Night Owls and Majestic Whales: Modeling Metaphor Comprehension as a Rational Speech Act over Vector Representations of Lexical Semantics

Anonymous ACL submission

Abstract

While they are some of the few computational models that directly capture pragmatic processes underlying language reasoning, current Rational Speech Act (RSA) models of metaphor are (1) not easily scalable, and (2) do not align well with contemporary accounts of metaphor comprehension. The following research project leverages GloVe word vectors to capture pragmatic language reasoning in metaphoric utterances using an updated RSA framework. This updated framework better aligns model predictions with Relevance Theoretic and Construction Grammatical theories of metaphor semantics. The model yields high posterior probabilities for attributes of metaphors that humans deem relevant in metaphoric utterances over erroneous ones in 89% of all cases, validating the methodology to generate prior probabilities for a RSA framework. When presented with biased priors like listeners are in many naturalistic conversations, the model accurately matches human judgements of the most topical attribute of a topic/target indicated by a metaphoric utterance 90% of the time.

1 Introduction

Metaphor serves as an incredibly poignant communicative device allowing speakers to highlight the specific attributes of a topic of conversation, or a *target*, by means of referencing a seemingly unrelated category or object in the world, or what's called the *source* (Lakoff and Johnson, 1980; Glucksberg et al., 1982). For example, if in the midst of an argument I wanted to pejoratively call my younger brother big, I could call him a "whale". While the statement is not literally true, a listener sufficiently tuned into the context of our argument would be able to infer that the word "whale" is here used to reference my brother's corporeal size.

Because of the complexity underlying metaphor comprehension, computational descriptions of *metaphor comprehension* are uncommon. Work in Natural Language Processing (NLP) is often more concerned with identifying whether or not an utterance contains metaphor at all (Stowe et al., 2019; Stowe and Palmer, 2018; Shaikh et al., 2014; Mosolova et al., 2018; Rai and Chakraverty, 2020; Shutova, 2010), and such models are unperturbed by questions of comprehension more broadly. Additionally, the few computational models of metaphor comprehension that do exist are either only useful in analyzing small, hand-curated datasets (Kao et al., 2014), or ignore the pragmatic processes and factors that are important to real-world metaphor comprehension (Dodge et al., 2015; Huang and Arnold, 2016; Rosen, 2018; Bizzoni and Lappin, 2018; Mohler et al., 2014; Bollegala and Shutova, 2013).

One way to scale computational models of language processes in domains outside of metaphor comprehension has been to leverage what are known as word vectors. Word Vectors have been a staple of NLP applications for some time. Word vectors represent the semantic meaning of a word by projecting words into an N-dimensional word vector space (Mikolov et al., 2013; Pennington et al., 2014; Levy and Goldberg, 2014). These vectors are generated using the correlation of words to their contexts–either a statistical model or an artificial neural network (ANN) is used to predict a word conditioned on its surrounding context, and a portion of the output of that model is then used to represent the meaning of that word as a vector of numbers.

Despite their ubiquity in NLP applications, quantitative models that map word vectors to actual human understanding are rare, making direct application of word vectors to psycholinguistic models onerous. One study conducted by Grand et al. shows that it is possible to arrange word vectors for adjectives into dipole dimensions of meaning, and then leverage these dimensions to extract judgements about adjective associations to nouns that are subsequently projected onto these dimensions. The basic intuition being that if one were to take a word and its antonym, and then two synonyms for each of these two, one could subtract the GloVe word vectors for each set of synonyms from the set of antonyms, average these subtractions, and create a stable dimension of meaning. From there, one can project the word vectors for various nouns onto these dimensions and their relative positions on the dimension of meaning will correlate with human judgements for adjective-noun pairings in the real world. This makes Grand et al.'s dimensions of meaning potentially useful in tasks where one needs to map word vector semantics to human judgements.

The Rational Speech Act (RSA) framework is a well attested framework for modeling pragmatic language comprehension broadly (Goodman and Frank, 2016; Frank and Goodman, 2012). In an RSA model, the process of language reasoning is described in terms of a pragmatic listener who assumes that a speaker will rationally select an utterance that is maximally informative and easy to unpack based on the assumed shared context between the speaker and listener. As an additional source of reasoning, in Question Under Discussion RSA (QUD-RSA) models, the listener also brings to bear their prior knowledge of what are the likely questions that a speaker might be trying to answer with their utterance, based on observations about the state of the world.

Within the QUD-RSA framework, one model of metaphor comprehension already exists. Due to myriad constraints it is difficult to "scale" beyond its single, experimental use-case, however. Furthermore it makes strong assumptions about how speakers and listeners reason about adjectives given an utterance that make it difficult to align with the most contemporary theories of metaphor comprehension. For example, in the model described in Kao et al., the utterance "whale" is associated specifically with three adjectives–"large", "graceful" and "majestic", which are in turn organized into a closed set of worlds containing a combination of 1-3 of these adjectives. Because of this, it is assumed that when reasoning about "whale", that all three of these adjectives need be jointly reasoned about, and results in the model assigning the highest probability to situations in which worlds containing more than 1 of these specific adjectives are almost always more likely. However, work in both Construction Grammatical (CG) descriptions of metaphor (Sullivan, 2009, 2014; Sikos et al., 2008) and Relevance Theoretic (RT) approaches to comprehension (Moreno, 2004; Carston, 2015) provide a slew of evidence that the adjectives invoked by a metaphor are much more variable than the assumptions made in Kao et al. (2014), highly context dependent, and are reasoned about independently from one another. Knowing this, static mappings of adjectives to a metaphoric source like those described in Kao et al. are insufficiently flexible to capture the ways that people reason about metaphoric utterances in daily communication. People do not seem to reason about "worlds" in a way that aligns well with the assumptions underlying the model proposed in (Kao et al., 2014). If people are reasoning about worlds at all, those worlds are certainly not composed of discrete, predetermined sets of adjectives.

So to recap: existing computational models of metaphor comprehension all appear to bite one of the following critiques: they either (1) do not factor in the pragmatic processes underlying metaphor comprehension in the real world, (2) can't be scaled to more than a few examples, or (3) make assumptions about how humans associate relative adjectives and descriptors with metaphor source domains that are not supported by empirical and contemporary theoretic accounts of metaphor comprehension.

While I agree that QUD-RSA models like Kao et al.'s are the best starting point for capturing the core, pragmatic reasoning that underlies human metaphor comprehension, it is imperative to update this existing framework to better match how humans reason about metaphors–with relaxed assumptions about how features are associated with metaphor source domains, and are more broadly generalizable (read: scalable). I believe that it is possible to accomplish this by retooling the framework to *reason about features along dimensions of meaning* like those described in Grand et al., while simultaneously relaxing the model to reason about *what dimensions of meaning are relevant* as the actual QUD. The result is a model that avoids the restrictive constraints that the original (Kao et al., 2014) model requires, and better matches what we know about metaphor comprehension from RT and CG perspectives. It also allows us to leverage NLP based tools like word vectors to scale such a model–

a welcome bonus for myriad reasons.

In section 2, I'll describe the various sources of data used in this study. Then, in section 3, I'll describe the the formal model I've built, and the results of applying the novel model to the original data from Kao et al. (2014). I'll conclude this report with a discussion of the results in section 4 as well as possible future extensions for this work.

2 Data Used

In total, three distinct sources of data were used in this research project. First and foremost, I used the same animal names as those described in Kao et al.. Second, from the same study I used the experimental data collected by the researchers in experimental conditions in experiment 2. In it, participants were presented with a simple, single sentence sentence scenario, followed by a single sentence containing a metaphor in which a fictional protagonist was called one of the animal names from the experimental stimuli. Participants were then asked to provide slider bar values for how much they believed one of six adjectives was being invoked by the metaphor–3 of those words were adjectives found to be associated with the animal name in a previous experiment, and the remaining 3 adjectives were antonyms of the associated ones. The values for the slider bars were then recorded as percentages indicating how relevant participants thought each of the provided adjectives were to the intended meaning of the metaphoric utterance. In some conditions (condition 2, also referred to as the uniform prior condition) no context was provided simulating a uniform prior on adjectives, whilst in others (condition 4, also refferred to as the biased prior or QUD condition) the researchers heavily implied through the scenario that one of the adjectives might be more relevant given the metaphor simulating a biased prior towards the relevance of one of the features. Note: all Person Identifying Information(PII) was scrubbed by the original collectors (Kao et al., 2014) prior to my accessing it.

Finally, not all of the adjectives used in Kao et al. (2014) correlated with one of the described dimensions of meaning in Grand et al. (2018). To augment the number of dimensions, then, I relied on synonyms and antonyms for adjectives scraped from thesaurus.com using a web scraper built in the ScraPy python package. With the web scraped synonyms and antonyms, I augmented the dimensions of meaning described in Grand et al. (2018) with new ones to cover all the adjectives and antonyms described in Kao et al. (2014).

3 The Cognitive Model

The model as described in this paper extends the logic described in Kao et al. (2014) to a distributed model of lexical semantics, and relaxes the restriction on the model from reasoning about discrete worlds containing a finite number of adjectives, to reasoning about dimensions of meaning relevant to an ongoing and potentially dynamic discourse. It does so by leveraging the operation of semantic projection onto dimensions of meaning as described in Grand et al. (2018). The model was implemented in PyTorch, though the GloVe embeddings leveraged were loaded in manually from a pre-trained GloVe repository (Pennington et al., 2014). The full code can be found at https://github.com/zaqari/NAACL2022-RSAMetaphor

To help visualize how this extension works, let's begin with a visualization. Let's pretend that we have projected word vectors for the names of animals onto a set of dimensions of meaning derived using the same methods described in Grand et al. (2018). This operation can be organized to yield a matrix of values where every row is coincides with one of the various animals in our data, and each column coincides with a particular dimension of meaning (i.e. large-small, majestic-inferior, etc.). Let us also assume that we have projected a set of adjectives onto each axis. On any axis, only a subset of all adjectives in our vocabulary will be useful on any axis. For simplicity we'll assume that the adjectives that are useful on an axis are a closed set and are restricted to only the six adjectives used to construct the axis as described in Grand et al. (2018). Our "game" is to get the listener to select the correct adjective from our vocabulary, using an animal name as a stimulus, and their prior knowledge of what dimensions of meaning are at play in a given dialogue. Visually, this is the same as selecting the correct rows from our matrix of adjectives projected onto dimensions of meaning based on the difference between the adjective f and utterance u on that dimension of meaning D, and our prior belief on which dimensions of meaning D are relevant.

The Literal Listener

Formally, the literal listener reasons about the *error* between the value for an adjective f on a dimension of meaning D and an utterance u on the same dimension of meaning. It is assumed that the smaller the error, the more probable is that the adjective f is part of the implied meaning of the utterance u. This process requires us to do the following: (1) Quantify the error between animals and adjectives on an adjective's respective dimension of meaning, and (2) quantify our belief that the distance of the animal to the adjective is significant in some meaningful way. To accomplish (1), we take the squared percent error of the utterance/animal term projected onto a dimension of meaning D_u and an adjective on the same dimensions of meaning D_f, and (2) to quantify our belief that this distance is meaningful we use a half-Gaussian from range $[0, \infty]$, with $\mu = 0$, and a single tune-able hyper-parameter for the scale of the half-Gaussian, σ. The half-Gaussian in this case is useful in that it directly captures the intuition that if the percent error between an animal and an adjective on the adjective's dimension of meaning is zero, then we would have maximum confidence that the animal is a good, easily understandable substitution for that specific adjective. We formalize these operations in equation 1. Let D_f be an the word vector for adjective in question projected on a dimension of meaning D and D_u be the word vector for the animal name/utterance projected onto the same dimension. We use the Dirac-delta function to return either a 1 or a 0 if the adjective f is useful on dimension of meaning D (with f being useful if it was used to construct D per Grand et al. (2018)).

$$L_0(f, D|u) = P_{\mathcal{N}_{[0,\infty]}}\left(\left(\frac{D_u - D_f}{D_f}\right)^2 \middle| 0, \sigma\right)\delta_{f \in D} \quad (1)$$

Note: in the remainder of this paper, I set the value of $\sigma = 1.25$. This value was found using a simplified grid-search algorithm to maximize the posterior probabilities output in section 3.

Utility and the Pragmatic Speaker

As mentioned in the overview the goal of the speaker is to convey an adjective f. The utility of an utterance u in evoking an adjective f on a dimension of meaning D is the negative surprisal that a listener would experience upon hearing u in lieu of f when reasoning about D. In other words, the utility is how expected an utterance u might be in lieu of an adjective f on the dimension of meaning D.

$$U_1(u|f, D) = logL_0(f, D|u) \quad (2)$$

A rational speaker wants to conjure up the correct adjective in the mind of a literal listener. It's assumed then that out of their vocabulary of utterances that they would pick the specific utterance u that has the highest utility in accomplishing this goal. We model the way a speaker would choose an utterance with the highest utility by using a softmax decision rule, which has been shown to describe an approximately rational agent (Sutton and Barto, 1998) in multi-choice tasks with varying rewards per choice.

$$S_1(u|f, D) \propto e^{\lambda U_1(u|f,D)} \quad (3)$$

Note: λ is an optimality parameter that sets the contrast between possible choices of alternative utterances.

The Pragmatic Listener

Now, recall that the goal of a rational speaker is to coerce a listener to select an appropriate adjective from a matrix of possible adjectives. However, the listener has prior knowledge that the topic of conversation is not literally whatever source the utterance refers to. If I refer to my younger brother as a "whale", I do not mean that my younger brother is literally a whale, but I do want the listener to pick some relevant adjective or descriptor associated with whales. The pragmatic listener thus needs to keep track of the following four bits of information to accomplish this–the first two have already been previously discussed at the top of this description.

1. Their belief about what conditions would lead a speaker to select the utterance u that the listener heard.

2. What dimensions of meaning D are relevant/probable during a dialogue.

3. The probability that the topic of conversation is either literally an example of the *source* evoked by u, or some other salient category.

4. The probability that an adjective f is a good descriptor for an entity that belongs to the *source* matching the utterance u *and* the actual category that the topic of conversation belongs to.

To formalize all these points, I need to introduce one final variable–the formal *category* c that can be either the *source* evoked by the utterance (i.e. literally a "whale" in "my younger brother is a whale") or some other category which the topic of conversation (i.e. my younger brother who I called a "whale") actually belongs to. We can thus formalize a pragmatic listener that outputs a posterior probability for adjectives f conditioned on categories c as shown in equation 4.

$$L_1(f, c|u, D) \propto P(c) \sum_D P(D)P(f, D|c)S_1(u|f, D) \tag{4}$$

3.1 Results

I test the model's output on the original human data collected in experiment 2 described in Kao et al. (2014). Specifically, I look to conditions 2 (the uniform prior condition) and 4 (the biased prior/QUD condition) from that experiment, corresponding to the uniform prior condition where none of the adjectives f are rendered more salient than another, and the QUD-biased condition where the top most popular adjective f is rendered more salient in experimental stimuli.

Identically to (Kao et al., 2014) in all instances, the model correctly predicts the correct category c–in zero instances does the model erroneously predict that the topic under discussion is literally an example of an animal as evoked by the utterance. This simple qualitative observation confirms that the model is indeed capable of figurative language reasoning.

As a sanity check to validate the underlying logic of the literal listener and speaker functions, I tested the percentage of instances in which the model yields a higher probability–both in terms of prior probabilities in L_0 and posterior probabilities in L_1–for adjectives attested to be associated with a metaphor source domain, as opposed to their antonyms on a dimension of meaning for which both are relevant. The literal listener yields higher probability for the correctly attributed over the antonym in 89% of all cases. This number is significant–randomly permuting the word vectors used to generate dimensions of meaning and source term locations on those dimensions yields 0 permutations out of 1000 that have higher accuracy ($p < 1e^{-5}$). This holds true as well for posterior probabilities generated by L_1, both uniform and biased prior conditions. Figure 1 shows plots for probabilities assigned to the correct adjective and its antonym for the literal listener, Pragmatic Listener in the uniform prior condition, and Pragmatic listener in the biased prior condition respectively.

For both uniform and biased prior conditions I tested model fit to participant data using the following three tests. (1) The percent time that the model's prediction for *the most probable adjective f matched human judgements for the most relevant* f as identified from a participant's slider responses. (2) The mean error between the rank for the probabilities of each adjective f generated by the model for a given condition, compared to the rank for the slider responses of adjectives f provided by a participant in the same condition. (3) The Pearson Correlation of the probabilities for all adjectives f provided by the model in a given condition and the slider-value probabilities for participants in the same condition.

With a uniform prior belief on dimensions of meaning, the model matches the adjective f that human annotators indicated as being the most relevant adjective 34% of the time. This is low, but not surprising. As noted in (Kao et al., 2014) "The predicted reliability of participants' ratings using the Spearman-Brown prediction formula is 0.828 (95% CI = [0.827, 0.829]), suggesting first that people do not agree perfectly on metaphorical interpretations". This may have been a significant confound to model results in the uniform prior condition–similarly to the results reported in (Kao et al., 2014). I then tested the error between ranks assigned to all adjectives f conditioned on an utterance/animal name u by the model when compared to the ranks assigned to the same f conditioned on u by human participants. The average error between the model rankings and participant rankings is .915 (median: 1.) indicating that on average the rank for the model's predicted values differs from the rank for participants' slider values by 1. Pearson R between adjective probabilities predicted by the model and slider values indicated by participants in the uniform condition indicates no relationship $(r(1175) = -0.03, p = .24)$.

In the biased prior condition the model performs exceptionally well. Following the example set in (Kao et al., 2014), I set the model's prior on the correct dimension of meaning to be higher than all other dimensions of meaning ($P(D_{correct}) = .7$), and allowed other dimensions of meaning to share

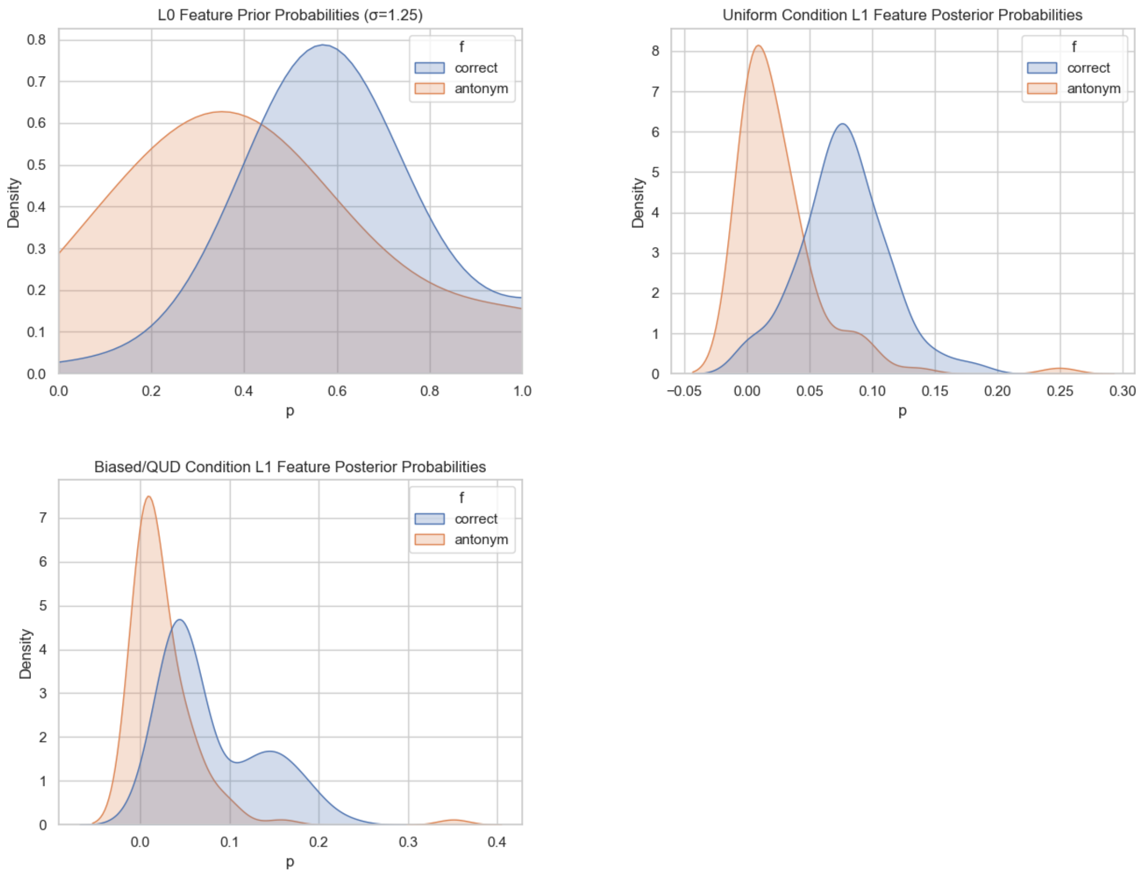


Figure 1: Plots for probabilities assigned to the correct adjective and its antonym for the literal listener, Pragmatic Listener in the uniform prior condition, and Pragmatic listener in the QUD/biased-prior condition respectively. Note that in the QUD condition, that correct adjectives that are not on the biased dimension of meaning have probabilities that are pushed closer to zero (though are still greater than that of the antonyms), whilst correct adjectives that are on the biased dimension of meaning get a boost beyond the max probabilities observed for correct adjectives in the uniform prior condition.

a uniform, non-zero prior probability for the remaining probability mass. The model's prediction for the most probable adjective f matched that of participants 90% of the time. The average error between the ranked probabilities for adjectives f provided by the model compared to the ranked slider values for adjectives f indicated by participants is .62. There is moderate, but statistically significant correlation between the model's posterior probabilities for adjectives f and the slider values indicated by participants ($r(1175) = .45, p < 1e^{-5}$).

LMPLots showing the distribution and slope of outputs from $L_1(f, c|u, D)$ for both uniform and biased prior conditions as well as summary table of results for both conditions is provided in in table 1.

4 Discussion

The results paint an interesting picture of the efficacy of the model. My objective in this section is to break down what the model tells us, as well as point to some potential confounds in the model. I'll conclude this section with a brief discussion of future directions for this line of research.

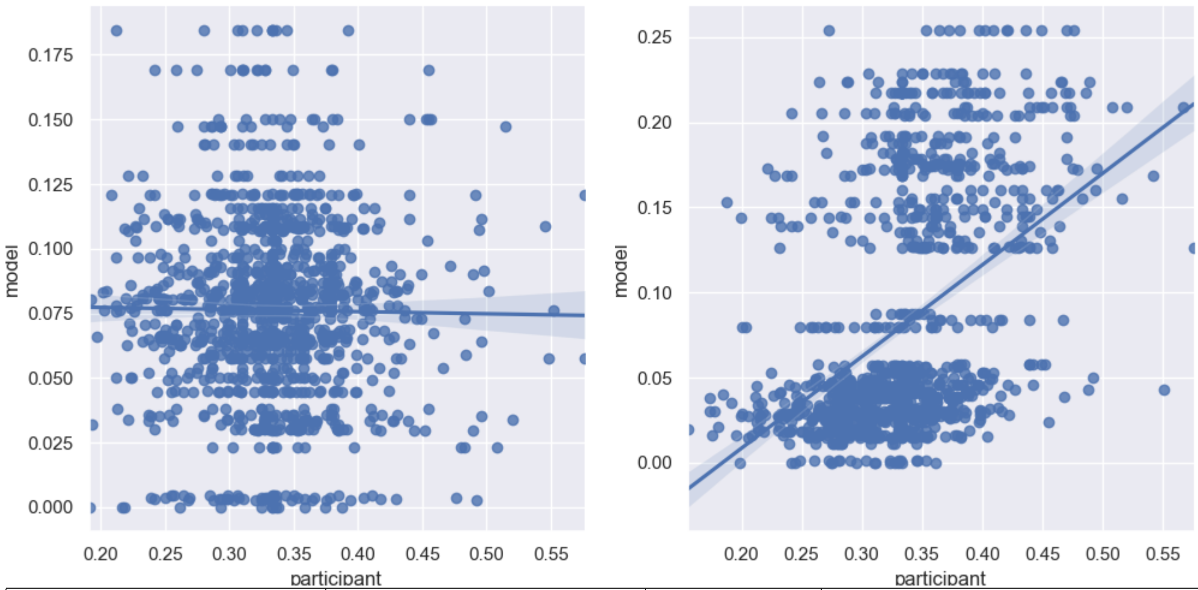
To begin with, the model qualitatively matches human judgements in the QUD condition (i.e. when there was a biased prior on what dimension of meaning was at play) and does so quite well. This is particularly heartening. As previously mentioned, treating metaphor comprehension as reasoning about a static, closed set of worlds doesn't align with current explanations of human metaphoric reasoning (Moreno, 2004; Carston, 2015; Sullivan, 2009, 2014; Sikos et al., 2008). The model I've described still leverages the QUD-RSA framework using utterances and prior beliefs to project onto relevant dimensions of meaning, but by reasoning about those dimensions of meaning directly rather than a static set of worlds it better matches what we know about human behavior in this regard. It does so reliably (based on its correlation and mean rank error) and with excellent accuracy.

The model performed below my expectations in the uniform prior condition however. Again, this isn't entirely shocking. Kao et al. described in their original write-up that there was indeed variation between participants themselves in how they assigned relevance to the adjectives they were presented with in the uniform prior condition, and this lead to confounds with their results as well. Why might this be the case at all? What explains the variation in human responses? Additional empirical research is required to adequately answer these questions.

I believe another potential confound in this case as well is the loose link between GloVe word vectors used and human reasoning. While Grand et al. show that their use of dimensions of meaning is indeed reliable at better matching human judgements of adjective attribution using word vectors, even they note that the correlation between the two is not perfect–correlation varied a lot between various conditions the researchers tested, ranging from .15 to .94. Similar to Kao et al., they also note that there is significant variation in individual responses provided by human participants. In sum, even when using the method for deriving dimensions of meaning described in Grand et al. (2018), mapping of word vectors to human judgement is messy for a multitude of reasons.

Despite the model's poor performance in matching participant's slider values in the uniform prior condition, the model did accurately prescribe higher probability for the correct adjectives as attested in (Kao et al., 2014) over their antonyms, however. Taken on balance, then, despite the fact that the model did not *completely* replicate human judgements, it did replicate human judgements that the correct adjectives were more likely than their antonyms. In a way, the model performs almost like another participant in this regard–its responses are as variable when compared to any study participant as the agreement would be between *any two* participants picked at random.

The model I've proposed is significantly more scalable than the original model proposed in (Kao et al., 2014). By using word vectors to generate prior probabilities for adjectives–however messy the mapping between word vectors and human judgements might be–there is feasibly no upper limit to the application of the model to new sets of source domains, adjectives, or even dimensions of meaning. A researcher need only define what source domains they're interested in studying (which is common already in studies of metaphor in humans), as well as a number of dimensions of meaning. Dimensions of meaning can be generated quickly either by hand or by using a simple web scraper to generate sets of adjectives that can be used to construct them. In fact, one could even extend this model to other languages–as long as you can generate a word vector model for that language, you have all that you need to leverage this model.



Condition	Prediction Accuracy	Rank Error	R
Condition 2 (uniform prior on adjectives)	.34	.91	$r(1175) = -0.03, p = .24$
Condition 4 (biased prior on adjectives)	.90	.62	$r(1175) = .45, p < 1e^{-5}$

Table 1: From left to right: LM-Plots for the correlation of posterior probabilities provided by $L_0(f, c|u, D)$ for the uniform and biased conditions respectively. X-axis value correspond to human slider values and Y-axis values correspond with model posterior probabilities. On bottom: Summary table of relevant statistics for evaluating model performance in the two conditions discussed.

To recap, the model I've described (1) captures the pragmatic roots of metaphor comprehension, (2) can be easily scaled to look at much broader sets of source domains, and (3) does not make the same hard assumptions about how one reasons about "worlds" as previous RSA models have (and thus aligns with what we know about human metaphor reasoning better).

At the same time, I genuinely believe that more work can be done to extend this model's utility. To start, in what ways could we make the model more context savvy? Using GloVe word vectors and dimension of meaning may capture some useful information about human judgements–the current model appears to demonstrate such. However, is it possible to retool how prior probabilities on adjectives (given an utterance) are generated using more contemporary, transformer models of lexical semantics? Models like BERT and GPT-3 both capture an exquisite amount of detail already about context (Devlin et al., 2019; Brown et al., 2020). Finding a means of leveraging these models to generate prior probabilities would decrease the need to worry about the prior on dimensions of meaning by already representing that information to some degree in the word vectors themselves.

While I focus on animal terms in this study, the model described can be efficiently applied to myriad other source domains. My decision to use the source domains I did was solely based on the availability of data and the need to validate that my model usefully extends Kao et al.'s existing QUD-RSA model. But extending this model further to look at non-animal metaphors in other social scenarios would be fascinating. As an example, it would be interesting to apply this model to metaphors surrounding the gun control or immigration debate in US politics as a means of capturing the subtle implicatures in political metaphor usage.

Plato once stated that "the greatest thing by far is to have command of metaphor. This alone cannot be imparted by another." But if the current research has shown anything, it is that it is not enough to have an "eye for resemblances" as Plato put it, but that part of the magic of a good metaphor is in the way that context mixes with those resemblances to render metaphor comprehensible and relevant.

References

Yuri Bizzoni and Shalom Lappin. 2018. Predicting Human Metaphor Paraphrase Judgments with Deep Neural Networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Danushka Bollegala and Ekaterina Shutova. 2013. Metaphor Interpretation Using Paraphrases Extracted from the Web. *PLoS ONE*, 8(9):e74304.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. ArXiv: 2005.14165.

Robyn Carston. 2015. Lexical pragmatics, ad hoc concepts and metaphor: A Relevance Theory perspective. *Ital. J. Linguist*, 22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. MetaNet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado. Association for Computational Linguistics.

Michael C. Frank and Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998.

Sam Glucksberg, Patricia Gildea, and Howard B. Bookin. 1982. On understanding nonliteral speech: Can people ignore metaphors? *Journal of Verbal Learning and Verbal Behavior*, 21(1):85–98.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2018. Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings. *arXiv:1802.01241 [cs]*. ArXiv: 1802.01241.

Yi Ting Huang and Alison R. Arnold. 2016. Word learning in linguistic context: Processing and memory effects. *Cognition*, 156:71–87.

J. T. Kao, Leon Bergen, and Noah D Goodman. 2014. Formalizing the Pragmatics of Metaphor Understanding. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago.

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Arxiv*.

Michael Mohler, Bryan Rink, David Bracewell, and Marc Tomlinson. 2014. A Novel Distributional Approach to Multilingual Conceptual Metaphor Recognition. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1752–1763, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Rosa Moreno. 2004. Metaphor interpretation and emergence. *UCL Working Papers in Linguistics*.

Anna Mosolova, Ivan Bondarenko, and Vadim Fomin. 2018. Conditional Random Fields for Metaphor Detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 121–123, New Orleans, Louisiana. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Sunny Rai and Shampa Chakraverty. 2020. A Survey on Computational Metaphor Processing. *ACM Computing Surveys*, 53(2):1–37.

Zachary Rosen. 2018. Computationally Constructed Concepts: A Machine Learning Approach to Metaphor Interpretation Using Usage-Based Construction Grammatical Cues. In *Proceedings of the Workshop on Figurative Language Processing*, pages 102–109, New Orleans, Louisiana. Association for Computational Linguistics.

Samira Shaikh, Tomek Strzalkowski, Kit Cho, Ting Liu, George Aaron Broadwell, Laurie Feldman, Sarah Taylor, Boris Yamrom, Ching-Sheng Lin, Ning Sa, Ignacio Cases, Yuliya Peshkova, and Kyle Elliot. 2014. Discovering Conceptual Metaphors using Source Domain Spaces. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 210–220, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Ekaterina Shutova. 2010. Models of metaphor in NLP. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 688–697, Uppsala, Sweden. Association for Computational Linguistics.

Les Sikos, Susan Windisch Brown, Albert E Kim, Laura A Michaelis, and Martha Palmer. 2008. Figurative Language: "Meaning" is Often More than Just a Sum of the Parts. page 6.

Kevin Stowe, Sarah Moeller, Laura Michaelis, and Martha Palmer. 2019. Linguistic Analysis Improves Neural Metaphor Detection. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 362–371, Hong Kong, China. Association for Computational Linguistics.

Kevin Stowe and Martha Palmer. 2018. Leveraging Syntactic Constructions for Metaphor Identification. In *Proceedings of the Workshop on Figurative Language Processing*, pages 17–26, New Orleans, Louisiana. Association for Computational Linguistics.

Karen Sullivan. 2009. Grammatical Constructions in Metaphoric Language. *Studies in Cognitive Corpus Linguistics*.

Karen Sullivan. 2014. *Frames and Constructions in Metaphoric Language*, volume 14.

Richard S Sutton and Andrew G Barto. 1998. *Reinforcement Learning: An Introduction*, 1 edition, volume 1. Cambridge University Press.