PROBABILISTIC MODELS OF PRAGMATICS FOR NATURAL LANGUAGE

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF LINGUISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Reuben Cohn-Gordon
May 2020

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Christopher Potts) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Daniel Lassiter)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Leon Bergen)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Daniel Jurafsky)

Approved for the Stanford University Committee on Graduate Studies

Preface

Grice (1975) puts forward a view of linguistic meaning in which conversational agents enrich the semantic interpretation of linguistic expressions by recourse to *pragmatic* reasoning about their interlocutors and world knowledge. As a simple example, on hearing my friend tell me that she read some of *War and Peace*, I reason that, had she read all of it, she would have said as much, and accordingly that she read only part.

It turns out that this perspective is well suited to a probabilistic formalization. In these terms, linguistic meaning is fully characterized by a joint probability distribution $P(W, U)$ between states of the world W and linguistic expressions U. The Gricean perspective described above corresponds to a factoring of this enormously complex distribution into a semantics $[\![u]\!](w) : U \to (W \to \{0, 1\})$, world knowledge $P(W)$ and a pair of agents which reason about each other on the assumption that both are cooperative and have access to a commonly known semantics.

This third component, of back and forth reasoning between agents, originates in work in game-theory (Franke, 2009; Lewis, 1969) and has been formalized in probabilistic terms by a class of models often collectively referred to as the *Rational Speech Acts* (RSA) framework (Frank and Goodman, 2012). By allowing for the construction of models which explain in precise terms how Gricean pressures like informativity and relevance interact with a semantics, this framework allows us to take an intuitive theory and explore its predictions beyond the limits of intuition.

But it should be more than a theoretical tool. To the extent that its characterization of meaning is correct, it should allow for the construction of computational systems capable of reproducing the dynamics of open-domain natural language. For instance, on the assumption that humans produce language pragmatically, one would expect systems which generate natural language to most faithfully reproduce human behavior when aiming to be not only truthful, but also informative to a hypothetical interlocutor. Likewise, systems which interpret language in a human-like way should perform best when they model language as being generated by an informative speaker.

Despite this, standard approaches to many natural language processing (NLP) tasks, like image captioning (Farhadi et al., 2010; Vinyals et al., 2015), translation (Brown et al., 1990; Bahdanau et al., 2014) and metaphor interpretation (Shutova et al., 2013), only incorporate pragmatic reasoning implicitly (in the sense that a supervised model trained on human data may learn to replicate pragmatic behavior).

The approach of this dissertation is to take models which capture dynamics of pragmatic language use and apply them to open-domain settings. In this respect, my work builds on research in this vein for referential expression generation (Monroe and Potts, 2015; Andreas and Klein, 2016a), image captioning (Vedantam et al., 2017) and instruction following (Fried et al., 2017), as well as work using neural networks as generative models in Bayesian cognitive architectures (Wu et al., 2015; Liu et al., 2018).

The content of the dissertation divides into two parts. The first (chapter 2) focuses on the *interpretation* of language (particularly non-literal language) using a model of non-literal language previously applied to hyperbole and metaphor interpretation in a setting with a hand-specified and idealized semantics. Here, the goal is to instantiate the same model, but with a semantics derived from a vector space model of word meaning. In this setting, the model remains unchanged, but states are points in an abstract *word embedding* space - a central computational linguistic representation of meaning (Mikolov et al., 2013; Pennington et al., 2014). The core idea here is that points in the space can be viewed as a continuous analogue of possible worlds, and that linear projections of a vector space are a natural way to represent the *aspect* of the world that is relevant in a conversation.

The second part of the dissertation (chapters 3 and 4) focuses on the *production* of language, in settings where the length of utterances (and consequently the set of all possible utterances) is unbounded. The core idea here is that pragmatic reasoning can take place *incrementally*, that is, midway through the saying or hearing of an utterance. This incremental approach is applied to neural language generation tasks, producing informative image captions and translations.

The result of these investigations is far from a complete picture, but nevertheless a substantial step towards Bayesian models of semantics and pragmatics which can handle the full richness of natural language, and by doing so provide both explanatory models of meaning and computational systems for producing and interpreting language.

Acknowledgments

Thanks to my brother, Kati. You are patient, kind, good-natured and always, *always* helpful. Every few weeks of my PhD, I messaged you with computer problems - sometimes very fiddly ones - and I don't think you ever got annoyed. I am very grateful. Thanks to my parents. To my mother Avra, who taught me everything (really), and my father Mike who always bought me books (not least the first book about linguistics I ever read), and whom I miss very much.

Most of the best, cleverest, funniest, and just downright brilliant people I ever met, I met at Stanford. Thanks to Shalini, the kindest friend and best content provider a person could have, to Brooke, who is endlessly smart, fun and understanding, to Mae, who is so insightful and compassionate, to Daisy, who has the best taste in everything, to Sebastian, who was always generous in helping me throughout my PhD (and baked me a delicious babka for my defense), to Poorvi, who always let me crash on her couch when I was too tired to walk home after doing problem sets, to CJ, who is consistently the best of company, to Chiara, who gives me excellent things to read, to Claire, who is thoughtful, perceptive, and so kind, and to Jeff, with whom I have spent many happy hours in faux British pubs and who did not complain when I forgot to return his books (I still have them),

Thanks also to Branden, who is Australian and likes Pavement, Brandon, who studies logic and likes Pavement, and Brendan, who is Australian and studies logic[1]. I would also like to extend my gratitude to all the other Brandens, Brendons, Brendans, Brendens and Brandons I have had the pleasure of meeting.

Thanks to Hiroto Udagawa[2], for both meaningful and meaningless conversation, sometimes at the same time, to Tarun, for hot takes on long books, to Josiah, for his perceptiveness, to Bitiya, for her unique sense of chaos (and lending me Everything is Illuminated, which I never returned), Som-Mai for lovely conversations over many years, to Amel, for great chats over many lunches, to Gabi, for wry pronouncements over many burritos, to Juliana, whose copy of Barthes I have still not returned, to Kate, for many fun escapades, as well as to Ed, Emily, Prerna, Lelia, Jenn, Matthias, Tyler, Jon, Helena, MH, Margarita, Shreya, Jason, Heeral, and Maria.

[1] I have refrained from including the corresponding Venn diagram.

[2] `https://www.linkedin.com/in/hiroto-udagawa/`

Thanks to Mansi, for poorly executed bad ideas, Alex, for well executed bad ideas, Pearly, for the hijinks, Nam, for the shenanigans, Zack, for the capers, and Sophia, for the adventures.

As for my academic life, thanks to my advisor Chris Potts, who is unusually tall. But who in addition to that sets the bar for professionalism and hard work, in absolutely everything that he does. Thanks also to Dan Lassiter, for his unfailingly insightful discussion, and to Tobias Gerstenberg and Dan Jurafsky. To Leon Bergen, I would like to extend a degree of gratitude commensurate with his behavior. Thanks also to Noah Goodman, for his endless creativity and provision of the intellectual landscape my thesis is based on, to Roger Levy, who is the kindest PI you could imagine, and taught me how to run experiments, to Beth Levin for her guidance and unfailing generosity with her time, and also to Ted Gibson, Tim O'Donnell, Judith Degen, Paul Kiparsky, and Cleo Condoravdi. Finally, thank you to the staff of the linguistics department, who have to do so much, and do it so well.

Contents

List of Tables

List of Figures

Chapter 1

Bayesian Models of Semantics and Pragmatics

> "...Now if you're ready, Oysters dear,
> We can begin to feed."
> "But not on us!" the Oysters cried,
> Turning a little blue.
> "After such kindness, that would be
> A dismal thing to do!"
> "The night is fine," the Walrus said.
> "Do you admire the view?"
>
> Lewis Carroll, *Through the Looking-Glass and What Alice Found There*

The meaning of an utterance is the information it communicates about the state of the world[1]. Some of this information is closely linked to the form of the utterance itself. *The night is fine*, for instance, communicates something about the weather. Other information, however, depends more significantly on the context of the utterance. In the context in which the Walrus comments on the night weather, in response to the Oysters' question about their fate, the fact that he eschews any denial carries a strong suggestion (both to the reader and presumably the terrified Oysters) that he is planning to eat them[2].

This information about the Walrus' intention is clearly not "built in" to the Walrus' utterance. For example, saying *The night is fine* as a response to *What would you like to eat?* at a cocktail party would communicate

[1] *Meaning* is of course a very broad term — this is the definition that best suits the content of this dissertation.

[2] Thanks to Brooke Husic for the example.

nothing about wanting to eat oysters. Rather, this particular meaning arises as a result both of the utterance, and the context in which it is said.

The study of natural language meaning accordingly divides into semantics - what is built into a linguistic expression - and pragmatics - what is communicated by the use of that expression in a particular context. While tools originating in the study of programming languages and logics have been fruitfully applied in the study of the semantics of natural language (Montague, 1973), pragmatic meaning has less clear counterparts in mathematical logic and computer science[3] and so poses a distinct challenge.

A core approach to understanding pragmatic meaning originates with Grice (1975), who proposes that the assumption of cooperativity between conversational participants leads to inferences being drawn when this assumption appears to be flouted. In the case of the Walrus, for instance, the flouting of the norm that questions (*Do you plan to eat us?*) should be answered informatively and relevantly leads the hearer (or reader) to find an explanation for the flouting, here that the true answer is one which the speaker prefers not to give (*Yes*).

In recent years, an approach to pragmatic meaning which builds on Grice's perspective, as well as the insight of (Lewis, 1969) that conversation can be viewed as a cooperative game, has emerged within the setting of Bayesian cognitive science (Tenenbaum et al., 2011). This approach, which is often referred to as the *Rational Speech Acts framework* (RSA), or as *probabilistic pragmatics*, derives a speaker's linguistic choices and a listener's inferences about these choices as the result of agents reasoning about their interlocutor, on the basis of a shared semantics and an assumption of cooperativity.

This dissertation is an attempt to show how this probabilistic perspective on pragmatics can be integrated with advances in natural language processing, where substantial progress is being made in tasks involving linguistic meaning. The thesis motivating this work is that probabilistic models of pragmatics are ideally positioned to bridge the gap between an appropriately general formal theory of meaning and the practical concerns of natural language processing.

Bridging this gap is a two way street. One the one hand, integrating an explicit model of pragmatic behavior into an NLP system yields an interpretable model in which semantic and pragmatic meaning can be properly isolated. On the other hand, and as will be seen throughout this dissertation, the requirements of real-world systems (unlimited potential utterances, continuous semantics) require conceptual advances leading to a better understanding of the models themselves.

The goal of this chapter is to describe the probabilistic perspective on pragmatic meaning, taking the very idealized setting of a *reference game* as a motivating example. This will serve as a basis for the main content of the dissertation, of scaling these models to less idealized settings.

[3] That said, concerns very similar to what linguists conceive of as pragmatics arise in problems of coordinated action in systems with distributed control. See for example (D'Andrea and Dullerud, 2003)

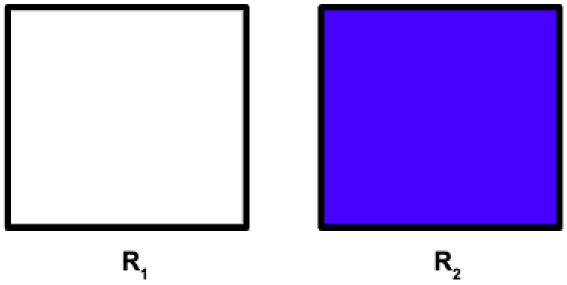


Figure 1.1: $U = \{$*square*, *blue square*$\}$, $W = \{R_1, R_2\}$

1.1 Reference Games

An idealized instance of communication between two agents known as a *reference game* serves as a useful example of the distinction between semantics and pragmatics. I begin by introducing a simple reference game, a model which qualitatively captures the reasoning humans employ when engaging in such a game, and then show how this example can be generalized to much richer instances of linguistic communication.

In a reference game, a speaker and listener see a set W of states[4]. The speaker is assigned one of these states as their target, and aims to communicate which state this is to the listener. The speaker does so by choosing an utterance from a set U of possible utterances. Figure 1.1 provides a concrete example.

Assuming that both the speaker and listener share a semantics, so that *blue square* can only refer to R_2 while *square* can refer to either, the most informative utterance for a speaker whose target is R_2 is *blue square*.

A listener who assumes that the speaker acts informatively in this way can draw an inference on hearing *square*. They can infer that R_1 is the target state, since had R_2 been the state, the speaker would have been more likely to say the more informative utterance *blue square*. We say that the use of the expression *square* carries the *implicature* that R_1 is the target state.

We now consider how to formally model the reasoning process just described of the speaker and listener. First of all, it is clear that an adherence to truth is not a sufficient requirement to derive either the behavior of the informative speaker described above or of the listener who reasons about this speaker. A speaker who only cares about producing true utterances will be equally inclined to say *blue square* or *square* when referring to R_2, since both are true. Likewise, a listener who only attends to semantics will be agnostic as to whether the target state is R_1 or R_2 on hearing *square*, since that utterance is compatible with either.

As such, what is needed is a model which formalizes the process of reasoning about one's interlocutor. The idea to use the tools of game theory in this endeavor originates with Lewis (1969) and is developed by Franke

[4] Depending on the context, these are alternatively referred to as referents, or possible worlds. I use the term *state* to emphasize the generality of the models presented here, but make use of these other terms when appropriate.

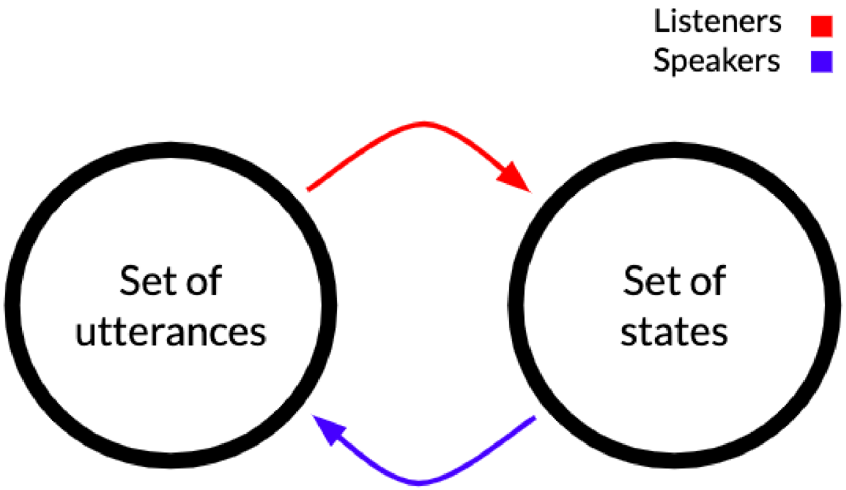


Figure 1.2: An overview of the RSA framework. Speakers and listeners are conditional probability distributions between a set of utterance and a set of states.

(2009) among others. The family of models used in what follows was introduced by Frank and Goodman (2012), and is often referred to as the *Rational Speech Acts* (RSA) framework.

1.1.1 A probabilistic perspective on reference games

RSA models are probabilistic, and view speakers and listeners as conditional probability distributions[5]. Speakers are of the form $P(u|w)$, distributions over which utterance u to say given a target state w, while listeners are of the form $P(w|u)$, distributions over the state w given a heard utterance u. $P_S(u)$ and $P_L(w)$ represent prior distributions over utterances and states respectively. This is summarized in figure 1.2.

First consider a model of a listener L_0 who only reasons about a semantics. For our purposes, a *truth-valued* semantics amounts to a function mapping linguistic expressions $u \in U$ to the set of states (or more generally states) W (often referred to in the linguistics literature as worlds, a term inherited from modal logic) with which u is compatible. This function is here represented in the form $[\![u]\!](w) : U \rightarrow W \rightarrow \{0,1\}$, i.e. as a relation between pairs of utterances and worlds[6].

[5] A (discrete) distribution $P(A)$ over a set A is the pair (A, f), where f is a function $A \rightarrow \mathcal{R}$, assigning each element of A a real-valued weight between 0 and 1, such that $\sum_{a \in A} f(a) = 1$. A conditional distribution $P(A|B)$ is a function $B \rightarrow Dist(A)$, where $Dist(A)$ is the set of all possible distributions on A. In other words, a conditional distribution takes (i.e. is conditioned on) $b \in B$ and returns a distribution over A.

[6] Note that $[\![\cdot]\!]$ is *curried*: it first takes an utterance, and then returns a function which takes a state and returns a truth-value (here 0 or 1, rather than F or T). This formulation aligns a truth-valued semantics with probabilistic generalizations, since a conditional probability distribution $P(U|W)$ can be seen as taking an utterance u and returning a distribution over W, which itself amounts to a function from states to real numbers between 0 and 1

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

This model can be understood as having a prior belief $P_L(w)$ about the target state and, on hearing an utterance u, updating this belief to exclude states incompatible with the semantics of u. Thus, on hearing *blue square*, L_0 is certain that the state is R_2 (since $L_0(R_2|\textit{blue square}) = 1.0$), but on hearing *square*, does not place any more weight on R_1 than it already had in its prior: $L_0(R_1|\textit{square}) = P_L(R_1)$ and $L_0(R_2|\textit{square}) = P_L(R_2)$.

The goal is to design a model L_1 which *does* prefer R_1 on hearing *square*. To this end, we first define a speaker S_1 which, given a state w, prefers utterances u which convey w to L_0, so that $S_1(\textit{blue square}|R_2) > S_1(\textit{square}|R_2)$ when there is equal prior preference for either utterance (i.e. when $P_S(\textit{square}) = P_S(\textit{blue square})$. Conceptually, S_1 is an agent who chooses among a set of actions (U) given a utility function[7] $T(u)$.

(2) $T(u) = \log(L_0(w|u))$

(3) $S_1(u|w) = \frac{e^{T(u))} \cdot P_S(u)}{\sum_{u' \in U} e^{T(u))} \cdot P_S(u')}$

Here, the equation simplifies to $S_1(u|w) = \frac{L_0(w|u) \cdot P_S(u)}{\sum_{u' \in U} L_0(w|u') \cdot P_S(u')}$, but expressing the utility explicitly is often conceptually useful, particularly in more complex models.

This puts us in a position to define L_1, capable of deriving the desired implicature by reasoning about what state S_1 must have had in order to have produced the heard utterance:

(4) $L_1(w|u) = \frac{S_1(u|w) \cdot P_L(w)}{\sum_{w' \in W} S_1(u|w') \cdot P_L(w')}$

Note that RSA equations will often be presented only up to proportionally, since their normalizing term can be recovered from context. For example, 4 would become $L_1(w|u) \propto S_1(u|w) \cdot P_L(w)$.

It will be useful to clearly distinguish the equations (1-4) from their *interpretation*[8] in a concrete setting, which provides particular sets U and W, a particular function $[\![\cdot]\!]$ of type $(U, W) \rightarrow \{0, 1\}$, and particular distributions $P_L(w)$ and $P_S(u)$.

For instance, one interpretation that models the reference game depicted in figure 1.1 is as follows:

- $W : \{R_1, R_2\}$
- $U : \{\textit{square}, \textit{blue square}\}$

[7] Note that here and throughout the dissertation, log is taken to be the natural logarithm.

[8] I use the word *interpretation* by analogy to the situation in mathematical logic, where a distinction is made between a formal language and its interpretation, which (typically) provides concrete sets corresponding the various symbols in the language. Similarly, an interpretation of the RSA equations provides sets and distributions corresponding to the relevant symbols, such as P_L and $[\![\cdot]\!]$.

- $P(w)$: $\{R_1 : 0.5, R_2 : 0.5\}$
- $P(u)$: $\{$*square* : 0.5, *blue square* : 0.5$\}$

The obvious semantics is then:

$$[\![u]\!](w) \mapsto \begin{cases} 0 & u = \textit{blue square}, w = R_1 \\ 1 & \textit{otherwise} \end{cases}$$

Under this interpretation, S_1 prefers the more informative utterance *blue square* over *square* when its target referent is R_2. As a consequence, L_1 assigns more probability to R_1 being the target state on hearing *square*, although R_2 is still a possibility: $L_1(R_1|\textit{square}) > L_1(R_2|\textit{square})$. In the RSA framework, this corresponds to the calculation of an implicature.

In the work discussed in this dissertation, we will see very different examples of interpretations, involving, among other things, infinite sets of utterances, and continuous state spaces. While these interpretations allow for the application of these models of reference games to complex domains, the core dynamics of the models often arise in simple interpretations, such as the above. For a striking example, see figure 2.4, which visualizes two implicatures in a continuous setting.

Implementation Probabilistic programming languages like Church (Goodman et al., 2012) or WebPPL (Goodman and Stuhlmüller, 2014) provide an easy way to prototype models of nested inference. RSA models can also be implemented in standard languages, which allows for more straightforward integration with other NLP tools; most of the models considered in the dissertation are implemented in Python's numerical computation library NumPy, or Tensorflow.

1.1.2 Variations of the basic model

The L_1 presented above is, more or less, the simplest possible model capable of the pragmatic reasoning involved in the above example of a reference game. However, it is worth briefly discussing several variations relevant to the work presented in the following chapters.

Higher recursion depth In principle, it is straightforward to define a speaker who reasons about L_1, a listener that reasons about that speaker, and so on. For more complex phenomena than the implicature discussed above, such higher order agents may be necessary (see (Bergen et al., 2016)).

Taking this further, at infinite depth of nesting we obtain a fixed point, in the form of a speaker and listener who make hard (non-probabilistic) choices of utterance and state respectively, at least in the case of the simplest RSA models.

If we view this fixed point as optimal behavior in some sense, it is natural to view models like L_1 as approximations which reflect how real language users behave, who only reason up to a small depth of nesting (Goodman and Stuhlmüller, 2013).

A literal speaker L_1 grounds out in a literal listener L_0. However, a comparable model could be defined in terms of a literal *speaker* S_0, in terms of which L_0 is defined, and so on. In certain cases, it is useful to use a model of this variety, particularly when integrating Bayesian pragmatics with a neural semantics which takes the form of a conditional distribution over utterances given states (see chapter 4).

Utterance cost Another term often included in RSA models is utterance cost, used to represent preferences or dispreferences for utterances, deriving among other factors from psycholinguistic or phonetic production difficulty.

Cost can be represented by a function $C(u)$ added (in log space) to the speaker's utility, and acts as an alternative to having a speaker prior:

(5) $S_1(u|w) = \frac{e^{ln(L_0(w|u))-C(u)}}{\sum_{u' \in U} e^{ln(L_0(w|u))-C(u)}}$

Note that the presence of the exponential derives from the formulation of agents in Bayesian decision theory as using the softmax (or Boltzmann) distribution over actions.

In the models I consider, no use of cost is made, since any effect of cost can be obtained through a change in the speaker's prior over utterances. For instance, rather than saying that the utterance *blue square* is costlier than *square*, on account of its length, we can assign lower probability to *blue square* than *square* under P_S.

Rationality The degree to which the speaker S_1 cares about informativity can be modulated by a parameter α (by default set to 1), included as follows[9]:

(6) $S_1(u|w) = \frac{e^{\alpha \cdot (ln(L_0(w|u))-C(u))}}{\sum_{u' \in U} e^{\alpha \cdot (ln(L_0(w|u))-C(u))}}$

As α tends to infinity, the speaker's decision becomes increasingly categorical, putting all mass on exactly one utterance, and the listener reasoning about such a speaker makes accordingly categorical inferences.

[9] In the interpretation of the softmax distribution in statistical mechanics, this parameter is designated $-\beta$ and referred to as the inverse temperature.

Speaker knowledge In the current model, speaker agents S_n have complete knowledge of the state of the world w. A relaxation of this assumption is that the speaker has uncertainty, represented as a distribution over states, much like the listener. In this setting, the speaker's aim might be to minimize the distance between their distribution over states and the listener's posterior distribution, as measured, for example by Kullback-Leibler divergence. However, the right utility function may also be more complex; if a cooperative speaker is very uncertain but suspects that their interlocutor is not, the speaker's goal is probably not to reduce the listener to the speaker's state of uncertainty. Instead, the speaker may aim to minimize the listener's expected entropy. As such, the proper way to extend RSA models to situations involving incomplete knowledge on the part of the speaker is an open question.

1.2 Reference games as models of natural language

The simple reference game described in section (1.1.1) serves as a way of thinking about language more broadly. In communication between two agents[10], both share some set of hypotheses about how the world might be (a common ground) and undertake speech acts by uttering linguistic expressions which inform their interlocutor about this world.[11] Pragmatic reasoning takes place in light of a semantics which is also commonly known to both participants.

One simple example relates to the existential quantifier *some*. In natural language, sentences like (7) often seem to carry the additional meaning of *not all*, e.g. that John ate some but not all of the cookies. We refer to this as an *implicature* of (7).

(7) John ate some of the cookies.

To account for this, we could either posit an operation in the grammar which changes the semantic form of (7), such as in (Chierchia et al., 2008), or alternatively posit a Gricean explanation (Grice, 1975), as in (8):

(8) If the speaker meant to convey a world state in which John ate all of the cookies, she would have said *John ate* all *of the cookies.*

The debate over whether and which implicatures are grammaticalized is complex (Potts et al., 2016). Grammatical accounts have the benefit of explaining embedded implicatures in a direct way: these are cases where an implicated meaning is subsequently involved in the compositional semantics, as in *John believes some of*

[10] Little work in this paradigm has focused on multiparty communication, where a range of more complex dynamics emerge. For example, while a single linguistic act in a two-party exchange involves just a speaker and a listener, in a three party setting, two parties can either be directly addressed at once, or one party can be a bystander. Richer pragmatic reasoning ensues as a result, as discussed by Schober and Clark (1989) among others.

[11] At least, this is one thing an agent may do - clearly there are cases where informative, let alone cooperative, behavior is not a reasonable assumption.

the students left, where it seems that *some* comes to mean *not all* before being the proposition whose meaning it determines is input to *believe*, so that *John believes that not all of the students left* is the resulting meaning.

Gricean accounts, on the other hand, are able to handle the effects of context on implicatures. As an example, suppose that a friend and I have just discovered that the cookies we made were poisonous (we are poor bakers). Our main concern is whether another acquaintance, John, has been poisoned. In this case, I might say (7) to my friend, even if John ate all, and in turn, my friend might not draw the *not all* implicature. This suggests that the implicature is a function of what *aspect* of the world is relevant (see chapter 2 for a discussion of how this can be modeled formally). The models discussed in this work take the Gricean viewpoint as their basis, although I return to the issue of whether these approaches are truly complementary in chapter 5.

To model the implicature in (7) with a Bayesian pragmatic model, first recall that the equations defining the model do not themselves specify how we should interpret U and W. In particular, modeling this implicature calls for viewing elements of U as speech acts. Among the set of sentences which can be uttered is (7). Meanwhile W is a set of *possible worlds* which includes the worlds in which John ate only some and John ate all of the cookies. A simple interpretation of L_1 in this vein is given below, where the state 0 designates the set of possible worlds in which John ate 0 cookies, and likewise for 1, 2 and 3. Note that each state here is an equivalence class of worlds.

- $P(w) : \{0 : \frac{1}{4}, 1 : \frac{1}{4}, 2 : \frac{1}{4}, 3 : \frac{1}{4}\}$
- $P(u)$: {*John ate some of the cookies*: $\frac{1}{3}$, *John ate all of the cookies*: $\frac{1}{3}$, *John ate none of the cookies*: $\frac{1}{3}$}

We assume a semantics in which *John ate all of the cookies* is compatible only with $w = 3$, *John ate some of the cookies* with $w > 0$ and *John ate none of the cookies* with $w = 0$. Then, hearing *John ate some of the cookies* causes L_1 to prefer the states in which John ate some *but not all* of the cookies ($L_1(w = 1|u = \textit{some})$= $L_1(w = 2|u = \textit{some})$ = $\frac{4}{9} > L_1(w = 3|u = \textit{some}) = \frac{1}{9}$). Note that in a model with a rationality parameter α, as α increases, L_1's inference on hearing *John ate some of the cookies* tends towards the conclusion that L_0 would draw purely from the semantics on hearing *John ate some but not all of the cookies*.

Recent work on probabilistic models of pragmatics has focused on expanding the framework from simple implicatures like the above to richer phenomena, including vagueness (Lassiter and Goodman, 2017, 2013), manner implicature (Bergen et al., 2016), embedded implicatures (Potts et al., 2016), focus effects (Bergen and Goodman, 2015b), inferences drawn from questions (Hawkins et al., 2015) and figurative uses of language (Kao et al., 2014b, a).

These extensions often make use of tools not only from probability theory and information theory, but also from logic, in convenient ways. For example, the model of question (Hawkins et al., 2015) treats questions as partitions over possible worlds, and derives implicatures from the assumption that the asker aims to *maximize expected information gain*. The model of figurative language (Kao et al., 2014b) makes use of uncertainty over an implicit *question under discussion* to define a listener which infers the question under discussion which best justifies a given utterance. The model of focus (Bergen and Goodman, 2015b) assumes that communication takes place on a *noisy channel*, that the speaker aims to mitigate loss of important information and that the listener accordingly draws inferences about the speaker's knowledge. Because of the Bayesian setting, it is straightforward to incorporate multiple sources of uncertainty, so that, for example, a listener must reason not only about how the world is, but also what the semantics of the language is like (Bergen et al., 2016). A similar joint inference drives the model of metaphor used in chapter 2.

The coverage of RSA A number of other proposals for modeling cooperative language use, in particular the Iterated Best Response framework (Franke, 2009), exist. The focus on the Rational Speech Acts framework in this dissertation is the result of two of its prominent features. The first that it is probabilistic - language production and interpretation are represented by conditional probability distributions - and as a result, is very amenable to integration with statistical models of semantics, a central theme of this dissertation. The second is that the RSA framework has been applied to a wide and increasing range of pragmatic phenomena beyond scalar implicatures, as discussed briefly above, and as such presents the best existing candidate for a general framework in which to formalize pragmatic meaning.

1.2.1 A probabilistic conception of meaning

In the traditional approach to natural language semantics (Montague, 1973) inherited from logic (Field, 1972), we think of the semantic meaning of an utterance u as the set of worlds, known as a proposition, that are compatible with u. Typically this is obtained by mapping an utterance u to a corresponding logical formula which in turn maps, under a semantics, to a proposition. This notion of semantic meaning does not depend on context, e.g. the listener's beliefs or the other utterances available to the speaker.

In the probabilistic setting, the semantic meaning of an utterance u admits a generalization: it is the *posterior distribution of the literal listener L_0 given u*. In other words, it is the belief that u induces on a listener who reasons in terms of the semantics and a prior. Note that this *is* context dependent: the L_0 posterior depends on the prior P_L. However, this dependence is bounded: if an utterance is incompatible with a state w under the semantics, L_0 will put no probability on it, no matter the prior. Moreover, the other utterances $u' \in U$ available to a hypothetical speaker have no influence on $L_0(w|u)$. In what follows, we use the term *semantic meaning* of u in this probabilistic sense, to refer to the posterior distribution of a listener reasoning about a semantics who hears u[12].

[12] We note that inferences made on the basis of contextual information are often included in the remit of pragmatics, but we reserve that term for inferences deriving from Gricean reasoning.

Note that the logical version of the semantic meaning of any u is a *subset* of W (traditionally, a set of possible worlds, which is a proposition), while the probabilistic version is a *distribution* over W, which is like a subset of W with additional information about degrees of belief in each $w \in W$.

One nice consequence of the probabilistic viewpoint is that pragmatic meaning takes a similar form to semantic meaning: the pragmatic meaning of an utterance u is the posterior of a *pragmatic* listener, i.e. one like L_1 that reasons about an informative speaker (or a more sophisticated model, capable of carrying out more complex inferences).

So under this probabilistic perspective, pragmatic and semantic meanings are objects of the same type, namely distributions over W. While the semantic meaning depends only on $[\![\cdot]\!]$ and P_L, the pragmatic meaning depends also on $P_S(U)$, and consequently on U. In plain language, the pragmatic meaning of an utterance in context depends not only on what states are probable, and on the semantics, but on what other utterances were available to the speaker.

A final important note is that this notion of meaning is abstract in the choice of W. In certain settings, W could be as simple as a two element set, so that a meaning, in the probabilistic sense, is a Bernoulli distribution. An example of this would be a reference game with two objects, or the meaning of the utterances *yes* and *no*. However, in another setting, W could be a continuous space, describing, for example, the location of a particle, or the state of a complex system. The same goes for U, which is finite in the application of probabilistic pragmatics in chapter 2 and countably infinite in chapter 4. One could even imagine a setting where "utterances" were vowels, and lived in a continuous space. As such, one of the themes emerging from this dissertation is that the probabilistic view of semantic and pragmatic meaning unifies a diverse range of phenomena, and is entirely compatible with logical approaches to meaning.

1.3 Applying models of pragmatics to real-world settings

The example of a probabilistic model of pragmatics in section 1.2 is a proof on concept, in the sense that it exhibits a qualitative behavior (exhaustification) which appears in human language use. Indeed, many more complex RSA models are similarly proofs of concept.

In an ideal world, however, we might envisage these models being used also as practical methods for calculating or producing implicatures in natural language. This would serve two purposes: allowing theoretical insights to be put to use in computational tasks, and providing a means to validate theoretical models in realistic settings. What would it take to achieve this goal? Unsurprisingly, a number of serious challenges emerge, outlined below.

First of all, a speaker's choice of utterance in the real world is not confined to a finite set of options. It is a basic premise of linguistics that sentences are structured recursively, allowing for an unbounded number of possible utterances. While it is possible to represent these utterances, for example as sequences of words or acoustic signals, there remains a question of what utterances are considered to be available *alternatives* in a particular context (Katzir, 2007). Considering every sentence as a possible alternative is both cognitively implausible and computationally intractable, as discussed in chapter 4.

Secondly, the contrived set of states, corresponding to referents, in the example in section 1.2 is too simple to capture all but the most basic real-world inferences. How states of the world, or at least their cognitive representations, should be treated is a question which extends far beyond linguistics alone, but at the very least, a means of automatically obtaining a state space is needed.

Thirdly, it is unclear how we should obtain a semantics $[\![\cdot]\!]$. This is both a theoretical and a practical question. Theoretically, it is necessary to decide what form the semantics should take, in particular whether it should be truth-valued. Practically, there is the question of how it should be obtained in an automatic fashion from data.

Fourthly, extensions are needed in order for an RSA listener model to be able to handle non-literal, vague, ironic or presuppositional language (Kao et al., 2014a; Lassiter and Goodman, 2017; Qing et al., 2016), speech acts like questions (Hawkins et al., 2015), commands, and implicatures which are embedded or involve the maxim of matter (Potts et al., 2016; Bergen et al., 2016), to name just a few limitations of the basic model outlined in sections 1.1.1 and 1.2.

Finally, even if we did have adequate interpretations of W and U, a semantics and a more sophisticated probabilistic model, it would still be impossible to make predictions from the model without a strategy for performing inference to obtain the posterior speaker or listener distribution, which may not be tractable to compute exactly.

These are all obstacles in making the insights of models of pragmatics useful in applications which involve natural language. They are also obstacles in validating these models in real world settings; presenting a model in an idealized setting is all well and good, but does little to support the claim that the model captures human behavior.

The project of this dissertation can be seen as moving on the cline of complexity from the simple case of the reference game described in section 1.1.1 towards a fully open-domain setting. As such the aim is to address, to some degree, each of the challenges (identifying an appropriate representation of states, generating alternative utterances, obtaining a semantics, enriching the model, performing tractable inference) outlined above.

In a fully open-domain setting, all of these challenges will need to be tackled at once. Thankfully however, it is possible to consider tasks where only some of these problems need to be solved. In particular, certain tasks involve a complex utterance space U but a simple state space W, while the opposite is true for others. Some tasks involve a relatively simple model, while others require more sophistication.

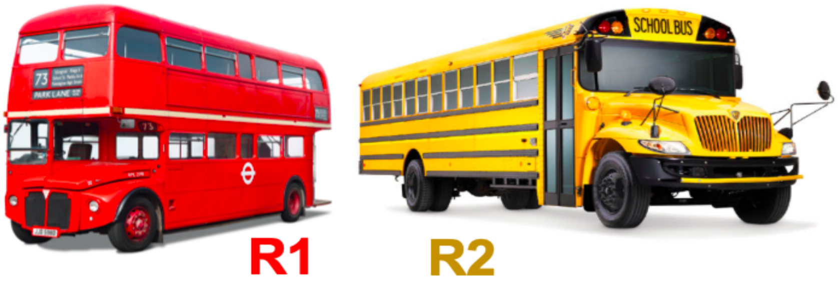


Figure 1.3

To see this, I now discuss a number of *natural language processing* (NLP) tasks involving pragmatic reasoning, and the respective challenges that the application of probabilistic models of pragmatics to them presents.

1.3.1 NLP tasks

Many NLP tasks take the form of conditional language generation, where the goal is to generate natural language which expresses some input data. One example is image captioning, where the input data is an image, and the desired output is a sentence describing that image. Another is translation, where the input data is a source language sentence, with the goal of producing a target language sentence that translates it.

Due to advances in machine learning methods in recent years (LeCun et al., 2015), both of these tasks can be performed quite well automatically (Karpathy and Fei-Fei, 2015; Bahdanau et al., 2014), by systems trained on large datasets of aligned data. Once trained, both take the form of conditional distributions, $P(caption|image)$ for image captioning and $P(translation|sentence)$ for translation.

Generating complex utterances

Image captioning For image captioning, producing a truthful caption is an obvious desideratum. That is, an image of a red bus should not be captioned with: *There is a yellow bus*. A further desideratum, however, is that the caption be informative; *This is a bus* is under-informative, particularly if the goal is to distinguish between an image of a red and a yellow bus, as shown in figure 1.3.

This leads to the task of *unambiguous image captioning*, where the goal is to accurately caption a target image in a way which doesn't refer to any of a set of distractor images. This amounts to playing a reference game in the domain of image captioning, where the states are images and the utterances are sequences of words.

The task of unambiguous image captioning provides a setting where the set of states W can be kept small, but the set of utterances U is complex, including any natural language sentence. As such, this is a good domain in

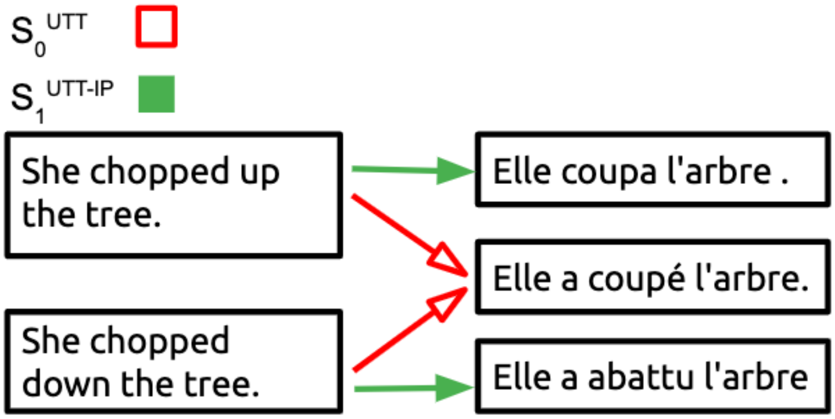


Figure 1.4: Many to one translation produced by a neural sequence to sequence model (S_0^{SNT}) and one to one translation produced by a neural model augmented with an informativity based utility at decoding time ($S_1^{\text{SNT-IP}}$)

which to investigate the challenge of producing informative utterances by using an S_1 model in a way which does not require considering a large set of possible worlds.

The approach I propose is to exploit the sequential structure of utterances, by making pragmatically informative decisions not at the level of whole utterances, but rather at the level of individual words.

Image captioning is also a domain in which the semantics is not truth-valued (i.e. a relation between utterances and states), but rather takes the form of a neurally learned conditional probability distribution. The approach taken here is to treat this neural captioning model as a literal speaker S_0, representing a conventional association between images and captions, and to design a Bayesian pragmatic model from this starting point.

Translation A similar production task to unambiguous image captioning arises in the domain of translation. Given a target sentence and a set of distractors, the goal is to translate it into another language in a way which does not amount to a translation of any of the distractors. Failure to do this results in "lossy" many-to-one translations, where multiple sentences with distinct meanings in one language map to a single sentence in another. An example is shown in figure 1.4. Here, the goal of being informative, imposed by S_1, amounts to preferring a one-to-one mapping from source to target language. In addition, translation offers a domain where the set of states, themselves being sequences of words, can have rich structure.

Interpreting abstract states

In AI and NLP tasks, it is common to learn mappings from language (as well as images) to vectors in an abstract representation space. For example, vectors of images derived from the final layers of a convolutional neural network tend to place conceptually similar images (like two different images of giraffes) close together in the representation space even if they are far away in the raw pixel space.

In computational linguistic tasks, mapping words (and more recently phrases and sentences (Peters et al., 2018; Devlin et al., 2018)) to vectors is common practice, as a way to provide semantic information useful for tasks including translation (Bahdanau et al., 2014), named entity recognition (Lample et al., 2016), and sentiment analysis (Dos Santos and Gatti, 2014).

One way to take advantage of these representations is to treat them as states of the world which a speaker is trying to convey and a listener is trying to infer. That is, a speaker is assigned a vector w and chooses a word which best conveys it to a listener. Correspondingly, a listener hears a word and infers what vector in the space the speaker was trying to communicate.

Metaphor interpretation For example, consider the setting where U is a set of predicates, and the goal of the speaker is to choose the predicate which best conveys the state of the predicated noun, denoted by a vector in a word embedding space. In particular, these predicates are metaphorical (e.g. *John is a shark*, *Time is a river*), so that the listener must *jointly* infer the state and a subspace about which the speaker cares to communicate.

This task targets a different set of challenges to image captioning and translation. In particular, it provides a good domain for addressing the challenge of enriching states beyond a small finite set W of entities, while allowing for a relatively small set of utterances U.

One challenge that both metaphor interpretation and translation/captioning address is to provide a non-truth-valued semantics. However, the approach here differs. For this task, the semantics comes from the geometry of the word embedding space, rather than a neural S_0.

Metaphor interpretation is also a domain where a more sophisticated model than L_1 is useful; a model of non-literal language proposed by Kao et al. (2014b) turns out to be adaptable to the task.

1.4 Summary of contents

In enriching idealized models of pragmatic reasoning to open-domain natural language, two main themes emerge. First, that by moving from a discrete set of states to a continuous space of latent representations, we can benefit from the dynamics that models of pragmatic reasoning already exhibit, but also inherit the power of modern statistical approaches to semantic representations.

Second, that a cognitively plausible and computationally tractable model of pragmatic reasoning should proceed incrementally. This applies not only to generation of captions and translations, but also to metaphor interpretation, where the meanings of phrases and sentences are derivable from their parts. In a sense discussed in chapter 5.4, this amounts to a proposal for a compositional pragmatics.

The content of the dissertation divides accordingly. In chapter 2, I show how a state space W that is also a vector space can be incorporated into a Bayesian model of pragmatics. I then apply one such model to the task of metaphor interpretation.

The theoretical goal is first to show that models introduced to handle non-literal meaning have a broad applicability to predication and modification generally. The technical goal is to show that Bayesian models of pragmatics make sense in a setting where W is continuous, and by means of approximate inference algorithms, allow for tractable inference. The empirical goal is to show that the interpretations of metaphors arising from an explicit model of pragmatic reasoning improve on non-pragmatic baselines, as measured by human judgments.

In chapter 3, I introduce an incremental model of pragmatic reasoning and investigate the quantitative differences in its behavior to existing models. In chapter 4, I apply this incremental model to image captioning and translation, and introduce methods to evaluate the behavior of the resulting systems.

The theoretical goal in these chapters is to show that a model of incremental pragmatic reasoning is able to account for human behaviors out of reach to a model of reasoning based on entire utterances, in particular, anticipatory implicatures (Sedivy, 2007) and certain types of over-informative behavior (Rubio-Fernández, 2016). The technical goal is to show that by reasoning incrementally, it is possible to circumvent the intractability of inference in an infinite space of utterances. The empirical goal is to demonstrate that incremental reasoning gives rise to globally informative utterances and that this improves a system's ability to produce informative language.

Chapter 2

Pragmatics with a distributional semantics

The work on metaphor interpretation discussed in this chapter is the product of joint work with Leon Bergen, as yet unpublished. Parts of the prose of that paper appears in this chapter.

Word embeddings, which map words to vectors in a high dimensional space, are part of the standard tool set of natural language processing (NLP), and are used in modern systems for translation (Bahdanau et al., 2014), sentiment analysis (Dos Santos and Gatti, 2014), image captioning (Karpathy and Fei-Fei, 2015), named entity recognition (Lample et al., 2016) and entailment detection (Bowman et al., 2015), among others.

Word embeddings present a very different perspective on meaning to the traditional approach of natural language semantics, where sentences are represented as logical formulas with truth values given a set theoretical model, and words are represented as lambda expressions which compose to form sentence type meanings. By contrast, a word embedding is difficult to interpret, save for its relation to other embeddings; standard embeddings tend to have the property that semantically similar words have embeddings which are close (in cosine distance) to each other.

The focus of this chapter is to address the following question: is a vector space representation of word (or indeed phrase and sentence) meaning compatible with the Gricean view of pragmatics? If it is, can we build models of pragmatic production and interpretation based on word embeddings? My aim is to lay out what this would look like in theory, and show that it is attainable in practice.

This is a desirable goal, since it offers a way to combine the dynamics of Bayesian pragmatic models, effective in idealized domains, with practical systems for using language. It would not only provide a means to obtain real-world systems which could reason pragmatically, but also a way to test Bayesian models of pragmatics on real world data.

2.1 Metaphor interpretation

As an NLP task involving word embeddings with which to investigate the use of pragmatic reasoning, I consider *metaphor interpretation*.

Metaphor interpretation is an example of a problem which is both of interest in NLP, where practical solutions involving word embeddings are common, and to linguistics (Lakoff and Johnson, 1980), philosophy (Black, 1955) and cognitive science (Rohrer, 2002). Viewed as an NLP task, the goal is for a computational system to take a metaphorical expression and arrive at a representation of its meaning.

For simplicity, I focus on metaphorical prediction (as in *Jane is a solider*) and metaphorical modification (as in *fiery temper*). In what follows, I refer to the predicated or modified noun phrase (*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).

A key property of metaphors is that only *some* aspects of the source are true of the target; if we know that Jane is a journalist, then Jane is presumably not like a solider with respect to owning a gun, but rather with respect to her ruthlessness or work ethic. A fiery temper is not fiery in the sense of having a high temperature - it is not even clear what this would mean - but rather with respect to its intensity or volatility.

There are several reasons for choosing this particular task as a test case for pragmatic reasoning in a distributional setting. First, the application of word embeddings to the problem is an area of active research (Shutova, 2016), where the rich lexical information present in the geometry of the representation space can be exploited. Second, there is a Bayesian pragmatic model of metaphor, which has previously been fruitfully applied in simple settings, see (Kao et al., 2014a). This model assumes that given a target word and an aspect of the target that they wish to convey, the speaker's task is to choose a source word. Meanwhile, the task of the listener is to infer two things: what *aspect* (or *aspects*) of the source are and are not relevant, and what the target is like with respect to those aspects. In the case of metaphorical modification, utterances are single adjectives in the reference game, and states are states of the head noun. For predicative metaphors, utterances are predicates and states are states of the noun corresponding to the subject. As an example, a listener may hear *solider* predicated of *man* and infer the aspect *ruthlessness*, and that the person being predicated scores highly along this aspect.

A third reason to choose this particular task is that the Bayesian pragmatic model of metaphor interpretation, which I refer to as L_1^{Q}, is a richer model than the standard L_1, and provides a good test case of a more complex pragmatic reasoning than has previously been applied to an NLP task. By happy coincidence, it turns out that this increase in model complexity results in a much simpler inference algorithm (see section 2.4).

The generality of the task While metaphor may seem to be of fairly limited importance to the broader project of building cognitively realistic and computationally useful models of language understanding, the proposed model has surprisingly general applicability. This is a result of the view of metaphor interpretation as a joint inference of relevant aspects and states of the predicated (modified) noun. From this perspective, many cases of predication and modification not traditionally treated as metaphorical can be captured by the model. As an example, when interpreting *red bus*, *red watermelon*, and *red room*, it seems as if each target noun is red in a different aspect (with respect to its exterior, its interior, and some things it contains, respectively). This line of argument is taken up in more detail in section 2.6.2.

Other views of metaphor On the other hand, the opposite criticism might also be leveled, that metaphor is of too broad importance to be treated from the perspective of L_1^Q. Indeed, some accounts of metaphor view it as a central cognitive process (Hofstadter and Sander, 2013; Lakoff and Johnson, 1980), concerning the human ability to link different domains that possess similar structures. In this respect, it is perhaps useful to note that the kind of metaphor of interest in this chapter is purely the linguistic phenomenon whereby an utterance is only literally true with respect to some partitioning of the world, but where an interlocutor is capable of inferring what this aspect is. We could think of this as metaphor in the sense of (Black, 1955), rather than any more general sense of analogy or metaphor at issue in cognitive science.

2.1.1 Pragmatic reasoning for metaphor

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, (9), said of a sleeping dog, could convey that it is unresponsive, but said of a large alert dog, could convey that it is heavy.

(9) The dog is a rock.

While certain metaphors are conventional - comparing someone to a lion tends to connote bravery - examples like (9) suggest that the interpretation of a metaphor is contextually dependent on what is known about the target. Even if the majority of metaphors become conventionalized over time (see for example the change in meaning of *fool* over time and the rarity of its original literal use), it is clear that humans are able to interpret novel metaphors. This poses a challenge which a theory relying entirely on idiomatic conventions could not resolve. In this regard, I concur with the view of MacWhinney and Fromm (2014), that “The fact that nearly all uses of metaphorical collocations are at least partially conventionalized should not obscure the fact that metaphorical language in general can be productive. ”

One pragmatic analysis of metaphor, inspired by Black (1955) and Grice (1975), posits that to interpret a metaphor, a listener must infer, based both on prior knowledge and considerations of their interlocutor’s goals, both what *aspect* of the target is being described by the source, and what the target is like with respect to that aspect.

The appeal of this view, as with Gricean explanations of language use more generally, is the ability to explain the productivity and dependence on context of metaphor, in a way which takes into account an underlying semantics (e.g. the conventional meanings of *dog* and *rock*).

I now introduce L_1^{Q}, a Bayesian model in the Rational Speech Acts framework which formalizes this process of reasoning, and discuss its application to metaphor in a hand-constructed setting, before turning to its integration with word embeddings for open-domain metaphor interpretation. The hope (which I show is borne out) is that L_1^{Q} will harness the expressiveness of a vectorial semantics but with the ability to reason pragmatically afforded by a Bayesian model.

2.1.2 A model of non-literal interpretation

L_1^{Q} can be understood as an extension of the model L_1, and in turn S_1, introduced in chapter 1.1.1. In L_1, a parameter was left implicit dictating which *aspects* of the world a speaker cares about conveying. For instance, the listener who hears "I ate some of the cookies." is modeled as drawing inferences about the number of cookies eaten, but not about whether it is raining in Timbuktu.

In other words, if we think of the full state space W as all possible worlds, in any particular model, we are partitioning W into cells, according to some *question under discussion* (QUD) (Roberts, 1996). For instance, the question *How many cookies did I eat?* partitions worlds $w \in W$ into cells according to the *number of cookies I ate* predicate.

We can make this dependence the model on a particular partitioning of the world explicit by replacing S_1 with S_1^{Q}. Here, $\delta_{a=b}$ is the delta function, equaling 1 if $a = b$, else 0.

(10) $S_1^{\mathrm{Q}}(u|w,q) \propto \sum_{w'} \delta_{q(w)=q(w')} \cdot L_0(w'|u) \cdot P_S(u)$

Surjective[1] functions $q : W \rightarrow A$ formalize the notion of a *question under discussion*. In previous RSA literature, they have themselves been referred to as questions under discussion, although I opt for the term *projection* which connects to the word vector setting I go on to discuss.

A simple example is as follows: suppose that $W = A \bigtimes B$, where $A = \{\text{Jane is hardworking, Jane is lazy}\}$ and $B = \{\text{Jane owns a gun, Jane doesn't own a gun}\}$, so that each $w \in W$ is a tuple (a, b) for $a \in A, b \in B$. Then two projections, which we could call $q_{work-ethic}$ and $q_{gun-ownership}$, are defined as $\lambda(x, y) : x$ and $\lambda(x, y) : y$ respectively. A third trivial projection is simply the identity function.

[1] A surjective function $q : A \rightarrow B$ is such that for every element b of its range B, there is some element $x \in A$ with $q(a) = b$.

Fixing some such projection q, the goal of S_1^{Q}, as with S_1, is to be informative, but now only up to the partition induced by q. For instance, for $q = q_{work-ethic}$, S_1^{Q} will prefer utterances which result in the listener model L_0 placing high weight on worlds which agree with the speaker's world on the *work ethic* dimension, regardless of the effect on the *gun ownership* dimension.

Likewise, a value of q which maps a world to all the worlds in which John ate exactly the same number of cookies will result in S_1^{Q} being informative, but only up to the goal of conveying the number of cookies. The S_1^{Q} may mislead the listener with respect to the weather in Timbuktu, for example, in the course of carrying out their goal.

A listener who reasons about the projection Since q is an explicit variable on which S_1^{Q} depends, one can create a listener L_1^{Q} which *jointly reasons* about the world w and the aspect of the world q which the speaker wishes to communicate.

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

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.

This is a direct consequence of the definition of S_1^{Q}, the model that L_1^{Q} reasons about, which is able to produce literally false utterances. Note by contrast that L_1 will assign no probability to a state w which is incompatible in the semantics with the utterance it receives. This follows from the fact that S_1 has negative infinite utility in saying an utterance u incompatible with w, because L_0 would then assign no probability to w.

Because of this property, L_1^{Q} can be used as a model of non-literal language, such as hyperbole (Kao et al., 2014b) and metaphor (Kao et al., 2014a), as I now discuss.

2.1.3 Applying L_1^{Q} to metaphor

We can model predicative metaphor using L_1^{Q} by framing it as the following communication game: a speaker wishes to communicate what an entity (or possibly type of entity) is like and chooses a predicate to do so. Conversely, a listener hears a predicate and updates their belief about what the predicated entity is like.

The reason the L_1^{Q} model is important here, as opposed to L_1 is to address the literal falsity of metaphors. *John is a shark*, for instance, if taken literally, would ascribe all properties of sharks to John. The intuition behind using L_1^{Q} is that a listener, in interpreting a metaphor, jointly reasons about which properties of sharks are relevant and what John is like with respect to those properties.

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*.

Metaphors are commonly combined with generic language (*Men are sharks*), but I treat generic language, for which separate work in the RSA framework exists (Tessler and Goodman, 2016), as an orthogonal issue.

Making concrete predictions To derive metaphor interpretations 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 $[\![\cdot]\!]$. (Assume throughout that the priors P_{L_Q} over Q and P_S over U are uniform, unless otherwise specified.)

One possible interpretation, similar to what is provided by (Kao et al., 2014a), treats points in the state space W as lists of truth values, each corresponding to a binary property. I refer to this as a *set theoretic* interpretation of L_1^{Q} and describe how it works below. As discussed in 2.2, word vectors will be incorporated into L_1^{Q} simply by changing the interpretation, but keeping the model itself the same.

Set theoretic interpretations of L_1^{Q} for metaphor Given a set of properties P, a state is a subset of P. Using the example of *John is a shark*[2], suppose $P = \{vicious, aquatic\}$, so that four states are possible: John is both vicious and aquatic, only vicious, only aquatic, or neither. Equivalently, we can think of a state as a list of truth-values, one for each property. For example, instead of writing $w = \{vicious\}$, we could write $(vicious = T, aquatic = F)$, or just (T, F).

Utterances are predicates, such as *shark*, although they need not be metaphorical; *vicious* could also be a predicate. Projections are functions $W \to \mathcal{P}(W)$ which map states to the set of states agreeing on a particular property or set of properties. For example, $q_{vicious}$ maps $\{vicious\}$ to $\{\{vicious, aquatic\}, \{vicious\}\}$, the set of states which agree with $\{vicious\}$ on the *vicious* property but may differ on the *aquatic* property. If we think of states as lists of truth-values, a projection is a function which drops some elements of the list. For instance, $q_{vicious} = (\lambda(x, y) : x)$.

An extremely minimal example of a set theoretic interpretation of L_1^{Q} is as follows:

- $P_L = \{(\text{vicious} = T, \text{aquatic} = T) : 0.075,$
  $(\text{vicious} = T, \text{aquatic} = F) : 0.675,$
  $(\text{vicious} = F, \text{aquatic} = T) : 0.025,$
  $(\text{vicious} = F, \text{aquatic} = F) : 0.225\}$
- $U = \{\textit{shark}, \textit{silence}\}$

[2] Granted, this usage of *shark* is somewhat conventionalized. The reader may substitute this predicate for a metaphor of their choosing if they prefer.

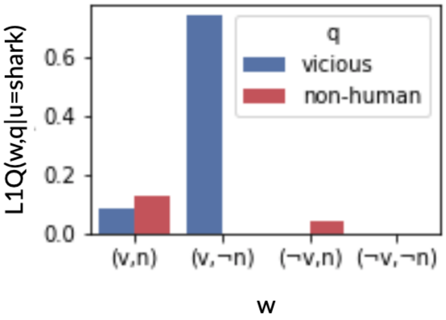


Figure 2.1: Figure showing the posterior distribution of L_1^Q on hearing *shark*, when *swimmner* is the other possible utterance, and *vicious* and *aquatic* are the QUDs. v and n abbreviate the Boolean variables *vicious* and *aquatic* respectively.

- $Q = \{q_{vicious}(\lambda(x,y) : x), \quad q_{aquatic}(\lambda(x,y) : y)\}$

The results of L_1^Q hearing *shark* are shown in figure 2.1. Note that a semantics $[\![\cdot]\!]$ is assumed in which *shark* is compatible only with (*vicious*, *aquatic*), and *silence* is compatible with every state. Also note that the projections map each tuple to its value at a single property. In theory, for larger n-tuples of properties, a projection could map to multiple properties, representing a speaker who wishes to communicate multiple aspects of the state w. I return to this point in section 2.6.4.

The key fact to observe about L_1^Q in this example is that the prior belief that John is not aquatic 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. In this respect, L_1^Q qualitatively differs from L_1: it can hear an utterance u and infer a world w which is not compatible with u in the semantics.

Importantly however, 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 like L_1: for instance, suppose we add a third property, *quickness*, so that *shark* is compatible only with $(\text{vicious} = T, \text{aquatic} = T, \text{quick} = T)$, and also add a third utterance, *hummingbird*, compatible with only $(\text{vicious} = F, \text{aquatic} = T, \text{quick} = T)$.

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.

We can also marginalize out the world variable by summing over it, to obtain a *marginal posterior distribution* over projections in Q. This tells us which projection is most likely, given that the listener heard *shark*. Similarly, we can obtain a marginal posterior over states in W.

With the probabilistic notion of meaning introduced in chapter 1.2.1 in mind, we can say that the posterior distribution of L_0 over W on hearing a metaphor captures its *literal* meaning, while the marginal posterior distribution of L_1^{Q} over W captures its full, metaphorical meaning. These two meanings are objects of the same type, which allows them to be directly compared. In the case of this example they differ in that the literal interpretation places no probability on John *not* being aquatic, while the pragmatic meaning does.

Given hand-selected utterances and states, as well as projection functions and a semantics, L_1^{Q} makes predictions which qualitatively reflect a Gricean story: on hearing a metaphor like *John is a shark*, a listener has a prior belief that only certain properties of sharks pertain to John, and jointly reasons about which aspects of John the speaker wishes to communicate and what John is like.

In this regard, the model displays qualitative behavior that appears to capture human metaphor interpretation well (Kao et al., 2014a), but to evaluate on arbitrary predicative metaphors, a hand-supplied semantics is required. This severely restricts the utility of the model as a computational system to provide interpretations of metaphors. It also makes evaluation of its predictions difficult (though possible through a crowd sourced semantics (Kao et al., 2014a)).

This isn't an inherent problem with the L_1^{Q} model, so much as with current interpretation, which requires a number of elements difficult to supply other than in a manual fashion. If these elements could instead be supplied in an automatic way, we would be able to apply L_1^{Q} to arbitrary predicative metaphors. I now introduce word embeddings, and propose a way to use them as a computationally obtained semantics for L_1^{Q}.

2.2 Word embeddings

A word embedding E is a mapping from words to points in a high-dimensional vector space[3]. By dimensionality reduction of a co-occurrence matrix (Pennington et al., 2014), or by extracting the weights of a statistical model (Mikolov et al., 2013; Peters et al., 2018; Devlin et al., 2018), embeddings can be obtained which are useful for downstream tasks (Dai and Le, 2015; Radford et al., 2018; Chen et al., 2016; He et al., 2017).

Insofar as they improve performance when used as a starting point for downstream tasks, it would seem that embeddings map words to vectors which capture semantic information. While it is difficult to decode this information explicitly, for instance by producing a basis of the vector space corresponding to concepts that compose word meanings, it has been observed that the geometry of the space has compelling properties with respect to word meanings.

For instance, in well-trained embeddings, human-judged semantic similarity of a pair of words a and b corresponds to a metric, such as cosine distance, between the vectors $\overrightarrow{a}$ and $\overrightarrow{b}$.

[3] A vector space is a set equipped with structure which allows elements to be added together, and multiplied by scalars. See Axler (1997) for a rigorous definition.

Metaphor is an obvious candidate for the use of word embeddings: 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).

2.2.1 Connecting word embeddings to models of pragmatics

The question now is how to use word embeddings in the context of a communication game, in particular the one described above for metaphor. The key here is to interpret states of the target as points in the word embedding space. From this perspective, for a given predicative metaphor (say *John is a shark*), every point in the word embedding space corresponds to a way John could be.

The goal of the speaker is to choose a source word in order 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 et al., 2010), in an abstract word embedding space.

For AN metaphors, a similar approach can be taken. To interpret the AN phrase *fiery temper*, a listener has prior uncertainty about what point in the space best describes the temper in question, and updates their beliefs on the basis of the "utterance" *fiery*.

The closest work to this perspective is (Kintsch, 2000), which proposes a single scheme for literal and metaphorical predication in a vector space setting.

Recalling the probabilistic notion of meaning introduced in chapter 1.2.1, we can say that the meaning of a metaphor, from this perspective, is a distribution over points in the vector space that a word embedding maps to (which word embedding we use is a hyperparameter of the model, but we consider it fixed for the purpose of this discussion). For example, we can say that the meaning of *temper*, and of *fiery temper*, are then both distributions over points in the word embedding space.

Note that word embedding spaces do not come equipped with interpretable dimensions[4]. That is to say, it is not that case that each dimension of the space corresponds to a property, let alone a property of John. This raises the question of what it means to treat points in the word embedding space as states.

However, by using cosine distance (or another metric) we can still make sense of a vector in the space representing the state of John. For instance, a given state represents John being dependable, clever or tall to the extent that its cosine distance to $\overrightarrow{dependable}$, $\overrightarrow{clever}$ or $\overrightarrow{tall}$ is low.

Of course, this is a crude representation of meaning at best. There is no distinction between words with clearly different types, such as verbs, adverbs and proper nouns, and no notion of truth. It is not clear what information the distance between many pairs of words, e.g. *idealism* and *of*, conveys.

[4] While the co-occurrence matrix that GLoVe is drawn from does have interpretable dimensions - each is a word, this is not the case once the dimensionality is reduced.

As such, we should view a word embedding under this interpretation of states as a rough, but useful representation of meaning, which has enough structure to be used for the task at hand.

I now give a concrete example of a vectorial interpretation of an RSA model, namely L_1^{Q}. For reasons discussed in section 2.4, the introduction of inference over projections at the level of the pragmatic listener, ostensibly adding additional complexity to L_1, turns out to simplify the inference procedure substantially, and to provide results which are simpler to interpret.

2.3 A vectorial interpretation of L_1^{Q}

The set-theoretic interpretation discussed in chapter 1 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 $U \to (W \to \{0,1\})$.

We now introduce a *vectorial* interpretation of L_1^{Q}. Importantly, this requires no modification to equations (10, 11). As in the example presented in section 2.1.2, U is a set of adjectives. The crucial difference is that the state space W is now not just a set, but a vector space determined by a word embedding $E : U \to W$, so that elements $\overrightarrow{w} \in W$ are vectors.

Note that this generalization is mathematically 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 $\overrightarrow{w}$ as a vector with entries 0s and 1s, or T and F, dictating the presence or absence of each property $p \in P$. From this perspective, it turns out that the only change made in introducing word embeddings is to move to a vector space over the real field, and to introduce a new semantics, as discussed below.

2.3.1 The listener's prior

In the set-theoretic interpretation of L_1^{Q}, with W finite, a discrete prior P_L over W sufficed. In the present case, where W is necessarily infinite (ranging continuously over real-valued vectors), a multivariate spherical Gaussian distribution is used, which can be parametrized by a vector μ for the mean and a single scalar σ (the value of every diagonal entry of the covariance matrix). The prior over projections P_{L_Q} is taken to be uniform.

$$(12)\quad P_L(w) = P_{\mathcal{N}}(w|\mu = E(\textit{target}), \sigma = \sigma_1)$$

The multidimensional Gaussian distribution weights most heavily those points nearest to its mean. By setting the mean of the prior as E(*target*) (where *target* is, for example, *temper* in the metaphor *fiery temper*), the model encodes 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, representing the degree of uncertainty the listener has, and consequently, their willingness to update their beliefs.

2.3.2 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 (11) requires only that the semantics $[\![\cdot]\!]$ be a function $U \rightarrow (W \rightarrow \mathbb{R})$. We can define such a function as follows, with σ_2 as a hyperparameter, and u the metaphorical adjective or *source*:

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

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. As with the definition of the prior in (12), the semantics introduces a hyperparameter, namely σ_2. $[\![u]\!](w)$ decreases with σ_2.

Note the analogy between a truth-conditional semantics and this probabilistic one: in the former, $[\![u]\!]$ is effectively a set of worlds, so that $[\![u]\!]$(w) iff $w \in [\![u]\!]$. In the latter, $[\![u]\!]$ is a distribution, so that $[\![u]\!]$(w) is the mass placed on w by $[\![u]\!]$.

2.3.3 The literal listener

As before, the literal listener is defined as in (14):

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

Substituting in the definitions of the prior and semantics, this gives us (15):

(15) $L_0(w|u) \propto P_{\mathcal{N}}(w|\mu = E(u), \sigma = \sigma_2) \cdot P_{\mathcal{N}}(w|\mu = E(\textit{target}), \sigma = \sigma_1)$

To give an intuition for the behavior of L_0 in the vectorial interpretation, figure 2.2 depicts the effect of hearing *shark* (denoted as the vector (1,1)) given a prior centered at *man* (denoted as (0,0)). The dimensions correspond to the two properties describing the state of the target used in the comparable set-theoretic interpretation of *The man is a shark* introduced in section 2.1.3.

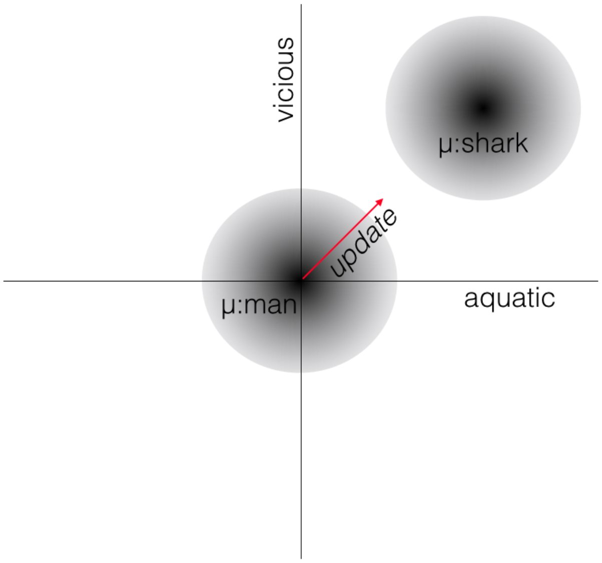


Figure 2.2: Illustration of literal listener L_0 given *The man is a shark*, with $\overrightarrow{man}$ = (0,0) and $\overrightarrow{shark}$ = (1,1). L_0's prior is centered at $\overrightarrow{man}$, and is updated towards $\overrightarrow{shark}$.

2.3.4 Projections

Finally, a notion of a projection function q that is defined on a vector space, and a set Q of such projections, is needed.

It has been argued (Pennington et al., 2014) that word embeddings such as GLoVe exhibit a degree of linear structure, in the sense that for various quadruples of words (A,B,C,D), such that A is to B as C is to D, the corresponding pretrained vectors approximately satisfy the equation $\overrightarrow{\text{A}} - \overrightarrow{\text{B}} = \overrightarrow{\text{C}} - \overrightarrow{\text{D}}$, where $\overrightarrow{\text{A}}$ is the word vector corresponding to the word A[5]. For instance, the nearest word vector by cosine distance to the point $(\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}})$ in the Word2Vec embedding space is $\overrightarrow{\text{queen}}$.

For this reason, the natural implementation of a projection in a vector space is as a linear projection, parametrized by some hyperplane, which maps from the full space to a lower dimensional subspace. Geometrically, it can be thought of as dropping a line from an input vector $\overrightarrow{w}$ at a right angle onto a vector (or hyperplane) $\overrightarrow{u}$, capturing the degree to which $\overrightarrow{w}$ extends along $\overrightarrow{v}$, and ignoring orthogonal dimensions. See figure 2.3 for a two dimensional example.

In the more general multidimensional case, given a hyperplane in our space we can obtain a function mapping points in the space to new positions, derived by dropping them perpendicularly onto the hyperplane. This projection is a linear transform to a subspace of the original vector space.

[5] But see, for example, (Schluter, 2018) for limitations of this perspective.

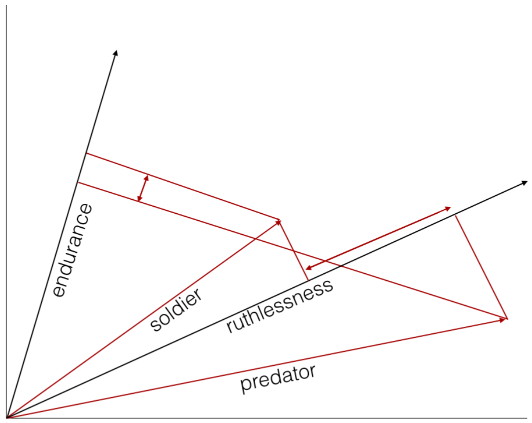


Figure 2.3: In this hand-constructed 2D example, vectors for $\overrightarrow{soldier}$ and $\overrightarrow{predator}$ are mapped onto subspaces given by $\overrightarrow{endurance}$ and $\overrightarrow{ruthlessness}$.

Mathematically, linear projections are the natural analogue of the projection functions used in the set-theoretic interpretation of L_1^{Q}; when viewed as vectors in a vector space over the Boolean field, projection functions are precisely linear projections.

What set of linear projections Q should we choose? In the set-theoretic interpretation, Q was simply the set of all one-dimensional projections, or equivalently, projections onto each of the standard basis vectors. In the word embedding case, however, the basis vectors of a word embedding E have no simple interpretation as properties, since, in a distributional setting, the axes of a world state vector do not neatly correspond to its attributes.

Instead, we have to find some other subspaces on which to project. 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 $\overrightarrow{vicious}$.

In practice, I 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?

2.3.5 The model as a whole

Putting all this together, we have a model in which a metaphor is interpreted as a distribution over points in a word embedding space, obtained by starting with a prior centered at the target (e.g. *man*) and moving towards the target (e.g. *shark*), but only along certain subspaces. Which subspaces these are, and in what direction one moves along them, is determined by reasoning about what an informative speaker would have been likely to have said, given any particular pair of subspace and world.

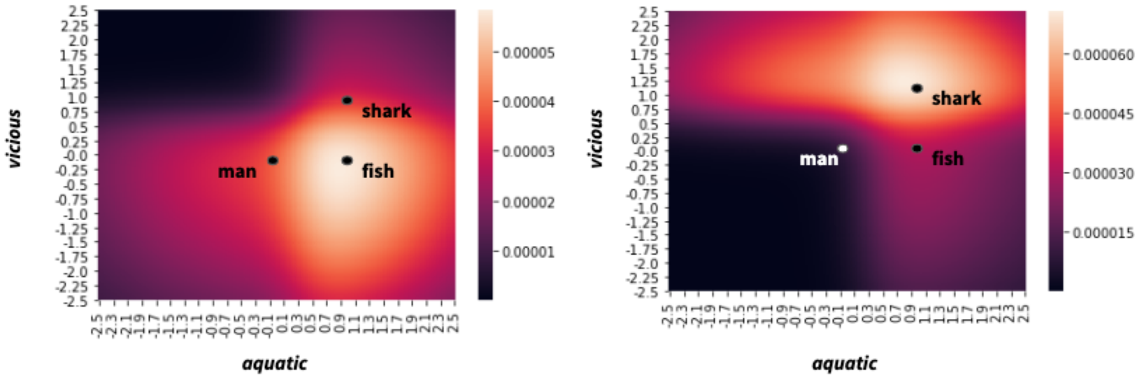


Figure 2.4: Heatmaps visualizing the inferred L_1^{Q} marginal posterior over worlds given *fish* (left) and *shark* (right), with $U = \{$*man*, *shark*, *fish*$\}$, hand-chosen denotations overlaid, and $\sigma_1 = 5.0, \sigma_2 = 0.5$.

It is useful to visualize this process with a two-dimensional example. Figure 2.4 provides a visualization of the L_1^{Q} posterior in a simple 2D case corresponding to the example used in section 2.1.3. Here, the semantics is again hand-chosen, with exactly the same denotations for each word and the same sets of utterances U and projections Q, but W is a continuous space. Brighter regions of the heatmap correspond to regions with greater probability mass. The listener's prior is a Gaussian ball around *man*, so the result of saying *fish* (shown in the left hand heatmap) is a shift of probability mass towards the vector $\overrightarrow{fish}$. However, more mass can be seen below $\overrightarrow{fish}$ than above. This is because L_1^{Q} has inferred that had a point above $\overrightarrow{fish}$ been the state that the speaker intended to communicate, *shark* would have been a preferable utterance. This can be understood as the computation of a scalar implicature, but in a continuous space.

Looking at the right hand heatmap, the model has shifted weight towards *shark*, but additionally has its probability density spread with higher variance along the x axis than the y axis. The reason for this is that L_1^{Q} has drawn an inference that *vicious*, the projection corresponding to variation along the y axis, is the relevant one, and that for this reason, it is relatively likely that the speaker is not intending to communicate states with high values on the x axis. Again, this is precisely the inference drawn by the set-theoretic interpretation of L_1^{Q}, but now in a continuous space.

Questions under discussion and linear projections The use of linear projections to model the ignoring of certain features mirrors intuitions observed elsewhere in the cognitive science and natural language processing literature. For instance, Kintsch (2000) notes:

> "Computing a meaning always involves activating context-appropriate features and inhibiting or deactivating inappropriate features."

A comparable point is made about adjective-noun (AN) composition by Grefenstette (2013), which resembles the intuition motivating L_1^{Q} quite closely:

> “In turn, through composition with its argument, I expect the function for such an adjective to *strengthen* the properties that characterise it in the representation of the object it takes as argument...When I apply “angry” to “dog” the vector for the compound “angry dog” should contain some of the information found in the vector for “dog”. But this vector should also have higher values for the basis weights of “fighting”, “aggressive” and “mean”, and correspondingly lower values for the basis weights of “passive”, “peaceful”, “loves”.”

2.4 Inference

The L_1^{Q} model, when instantiated in the setting of word vectors, poses a problem for inference. The standard method of enumerating all possible outcomes is clearly not viable at L_0 when W is infinite. The problem becomes yet more complicated at L_1^{Q}, where a joint inference between a discrete set Q and an infinite W takes place.

Either analytic or approximate methods are required. In practice, a mix of the two is used; the L_0 and S_1 posteriors can be calculated analytically, while L_1^{Q} requires the development of an approximate inference algorithm.

For reasons of efficiency, I only consider projections along a vector, rather than a larger subspace. This means that projections correspond to single adjectives (although see section 2.6.4 for a discussion of the benefits of multidimensional projections).

I now describe this algorithm in parts, working up from the L_0.

2.4.1 $\mathbf{L_0}$

Intuitively, the vectorial interpretation of L_0 amounts to the process shown in figure 2.2, where a ball, corresponding to the prior, is moved in the direction of the point corresponding 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 12, and the semantic interpretation is defined as in Equation 13, then conjugacy implies that the listener posterior is given by:

$$(16) \quad L_0(w|u) = P_{\mathcal{N}}(w|\mu = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \cdot (\frac{\textit{E(target)}}{\sigma_1^2} + \frac{\textit{E(u)}}{\sigma_2^2}), \sigma = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2})$$

2.4.2 $\mathbf{S_1{}^Q}$

The speaker is defined by Equation 10, which in the continuous case can be rewritten as:

(17) $S_1^Q(u|w,q) \propto \int_{w'} \delta_{q(w)=q(w')} \cdot L_0(w'|u)$

This integral is computing the marginal probability of w_q, the projection of world w onto QUD vector q. From Equation 16, $L_0(\cdot|u)$ 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.

2.4.3 L_1^Q

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

$$\begin{aligned}
&L_1(q|u)\\
&= \int_w L_1(w,q|u)\\
&= \frac{1}{K} P_{L_Q}(q) \int_w P_L(w) S_1^Q(u|w,q)\\
&= \frac{1}{K} P_{L_Q}(q) \int_w P_L(w_q, w^{\perp}) S_1^Q(u|w_q,q)\\
&= \frac{1}{K} P_{L_Q}(q) \int_w P_L(w_q) P_L(w^{\perp}) S_1^Q(u|w_q,q)\\
&= \frac{1}{K} P_{L_Q}(q) \int_{w^{\perp}\in Q^{\perp}} P_L(w^{\perp}) \int_{w_q \in Q} P_L(w_q) S_1^Q(u|w_q,q)\\
&= \frac{1}{K} P_{L_Q}(q) \int_{w_q \in Q} P_L(w_q) S_1^Q(u|w_q,q)
\end{aligned}$$

I verify the correctness of this algorithm in the 2 dimensional case by comparison to the exact posterior, which is numerically derivable in 2 dimensions (by discretization of the continuous space).

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. The constant K can be found from the constraint $\sum_q L_1(q|u) = 1$.

2.4.4 Implementation Details

Implementing this algorithm requires vectorization of the code. This refers to the standard practice in machine learning models of avoiding loops and making use of matrix operations which can be calculated more efficiently and benefit from GPU speed up. In particular, the loop at the S_1^Q over utterances can be vectorized.

The model was implemented both in Tensorflow and NumPy. The motivation for the former was the ability to make use of inference algorithms which require gradients to be calculated automatically (such as Hamiltonian Monte Carlo and Variational Inference) - however, as it turned out, the L_1^{Q} inference algorithm does not require this, since the problem reduces to a one dimensional integral, as shown in the previous section.

2.5 Evaluating the model of metaphor interpretation

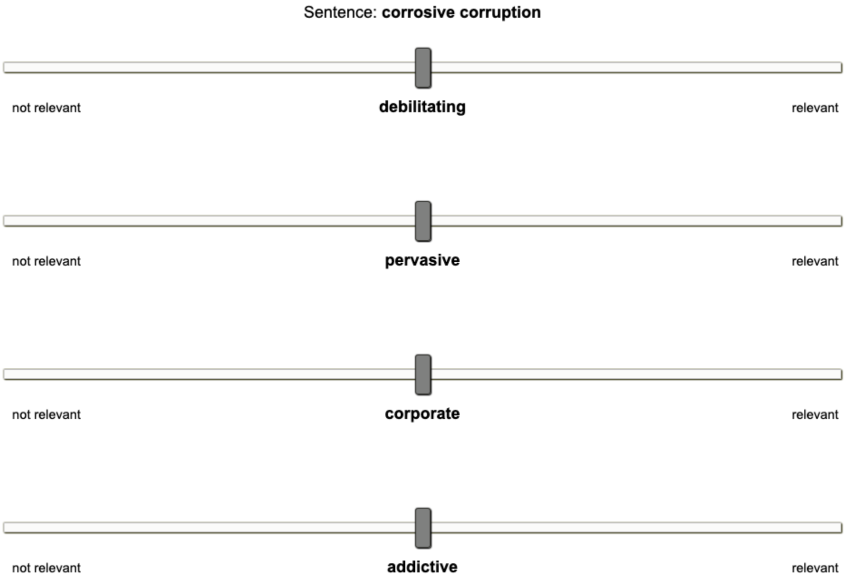
In order to inspect the behavior of L_1^{Q}, a method for transforming its posterior distribution into an interpretable prediction is needed. Points $w \in W$ are difficult to interpret on their own, but projections $q \in Q$ on the other hand correspond directly to adjectives representing the aspect of the subject that the speaker wishes to convey.

For this reason, I use the marginal posterior over Q to generate predictions from the model. In particular, the marginal posterior over Q generates a ranking on the set of adjectives used to supply the projections, so that the best interpretations of a metaphor can be taken to be the highest ranked adjectives under this distribution.

I now discuss initial attempts to measure the quality of these predictions. While predictions from L_1^{Q} appear to be qualitatively reasonable, we encountered difficulty outperforming a baseline model which uses word vectors but no explicit model of pragmatics.

Random baseline The random baseline model is defined as follows: for a given metaphor of the form $(a$ $n)$, we take the mean of the embeddings of the adjective *a* and noun *n*. We then randomly select two adjectives from the top 100 hundred adjectives nearest to this mean and use these as the baseline interpretations for the metaphor. The mean (which is a weighted sum) is used in light of the effectiveness of vector addition in deriving representations of phrasal and sentence meanings from constituent words, see (Mitchell and Lapata, 2010; Grefenstette, 2013; Socher et al., 2013). Cosine distance is a standard metric of similarity used for word embeddings (Pennington et al., 2014).

Stronger baseline The stronger baseline, rather than choosing randomly from the 100 nearest adjectives, selects the top 2 nearest.



Sentence: corrosive corruption

Figure 2.5: An item in the experiment. Item order, and in-item order of the 4 adjectives from L_1^{Q} and baseline models is randomized.

Experiment Tsvetkov et al. (2014) provides a corpus of ∼800 AN metaphors, gathered by human annotators, of which I select ∼100 of the least frequent by bigram count[6] for the experiment, in order to filter out conventionalized metaphors. Our full set of 109 metaphors is shown in figure 2.6.

In the 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. Figure 2.6 shows the results from a comparison of L_1^{Q} and the *random* baseline described above. An example is shown in figure 2.5.

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 (although this affects results very little, which remain significant without the exclusion).

2.5.1 L_1^{Q} hyperparameters

The 300 dimensional GloVe vectors trained on Wikipedia 2014 and Gigaword 5 are used as the word embedding E. For each AN metaphor $(a\ n)$, U is the 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

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

L_1^Q adjectives	random 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

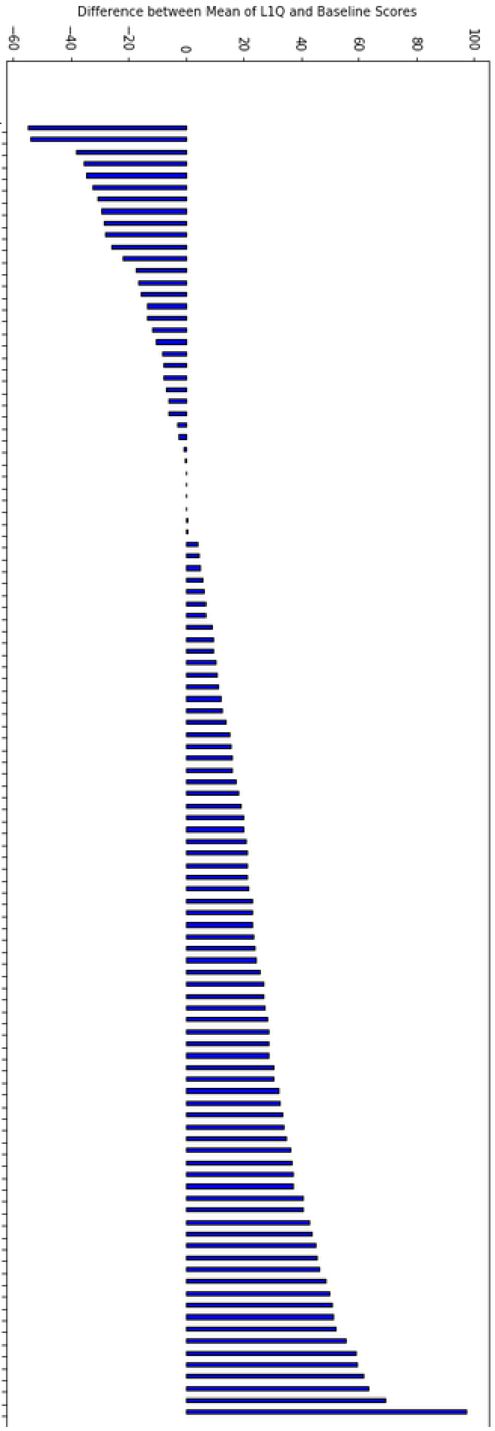


Figure 2.6: The 109 metaphors used in the experiment, and corresponding L_1^Q proposals and random baseline, plotted against mean rating given to that metaphor under these two models.

adjectives with concreteness ranking > 3.0 in the concreteness corpus of Brysbaert et al. (2014), to exclude abstract nouns.

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

By tuning on a validation set of hand-selected metaphors, $\sigma_1 = \sigma_2 = 0.1$ are chosen as the best values of these hyperparameters. The adjectives corresponding to the two projections with highest marginal posterior mass under L_1^{Q} are selected as the interpretations provided from the model in the experiment.

2.5.2 Results

The results of the comparison between L_1^{Q} and the random baseline, shown in Figure 2.6, 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).

Failure to beat a stronger baseline We also try comparing L_1^{Q} with the strong baseline, where the top 2 nearest adjectives to the mean of the adjective and noun are the baseline proposal. In this case, a significant preference for the baseline model was found. This means that there is little to no evidence that pragmatic reasoning, as embodied by the L_1^{Q} model, is valuable for this task, since the random baseline is very weak indeed.

2.6 Discussion

The system for metaphor interpretation proposed in this chapter forms part of a more general conviction that the way forward for cognitively accurate models of natural language interpretation and production should involve an empirically learned semantics (which may be in part, or entirely non-truth conditional) and an explicit model of interagent reasoning.

The approach taken here has been to use a distributional semantics in concert with a Bayesian model of pragmatic reasoning, to tackle the task of metaphor interpretation. The core steps which allowed this were to take points in a vector space to represent possible states of a predicated or modified target noun, to use linear projections to represent *aspects* of this state, and to develop an inference algorithm capable of approximating the posterior distribution of L_1^{Q}. This demonstrates that a continuous state space is not an impossible challenge for an RSA model.

I take the compatibility of Gricean pragmatics and vector space models of meaning to be a separate question to whether the latter is an accurate or even effective representation of meaning. That said, the compatibility of a Gricean theory of pragmatics with a non-truth-valued semantics shows that a truth-valued semantics is not essential to the Gricean enterprise - a useful technical point.

As well as being a computational linguistic tool, the fact that L_1^{Q} can handle arbitrary AN metaphors allows for more thorough empirical investigation of the utility of pragmatic reasoning in performing natural language tasks.

I conclude the chapter with a discussion of the place of the proposed metaphor interpretation system in the more general goal of pragmatic reasoning about natural language. First, I outline some of the implicit assumptions the system relies on and how they can be relaxed where needed (section 2.6.1). I then discuss the possibility that, while proposed as a system for metaphor interpretation, L_1^{Q} serves to capture adjectival modification and predication more generally (section 2.6.2). Extending the argument, I note that L_1^{Q}'s process of AN phrase interpretation, which begins (in the prior) with the meaning of a noun, and results (in the posterior) in the meaning of a noun phrase, is suggestively compositional; I hypothesize that this is indeed a form of compositionality for word vectors, a desirable practical and theoretical goal (section 2.6.3). Finally, I outline key directions of future work to develop the system, towards a fully open-domain model of language interpretation (section 2.6.4).

2.6.1 Assumptions of the model

There are several assumptions about word embeddings that L_1^{Q} relies on which are approximations at best. The first comes from the fact that the vectorial interpretation of L_1^{Q} uses a word embedding as a semantics. This assumes that a word embedding, specifically GLoVe, contains only semantic information.

In practice, however, there is every reason to suspect that a model such as GLoVe or Word2Vec, which is trained on real corpora, learns to incorporate both semantic and pragmatic information.

As a concrete example, the co-occurrence of *shark* and *man* arising from instances of the metaphor “The man is a shark.” in GloVe’s training data, influences the position of the vectors for *shark* and *man*. Thus, in some sense, GLoVE is already representing information which it is the task of L_1^{Q} to infer. For this reason, it would seem that pragmatics is being *double counted* by the model, in the sense that the semantics already encodes pragmatic information.

To avoid this double counting, the ideal solution is to obtain a word embedding which only represents semantic information. Bayesian modeling offers a principled way to do this: infer a semantics (in the form of a word embedding) from observations of real world data on the assumption that these data are produced by pragmatic reasoning as modeled by L_1^{Q}. In other words, the solution to infer what a semantics must be like such that, when used as the basis of a model of pragmatic reasoning, it would produce language as observed in corpora.

One way to approach this challenging inference problem is to *back-propagate* through an RSA model based on observations of pragmatically generated data. This approach is taken by Monroe and Potts (2015), where a semantics and pragmatics are jointly learned.

In the present case, the idea would be to simultaneously update the values of the word embeddings while drawing inference about the meanings of particular phrases or sentences. Aside from working out the conceptual details much more concretely, the barrier to such an approach is the speed of the inference algorithm; in order to train on a large scale dataset, the process of deriving pragmatic interpretations would have to be implemented in a significantly faster way. This challenge presents an important avenue for future work.

Non-linearity The use of linear projections in the vectorial interpretations of L_1^{Q} and S_1^{Q} exploits the assumed linear structure of the embedding space (Pennington et al., 2014). This linearity is approximate at best (see (Linzen, 2016; Finley et al., 2017) for potential caveats). A natural question, therefore, is whether a non-linear notion of projection is more suitable; the difficulty here is that linear projections are both simple to compute and easy to generate from the vectors of arbitrary words.

With respect to the discussion above, of inferring a suitable semantic representation, another approach is available. By using a pragmatic model which makes the assumption of linearity, it would be possible to back out a word embedding which is linear in the desired sense *by design*.

2.6.2 The scope of the model

Prima facie, it seems that some predications and modifications are metaphorical, while others are literal. As far as L_1^{Q} is concerned, metaphorical meaning is distinguished from literal meaning by involving a projection which determines some aspects of the world which the speaker cares to communicate. Viewed from this perspective, metaphorical language is arguably very pervasive; in fact, there is a case to be made that the mechanism of L_1^{Q} should be taken as an account for modification and predication generally.

Subsectivity and metaphor An argument of this form has been made in the context of intersective and non-intersective adjectives.

In truth-conditional approaches to semantics, intersective adjectives are defined by the following property: for an intersective adjective A, if A_x is the set of things which are A, and for a noun N, if N_x is the set of things in the extension of that noun, then the intersection of these two sets is the extension of $[AN]$. For example, if one believes that *red* is an intersective adjective, then the set of red apples would be the intersection of the set of red things and the set of apples.

By contrast, a *non-intersective* adjective is one like *skillful*, where a skillful cook might be skillful in a different way to a skillful dancer. As such, if one were to posit an intersective analysis, on which the extension *skillful*$_x$ consisted of the set of skillful things, it would trigger the unwanted entailment that a skillful dancer who was a cook was also a skillful cook. Instead, one can say that a non-intersective adjective is a function, taking a noun and returning a new extension.

The intuition connecting intersectivity to metaphor is this: non-intersective adjectives modify their noun only with respect to certain aspects, much like how metaphorical modification requires the identification of the aspect of the noun that is being modified. As such, I put forward the following hypothesis: metaphorical and non-intersective modification can both be modeled by the joint inference over projections in L_1^{Q}. Exploring this hypothesis, by applying the model to non-intersective adjective modification, is an avenue of future work.

The extent of metaphor The question of the extent of metaphor closely parallels the discussion raised early in the field of truth-conditional semantics, of the extent of non-intersectivity.

Quine, for instance notes that “a red apple is red on the outside while a pink grapefruit is pink on the inside” (Lahav, 1989). Though not framed in these terms by Quine, this suggests that an intersective treatment may not be appropriate for color modifiers, since the way in which a given object is a certain color depends on the object, much in the same way as skillfulness is object dependent. That is, an intersective treatment would force us to say that a pink grapefruit is also a pink object, rather than merely being pink for a grapefruit.

Partee draws an explicit connection between this observation and metaphorical modification (see (Lahav, 1989)), by suggesting that cases like *red apple* could be part of the same phenomenon as *flat note*, *flat beer* and *flat tire*, where each is flat in a different way. This leads to a second hypothesis, that all adjectival modification is metaphorical, in the sense of metaphor specified by L_1^{Q}.

To the extent that they are true, the consequence of these two hypotheses is that the mechanism used in L_1^{Q} has general purpose applicability to adjectival modification.

2.6.3 Compositional distributional semantics

Distributional semantics models rich lexical information in a way which is demonstrably useful for NLP tasks. However, it offers no canonical method of composition[7], by which word meanings (represented as

[7] Or rather, it does in the form of the matrix-vector product, but this is linear and for that reason seems too weak a form of composition.

vectors) can be assembled into meanings for phrases or sentences.

This is a stark contrast with traditional approaches in semantics (Montague, 1973), which place relatively little information in lexical items, but offer a rich theory of composition, by way of the simply-typed lambda calculus (Loader, 1998).

In the past, attempts have been made to create a *compositional distributional* theory, either through a direct analogue of lambda calculus (Coecke et al., 2010) or through statistical models which respect a structure determined by the syntax (Socher et al., 2013). This is a desideratum for both NLP applications and linguistic theory (Baroni et al., 2014), since it provides a modular and principled way to reduce the complexity of a sentence meanings to its simpler constituents.

As an alternative to these proposals, Bayesian models of language interpretation offer a notion of composition which, in the context of a vectorial interpretation, is also distributional.[8]

The application of L_1^{Q} to metaphorical AN phrase interpretation provides an example: here, the posterior distribution of L_1^{Q} (or more precisely, the marginal posterior over W) represents the meaning of an AN phrase. This meaning is determined by the head noun, which parametrizes the L_1^{Q} prior, and a received utterance (the adjective), on the basis of which L_1^{Q} updates the prior to the posterior. To put it another way, the modifying effect of the metaphorical adjective on the noun here is represented by the update from the L_1^{Q} prior to posterior.

The case of L_1^{Q} can be generalized in two ways, to obtain a much more general claim about composition. First, in light of the discussion of the generality of L_1^{Q} as a model of modification in section 2.6.2, a general approach to adjective noun phrase composition can be proposed:

(18) If the meaning of [NOUN] is the prior of a listener model, then the meaning of [ADJECTIVE NOUN] is the posterior of a listener model (L_1^{Q} or another appropriate model) after receiving [ADJECTIVE].

The second generalization is from adjective noun phrases to composition more broadly, between any head and modifier, in a recursive manner. To obtain a meaning for *unpredictable fiery temper*, the posterior distribution obtained from *fiery temper* can subsequently be used as a prior, to be updated by *unpredictable*. Since the compositional process proposed here can be performed recursively, it is possible to obtain the meaning of any phrase or sentence as a distribution over a word embedding space, by iteratively composing the subtrees of the sentence in the normal fashion. In other words, the move from L_1 prior to posterior (the process of Bayesian inference) can be understood as the function by which an head and modifier combine, generalizing the mechanism of function application used in a truth-conditional setting.

[8] The term *distributional* is unfortunate here, since it bears no relation to the probability distribution over U or W used in the models themselves.

This compositional distributional proposal differs from (Coecke et al., 2010) in returning a *distribution* over points in the word embedding space as the meaning for a phrase, rather than a single point. It also differs, more substantially, by incorporating pragmatic reasoning at each stage of the composition. Note, however, that an analogous version of this proposal, using L_0 instead of L_1^{Q}, would not involve Gricean pragmatic reasoning (but see chapter (5) for a discussion of the relation of this proposal for compositional distributional semantics to incremental pragmatic reasoning as introduced in chapter 3).

2.6.4 Future work

I now discuss a number of possible extensions of the system proposed in this chapter, to the end of a general model of pragmatic reasoning for natural language.

Validation of the model While the model predictions, for example those shown in section 2.5 appear qualitatively promising, what is needed in future work is a clear demonstration of the ability of L_1^{Q} to outperform baseline models which use a vectorial semantics but have no explicit pragmatic reasoning. Current evaluations, as discussed in section 2.5, have shown a human judgment preference for a baseline model.

Multidimensional projections The notion of projection, either in the set-theoretic or vectorial interpretation, maps a state of the world to a single aspect or dimension of that state. In principle, the projection can map to multiple dimensions. In the vectorial case, this amounts to a projection onto a plane (two dimensions) or a hyperplane ($n > 2$ dimensions).

This would allow the model to interpret a metaphor as carrying information along several dimensions of meaning at once. An interesting hypothesis is that metaphors are useful over and above literal language precisely because they enable a speaker to communicate along many dimensions simultaneously. This could be investigated for a model with multidimensional projections. In fact, the model proposed in this chapter is equipped to reason about multidimensional subspaces - the difficulty is that inference in this setting becomes much more expensive, because of the discretization of what currently is a single-dimensional subspace, but would become multidimensional. However, performing efficient quadrature for a low-dimensional integral is not, in theory, a prohibitively difficult task.

Metaphor production In some sense, S_1^{Q} can be taken as a model of metaphor production. However, a more accurate model is one which is capable of reasoning about a listener who is capable of metaphor interpretation. Such a model, S_2^{Q}, would be defined in terms of L_1^{Q}, and would produce utterances either in terms of a state w and projection q, or a state w alone (by marginalizing over q).

The importance of such a model is to investigate the contexts in which metaphorical language is useful, to underlie a system for metaphor production, and to provide a different form of testable prediction to L_1^{Q}.

Sentence embeddings Vectorial representations of words in context, and sentences, are an increasingly prevalent NLP tool in light of recent advances (Devlin et al., 2018; Peters et al., 2018).

A direction of future work which these representations make possible relates to the dependence of a metaphor's interpretation on context which, though a key feature of L_1^{Q}, is never directly exploited. That is, L_1^{Q} predicts that the interpretation of a metaphor should depend heavily on prior information about the target noun. For instance, example (9), *The dog is a rock*, has different meanings when said of a sleeping or heavy dog.

Contextualized word embeddings, where the vector corresponding to a word depends on its local context, provide a straightforward way to incorporate context, simply by taking the listener's prior to be a context dependent word vector, rather than a word vector in isolation.

Complex utterances One of the advantages of working with word vectors was the simplification of the utterance space U, to a relatively small finite set, placing the focus on the complexity of the state space W. The focus of chapters 3 and 4 is on more complex utterances spaces. Therefore, in chapter 5, I discuss the natural third step: combining the approaches introduced in this dissertation to form a single model with both a complex state space and a complex utterance space.

Chapter 3

Incremental Pragmatics

The model of incremental pragmatics discussed in this chapter is the product of joint work with Chris Potts and Noah Goodman, as published in (Cohn-Gordon et al., 2018b). Parts of the prose of that paper appears in this chapter.

In the applications of models of pragmatics considered so far, the utterances $u \in U$ have been atomic, in the sense of having no internal structure. In the example discussed in chapter 1.1.1, U consisted of two hand-chosen labels, while in the case of metaphor interpretation discussed in chapter 2.3, U consisted of an automatically chosen set of adjectives, viewed as "utterances" in the context of a noun they predicated.

By contrast, expressions[1] in natural languages have rich structure - indeed, this structure is a central object of study in linguistics. Consequently, the following questions arise when applying models of pragmatic reasoning to natural language:

- If, as a result of being recursively generated, the utterance set U consists of an infinite number of expressions, how do the models of pragmatics considered so far (S_1 and L_1) fare in such a setting? (*the question of unbounded utterances*)
- How do we explain inferences made during the interpretation of an utterance? (*the question of anticipatory implicatures*)

In answering both of these questions, it turns out to be useful to proceed *incrementally*, choosing each successive word pragmatically, rather than reasoning pragmatically on the level of the sentence as a whole. The

[1] Following chapter 1, we distinguish between expressions, which are linguistic objects, and utterances, which really designate speech acts, i.e. the action of uttering an expression. In theory, an utterance could be more general, corresponding to any action bearing semiotic significance, such as hand gestures, or even clothing choice, but in this dissertation the utterances correspond to the production of linguistic expressions. This distinction is particularly important in the present chapter, since an utterance, depending on whether a speaker model is global or incremental, may be a word or a sentence.

focus of this chapter is to propose just such a model of *incremental pragmatic reasoning*, which takes place at the level of words (or other segments) during the interpretation or production of an utterance. A speaker model which proceeds in this incremental way is capable of handling an unbounded set of utterances (a point which is developed in the context of natural language processing systems for image captioning and translation in chapter 4). Meanwhile, an incrementally pragmatic listener can draw implicatures before the completion of an utterance. I now describe these issues of production and interpretation in more detail, showing the difficulties they pose for the standard RSA model.

The problem with unbounded utterances The set of sentences available in a natural language is not finite[2]. Instead, recursively applicable rules generate an infinite set of sentences of potentially unbounded length (Chomsky, 1957). Suppose we attempt to use L_0, S_1, and L_1, or more complex variants of the same, as models of pragmatic reasoning in such a setting. Mathematically speaking, these models are still well defined. A literal listener L_0, armed with a compositional semantics, can handle any one of a recursively generated set of expressions. For S_1 given a state w, it makes sense to ask what utterance is most informative (even out of an infinite alternative set U) with respect to a literal listener, and for L_1, to ask what state S_1 intended to communicate.

However, actually calculating the posterior distribution of S_1 (i.e. performing inference) is now intractable. The reason is simply that S_1 includes a normalizing constant with a sum over U. Moreover, finding the *maximum a posteriori* utterance (that is, the utterance with the highest probability under the S_1 posterior) would require a search through every one of an infinite set. Since L_1 is defined in terms of S_1, it inherits this intractability.

This poses a challenge both for applying Bayesian models of pragmatics to natural language processing tasks and for their plausibility as a cognitive model of informative language production and interpretation.

The problem with anticipatory implicatures

Sedivy (2007) provides compelling empirical evidence that humans draw pragmatic inferences partway through utterances. For instance, when shown a scene with a tall cup, a tall pitcher, a short cup, and a key, a listener who hears "Give me the tall–" will fixate on the tall cup before the utterance is complete.

Intuitively, the reasoning the listener performs is clear: on hearing "Give me the tall–", the listener reasons that, had the speaker intended to refer to the pitcher, no modifier would have been needed (given the absence of a short pitcher), while if the tall cup had been the referent, *tall* would have been an informative modifier. As a consequence, the listener infers that the tall cup is the most likely intended referent.

A model which reasons about whole utterances like L_1 is incapable of deriving this inference, simply because it requires a full utterance, not a partial one.

[2] Or at least, the size of the set grows exponentially with the maximum sentence length.

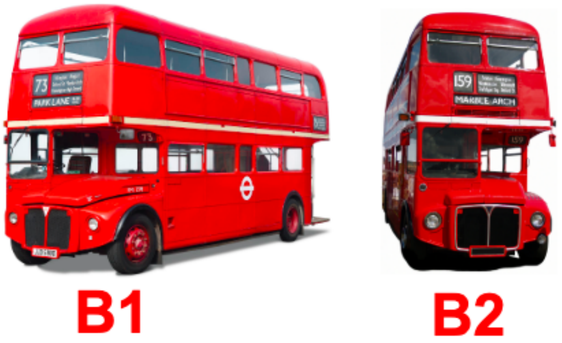


Figure 3.1: Two similar buses. An informative caption in this context must be very specific.

Subsampling an infinite utterance set One approach to the problem of unbounded utterances is to approximate U with a finite set of samples from U. For instance, a speaker might first identify a number of low cost semantically allowable utterances from an infinite U, and then treat this set as U for the purpose of pragmatic reasoning. Previous attempts to apply pragmatic reasoning to an infinite utterance space, for the purpose of designing NLP systems, have employed this method (Andreas and Klein, 2016b; Mao et al., 2016).

One problem with this approach is that it relies on the assumption that there is a reasonable probability of an utterance which is low cost and truthful also being pragmatically informative. As an example of where this assumption fails, consider a reference game in which the goal is to refer to B_1 in the context of B_2, as shown in figure 3.1. In this case, a description which is informative in the context, like *The 73 bus* or *The bus showing its left hand side*, might well be costly enough that it is very unlikely to be one of a sampled subset of U selected for reasons other than informativity. The consequence is that it would then not be an utterance S_1 could produce.

Subsampling also fails to provide an explanation of anticipatory implicatures, since it still takes the approach of treating utterances as atomic units.

Reasoning incrementally I propose a different approach. Rather than sampling a set of full expressions from U and then performing pragmatics, the speaker performs pragmatics *while* sampling an expression. Similarly, a listener performs pragmatics partway through interpreting an expression.

This approach relies on the assumption that the utterance set consists of expressions with recursive structure, so that the distribution over utterances can be decomposed into a product of simpler distributions. A simple decomposition (corresponding to a practically useful but linguistically unmotivated right branching recursive structure) is the one used in language models for NLP, where the distribution over word n of an utterance depends on words 1 to $(n-1)$. This is the case discussed in this chapter, although see section 3.4 for a discussion of how this approach could be extended to a structure which respected syntactic constitutents, like a probabilistic context free grammar (PCFG).

The advantage of this incremental approach is two fold. Firstly, it avoids the problem of an infinite utterance space, by reasoning about alternatives at the level of words, or subword segments, of which only a finite number exist. Secondly, it provides a means to calculate pragmatic inferences on the basis of an incomplete utterance, like "Give me the tall–". As such, it constitutes a more cognitively plausible theory of pragmatic reasoning than a global model, while preserving the core approach.

Aside from its usefulness as an NLP tool (the focus of chapter 4), it is important to consider the plausibility of incremental pragmatics as a linguistic theory: do humans in fact make use of this sort of incremental reasoning, either in production or interpretation? What predictions would an incremental theory of pragmatics make differently to a global one? What empirical evidence can be brought to bear on this claim?

The rest of the chapter is structured as follows. I first introduce an incremental model of pragmatics in formal detail. I then discuss its mathematical difference to a non-incremental, or *global* model. I then explain two natural language phenomena, anticipatory implicatures and cross linguistic variation in over-informative language (Rubio-Fernández, 2016), as an effect of this difference, in order to motivate incremental pragmatic reasoning as a linguistic theory. The first of these requires a model of language interpretation, while the second requires a model of language production. As such this pair of case studies serves to demonstrate both aspects of incremental pragmatics. As a further exploration, I present a comparison of the behavior of an incremental pragmatic speaker against human language production. Finally, I discuss several of the theoretical implications of incremental pragmatics, as relates to compositionality and a theory of alternatives.

3.1 Incremental models

Standard RSA models are global in the sense that the pragmatic reasoning is defined over complete utterances. Speakers are conditional distributions of the form $P(u|w)$, while listeners are of the form $P(w|u)$, for an utterance u and state w.

To avoid confusion, in this chapter and the next, I refer to global speakers and listeners with the superscript *SNT* for *sentence*, e.g. L_0^{SNT}.

The core idea of incremental pragmatics is to consider models of speakers which choose the next word u given a state w and a context c in the form of a previous sequence of words, and models of listeners which draw an inference about w given c and u. Precisely as in the sentence level models, these word level models are nested - in fact their definition is identical, save for the presence of a fixed sequence of previous words c. This is illustrated in figure 3.2.

While I will by default talk about word level incrementally in examples throughout this chapter, it is important to note that this is not a commitment of the model: the proposed approach could apply both to smaller units

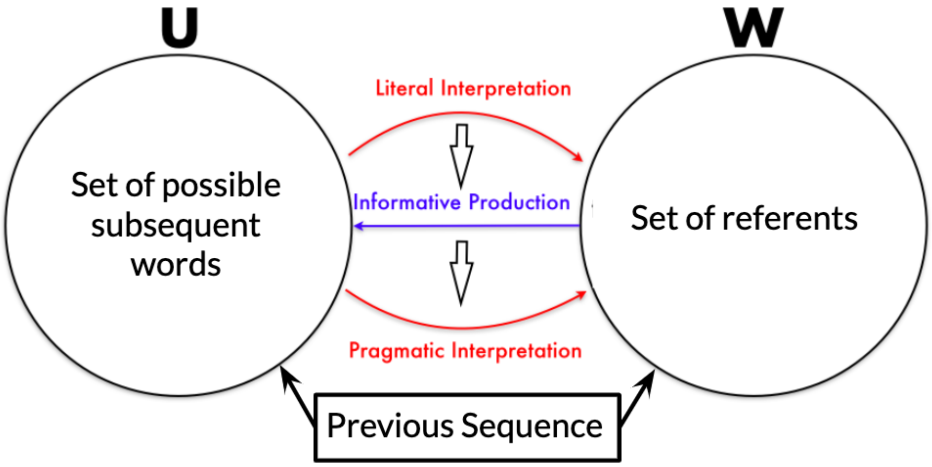


Figure 3.2: The crucial change from global to incremental RSA is to parametrize the model with a context, here a sequence of previous words (or segments), and have U be the set of possible subsequent words (segments) which could follow. In all other respects, the RSA model remains unchanged.

(e.g. sub-word segments, or even characters, as in chapter 4.2.3) and larger ones (sequences of sentences - see section 3.4).

The intuition behind these incremental models is that the choice of a word may be made by a speaker in order to be informative, rather than the choice of a whole sentence. Consequently, a listener may draw an inference without the received utterance being complete. I argue that this is the case for anticipatory implicatures.

Formally, the first step is to define word level analogues to the standard RSA agents, which I refer to as $L_0^{\text{WORD}}(w|u,c)$, $S_1^{\text{WORD}}(u|w,c)$, and $L_1^{\text{WORD}}(w|u,c)$. Note that in these models, utterances u correspond to *words* not sentences. c is a sequence of words. We start with L_0^{WORD}.

In this chapter, where the focus is on direct comparison between global and incremental models, the goal is to define L_0^{WORD} as similarly to L_0^{SNT} as possible. This motivates defining an incremental semantics $[\![u]\!](w) : U \rightarrow C \rightarrow [0,1]$ in terms of a global semantics $[\![u]\!](w) : U \rightarrow W \rightarrow \{0,1\}$ and a set of full expressions.

For any partial sequence c and set of states W, $[\![c]\!](w) \in [0,1]$ is the number of full expression extensions of c which are compatible with w divided by the total number of possible extensions of c into full expressions. Here, a possible extension is determined by a prespecified set of full expressions, but one could also envision a grammar determining which extensions are valid. Where c is a full expression, $[\![c]\!](w) \in \{0,1\}$ is as in the global model; where c is a partial expression, $[\![c]\!](w)$ is real-valued in the interval $[0,1]$, representing the biases created by c.

Note that the return value of the incremental semantics is the continuous interval $[0,1]$, rather than the discrete set $\{0,1\}$; it is a real number. Also note that the process to convert a global semantics into an incremental one requires a fixed set U of full expressions to be specified explicitly.

The advantage of this approach is that a direct comparison of both incremental and global models, on the basis of a single semantics, is possible. However, this move undermines one proposed goal of incrementality, namely to create a model of a speaker which does not have to reason about an infinite utterance set. In the case where an incremental semantics is defined in terms of a global one, even L_0^{WORD} is intractable to compute when U is infinite. For this reason, chapter 4 focuses on tractable variants of S_1^{WORD} and L_1^{WORD}, where the intractable conversion from a global semantics is avoided. The goal here, then, is a theoretical one: to understand the dynamics of incremental models of pragmatic reasoning, as they differ from their global counterparts.

With the definition of an incremental semantics in place, the equations for the incremental literal listener L_0^{WORD} and incremental pragmatic speaker and listener S_1^{WORD} and L_1^{WORD}, are straightforward:

$$L_0^{\text{WORD}}(w|c, \textit{word}) \propto [\![c + \textit{word}]\!](w) \cdot P_L(w) \tag{3.1}$$

$$S_1^{\text{WORD}}(\textit{word}|c, w) \propto L_0^{\text{WORD}}(w|c, \textit{word}) \cdot P_S(\textit{word}) \tag{3.2}$$

$$L_1^{\text{WORD}}(w|c, \textit{word}) \propto S_1^{\text{WORD}}(\textit{word}|c, w) \cdot P_L(w) \tag{3.3}$$

By contrast, compare to the definitions of the global literal listener, pragmatic speaker and pragmatic listener, introduced in chapter 1.1.1:

$$L_0^{\text{SNT}}(w|u) \propto [\![u]\!](w) \cdot P_L(w) \tag{3.4}$$

$$S_1^{\text{SNT-GP}}(u|w) \propto L_0^{\text{SNT}}(w|u) \cdot P_S(u) \tag{3.5}$$

$$L_1^{\text{SNT}}(w|u) \propto S_1^{\text{SNT-GP}}(u|w) \cdot P_L(w) \tag{3.6}$$

Note that the previously introduced global S_1 is now termed $S_1^{\text{SNT-GP}}$, to distinguish it not only from a word level speaker S_1^{WORD}, but from a sentence level speaker $S_1^{\text{SNT-IP}}$ defined in terms of S_1^{WORD}, which will be introduced shortly.

Figure 3.3 presents a running illustrative example, to understand the behavior of S_1^{WORD}. We imagine there are three referents, a red dress (R_1), a blue dress (R_2), and a red hat (R_3). We have a simple language composed of three utterances, *dress*, *red dress*, and *red object*, each with its expected semantics.

I represent the end of an expression as a STOP token, so that the choice of STOP as the next "word" represents the decision that the expression is complete. Costs are taken to be 0, unless specified otherwise; see chapter 1.1.2 for how to incorporate cost into S_1 models.

An important edge case There may be cases in which there is no possible true continuation of a sequence of words into a true utterance. For instance, no continuation of *red* constitutes a truthful description of

R_2. In such situations, we say that probability is evenly distributed over all choices of word, so that $S_1^{\text{WORD}}(dress|c = [red], w = R_2) = S_1^{\text{WORD}}(object|c = [red], w = R_2) = 0.5$. The reason to specify this case explicitly is to avoid a division by zero, which would otherwise occur here, since all choices have zero probability mass.

Behavior of word level pragmatic models Figure 3.3c summarizes the reasoning of the incremental pragmatic speaker S_1^{WORD}, assuming 0 cost on all words for simplicity. This agent prefers *red* as a first word when conveying R_1: $S_1^{\text{WORD}}(\textit{red}|c = [], w = \text{R}_1) = 0.57$. However, if R_3 is the intended referent, the agent *must* begin with *red* (since *hat* is not an available word in this simple example). As shown in figure 3.3d, this fact allows the pragmatic listener to infer from hearing *red* that the referent is most likely R_3: $L_1^{\text{WORD}}(\text{R}_3|c = [], \textit{red}) = 0.64$. This is closely related to the core dynamic underlying anticipatory implicatures, as discussed in section 3.2.2.

More generally, the behavior of S_1^{WORD} is to prefer informativity at each choice of subsequent word. It behaves just like $S_1^{\text{SNT-GP}}$ in this regard, but for each choice of the next word rather than the expression as a whole.

3.1.1 An unrolled incremental speaker

From the word level agent S_1^{WORD}, one can use the chain rule to obtain $S_1^{\text{SNT-IP}}$, a *sentence-level speaker*[3] whose values are the result of *incrementally pragmatic* inferences:[4]

$$S_1^{\text{SNT-IP}}(u|w) = \prod_{i=0}^{n-1} S_1^{\text{WORD}}(u_i|c = u[:i], w) \tag{3.7}$$

To sample an utterance from $S_1^{\text{SNT-IP}}$ given w, we choose the first word $word_1$ by sampling from $S_1^{\text{WORD}}(word|w = w, c = [])$, and this decision then becomes part of the context for sampling the second word from $S_1^{\text{WORD}}(word|w = r, c = [word_1])$. Similarly for the nth word. Whereas $S_1^{\text{SNT-GP}}$ in (3.5) makes pragmatic calculations on the basis of whole utterances, $S_1^{\text{SNT-IP}}$ makes incremental pragmatic decisions about each choice of word, which together also give rise to a distribution over utterances.

Importantly, $S_1^{\text{SNT-IP}}$ can plan ahead, in the sense of finding the sequence of words which maximizes the probability of S_1^{WORD} at each step. However, this planning does not involve pragmatics: there is never comparison of a sequence's informativity with the informativity of other sequences. In this sense, pragmatics only happens at the word level in $S_1^{\text{SNT-IP}}$, even though it is a sentence level model.

[3] In examples that follow, the "sentences" produced by $S_1^{\text{SNT-IP}}$ are really noun phrases, so rather than "sentence-level speaker" it is perhaps more appropriate to describe the model as a speaker which produces multiword phrases.

[4] I use $u[n]$ for the $(n-1)$th element of a list u, and $u[:n]$ for the sublist of u up to but not including $u[n]$.

$[\![\cdot]\!]$	R_1	R_2	R_3
dress	1	1	0
red dress	1	0	0
red object	1	0	1

cost	
dress	0
red dress	0
red object	0

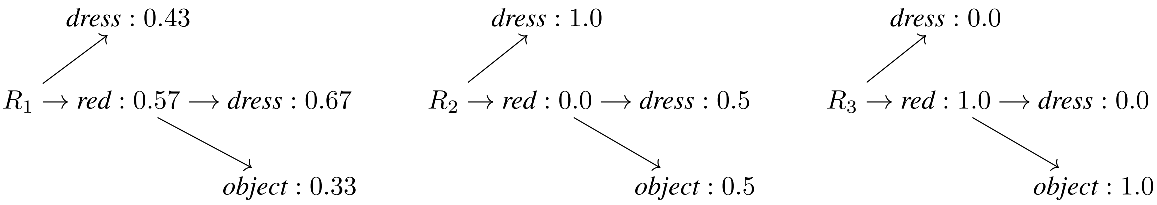
(a) Reference game.

L_0^{SNT}	R_1	R_2	R_3
dress	0.5	0.5	0.0
red dress	1.0	0.0	0.0
red object	0.5	0.0	0.5

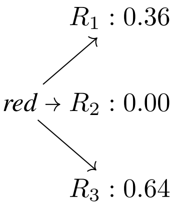
$S_1^{\text{SNT-GP}}$	*dress*	*red dress*	*red object*
R_1	0.25	0.5	0.25
R_2	1.0	0.0	0.0
R_3	0.0	0.0	1.0

L_1^{SNT}	R_1	R_2	R_3
dress	0.2	0.8	0
red dress	1.0	0.0	0.0
red object	0.2	0.0	0.8

(b) Global RSA.


dress : 0.43
R1 → red : 0.57 → dress : 0.67
object : 0.33
dress : 1.0
R2 → red : 0.0 → dress : 0.5
object : 0.5
dress : 0.0
R3 → red : 1.0 → dress : 0.0
object : 1.0

(c) Incremental RSA speaker predictions.



(d) Incremental RSA listener predictions upon hearing *red*.

$S_1^{\text{SNT-IP}}$	*dress*	*red dress*	*red object*
R_1	0.42	0.38	0.20
R_2	1.0	0.0	0.0
R_3	0.0	0.0	1.0

(e) Incremental utterance-level predictions from $S_1^{\text{SNT-IP}}$.

Figure 3.3: Illustrative example comparing global and incremental RSA. For ease of comparison to the global model, the STOP token for the incremental model is not depicted.

$$\begin{tikzcd}
full : sem \arrow[r, green] \arrow[d, leftrightarrow] & L_0^{\text{SNT}} \arrow[r, green] & S1^{\text{UTT-}\mathbf{X}} \\
inc : sem \arrow[r, red, dashed] & L_0^{\text{WORD}} \arrow[r, red, dashed] & S_1^{\text{WORD}} \arrow[u, red, dashed]
\end{tikzcd}$$

Figure 3.4: Two ways of constructing an utterance-level pragmatic speaker from a semantics. The solid green path is to obtain a literal listener over full utterances and then perform pragmatics, which gives rise to $S_1^{\text{SNT-GP}}$ while the dashed red path is to obtain an incremental literal listener, use it to construct a word-level pragmatic speaker from L_0^{WORD} and then use this to define an utterance-level pragmatic speaker, $S_1^{\text{SNT-IP}}$.

Efficient search for optimal utterances In the examples considered in this chapter, where the set of possible utterances is still finite, it is possible to compute to *maximum a posteriori* (MAP) utterance u given a state w for $S_1^{\text{SNT-IP}}$. This is simply the process of searching over all possible full utterances and choosing the one which maximizes the product of S_1^{WORD} at each step (*full search*). In infinite settings, or even settings with a large set of possible utterances, this is no longer possible. This is not a result of the pragmatic reasoning in $S_1^{\text{SNT-IP}}$, but merely of the fact that searching over the space of all sequences is in general intractable.

Chapter 4 discusses the sampling strategies such as *greedy search* and *beam search* needed in such cases. Insofar as $S_1^{\text{SNT-IP}}$ is a cognitively plausible model for real natural language, where the set of utterances is indeed infinite, it is reasonable to assume that some form of search other than full search must be taking place. However, for present purposes, I compare the behavior of $S_1^{\text{SNT-IP}}$ with full search to $S_1^{\text{SNT-GP}}$.

3.2 The consequences of incremental pragmatic reasoning

Importantly, while $S_1^{\text{SNT-GP}}$ and $S_1^{\text{SNT-IP}}$ are of the same *type*, in the sense of being conditional probability distributions over full utterances, they are not the same distribution. Another way to put this is to say that the operations of pragmatic reasoning and of unfolding a sentence word by word do not commute - their order matters. This raises two immediate questions: in ways ways do the predictions of the two models differ qualitatively, and are there empirical data which are better modeled by one over the other?

Core differences Figure 3.4 depicts the core relationships between the global and local models, focusing on the pragmatic speaker. The agents along the solid green path define the global model $S_1^{\text{SNT-GP}}$, while those along the dashed red path define the incremental model as given by $S_1^{\text{SNT-IP}}$.

The predictions of $S_1^{\text{SNT-IP}}$ for our illustrative example are given in figure 3.3e. Comparing them with the global pragmatic speaker predictions in figure 3.3b, we see that the two make substantively different predictions. In the global model, the speaker who wishes to refer to R_1 prefers *red dress*. In contrast, in the incremental model, the speaker referring to R_1 prefers *dress*. The reason for $S_1^{\text{SNT-IP}}$ having these values is that saying *dress* ensures the termination of the utterance (given the set of utterances that are available in this example), which therefore has probability 1.0 at the next time step, while saying *red* leaves two options, *dress* and *object*.

Thus, $S_1^{\text{SNT-GP}}$ and $S_1^{\text{SNT-IP}}$ are not only quantitatively different, but even differ in their predictions about which utterances are optimal. This particular case highlights the propensity of $S_1^{\text{SNT-IP}}$ to prefer utterances which at each step minimize the uncertainty at future steps.

Figure 3.5 provides an abstract example of the difference between $S_1^{\text{SNT-IP}}$ and $S_1^{\text{SNT-GP}}$. Moving from left to right, the numbers in green depict the S_1^{WORD} probabilities at each of the two steps in the generation of a complete two segment expression, when the target reference is W_1 instead of distractor W_2. The probability of the full utterance at $S_1^{\text{SNT-IP}}$ is the product of the two S_1^{WORD} steps.

When referring to W_1, $S_1^{\text{SNT-GP}}$ gives equal weight to *AA*, *BA* and *BB*. $S_1^{\text{SNT-IP}}$, however, first chooses between *A* and *B*: in this decision, *B* is preferred, since one of the two continuations of *A*, namely *AB*, is not compatible with W1. However, if *A* is chosen, the subsequent choice is fully determined to be *A*: ($p(A|[A], \mathrm{W1} = 1.0)$. This results in a preference for *AA*.

I now turn to an empirical question: do these differences in $S_1^{\text{SNT-IP}}$ from $S_1^{\text{SNT-GP}}$ bear any relation the data from natural language? I argue that they do, and that the incrementality of $S_1^{\text{SNT-IP}}$ allows it to explain data which cannot be explained by a global model.

3.2.1 Cross-linguistic discrepancies in over-informative language

It has been observed that, when generating referring expressions (REs), humans often provide more information than necessary to refer unambiguously (Engelhardt et al., 2006; Herrmann and Deutsch, 1976). For instance, Rubio-Fernández (2016) shows that English speakers often use redundant color terms (e.g., *the red dress*) in a scene with only a single dress, where the shorter utterance *dress* would suffice. However, Rubio-Fernández (2016) also notes that Spanish speakers are less likely to over-describe with the analogous referring expression, *el vestido rojo*, in the same situation. This difference is a challenge for non-incremental pragmatic accounts, since, *ceteris paribus*, we would expect semantically equivalent Spanish and English REs to have the same production probability.

Using incremental pragmatics, one can model an idealized version of the English case as follows: let the referents be a red dress (R_1) and a blue hat (R_2), and the possible full utterances be *dress*, *red dress*, *hat*, and *blue hat*, with the obvious semantics.

I make the following assumption regarding the *cost* term: assume a cost of 1.0 for all words but a cost of 0.0 for the STOP token. Further assume that a full expression's cost is the sum of the cost of its words. The effect of this cost term is to penalize longer expressions, all else being equal.

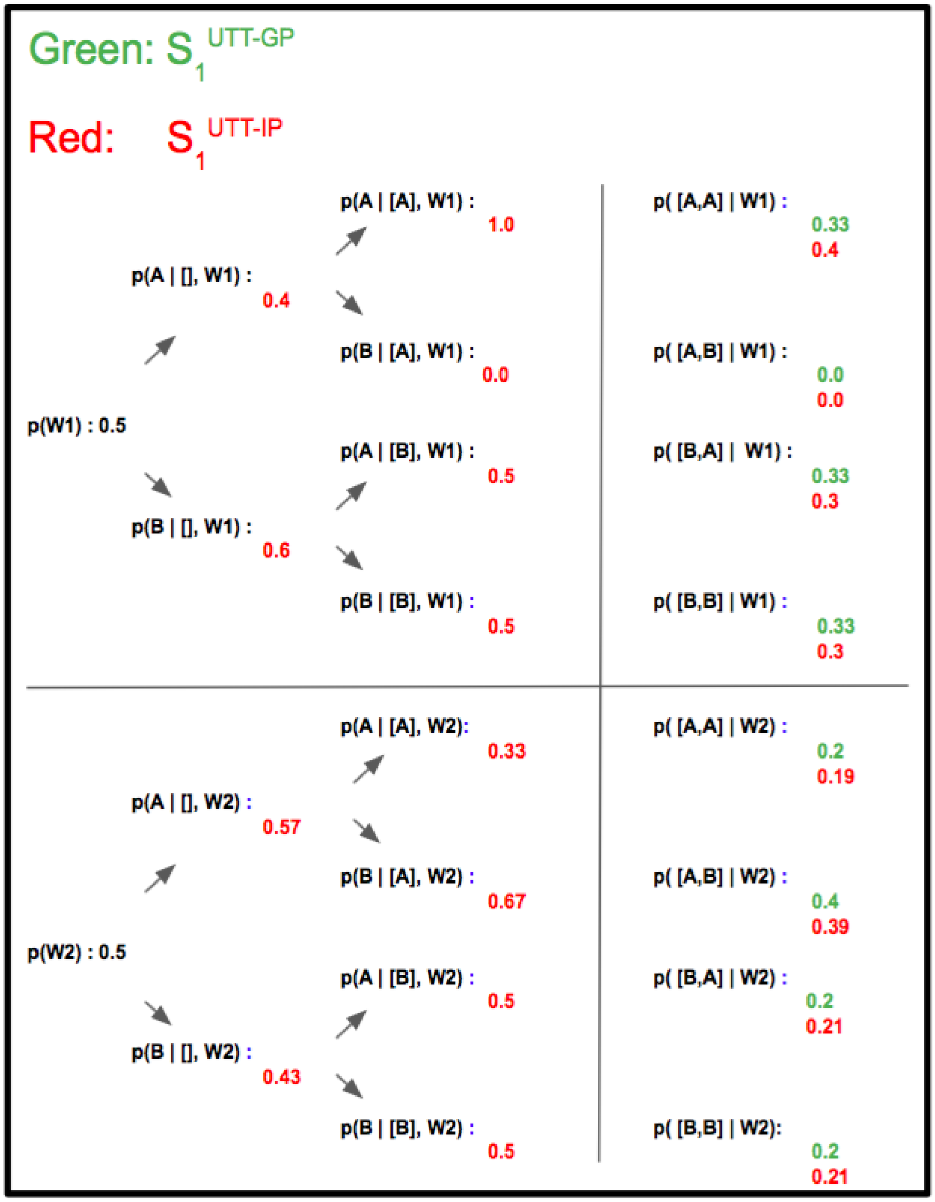


Figure 3.5: A depiction of the probabilities of the $S_1^{\text{SNT-IP}}$ in red and $S_1^{\text{SNT-GP}}$ in green, for a simple abstract example. The four full utterances are *AA*, *AB*, *BA* and *BB*, while the two worlds are W1 and W2. The semantics assigns u = W1 and w = *AB* to 0 but all other utterance–world pairs to 1. Given a world w, the incremental speaker first chooses the first letter to be *A* or *B*, and then chooses the second letter conditioned jointly on w and the first letter, to obtain a full utterance. The resulting full utterance probabilities are compared with the predictions of the global $S_1^{\text{SNT-GP}}$. As can be seen, the incremental and global speakers assign different probabilities to each utterance and are consequently distinct from each other.

On these assumptions, the globally pragmatic speaker $S_1^{\text{SNT-GP}}$ prefers *dress* to *red dress*, since both are fully informative but the latter is costlier: $S_1^{\text{SNT-GP}}(\textit{dress}|R_1) = 0.73 > S_1^{\text{SNT-GP}}(\textit{red dress}|R_1) = 0.27$. Meanwhile the incremental pragmatic speaker $S_1^{\text{SNT-IP}}$ is undecided: $S_1^{\text{SNT-IP}}(\textit{dress}|R_1) = S_1^{\text{SNT-IP}}(\textit{red dress}|R_1) = 0.5$. The increase in mass on the over-informative RE *red dress* in $S_1^{\text{SNT-IP}}$ as compared to $S_1^{\text{SNT-GP}}$ is the result of incremental processing: the decision between *red* and *dress* is made on the basis of informativity, and both words are equally informative. However, if *red* is chosen, the subsequent, now over-informative word *dress* has to follow, since *red* on its own is not a full expression.

I explore the generality of this dynamic – that incremental pragmatics may lead to the language model being compelled to produce longer utterances – in section (3.3), where I apply the model to real-world data, in the form of the TUNA corpus.

However, this effect does not obtain in Spanish, where adjectives are post-nominal. In the Spanish case, let our utterances be *vestido*, *vestido rojo*, *sombrero*, and *sombrero azul*, with the same referents and costs as before. Then there is no difference between the global and incremental models:

$$S_1^{\text{SNT-GP}}(\textit{vestido}|R_1) = 0.73 > S_1^{\text{SNT-GP}}(\textit{vestido rojo}|R_1) = 0.27$$
$$S_1^{\text{SNT-IP}}(\textit{vestido}|R_1) = 0.73 > S_1^{\text{SNT-IP}}(\textit{vestido rojo}|R_1) = 0.27$$

When choosing the word to follow *vestido*, the incremental pragmatic speaker has no need to say *rojo* rather than STOP, since the goal of communicating the referent has already been completed by *vestido*. As a result, the speaker chooses the less costly option, STOP. The relevant difference here from the English case is that it is grammatical to stop after the first word (since the first word is a noun, not an adjective as in English).

A qualitative property which this example illustrates is a dislike in $S_1^{\text{SNT-IP}}$ for expressions which begin with a sequence of words which would mislead the incremental literal listener L_0^{WORD}. This is the basis on which anticipatory implicatures are formed, as discussed in section 3.2.2. This is not a hard constraint: utterances which would initially mislead an incremental listener are not categorically ruled out. However, the question of whether this behavior is empirically justified is a worthwhile topic for future investigation.

3.2.2 Anticipatory implicatures

The case of anticipatory implicatures introduced by Sedivy (2007) involve a listener making a pragmatic inference on the basis of a partial utterance. In particular, “Pass me the tall –” triggers an inference that the referent has a short counterpart, motivating the speaker to have used the modifier *tall*.

I take this as motivation for a model capable of calculating implicatures based on partial utterances. In particular, I propose the incremental pragmatic listener L_1^{WORD} is suitable for this task, which can calculate an implicature by reasoning that, had the speaker intended to refer to the pitcher, they would not have had

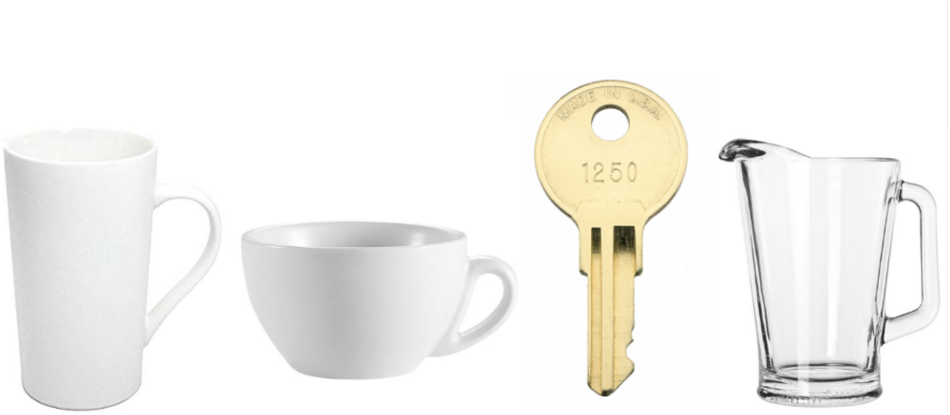


Figure 3.6

any motivation to say “tall”. By contrast, on the assumption that the speaker’s referent is the tall cup, the contrastive modifier serves to distinguish the intended referent from the short cup.

To model this implicature formally, I make the simplifying assumption that the possible full utterances are *tall cup*, *short cup*, *tall pitcher*, *cup*, *pitcher*, and *key*, with the referents shown in figure 3.6. For consistency with the previous example, assume the additive $cost$ function from section 3.2.1.

On hearing *tall* as the first word of an utterance, L_1^{WORD}, the incremental pragmatic listener, can draw the following inference: the intended referent is likely to have been the tall cup, since had it been the tall pitcher, there would have been no need to use the contrastive modifier *tall*: $L_1^{\text{WORD}}(\text{the pitcher}|c = [], \textit{tall}) = 0.4$ while $L_1^{\text{WORD}}(\text{the tall glass}|c = [], \textit{tall}) = 0.6$.

This implicature is cancelable, and indeed, were the next word to be *pitcher*, we would exclude all referents but the pitcher. In this respect, the model’s inference represents the confusion created by uttering “tall pitcher”, where after the first word of the utterance, the majority of probability mass is on a referent (*the tall cup*) which, after the second word (*pitcher*), has no probability mass.

3.3 Comparison to human behavior

In order to observe the behavior of the incremental pragmatic model on real data, I make use of the TUNA corpus (van Deemter et al., 2006). TUNA is built around a referring expression task grounded in images. The images are coded using a fixed set of attributes, and the human-produced utterances are coded using the same attributes. Thus, TUNA lets us study the core content of naturally produced referring expressions without the necessity of confronting the full complexity of natural language.

The goal is to show that, when a cost is imposed which prefers shorter utterances, the incremental model

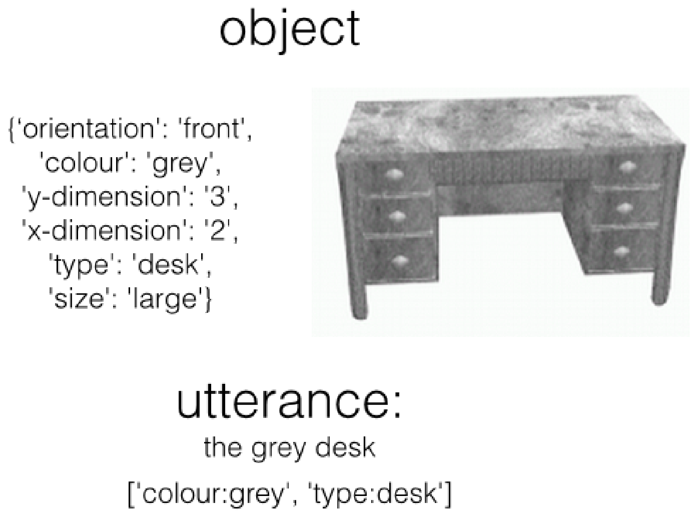


Figure 3.7: An example target entity from the *furniture* domain, along with its coding as a dictionary, and the human generated referring expression for it in a context of other images.

$S_1^{\text{SNT-IP}}$ is less affected, and on average produces more two-word utterances than $S_1^{\text{SNT-GP}}$.

I hypothesize this on the basis of the preference of $S_1^{\text{SNT-IP}}$ for utterances where the choice of each word is made with high certainty, as discussed in section 3.2. This means that informative one-word utterances which have reasonable probability of being extended with a second word will score lower than two-word utterances where the choice of the first word largely determines the choice of the second. Since most one-word utterances admit the possibility of an extension to a second word, this would result in a preference for the longer, two-word utterances. This would provide further evidence that the propensity for overinformativity described in section 3.2.1 is indeed an effect of the $S_1^{\text{SNT-IP}}$ model on less idealized data.

Data The TUNA corpus defines a reference game in the sense of figure 3.3a. Each *trial* contains a set of images (entities), of which one or several are the target, and a human-generated referring expression for the target in the context of all the images. I refer to the full set of target and non-target entities as the *context set*.

Both images and utterances are coded as sets of attributes (figure 3.7). This coding defines a semantics. For instance, in figure 3.7, the utterance “the grey desk” is true of the entity, since *type:desk* and *colour:grey* are included in its attributes. For the *furniture* domain, attributes such as color, object type, and size are coded. The *people* domain is more complex, coding for more attributes, including age, clothing, hair color, glasses, and orientation. Both domains also code for the position of the image relative to the other images in the context set.

3.3.1 Methods

For simplicity, I restrict the model to the subset of the *furniture* and *people* domains where only a single referent is provided, and consider only utterances of two words or fewer. These constitute 32% of the total utterances in the single referent corpora, and to our knowledge are not distinct in other ways than their length.

For each trial, the possible utterances are those from the set of all two-word utterances across the entire corpus (either of furniture or people) which are compatible with at least one of the entities in the trial. I calculate the set of optimal utterances (since there may be more than one utterance with maximum probability) for both $S_1^{\text{SNT-GP}}$ and $S_1^{\text{SNT-IP}}$. For our cost function, we assume all words have a cost of 1.0 except the STOP token, which has cost 0.0. Utterances cost the sum of their words. This has the effect of penalizing longer utterances.

For each trial, we have a set of entities as referents, with the designated target identified among these entities. In addition, we can define the set of all possible true utterances for a given trial. Thus, it is possible to make predictions according to both $S_1^{\text{SNT-GP}}$ and $S_1^{\text{SNT-IP}}$ for each trial without having to enrich the TUNA dataset in any way.

3.3.2 Results

As expected, a preference is found for longer utterances; out of the 114 *people* trials, $S_1^{\text{SNT-GP}}$ identifies 120 two-word utterances as optimal, compared to 287 for $S_1^{\text{SNT-IP}}$. In the 83 trials of the *furniture* domain, $S_1^{\text{SNT-GP}}$ marks 88 two word utterances as optimal, compared to 149 for $S_1^{\text{SNT-IP}}$. (More than one utterance may be optimal for a given trial, in the event that multiple utterances have the same, maximal probability of being chosen.)

An example of a representative case is the trial where the entity in figure 3.7 is the target, and no other distractors are grey, although others are desks. In this case, both "grey" and "a grey desk" are fully informative, in the sense of only being compatible with the target. With the cost term having the effect of penalizing longer utterances, $S_1^{\text{SNT-GP}}$ chooses "grey" as optimal. For $S_1^{\text{SNT-IP}}$, however, neither of these utterances are optimal, because probability is divided between stopping after "grey" and continuing with "desk". Instead, the optimal utterance, "right middle", describes the position of the target among the images of the context set[5]. "Right" is not an available full utterance (as it is not attested in the data) and so no probability mass is lost on the possibility of stopping at this point in the production of the utterance.

While this result offers a possible motivation for over-informative behavior, the nature of the relation between the observed behavior of the model and overinformativity is not straightforward and merits further work –

[5] Note that "right middle" counts as a full utterance in this dataset, as it is produced as a referring expression by human annotators.

for one thing, two-word utterances are not always more informative than one-word utterances. In particular, it would be desirable to use a more direct proxy for over-informativity than preference for longer utterances.

3.4 Discussion

The core proposal of this chapter was a model of incremental pragmatic reasoning, consisting of S_1^{WORD} and L_1^{WORD}. The former is a model of a speaker who tries to maximize informativity at each choice of word or segment during the generation of an utterance, while the latter is a model of a listener who, having heard a partial utterance, assumes that the next word of it was generated by the informative speaker S_1^{WORD}.

The broader idea on offer is of pragmatic reasoning performed during a recursive process, here the unrolling of a sentence word by word. I now discuss two ways in which the core notion of incremental pragmatic reasoning could be differently applied.

Discourse level pragmatics The apparatus of pragmatic inferences made incrementally given a context c, could be adapted to discourse level phenomena. For instance, consider the situation of a speaker who produces a speech. To the extent that such a speaker is pragmatic, it is clearly not by considering the whole space of possible speeches, and choosing the most informative one. A much more reasonable proposal is that the speaker chooses each sentence pragmatically, or even each phrase or word.

In this sense, even a sentence level model of pragmatic reasoning, like $S_1^{\text{SNT-GP}}$, implicitly concedes that pragmatics is incremental on the level of sentences (although no mechanism for multisentence language production is specified), presumably with greedy unrolling, so that a speaker chooses what to say sentence by sentence.

A direction of future research is to investigate whether the incremental model $S_1^{\text{SNT-IP}}$, when applied at the sentence level over the course of a multiple sentence utterance, acts in accordance with human behavior. For instance, a consequence of the incremental definition of $S_1^{\text{SNT-IP}}$ is a preference for putting the most informative information early. Is this borne out in the setting of multi-sentence utterances?

Pragmatic compositionality As well as applying the incremental approach to smaller or larger units than words, a possible extension is to more complex recursive structure.

Abstractly, the incremental approach to pragmatics presented in this chapter exploits the recursive structure of lists, where the probability of the nth item is a function of the previous $n-1$ items. Rather than unfolding an utterance (or rather, a distribution over utterances), and then performing pragmatic reasoning, the approach is to perform pragmatic reasoning *inside* the unfolding (or in the case of interpretation, folding) of an utterance.

There is a sense, therefore, in which this approach is compositional: pragmatic meanings are deriving for subparts of an expression, and then assembled into a whole. Of course, compositionality in language, at least in the standard sense of semantic compositionality, follows the recursive structure given by a syntax, which is richer than the list recursion used here.

However, there is no reason that exactly the same strategy, of calculating pragmatic enrichments during a recursive process, could not be applied in a more complex case. As such, an avenue of future work is to investigate the behavior of a comparable model to $S_1^{\text{SNT-IP}}$ for a set of utterances defined with a probabilistic context free grammar (Jelinek et al., 1992), which more closely resembles the constituent structure of natural language. In this case, inferences would be computed at each node, e.g. during the production or interpretation of a *VP* node.

Chapter 4

Informative Language Generation

The work on image captioning discussed in this chapter is the product of joint work with Chris Potts and Noah Goodman, as published in (Cohn-Gordon et al., 2018a). The work on translation is the product of joint work with Noah Goodman, as published in (Cohn-Gordon and Goodman, 2019). Parts of the prose of those two papers appears in this chapter.

Chapter 3 focused on idealized interpretations of S_1^{WORD} and L_1^{WORD}, and the linguistic relevance of the differences between incremental and global models of pragmatic reasoning. By contrast, the focus of the present chapter is the application of Bayesian models of pragmatics to computational tasks involving natural language generation.

Recent years have seen a significant improvement in the quality of statistical models for AI tasks. In particular, *deep* architectures have been used to great success in both vision and language (LeCun et al., 2015). Combining statistical models of natural language and vision with the RSA framework for pragmatic reasoning is an appealing prospect, since it promises a way to utilize the dynamics of pragmatic reasoning in real world settings. On the assumption that RSA models accurately capture human behavior, this is a means to improve AI systems for language generation and understanding, in a way which makes use of an interpretable model.

Thematically, the approach here parallels chapter 2. The goal is to use a statistical model as a semantics on the basis of which pragmatic reasoning can be performed. The difference here is that rather than using word embeddings as in chapter 2, the present focus is on neural models of grounded language generation, i.e. models which take the form $P(u|w)$ for an utterance u and a state of some kind w. This encompasses the two tasks discussed in detail in this chapter, image captioning and translation. For image captioning, w is an image, and u is a caption which describes that image. For translation, w is a source language sentence, and u is a translation into a target language.

Reference Games, Meaning and Natural Language Generation

In the context of the RSA framework, a speaker model is a conditional probability distribution $P(u|w)$ which can be interpreted as a system which takes a state w and (stochastically) produces an utterance u. A state is to be understood as a way that the world can be (or more precisely, an equivalence class of ways the world can be).

Recall also that a listener model takes an utterance u and stochastically produces a state, or equivalently: deterministically produces a distribution over states. In chapter 1.2.1, we equated this output distribution with the meaning of u.

These ideas cohere naturally with the task of image captioning. We can think of images as representations of the state of the world. A captioner is then a system which translates the information inherent in that state into a natural language expression.

A similar idea applies to translation. Rather than representing the state of the world with an image, we represent it with a natural language sentence. For example, we can think of a translation system which takes the sentence “It is raining.” as producing a (French) natural language expression to express the state of the world represented by this English sentence.

Informativity While RSA models are capable of capturing a range of behaviors, both in the production and interpretation of language, I focus here on one of the simpler dynamics of the framework, namely the preference for informativity of S_1, or in Gricean terms, the adherence to the maxim of Quantity. As such, the approach I discuss is to instantiate S_1 in the context of a neural language production model, and to use it as a means of compelling that system to be informative.

Here, a pragmatically informative speaker is one which produces an utterance — out of an unlimited set of utterances — which not only describes the state of the world, but takes into account their interlocutor’s beliefs about what other states in W are possible. The thesis here is that this behavior is eminently desirable in real-world systems, and imitative of human behavior. The reason for this is that a lack of informativity has turned out to be a failing of language generation systems on a variety of tasks, such as translation (Li and Jurafsky, 2016) and dialog (Jiang and de Rijke, 2018), where systems often produce generic uninformative language.

The challenge of an infinite utterance space The set of possible utterances in natural language is at least countably infinite (Chomsky, 1957; Langendoen et al., 1984). When U is an infinite set of sentences, S_1, or more specifically $S_1^{\text{SNT-GP}}$ (as introduced in chapter 3) is intractable to compute. This is a practical concern, but also a conceptual and cognitive one: how do humans reason over possible alternatives when the set of utterances that could be chosen as alternatives is unbounded? Said differently, how to we upgrade contrived models of pragmatic reasoning in which U is small and finite to a setting where U can be any sentence?

For image captioning and translation (as well as other natural language generation tasks), the models that result from training typical deep learning systems (see section 4.1) can produce any sequence of words (or in some cases, characters). As such, they provide a good setting to explore this problem, by taking U to be *the set of all possible sequences of words*, and using the distribution produced by a neurally trained captioning or translation model over these sequences as our semantics.

What was introduced as a theoretically appealing model in chapter 2, namely the incrementally pragmatic speaker $S_1^{\text{SNT-IP}}$, serves the purpose of providing a tractable algorithm for informative language production in natural language generation tasks. The reason for this algorithmic advantage is that $S_1^{\text{SNT-IP}}$ reasons about a space of alternatives consisting of the next generated segment, which in the case of word level incremental pragmatics has the size of the vocabulary.

I now introduce the neural architectures relevant to the approach of this chapter (section 4.1), the application of Bayesian pragmatics to those architectures (section 4.2), and evaluations on the tasks of image captioning and translation (sections 4.3 and 4.4).

For the latter, I introduce an extension in which the use of both a neural speaker and listener model allows for an informative speaker to be defined without the explicit specification of a set of states (section 4.5). This returns us to the question raised in the previous chapter, namely: can we find ways to specify a state space (and a prior over that state space) automatically?

4.1 Sequence models

Many NLP tasks require a system to produce text, consisting of a sequence of words, subwords or characters, conditional on some input. For image captioning (Farhadi et al., 2010; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015), the input is an image, with the sequence output being a description of that image. For translation (Bahdanau et al., 2014; Gehring et al., 2016; Vaswani et al., 2017), the input is also a sequence, but in a different language to the output. For summary (Rush et al., 2015), both the input and output are in the same language.

Deep learning Recent deep learning approaches, which have yielded unprecedented success at a range of visual and linguistic tasks train an *end-to-end* architecture directly from input to output. For example, an image captioning architecture takes an image, represented as a vector, performs a number of operations to arrive at an encoding of the image, and a further series to translate that encoding into a prediction about a sequence of words.

These operations are parametrized by a vector θ, which is learned from data[1]. More concretely, one wants to find a value of θ for which a loss function, roughly corresponding to the probability of producing the output sequence from the input sequence or caption, is low on average for a large *training dataset*. Using stochastic gradient descent, one can obtain a value of θ which is a local minimum for the loss function. The procedure is similar for translation and captioning, although the details of the architecture (i.e. the function mapping from inputs to outputs) differs. What is important is that all the operations in the architecture are differentiable, so that gradients can be easily calculated. Tools for computing this gradient automatically (Abadi et al., 2016; Paszke et al., 2017) allow for the rapid prototyping of a wide variety of models, and have contributed to the speed of the development of deep learning in recent years.

Sequential decoding Once trained (i.e. once a value of θ has been obtained on the training dataset), the model can be used to make predictions. A feature of language generation models, ranging from recurrent neural networks (Mikolov et al., 2010), to convolutional networks (Kalchbrenner et al., 2014) to transformers (Vaswani et al., 2017), is that their predictions, at decoding time, amount to a probability distribution over the next token, given an input (such as an image or source sentence) and a previous sequence of tokens. Here, a token could be a word, subword or character, depending of the architecture of the model in question. As in chapter 3, I use words as the default token when describing these models. This assumption that a sentence should be generated left to right, token by token, in not a necessity, but is common in practice.

What this means, is that a trained image captioning or translation model, across practically all commonly used architectures, amounts to a distribution $S_0^{\text{WORD}}(token|w,[token])$, where $[token]$ is the sequence of previous tokens in the sentence generated so far. I refer to this distribution as S_0^{WORD} to suggest a connection to the RSA framework which will shortly be exploited, namely that a trained neural language generation model can be viewed as an RSA speaker model. Since this speaker incorporates no explicit pragmatic reasoning, it can be viewed as a "literal" speaker, and receives the 0 index accordingly.

From token level to sentence level distributions S_0^{WORD} yields a distribution S_0^{SNT} over full sentences, similar to what was introduced in chapter 3. Python list indexing conventions are used here, with "+" meaning concatenation of list to token or list of tokens, and zero-indexing assumed:

$$S_0^{\text{SNT}}(u|w,k) = \prod_{i=0}^{n-1} S_0^{\text{WORD}}(u = u[i]|w = w, c = k + u[:i]) \tag{4.1}$$

Here k is a sequence of tokens produced already (generalizing from 3.7 in chapter 3.1 where k is assumed to be the empty list). S_0^{SNT} is then a distribution over sequences u which extend k (but note that our convention here is that u does not contain c).

[1] In practice, θ is a set of multidimensional arrays, but can be treated as a single vector, consisting of all the parameters concatenated.

S_0^{SNT} returns a distribution over full target language sentence continuations. In what follows, we omit *k* when it is empty, so that $S_0^{\mathrm{SNT}}(u|w)$ is shorthand for $S_0^{\mathrm{SNT}}(u = u|w = w, k = [])$. Informally, we will refer to the transformation that yields the distribution S_0^{SNT} from S_0^{WORD} as the *unrolling* of S_0^{SNT}[2].

Importantly, it is not obviously the case that language quality (as measured by human judgment) and probability of a sequence under S_0^{SNT} are linearly correlated. Indeed, for the case of text generation from language models, Holtzman et al. (2019) find that very high probability sequences are often very linguistically deficient, although these high probability sequences are rare (among all possible sequences).

Thus, instead of obtaining a literal speaker or listener model through a handwritten semantics, we rely on a pretrained neural model S_0^{SNT}. Note that this model will assign some probability to untrue utterances as well as ungrammatical sequences of words, and perhaps encode some preference for informative or salient utterances already, unlike an S_0^{SNT} model defined in terms of a hand constructed semantics. For example, an image captioning system will assign some probability (even if very little) to “This is a dog” given an image of a red bus, and may already prefer “This is a red bus” to “This is a bus” even without explicit pragmatic reasoning.

Model architectures For image captioning, a standard architecture is a convolutional neural network, following by a recurrent network such as an LSTM (Gers et al., 1999). Often attention is used (Xu et al., 2015), which provides a mechanism for determining which part of the image to focus on at each point in the process of producing the caption.

BiLSTMs with attention (Bahdanau et al., 2014), and more recently CNNs (Gehring et al., 2016) and entirely attention based models (Vaswani et al., 2017) constitute the state-of-the-art architectures in neural machine translation.

For our purposes, the details of these architectures are irrelevant. The pragmatic reasoning takes place once the model is already trained, at which point it is simply a black box conditional probability distribution over the next token, given a sequence of previous tokens and a state.

4.2 Bayesian pragmatics in the context of neural language generation

Up to this point, we have discussed S_0^{WORD} and S_0^{SNT} in a way which was deliberately agnostic as to whether the underlying model was for translation, image captioning, or another grounded natural language generation task. We do this to emphasize that the approach taken in this chapter applies to all such tasks. For the purposes of motivating the use of pragmatics in these settings, we now consider two concrete examples.

[2] This unrolling can be understood as a list *unfold* in the sense of functional programming, although this approach was not taken in the Python implementation used here.

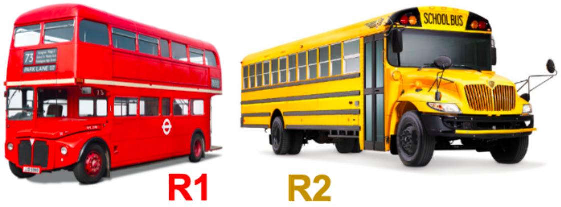


Figure 4.1: Two images, of a red London bus and a yellow school bus.

Informative image captioning In order to perform image captioning well, a computational system must have information about how expressions in natural language (represented as lists of words or even characters) correspond to images (represented as arrays of pixels). Recent work has given rise to systems which perform well at this task (in the sense of providing accurate captions for unseen images) in the form of end-to-end neural architectures (see section 4.1), trained on large datasets of pairs of images and captions (Lin et al., 2014; Krishna et al., 2017).

Good captions ought not only be true, but informative. For example, a caption of a picture of a horse in a field should not be: "Blades of grass", since this fails to distinguish it from images of grass alone. Neural image captioning models perform well with respect to producing true captions but often less well with respect to informativity. In other words, the captions they generate are often too generic, omitting important details. This is made apparent when two images which are different in important ways receive the same caption. As an idealized example, suppose that *A bus* was the caption for both images in figure 4.1.

The goal of this task is to take a set of images, of which one is the target and the rest are distractors, and return a caption which unambiguously identifies the target and not the distractors. This can be understood as an instance of referential expression generation, which suggests a relation between this task and the task performed by an S_1 informative speaker model.

For instance, consider R_1 and R_2 displayed in Figure 4.1. If the task of the captioning model is to refer to R_1 unambiguously in the context of R_1 and R_2, the utterance *This is a bus.* would be uninformative.

This behavior would amount to a failure to distinguish between R_1 and R_2 in a reference game, where the goal of the speaker was to communicate their target, say R_1. The RSA informative speaker S_1 is designed precisely to succeed at this reference game. As such, one imagines that if we could use a trained neural image captioner S_0 as the basis for a pragmatic image captioner S_1, we would have a system which would describe features of the target image which differed from the other images in W.

The result will be that the preferred utterances of the S_1 for R_1 will mention aspects in which R_1 differs from R_2, insofar as these aspects can be detected by the convolutional neural net being used. This should result in a system with the semantic power of a deep learning architecture, but the ability to be strategic in language use of a Bayesian pragmatic model. For example, an S_1 model would prefer utterances which were not only true but also informative, and so an utterance like *A red bus.* would be preferred as a caption for R_1 in the context of R_2 (that is, when $W = \{R_1, R_2\}$).

Informative translation Very similar considerations arise for translation as do for image captioning. Here, many sentences which differ significantly in meaning are translated, by state of the art systems, to the same target language sentence.

For instance, “I cut my finger.” and “I cut my finger off.” describe different states of the world, but are translated, under state-of-the-art systems[3] to a single French sentence: “Je me suis coupé le doigt.”

Again, one can envision the use of S_1 built in terms of a neural translation model which produces coherent translations for a target sentence (e.g. “I cut my finger.”) which are distinct from translations for any distractor sentence (e.g. “I cut my finger off.”).

With the two cases in mind as motivation, I now describe the process of incorporating pragmatic reasoning, in particular, informativity, into a neurally trained language generation system.

In our previous examples of RSA models, the nesting has begun with a listener L_1^{SNT} with an explicit semantics. However, with some simple alterations, a similar RSA model can be built on top of S_0^{SNT}. To do so, we define $S_1^{\text{SNT-GP}}$ in two steps, first defining L_1^{SNT} in terms of S_0^{SNT} and then $S_1^{\text{SNT-GP}}$ in terms of L_1^{SNT}. Note that we refer to this version of S_1 as $S_1^{\text{SNT-GP}}$, in accordance with the naming scheme introduced in chapter 3.

(19) $L_1^{\text{SNT}}(w|u) \propto S_0^{\text{SNT}}(u|w) \cdot P_L(w)$

(20) $S_1^{\text{SNT-GP}}(u|w) \propto S_0^{\text{SNT}}(u|w) \cdot L_1^{\text{SNT}}(w|u)^{\alpha}$

One noteworthy difference between equation (20) and the vanilla RSA S_1 of equation (3), introduced in chapter 1 is that now, the speaker’s prior over utterances is supplied by S_0^{SNT} conditioned on w, whereas before, this prior was determined by a separate distribution $P_S(u)$ not dependent on w and by default assumed to be uniform. This is a design choice which encourages $S_1^{\text{SNT-GP}}$ to produce language similar to S_0^{SNT}. It is very similar to the solution used by Vedantam et al. (2017), which employs a weighted sum of an informative (there termed introspective) and literal speaker. This weighting amounts to a prior, similar to the Bayesian interpretation of a regularization term as a prior.

The motivation for using S_0^{SNT} as a prior is that, when the prior is uniform over U, the model tends to produce ill-formed language. This is observed in figure 4 of (Vedantam et al., 2017) (noting that a uniform prior in the Bayesian setting is equivalent to placing all the weight on the informative speaker.) To see why this is, note that the term L_1^{SNT} in the definition of $S_1^{\text{SNT-GP}}$ causes the model to favor captions which are more likely to be produced by the target image than the distractors, regardless of their truth. So for example, the caption

[3] Both Google Translate and Fairseq’s pretrained English-French model exhibit this property.

A red brick is an untrue caption for a red bus, but is more likely to be produced for a red bus than a yellow one. Even worse, if a nonsensical sequence of characters is more likely to be produced for the target than the distractors, it too will score highly under L_1^{SNT}. Thus, using S_0^{SNT} as a prior constrains the behavior of the system.

Another possibility is to use a language model $P_{langmod}(u)$ as a prior, rather than S_0^{SNT}. This helps rule out ungrammatical language, but does not resolve the issue of the model generating false but distinctive captions. However, in section 4.3.2, I suggest that the real cause of untruthful captions is an impoverished space of possible states (here, images).

Another feature to note of equation (20) is the presence of α (introduced in chapter 1.1.2), which controls the degree of informativity of the model. As α increases, $S_1^{\text{SNT-GP}}$'s distribution becomes more concentrated on the utterance which maximizes $L_1^{\text{SNT}}(w|u)$. We can therefore view it as determining the degree to which the model cares about being informative.

We first illustrate the behavior of $S_1^{\text{SNT-GP}}$ with an idealized example where S_0^{SNT} is hand-specified, and only two states and utterances are considered. These states are R_1 and R_2, with utterances *bus* and *red bus*

- $W = \{R_1, R_2\}$
- $P(w)$: uniform distribution over W
- $U = \{\textit{bus}, \textit{red bus}\}$
- S_0^{SNT} :
  - $S_0^{\text{SNT}}(\textit{bus}|R_1) = 0.5$
  - $S_0^{\text{SNT}}(\textit{red bus}|R_1) = 0.5$
  - $S_0^{\text{SNT}}(\textit{bus}|R_2) = 1.0$
  - $S_0^{\text{SNT}}(\textit{red bus}|R_2) = 0.0$

This example is, intentionally, very artificial. It assumes a very restricted set of utterances (*yellow bus*, for instance, is not available). It also assumes that *red bus* and *bus* are equally probable under S_0^{SNT} given R_1. The point is to show that under these assumptions, $S_1^{\text{SNT-GP}}$ breaks the symmetry between the utterances when referring to R_1. The more general point is that like the versions of S_1 introduced in chapters 1, 2 and 3, $S_1^{\text{SNT-GP}}$ as defined here exhibits a preference for utterances which refer to the target reference and not the other referents in W.

Indeed, we find that $S_1^{\text{SNT-GP}}(\textit{red bus}|R_1) = \frac{2}{3} > \frac{1}{3} = S_1^{\text{SNT-GP}}(bus|R_1)$, as claimed.

To move from the idealized example above to real image captioning, we simply replace the hand-specified S_0^{SNT} with its neural counterpart, as described in section 4.1. No changes need to be made to equation (20).

This is all well and good when U is finite, as in the example above, but the whole point of an image captioning system is that it does not require a pre-specified set of utterances. Rather, the whole set of possible sequences of words forms the utterance set U.

The problem Unfortunately for us, while $S_1^{\text{SNT-GP}}$ is a perfectly well defined distribution over an infinite set U given a neural model S_0^{SNT}, actually using it, in the sense of sampling from it, finding the maximum a posteriori caption, or determining the probability of captions conditioned on particular images is impossible.

The reason is simple: $S_1^{\text{SNT-GP}}$ has a normalizing term $Z = \sum_{u' \in U} S_0^{\text{SNT}}(u'|w) \cdot L_1^{\text{SNT}}(w|u')$, where U is the infinite set of all possible sequences of tokens of any length. Even when bounded to a finite length, U is exponentially large in the maximum sequence length. As such, computing Z exactly is impossible, and all we have is an energy model, i.e. a distribution up to a normalization constant.

All we can do is rank any set of captions, since this does not require the normalizing term. This allows us to iterate through a set of possible captions and determine which is best. This would be a way of finding the maximum a posteriori caption under $S_1^{\text{SNT-GP}}$ (although this may not be the best caption - see the note in section (4.1), if not, once again, for the infinitude of U.

The solution employed by Monroe and Potts (2015) and Andreas and Klein (2016a) to the intractabiliity of inference for $S_1^{\text{SNT-GP}}$ when U is infinite is to sample a small subset of probable utterances from the S_0, as a finite version of U, which I refer to as U', upon which exact inference at $S_1^{\text{SNT-GP}}$ can be performed. While tractable, this approach has the shortcoming of only considering a small region of the true set U, which decreases the extent to which pragmatic reasoning will be able to apply. In particular, if a useful caption never appears in the sampled prior, it cannot appear in the posterior. Furthermore, as the maximum caption length increases, the number of possible utterances $|U|$ increases, and the fraction of captions u in U that are in U' decreases.

The method I employ here is the incremental pragmatic reasoning introduced in chapter 3. A simple version of this approach in the context of image captioning is employed by Vedantam et al. (2017), where it is termed the "emittor-suppressor" method. There, the approach is restricted to pragmatics for a pair of images, but is easily extensible to the full generality of the RSA framework.

Applying incremental pragmatics to neural sequence models takes advantage of the fact that state-of-the-art models generate language word by word, usually in a left-to-right fashion, as described in section 4.1. In other words, the distribution S_0^{SNT} is fully determined by S_0^{WORD}. This applies not only to RNN architectures like the LSTM, but also to more recent architectures like the Transformer (Vaswani et al., 2017), which is used for translation in section 4.4.

$$\begin{tikzcd}
S_0^{\text{SNT}} \arrow[r] & S1^{\text{UTT-}\mathbf{X}} \\
S_0^{\text{WORD}} \arrow[u] \arrow[r, dashed] & S_1^{\text{WORD}} \arrow[u, dashed]
\end{tikzcd}$$

Figure 4.2: Two ways of constructing an utterance-level pragmatic speaker from S_0^{WORD}. The solid green path is to construct a literal speaker S_0^{SNT} over full utterances and then perform pragmatics, which gives rise to $S_1^{\text{SNT-GP}}$ while the dashed red path is to construct a word-level pragmatic speaker S_1^{WORD} from S_0^{WORD} and then use this to define an utterance-level pragmatic speaker, $S_1^{\text{SNT-IP}}$.

Because S_0^{SNT} decomposes into S_0^{WORD}, we can define S_1^{WORD} in terms of S_0^{WORD} and obtain a model $S_1^{\text{SNT-IP}}$ from S_1^{WORD}. We can think of this as performing pragmatic reasoning on the level of words. As depicted in figure 4.2, this amounts to exchanging the order of two operations: instead of first unrolling a word level speaker and then performing pragmatics (the global model $S_1^{\text{SNT-GP}}$), we can first perform pragmatics and then unroll a world level speaker (the incremental model $S_1^{\text{SNT-IP}}$).

My implementations[4] of incremental pragmatics take advantage of this viewpoint, by defined two functions on objects representing distributions, namely *unfold* and *pragmatics*, which can be applied in either order, to yield $S_1^{\text{SNT-GP}}$ and $S_1^{\text{SNT-IP}}$.

S_1^{WORD} is defined in terms of S_0^{WORD} as follows, analogous to the definition of $S_1^{\text{SNT-GP}}$ in terms of S_0^{SNT}. $S_1^{\text{SNT-IP}}$ is then the *unrolling* of S_1^{WORD}, analogous to the unrolling of S_0^{SNT} from S_0^{WORD}:

(21) $L_1^{\text{WORD}}(w|u,c) \propto S_0^{\text{WORD}}(u|w,c) \cdot P_L(w)$

(22) $S_1^{\text{WORD}}(u|w,c) \propto S_0^{\text{WORD}}(u|w,c) \cdot L_1^{\text{WORD}}(w|u,c)^{\alpha}$

(23) $S_1^{\text{SNT-IP}}(u|w,k) = \prod_{i=0}^{n-1} S_1^{\text{WORD}}(u = u[i]|w = w, c = k + u[:i])$

By default, we take P_L to be a uniform distribution over W, although see section 4.2.1 for a discussion of an alternative approach.

4.2.1 Unrolling strategy

A greedy sampling strategy for a recurrent model $P(wd|w,c)$ (i.e. a model conditioned on an input state and a sequence of words) and an input w, samples by starting with an empty list [], choosing the highest probability word wd^* from $P(\cdot|w,[])$, and then choosing the next word as the highest probability choice from $P(\cdot|w,[wd])$, and so on. It stops when it samples an end token, such as a full stop.

[4] Available at https://github.com/reubenharry/pragmatic-translation

Greedy sampling is not guaranteed to sample the most probable sequence under the unrolled distribution corresponding to $P(wd|w, c)$. It will fail to do so if the most probable sequence ever involves the production of a word which is not optimal. For example, "The red bus." may be a more probable full caption than "Bus is red" for an image captioning system given an image of a red bus, but the latter may be produced by greedy sampling due to the model's preference for *bus* as the initial word.

The solution to this problem is *lookahead*, i.e. the ability to consider a sequence following the immediate choice before deciding on the next word in the sequence. One way to achieve this is with a beam search, a common technique for decoding from recurrent models which keeps track of a *beam* of n candidate sequences. At each time step of the decoding, each of the sequences in the beam is extended, yielding a new, larger beam, from which only the top n sequences with highest probability are kept. For a sufficiently high value of n, this amounts to *full* decoding (i.e. an exhaustive exploration of all possible sequences) and for $n = 1$, it amounts to a greedy search. Values of n around 10 represent a tractable compromise between full and greedy search.

Updating the prior at each timestep In the version of $S_1^{\text{SNT-IP}}$ presented above, the decision of the next word at each time step depends on S_1^{WORD} and in turn on L_1^{WORD}, a model which has P_L as its (by default uniform) prior. Put more straightforwardly, $S_1^{\text{SNT-IP}}$ aims to be informative on the assumption that it is communicating with a listener which, at each time step, has a uniform distribution over which item in W is the referent.

One can imagine a model in which this assumption is changed, so that the listener's prior evolves over the course of the utterance. For instance, a listener who has heard *The red*, might already strongly suspect that R_1 is the referent. If this is the case, the speaker should then have less cause to continue to produce informative language, and might prefer to conclude the caption with *bus*, rather than, for example, *double decker bus*.

To put this into practice, we can introduce a new unrolling procedure in which, at timestep t of the unrolling, the listener L_1^{WORD} takes as its prior over images the L_1^{WORD} posterior from timestep $(t-1)$. We can introduce a prior distribution *ip* over states for the listener, and use superscript time indexes, to make this more precise:

$$L_1^{WORD:t}(w|u, ip^t, pc^t) \propto S_0^{WORD:t}(u|w, pc^t) \cdot ip^t(w) \tag{4.2}$$

$$S_1^{WORD:t}(u|w, ip^t, pc^t) \propto S_0^{WORD:t}(u|w, pc^t) \cdot L_1^{WORD:t}(w|u, ip^t, pc^t)^{\alpha} \tag{4.3}$$

$$ip^0(w) = \frac{1}{|W|} \tag{4.4}$$

$$ip^t(w) = L_1^{WORD:t}(w|u, ip^{t-1}, pc^t) \tag{4.5}$$

4.2.2 Effect of incrementality

Incrementality, while an excellent way to explore the space of possible pragmatic utterances, has its limitations. In particular, it can result in an overeagerness to be informative on the part of $S_1^{\text{SNT-IP}}$. As an example, suppose that the caption "The bus is yellow." is preferable to "Yellow is the bus" on grounds of linguistic naturalness. If S_0^{WORD} encodes this preference, $S_1^{\text{SNT-IP}}$ will inherit it too, but supposing that its target is a yellow bus, and the only other image in W is a red bus, it will also be compelled to begin its caption with *yellow*. More generally, while beam search allows lookahead of a kind, it is not a solution to the S_1^{WORD}'s inherent preference for local informativity. This is an intrinsic consequence of incremental pragmatic reasoning (see chapter 3 for discussion of evidence for comparable human behavior).

Despite this drawback, it is possible to establish some simple commonalities between $S_1^{\text{SNT-GP}}$ and $S_1^{\text{SNT-IP}}$. Given a state w, call an utterance u *weakly informative* if $L_1^{\text{SNT}}(w|u) \geq \frac{1}{|W|}$, where W is the set of possible states. In other words, given u, the literal listener L_1^{SNT} will guess the correct state with probability at least at chance (when costs are 0). We note that the utterance u_w^* obtained by greedy unrolling at each step of generation is weakly informative. To see this, observe that the nth word of u_w^* is $argmax_{word} S_1^{\text{WORD}}(word|w = r, c = u_w^*[: n])$. Since at each step S_1^{WORD} produces a word which, at worst, does not rule out any states for L_1^{WORD}, the resulting sentence u^* at worst gives $L_1^{\text{SNT}}(w|u_w^*) \geq \frac{1}{|W|}$. In other words, greedily unrolling the incremental speaker will produce an utterance which is at least as informative as chance.

This result suggests that the strategy of choosing the most informative word (or syntactic unit) at each point in the generation of an utterance can be used as a substitute for choosing, from all utterances, the one which is most informative. Whether it is actually a substitute that works well, or reproduces human behavior is a question returned to in section 4.3.

4.2.3 Character level incrementality

As well as reasoning incrementally on the level of words, it is possible to do so on the level of characters. This possibility arises in the context of a character level recurrent network (Chung et al., 2016), where each successive character, rather than word, is determined by the previous sequence of characters. Despite the intuition importance of words as linguistic units, character level models are able to produce grammatical expressions, suggesting that they are capable of learning linguistic structure. The practical advantage of character level pragmatics is that U is much smaller (≈30 vs. ≈$20,000$), making the ensuing RSA model much more computationally efficient, and allowing a larger beam search. Because of the way $S_1^{\text{SNT-IP}}$ is defined, no conceptual effort is required in changing from characters to words.

However, it is not a-priori obvious that incremental reasoning on the level of characters should yield anything comparable to global informativity, given the distinctions between $S_1^{\text{SNT-GP}}$ and $S_1^{\text{SNT-IP}}$ discussed above.

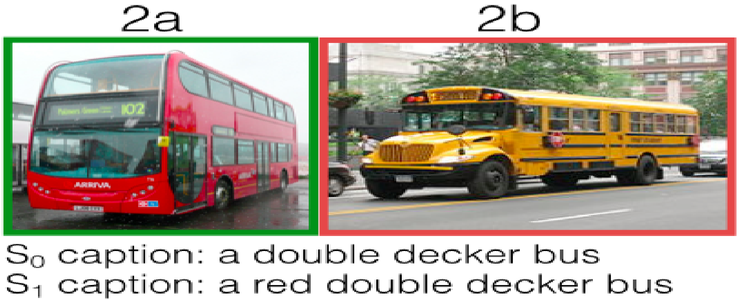

2a
2b
S0 caption: a double decker bus
S1 caption: a red double decker bus

Figure 4.3: Captions for the target image (in green).

4.3 Evaluating pragmatic image captioning

I now describe experiments evaluating the use of $S_1^{\text{SNT-IP}}$ in the domain of image captioning. These experiments are conducted both for word level and character level CNN-LSTM models.

Figure 4.3 shows an example of an S_0 and S_1 character level caption for a target image in the context of a set of images W. Qualitatively, we see that S_1 produces an unambiguous caption, while S_0 produces an ambiguous one. This suggests that even at the level of characters, the preference for pragmatic choices yields caption level informativity. This behavior, if robust across examples, is both surprising and desirable, since it suggests that informative language can be produced without consideration of caption level alternative utterances.

The informativity of a caption can be measured by playing the reference game implicit in the referential caption generation task: is a listener able to recover the speaker's intended target? This is the form of evaluation we undertake.

To this end, a listener $L_{eval}(image|caption) \propto P_{S_0^{\text{SNT}}}(caption|image)$ is defined, where $P_{S_0^{\text{SNT}}}(caption|image)$ is the total probability of S_0 incrementally generating *caption* given *image*. In other words, L_{eval} is just the global listener which uses Bayes' rule to obtain from S_0^{SNT} the posterior probability of each image w given a full caption u.

The neural S_0^{WORD} used in the definition of L_{eval} must be trained on separate data to the neural S_0^{WORD} used for the $S_1^{\text{SNT-IP}}$ model which produces captions, since otherwise this $S_1^{\text{SNT-IP}}$ production model effectively has access to the system evaluating it. As Mao et al. (2016) note, "a model might 'communicate' better with itself using its own language than with others". In evaluation, the training data is therefore divided in half, with one part for training the S_0^{WORD} used in the caption generation model $S_1^{\text{SNT-IP}}$ and one part for training the S_0^{WORD} used in the caption evaluation model L_{eval}.

We say that the caption succeeds as a referring expression if the target has more probability mass under the distribution $L_{eval}(image|caption)$ than any distractor. In other words, if the system makes a hard choice by choosing the most probable referent, it succeeds when it guesses the correct referent.

To summarize, the success of $S_1^{\text{SNT-IP}}$ at being informative is measured in the natural way: by the ability of a listener (here a separately trained neural agent, but potentially a human) to pick the intended target image out of W.

Dataset I train the production and evaluation models on separate sets consisting of regions in the Visual Genome dataset (Krishna et al., 2017) and full images in MSCOCO (Lin et al., 2014). This is carried out both for a character-level and a word level model, to allow for comparison between the two. Both datasets consist of over 100,000 images of common objects and scenes. MSCOCO provides captions for whole images, while Visual Genome provides captions for regions within images.

The test sets consist of clusters of 10 images. For a given cluster, each image in it is set as the target, in turn. Two test sets are used. Test set 1 (TS1) consists of 100 clusters of images, 10 for each of the 10 most common objects in Visual Genome.[5]

Test set 2 (TS2) consists of regions in different Visual Genome images whose ground truth captions have high word overlap, an indicator that they are similar. We again select 100 clusters of 10. Both test sets have 1000 items in total (10 potential target images for each of 100 clusters).

Hyperparameters I use a beam search with width 10 to produce captions, and a rationality parameter of $\alpha = 5.0$ for the S_1.

4.3.1 Results

As shown in Table 3.7, the character-level S_1 obtains higher accuracy (68% on TS1 and 65.9% on TS2) than the S_0 (48.9% on TS1 and 47.5% on TS2), demonstrating that S_1 is better than S_0 at referring.

Advantage of Incremental RSA 66% percent of the times in which the S_1 caption is referentially successful and the S_0 caption is not, for a given image, the S_1 caption is not one of the top 50 S_0 captions, as generated by the beam search unrolling at S_0. This means that in these cases the non-incremental method discussed in section 4.2 could not have generated the S_1 caption, if these top 50 S_0 captions were the support of the prior over utterances. This is an indication of the value of incremental pragmatic reasoning, but more systematic experimentation would be valuable here.

[5] Namely, *man*, *person*, *woman*, *building*, *sign*, *table*, *bus*, *window*, *sky*, and *tree*.

Model	**TS1**	**TS2**
Char S_0	48.9	47.5
Char S_1	$\mathbf{68.0}$	$\mathbf{65.9}$
Word S_0	57.6	53.4
Word S_1	60.6	57.6

Table 4.1: Accuracy on both test sets.

Comparison to Word-Level $S_1^{\text{SNT-IP}}$ I compare the performance of the character-level model to a word-level model.[6] It is evaluated with an L_{eval} model that also operates on the word level.

Though the word S_0 performs better on both test sets than the character S_0, the character S_1 outperforms the word S_1, demonstrating the advantage of a character-level model for pragmatic behavior.

Variants of the Model I further explore the effect of two design decisions in the character-level model. First, I consider a variant of S_1 which has a prior over utterances determined by an LSTM language model trained on the full set of captions. This achieves an accuracy of 67.2% on TS1. Second, I consider our standard S_1 but with unrolling such that the L_0 prior is drawn uniformly at each time step rather than determined by the L_0 posterior at the previous step. This achieves an accuracy of 67.4% on TS1. This suggests that neither this change of S_1 nor L_0 priors has a large effect on the performance of the model.

4.3.2 Shortcomings of the current approach

The integration of $S_1^{\text{SNT-IP}}$ with a neural sequence model is an exciting step towards marrying theoretical models of linguistic behavior with statistical models capable of handling some of the complexity of real data.

Three shortcomings of the present approach are that the set of states W is small, leading to language which distinguishes the referents but is inaccurate, that W has to be explicitly specified, whereas ideally our model would also supply W and P_L automatically, and that using images as representations of states is cumbersome when we want fine-grained control over states.

I now discuss these three problems in more detail, with the aim of motivating the shift to a domain where a simultaneous solution to all three is possible.

Problem 1: Strange behavior when W is small The goal of S_1 (whether $S_1^{\text{SNT-IP}}$ or $S_1^{\text{SNT-GP}}$), in the context of image captioning, is to being informative in the sense of distinguishing the target image from the other images in W. In the case where W consists of very few, say two, images, and the rationality parameter

[6] Here, we use greedy unrolling, for reasons of efficiency due to the size of U for the word-level model, and set $\alpha = 1.0$ from tuning on validation data. For comparison, we note that greedy character-level S_1 achieves an accuracy of 61.2% on TS1.

α is high, the S_1 is compelled to be informative at the cost of producing accurate or natural captions. For example, it might be the case that R_1 in figure 4.1 induces a neural image captioner S_0 to produce "brick" as a caption with a much higher probability than it has to produce "brick" for R_2. The consequence of this fact is an ensuing desire in S_1 to say "brick", even though in absolute terms, "brick" has low probability under S_0^{SNT} for either R_1 or R_2.

One way of interpreting the problem here is that S_1 never has to worry about mistakenly communicating to their imagined listener that their target image is a brick. This is because neither of the images in W are in fact bricks. If a third image of a brick *were* added to W, producing "brick" as an S_1 caption for R_1 would suddenly lose almost all of its probability.

In other words, S_1 only cares about being informative up to the uncertainty of its imagined listener, and if this uncertainty is constrained to a very specific set of hypotheses, there can be strange consequences for the behavior of S_1. This problem bears a resemblance to the idea of a question under discussion or projection, as discussed in chapter 2; an overly small set of states can be seen as a too-coarsely grained partition over possible worlds.

A natural way to address this problem is to make the state space W larger, so that it contains images corresponding to a larger set of possible captions. If P_L consequently put some probability mass on a huge variety of possible images, the problem would be alleviated: the informativity of "brick" for conveying R_1 over R_2 would be counteracted by its much greater informativity for some R_n, containing an actual brick.

Problem 2: The need for explicitly specified states If our aim is to model human linguistic behavior in the domain of image captioning, the artificial setting of a reference game, in which a set W of a target and accompanying distractors are selected by hand, is limited. One particular limitation that has not been addressed so far is that an explicit set of distractor images needs to be provided. Or to pose the problem more generally, a prior P_L, giving both a set of distractors and according probabilities must be provided. So far, we supplied the distractors manually, and took the distribution over them to be uniform.

Ideally, we would want to be able to produce this distribution automatically, so that given a single image, a system can produce a caption which is informative in the sense of distinguishing the target image from these distractors.

In theory, this distribution could contain very rich information. For example a human captioner knows that it is unlikely that grass is any other color than green, and so that it is unnecessary to stipulate it is green - doing so has little utility in informing an interlocutor. As another example, when describing the image in figure 4.4, I might say *A bus without a driver*. I choose to mention this detail, instead of saying *A bus with round wheels* precisely because the former is much more surprising under a distribution over natural images, or alternatively, much more surprising under a distribution over states of the world. Equivalently, choosing

Figure 4.4: Image of a driverless electric bus on a road.

this detail gives the listener a high probability of being able to recover the image I want to communicate (or the state of the world this image corresponds to).

One appealing solution is to use a neural model to provide this distribution. For example, a variational autoencoder can explicitly learn a distribution over images, which could be used as P_L. Taking this approach would even remove the need to explicitly specify a set of images W, since in this case, W would be the space of all possible images (of a certain size), each assigned a probability by P_L. Note that this also addresses *problem 1*, since a continuous distribution like this would not be unnaturally restricted.

The drawback of this approach is that probabilistic generative models of images, while an active area of research, are not (at the time of writing) able to produce particularly high quality images, outside of constrained domains like digit or face generation. It is therefore unclear whether pursuing this approach would yield good results. Furthermore, computing the posterior given a continuous prior of this form would be a challenging endeavor.

Problem 3: The awkwardness of images as representations of states In the current approach, images act as a stand-in for states of the world. In other words, the visual information present in an image is what determines the internal representation fed to the recurrent neural network, and so, is what determines what language is produced.

If we want to control the language a captioning system produces, we therefore have to change the input. For example, to get our system to describe a state of the world in which a bus is driving off a cliff, we would need to first obtain an image of that scenario.

Resolving these problems by recourse to a new domain Viewed from the perspective of a reference game, translation is a setting where W consists of source language sentences, and U consists of target language translations. A speaker model produces a translation given a sentence in the source language, and a listener, symmetrically, produces a source language translation given a target language sentence. As discussed in section 4.4, there is an intuitive use case for an S_1 model in the context of such a reference game, to provide translations which preserve semantic distinctions in the source language.

With respect to problem 3, having sentences as states offers an immediate solution. Rather than having to find or create an image which represents some particular state of the world, we can simply describe that state by a natural language expression in the source language.

Moreover, in section 4.5 I will show that we can exploit the fact that translation is feasible in both directions (source to target language, and target to source language) to provide an informative translation model which does not require the explicit specification of a set of distractors or a prior over them. This will avoid the problems incurred by having a small set of manually specified states (problems 1 and 2). Further, I show that, when implemented with an efficient inference algorithm, this approach improves both cycle-consistency and translation quality measured by standard metrics.

4.4 Translation

Languages differ in what meaning distinctions they must mark explicitly. As such, translations risk mapping from a form in one language to a more ambiguous form in another. For example, the definite (4.6) and indefinite (4.7) both translate (under *Fairseq* and *Google Translate*) to (4.8) in French, which is ambiguous in definiteness.

That is, English always distinguishes between definite and indefinite nouns (by the presence or absence of a determiner) whereas in many Romance languages, like French, a noun without a determiner can be either definite or indefinite.

The animals run fast. (4.6)

Animals run fast. (4.7)

Les animaux courent vite (4.8)

More generally, when viewed as a function from a set of source language sentences to the set of target language sentences, state-of-the-art translation systems are not injective (one-to-one). There is no reason to think that translation *should* be one-to-one, or even a well defined function (i.e. having a unique translation for each input sentence), particularly at the level of sentences, where contextual information that is not explicitly present in the sentence may be critical for determining the translation.

However, there are practical situations in which it may be desirable to avoid a particular ambiguity. In the above English sentences and translation, for instance, the French sentence does not convey whether a definite group of animals is being referred to, or whether instead the sentence is meant in a generic sense. If this distinction was important in the original English sentence, its absence from the translation might be undesirable.

A symptom of many-to-one translation is loss of cycle-consistency. Formally, say that a pair of functions $f : A \to B$, $g : B \to A$ is cycle-consistent (i.e. is an isomorphism) if $g \cdot f = id$, the identity function. If f is *not* one-to-one, then (f, g) is not cycle-consistent.[7] For the case of translation, this equates to the ability of a translation system from target language to source language to recover the original sentence from its translation. When two sentences are mapped to one, they cannot both be restored with a target-source translator to their respective original sentences: information has been lost.

4.4.1 Avoiding many-to-one translation

While languages differ in what distinctions they are *required* to express, all are usually capable of expressing any given distinction when desired. As such, meaning loss of the kind discussed above is, in theory, avoidable.

Instantiating the informative speaker model of RSA in the domain of translation provides a natural way to encourage a trained translation system to produce translations which distinguish the source language sentence from distractors, and so, to avoid many-to-one mappings. That is, we play a reference game in which the states are a finite set W of source language sentences (such as the pair 4.6 and 4.7) and the utterances are the infinite set of target language utterances U.

To use $S_1^{\text{SNT-IP}}$, we specify a set of distractor sentences in the source language, of which one is the target, and proceed almost exactly as in the case of image captioning, using beam search to decode from $S_1^{\text{SNT-IP}}$ as an approximation of $S_1^{\text{SNT-GP}}$.

As an example of the behavior of the system, figure 4.5 shows the S_0^{SNT} and $S_1^{\text{SNT-IP}}$ translations for sentences A and B which jointly compose the set of states ($W = \{A, B\}$). The key property of this model is that, for $W = \{A, B\}$, when translating A, $S_1^{\text{SNT-GP}}$ prefers translations of A that are unlikely to be good translations of B. So for pairs like (4.6) and (4.7), $S_1^{\text{SNT-GP}}$ is compelled to produce a translation for the former that reflects its difference from the latter, and vice versa. For the French example above, $S_1^{\text{SNT-IP}}$'s translation of (4.6) when $W = \{$(4.6), (4.7)$\}$ is “Ces animaux courent vite” (*These* animals run fast.).

The examples here are from only two European languages and the evaluation in section 4.6 is only for German. However, the technique is general to any language pair. Exploration of more distant language pairs is a valuable topic for future research.

[7] Note however that when A and B are infinite, the converse does not hold: even if f and g are both one-to-one, (f, g) need not be cycle-consistent, i.e. $f \circ g$ need not be the identity. Consider, for example, the case where $f(x) = g(x) = 2x$. f and g are bijections on the reals, but e.g. $f(1) \neq 1$.

A	He is wearing glasses.
B	He wears glasses.
$S_0^{\text{SNT}}(A)$	Er trägt eine Brille.
$S_0^{\text{SNT}}(B)$	Er trägt eine Brille .
$S_1^{\text{SNT-IP}}(A)$	Er trägt jetzt eine Brille.
$S_1^{\text{SNT-IP}}(B)$	Er hat eine Brille.

Figure 4.5: Similar to Figure 4.1, S_0^{SNT} collapses two English sentences into a single German one, whereas $S_1^{\text{SNT-IP}}$ distinguishes the two in German.

4.4.2 When is translation many-to-one?

It is natural to wonder how often many-to-one translations occur. To explore this question for a single language pair, I create a corpus of 500 pairs of distinct English sentences which map to a single German one (the evaluation language in section 4.6). This is done by selecting short sentences from the Brown corpus (Francis and Kucera, 1964), translating them to German, and taking the best two candidate translations back into English, if these two themselves translate to a single German sentence. Translation in both directions was done with Fairseq. I identify a number of common causes for the many-to-one maps. Two frequent types of verbal distinction lost when translating to German are tense (54 pairs, e.g. "...others {were, have been} introduced .") and modality (16 pairs, e.g. "...prospects for this year {*could*, *might*} be better."), where German "können" can express both epistemic and ability modality, distinguished in English with "might" and "could" respectively. Owing to English's large vocabulary, lexical difference in verb (31 pairs, e.g. "arise" vs. "emerge"), noun (56 pairs, e.g. "mystery" vs. "secret"), adjective (47 pairs, e.g. "unaffected" vs. "untouched") or deictic/pronoun (32 pairs, usually "this" vs "that") are also common.

While the dataset is by no means representative, since only differences that appear in the beam of a German-English translator are observed, it reveals some common classes of distinction English makes more than German. The most common pairwise differences are lexical (220), either by choice of verb (e.g. "arise" vs. "emerge"), noun (e.g. "mystery" vs. "secret"), adjective (e.g. "unaffected" vs. "untouched"), adverb (e.g. "seldom" vs. "rarely") or deictic/pronoun (very commonly "this" vs "that"). A large number of the pairs differ instead either orthographically, or in other ways that do not correspond to a clear semantic distinction (e.g. "She had {*taken*, *made*} a decision."). 29 differ by the presence or absence of a determiner (e.g. "This makes (the) order of entries variable.").

English has a particularly large lexicon, so it is unsurprising that distinctions between lexical items (a difference in a single noun, for example) will often be lost. While a large number of differences are lexical , certain semantic distinctions, in particular tense and modal force, consistently occur.

4.5 Generalizing to an unbounded state space

While $S_1^{\text{SNT-IP}}$ can disambiguate between pairs of sentences, it has two shortcomings. First, it requires one (or more) distractors to be provided, so translation is no longer fully automatic. Second, because the distractor set W consists of only a pair (or finite set) of sentences, $S_1^{\text{SNT-IP}}$ only cares about being informative up to the goal of distinguishing between these sentences. This mirrors the problem discussed in section 4.3.2, where an image caption is inaccurate because the goal of being informative outweighs the probabilistic preference for true captions.

In section 4.3.2, I conjectured that by expanding W, this problem could be lessened. I proposed an extreme version, in which W was the set of all images, and P_L a neurally learned distribution over them. This was judged to be infeasible, due to the difficulty of generating natural images (although versions for simpler domains, like MNIST) are conceivable.

However, in the domain of translation, a distribution over W is easy to obtain. In particular, a translator from target to source language constitutes a conditional distribution $P(w|u)$. Using the convention that a 0 index refers to a model with no explicit pragmatic reasoning, I term such a target-source translation model $L_0^{\text{SNT}}(w|u)$. With that, we can introduce a simple variation of the informative speaker, $S_1^{\text{SNT-CGP}}$ (for “cyclic global pragmatics”), as follows:

$$S_1^{\text{SNT-CGP}}(u|w) \propto S_0^{\text{SNT}}(u|w) L_0^{\text{SNT}}(w|u)^{\alpha} \tag{4.9}$$

What does this mean? First note that L_1^{SNT}, which previously performed Bayesian inference to obtain a posterior distribution over a finite set W of source language sentences has been replaced by a neural model L_0^{SNT}, which is a distribution over an infinite set of source language sentences.

It is also worth emphasizing that the listener/speaker terminology has no conceptual significance in this setting, other than to remain consistent with other applications of RSA. S_0^{SNT} and L_0^{SNT} are symmetric - we could just as well switch their names.

The resulting behavior is that, given a source sentence w, $S_1^{\text{SNT-CGP}}$ prefers translations u which have high probability under S_0^{SNT}, but *also* are likely to be recovered, in the sense of causing L_0^{SNT} to have high probability of producing translation w when given u.

$S_1^{\text{SNT-CGP}}$ is like $S_1^{\text{SNT-GP}}$, but its goal is to produce a translation which allows a listener model (now L_0^{SNT}) to infer the original sentence, not among a small set of presupplied possibilities, but among *all source language sentences*. As such, an optimal translation u of w under $S_1^{\text{SNT-CGP}}$ has high probability of being generated by S_0^{SNT}, high probability of being translated back to w by L_0^{SNT} and consequently, lower probability of being translated to any other sequence of English words. $S_1^{\text{SNT-CGP}}$ is very closely related to reconstruction methods, e.g. (Tu et al., 2017), and conceptually related to uses of a cyclic loss in non-linguistic domains, e.g. (Zhu et al., 2017).

Intuitively, total meaning preservation is achieved by a translation which distinguishes the source sentence w from every other sentence in the source language which differs in meaning. The result, for sufficiently high values of α, is an injective map from source to target language. Note, of course, that this injective map may correspond to nonsense translations: as before, raising α too high results in a preference for informativity over quality.

4.5.1 An incremental inference algorithm for $S_1^{\text{SNT-CGP}}$

Exact inference is again intractable, though as with $S_1^{\text{SNT-GP}}$, it is possible to approximate by subsampling from S_0^{SNT}. This is very close to the approach taken by Li and Jurafsky (2016), who find that reranking a set of outputs by probability of recovering input "dramatically decreases the rate of dull and generic responses." in a question-answering task. However, because the subsample is small relative to U, they use this method in conjunction with a diversity increasing decoding algorithm.

As in the case with explicit distractors, we instead opt for an incremental model, now $S_1^{\text{SNT-CIP}}$ which approximates $S_1^{\text{SNT-CGP}}$. The definition of $S_1^{\text{SNT-CIP}}$ (4.11) is more complex than the incremental model with explicit distractors ($S_1^{\text{SNT-IP}}$) since L_0^{WORD} must receive complete sentences, rather than partial ones like L_1^{WORD}. As such, we need to marginalize over continuations k of partial sentences in the target language:

$$\begin{aligned} S_1^{\text{WORD-C}}(wd|w,c) &\propto S_{\text{o}}^{\text{WORD}}(wd|w,c)\cdot \\ &\sum_k (L_0^{\text{SNT}}(w|c+wd+k) S_0^{\text{SNT}}(k|w,c+wd)) \end{aligned} \tag{4.10}$$

$$S_1^{\text{SNT-CIP}}(u|w,c) = \prod_t S_{\text{1}}^{\text{WORD-C}}(u[t]|w,c+u[:t]) \tag{4.11}$$

Since the sum over continuations of c in (4.10) is intractable to compute exactly - it is a sum over all possible continuation sequences, we approximate it by a single continuation, obtained by greedily unrolling S_0^{SNT}.

The whole process of generating a new word wd of the translation from a sequence c and a source language sentence w is as follows: first use S_0^{WORD} to generate a set of candidates for the next word (in practice, we only consider two, for efficiency). For each of these, use S_0^{SNT} to greedily unroll a full target language sentence from $c+wd$, namely $c+wd+k$, and rank each wd by the probability $L_0^{\text{SNT}}(w|c+wd+k)$.

The following pseudocode resembles the Python code[8] implementing $S_1^{\text{WORD-C}}$. In practice, we fix WIDTH=2:

[8] Note the use of Python indexing conventions, and Numpy (numerical Python) broadcasting.

```
def S1-WD-C.fwd(src=s,c=[]):
  support,logprobs = S0-WD.fwd(s)
  scores = []
  for wd in support[:WIDTH]:
    ext=S0-SNT.fwd(src=s,c=c+[wd])
    sc=L0-SNT.logprob(tgt=s,src=ext)
    scores.append(sc)
  unnorm=logprobs+scores
  next_word=support[argmx(unnorm)]
  return next_word

def S1-SNT-CPG.fwd(src=s,c=[]):
  next_word = None
  out = []
  while next_word=!STOP_TOKEN:
    support,logprobs = S0-WD.fwd(s)
    scores = []
    for wd in support[:WIDTH]:
    ext=S0-SNT.fwd(src=s,c=c+[wd])
    s=L0-SNT.logprob(tgt=s,src=ext)
    scores.append(s)
    unnorm=logprobs+scores
    next_word=support[argmx(unnorm)]
    out.append(next_word)
  return out
```

An example of the behavior of $S_1^{\text{SNT-CIP}}$ and S_0^{SNT} on a sentence from the test set we use for evaluation (see section 4.6) is shown below; S_0^{SNT} is able to preserve the phrase "world's eyes", which S_0^{SNT} translates merely as "world":

- Source sentence: Isolation keeps the world's eyes off Papua.
- Reference translation: Isolation hält die Augen der Welt fern von Papua.
- S_0^{SNT}: Die Isolation hält die Welt von Papua fern.
- $S_1^{\text{SNT-CIP}}$: Die Isolation hält die Augen der Welt von Papua fern.

Efficiency $S_1^{\text{SNT-CIP}}$ is much less than $S_1^{\text{SNT-IP}}$, for the reason that at *each time step* of the production of a sentence, it is necessary to greedily extend a partial sequence in the target language into a full sentence, and to translate it back into the source language. As such, this is not intended as the basis for a practical translation system. Rather, it is intended to show that cycle-consistency is important for translation, and should be taking into account in the design of real-world systems.

4.6 Evaluating $S_1^{\text{SNT-CIP}}$

The method of evaluation used for $S_1^{\text{SNT-IP}}$ in the domain of image captioning was to measure the accuracy of a separately trained listener model at recovering the intended referent. Similarly for $S_1^{\text{SNT-CIP}}$, we can measure cycle-consistency, the ability of a separately trained model to recover the original sentence. I view cycle-consistency as an indirect measure of meaning preservation, since the ability to recover the original sentence requires meaning distinctions not to be collapsed.

As with the evaluation of image captioning in section 4.3, in evaluating cycle-consistency it is important to use a separate target-source translation mechanism than the one used to define $S_1^{\text{SNT-CIP}}$. Otherwise, the system has access to the model which evaluates it and may improve cycle-consistency without producing meaningful target language sentences. For this reason, we translate German sentences (produced by S_0^{SNT} or $S_1^{\text{SNT-CIP}}$) back to English with *Google Translate*. To measure cycle-consistency, we use the BLEU metric (implemented with sacreBLEU (Post, 2018)), with the original sentence as the reference.

However, this improvement of cycle consistency, especially with a high value of α, may come at the cost of translation quality. Moreover, it is unclear whether BLEU serves as a good metric for evaluating "translation" of sentences when the source and target language are the same. To further ensure that translation quality is not compromised by $S_1^{\text{SNT-CIP}}$, we evaluate BLEU scores of the German sentences it produces. This requires evaluation on a corpus of aligned sentences, unlike the sentences collected from the Brown corpus in section 4.4.2[9].

Note that this method of evaluation was not possible for the use of $S_1^{\text{SNT-IP}}$ in the context of image captioning, since there was no expectation that the captions generated would resemble the reference captions, which were produced without the goal of distinguishing the target image from an explicitly specified set of distractors.

We conduct our evaluations on English to German translation, making use of publicly available pre-trained English-German and German-English Fairseq models.

We perform both evaluations (cycle-consistency and translation) on 750 sentences[10] of the 2018 English-German WMT News test-set.[11] We use greedy unrolling in all models (using beam search is a goal for future work). For α (which represents the trade-off between informativity and translation quality) we use 0.1,

[9] While we find that $S_1^{\text{SNT-CIP}}$ improves cycle-consistency for the Brown corpus over S_0^{SNT}, we have no way to establish whether this comes at the cost of translation quality.

[10] Our implementation of $S_1^{\text{SNT-CIP}}$ was not efficient, and we could not evaluate on more sentences for reasons of time.

[11] `http://www.statmt.org/wmt18/translation-task.html`

Model	Cycle	Translate
S_0^{SNT}	43.35	37.42
$S_1^{\text{SNT-CIP}}$	**47.34**	**38.29**

Table 4.2: BLEU score on cycle-consistency (*c*) and translation (*t*) for WMT, across baseline and informative models. Greedy unrolling and $\alpha = 0.1$

obtained by tuning on validation data. Note that this is a low value compared to what was used for the image captioning system.

Results As shown in table (4.2), $S_1^{\text{SNT-CIP}}$ improves over S_0^{SNT} not only in cycle-consistency, but in translation quality as well. This suggests that the goal of preserving information, in the sense defined by $S_1^{\text{SNT-CGP}}$ and approximated by $S_1^{\text{SNT-CIP}}$, is important for translation quality.

4.7 Discussion

The achievement of this chapter was to take the incremental model of pragmatics proposed in chapter 3, in particular the model of informative language production $S_1^{\text{SNT-IP}}$, and apply it to natural language generation tasks (image captioning and translation) where incrementality serves as a way to overcome the intractability inherent in reasoning over an infinite set of possible utterances.

In most of these cases, I considered a definition of the informative speaker $S_1^{\text{SNT-GP}}$ in terms of a model L_1^{SNT} which has a uniform prior over an explicitly specified set of images W, and performs Bayes rule with S_0^{SNT} as the likelihood. $S_1^{\text{SNT-IP}}$ then constitutes an acceptable approximation of $S_1^{\text{SNT-IP}}$, as shown by the evaluations in 4.3.

A surprising finding of the application of incremental pragmatics (i.e. the use of $S_1^{\text{SNT-IP}}$ in place of $S_1^{\text{SNT-GP}}$) was that not only did character level incrementality produce globally informative language, but that it produced better results than word level incrementality. It is hard to say why this is; one likely reason is that the small number of characters ($\sim$ 60) allowed for a wider beam search. Intuitively, the character level model allows for the exploration of more paths before committing to a single one.

A natural question this raises is whether there is a possibility of achieving improved results by choosing a level of incrementality between the word and character level. The word piece segmentation used in the models explored in section 4.4 is one example. More generally, it might well be the case that an optimal unrolling strategy is to choose certain parts of the sentence in which to increase or decrease the granularity of incrementality, according to heuristics.

In section 4.5 I considered a second model, $S_1^{\text{SNT-CGP}}$ which instead of reasoning about L_1^{SNT}, reasons about a neural model L_0^{SNT}. This approach removes the need for a hand-chosen distractor set W, since L_0^{SNT} is a distribution over all source language sentences. $S_1^{\text{SNT-CIP}}$ is then an approximation to $S_1^{\text{SNT-CGP}}$. Unlike back-translation to augment data during training (Sennrich et al., 2015), our model uses pretrained translators.

Models of this variety differ in an important way from standard RSA models. While standard models have a single "base case", i.e. an agent L_0 or S_0 which ends the nesting of agents reasoning about each other, there are now two base cases[12], a listener and a speaker. Previously, we viewed the semantics as S_0, now it is separated into two components.

The apparent success of $S_1^{\text{SNT-CIP}}$ on improving translation quality on a pair of two fairly similar languages raises the question of whether improvements will increase for more distant language pairs, in which larger scale differences exist in what information is obligatorily represented - this is a direction for future work.

A broader kind of extension to the work proposed in this chapter concerns richer models of pragmatics; $S_1^{\text{SNT-GP}}$ is a very simple model, but the whole attraction of integrating Bayesian pragmatics with neural models is the possibility of the complex reasoning patterns seen in idealized models being extended to real data.

For instance, integrating a neural model with a Bayesian pragmatic model of figurative language (Kao et al., 2014b), question asking (Hawkins et al., 2015), or focus (Bergen and Goodman, 2015b) has the potential to yield a model capable of generating much more controlled language. The takeaway message of the chapter should therefore be that the unbounded utterance space is not a serious barrier to the implementation of more complex models.

[12] I use the term *base case* slightly misleadingly here since the listener and speaker agents in question are not defined recursively in the present setting. However, one can easily do so, by defining L_n in terms of S_n and S_n in terms of L_{n-1}. Then the term base case applies accurately.

Chapter 5

A Unified Perspective

This dissertation builds on a perspective in which the intuitions about pragmatics laid out by Grice (1975), and the notion of a convention proposed by Lewis (1969) are formalized in probabilistic models of *nested reasoning* between a speaker and a listener, known as the Rational Speech Acts framework (RSA).

Under this perspective, the meaning of an utterance corresponds to the belief (i.e. distribution over states) that it induces in a listener. The literal meaning corresponds to the belief induced in an agent who does not reason about their interlocutor, by contrast to the pragmatic meaning, which is the belief of an agent who does, and moreover assumes that their interlocutor is reasoning about them.

This perspective on meaning is compatible with a truth-conditional semantics, which can appear in the definition of the literal listener, although it does not require it. It makes a clear distinction between meaning inherent in an utterance (from the semantics), meaning derived from the context (prior knowledge, present already in the literal listener L_0), and meaning derived from Gricean reasoning (only present in L_1 and higher). Furthermore, it captures the notion of a convention as the information that the S_n assumes that the L_{n-1} assumes S_{n-1} assumes ... that L_0 knows. As the depth of nested reasoning tends to infinity, this converges to common knowledge, which can be seen as the theoretical characterization of convention.

What is nice about this vision is that it unifies several diverse perspectives on meaning. In particular, the logical notion of meaning developed by Tarski (1983) and adapted for natural language by Montague (1973) is united with the probabilistic viewpoint of a meaning as a distribution over states of the world, and the understanding of a language user as a Bayesian agent.

What this dissertation adds to the story The contribution of this dissertation is to show how RSA models behave when the state space W and the utterance space U have richer structure than just being sets of states and linguistic expressions respectively. Chapter 2 shows how the additional structure of a vector space plays nicely with a model capable of handling figurative language, L_1^{Q}, with the elegant consequence that *questions under discussion* amount to linear projections. Chapter 3 shows how a recursively generated utterance space U allows for pragmatic reasoning to be factored, so that pragmatics takes place at the level of words.

In both cases, performing inference efficiently is a challenge, but not an intractable one. A Gaussian formulation of a vector space semantics allows for a closed form solution to L_0, while L_1^{Q} itself can be approximated tractably as a result of the projection of w onto a subspace corresponding to a projection q. In the case of $S_1^{\text{SNT-GP}}$, the sequential structure of U allows for an approximation, $S_1^{\text{SNT-IP}}$, which I propose as a cognitively more realistic model of informative language generation (chapter 3).

These theoretical contributions are what enable the more practical contribution of the dissertation, of scaling RSA models to open domain natural language. The interpretation of L_1^{Q} over a vector space W allows it to be applied to a semantics defined in terms of word vectors. As a consequence, it is possible to apply RSA to arbitrary adjective-noun phrases, yielding both an NLP tool and a means of validations L_1^{Q} as a model of non-literal language interpretation. Similarly, the incremental approach to informativity, made possible when utterances $u \in U$ are sequences of words, allows for the application of S_1 (in particular, $S_1^{\text{SNT-IP}}$) to the natural language generation tasks of image captioning and translation (chapter 4).

In the course of addressing these theoretical and practical challenges, several questions emerge, which I'll consider in turn.

- Across the dissertation, several variations of the RSA framework have emerged, as dictated by the needs of each situation. What lessons can be taken away about the general architecture, particularly as concerns the "literal" speaker S_0 and listener L_0?
- What do the Bayesian models of pragmatics discussed in this dissertation have to tell us about what a semantics should look like?
- Chapters 2 and 4 share the approach of taking a *pretrained* statistical model and using it as a semantics. How viable is it to jointly *learn* a model (either a word embedding as in chapter 2 or an image captioning or translation model as in chapter 4) and perform pragmatic reasoning? What would this gain us?
- Both in the application of the vectorial L_1^{Q} to adjective-noun phrases, and the design of $S_1^{\text{SNT-IP}}$ and $S_1^{\text{SNT-CIP}}$, pragmatic reasoning takes place before the computation of full sentences. Is there a meaningful sense in which we can talk of a compositional pragmatics?
- The alternative utterance set: the set of utterances available as alternatives has been a point of controversy for Gricean theories of pragmatics over several decades (Katzir, 2007). What sort of theory does the approach taken in chapter 4 — namely to consider all *words* at each timestep as alternatives — correspond to? Is this a way to resolve the tension between grammaticalized and Gricean theories of

implicature?

- What is the natural next step in this research program? How can we build on the successful elements of the work presented here?

5.1 The architecture of Bayesian models of pragmatics

Chapter 1.1.2 outlines a number of choice points in the vanilla Rational Speech Acts model. Some are subsumed in existing parameters (for example, the effects of introducing a cost term amount to little more than the utterance prior) while others (like the rationality parameter α) do not change the fundamental dynamic of the model, but increase the strength of certain effects.

However, other architectural choices seem meaningful but arbitrary. In particular, why choose to begin the series of nested of speaker and listener models with L_0 rather than S_0? In the application of S_1 to language generation using neural models, it was clear that an S_0 was the preferred starting point, since it corresponds to the form of a neurally trained conditional language model. However, in the application of L_1^{Q} to language interpretation, it turned out to be more straightforward to have an L_0.

There is a generalization of the S_0 first and L_0 first versions of the vanilla RSA model, which involves both L_0 *and* S_0, shown as follows:

(24) $L_n(w|u) \propto S_{n-1}(u|w) \cdot L_0(w|u)$

(25) $S_n(u|w) \propto L_{n-1}(w|u) \cdot S_0(u|w)$

If we define S_0 and L_0 with the same semantics and respective uniform priors (as in equations 26 and 27), we obtain something equivalent to the vanilla RSA model sketched in chapter 1.1.1.

(26) $L_0(w|u) \propto [\![u]\!](w)$

(27) $S_0(u|w) \propto [\![u]\!](w)$

To make this identification, what is now L_2 corresponds to L_1 in the previous model: that is, L_2 here is the listener that reasons about a speaker who reasons about a literal listener. The change in the nth order listener in 24 and nth order speaker in 25 is that their priors are defined in terms of L_0 and S_0 respectively. Note that in this formulation, L_1 and S_1 are symmetric: substituting L for S changes from (24) to (25), and vice versa.

This *symmetric* variant of RSA is essentially the model used in the application of RSA to translation in chapter 4.5, although in that case, S_0 and L_0 do not have explicit semantics, but rather are neural models, and the nesting extends only one level, to S_1.

What this symmetric RSA model allows for is the possibility of a separation of the conventional relation between form and meaning (the semantics) into two parts, one coming from L_0 and one from S_0. This turned out to be useful for practical purposes in chapter 4.5, allowing for an efficient approximation of a sentence level informative speaker whose distractor set W is unbounded, but whether it is motivated from a theoretical perspective is a question for future work.

5.2 The nature of the semantics

There are two different ways of characterizing a semantics for natural language. The first is that a semantics is a logical relation between expressions $u \in U$ and states $w \in W$ (Montague, 1973). This is the sense of semantics inherited from Tarski's notion of a semantics in mathematical logic (Tarski, 1983).

The second notion is that the semantics is some function of U and W which is a *convention*, in the sense of Lewis. In other words, the semantics is the relationship between states and expressions which is common knowledge in a community of practice. Note that in this sense, all features of the base level listener L_0 and (optionally) speaker S_0 are part of the conventional knowledge that language users have, including not only the semantics, but the prior on states and utterances. I refer to these two notions of what a semantics is as the *Montagovian* and *Lewisian* views, respectively.

The semantics in vanilla RSA is in line with both of these viewpoints. L_n assumes that S_n assumes that $L_{n-1} \ldots$ assumes that L_0 has access to a relational semantics $[\![\cdot]\!]$. If the highest model is L_n for some n, this logical semantics is nth order higher order knowledge, which becomes common knowledge as n tends to infinity.

Mathematically speaking, the RSA framework only enforces the *Lewisian* view; in theory, there is no mathematical reason to prevent the semantics itself being soft. For instance, one could replace the semantics used to define L_0 in 28 with any positive real valued function $f(u, w)$ which assigns a score between 0 and 1 for each pair of utterance u and state w, as in 29.

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

(29) $L_0(w|u) = \frac{f(u,w) \cdot P_L(w)}{\sum_{w' \in W} f(u,w) \cdot P_L(w')}$

This approach is taken in the application of L_1^{Q} to word embeddings in chapter 2. Another option is to forgo a semantics altogether, and simply have S_0 or L_0 (or both) be conditional distributions, as in the neurally parametrized distributions for S_0 and L_0 in chapter 4.

I view this flexibility, on the part of the RSA framework, as a good thing. While a logical semantics is appropriate for modeling many aspects of natural language meaning, for many others it is not. One such example is use-conditional meaning (Gutzmann, 2015), such as the meaning of *sorry* or *ouch*, where a probability distribution $P(u|w)$ describing which states are likely to cause which utterances seems far more appropriate than a truth-conditional semantics. Similar concerns apply to sociolinguistic meaning, such as conventional information about the social identities of speakers who use certain phonetic features. For both of these cases, Qing and Cohn-Gordon (2018) present an RSA model grounded in a model $S_0(u|w)$ which represents use conditions and conventional knowledge about which social identities produce which sociolinguistic markers as a probability distribution. My view, therefore, is that probabilistic representations of a semantics are motivated not only on practical grounds, but on theoretical ones, although in many settings, a logical semantics is needed.

5.3 Jointly learning a semantics and reasoning about pragmatics

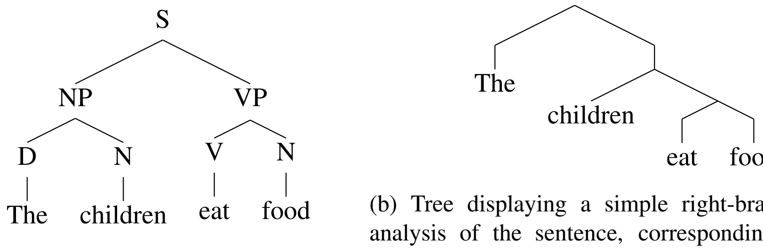
In chapter 2, the semantics takes the form of a word embedding, as discussed above. However, these embeddings are learned from large scale corpora in which language is used pragmatically. For example, the vector $\overrightarrow{shark}$ is learned from data in which *shark* is used metaphorically. As noted in chapter 2, it should be possible to learn word embeddings while taking into account pragmatic reasoning. For example, L_1^{Q} could, in theory, perform a triply joint inference, over states w, projections q and the semantics itself, i.e. the word embedding E. In practice, this would require a far more efficient inference algorithm than what is use in chapter 2 for L_1^{Q}, and more importantly, would require a task, perhaps the neighboring word prediction task on which Word2Vec is trained, against which the system could be trained.

An analogous consideration arises for the model considered in chapter 4, where the semantics takes the form of a conditional probability distribution over utterances given states (parametrized by a neural net).

In the setting of neural architectures, where end-to-end learning is generally the favored approach, a salient alternative is to view L_1 and S_1 as layers on top of the neural architecture, and train the model jointly by backpropagating through the RSA layers. This is the approach pursued by Mao et al. (2016) for images, by Monroe and Potts (2015) for idealized reference games, and by McDowell and Goodman (2019) for color patch description, all using the subsampling approach to $S_1^{\text{SNT-GP}}$ inference of Andreas and Klein (2016b).

The difficulty is that this requires a data set in which the observations are pairs of states (e.g. images) *in context* and utterances. In other words, the model must be trained on data in which the set of distractor images for each target image is provided. In Mao et al. (2016) for example, a special dataset is assembled for this purpose. A further challenge is to train an end-to-end *incremental* model, and to investigate the nature of the semantics which emerges in such an end-to-end system.

Models capable of inferring a semantics in this way are an important direction for future work: they allow quantitative answers to the question: what must the conventional knowledge of the relationship between



(a) Tree displaying the constituent structure of a sentence.

(b) Tree displaying a simple right-branching analysis of the sentence, corresponding to a linked list.

forms and meanings in a linguistic community be like in order to have produced the observed linguistic data? Since this is the inferential task faced by humans in the real world, to infer the conventions of language from data in which those conventions are not directly exhibited but rather give rise to pragmatic behavior, attempting to handle this problem is a core part of scaling RSA.

5.4 Compositional pragmatics

The approach inherent in vanilla RSA, inherited from Grice and subsequent work in linguistics, is to view pragmatics as a module which takes a fully formed semantic representation as an input. In other words, whatever happens in the compositional semantics to assemble a sentence meaning is irrelevant to the ensuing process of Gricean reasoning.

This assumption is challenged both by the application of L_1^{Q} to adjective-noun metaphors in chapter 2 and the incremental models S_1^{WORD} and L_1^{WORD} proposed in chapter 3.

In the case of L_1^{Q} and adjective-noun (AN) phrases, the meaning of a given AN phrase corresponds to the posterior distribution of L_1^{Q} on having heard the adjective when the noun gave rise to the prior. This was discussed in chapter 2.6.3, where a comparison was made to the classical treatment in formal semantics of adjective-noun composition as function application (with the adjective as a function taking the noun). What is novel about this view of compositionality is not just that composition is treated as a probabilistic inference, but that this inference can include pragmatic reasoning. In other words, pragmatics is included in the compositional process of deriving the meaning of a phrase from the meaning of its parts.

The incremental listener L_1^{WORD} can be seen as a comparable model of compositional pragmatics, with the stipulation that the syntactic structure which dictates the process of composition is a binary right branching tree, i.e. a list. (These trees are right branching because the proposal for incremental pragmatics is intended to model the sequential process of hearing an utterance word by word, rather than the process of semantic composition according to a constituency parse of a sentence. The difference between the two trees is illustrated in figures 5.1a and 5.1b. Also note that right-branching composition corresponds to incremental processing from right to left - but similar remarks apply to a left-branching structure.)

In particular, we can model the incremental processing of a sentence with L_1^{WORD} by applying the model to the first word of the sentence, and using the resulting posterior distribution when applying the model to the second word, and so on (see chapter 4.2.1 for a related approach using S_1^{WORD}).

The important point is that the notion of performing pragmatics incrementally, and using the posterior distribution at one node as the prior for another, is orthogonal to the choice of syntactic structure over which semantic composition takes place. This means that a compositional pragmatics which respected the syntactic structure of a sentence would work in (broadly) the same way as the case of L_1^{Q} and L_1^{WORD} discussed above. This is an avenue of future work; the salient question is whether there are theoretical or empirical reasons to favor a theory of pragmatics which interacts with the syntactic structure of the sentence.

5.5 Choosing alternatives

A key question in Gricean accounts of pragmatic reasoning is how the set of utterances available to a speaker should be chosen (Katzir, 2007; Chierchia et al., 2008). This set, often referred to as the *alternatives*, corresponds to U (or rather, U without the chosen utterance u) in the Bayesian pragmatic framework of RSA.

It is clear that varying the set of alternatives can strongly affect the model's predictions. For instance, if *I ate all of the cookies* is an alternative to *I ate some of the cookies*, then the latter may create the implicature of *some but not all*. However, if *I ate some but not all of the cookies* is itself in the alternative set, this implicature is, at least, diminished. For this reason, a principled method of deriving alternatives needs to be part of any Gricean theory of pragmatic reasoning.

In the context of a probabilistic account of Gricean reasoning, the issue can be generalized to the question of how we can obtain not only a set U, but also a *distribution* over U.

One sort of proposal views alternative generation (and pragmatic exhaustification) as a grammatical process (Chierchia et al., 2008) where alternative sentences are produced by substituting a given word with its word-level alternatives.

Another answer is to assume that every sentence is a possible alternative for every other sentence, with a probability distribution which is part of the linguistic convention making up the language in question dictating the weight given to difference utterances in U.

This is the approach taken in the application of $S_1^{\text{SNT-GP}}$ to the tasks of image captioning and translation, where

U is the set of all sequences of words. The choice of prior over U here is either supplied by a language model or by the neurally trained S_0 itself, conditioned on the image being captioned or sentence being translated.

What picture does the incrementally informative speaker $S_1^{\text{SNT-IP}}$ subscribe to? On the one hand, this model considers all possible alternatives, weighted by a learned probability distribution which is taken to represent conventional knowledge. On the other, the alternatives are at the word level.

In this way, $S_1^{\text{SNT-IP}}$ represents a kind of compromise between the two positions described above. Alternatives are generated locally, and the process by which they are chosen is part of either the grammar, the semantics or both. A salient question for future work is whether a model like $S_1^{\text{SNT-IP}}$ is capable of making correct predictions about embedded implicatures (Potts et al., 2016), which is a crucial motivation for a grammaticalized theory of implicature.

5.6 The larger project

It seems appropriate to conclude by considering where the work discussed in this dissertation, interdisciplinary as it is, fits into larger research programs in the fields of linguistics, cognitive science and artificial intelligence.

In terms of linguistic theory, RSA is a formalization of Gricean pragmatic reasoning. As discussed in chapter (1.2.1), it presents a probabilistic model in which both the semantic and pragmatic meaning of an expression are the belief the expression induces in a (literal/pragmatic) listener. Regarding RSA as a linguistic theory, the aim of the research presented here was to show that it is compatible with non-trivial linguistic structure (see chapter 3), although there is a long way to go here, exploring the interaction of pragmatics as formalized by RSA with complex syntactic structure.

While RSA is compatible with a non-probabilistic semantics, it does not require it, and the relaxation to a purely probabilistic S_0/L_0, without a truth-conditional semantics, turned out to be useful in practical applications (chapter 4). The take home message is that a probabilistic model in which language interpretation corresponds to an inference about the state of the world is a comfortable fit both for symbolic approaches to meaning, and statistical ones.

As a contribution to cognitive science, the Rational Speech Acts framework can be seen as an attempt to provide a computational level theory (Bechtel, 1994) of language interpretation and production (Bergen and Goodman, 2015a). In this regard, the extensions of RSA proposed throughout this dissertation are comparable to recent work which combines Bayesian models of cognitive processes (such as intuitive theories of physics, or interpretation of scenes) with deep learning techniques (Wu et al., 2015; Liu et al., 2018), in order to harness the interpretability and explanatory utility of the former and the modeling power of the latter.

Finally, the models proposed in this dissertation function as NLP systems in their own right, for interpreting and generating language. This is part of a larger project to bridge the gap between theoretical insights in linguistics and cognitive science, and real world applications, both as a means of validating those theories, and putting them to fruitful use.

Bibliography

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182. Association for Computational Linguistics, 2016a. URL `http://aclweb.org/anthology/D16-1125`.

Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182. Association for Computational Linguistics, 2016b. URL `http://aclweb.org/anthology/D16-1125`.

Sheldon Jay Axler. *Linear algebra done right*, volume 2. Springer, 1997.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, 9, 2014.

William Bechtel. Levels of description and explanation in cognitive science. *Minds and Machines*, 4(1): 1–25, 1994.

Leon Bergen and Noah D Goodman. The strategic use of noise in pragmatic reasoning. *Topics in cognitive science*, 7(2):336–350, 2015a.

Leon Bergen and Noah D Goodman. The strategic use of noise in pragmatic reasoning. *Topics in cognitive science*, 7(2):336–350, 2015b.

Leon Bergen, Roger Levy, and Noah D. Goodman. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9(20), 2016.

Max Black. Xii-metaphor. In *Proceedings of the Aristotelian Society*, volume 55, pages 273–294. The Oxford University Press, 1955.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2), 1990.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911, 2014.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*, 2016.

Gennaro Chierchia, Danny Fox, and Benjamin Spector. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Unpublished manuscript*, 2008.

Noam Chomsky. *Syntactic structures*. Walter de Gruyter, 1957.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147*, 2016.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*, 2010.

Reuben Cohn-Gordon and Noah Goodman. Lost in machine translation: A method to reduce meaning loss. *arXiv preprint arXiv:1902.09514*, 2019.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443. Association for Computational Linguistics, 2018a. URL `http://aclweb.org/anthology/N18-2070`.

Reuben Cohn-Gordon, Noah D Goodman, and Christopher Potts. An incremental iterated response model of pragmatics. *arXiv preprint arXiv:1810.00367*, 2018b.

Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087, 2015.

Raffaello D'Andrea and Geir E Dullerud. Distributed control design for spatially interconnected systems. *IEEE Transactions on automatic control*, 48(9):1478–1495, 2003.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Cicero Dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, 2014.

Paul E Engelhardt, Karl GD Bailey, and Fernanda Ferreira. Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, 54(4):554–573, 2006.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*, pages 15–29. Springer, 2010.

Hartry Field. Tarski's theory of truth. *The Journal of Philosophy*, 69(13):347–375, 1972.

Gregory Finley, Stephanie Farmer, and Serguei Pakhomov. What analogies reveal about word vectors and their compositionality. In *Proceedings of the 6th joint conference on lexical and computational semantics (* SEM 2017)*, pages 1–11, 2017.

W Nelson Francis and Henry Kucera. Brown corpus. *Department of Linguistics, Brown University, Providence, Rhode Island*, 1, 1964.

Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336 (6084):998, 2012.

Michael Franke. *Signal to Act: Game Theory in Pragmatics*. ILLC Dissertation Series. Institute for Logic, Language and Computation, University of Amsterdam, 2009.

Daniel Fried, Jacob Andreas, and Dan Klein. Unified pragmatic models for generating and following instructions. *arXiv preprint arXiv:1711.04987*, 2017.

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*, 2016.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419. Association for Computational Linguistics, 2010.

Noah Goodman, Vikash Mansinghka, Daniel M Roy, Keith Bonawitz, and Joshua B Tenenbaum. Church: a language for generative models. *arXiv preprint arXiv:1206.3255*, 2012.

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

Noah D Goodman and Andreas Stuhlmüller. The design and implementation of probabilistic programming languages, 2014.

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

H. Paul Grice. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics*, volume 3: Speech Acts, pages 43–58. Academic Press, New York, 1975.

Daniel Gutzmann. *Use-conditional meaning: Studies in multidimensional semantics*, volume 6. OUP Oxford, 2015.

Robert XD Hawkins, Andreas Stuhlmüller, Judith Degen, and Noah D Goodman. Why do you ask? good questions provoke informative answers. In *CogSci*. Citeseer, 2015.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 473–483, 2017.

Theo Herrmann and Werner Deutsch. *Psychologie der objektbenennung*. Huber, 1976.

Douglas Hofstadter and Emmanuel Sander. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic Books, 2013.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Frederick Jelinek, John D Lafferty, and Robert L Mercer. Basic methods of probabilistic context free grammars. In *Speech Recognition and Understanding*, pages 345–360. Springer, 1992.

Shaojie Jiang and Maarten de Rijke. Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots. *arXiv preprint arXiv:1809.01941*, 2018.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

Justine T. Kao, Leon Bergen, and Noah D. Goodman. Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 719–724, Wheat Ridge, CO, July 2014a. Cognitive Science Society.

Justine T. Kao, Jean Y. Wu, Leon Bergen, and Noah D. Goodman. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007, August 2014b.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

Roni Katzir. Structurally-defined alternatives. *Linguistics and Philosophy*, 30(6):669–690, 2007.

Walter Kintsch. Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 7(2): 257–266, 2000.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Ran Lahav. Against compositionality: the case of adjectives. *Philosophical studies*, 57(3):261–279, 1989.

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

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

D Terence Langendoen, Paul Martin Postal, et al. *The vastness of natural languages*. Blackwell Oxford, 1984.

Daniel Lassiter and Noah D Goodman. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and linguistic theory*, volume 23, pages 587–610, 2013.

Daniel Lassiter and Noah D Goodman. Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 194(10):3801–3836, 2017.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 1969.

Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*, 2016.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Tal Linzen. Issues in evaluating semantic spaces using word analogies. *arXiv preprint arXiv:1606.07736*, 2016.

Yunchao Liu, Zheng Wu, Daniel Ritchie, William T Freeman, Joshua B Tenenbaum, and Jiajun Wu. Learning to describe scenes with programs. 2018.

Ralph Loader. *Notes on simply typed lambda calculus*. University of Edinburgh, 1998.

Brian MacWhinney and Davida Fromm. Two approaches to metaphor detection. In *LREC*, pages 2501–2506, 2014.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

Bill McDowell and Noah Goodman. Learning from omission. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 619–628, 2019.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34 (8):1388–1429, 2010.

Will Monroe and Christopher Potts. Learning in the Rational Speech Acts model. In *Proceedings of 20th Amsterdam Colloquium*, Amsterdam, December 2015. ILLC.

Richard Montague. The proper treatment of quantification in ordinary english. In *Approaches to natural language*, pages 221–242. Springer, 1973.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018.

Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C Frank. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, 33(4):755–802, 2016.

Ciyang. Qing and Reuben. Cohn-Gordon. Use-conditional meaning in rational speech act models. *Sinn und*

Bedeutung, 111, 2018.

Ciyang Qing, Noah D Goodman, and Daniel Lassiter. A rational speech-act model of projective content. In *CogSci*, 2016.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI, 2018.

Craige Roberts. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136, 1996.

Tim Rohrer. The cognitive science of metaphor from philosophy to neuroscience. *Theoria et Historia Scientiarum*, 6(1):27–42, 2002.

Paula Rubio-Fernández. How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in psychology*, 7:153, 2016.

Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

Natalie Schluter. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2039. URL `https://www.aclweb.org/anthology/N18-2039`.

Michael F Schober and Herbert H Clark. Understanding by addressees and overhearers. *Cognitive psychology*, 21(2):211–232, 1989.

Julie C Sedivy. Implicature during real time conversation: A view from language processing research. *Philosophy compass*, 2(3):475–496, 2007.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

Ekaterina Shutova. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 2016.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353, 2013.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642, 2013.

Alfred Tarski. *Logic, semantics, metamathematics: papers from 1923 to 1938*. Hackett Publishing, 1983.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Michael Henry Tessler and Noah D Goodman. Communicating generalizations about events. In *CogSci*, 2016.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), volume 1, pages 248–258, 2014.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. Neural machine translation with reconstruction. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 130–132. Association for Computational Linguistics, 2006.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems*, pages 127–135, 2015.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.