Translating Neuralese

Jacob Andreas **Anca Dragan** **Dan Klein**
Computer Science Division
University of California, Berkeley
{jda,anca,klein}@cs.berkeley.edu

Abstract

Several approaches have recently been proposed for learning decentralized deep multiagent policies that coordinate via a differentiable communication channel. While these policies are effective for many tasks, interpretation of their induced communication strategies has remained a challenge. Here we propose to interpret agents' messages by translating them. Unlike in typical machine translation problems, we have no parallel data to learn from. Instead we develop a translation model based on the insight that agent messages and natural language strings mean the same thing *if they induce the same belief about the world in a listener*. We present theoretical guarantees and empirical evidence that our approach preserves both the semantics and pragmatics of messages by ensuring that players communicating through a translation layer do not suffer a substantial loss in reward relative to players with a common language.[1]

1 Introduction

Several recent papers have described approaches for learning *deep communicating policies* (DCPs): decentralized representations of behavior that enable multiple agents to communicate via a differentiable channel that can be formulated as a recurrent neural network. DCPs have been shown to solve a variety of coordination problems, including reference games (Lazaridou et al., 2016b), logic puzzles (Foerster et al., 2016), and simple control (Sukhbaatar et al., 2016). Appealingly, the agents' communication protocol can be learned via direct backpropagation through the communication channel, avoiding many of the challenging inference problems associated with learning in classical decentralized decision processes (Roth et al., 2005).

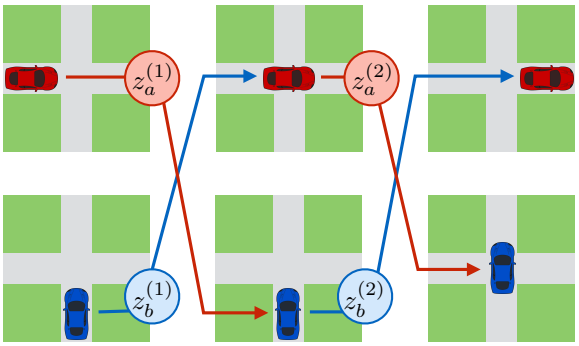


Figure 1: Example interaction between a pair of agents in a deep communicating policy. Both cars are attempting to cross the intersection, but cannot see each other. By exchanging message vectors $z^{(t)}$, the agents are able to coordinate and avoid a collision. This paper presents an approach for *understanding* the contents of these message vectors by translating them into natural language.

But analysis of the strategies induced by DCPs has remained a challenge. As an example, Figure 1 depicts a driving game in which two cars, which are unable to see each other, must both cross an intersection without colliding. In order to ensure success, it is clear that the cars must communicate with each other. But a number of successful communication strategies are possible—for example, they might report their exact (x, y) coordinates at every timestep, or they might simply announce whenever they are entering and leaving the intersection. If these messages were communicated in natural language, it would be straightforward to determine which strategy was being employed. However, DCP agents instead communicate with an automatically induced protocol of unstructured, real-valued recurrent state vectors—an artificial language we might call "neuralese," which superficially bears little resemblance to natural language, and thus frustrates attempts at direct interpretation.

[1] We have released code and data at http://github.com/jacobandreas/neuralese.

We propose to understand neuralese messages by *translating* them. In this work, we present a simple technique for inducing a dictionary that maps between neuralese message vectors and short natural language strings, given only examples of DCP agents interacting with other agents, and humans interacting with other humans. Natural language already provides a rich set of tools for describing beliefs, observations, and plans—our thesis is that these tools provide a useful complement to the visualization and ablation techniques used in previous work on understanding complex models (Strobelt et al., 2016; Ribeiro et al., 2016).

While structurally quite similar to the task of machine translation between pairs of human languages, interpretation of neuralese poses a number of novel challenges. First, there is no natural source of parallel data: there are no bilingual "speakers" of both neuralese and natural language. Second, there may not be a direct correspondence between the strategy employed by humans and DCP agents: even if it were constrained to communicate using natural language, an automated agent might choose to produce a different message from humans in a given state. We tackle both of these challenges by appealing to the grounding of messages in gameplay. Our approach is based on one of the core insights in natural language semantics: messages (whether in neuralese or natural language) have similar meanings *when they induce similar beliefs about the state of the world*.

Based on this intuition, we introduce a translation criterion that matches neuralese messages with natural language strings by minimizing statistical distance in a common representation space of distributions over speaker states. We explore several related questions:

- What makes a good translation, and under what conditions is translation possible at all? (Section 4)
- How can we build a model to translate between neuralese and natural language? (Section 5)
- What kinds of theoretical guarantees can we provide about the behavior of agents communicating via this translation model? (Section 6)

Our translation model and analysis are general, and in fact apply equally to human–computer and human–human translation problems grounded in gameplay. In this paper, we focus our experiments specifically on the problem of interpreting communication in deep policies, and apply our approach to the driving game in Figure 1 and two reference games of the kind shown in Figure 2. We find that this approach outperforms a more conventional machine translation criterion both when attempting to interoperate with neuralese speakers and when predicting their state.

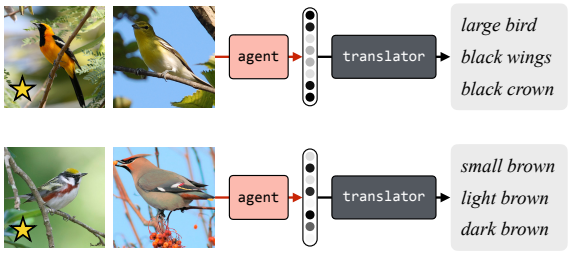


Figure 2: Overview of our approach—best-scoring translations generated for a reference game involving images of birds. The speaking agent's goal is to send a message that uniquely identifies the bird on the left. From these translations it can be seen that the learned model appears to discriminate based on coarse attributes like size and color.

2 Related work

A variety of approaches for learning deep policies with communication were proposed essentially simultaneously in the past year. We have broadly labeled these as "deep communicating policies"; concrete examples include Lazaridou et al. (2016b), Foerster et al. (2016), and Sukhbaatar et al. (2016). The policy representation we employ in this paper is similar to the latter two of these, although the general framework is agnostic to low-level modeling details and could be straightforwardly applied to other architectures. Analysis of communication strategies in all these papers has been largely ad-hoc, obtained by clustering states from which similar messages are emitted and attempting to manually assign semantics to these clusters. The present work aims at developing tools for performing this analysis automatically.

Most closely related to our approach is that of Lazaridou et al. (2016a), who also develop a model for assigning natural language interpretations to learned messages; however, this approach relies on supervised cluster labels and is targeted specifically towards referring expression games. Here we attempt to develop an approach that can handle general multiagent interactions without assuming a prior discrete structure in space of observations.

The literature on learning decentralized multiagent policies in general is considerably larger (Bernstein et al., 2002; Dibangoye et al., 2016). This includes work focused on communication in multiagent settings (Roth et al., 2005) and even communication using natural language messages (Vogel et al., 2013b). All of these approaches employ structured communication schemes with manually engineered messaging protocols; these are, in some sense, automatically interpretable, but at the cost of introducing considerable complexity into both training and inference.

Our evaluation in this paper investigates communication strategies that arise in a number of different games, including reference games and an extended-horizon driving game. Communication strategies for reference games were previously explored by Vogel et al. (2013a), Andreas and Klein (2016) and Kazemzadeh et al. (2014), and reference games specifically featuring end-to-end communication protocols by Yu et al. (2016). On the control side, a long line of work considers nonverbal communication strategies in multiagent policies (Dragan and Srinivasa, 2013).

Another group of related approaches focuses on the development of more general machinery for interpreting deep models in which messages have no explicit semantics. This includes both visualization techniques (Zeiler and Fergus, 2014; Strobelt et al., 2016), and approaches focused on generating explanations in the form of natural language (Hendricks et al., 2016; Vedantam et al., 2017).

3 Problem formulation

Games Consider a cooperative game with two players a and b of the form given in Figure 3. At every step t of this game, player a makes an observation $x_a^{(t)}$ and receives a message $z_b^{(t-1)}$ from b. It then takes an action $u_a^{(t)}$ and sends a message $z_a^{(t)}$ to b. (The process is symmetric for b.) The distributions $p(u_a|x_a, z_b)$ and $p(z_a|x_a)$ together define a policy π which we assume is shared by both players, i.e. $p(u_a|x_a, z_b) = p(u_b|x_b, z_a)$ and $p(z_a|x_a) = p(z_b|x_b)$. As in a standard Markov decision process, the actions $(u_a^{(t)}, u_b^{(t)})$ alter the world state, generating new observations for both players and a reward shared by both.

The distributions $p(z|x)$ and $p(u|x, z)$ may also be viewed as defining a *language*: they specify how a speaker will generate messages based on world states, and how a listener will respond to these messages. Our goal in this work is to learn to translate between pairs of languages generated by different policies. Specifically, we assume that we have access to two policies for the same game: a "robot policy" π_r and a "human policy" π_h. We would like to use the representation of π_h, the behavior of which is transparent to human users, in order to *understand* the behavior of π_r (which is in general an uninterpretable learned model); we will do this by inducing bilingual dictionaries that map message vectors z_r of π_r to natural language strings z_h of π_h and vice-versa.

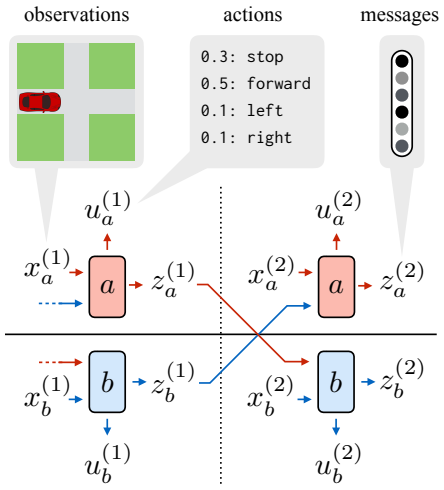


Figure 3: Schematic representation of communication games. At every timestep t, players a and b make an observation $x^{(t)}$ and receive a message $z^{(t-1)}$, then produce an action $u^{(t)}$ and a new message $z^{(t)}$.

Learned agents π_r Our goal is to present tools for interpretation of learned messages that are agnostic to the details of the underlying algorithm for acquiring them. We use a generic DCP model as a basis for the techniques developed in this paper. Here each agent policy is represented as a deep recurrent Q network (Hausknecht and Stone, 2015). This network is built from communicating cells of the kind depicted in Figure 4. At every timestep, this agent receives three pieces of information: an

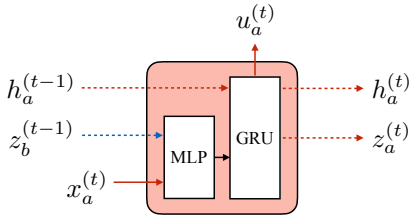


Figure 4: Cell implementing a single step of agent communication (compare with Sukhbaatar et al. (2016) and Foerster et al. (2016)). *MLP* denotes a multilayer perceptron; *GRU* denotes a gated recurrent unit (Cho et al., 2014). Dashed lines represent recurrent connections.

observation of the current state of the world, the agent's memory vector from the previous timestep, and a message from the other player. It then produces three outputs: a predicted Q value for every possible action, a new memory vector for the next timestep, and a message to send to the other agent.

Sukhbaatar et al. (2016) observe that models of this form may be viewed as specifying a single RNN in which weight matrices have a particular block structure. Such models may thus be trained using the standard recurrent Q-learning objective, with communication protocol learned end-to-end.

Human agents π_h The translation model we develop requires a representation of the distribution over messages $p(z_a|x_a)$ employed by human speakers (without assuming that humans and agents produce equivalent messages in equivalent contexts). We model the human message generation process as categorical, and fit a simple multilayer perceptron model to map from observations to words and phrases used during human gameplay.

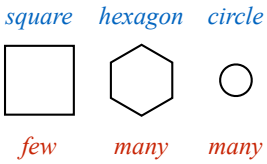
4 What's in a translation?

What does it mean for a message z_h to be a "translation" of a message z_r? In standard machine translation problems, the answer is that z_h is likely to co-occur in parallel data with z_r; that is, $p(z_h|z_r)$ is large. Here we have no parallel data: even if we could observe natural language and neuralese messages produced by agents in the same state, we would have no guarantee that these messages actually served the same function. Our answer must instead appeal to the fact that both natural language and neuralese messages are grounded in a common environment. For a given neuralese message z_r, we will first compute a grounded representation of that message's meaning; to translate, we find a natural-language message whose meaning is most similar. The key question is then what form this grounded meaning representation should take. The existing literature suggests two broad approaches:

Semantic representation The meaning of a message z_a is given by its denotations: that is, by the set of world states of which z_a may be felicitously predicated, given the existing context available to a listener. In probabilistic terms, this says that the meaning of a message z_a is represented by the distribution $p(x_a|z_a, x_b)$ it induces over speaker states. Examples of this approach include Guerin and Pitt (2001) and Pasupat and Liang (2016).

Pragmatic representation The meaning of a message z_a is given by the behavior it induces in a listener. In probabilistic terms, this says that the meaning of a message z_a is represented by the distribution $p(u_b|z_a, x_b)$ it induces over actions given the listener's observation x_b. Examples of this approach include Vogel et al. (2013a) and Gauthier and Mordatch (2016).

These two approaches can give rise to rather different behaviors. Consider the following example:



The top language (in blue) has a unique name for every kind of shape, while the bottom language (in red) only distinguishes between shapes with few sides and shapes with many sides. Now imagine a simple reference game with the following form: player a is covertly assigned one of these three shapes as a reference target, and communicates that reference to b; b must then pull a lever labeled `large` or `small` depending on the size of the target shape. Blue language speakers can achieve perfect success at this game, while red language speakers can succeed at best two out of three times.

How should we translate the blue word *hexagon* into the red language? The semantic approach suggests that we should translate *hexagon* as *many*: while *many* does not uniquely identify the hexagon, it produces a distribution over shapes that is closest to the truth. The pragmatic approach instead suggests that we should translate *hexagon* as *few*, as this is the only message that guarantees that the listener will pull the correct lever `large`. So in order to produce a correct listener action, the translator might have to "lie" and produce a maximally inaccurate listener belief.

If we were exclusively concerned with building a translation layer that allowed humans and DCP agents to interoperate as effectively as possible, it would be natural to adopt a pragmatic representation strategy. But our goals here are broader: we also want to facilitate *understanding*, and specifically to help users of learned systems form true beliefs about the systems' computational processes and representational abstractions. The example above demonstrates that "pragmatically" optimizing directly for task performance can sometimes lead to translations that produce inaccurate beliefs.

We instead build our approach around semantic representations of meaning. By preserving semantics, we allow listeners to reason accurately about the content and interpretation of messages. We might worry that by adopting a semantics-first view, we have given up all guarantees of effective interoperation between humans and agents using a translation layer. Fortunately, this is not so: as we will see in Section 6, it is possible to show that players communicating via a semantic translator perform only boundedly worse (and sometimes better!) than pairs of players with a common language.

5 Translation models

In this section, we build on the intuition that messages should be translated via their semantics to define a concrete translation model—a procedure for constructing a natural language $\leftrightarrow$ neuralese dictionary given agent and human interactions.

We understand the meaning of a message z_a to be represented by the distribution $p(x_a|z_a, x_b)$ it induces over speaker states given listener context. We can formalize this by defining the belief distribution β for a message z and context x_b as:

$$\beta(z_a, x_b) = p(x_a|z_a, x_b) = \frac{p(z_a|x_a)p(x_b|x_a)}{\sum_{x'_a} p(z_a|x'_a)p(x_b|x'_a)}$$

Here we have modeled the listener as performing a single step of Bayesian inference, using the listener state and the message generation model (by assumption shared between players) to compute the posterior over speaker states. While in general neither humans nor DCP agents compute explicit representations of this posterior, past work has found that both humans and suitably-trained neural networks can be modeled as Bayesian reasoners (Frank et al., 2009; Paige and Wood, 2016).

This provides a context-specific representation of belief, but for messages z and z' to have the same semantics, they must induce the same belief over *all* contexts in which they occur. In our probabilistic formulation, this introduces an outer expectation over contexts, providing a final measure q of the quality of a translation from z to z':

$$\begin{aligned} q(z, z') &= \mathbb{E}\big[\mathcal{D}_{\text{KL}}(\beta(z, X_b) \mid\mid \beta(z', X_b)) \mid z, z'\big] \\ &= \sum_{x_a, x_b} p(x_a, x_b|z, z')\mathcal{D}_{\text{KL}}(\beta(z, x_b) \mid\mid \beta(z', x_b)) \\ &\propto \sum_{x_a, x_b} p(x_a, x_b) \cdot p(z|x_a) \cdot p(z'|x_a) \\ &\qquad \cdot \mathcal{D}_{\text{KL}}(\beta(z, x_b) \mid\mid \beta(z', x_b)); \end{aligned} \tag{1}$$

Algorithm 1 Translating messages

given: a phrase inventory L
function TRANSLATE(z)
  return $\arg\min_{z' \in L} \hat{q}(z, z')$

function $\hat{q}(z, z')$
  // sample contexts and distractors
  $x_{ai}, x_{bi} \sim p(X_a, X_b)$ for $i = 1..n$
  $x'_{ai} \sim p(X_a|x_{bi})$
  // compute context weights
  $\tilde{w}_i \leftarrow p(z|x_{ai}) \cdot p(z'|x_{ai})$
  $w_i \leftarrow \tilde{w}_i / \sum_j \tilde{w}_j$
  // compute divergences
  $k_i \leftarrow \sum_{x \in \{x_a, x'_a\}} p(z|x) \log \frac{p(z|x)}{p(z'|x)}$
  return $\sum_i w_i k_i$

recalling that in this setting

$$\begin{aligned} \mathcal{D}_{\text{KL}}(\beta \mid\mid \beta') &= \sum_{x_a} p(x_a|z, x_b) \log \frac{p(x_a|z, x_b)}{p(x_a|z', x_b)} \\ &\propto \sum_{x_a} p(x_a|x_b)p(z|x_a) \log \frac{p(z|x_a)}{p(z'|x_a)} \end{aligned} \tag{2}$$

which is zero when the messages z and z' give rise to identical belief distributions and increases as they grow more dissimilar. To translate, we would like to compute $tr(z_r) = \arg\min_{z_h} q(z_r, z_h)$ and $tr(z_h) = \arg\min_{z_r} q(z_h, z_r)$. Intuitively, Equation 1 says that we will measure the quality of a proposed translation $z \mapsto z'$ by asking the following question: in contexts where z is likely to be used, how frequently does z' induce the same belief about speaker states as z?

While this translation criterion directly encodes the semantic notion of meaning described in Section 4, it is doubly intractable: the KL divergence and outer expectation involve a sum over all observations x_a and x_b respectively; these sums are not in general possible to compute efficiently. To avoid this, we approximate Equation 1 by sampling. We draw a collection of samples (x_a, x_b) from the prior over world states, and then generate for each sample a sequence of distractors (x'_a, x_b) from $p(x'_a|x_b)$ (we assume access to both of these distributions from the problem representation). The KL term in Equation 1 is computed over each true sample and its distractors, which are then normalized and averaged to compute the final score.

Sampling accounts for the outer $p(x_a, x_b)$ in Equation 1 and the inner $p(x_a|x_b)$ in Equation 2.

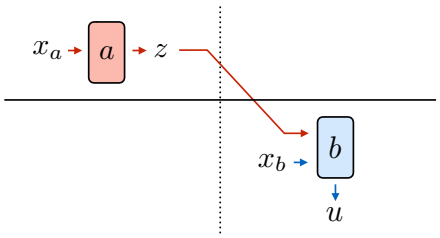


Figure 5: Simplified game representation used for analysis in Section 6. A speaker agent sends a message to a listener agent, which takes a single action and receives a reward.

The only quantities remaining are of the form $p(z|x_a)$. In the case of neuralese, this distribution already is part of the definition of the agent policy π_r and can be reused directly. For natural language, we use transcripts of human interactions to fit a model that maps from world states to a distribution over frequent utterances as discussed in Section 3. Details of these model implementations are provided in Appendix B, and the full translation procedure is given in Algorithm 1.

6 Belief and behavior

The translation criterion in the previous section makes no reference to listener actions at all. The shapes example in Section 4 shows that some model performance might be lost under translation. It is thus reasonable to ask whether this translation model of Section 5 can make any guarantees about the effect of translation on behavior. In this section we explore the relationship between belief-preserving translations and the behaviors they produce, by examining the effect of belief accuracy and strategy mismatch on the reward obtained by cooperating agents.

To facilitate this analysis, we consider a simplified family of communication games with the structure depicted in Figure 5. These games can be viewed as a subset of the family depicted in Figure 3; and consist of two steps: a listener makes an observation x_a and sends a single message z to a speaker, which makes its own observation x_b, takes a single action u, and receives a reward. We emphasize that the results in this section concern the theoretical properties of idealized games, and are presented to provide intuition about high-level properties of our approach. Section 8 investigates empirical behavior of this approach on real-world tasks where these ideal conditions do not hold.

Our first result is that translations that minimize semantic dissimilarity q cause the listener to take near-optimal actions:[2]

[2]Proof is provided in Appendix A.

Proposition 1.
Semantic translations reward rational listeners. Define a *rational listener* as one that chooses the best action in expectation over the speaker's state:

$$U(z, x_b) = \arg\max_u \sum_{x_a} p(x_a|x_b, z) r(x_a, x_b, u)$$

for a reward function $r \in [0, 1]$ that depends only on the two observations and the action.[3] Now let a be a speaker of a language r, b be a listener of the same language r, and b' be a listener of a different language h. Suppose that we wish for a and b' to interact via the translator $tr : z_r \mapsto z_h$ (so that a produces a message z_r, and b' takes an action $U(z_h = tr(z_r), x_{b'})$). If tr respects the semantics of z_r, then the bilingual pair a and b' achieves only boundedly worse reward than the monolingual pair a and b. Specifically, if $q(z_r, z_h) \leq D$, then

$$\mathbb{E}r(X_a, X_b, U(tr(Z))) \geq \mathbb{E}r(X_a, X_b, U(Z)) - \sqrt{2D} \quad (3)$$

So as discussed in Section 4, even by committing to a semantic approach to meaning representation, we have still succeeded in (approximately) capturing the nice properties of the pragmatic approach.

Section 4 examined the consequences of a mismatch between the set of primitives available in two languages. In general we would like some measure of our approach's robustness to the lack of an exact correspondence between two languages. In the case of humans in particular we expect that a variety of different strategies will be employed, many of which will not correspond to the behavior of the learned agent. It is natural to want some assurance that we can identify the DCP's strategy as long as *some* human strategy mirrors it. Our second observation is that it is possible to exactly recover a translation of a DCP strategy from a mixture of humans playing different strategies:

Proposition 2.
Semantic translations find hidden correspondences. Consider a fixed robot policy π_r and a set of human policies $\{\pi_{h1}, \pi_{h2}, \ldots\}$ (recalling from Section 3 that each π is defined by distributions

[3]This notion of rationality is a fairly weak one: it permits many suboptimal communication strategies, and requires only that the listener do as well as possible given a fixed speaker—a first-order optimality criterion likely to be satisfied by any richly-parameterized model trained via gradient descent.

$p(z|x_a)$ and $p(u|z, x_b)$). Suppose further that the messages employed by these human strategies are *disjoint*; that is, if $p_{hi}(z|x_a) > 0$, then $p_{hj}(z|x_a) = 0$ for all $j \neq i$. Now suppose that all $q(z_r, z_h) = 0$ for all messages in the support of some $p_{hi}(z|x_a)$ and > 0 for all $j \neq i$. Then every message z_r is translated into a message produced by π_{hi}, and messages from other strategies are ignored.

This observation follows immediately from the definition of $q(z_r, z_h)$, but demonstrates one of the key distinctions between our approach and a conventional machine translation criterion. Maximizing $p(z_h|z_r)$ will produce the natural language message most often produced in contexts where z_r is observed, regardless of whether that message is useful or informative. By contrast, minimizing $q(z_h, z_r)$ will find the z_h that corresponds most closely to z_r even when z_h is rarely used.

The disjointness condition, while seemingly quite strong, in fact arises naturally in many circumstances—for example, players in the driving game reporting their spatial locations in absolute vs. relative coordinates, or speakers in a color reference game (Figure 6) discriminating based on lightness vs. hue. It is also possible to relax the above condition to require that strategies be only *locally* disjoint (i.e. with the disjointness condition holding for each fixed x_a), in which case overlapping human strategies are allowed, and the recovered robot strategy is a context-weighted mixture of these.

7 Evaluation

7.1 Tasks

In the remainder of the paper, we evaluate the empirical behavior of our approach to translation. Our evaluation considers two kinds of tasks: reference games and navigation games. In a reference game (e.g. Figure 6a), both players observe a pair of candidate referents. A speaker is assigned a target referent; it must communicate this target to a listener, who then performs a choice action corresponding to its belief about the true target. In this paper we consider two variants on the reference game: a simple color-naming task, and a more complex task involving natural images of birds. For examples of human communication strategies for these tasks, we obtain the XKCD color dataset (McMahan and Stone, 2015; Monroe et al., 2016) and the Caltech Birds dataset (Welinder et al., 2010) with accompanying natural language descriptions (Reed et al., 2016). We use standard train / validation / test splits for both of these datasets.

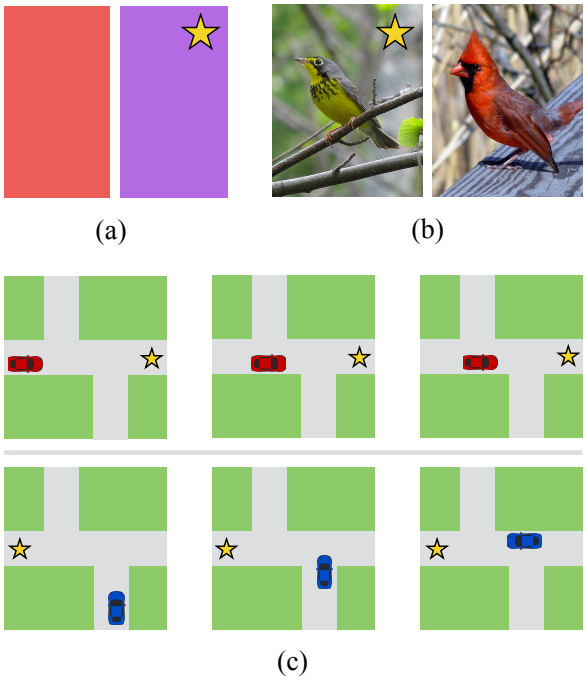


Figure 6: Tasks used to evaluate the translation model. (a–b) Reference games: both players observe a pair of reference candidates (colors or images); Player a is assigned a target (marked with a star), which player b must guess based on a message from a. (c) Driving game: each car attempts to navigate to its goal (marked with a star). The cars cannot see each other, and must communicate to avoid a collision.

The final task we consider is the driving task (Figure 6c) first discussed in the introduction. In this task, two cars, invisible to each other, must each navigate between randomly assigned start and goal positions without colliding. This task takes a number of steps to complete, and potentially involves a much broader range of communication strategies. To obtain human annotations for this task, we recorded both actions and messages generated by pairs of human Amazon Mechanical Turk workers playing the driving game with each other. We collected close to 400 games, with a total of more than 2000 messages exchanged, from which we held out 100 game traces as a test set.

7.2 Metrics

A mechanism for understanding the behavior of a learned model should allow a human user both to correctly infer its beliefs and to successfully interoperate with it; we accordingly report results of both "belief" and "behavior" evaluations.

To support easy reproduction and comparison (and in keeping with standard practice in machine

translation), we focus on developing automatic measures of system performance. We use the available training data to develop simulated models of human decisions; by first showing that these models track well with human judgments, we can be confident that their use in evaluations will correlate with human understanding. We employ the following two metrics:

Belief evaluation This evaluation focuses on the denotational perspective in semantics that motivated the initial development of our model. We have successfully understood the semantics of a message z_r if, after translating $z_r \mapsto z_h$, a human listener can form a correct belief about the state in which z_r was produced. We construct a simple state-guessing game where the listener is presented with a translated message and two state observations, and must guess which state the speaker was in when the message was emitted.

When translating from natural language to neuralese, we use the learned agent model to directly guess the hidden state. For neuralese to natural language we must first construct a "model human listener" to map from strings back to state representations; we do this by using the training data to fit a simple regression model that scores (state, sentence) pairs using a bag-of-words sentence representation. We find that our "model human" matches the judgments of real humans 83% of the time on the colors task, 77% of the time on the birds task, and 77% of the time on the driving task. This gives us confidence that the model human gives a reasonably accurate proxy for human interpretation.

Behavior evaluation This evaluation focuses on the cooperative aspects of interpretability: we measure the extent to which learned models are able to interoperate with each other by way of a translation layer. In the case of reference games, the goal of this semantic evaluation is identical to the goal of the game itself (to identify the hidden state of the speaker), so we perform this additional pragmatic evaluation only for the driving game. We found that the most data-efficient and reliable way to make use of human game traces was to construct a "deaf" model human. The evaluation selects a full game trace from a human player, and replays both the human's actions and messages exactly (disregarding any incoming messages); the evaluation measures the quality of the natural-language-to-neuralese translator, and the extent to which the learned agent model can accommodate a (real) human given translations of the human's messages.

(a)

as listener	as speaker: R	as speaker: H	
R	1.00	0.50	random
		0.70	direct
		0.73	belief (ours)
H*	0.50	0.83	
	0.72		
	0.86		

(b)

as listener	as speaker: R	as speaker: H	
R	0.95	0.50	random
		0.55	direct
		0.60	belief (ours)
H*	0.50	0.77	
	0.57		
	0.75		

Table 1: Evaluation results for reference games. (a) The colors task. (b) The birds task. Whether the model human is in a listener or speaker role, translation based on belief matching outperforms both random and machine translation baselines.

Baselines We compare our approach to two baselines: a *random* baseline that chooses a translation of each input uniformly from messages observed during training, and a *direct* baseline that directly maximizes $p(z'|z)$ (by analogy to a conventional machine translation system). This is accomplished by sampling from a DCP speaker in training states labeled with natural language strings.

8 Results

In all below, "R" indicates a DCP agent, "H" indicates a real human, and "H*" indicates a model human player.

Reference games Results for the two reference games are shown in Table 1. The end-to-end trained model achieves nearly perfect accuracy in both

Figure 7: Best-scoring translations generated for color task.

as listener	as speaker R	as speaker H	
R	0.85	0.50	random
		0.45	direct
		0.61	belief (ours)
H*	0.5	0.77	
	0.45		
	0.57		

Table 2: Belief evaluation results for the driving game. Driving states are challenging to identify based on messages alone (as evidenced by the comparatively low scores obtained by single-language pairs) . Translation based on belief achieves the best overall performance in both directions.

R / R	H / H	R / H	
1.93 / 0.71	— / 0.77	1.35 / 0.64	random
		1.49 / **0.67**	direct
		1.54 / 0.67	belief (ours)

Table 3: Behavior evaluation results for the driving game. Scores are presented in the form "reward / completion rate". While less accurate than either humans or DCPs with a shared language, the models that employ a translation layer obtain higher reward and a greater overall success rate than baselines.

cases, while a model trained to communicate in natural language achieves somewhat lower performance. Regardless of whether the speaker is a DCP and the listener a model human or vice-versa, translation based on the belief-matching criterion in Section 5 achieves the best performance; indeed, when translating neuralese color names to natural language, the listener is able to achieve a slightly higher score than it is natively. This suggests that the automated agent has discovered a more effective strategy than the one demonstrated by humans in the dataset, and that the effectiveness of this strategy is preserved by translation. Example translations from the reference games are depicted in Figure 2 and Figure 7.

Driving game Behavior evaluation of the driving game is shown in Table 3, and belief evaluation is shown in Table 2. Translation of messages in the driving game is considerably more challenging than in the reference games, and scores are uniformly lower; however, a clear benefit from the belief-matching model is still visible. Belief matching leads to higher scores on the belief evaluation in both directions, and allows agents to obtain a higher reward on average (though task completion rates remain roughly the same across all agents). Some example translations of driving game messages are shown in Figure 8.

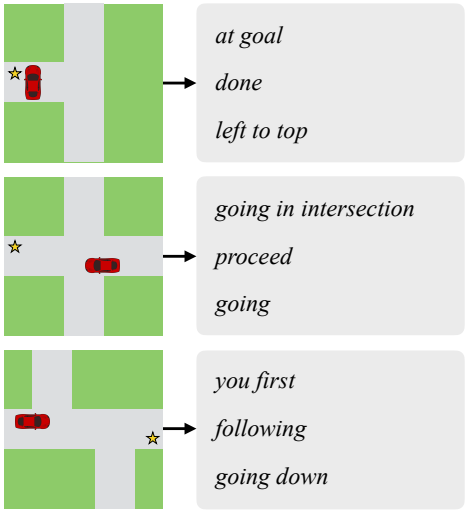


Figure 8: Best-scoring translations generated for driving task generated from the given speaker state.

9 Conclusion

We have investigated the problem of interpreting message vectors from deep networks by translating them. After introducing a translation criterion based on matching listener beliefs about speaker states, we presented both theoretical and empirical evidence that this criterion outperforms a conventional machine translation approach at recovering the content of message vectors and facilitating collaboration between humans and learned agents.

While our evaluation has focused on understanding the behavior of deep communicating policies, the framework proposed in this paper could be much more generally applied. Any encoder–decoder model (Sutskever et al., 2014) can be thought of as a kind of communication game played between the encoder and the decoder, so we can analogously imagine computing and translating "beliefs" induced by the encoding to explain what features of the input are being transmitted. The current work has focused on learning a purely categorical model of the translation process, supported by an unstructured inventory of translation candidates, and future work could explore the *compositional* structure of messages, and attempt to synthesize novel natural language or neuralese messages from scratch. More broadly, the work here shows that the denotational perspective from formal semantics provides a framework for precisely framing the demands of interpretable machine learning (Wilson et al., 2016), and particularly for ensuring that human users without prior exposure to a learned model are able to interoperate with it, predict its behavior, and diagnose its errors.

Acknowledgments

JA is supported by a Facebook Graduate Fellowship and a Berkeley AI / Huawei Fellowship. We are grateful to Lisa Anne Hendricks for assistance with the Caltech Birds dataset.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research* 27(4):819–840.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* .

Jilles Steeve Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. 2016. Optimally solving Dec-POMDPs as continuous-state MDPs. *Journal of Artificial Intelligence Research* 55:443–497.

Anca Dragan and Siddhartha Srinivasa. 2013. Generating legible motion. In *Robotics: Science and Systems*.

Jakob Foerster, Yannis M Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*. pages 2137–2145.

Michael C Frank, Noah D Goodman, Peter Lai, and Joshua B Tenenbaum. 2009. Informative communication in word production and word learning. In *Proceedings of the 31st annual conference of the cognitive science society*. pages 1228–1233.

Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. 2016. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 317–326.

Jon Gauthier and Igor Mordatch. 2016. A paradigm for situated and goal-driven language learning. *arXiv preprint arXiv:1610.03585* .

Frank Guerin and Jeremy Pitt. 2001. Denotational semantics for agent communication language. In *Proceedings of the fifth international conference on Autonomous agents*. ACM, pages 497–504.

Matthew Hausknecht and Peter Stone. 2015. Deep recurrent q-learning for partially observable mdps. *arXiv preprint arXiv:1507.06527* .

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*. Springer, pages 3–19.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 787–798.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016a. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182* .

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2016b. Towards multi-agent communication-based language learning. *arXiv preprint arXiv:1605.07133* .

Brian McMahan and Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics* 3:103–115.

Will Monroe, Noah D Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. *arXiv preprint arXiv:1606.03821* .

Brooks Paige and Frank Wood. 2016. Inference networks for sequential monte carlo in graphical models. volume 48.

Panupong Pasupat and Percy Liang. 2016. Inferring logical forms from denotations. *arXiv preprint arXiv:1606.06900* .

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 49–58.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pages 1135–1144.

Maayan Roth, Reid Simmons, and Manuela Veloso. 2005. Reasoning about joint beliefs for execution-time communication decisions. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. ACM, pages 786–793.

Hendrik Strobelt, Sebastian Gehrmann, Bernd Huber, Hanspeter Pfister, and Alexander M Rush. 2016. Visual analysis of hidden state dynamics in recurrent neural networks. *arXiv preprint arXiv:1606.07461* .

Sainbayar Sukhbaatar, Rob Fergus, et al. 2016. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*. pages 2244–2252.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. pages 3104–3112.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. *arXiv preprint arXiv:1701.02870* .

Adam Vogel, Max Bodoia, Christopher Potts, and Daniel Jurafsky. 2013a. Emergence of Gricean maxims from multi-agent decision theory. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. pages 1072–1081.

Adam Vogel, Christopher Potts, and Dan Jurafsky. 2013b. Implicatures and nested beliefs in approximate Decentralized-POMDPs. In *ACL (2)*. pages 74–80.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Andrew Gordon Wilson, Been Kim, and William Herlands. 2016. Proceedings of nips 2016 workshop on interpretable machine learning for complex systems. *arXiv preprint arXiv:1611.09139* .

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2016. A joint speaker-listener-reinforcer model for referring expressions. *arXiv preprint arXiv:1612.09542* .

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, pages 818–833.

A Proofs

Proof of Proposition 1 We know that

$$U(z, x_b) := \arg\max_u \sum_{x_a} p(x_a|x_b, z) r(x_a, x_b, z)$$

and that for all translations $(z, z' = t(r))$

$$D \geq \sum_{x_b} p(x_b|z, z') \mathcal{D}_{\mathrm{KL}}(\beta(z, x_b) \,||\, \beta(z', x_b)) \; .$$

Applying Pinsker's inequality:

$$\geq 2 \sum_{x_b} p(x_b|z, z') \delta(\beta(z, x_b), \beta(z', x_b))^2$$

and Jensen's inequality:

$$\geq 2 \Bigg(\sum_{x_b} p(x_b|z, z') \delta(\beta(z, x_b), \beta(z', x_b))) \Bigg)^2$$

so

$$\sqrt{D/2} \geq \sum_{x_b} p(x_b|z, z') \delta(\beta(z, x_b), \beta(z', x_b)) \; .$$

The next step relies on the following well-known property of the total variation distance: for distributions p and q and a function f bounded by $[0, 1]$,

$$|\mathbb{E}_p f(x) - \mathbb{E}_q f(x)| \leq \delta(p, q) \; . \qquad (*)$$

For convenience we will write

$$\delta := \delta(\beta(z, x_b), \beta(z', x_b)) \; .$$

A listener using the speaker's language expects a reward of

$$\sum_{x_b} p(x_b) \sum_{x_a} p(x_a|x_b, z) r(x_a, x_b, U(z, x_b))$$
$$\leq \sum_{x_b} p(x_b) \Bigg(\sum_{x_a} p(x_a|x_b, z') r(x_a, x_b, U(z, x_b)) + \delta \Bigg)$$

via (*). From the assumption of player rationality:

$$\leq \sum_{x_b} p(x_b) \Bigg(\sum_{x_a} p(x_a|x_b, z') r(x_a, x_b, U(z', x_b)) + \delta \Bigg)$$

using (*) again:

$$\leq \sum_{x_b} p(x_b) \Bigg(\sum_{x_a} p(x_a|x_b, z) r(x_a, x_b, U(z', x_b)) + 2\delta \Bigg)$$
$$\leq \sum_{x_a, x_b} p(x_a, x_b|z) r(x_a, x_b, U(z', x_b)) + \sqrt{2D} \; .$$

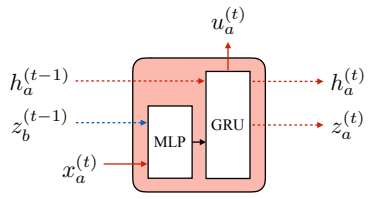
So the true reward achieved by a z'-speaker receiving a translated code is only additively worse than the native z-speaker reward:

$$\Bigg(\sum_{x_a, x_b} p(x_a, x_b|z) r(x_a, x_b, U(z, x_b)) \Bigg) - \sqrt{2D} \qquad \square$$

B Implementation details

B.1 Agents

Learned agents have the following form:



where h is a hidden state, z is a message from the other agent, u is a distribution over actions, and x is an observation of the world. A single hidden layer with 256 units and a $\tanh$ nonlinearity is used for the MLP. The GRU hidden state is also of size 256, and the message vector is of size 64.

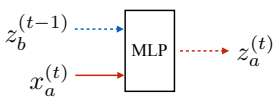
Agents are trained via interaction with the world as in Hausknecht and Stone (2015) using the ADAM optimizer (Kingma and Ba, 2014) and a discount factor of 0.9. The step size was chosen as 0.003 for reference games and 0.0003 for the driving game. An ϵ-greedy exploration strategy is employed, with the exploration parameter for timestep t given by:

$$\epsilon = \max \begin{cases} (1000 - t)/1000 \\ (5000 - t)/50000 \\ 0 \end{cases}$$

As in Foerster et al. (2016), we found it useful to add noise to the communication channel: in this case, isotropic Gaussian noise with mean 0 and standard deviation 0.3. This also helps smooth $p(z|x_a)$ when computing the translation criterion.

B.2 Representational models

As discussed in Section 5, the translation criterion is computed based on the quantity $p(z|x)$. The policy representation above actually defines a distribution $p(z|x, h)$, additionally involving the agent's hidden state h from a previous timestep. While in principle it is possible to eliminate the dependence on h by introducing an additional sampling step into Algorithm 1, we found that it simplified inference to simply learn an additional model of $p(z|x)$ directly. This model is trained alongside the learned agent to imitate its decisions, but does not get to observe the recurrent state, like so:



Here the multilayer perceptron has a single hidden layer with $\tanh$ nonlinearities and size 128. It is also trained with ADAM and a step size of 0.0003.

We use exactly the same model and parameters to implement representations of $p(z|x)$ for human speakers, but in this case the vector z is taken to be a distribution over messages in the natural language inventory, and the model is trained to maximize the likelihood of labeled human traces.

B.3 Tasks

Colors We use the version of the XKCD dataset prepared by McMahan and Stone (2015). Here the input feature vector is simply the LAB representation of each color, and the message inventory taken to be all unigrams that appear at least five times.

Birds We use the dataset of Welinder et al. (2010) with natural language annotations from Reed et al. (2016). The model's input feature representations are a final 256-dimensional hidden feature vector from a compact bilinear pooling model (Gao et al., 2016) pre-trained for classification. The message inventory consists of the 50 most frequent bigrams to appear in natural language descriptions; example human traces are generated by for every frequent (bigram, image) pair in the dataset.

Driving Driving data is collected from pairs of human workers on Mechanical Turk. Workers received the following description of the task:

> Your goal is to drive the red car onto the red square. Be careful! You're driving in a thick fog, and there is another car on the road that you cannot see. However, you can talk to the other driver to make sure you both reach your destinations safely.

Players were restricted to messages of 1–3 words, and required to send at least one message per game. Each player was paid $0.25 per game. 382 games were collected with 5 different road layouts, each represented as an 8x8 grid presented to players as in Figure 8. The action space is discrete: players can move forward, back, turn left, turn right, or wait. These were divided into a 282-game training set and 100-game test set. The message inventory consists of all messages sent more than 3 times. Input features consists of indicators on the agent's current position and orientation, goal position, and map identity. Data is available for download at `http://github.com/jacobandreas/neuralese`.