]$
- $Q = \{\lambda(x, y) : x, \lambda(x, y) : y\}$

TODO: add details and predictions

The literal listener L_0 represents a model of an agent that updates their belief about the world strictly in accordance with the semantics $[\![\cdot]\!]$. This semantics represents the conventional association between states w and utterances u which the speaker and listener take as given. P_L represents the listener's prior belief.

Projections Functions $q \in Q$ formalize the notion of picking a particular *aspect* or *dimension* of w. Formally, they are surjective functions out of W. A simple example is as follows: suppose that $W = AxB$, where $A = \{\text{Jane is hardworking}, \text{Jane is lazy}\}$ and $B = \{\text{Jane owns a gun}, \text{Jane doesn't own a gun}\}$, so that each $w \in W$ is a tuple (a, b) for $a \in A, b \in B$. Then the two projections, which we could call $q_{work-ethic}$ and $q_{gun-ownership}$, are defined as $\lambda(x, y) : x$ and $\lambda(x, y) : y$ respectively. A third trivial projection is simply the identity function.

The informative speaker S_1 has a state w, and reasons about L_0, preferring utterances u which maximize the L_0 posterior probability on w, up to the aspect of w specified by q. When q is the identity, $S_1(u|w) \propto L_0(w|u)$, and is thus a model of a speaker who prefers to choose the most informative utterance available.

For example, if trying to communicate state w, S_1 will prefer an utterance u which is compatible under the semantics with only w, over an utterance u' which is compatible with both w and w'.

The pragmatic listener L_1^Q jointly infers values for w and q, here with the assumption of a uniform prior over Q (which can be relaxed). The key dynamic is that the listener may hear an utterance u and infer a pair (w, q) where u is semantically incompatible with w (i.e. $[\![u]\!](w) = 0$) but where u conveys some aspect of w as determined by q.

Interpreting L_1^Q TODO: HMMMM, so nouns map to distributions over properties??? this is getting so so so messy

We begin by describing an interpretation of L_1^Q where nouns and adjectives denote sets of properties, similar to the interpretation in (Kao et al., 2014). Let the set of all properties under consideration be P, and call the function $E : U \to P$ take utterances to their corresponding sets.

TODO: note that in standard compositional semantic analyses of adjectival modification, the meaning of $[a \quad n]$ is a truth-valued function on entities. say: note as different as it seems

TODO: general definition of q, because, could be multiple dimensions

We refer to this as a *set-theoretic* interpretation of L_1^Q, in anticipation of the *vectorial* interpretation introduced in section .

TODO: A concrete example is as follows:

DISTRIBUTIONAL SEMANTICS

From a linguistic corpus, it is possible to obtain a mapping from words to points in a high-dimensional vector space, which has the property that semantic similarity of a pair of words a and b corresponds to metrics, such as cosine distance, between the vectors $\overrightarrow{a}$ and $\overrightarrow{b}$.

Mappings of this sort, commonly referred to as *word embeddings* or a *distributional model of word meaning*, can be obtained either by dimensionality reduction of a co-occurrence matrix (Pennington, Socher, & Manning, 2014), or by extracting the weights of a statistical model (Devlin, Chang, Lee, & Toutanova, 2018; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Peters et al., 2018).

In either case, word embeddings provide a way to empirically obtain fine grained connotations of lexical items (Mikolov et al., 2013), and have been used effectively in a number of NLP tasks (Dai & Le, 2015; Radford, Narasimhan, Salimans, & Sutskever, 2018). They have also been used to compute vectorial representations of phrases and sentences ((Socher et al., 2013; **?**)).

Metaphor is an obvious candidate for approaches that use distributional semantics: a wide variety of attempts have been made to leverage the information inherent in pre-trained word vectors for the detection, interpretation and paraphrase of metaphor (see Shutova (2016) for an overview of proposed systems.). TODO: needs your own citations

Our hypothesis is that, while the information in high quality word embeddings captures important aspects of meaning, a cognitively realistic model of metaphor interpretation should also incorporate Gricean reasoning, of the sort formalized in section . We now explain how the L_1^Q model can be combined with a distributional model of word meaning.

BAYESIAN PRAGMATICS WITH A DISTRIBUTIONAL SEMANTICS

The set-theoretic interpretation of L_1^Q takes states $w \in W$ to be sets of properties describing the source of the metaphor in question, and projections $q \in Q$ to be surjective functions out of W. The semantics maps utterances to functions from worlds to Boolean values, or equivalently, maps a pair (u, w) to 1 if they are compatible, and 0 otherwise.

We now introduce a *vectorial* interpretation of L_1^Q. Importantly, this requires no modification to equations (4-6). The crucial difference is that our state space W is now not just a set, but a vector space determined by a word embedding E, so that elements $\vec{w} \in W$ are vectors.[4]

As before, U is a set of adjectives.

The listener's prior In the set-theoretic interpretation of L_1^Q, where W can be finite, a discrete prior over W sufficed. In the present case, where W is necessary infinite (ranging continuously over real-valued vectors), we use a multivariate spherical Gaussian distribution, which can be parametrized by a vector $\vec{\mu}$ for the mean and a single scalar σ (the value of every diagonal entry of the covariance matrix).

We can view the listener's prior P_L as representing uncertainty over the position of the entity or concept that the source noun represents.[5] The goal of the speaker is to convey a position in the space (which they think represents the nature of the source noun's denotation) to the listener, and the goal of the listener is to infer what this position is. In this sense, a spatial reference game (Golland, Liang, & Klein, 2010) is being played, in an abstract word embedding space.

(7) $P_L(w) = P_{\mathcal{N}}(w|\mu = \overrightarrow{source}, \sigma = \sigma_1)$

The multidimensional Gaussian distribution weights most heavily those points nearest to its mean. By setting the mean at the image of the source noun under E, we encode the assumption that the speaker's point is in the neighborhood of the source noun. σ_1 is a hyperparameter of the model (see section)

The semantics A word embedding space has no explicit representation of truth. That is to say, while we can compare the similarity of a noun and an adjective according to a variety of metrics, we do not have a means of categorically determining the compatibility of that adjective and noun.

Mathematically speaking, this is not a problem, since the definition of L_0 in (4) requires only that the semantics $[\![u]\!]$ be a function $W \to \mathbb{R}$. We can define such a function as follows:

(8) $[\![u]\!](w) = P_{\mathcal{N}}(w|\mu = \overrightarrow{predicate_u}, \sigma = \sigma_2)$

The result of this definition is that the value of $[\![u]\!](w)$ is a real number which scales quadratically with the Euclidean distance between u and w. TODO: confirm

The advantage of defining a semantics in this way is that both the prior of L_0, shown in (7) and the likelihood, namely the semantics shown in (8), have the form of Gaussian distributions. As discussed in section , this allows for a closed form solution of L_0. As with the definition of the prior in (7), the semantics introduces a hyperparameter σ_2.

[4] We note that this generalization is natural, since the set-theoretic interpretation of L_1^Q can be viewed as a special case of the vectorial interpretation, for a vector space over the Boolean field (rather than the real field). That is, given a basis set of properties $P = P_0 ... P_n$, a vector of n 0s and 1s, with 1s corresponding to present properties and 0s corresponding to absent ones, is equivalent to a subset of P, i.e. the denotation of a word.

[5] Note that this prior does not represent lexical uncertainty over the meaning of the subject word, but rather uncertainty over what the entity or concept that the subject denotes is like.

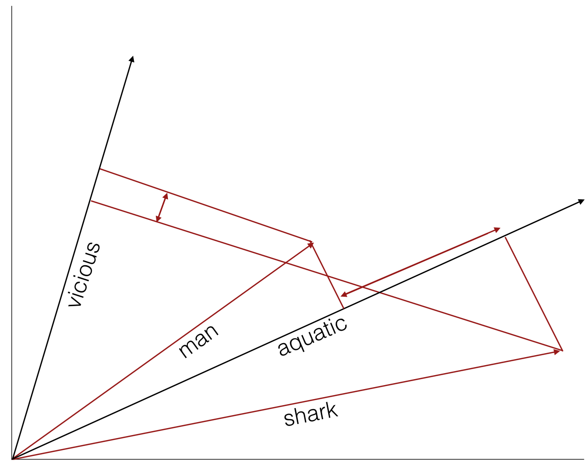

Figure 1: In this example, the vectors for *vicious* and *aquatic* each parametrize a QUD mapping all the points in the space (such as *man* and *shark*) to new points. These new points can be thought of as the positions of *man*, *shark* and so on in a new space which cares only about the position of the points with respect to the *vicious* vector.

Projections Finally, we need to supply an notion of a projection function q that is defined on our vector space, and to specify a set Q of projections.

Here, the natural counterpart of a partition function is the linear projection along a vector (or hyperplane) $\overrightarrow{v}$. This is a linear mapping from W to a subspace given by $\overrightarrow{v}$, which captures the degree to which $\overrightarrow{w}$ extends along a vector (or hyperplane) $\overrightarrow{v}$, and ignores orthogonal dimensions. Geometrically, it can be thought of as dropping a line from an input vector $\overrightarrow{w}$ at a right angle onto $\overrightarrow{u}$, as is depicted in figure 1.

To see why this is the natural analogue of the projection functions used in the set-theoretic interpretation of L_1^Q, note that when viewed as vectors in a vector space over the Boolean field, projection functions are precisely linear projections. Alternatively, note that in either case, projections q are idempotent maps $(W \to W)$.

As in the set-theoretic interpretation, we restrict ourselves to projections to a single dimension, i.e. projections along a vector, rather than a hyperplane. However, we no longer have an obvious set of projections Q, corresponding to an explicit set of properties P. This is because the basis vectors of a word embedding space such as GloVe or Word2Vec do not correspond to easily interpretable attributes of the words in the space.

To obtain such a set Q, we first note that since the denotations of words are vectors in W, any word parametrizes a linear projection q. For instance, we can think of *vicious* as parameterizing a *viciousness* projection, which measures how far the denotations of all other points in the space fall along $\overrightarrow{vicious}$.

As such, we can specify Q as a set of words. In general, it makes sense to choose a set of gradable adjectives, so that the projection of a noun n onto $\overrightarrow{v}$ amounts to asking: to what extent is n v? We discuss the exact choice of Q in section .

INFERENCE

Calculating the posterior of L_1^Q given an utterance u is far more difficult in the vectorial interpretation than the set-theoretic one. Because $P_L(W)$ is now a continuous distribution, inference by enumeration is no longer possible, and either analytic or approximate methods are required. We employ a mix of the two; the L_0 and S_1 posteriors can be calculated analytically, while L_1^Q requires us to develop an approximate inference algorithm. Our inference algorithm for L_1^Q is implemented in Tensorflow, and will be made publicly available. We describe this algorithm in parts, working up from the L_0.

L$_0$ Inference Intuitively, the vectorial interpretation of L_0 amounts to the process shown in figure 2, where a ball, corresponding to the prior, is moved in the direction of another point, the received utterance. To calculate L_0 analytically, we make use of Gaussian self-conjugacy, which allows a distribution with a Gaussian prior and likelihood term to be rewritten as a single Gaussian of different parameters.

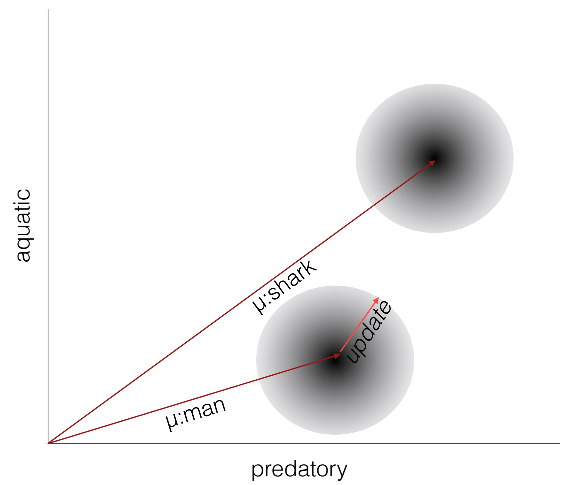

Figure 2: 2D Visualization of L_0

S$_1$ Inference Since U remains a finite set of utterances, its posterior can be computed exactly, as for the set-theoretic interpretation of S_1.

L$_1$ Definition The L_1 posterior is a joint distribution over one continuous and one discrete random variable. We are unable to use conjugacy to compute it analytically, but equally unable to compute it exactly.

Moreover, we find that Mean Field Variational Inference TODO CITATION is not well suited since the posterior is non-Gaussian (see figure 4).

Why not HMC?

details:

We devise the following approximate inference algorithm.

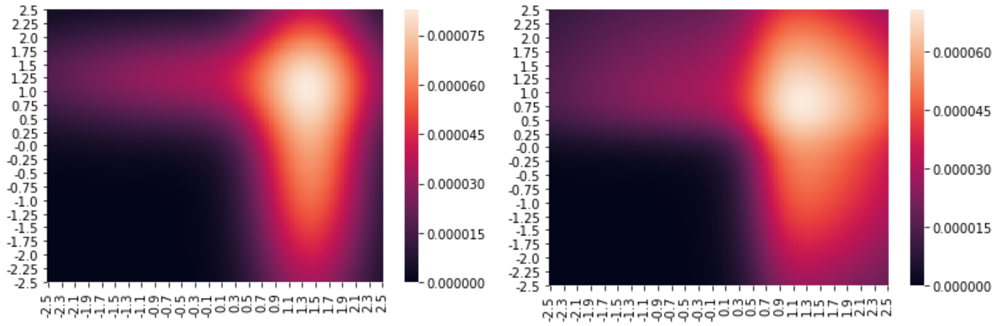

Figure 3: Comparison between the L_1^Q posterior under our inference algorithm (left) and under the true posterior (right), which is calculable in this two dimensional case (up to discretization of the space).

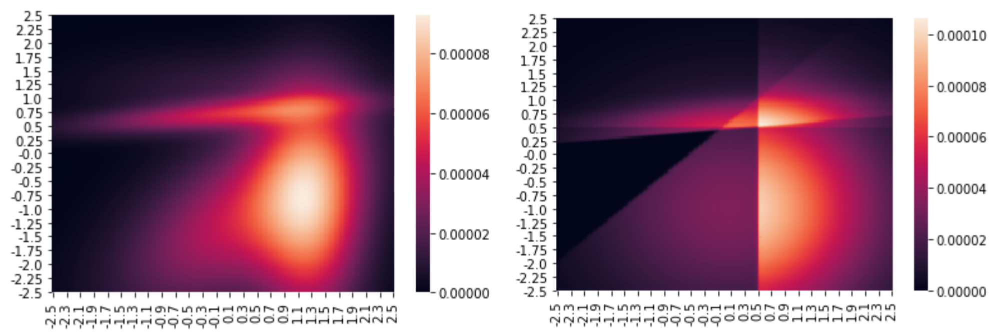

Figure 4: TODO Comparison between the L_1^Q posterior under our inference algorithm (left) and under the true posterior (right), which is calculable in this two dimensional case (up to discretization of the space).

To compute the joint posterior over W and Q, we first compute the conditional posterior $P(w|q,u)$ for each $q \in Q$, and then determine the marginal posterior on Q, $P(Q)$ TODO HOW THO?

by the following observation: for an orthonomal basis of W including q, the mean and variance of this posterior is the same as the prior in every dimension except q. In the single dimension of interest, we approximate the conditional posterior by performing gradient descent to find the mean of the posterior, and discretizing the now one-dimensional prior.

with compute variance of this dimension by... this allows us to compute the marginal distribution

To calculate the mixture weights (i.e. $P_c(q)$), we use the fact that the determinant of the covariance matrix UM TODO: true? is proportionally to the integral of

Interpreting results from L_1^Q

We now have an algorithm for approximating the posterior distribution of L_1^Q after hearing a metaphor u. One way of converting this into an interpretable prediction is to examine the marginal distribution over Q, and take adjectives q which have high probability under this distribution to be interpretations of the metaphor u that the model considers likely.

For example, on the metaphor TODO example

The model assigns BLAH TODO probability to ETC TODO

EXPERIMENTAL VALIDATION

In order to test our model on human judgments, we design an experiment which compares the L_1^Q interpretations of metaphors to a baseline model which uses word embeddings but no pragmatic reasoning.

Experimental Design Choice of metaphors **?** provides a list of $\tilde{8}00$ AN metaphors, like EXAMPLES TODO, gathered by TODO how. We take the least frequent (by COCA[6] bigram count). We do this to filter out conventionalized metaphors, like TODO EXAMPLE.

In our experiment, each participant is shown a series of 11 TODO CHECK metaphors. For each metaphor, they are asked to rate on a slider four adjectives representing interpretations of the metaphor, of which two are selected by L_1^Q and two from a baseline model (details below).

SHOW SLIDE: TODO An example of one item in the experiment is shown in figure **??**

The experiment was run on Mechanical Turk, with TODO HOW MANY participants, all of whom are native English speakers. Participants who fail a simple preliminary instruction following task are excluded.

[6] TODO: CITATION INFO: Davies, Mark. (2011) N-grams data from the Corpus of Contemporary American English (COCA). Downloaded from http://www.ngrams.info on January 24, 2019.

Baseline model The aim of our experiment is to determine whether pragmatic reasoning results in better interpretations of metaphors, according to human judgments. As such, a natural baseline model to compare against is a model with a distributional semantics that does not make use of the pragmatic reasoning process inherent in L_1^Q.

model: given Q take the mean of the *subject* and *predicate* (or noun and adjective) and rank $q \in Q$ by cosine distance to this mean. We use mean (which is a weighted sum) in light of the EXAMPLES and cosine distance as the TODO FINISH

This proves to be a very effective baseline: EXAMPLES

justification for: cosine distance mean: weighted sum

$\boldsymbol{L_1^Q}$ **hyperparameters** We use the largest available (300 dimensional) GloVe vectors, as our word embedding E. For a given metaphor of the form *adjective noun*, we select a set of 100

as adj_{alt} *noun* for the adjectives closest by cosine distance to $\overrightarrow{noun}$ BE CAREFUL BECAUSE OF YOUR NEW NOTION OF ALTERNATIVES

Similarly, we select a set of projections corresponding to the hundred closest adjectives to the mean of the subject and predicate (as in the baseline model).

mean centering

By tuning on a validation set of hand-selected metaphors, we find that $\sigma_1 = \sigma_2 = 0.1$ is the best value for the variances of the Gaussians used in the prior and semantics respectively.

IMPORT stuff

Analysis Statistical analysis: To analyze the results of the experiment, we ran a linear mixed effects model, predicting slider rating from baseline vs. L_1 metaphor, with by-metaphor and by-participant random intercepts. We find a significant correlation TODO DATA

TODO: figure: each metaphor, histogram

TODO: consider truth of:

DISCUSSION

valuable evidence that Gricean reasoning about informativity and relevance is key to the interpretation of metaphor, and language more generally.

We have also shown that the technical challenges in adapting a nested Bayesian model of pragmatics to a continuous setting, while significant, are not insurmountable.

DISCUSS UTTERANCES being finite set: The prior for S_1, which is over possible utterances, is finite, and therefore straightforward to compute[7]. say that other RSA work attempts to relax this assumption, and that fusing the two is a direction for future work

[7] As discussed in section , having a finite set of utterances is theoretically objectionable, but for the time being, a necessary modeling assumption.

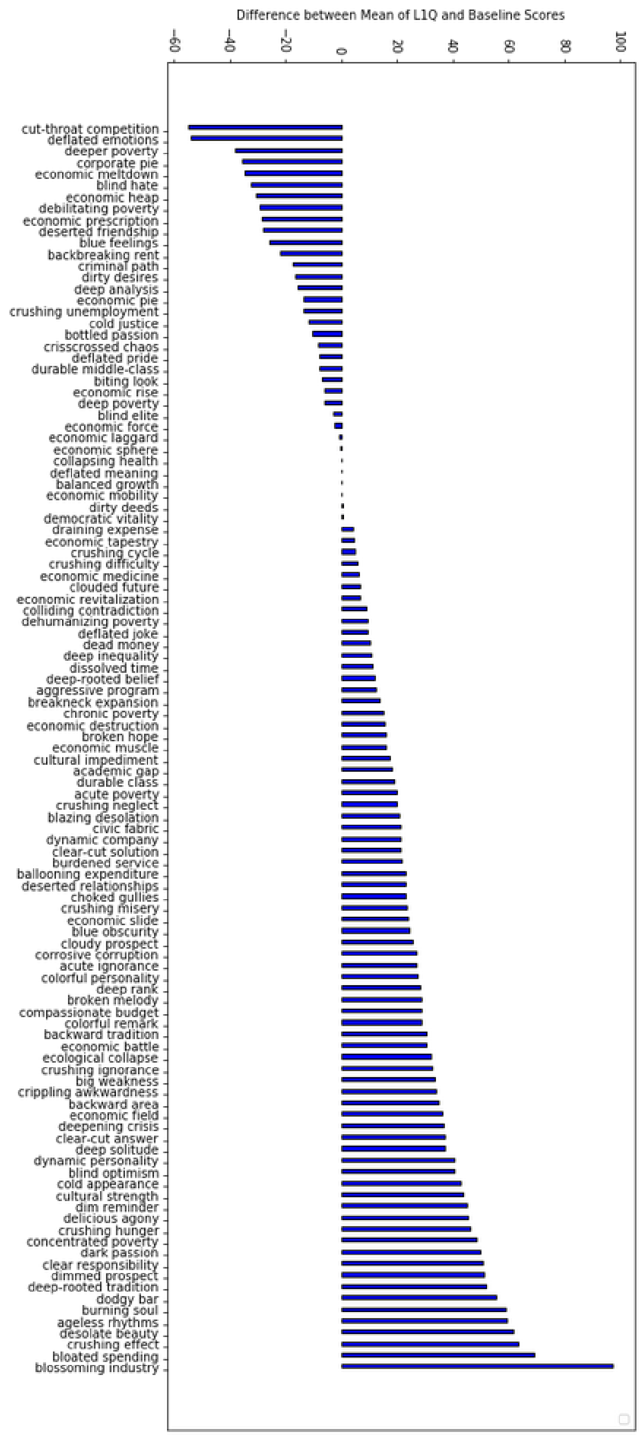

Figure 5: This figure shows all 109 metaphors used in the experiment, with the height of the bar corresponding to the difference of the mean rating given to that metaphor under the LL_1^Q model and baseline model. This shows that for roughly 75% of the metaphors, the LL_1^Q interpretation is preferred.

optimization needed to make this nlp tool

TODO: discuss multidimensional interpretations However, the model is extensible to the more general case, which we regard as an important direction for future work, since it allows us to capture the ability of metaphors to talk about many dimensions of meaning simultaneously.

REFERENCES

Dai, A. M., & Le, Q. V. (2015). Semi-supervised sequence learning. In *Advances in neural information processing systems* (pp. 3079–3087).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Golland, D., Liang, P., & Klein, D. (2010). A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 410–419).

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Kao, J. T., Bergen, L., & Goodman, N. (2014). Formalizing the pragmatics of metaphor understanding. In *Cogsci.*

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. *Chicago, IL: University of*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Emnlp* (Vol. 14, pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding with unsupervised learning* (Tech. Rep.). Technical report, OpenAI.

Shutova, E. (2016). Design and evaluation of metaphor processing systems. *Computational Linguistics*.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., … others (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (emnlp)* (Vol. 1631, p. 1642).