Metaphor and Linguistic Creativity: Pragmatic Reasoning with Distributional Semantics

Reuben Cohn-Gordon[a] **and Leon Bergen**[*,b]

[a]Stanford University; [b]University of California San Diego

This manuscript was compiled on June 1, 2019

Humans create and interpret novel metaphors, like *Time is a thief* or *My lawyer is a shark*, with relative ease, incorporating world knowledge to determine which aspects of the predicate (*thief, shark*) are true of the subject (*time, my lawyer*). Here we present a computational theory of metaphor, according to which metaphorical interpretations arise from joint, cooperative reasoning between a speaker and listener. We combine a Bayesian model of this reasoning process with empirically learned *word embeddings* which are used to provide an underlying representation of word meaning. This allows for open-domain interpretation of predicative and adjectival metaphors. We find a significant preference in human judgments for our model over a system which uses word embeddings without a explicit representation of inter-agent reasoning, providing evidence that reasoning about an informative and relevant speaker is key to understanding non-literal language.

metaphor | informativity | Bayesian pragmatics | distributional semantics

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 journalist, 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 *pragmatic* view of metaphor, proposed by Grice (1), takes the meaning conveyed by sentence (1) to be the result of joint, cooperative reasoning between a speaker and listener. That is, the speaker has some information about Jane, and wants to communicate some aspects of this information to the listener using the predicate *soldier*. The listener, in turn, jointly infers what Jane must be like and what aspects of Jane are relevant for the speaker.

Interpreting figurative language We build on a previous model of metaphor interpretation (2), developed as part of a probabilistic framework for pragmatic reasoning (3), which uses *projection functions* to determine the dimension of the world that the speaker cares about communicating. In this model, a listener jointly reasons about the state of the world (e.g. what Jane is like) and a projection function, corresponding to the aspect of the world the speaker cares to communicate (e.g. Jane's determination). This listener assumes an informative speaker - one whose choice of utterance maximizes the probability of communicating the state of the world - but only up to a projection which dictates the relevant dimension of the world.

This can be used to give 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, it is necessary to provide a semantics, specifying the literal meaning of each utterance (for example, that *soldier* literally describes an individual who serves in a military). Previous work has hand-constructed these literal interpretations, restricting the scalability of the models, and their applicability to previously unseen metaphors.

Our contribution We develop a model of pragmatic reasoning which uses empirically learned word-embeddings (4, 5) to represent word meanings, obtaining a system capable of interpreting open-domain predicative and adjectival metaphors without the need for hand-specified semantics. This adaptation requires a generalization of projection functions to linear projections in a vector space, and a novel inference algorithm to calculate metaphor interpretations. Constructing this system permits what is to our knowledge the first open-domain evaluation of a Bayesian model of pragmatic reasoning. Evaluated on human judgments, our model significantly outperforms a baseline which uses a word embedding semantics without explicit pragmatic reasoning. This suggests that the information in word embeddings alone is not sufficient to capture the creativity of metaphorical language, but that an explicit model of pragmatic reasoning is also key.

1. Overview of metaphor

Metaphor exists in many syntactic forms (6), and has been extensively studied in cognitive science (7–9), linguistics (10, 11) and other disciplines (12, 13).

For present purposes, we focus on metaphors involving copular predicates (e.g. *Jane is a soldier*) and AN noun

Significance Statement

Linguistic creativity — the ability to combine existing representations to create new meanings — is a distinctive trait of human cognition. Metaphor provides a general vehicle for creative transfer and reuse of concepts. Here, we develop a system for open-domain interpretation of metaphor. Our system integrates world knowledge automatically induced from large text corpora, with reasoning about the social goals of the speaker. The approach provides a general architecture for composing semantic knowledge with social reasoning, providing insight into the origins of linguistic creativity.

R.C.G. and L.B. designed the model, planned the experiment, performed analyses, and wrote the manuscript. R.C.G. carried out the experiment.

[*]To whom correspondence should be addressed. E-mail: LBergen@ucsd.edu

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 (14) 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 both world knowledge and the semantics of the phrases. For instance, the use of the source *river* in (2) may convey that time flows in a single direction, whereas the same source in (3) may convey that the basement is flooded.

(2) Time is a river.

(3) The basement is a river.

While certain metaphors are conventional - comparing someone to a lion tends to connote bravery - examples like (2) and (3) suggest that the interpretation of a metaphor is dependent on semantic and world knowledge, factors which are naturally incorporated in a pragmatic model.

2. A Bayesian model of metaphor interpretation

The Rational Speech Acts framework (RSA) provides an elegant and practical way of formalizing pragmatic reasoning (3). In this framework, listeners and speakers are represented as conditional probability distributions. Speakers are represented as distributions over possible utterances given worlds, and listeners as distributions over possible worlds given utterances. The most basic version of RSA (3) is incapable of interpreting metaphors, due to the strict assumption that the speaker's utterances are literally true. To address this, Kao et al. (2) propose a model L_1^Q, shown in (6), which in turn is defined in terms of S_1 (5) and L_0 (4).

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

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

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

The literal listener L_0 represents a model of a listener that, given an utterance $u \in U$, updates their belief about the world $w \in W$ by filtering out all worlds that are semantically incompatible with u. The term $[\![\cdot]\!]$ is a function $U \to (W \to \{0,1\})$, representing the semantics of the language. $P_L(w)$ is the prior probability of world w.

Projections Functions $q \in Q$ formalize the notion of picking a particular *aspect* or *dimension* of w. Formally, they are functions $W \to D$, for some set D.

The informative speaker S_1 has a state w they want to communicate to the listener L_0, and prefers utterances u which maximize the probability that L_0 assigns to w, up to the dimension of w specified by q. $\delta_{a=b}$ is an indicator function, and is equal to 1 if $a = b$, and equal to 0 otherwise. If q is the identity function, then $S_1(u|w) \propto L_0(w|u)$, and S_1 is thus a model of a speaker who prefers to choose the most informative utterance available.

The pragmatic listener The full model, L_1^Q, hears an utterance u, and jointly infers values for w and q by reasoning about S_1. 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. $[\![u]\!](w) = 0$); this will occur when u conveys effectively some feature of world w as determined by q. $P_{L_Q}(q)$ is the prior probability of projection q.

L_1^Q functions as a model of metaphor interpretation. For instance, using the metaphor in (7), the listener infers both a state w (representing what John is like) and a feature q (representing which aspects of John are relevant).

As an example in a hand-constructed setting, we could take John to be fully characterized by two features, whether he is vicious and whether he is aquatic, so that a state w is a value (true or false) for both of these predicates. The projections $q \in Q$ are then the functions mapping a state to its value on viciousness ($q_{vicious}$) or aquaticness ($q_{aquatic}$) respectively. Further, we assume that *shark* is semantically compatible only with the state in which John is both vicious and aquatic.

(7) John is a shark.

On hearing (7), the prior belief that John is not literally an aquatic animal leads L_1^Q to conclude that the speaker cares about conveying the viciousness dimension (i.e. has projection $q_{vicious}$), and that John is vicious. See (2) for quantitative examples.

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 the state in which John is quick, aquatic and vicious, and also add a third utterance, *dolphin*, compatible only with John being quick, aquatic and *not* vicious.

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 *dolphin*. The utterance *shark* is therefore more likely to have been produced by the speaker trying to communicate John's viciousness.

L_1^Q can model AN metaphors in a similar way. For a phrase like *John's fiery temper*, the listener infers the features of John's temper that would explain why the speaker described it with *fiery*.

3. Distributional Semantics

Word embeddings, or *distributional semantic models*, provide a representation of word meanings that can be learned from large corpora of language data. In these models, word meanings are mapped to points in a high-dimensional vector space, such that words with similar meanings are mapped to nearby points in the space. The embeddings can be obtained either by dimensionality reduction of a word co-occurrence matrix (5) estimated from a corpus, or by extracting the weights of a statistical model (4, 15, 16) trained on a separate task. In both cases, word embeddings provide a way to empirically obtain fine grained connotations of lexical items (4), and have been used effectively in a number of NLP tasks (17–19).

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 (20) for an overview of proposed systems).

We hypothesize that, while the information in high quality word embeddings captures important aspects of meaning, a cognitively realistic model of metaphor interpretation should also incorporate pragmatic reasoning, of the sort formalized in the RSA framework. We now explain how the L_1^Q model described above can be combined with a distributional model of word meaning.

4. Bayesian pragmatics with distributional semantics

We now introduce a *vector* 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, so that elements $w \in W$ are vectors. A word embedding maps words to vectors ($E : U \to W$). For our application of the model, we assume the set of utterances U is a set of adjectives.

The listener's prior To define a prior distribution P_L over the vector space W, we use a multivariate spherical Gaussian distribution $P_{\mathcal{N}}$, which can be parametrized by a vector μ for the mean and a single scalar σ (the covariance matrix is assumed to be $\sigma^2 I$). We define the prior over projections P_{L_Q} to be uniform (the set of projections is discussed below).

$$(8) \quad P_L(w) = P_{\mathcal{N}}(w|\mu = E(target), \sigma = \sigma_1)$$

We can view the prior 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. 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, the speaker and listener are playing a spatial reference game (21), in an abstract word embedding space. Our vector semantics bears comparison to the *conceptual space* semantics of (22), as well as the proposal for metaphor comprehension of (23).

The prior distribution places more probability mass on points closer 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 which determines the extent of the listener's prior uncertainty.

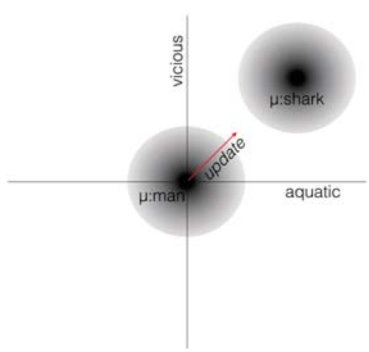

Fig. 1. Illustration of literal listener L_0 given *The man is a shark*, with $\overrightarrow{man}$ = (0,0) and $\overrightarrow{shark}$ = (1,1). L_0's prior is centered at $\overrightarrow{man}$, and is updated towards $\overrightarrow{shark}$.

The semantics Word embedding spaces allow us to compare the similarity of words (e.g., a noun and an adjective) according to different measures of distance in the space. However, they do not provide a means of categorically determining the compatibility of that adjective and noun, as previous pragmatic models have required (described in Section 2). We observe, however, that the definition of L_0 in (4) only mathematically requires 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:

$$(9) \quad [\![u]\!](w) = P_{\mathcal{N}}(w|\mu = E(source), \sigma = \sigma_2)$$

The value of $[\![u]\!](w)$ is a real number which decreases with the Euclidean distance between u and w. The advantage of defining the semantics in this way is that both the prior of L_0, shown in (8), and the likelihood, in (9), are Gaussian distributions, which allows for a closed form solution of L_0, described in *Materials and Methods*.

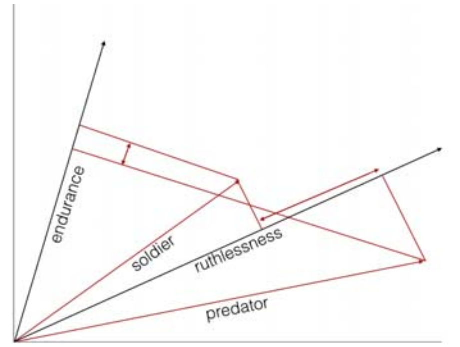

Fig. 2. In this hand-constructed 2D example, vectors for $\overrightarrow{soldier}$ and $\overrightarrow{predator}$ are projected onto subspaces given by $\overrightarrow{endurance}$ and $\overrightarrow{ruthlessness}$. Soldiers have greater endurance than predators, while predators are more ruthless.

Projections Finally, we need to supply a 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 along a vector (or hyperplane) v capturing the degree to which each w extends along v, ignoring orthogonal dimensions. Geometrically, this amounts to dropping a line from an input vector w at a right angle onto v, as depicted in figure 2. These projections exploit the linear structure of the embedding space (5), which is documented more extensively in the setting of word vectors than sentence vectors (15, 16), though see (24, 25) for potential caveats.

In practice, we restrict ourselves to projections along a vector, rather than a larger subspace. To obtain a set Q of projections, we note that since word meanings are vectors in W, any word parametrizes a linear projection q. For instance, we can think of the word *vicious* as defining a *viciousness* projection, which measures how far 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 onto v amounts to asking: to what extent does the noun have property v? Figure 4 provides a visualization of the L_1^Q posterior in a simple two-dimensional case corresponding to the example discussed in section 2.

Interpreting the output of L_1^Q The *Materials and Methods* describes how to calculate the interpretation of a metaphor u given these assumptions. In particular, it shows how to compute $L_1^Q(w, q|u)$, the joint distribution over states and projections after hearing a metaphor u. Unlike points $w \in W$, projections $q \in Q$ are readily interpretable, since they correspond to adjectives, describing the aspect of the metaphorical

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 unemployment
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

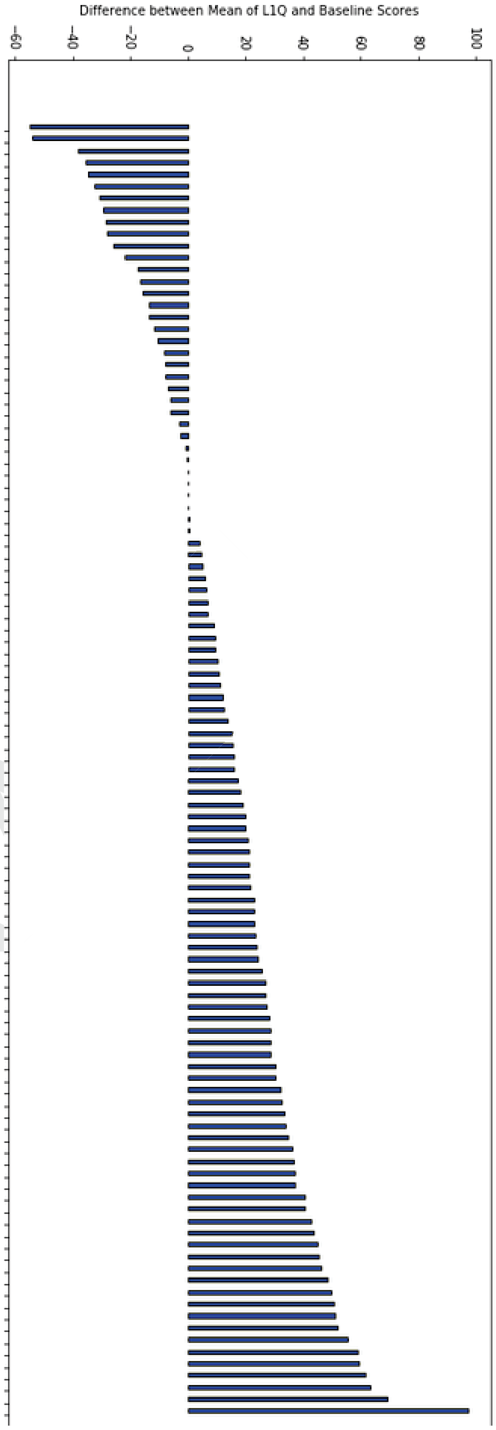

Fig. 3. The 109 metaphors used in the experiment, and baseline and L_1^Q interpretations. Bar positions indicate difference between judgments of L_1^Q and baseline proposals, averaged across participants and across both proposals of each model. Bars right of center indicate a preference for the pragmatic model, showing that for roughly 75% of the metaphors, the L_1^Q interpretation is preferred.

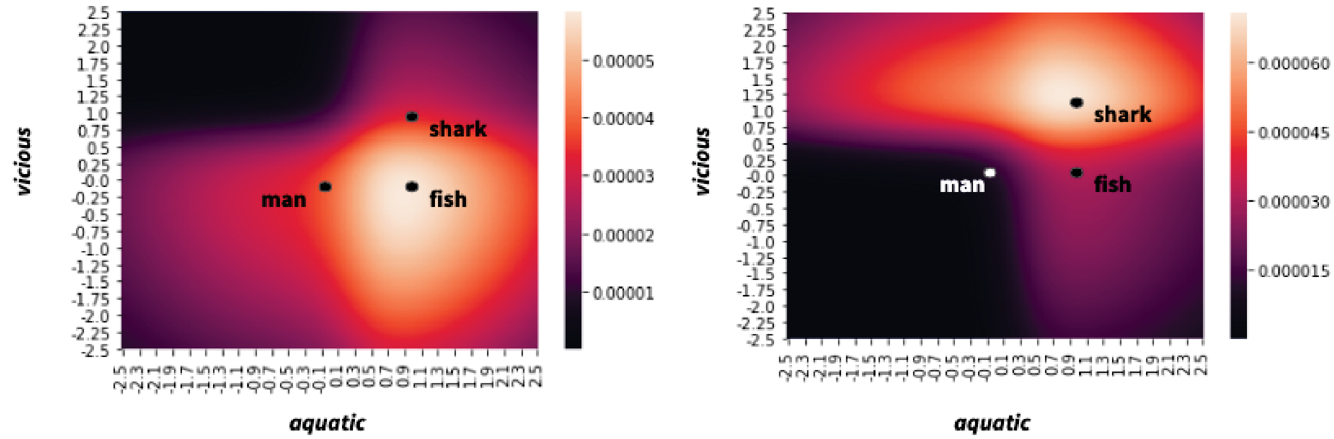

Fig. 4. Heatmaps visualizing the L_1^Q marginal posterior over states in a hand-constructed, two-dimensional case. The listener hears *The man is a fish* in the left panel, and *The man is a shark* in the right. The set of utterances $U = \{man, shark, fish\}$. For simplicity, the set of QUDs Q consists of orthogonal vectors, one along the x-axis (*aquatic*) and one along the y-axis (*vicious*). After hearing *fish*, the listener has lower uncertainty along the aquatic dimension, and higher uncertainty along the vicious dimension (left panel); after hearing *shark*, the situation is reversed (right panel).

adjective or predicate that is inferred to be relevant. For this reason, we use the marginal posterior over Q to generate predictions from the model. The top two L_1^Q marginal posterior projections q for each metaphor, which we use in our experiment, are shown in the leftmost column of Figure 3.

5. Experimental Evaluation

In order to evaluate whether pragmatic reasoning results in metaphor interpretations that better capture human judgments, we designed an experiment comparing L_1^Q interpretations of metaphors to a baseline model which uses word embeddings but no pragmatic reasoning.

Experimental Design. In the experiment, each participant was shown a series of 12 adjectival metaphors, selected randomly from a total of 109. For each metaphor, they were asked to rate four candidate interpretations of the metaphor on a slider bar. These four candidate interpretations consist of the best and second best adjective generated by L_1^Q, and similarly for a baseline model. The baseline model selects adjectives without pragmatic reasoning, using a standard procedure from the word embeddings literature (see *Materials and Methods*). An example is shown in Figure 5.

Analysis. The results, shown in Figure 3, were analyzed using mixed-effects models with random slopes and intercepts for items and participants. Participants rated four interpretations for each metaphor: the best and second-best interpretations, as output by each of the target and baseline models. Participants rated the target interpretations significantly higher than the baseline interpretations (β=13.8, t=5.3, p< 10^{-7}) in a combined analysis. The results were similar when the best target interpretations were compared to the best baseline interpretations (β=16.4, t=4.8, p< 10^{-5}) and when the second-best interpretations were compared (β=11.1, t=3.2, p<0.005).

6. Discussion

We have shown that it is possible to scale Bayesian pragmatic reasoning to distributional semantics, and using this to obtain a model of metaphor interpretation. Our evaluation, the first open-domain evaluation of a Bayesian model of pragmatic language 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.

Materials and Methods

Model inference. We employ a mix of analytic and approximate methods to compute the L_1^Q distribution. We first present the approach for computing L_0 and S_1 posteriors, which can be done analytically, and then present the approximate inference algorithm for L_1^Q. The implementation, written in TensorFlow, will be made publicly available.

L_0 Inference The vector interpretation of L_0 is illustrated in Figure 1, where a ball, corresponding to the prior, is moved in the direction of the point corresponding to the perceived utterance. To calculate L_0 analytically, we make use of Gaussian conjugacy. When the prior P_L is defined as in Equation 8, and the semantic interpretation is defined as in Equation 9, then conjugacy implies that the listener posterior is given by:

$$(10) \quad L_0(w|u) = P_{\mathcal{N}}(w|\mu{=}\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}(\frac{E(target)}{\sigma_1^2}+\frac{E(source)}{\sigma_2^2}), \sigma{=}\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2})$$

S_1 Inference The speaker is defined by Equation 5, which in the continuous case can be rewritten as:

$$(11) \quad S_1(u|w,q) \propto \int_{w'} \delta_{q(w)=q(w')} \cdot L_0(w'|u)$$

Here $q(w)$ is the projection of state w onto the subspace spanned by projection vector q. This integral computes the marginal probability of all states that are projected to the same location as w along q. From Equation 10, $L_0(\cdot|u)$ is a normally distributed random variable, and therefore the projection of this random variable onto a linear subspace is also normally distributed, providing a closed-form solution to S_1.

L_1^Q 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 projections in Q, derived as follows:

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

Here K is a normalizing constant, $w, q \in \mathbb{R}^n$, and w_q is the projection of w onto the vector q. D is the subspace of $\mathbb{R}^n$ spanned by the vector q, and $D^\perp$ is the orthogonal complement of D. The vector $w_\perp$ is the projection of vector w onto the subspace $D^\perp$. The final equation is a one-dimensional integral, and can be easily approximated. We use a Gaussian approximation, which easily generalizes to the setting of multi-dimensional projections. The constant K can be found from the constraint $\sum_q L_1(q|u) = 1$.

Experiment. The aim of our experiment is to determine whether pragmatic reasoning results in better interpretations of metaphors, according to human judgments. We compare against a lesioned model, with a distributional semantics that does not make use of pragmatic reasoning.

Baseline model Our baseline model is defined as follows: for a given metaphor of the form (a n), we take the mean of the adjective word embedding $E(a)$ and the noun word embedding $E(n)$. The two nearest adjectives q to this mean (measured by cosine distance) are the baseline interpretations for the metaphor. Taking the mean of word vectors is a standard technique for computing phrase and sentence meanings from constituent words (19, 26, 27), while cosine distance is commonly used to find words with the most similar meaning (5).

$\mathbf{L_1^Q}$ hyperparameters We use the largest available (300 dimensional) GloVe vectors, as our word embedding E. For each Adjective-Noun 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 (28), 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 metaphors, we choose $\sigma_1 = \sigma_2 = 0.1$; all model parameters and features of the architecture were frozen prior to the experiment. Metaphor interpretations are generated by selecting the two projections with highest marginal posterior mass under L_1^Q. We choose two rather than one since the model tends to distribute most of its probability mass to at least two projections, intuitively reflecting the fact that there is usually more than one good interpretation of a metaphor.

Experimental Methods Tsvetkov et al. (29) provide a corpus of ∼800 AN metaphors, gathered by human annotators, from which we select the least frequent by bigram count (n-gram data from the Corpus of Contemporary American English (30)) to filter out conventionalized metaphors. Our full set of 109 metaphors is shown in figure 3. The data will be made available online. The experiment was run on Mechanical Turk, with 99 native English speakers. Participants who failed to follow instructions on a test item were excluded, leaving 60 participants (analyses remain significant with all participants included). Participants are shown a metaphor, as in figure 5 and asked to judge how relevant each proposed adjective (here, *debilitating*, *pervasive*, *corporate*, *addictive*) is to the metaphorical meaning of the AN phrase. In a test example, they are told to rate *intense* as relevant to *fiery temper* "because a fiery temper is an intense temper" but rate *warm* as irrelevant.

1. Grice HP (1975) Logic and conversation. *1975* pp. 41–58.

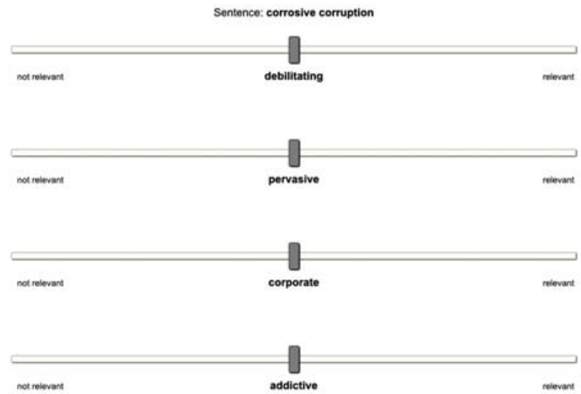

Fig. 5. An item in the experiment. Item order, and order of the 4 candidate adjectives are randomized.

2. Kao JT, Bergen L, Goodman N (2014) Formalizing the pragmatics of metaphor understanding. in *CogSci*.
3. Frank MC, Goodman ND (2012) Predicting pragmatic reasoning in language games. *Science* 336(6084):998–998.
4. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality in *Advances in neural information processing systems*. pp. 3111–3119.
5. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. in *EMNLP*. Vol. 14, pp. 1532–1543.
6. Black M (1954) Metaphor.
7. Tourangeau R, Sternberg RJ (1981) Aptness in metaphor. *Cognitive psychology* 13(1):27–55.
8. Roberts RM, Kreuz RJ (1994) Why do people use figurative language? *Psychological science* 5(3):159–163.
9. Camp E (2006) Metaphor in the mind: The cognition of metaphor 1. *Philosophy Compass* 1(2):154–170.
10. Glucksberg S, Keysar B (1993) How metaphors work.
11. Lakoff G (1993) The contemporary theory of metaphor.
12. Martin JH (1990) *A computational model of metaphor interpretation*. (Academic Press Professional, Inc.).
13. Davidson D (1978) What metaphors mean. *Critical inquiry* 5(1):31–47.
14. Lakoff G, Johnson M (1980) Metaphors we live by. *Chicago, IL: University of*.
15. Peters ME, et al. (2018) Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
16. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
17. Dai AM, Le QV (2015) Semi-supervised sequence learning in *Advances in neural information processing systems*. pp. 3079–3087.
18. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding with unsupervised learning, (Technical report, OpenAI), Technical report.
19. Socher R, et al. (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.
20. Shutova E (2016) Design and evaluation of metaphor processing systems. *Computational Linguistics*.
21. 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*. (Association for Computational Linguistics), pp. 410–419.
22. Gärdenfors P (2004) *Conceptual spaces: The geometry of thought*. (MIT press).
23. Kintsch W (2000) Metaphor comprehension: A computational theory. *Psychonomic bulletin & review* 7(2):257–266.
24. Linzen T (2016) Issues in evaluating semantic spaces using word analogies. *arXiv preprint arXiv:1606.07736*.
25. Finley G, Farmer S, Pakhomov S (2017) What analogies reveal about word vectors and their compositionality in *Proceedings of the 6th joint conference on lexical and computational semantics (* SEM 2017)*. pp. 1–11.
26. Mitchell J, Lapata M (2010) Composition in distributional models of semantics. *Cognitive science* 34(8):1388–1429.
27. Grefenstette E (2013) Category-theoretic quantitative compositional distributional models of natural language semantics. *arXiv preprint arXiv:1311.1539*.
28. Brysbaert M, Warriner AB, Kuperman V (2014) Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods* 46(3):904–911.
29. 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.
30. Davies M (2011) Word frequency data: Corpus of contemporary american english. *Provo, UT: COCA*.