OPEN MIND
Discoveries in Cognitive Science

an open access journal

Open-Domain Metaphor Interpretation: Distributional Semantics for Bayesian Pragmatics

Anonymized

Keywords: metaphor, informativity, Bayesian pragmatics, distributional semantics

ABSTRACT

Humans interpret metaphors, like *Time is a thief* or *My lawyer is a shark*, with relative ease, incorporating contextual knowledge to determine which aspects of the predicate (*thief, shark*) are true of the subject (*time, my lawyer*). One theory of the process underlying metaphor interpretation, proposed by Grice (1975), is that a listener reasons about an informative speaker (who in turn reasons about the listener) to update their beliefs about the subject and the relevant dimensional of meaning.

This reasoning process can be modeled probabilistically in the *Rational Speech Acts framework* (Frank & Goodman, 2012). 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 open domain predicative and adjectival metaphors without manually stipulating the denotations of the words they contain. We find a significant preference in human judgments for our model over a comparable word embedding model without explicit pragmatic reasoning.

Citation:

DOI:
http://dx.doi.org/

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

Received:
Accepted:
Published:

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

Corresponding Author:
Anonymized

The MIT Press

INTRODUCTION

Metaphor presents a compelling theoretical challenge for the understanding of meaning in natural language. 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 of 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 formalization of Gricean reasoning, the *Rational Speech Acts* (RSA) framework (Frank & Goodman, 2012). 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 to the speaker's model of the listener, up to the projection q.

This model provides an account of predicative metaphors (those of the form *A is a B*) and adjective-noun (AN) metaphors (like *fiery temper*). However, in order to generate predictions from the model, previous instantiations of the model have required the manual construction of a semantics for the words involved, restricting the model's scalability.

Our contribution By adapting an RSA model of non-literal meaning to a word embedding (distributional) semantics, we obtain a system capable of interpreting open domain predicative and adjectival metaphors, without the need for hand-specified word meanings. This adaption requires a generalization of projection functions to a vector space setting, and a novel inference algorithm to calculate the now continuous posterior distribution of L_1^Q. This permits what is to our knowledge the first open domain evaluation of an RSA model of pragmatic reasoning. We show that our model of metaphor interpretation significantly outperforms a baseline which uses a word embedding semantics without explicit pragmatic reasoning.

OVERVIEW OF METAPHOR

Metaphor exists in many syntactic forms. For present purposes, we focus on copular predicates (e.g. *Jane is a soldier*) and AN noun phrases (e.g. *fiery temper*). We refer to the predicated or modified noun (*Jane*, *temper*) as the *target* of the metaphor and the predicate or adjective (*soldier*, *fiery*) as the *source* (see Lakoff and Johnson (1980) for the more general sense of these terms).

For a given metaphor, only certain properties of the target are described by the source, and which these are depend on the metaphor and the context. For instance, (2), said of a sleeping dog, could convey that it is unresponsive, but said of a large alert dog, could convey that it is heavy.

(2) The dog is a rock.

While certain metaphors are conventional - comparing someone to a lion tends to connote bravery - examples like (2) 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 *dog* and *rock*).

[1] These are often referred to in RSA literature as *Questions under Discussion*.

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 & Goodman, 2012).

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 (Frank & Goodman, 2012) is incapable of interpreting metaphors, due to the strict assumption that the speaker's utterances are literally true. To address this, Kao et al. (2014) proposes a model L_1^Q, defined in (5).

(3) $L_0(w|u) \propto [\![u]\!](w) \cdot P_L(w)$

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

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

The literal listener L_0 represents a model of a listener that, given an utterance $u \in U$ updates their belief about the world $w \in W$ in accordance with the semantics $[\![\cdot]\!]$ and their prior P_L. This semantics, usually a function $U \to (W \to \{0,1\})$, represents the conventional association between states w and utterances u which the speaker and listener take as given.

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.

The informative speaker S_1 has a state w they want to communicate, 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. If q is the identity function, $S_1(u|w) \propto L_0(w|u)$, and is thus a model of a speaker who prefers to choose the most informative utterance available.

The pragmatic listener L_1^Q jointly infers values for w and q. 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.

We can view L_1^Q as a model of metaphor interpretation as follows, using the example in (6). The goal of the speaker is to communicate a state w along dimension q, representing what John is like along a certain dimension. To this end, they choose a predicate u, here *shark*. The goal of the listener on hearing (6) is to infer w as well as the *dimension* of w that the speaker cares about.

(6) John is a shark.

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 $[\![\cdot]\!]$. (We assume throughout that the prior P_{L_Q} over Q is uniform.) Jointly, we say that these determine an *interpretation* of L_1^Q.

One possible interpretation, similar to what is provided by Kao et al. (2014), treats points in the state space W as n-tuples of truth values, corresponding to Boolean properties

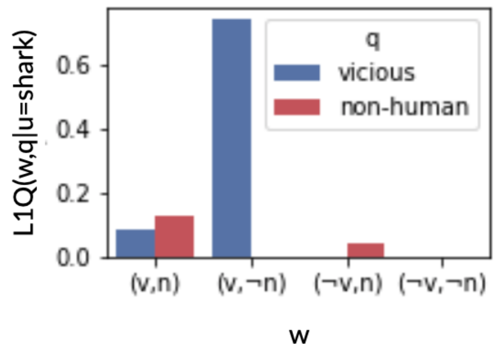

Figure 1: Figure showing the posterior distribution of L_1^Q on hearing *shark*. v and n abbreviate the Boolean variables *vicious* and *non-human* respectively.

$w \subset P$. We refer to this as a *set theoretic* interpretation of L_1^Q. We give a very simple example below, for $P = \{\text{vicious}, \text{non-human}\}$:

- $P_L = \{(\text{vicious} = T, \text{non-human} = T) : 0.075,$
  $(\text{vicious} = T, \text{non-human} = F) : 0.675,$
  $(\text{vicious} = F, \text{non-human} = T) : 0.025,$
  $(\text{vicious} = F, \text{non-human} = F) : 0.225\}$
- $U = \{shark, silence\}$
- $Q = \{q_{vicious} = \lambda(x, y) : x, \quad q_{non-human} = \lambda(x, y) : y)$

We assume a semantics $[\![\cdot]\!]$ in which *shark* is compatible only with (vicious, non-human), and *silence* is compatible with every state. Note that the projections map each tuple to its value at a single property.

The results of L_1^Q hearing *shark* are shown in figure 1. The key fact to note is that the prior belief that John is not literally an animal leads L_1^Q to conclude that the speaker cares about conveying the viciousness dimension (i.e. projection $q_{vicious}$), and that John is vicious.

Importantly, L_1^Q can do more than simply using prior knowledge to interpret literally false statements in a flexible way. It is also capable of reasoning about alternative utterances: for instance, suppose we add a third property, *quickness*, so that *shark* is compatible only with $(\text{vicious} = T, \text{non-human} = T, \text{quick} = T)$, and also add a third utterance, *hummingbird*, compatible with only $(\text{vicious} = F, \text{non-human} = T, \text{quick} = T)$.

In this second example, when L_1^Q hears *shark*, it infers that John is more likely vicious than quick. This is because a speaker who wanted to communicate that John is vicious would only be able to use the utterance *shark*, whereas a speaker who wanted to communicate that John is quick would be able to choose between either *shark* or *hummingbird*. The utterance *shark* is therefore more likely to have been produced by the speaker trying to communicate John's viciousness.

L_1^Q can also model AN metaphors in a similar way. For instance, for a phrase like *fiery temper*, we say that the goal of a listener is to decide what is true of the temper in question given that the speaker has modified it with *fiery*.

DISTRIBUTIONAL SEMANTICS

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

Mappings of this sort, commonly referred to as *word embeddings* or *distributional models 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) trained on a separate task.

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 (Coecke, Sadrzadeh, & Clark, 2010; 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.).

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 the RSA framework. We now explain how L_1^Q 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 given metaphor, and a semantics to be a function $U \rightarrow (W \rightarrow \{0,1\})$.

We now introduce a *vectorial* interpretation of L_1^Q. Importantly, this requires no modification to equations (3-5). The crucial difference is that our state space W is now not just a set, but a vector space determined by a word embedding $E : U \rightarrow W$, so that elements $\vec{w} \in W$ are vectors.[2] For our application of the model, we assume U is a set of adjectives.

The listener's prior In the set-theoretic interpretation of L_1^Q, with finite W, a discrete prior P_L over W sufficed. In the present case, where W is necessarily infinite (ranging over all 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 define the prior over projections P_{L_Q} to be uniform.

$$P_L(w) = P_{\mathcal{N}}(w|\mu = E(target), \sigma = \sigma_1) \quad (7)$$

[2] 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, consider an n-tuple of Booleans w as a vector of 0s and 1s, with P providing the basis of the space.

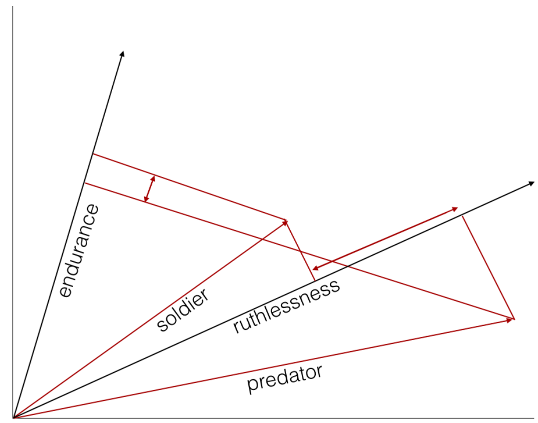

Figure 2: In this hand-constructed 2D example, vectors for $\overrightarrow{soldier}$ and $\overrightarrow{predator}$ are mapped onto subspaces given by $\overrightarrow{endurance}$ and $\overrightarrow{ruthlessness}$.

We can view P_L as representing uncertainty over the position of the entity or concept that the target noun (e.g. *man* in "The man is a shark") represents.[3] The goal of the speaker is to convey a position in the space to the listener, and the goal of the listener is to infer what this position is. In this sense, a spatial reference game is being played (Golland, Liang, & Klein, 2010), in an abstract word embedding space.

The multidimensional Gaussian distribution weights most heavily those points nearest to its mean. By setting the mean of the prior as $E(target)$, we encode the listener's assumption that the meaning the speaker wishes to communicate is in the neighborhood of the source noun. σ_1 is a hyperparameter of the model.

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.

As far as our model is concerned, this is not a problem since the definition of L_0 in (3) requires only that the semantics $[\![\cdot]\!]$ be a function $U \to (W \to \mathbb{R})$. We can define such a function as follows, with σ_2 as a hyperparameter:

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

The result of this definition is that the value of $[\![u]\!](w)$ is a real number which decreases with the Euclidean distance between u and w. 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, which allows for a closed form solution of L_0.

[3] This bears comparison to the *conceptual space* semantics of Gärdenfors (2004).

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 such projections. For this, we use linear projections[4] along a vector (or hyperplane) $\overrightarrow{v}$ capturing the degree to which each $\overrightarrow{w}$ extends along $\overrightarrow{v}$, ignoring 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 depicted in figure 2.

In practice, we restrict ourselves to projections along a vector, rather than a larger subspace. To obtain a set Q of projections, 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 the word *vicious* as parameterizing a *viciousness* projection, which measures how far the denotations of all other points in the space fall along $\overrightarrow{vicious}$. We choose Q as 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?

INFERENCE

Because $P_L(w)$ is a continuous distribution in the vectorial interpretation of L_1^Q, 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.[5] We describe this algorithm in parts, working up from the L_0.

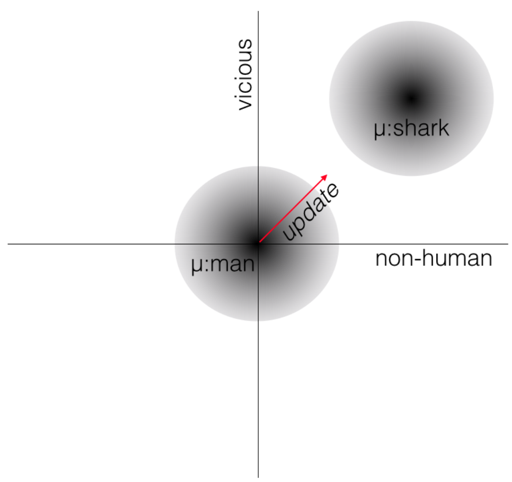

Figure 3: 2D depiction of vectorial L_0, for $\overrightarrow{man}$ = (0,0) and $\overrightarrow{shark}$ = (1,1).

L_0 Inference Intuitively, the vectorial interpretation of L_0 amounts to the process shown in figure 3, where a ball, corresponding to the prior, is moved in the direction of the point cor-

[4] 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 $(W \to W)$.

[5] Inference for all our models is implemented in Tensorflow, and will be made publicly available.

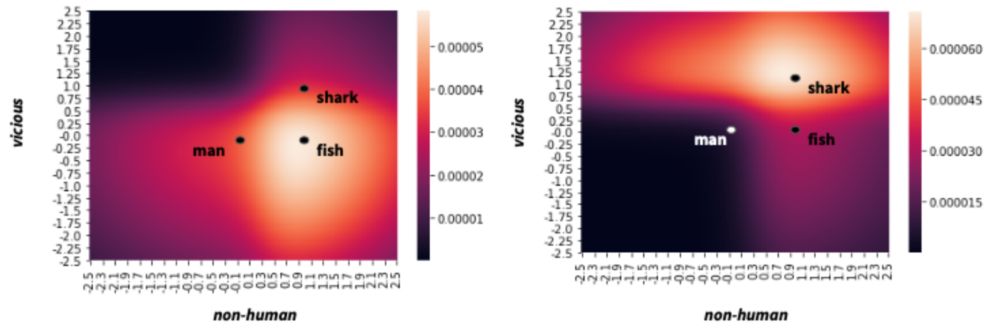

Figure 4: Heatmaps visualizing the inferred L_1^Q marginal posterior over worlds given *fish* (left) and *shark* (right), with $U = \{man, shark, fish\}$, hand-chosen denotations overlaid, and $\sigma_1 = 5.0, \sigma_2 = 0.5$.

responding to the received utterance. To calculate L_0 analytically, we make use of Gaussian conjugacy. When the prior P_L is defined as in Equation 7, and the semantic interpretation is defined as in Equation 8, then conjugacy implies that the listener posterior is given by:

$$(9) \quad L_0(w|u) = P_{\mathcal{N}}(w|\mu = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \cdot (\frac{\text{E(target)}}{\sigma_1^2} + \frac{\text{E(source)}}{\sigma_2^2}), \sigma = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2})$$

S_1 Inference The speaker is defined by Equation 4, which in the continuous case can be rewritten as:

$$(10) \quad S_1(u|w,q) \propto \int_{w'} \delta_{q(w)=q(w')} \cdot L_0(w'|u)$$

This integral is computing the marginal probability of w_q, the projection of world w onto QUD vector q. From Equation 9, $L_0(\cdot|u)$ is a normally distributed random variable, and therefore projection of this random variable onto a linear subspace is also normally distributed, providing a closed-form solution to S_1.

L_1 Inference The L_1 posterior is a joint distribution over one continuous and one discrete random variable. Because of the linear structure of the problem, we are able to devise a near-exact inference algorithm for the marginal distribution over QUDs in Q, derived as follows:

$$\begin{aligned}
L_1(q|u) &= \int_{\mathbb{R}^n} L_1(w,q|u)dw = \frac{1}{K}P_{L_Q}(q)\int_{\mathbb{R}^n} P_L(w)S_1(u|w,q)dw \\
&= \frac{1}{K}P_{L_Q}(q)\int_{\mathbb{R}^n} P_L(w_q,w^{\perp})S_1(u|w_q,q)dw = \frac{1}{K}P_{L_Q}(q)\int_{\mathbb{R}^n} P_L(w_q)P_L(w^{\perp})S_1(u|w_q,q)dw \\
&= \frac{1}{K}P_{L_Q}(q)\int_{Q^{\perp}} P_L(w^{\perp})dw^{\perp}\int_Q P_L(w_q)S_1(u|w_q,q)dw_q \\
&= \frac{1}{K}P_{L_Q}(q)\int_Q P_L(w_q)S_1(u|w_q,q)dw_q
\end{aligned}$$

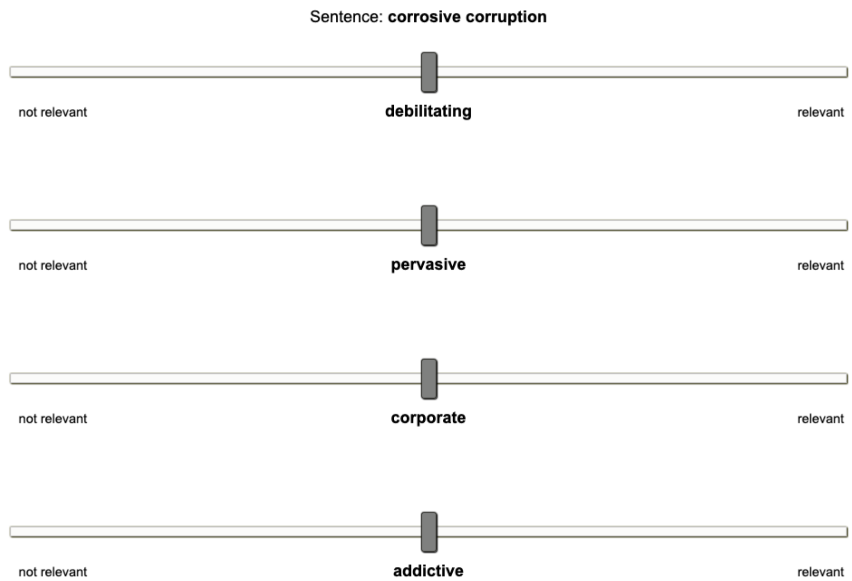

Figure 5: An item in the experiment. Item order, and in-item order of the 4 adjectives from L_1^Q and baseline models is randomized.

Here K is a normalizing constant, $w, q \in \mathbb{R}^n$, and w_q is the projection of w onto the vector q. Q is the subspace of $\mathbb{R}^n$ spanned by the vector q, and $Q^\perp$ is the orthogonal complement of Q. The vector $w^\perp$ is the projection of vector w onto the subspace $Q^\perp$. The final equation is a one-dimensional integral, and can be computed using a discrete approximation.[6] The constant K can be found from the constraint $\sum_q L_1(q|u) = 1$.

Figure 4 provides a visualization of the L_1^Q posterior in a simple 2D case corresponding to the example used for the set-theoretic example in figure 1.

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. We convert this into an interpretable prediction by selecting the adjectives $q \in Q$ which have high probability under the marginal posterior distribution over Q as interpretations of the metaphor.

EXPERIMENTAL EVALUATION

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.

[6] We use a Gaussian approximation, which easily generalizes to the setting of multi-dimensional QUDs.

L_1^Q adjectives	BASELINE adjectives	metaphor
undistorted, faceless	wolfish, pitiless	cut-throat competition
euphoric, giddy	despondent, chagrined	deflated emotions
greater, moral	alarming, impoverished	deeper poverty
traditional, fresh	homemade, organizational	corporate pie
imminent, speculative	recessionary, substantial	economic meltdown
evil, suicidal	angry, nasty	blind hate
stellar, untold	humiliating, regional	economic heap
unbearable, mental	afflicted, worse	debilitating poverty
bipartisan, anti	political, affordable	economic prescription
quiet, idyllic	important, everlasting	deserted friendship
beautiful, particular	fearful, sincere	blue feelings
unproductive, crummy	ungodly, undoable	backbreaking rent
dangerous, possible	alleged, guilty	criminal path
cynical, vulgar	lustful, insatiable	dirty desires
difficult, obvious	preliminary, analytical	deep analysis
easy, traditional	possible, social	economic pie
successive, miserable	fiscal, brutal	crushing unemploymen
wrong, new	warmer, appropriate	cold justice
refreshing, copious	unopened, divine	bottled passion
awash, treacherous	undulating, unprecedented	crisscrossed chaos
jubilant, galvanized	buoyant, wobbly	deflated pride
desirable, modest	residential, fragile	durable middle-class
amusing, cynical	nasty, afraid	biting look
recent, likely	slight, annual	economic rise
enduring, rich	afflicted, global	deep poverty
gifted, renowned	illiterate, national	blind elite
federal, necessary	southern, aggressive	economic force
cyclical, breakneck	zippy, sluggish	economic laggard
immense, intellectual	overall, analogous	economic sphere
precarious, creaking	adequate, prolonged	collapsing health
stale, unrealistic	wobbly, metaphorical	deflated meaning
optimal, budgetary	potential, adequate	balanced growth
entrepreneurial, greater	agile, possible	economic mobility
wicked, dishonest	virtuous, righteous	dirty deeds
cultural, genuine	astonishing, parliamentary	democratic vitality
rewarding, excess	dry, lavish	draining expense
vibrant, medieval	worse, grotesque	economic tapestry
relentless, debilitating	battered, periodic	crushing cycle
conquering, immense	embarrassing, overwhelming	crushing difficulty
academic, practical	advanced, recent	economic medicine
pessimistic, bleak	worried, serious	clouded future
ambitious, vibrant	strategic, postwar	economic revitalization
accidental, divergent	unspoken, intolerable	colliding contradiction
appalling, wretched	alarming, inhuman	dehumanizing poverty
overblown, depressing	bizarre, depressed	deflated joke
alone, unclear	male, additional	dead money
enduring, evident	homogeneous, big	deep inequality
supreme, founding	constitutional, socialist	dissolved time
longstanding, unshakeable	baneful, underlying	deep-rooted belief
ambitious, innovative	outspoken, risky	aggressive program
unchecked, continual	phenomenal, zippy	breakneck expansion
rampant, worsening	acute, fatal	chronic poverty
immediate, immense	atomic, illegal	economic destruction
unable, wrong	heavy, anxious	broken hope
internal, sustained	neurological, apparent	economic muscle
inherent, legitimate	ancient, glaring	cultural impediment
ideological, fiscal	philosophical, administrative	academic gap
good, higher	poor, efficient	durable class
worsening, prevalent	adverse, global	acute poverty
successive, heartbreaking	aggravated, gigantic	crushing neglect
unrelenting, magnificent	windy, torrid	blazing desolation
moral, secular	parliamentary, flexible	civic fabric
interactive, competitive	nonlinear, real	dynamic company
principled, definite	lasting, flexible	clear-cut solution
subsidized, costly	senior, insolvent	burdened service
unsustainable, voluminous	macroeconomic, insufficient	ballooning expenditure
lonely, idyllic	monogamous, featureless	deserted relationships
disoriented, contorted	unexplored, undulating	choked gullies
conquering, heartbreaking	appalling, gigantic	crushing misery
dramatic, depressing	recent, greater	economic slide
quiet, nondescript	forlorn, patterned	blue obscurity
foreseeable, pessimistic	windless, worrisome	cloudy prospect
pervasive, debilitating	addictive, corporate	corrosive corruption
widespread, alarming	abject, escalating	acute ignorance
flamboyant, humorous	multicolored, garish	colorful personality
certain, other	painful, huge	deep rank
minor, whole	good, numerous	broken melody
optimistic, generous	national, ambitious	compassionate budget
provocative, unflattering	memorable, charming	colorful remark
feudal, authoritarian	great, aristocratic	backward tradition
civil, disastrous	sluggish, allied	economic battle
consequent, structural	apparent, freshwater	ecological collapse
terrible, oppressive	absolute, innate	crushing ignorance
potential, tremendous	ready, mild	big weakness
unbearable, nagging	escalating, playful	crippling awkwardness
impoverished, prone	parallel, important	backward area
scientific, theoretical	potential, advanced	economic field
economic, ongoing	bilateral, unprecedented	deepening crisis
succinct, emphatic	goddamned, possible	clear-cut answer
spiritual, immense	otherworldly, heavy	deep solitude
intelligent, creative	narcissistic, aggressive	dynamic personality
naive, suicidal	poor, bald	blind optimism
usual, strange	wet, unlikely	cold appearance
enduring, emotional	aesthetic, newfound	cultural strength
bleak, glum	ghastly, enduring	dim reminder
bittersweet, unimaginable	salty, baked	delicious agony
crippling, debilitating	afflicted, nationwide	crushing hunger
productive, scarce	active, hydrochloric	concentrated poverty
romantic, pure	youthful, pale	dark passion
aware, direct	administrative, unable	clear responsibility
pessimistic, bleak	enticing, heightened	dimmed prospect
longstanding, generational	ingrown, hallowed	deep-rooted tradition
sleazy, seedy	new, formal	dodgy bar
sacred, holy	yellow, kindred	burning soul
timeless, immortal	rejuvenated, mellow	ageless rhythms
serene, otherworldly	overgrown, lyrical	desolate beauty
disastrous, debilitating	fierce, avenging	crushing effect
wasteful, unsustainable	lifeless, tough	bloated spending
thriving, vibrant	pigtailed, unraveled	blossoming industry

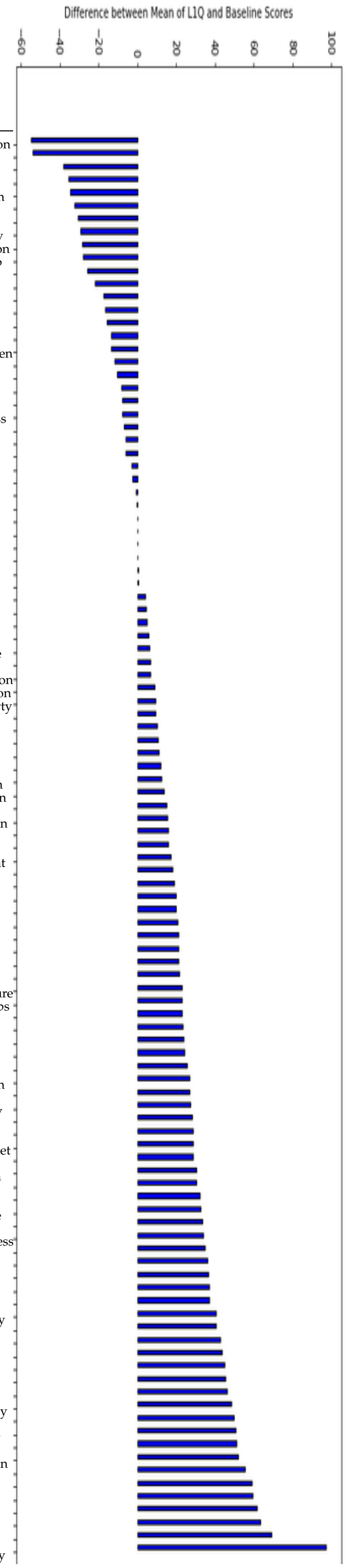

Figure 6: The 109 metaphors used in the experiment, and baseline and L_1^Q interpretations. Bar positions right of center indicate a preference for the pragmatic model.

Experimental Design Tsvetkov, Boytsov, Gershman, Nyberg, and Dyer (2014) provides a corpus of ~800 AN metaphors, gathered by human annotators, from which we select ~100 of the least frequent by bigram count[7] for our experiment, in order to filter out conventionalized metaphors. Our full set of 109 metaphors is shown in figure 6.

In our experiment, each participant is shown a series of 12 metaphors, selected randomly from the total 109. 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). An example is shown in figure 5.

The experiment was run on Mechanical Turk, with 99 participants, all of whom are native English speakers. Participants who failed to follow instructions on a test item were excluded, leaving 60 participants (analyses remain significant if all participants are included).

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.

Our baseline model is defined as follows: for a given metaphor of the form $(a\ n)$, we take the mean of the adjective a and noun n. The two nearest (measured by cosine distance) adjectives q to this mean are our baseline interpretations for the metaphor. We use the mean (a weighted sum) in light of the effectiveness of vector addition in deriving representations of phrasal and sentence meanings from constituent words (Grefenstette, 2013; Mitchell & Lapata, 2010; Socher et al., 2013). Cosine distance is a standard metric of similarity used for word embeddings (Pennington et al., 2014).

L_1^Q hyperparameters We use the largest available (300 dimensional) GloVe vectors, as our word embedding E. For each AN metaphor $(a\ n)$, we specify U as a set of 101 alternative utterances, consisting of a and 100 of the nearest adjectives (by cosine distance) to n. These adjectives are chosen from the set of the 1425 adjectives with concreteness ranking > 3.0 in the concreteness corpus of (Brysbaert, Warriner, & Kuperman, 2014), to exclude abstract nouns.

Similarly, we select a set Q of projections corresponding to the hundred closest adjectives to the mean of the subject and predicate (the method of adjective choice in the baseline model), and take P_{L_Q} to be a uniform distribution over Q.

By tuning on an independent validation set of hand-selected metaphors, we choose $\sigma_1 = \sigma_2 = 0.1$. We select the adjectives corresponding to the two projections with highest marginal posterior mass under L_1^Q as the interpretations provided from our model in the experiment.

Analysis The results, shown in figure 6, were analyzed using a mixed-effects model with random slopes and intercepts for items and participants. The target interpretations were rated significantly higher than the baseline interpretations (β=13.8, t=5.3, p<0.001).

[7] N-gram data from the Corpus of Contemporary American English (Davies, 2011).

DISCUSSION

We have shown that it is possible to scale Bayesian pragmatic reasoning to a distributional semantics and by so doing to obtain a model of metaphor interpretation. Our evaluation, the first such open domain test of an RSA model of interpretation, indicates that the principles of pragmatic reasoning continue to operate at this scale, and are key to obtaining human-like interpretations of metaphors. We see this as an important step towards a cognitively accurate and computationally tractable model of pragmatic language interpretation and production in general.

REFERENCES

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*(3), 904–911.

Coecke, B., Sadrzadeh, M., & Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

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

Davies, M. (2011). Word frequency data: Corpus of contemporary american english. *Provo, UT: COCA*.

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.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

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).

Grefenstette, E. (2013). Category-theoretic quantitative compositional distributional models of natural language semantics. *arXiv preprint arXiv:1311.1539*.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

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).

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, *34*(8), 1388–1429.

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., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642).

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., & Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (Vol. 1, pp. 248–258).