A Game-Theoretic Approach to Generating Spatial Descriptions

Dave Golland
UC Berkeley
Berkeley, CA 94720
dsg@cs.berkeley.edu

Percy Liang
UC Berkeley
Berkeley, CA 94720
pliang@cs.berkeley.edu

Dan Klein
UC Berkeley
Berkeley, CA 94720
klein@cs.berkeley.edu

Abstract

Language is sensitive to both semantic and pragmatic effects. To capture both effects, we model language use as a cooperative game between two players: a speaker, who generates an utterance, and a listener, who responds with an action. Specifically, we consider the task of generating spatial references to objects, wherein the listener must accurately identify an object described by the speaker. We show that a speaker model that acts optimally with respect to an explicit, embedded listener model substantially outperforms one that is trained to directly generate spatial descriptions.

1 Introduction

Language is about successful communication between a speaker and a listener. For example, if the goal is to reference the target object O1 in Figure 1, a speaker might choose one of the following two utterances:

(a) *right of* O2 $\qquad$ (b) *on* O3

Although both utterances are semantically correct, (a) is ambiguous between O1 and O3, whereas (b) unambiguously identifies O1 as the target object, and should therefore be preferred over (a). In this paper, we present a game-theoretic model that captures this communication-oriented aspect of language interpretation and generation.

Successful communication can be broken down into semantics and pragmatics. Most computational work on interpreting language focuses on compositional semantics (Zettlemoyer and Collins, 2005; Wong and Mooney, 2007; Piantadosi et al., 2008), which is concerned with verifying the truth of a sentence. However, what is missing from this truth-oriented view is the pragmatic aspect of language—that language is used to accomplish an end goal, as exemplified by speech acts (Austin, 1962). Indeed, although both utterances (a) and (b) are semantically valid, only (b) is pragmatically felicitous: (a) is ambiguous and therefore violates the Gricean maxim of manner (Grice, 1975). To capture this maxim, we develop a model of pragmatics based on game theory, in the spirit of Jäger (2008) but extended to the stochastic setting. We show that Gricean maxims

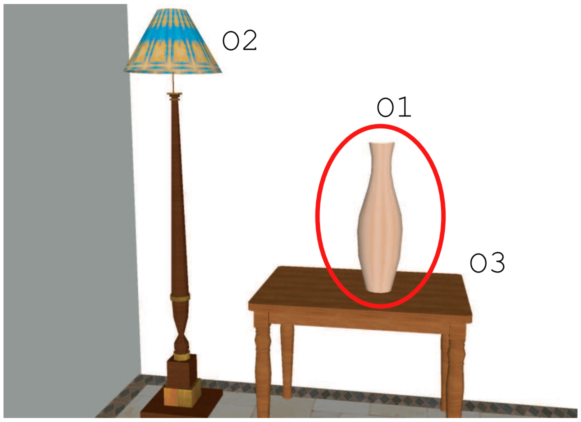


Figure 1: An example of a 3D model of a room. The *speaker*'s goal is to reference the target object O1 by describing its spatial relationship to other object(s). The *listener*'s goal is to guess the object given the speaker's description.

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 410–419,
MIT, Massachusetts, USA, 9-11 October 2010. ©2010 Association for Computational Linguistics

fall out naturally as consequences of the model.

An effective way to empirically explore the pragmatic aspects of language is to work in the grounded setting, where the basic idea is to map language to some representation of the non-linguistic world (Yu and Ballard, 2004; Feldman and Narayanan, 2004; Fleischman and Roy, 2007; Chen and Mooney, 2008; Frank et al., 2009; Liang et al., 2009). Along similar lines, past work has also focused on interpreting natural language instructions (Branavan et al., 2009; Eisenstein et al., 2009; Kollar et al., 2010), which takes into account the goal of the communication. This work differs from ours in that it does not clarify the formal relationship between pragmatics and the interpretation task. Pragmatics has also been studied in the context of dialog systems. For instance, DeVault and Stone (2007) present a model of collaborative language between multiple agents that takes into account contextual ambiguities.

We present our pragmatic model in a grounded setting where a speaker must describe a target object to a listener via spatial description (such as in the example given above). Though we use some of the techniques from work on the semantics of spatial descriptions (Regier and Carlson, 2001; Gorniak and Roy, 2004; Tellex and Roy, 2009), we empirically demonstrate that having a model of pragmatics enables more successful communication.

2 Language as a Game

To model Grice's cooperative principle (Grice, 1975), we formulate the interaction between a speaker S and a listener L as a cooperative game, that is, one in which S and L share the same utility function. For simplicity, we focus on the production and interpretation of single utterances, where the speaker and listener have access to a shared context. To simplify notation, we suppress writing the dependence on the context.

The Communication Game

1. In order to communicate a *target* o to L, S produces an *utterance* w chosen according to a strategy $p_{\mathrm{S}}(w \mid o)$.
2. L interprets w and responds with a *guess* g according to a strategy $p_{\mathrm{L}}(g \mid w)$.
3. S and L collectively get a utility of $U(o, g)$.

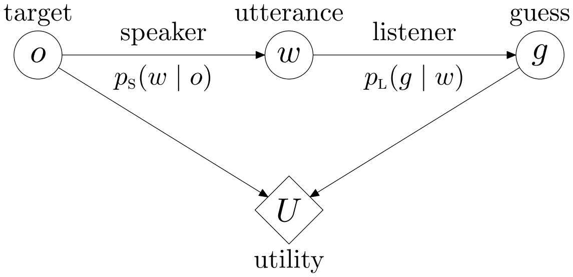


Figure 2: Diagram representing the communication game. A target, o, is given to the speaker that generates an utterance w. Based on this utterance, the listener generates a guess g. If $o = g$, then both the listener and speaker get a utility of 1, otherwise they get a utility of 0.

This communication game is described graphically in Figure 2. Figure 3 shows several instances of the communication game being played for the scenario in Figure 1.

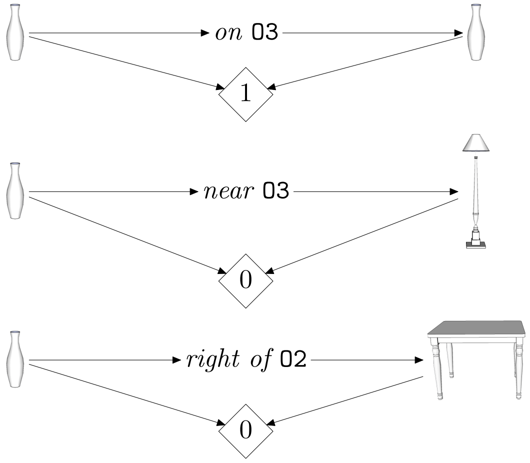


Figure 3: Three instances of the communication game on the scenario in Figure 1. For each instance, the target o, utterance w, guess g, and the resulting utility U are shown in their respective positions. A utility of 1 is awarded only when the guess matches the target.

Grice's maxim of manner encourages utterances to be unambiguous, which motivates the following utility, which we call *(communicative) success*:

$$U(o, g) \overset{\mathrm{def}}{=} \mathbb{I}[o = g], \qquad (1)$$

where the indicator function $\mathbb{I}[o = g]$ is 1 if $o = g$ and 0 otherwise. Hence, a utility-maximizing speaker will attempt to produce unambiguous utterances because they increase the probability that the listener will correctly guess the target.

Given a speaker strategy $p_{\mathrm{S}}(w \mid o)$, a listener strategy $p_{\mathrm{L}}(g \mid w)$, and a prior distribution over targets $p(o)$, the expected utility obtained by S and L is as follows:

$$\begin{aligned} \mathrm{EU}(\mathrm{S},\mathrm{L}) &= \sum_{o,w,g} p(o)p_{\mathrm{S}}(w|o)p_{\mathrm{L}}(g|w)U(o,g) \\ &= \sum_{o,w} p(o)p_{\mathrm{S}}(w|o)p_{\mathrm{L}}(o|w). \end{aligned} \tag{2}$$

3 From Reflex Speaker to Rational Speaker

Having formalized the language game, we now explore various speaker and listener strategies. First, let us consider *literal* strategies. A literal speaker (denoted S:LITERAL) chooses uniformly from the set of utterances consistent with a target object, i.e., the ones which are semantically valid;[1] a literal listener (denoted L:LITERAL) guesses an object consistent with the utterance uniformly at random.

In the running example (Figure 1), where the target object is O1, there are two semantically valid utterances:

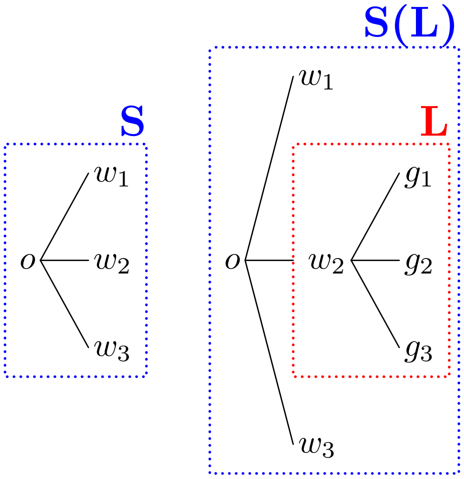
(a) *right of* O2 $\qquad (b)$ *on* O3

S:LITERAL selects (a) or (b) each with probability $\frac{1}{2}$. If S:LITERAL chooses (a), L:LITERAL will guess the target object O1 correctly with probability $\frac{1}{2}$; if S:LITERAL chooses (b), L:LITERAL will guess correctly with probability 1. Therefore, the expected utility EU(S:LITERAL, L:LITERAL) $= \frac{3}{4}$.

We say S:LITERAL is an example of a *reflex* speaker because it chooses an utterance without taking the listener into account. A general reflex speaker is depicted in Figure 4(a), where each edge represents a potential utterance.

Suppose we now have a model of some listener L. Motivated by game theory, we would optimize the expected utility (2) given $p_{\mathrm{L}}(g \mid w)$. We call the resulting speaker S(L) the *rational* speaker with respect to listener L. Solving for this strategy yields:

$$\begin{aligned} p_{\mathrm{S(L)}}(w \mid o) &= \mathbb{I}[w = w^*], \text{ where} \\ w^* &= \operatorname*{argmax}_{w'} p_{\mathrm{L}}(o \mid w'). \end{aligned} \tag{3}$$



(a) Reflex speaker (b) Rational speaker

Figure 4: (a) A reflex speaker (S) directly selects an utterance based only on the target object. Each edge represents a different choice of utterance. (b) A rational speaker (S(L)) selects an utterance based on an embedded model of the listener (L). Each edge in the first layer represents a different choice the speaker can make, and each edge in the second layer represents a response of the listener.

Intuitively, S(L) chooses an utterance, w^*, such that, if listener L were to interpret w^*, the probability of L guessing the target would be maximized.[2] The rational speaker is depicted in Figure 4(b), where, as before, each edge at the first level represents a possible choice for the speaker, but there is now a second layer representing the response of the listener.

To see how an embedded model of the listener improves communication, again consider our running example in Figure 1. A speaker can describe the target object O1 using either $w_1 =$ *on* O3 or $w_2 =$ *right of* O2. Suppose the embedded listener is L:LITERAL, which chooses uniformly from the set of objects consistent with the given utterance. In this scenario, $p_{\mathrm{L:LITERAL}}(\mathrm{O1} \mid w_1) = 1$ because w_1 unambiguously describes the target object, but $p_{\mathrm{L:LITERAL}}(\mathrm{O1} \mid w_2) = \frac{1}{2}$. The rational speaker S(L:LITERAL) would therefore choose w_1, achieving a utility of 1, which is an improvement over the reflex speaker S:LITERAL's utility of $\frac{3}{4}$.

[1] Semantic validity is approximated by a set of heuristic rules (e.g. *left* is all positions with smaller x-coordinates).

[2] If there are ties, any distribution over the utterances having the same utility is optimal.

4 From Literal Speaker to Learned Speaker

In the previous section, we showed that a literal strategy, one that considers only semantically valid choices, can be used to directly construct a reflex speaker S:LITERAL or an embedded listener in a rational speaker S(L:LITERAL). This section focuses on an orthogonal direction: improving literal strategies with learning. Specifically, we construct learned strategies from log-linear models trained on human annotations. These learned strategies can then be used to construct reflex and rational speaker variants—S:LEARNED and S(L:LEARNED), respectively.

4.1 Training a Log-Linear Speaker/Listener

We train the speaker, S:LEARNED, (similarly, listener, L:LEARNED) on training examples which comprise the utterances produced by the human annotators (see Section 6.1 for details on how this data was collected). Each example consists of a 3D model of a room in a house that specifies the 3D positions of each object and the coordinates of a 3D camera. When training the speaker, each example is a pair (o, w), where o is the input target object and w is the output utterance. When training the listener, each example is (w, g), where w is the input utterance and g is the output guessed object.

For now, an utterance w consists of two parts:

- A spatial preposition $w.r$ (e.g., *right of*) from a set of possible prepositions.[3]
- A reference object $w.o$ (e.g., O3) from the set of objects in the room.

We consider more complex utterances in Section 5.

Both S:LEARNED and L:LEARNED are parametrized by log-linear models:

$$p_{\text{S:LEARNED}}(w|o;\theta_{\text{S}}) \propto \exp\{\theta_{\text{S}}^{\top}\phi(o,w)\} \quad (4)$$

$$p_{\text{L:LEARNED}}(g|w;\theta_{\text{L}}) \propto \exp\{\theta_{\text{L}}^{\top}\phi(g,w)\} \quad (5)$$

where $\phi(\cdot,\cdot)$ is the feature vector (see below), θ_{S} and θ_{L} are the parameter vectors for speaker and listener. Note that the speaker and listener use the same set of features, but they have different parameters. Furthermore, the first normalization sums over possible utterances w while the second normalization sums over possible objects g in the scene. The two parameter vectors are trained to optimize the log-likelihood of the training data under the respective models.

Features We now describe the features $\phi(o, w)$. These features draw inspiration from Landau and Jackendoff (1993) and Tellex and Roy (2009).

Each object o in the 3D scene is represented by its bounding box, which is the smallest rectangular prism containing o. The following are functions of the camera, target (or guessed object) o, and the reference object $w.o$ in the utterance. The full set of features is obtained by conjoining these functions with indicator functions of the form $\mathbb{I}[w.r = r]$, where r ranges over the set of valid prepositions.

- *Proximity functions* measure the distance between o and $w.o$. This is implemented as the minimum over all the pairwise Euclidean distances between the corners of the bounding boxes. We also have indicator functions for whether o is the closest object, among the top 5 closest objects, and among the top 10 closest objects to $w.o$.
- *Topological functions* measure containment between o and $w.o$: $vol(o \cap w.o)/vol(o)$ and $vol(o \cap w.o)/vol(w.o)$. To simplify volume computation, we approximate each object by a bounding box that is aligned with the camera axes.
- *Projection functions* measure the relative position of the bounding boxes with respect to one another. Specifically, let v be the vector from the center of $w.o$ to the center of o. There is a function for the projection of v onto each of the axes defined by the camera orientation (see Figure 5). Additionally, there is a set of indicator functions that capture the relative magnitude of these projections. For example, there is a indicator function denoting whether the projection of v onto the camera's x-axis is the largest of all three projections.

[3] We chose 10 prepositions commonly used by people to describe objects in a preliminary data gathering experiment. This list includes multi-word units, which function equivalently to prepositions, such as *left of*.

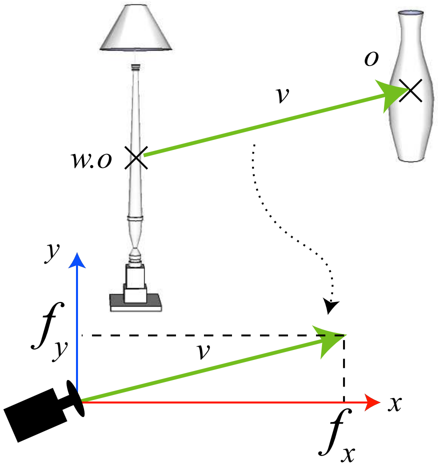


Figure 5: The projection features are computed by projecting a vector v extending from the center of the reference object to the center of the target object onto the camera axes f_x and f_y.

5 Handling Complex Utterances

So far, we have only considered speakers and listeners that deal with utterances consisting of one preposition and one reference object. We now extend these strategies to handle more complex utterances. Specifically, we consider utterances that conform to the following grammar:[4]

[noun]	N	→	*something* | O1 | O2 | ···
[relation]	R	→	*in front of* | *on* | ···
[conjunction]	NP	→	N RP*
[relativization]	RP	→	R NP

This grammar captures two phenomena of language use, conjunction and relativization.

- Conjunction is useful when one spatial relation is insufficient to disambiguate the target object. For example, in Figure 1, *right of* O2 could refer to the vase or the table, but using the conjunction *right of* O2 *and on* O3 narrows down the target object to just the vase.
- The main purpose of relativization is to refer to objects without a precise nominal descriptor. With complex utterances, it is possible to chain relative prepositional phrases, for example, using *on something right of* O2 to refer to the vase.

Given an utterance w, we define its *complexity* $|w|$ as the number of applications of the relativization rule, RP → R NP, used to produce w. We had only considered utterances of complexity 1 in previous sections.

5.1 Example Utterances

To illustrate the types of utterances available under the grammar, again consider the scene in Figure 1.

Utterances of complexity 2 can be generated either using the relativization rule exclusively, or both the conjunction and relativization rules. The relativization rule can be used to generate the following utterances:

- *on something that is right of* O2
- *right of something that is left of* O3

Applying the conjunction rule leads to the following utterances:

- *right of* O2 *and on* O3
- *right of* O2 *and under* O1
- *left of* O1 *and left of* O3

Note that we inserted the words *that is* after each N and the word *and* between every adjacent pair of RPs generated via the conjunction rule. This is to help a human listener interpret an utterance.

5.2 Extending the Rational Speaker

Suppose we have a rational speaker S(L) defined in terms of an embedded listener L which operates over utterances of complexity 1. We first extend L to interpret arbitrary utterances of our grammar. The rational speaker (defined in (2)) automatically inherits this extension.

Compositional semantics allows us to define the interpretation of complex utterances in terms of simpler ones. Specifically, each node in the parse tree has a *denotation*, which is computed recursively in terms of the node's children via a set of simple rules. Usually, denotations are represented as lambda-calculus functions, but for us, they will be distributions over objects in the scene. As a base case for interpreting utterances of complexity 1, we can use either L:LITERAL or L:LEARNED (defined in Sections 3 and 4).

[4] Naturally, we disallow direct reference to the target object.

Given a subtree w rooted at $u \in \{\text{N}, \text{NP}, \text{RP}\}$, we define the denotation of w, $[\![w]\!]$, to be a distribution over the objects in the scene in which the utterance was generated. The listener strategy $p_{\text{L}}(g|w) = [\![w]\!]$ is recursively as follows:

- If w is rooted at N with a single child x, then $[\![w]\!]$ is the uniform distribution over $\mathcal{N}(x)$, the set of objects consistent with the word x.
- If w is rooted at NP, we recursively compute the distributions over objects g for each child tree, multiply the probabilities, and renormalize (Hinton, 1999).
- If w is rooted at RP with relation r, we recursively compute the distribution over objects g' for the child NP tree. We then appeal to the base case to produce a distribution over objects g which are related to g' via relation r.

This strategy is defined formally as follows:

$$p_{\text{L}}(g \mid w) \propto \begin{cases} \mathbb{I}[g \in \mathcal{N}(x)] & w = (\text{N}\, x) \\ \prod_{j=1}^{k} p_{\text{L}}(g \mid w_j) & w = (\text{NP}\, w_1 \dots w_k) \\ \sum_{g'} p_{\text{L}}(g \mid (r, g')) p_{\text{L}}(g' \mid w') & w = (\text{RP}\, (\text{R}\, r)\, w') \end{cases} \tag{6}$$

Figure 6 shows an example of this bottom-up denotation computation for the utterance *on something right of* O2 with respect to the scene in Figure 1. The denotation starts with the lowest NP node $[\![\text{O2}]\!]$, which places all the mass on O2 in the scene. Moving up the tree, we compute the denotation of the RP, $[\![$*right of* O2$]\!]$, using the RP case of (6), which results in a distribution that places equal mass on O1 and O3.[5] The denotation of the N node $[\![$*something*$]\!]$ is a flat distribution over all the objects in the scene. Continuing up the tree, the denotation of the NP is computed by taking a product of the object distributions, and turns out to be exactly the same split distribution as its RP child. Finally, the denotation at the root is computed by applying the base case to *on* and the resulting distribution from the previous step.

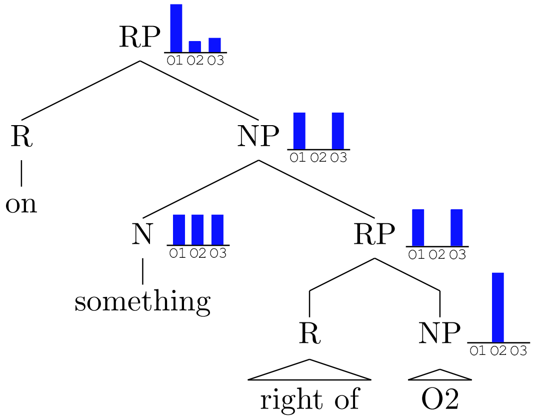


Figure 6: The listener model maps an utterance to a distribution over objects in the room. Each internal NP or RP node is a distribution over objects in the room.

Generation So far, we have defined the listener strategy $p_{\text{L}}(g \mid w)$. Given target o, the rational speaker S(L) with respect to this listener needs to compute $\text{argmax}_w\, p_{\text{L}}(o \mid w)$ as dictated by (3). This maximization is performed by enumerating all utterances of bounded complexity.

5.3 Modeling Listener Confusion

One shortcoming of the previous approach for extending a listener is that it falsely assumes that a listener can reliably interpret a simple utterance just as well as it can a complex utterance.

We now describe a more realistic speaker which is robust to listener confusion. Let $\alpha \in [0, 1]$ be a *focus parameter* which determines the confusion level. Suppose we have a listener L. When presented with an utterance w, for each application of the relativization rule, we have a $1 - \alpha$ probability of losing focus. If we stay focused for the entire utterance (with probability $\alpha^{|w|}$), then we interpret the utterance according to p_{L}. Otherwise (with probability $1 - \alpha^{|w|}$), we guess an object at random according to $p_{\text{rnd}}(g \mid w)$. We then use (3) to define the rational speaker S(L) with respect the following "confused listener" strategy:

$$\tilde{p}_{\text{L}}(g \mid w) = \alpha^{|w|} p_{\text{L}}(g \mid w) + (1 - \alpha^{|w|}) p_{\text{rnd}}(g \mid w). \tag{7}$$

As $\alpha \to 0$, the confused listener is more likely to make a random guess, and thus there is a stronger penalty against using more complex utterances. As

[5] It is worth mentioning that this split distribution between O1 and O3 represents the ambiguity mentioned in Section 3 when discussing the shortcomings of S:LITERAL.

$\alpha \to 1$, the confused listener converges to p_{L} and the penalty for using complex utterances vanishes.

5.4 The Taboo Setting

Notice that the rational speaker as defined so far does not make full use of our grammar. Specifically, the rational speaker will never use the "wildcard" noun *something* nor the relativization rule in the grammar because an NP headed by the wildcard *something* can always be replaced by the object ID to obtain a higher utility. For instance, in Figure 6, the NP spanning *something right of* O2 can be replaced by O3.

However, it is not realistic to assume that all objects can be referenced directly. To simulate scenarios where some objects cannot be referenced directly (and to fully exercise our grammar), we introduce the *taboo setting*. In this setting, we remove from the lexicon some fraction of the object IDs which are closest to the target object. Since the tabooed objects cannot be referenced directly, a speaker must resort to use of the wildcard *something* and relativization.

For example, in Figure 7, we enable tabooing around the target O1. This prevents the speaker from referring directly to O3, so the speaker is forced to describe O3 via the relativization rule, for example, producing *something right of* O2.

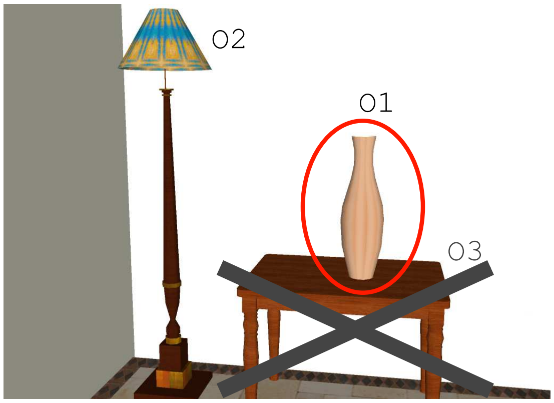


Figure 7: With tabooing enabled around O1, O3 can no longer be referred to directly (represented by an X).

6 Experiments

We now present our empirical results, showing that rational speakers, who have embedded models of listeners, can communicate more successfully than reflex speakers, who do not.

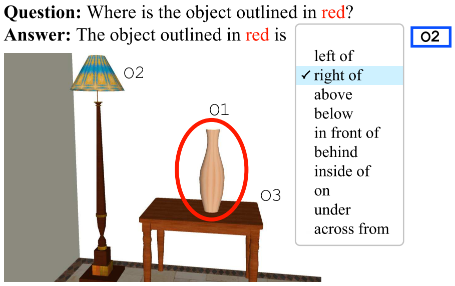

Question: Where is the object outlined in red?
Answer: The object outlined in red is

Figure 8: Mechanical Turk speaker task: Given the target object (e.g., O1), a human speaker must choose an utterance to describe the object (e.g., *right of* O2).

6.1 Setup

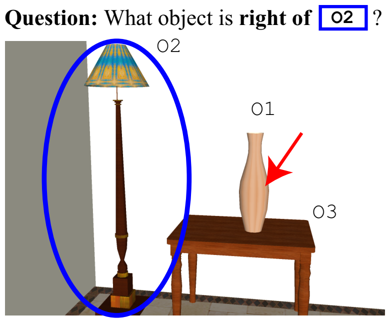
We collected 43 scenes (rooms) from the Google Sketchup 3D Warehouse, each containing an average of 22 objects (household items and pieces of furniture arranged in a natural configuration). For each object o in a scene, we create a *scenario*, which represents an instance of the communication game with o as the target object. There are a total of 2,860 scenarios, which we split evenly into a training set (denoted Tr) and a test set (denoted Ts).

We created the following two Amazon Mechanical Turk tasks, which enable humans to play the language game on the scenarios:

Speaker Task In this task, human annotators play the role of speakers in the language game. They are prompted with a target object o and asked to each produce an utterance w (by selecting a preposition $w.r$ from a dropdown list and clicking on a reference object $w.o$) that best informs a listener of the identity of the target object.

For each training scenario o, we asked three speakers to produce an utterance w. The three resulting (o, w) pairs are used to train the learned reflex speaker (S:Literal). These pairs were also used to train the learned reflex listener (L:Literal), where the target o is treated as the guessed object. See Section 4.1 for the details of the training procedure.

Listener Task In this task, human annotators play the role of listeners. Given an utterance generated by a speaker (human or not), the human listener must



Question: What object is **right of** O2 ?

Figure 9: Mechanical Turk listener task: a human listener is prompted with an utterance generated by a speaker (e.g., *right of* O2), and asked to click on an object (shown by the red arrow).

guess the target object that the speaker saw by clicking on an object. The purpose of the listener task is to evaluate speakers, as described in the next section.

6.2 Evaluation

Utility (Communicative Success) We primarily evaluate a speaker by its ability to communicate successfully with a human listener. For each test scenario, we asked three listeners to guess the object. We use $p_{\text{L:HUMAN}}(g \mid w)$ to denote the distribution over guessed objects g given prompt w. For example, if two of the three listeners guessed O1, then $p_{\text{L:HUMAN}}(\texttt{O1} \mid w) = \frac{2}{3}$. The expected utility (2) is then computed by averaging the utility (communicative success) over the test scenarios TS:

$$\begin{aligned}\text{SUCCESS}(\text{S}) &= \text{EU}(\text{S}, \text{L:HUMAN}) \qquad (8)\\ &= \frac{1}{|\text{TS}|}\sum_{o\in \text{TS}}\sum_{w} p_{\text{S}}(w|o)p_{\text{L:HUMAN}}(o|w).\end{aligned}$$

Exact Match As a secondary evaluation metric, we also measure the ability of our speaker to exactly match an utterance produced by a human speaker. Note that since there are many ways of describing an object, exact match is neither necessary nor sufficient for successful communication.

We asked three human speakers to each produce an utterance w given a target o. We use $p_{\text{S:HUMAN}}(w \mid o)$ to denote this distribution; for example, $p_{\text{S:HUMAN}}(\textit{right of}\ \texttt{O2} \mid o) = \frac{1}{3}$ if exactly one of the three speakers uttered *right of* O2. We then define the *exact match* of a speaker S as follows:

$$\text{MATCH}(\text{S}) = \frac{1}{|\text{TS}|}\sum_{o\in \text{TS}}\sum_{w} p_{\text{S:HUMAN}}(w \mid o)p_{\text{S}}(w \mid o). \tag{9}$$

Speaker	Success	Exact Match
S:LITERAL [reflex]	4.62%	1.11%
S(L:LITERAL) [rational]	33.65%	2.91%
S:LEARNED [reflex]	38.36%	5.44%
S(L:LEARNED) [rational]	**52.63%**	14.03%
S:HUMAN	41.41%	**19.95%**

Table 1: Comparison of various speakers on communicative success and exact match, where only utterances of complexity 1 are allowed. The rational speakers (with respect to both the literal listener L:LITERAL and the learned listener L:LEARNED) perform better than their reflex counterparts. While the human speaker (composed of three people) has higher exact match (it is better at mimicking itself), the rational speaker S(L:LEARNED) actually achieves higher communicative success than the human listener.

6.3 Reflex versus Rational Speakers

We first evaluate speakers in the setting where only utterances of complexity 1 are allowed. Table 1 shows the results on both success and exact match. First, our main result is that the two rational speakers S(L:LITERAL) and S(L:LEARNED), which each model a listener explicitly, perform significantly better than the corresponding reflex speakers, both in terms of success and exact match.

Second, it is natural that the speakers that involve learning (S:LITERAL and S(L:LITERAL)) outperform the speakers that only consider the literal meaning of utterances (S:LEARNED and S(L:LEARNED)), as the former models capture subtler preferences using features.

Finally, we see that in terms of exact match, the human speaker S:HUMAN performs the best (this is not surprising because human exact match is essentially the inter-annotator agreement), but in terms of communicative success, S(L:LEARNED) achieves a higher success rate than S:HUMAN, suggesting that the game-theoretic modeling undertaken by the rational speakers is effective for communication, which is ultimate goal of language.

Note that exact match is low even for the "human speaker", since there are often many equally good

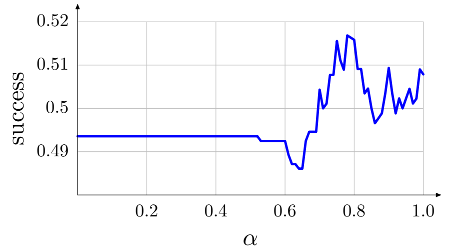


Figure 10: Communicative success as a function of focus parameter α without tabooing on TSDEV. The optimal value of α is obtained at 0.79.

ways to evoke an object. At the same time, the success rates for all speakers are rather low, reflecting the fundamental difficulty of the setting: sometimes it is impossible to unambiguously evoke the target object via short utterances. In the next section, we show that we can improve the success rate by allowing the speakers to generate more complex utterances.

6.4 Generating More Complex Utterances

We now evaluate the rational speaker S(L:LEARNED) when it is allowed to generate utterances of complexity 1 or 2. Recall from Section 5.3 that the speaker depends on a focus parameter α, which governs the embedded listener's ability to interpret the utterance. We divided the test set (TS) in two halves: TSDEV, which we used to tune the value of α and TSFINAL, which we used to evaluate success rates.

Figure 10 shows the communicative success as a function of α on TSDEV. When α is small, the embedded listener is confused more easily by more complex utterances; therefore the speaker tends to choose mostly utterances of complexity 1. As α increases, the utterances increase in complexity, as does the success rate. However, when α approaches 1, the utterances are too complex and the success rate decreases. The dependence between α and average utterance complexity is shown in Figure 11.

Table 2 shows the success rates on TSFINAL for $\alpha \to 0$ (all utterances have complexity 1), $\alpha = 1$ (all utterances have complexity 2), and α tuned to maximize the success rate based on TSDEV. Setting α in this manner allows us to effectively balance complexity and ambiguity, resulting in an improvement in the success rate.

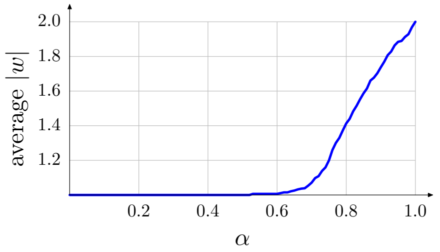


Figure 11: Average utterance complexity as a function of the focus parameter α on TSDEV. Higher values of α yield more complex utterances.

Taboo Amount	Success ($\alpha \to 0$)	Success ($\alpha = 1$)	Success ($\alpha = \alpha^*$)	α^*
0%	51.78%	50.99%	54.53%	0.79
5%	38.75%	40.83%	43.12%	0.89
10%	29.57%	29.69%	30.30%	0.80
30%	12.40%	13.04%	12.98%	0.81

Table 2: Communicative success (on TSFINAL) of the rational speaker S(L:LEARNED) for various values of α across different taboo amounts. When the taboo amount is small, small values of α lead to higher success rates. As the taboo amount increases, larger values of α (resulting in more complex utterances) are better.

7 Conclusion

Starting with the view that the purpose of language is successful communication, we developed a game-theoretic model in which a *rational* speaker generates utterances by explicitly taking the listener into account. On the task of generating spatial descriptions, we showed the rational speaker substantially outperforms a baseline reflex speaker that does not have an embedded model. Our results therefore suggest that a model of the pragmatics of communication is an important factor to consider for generation.

Acknowledgements This work was supported by the National Science Foundation through a Graduate Research Fellowship to the first two authors. We also would like to acknowledge Surya Murali, the designer of the 3D Google Sketchup models, and thank the anonymous reviewers for their comments.

References

J. L. Austin. 1962. *How to do Things with Words: The William James Lectures delivered at Harvard Univer-*

sity in 1955. Oxford, Clarendon, UK.

S. Branavan, H. Chen, L. S. Zettlemoyer, and R. Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Singapore. Association for Computational Linguistics.

D. L. Chen and R. J. Mooney. 2008. Learning to sportscast: A test of grounded language acquisition. In *International Conference on Machine Learning (ICML)*, pages 128–135. Omnipress.

David DeVault and Matthew Stone. 2007. Managing ambiguities across utterances in dialogue.

J. Eisenstein, J. Clarke, D. Goldwasser, and D. Roth. 2009. Reading to learn: Constructing features from semantic abstracts. In *Empirical Methods in Natural Language Processing (EMNLP)*, Singapore.

J. Feldman and S. Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain and Language*, 89:385–392.

M. Fleischman and D. Roy. 2007. Representing intentions in a cognitive model of language acquisition: Effects of phrase structure on situated verb learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, Cambridge, MA. MIT Press.

M. C. Frank, N. D. Goodman, and J. B. Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–585.

Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. In *Journal of Artificial Intelligence Research*, volume 21, pages 429–470.

H. P. Grice. 1975. Syntax and Semantics; Logic and Conversation. 3:Speech Acts:41–58.

G. Hinton. 1999. Products of experts. In *International Conference on Artificial Neural Networks (ICANN)*.

G. Jäger. 2008. Game theory in semantics and pragmatics. Technical report, University of Tübingen.

T. Kollar, S. Tellex, D. Roy, and N. Roy. 2010. Toward understanding natural language directions. In *Human-Robot Interaction*, pages 259–266.

Barbara Landau and Ray Jackendoff. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2spatial prepositions analysis, cross linguistic conceptual similarities; comments/response):217–238.

P. Liang, M. I. Jordan, and D. Klein. 2009. Learning semantic correspondences with less supervision. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Singapore. Association for Computational Linguistics.

S. T. Piantadosi, N. D. Goodman, B. A. Ellis, and J. B. Tenenbaum. 2008. A Bayesian model of the acquisition of compositional semantics. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*.

T Regier and LA Carlson. 2001. Journal of experimental psychology. general; grounding spatial language in perception: an empirical and computational investigation. 130(2):273–298.

Stefanie Tellex and Deb Roy. 2009. Grounding spatial prepositions for video search. In *ICMI*.

Y. W. Wong and R. J. Mooney. 2007. Learning synchronous grammars for semantic parsing with lambda calculus. In *Association for Computational Linguistics (ACL)*, pages 960–967, Prague, Czech Republic. Association for Computational Linguistics.

C. Yu and D. H. Ballard. 2004. On the integration of grounding language and learning objects. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 488–493, Cambridge, MA. MIT Press.

L. S. Zettlemoyer and M. Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Uncertainty in Artificial Intelligence (UAI)*, pages 658–666.