The spontaneous emergence of discrete and compositional messages

(1) Thoughts on the title? (2) It's hidden on submission, but we need to decide author order as well.

Anonymous ACL submission

Abstract

We propose a general framework to study language emergence through signaling games with neural agents. Using a continuous latent space, we are able to (i) train using backpropagation, (ii) show that discrete messages nonetheless naturally emerge. We discuss various ways to test classic properties of the emergent language in production and perception.

1 Introduction

In a signaling game, artificial agents learn to communicate to achieve a common goal: a sender sees some piece of information and produces a message, this message is then sent to a receiver that must take some action (Lewis, 1969; Skyrms, 2010). If the action is coherent with the sender's initial piece of information, the whole communication stream —the choice of the message and its interpretation— is reinforced. For instance, in a referential game, sender and receiver see a set of objects, and the sender knows which of these the receiver must pick; the sender thus sends a message to the receiver, who must interpret it to pick up the right object (Lazaridou et al., 2017, 2018; Havrylov and Titov, 2017; Chaabouni et al., 2019).

This setting has been used to study the factors influencing the emergence of various fundamental properties of natural language, such as *compositionality* (Kirby et al., 2015; Franke, 2016; Steinert-Threlkeld, 2016; Mordatch and Abbeel, 2018; Lazaridou et al., 2018; Choi et al., 2018). In this paper, we add focus on two other so-called 'design features' of natural language (Hockett, 1960): *discreteness* (i.e. words form clusters in acoustic space), and *displacement* (i.e. efficient communication can occur about objects and facts beyond the immediate context of the conversation).

From an implementation point of view, we follow the recent literature which has shown that a signaling game is essentially an autoencoder setting, with the encoder playing the role of the sender, and the decoder the role of the receiver (see Fig. 1). In this literature, however, the discreteness of the communication protocol is assumed, since the networks then traditionally use a (normally sequential and) discrete latent space (Havrylov and Titov, 2017; Chaabouni et al., 2019; Kharitonov et al., 2019).

Our main contribution is a generalization of the current implementation of signaling games as autoencoders. Our implementation covers a broader variety of signaling games, and it crucially incorporates the possibility of displacement and no *a priori* assumption of discreteness. Our main result is that under appropriate conditions, discreteness emerges spontaneously, that is, if the latent space is thought about as a continuous acoustic space, then trained messages form coherent clusters, just like regular words do.

In addition to contributing to our understanding of the emergence of communication protocols with features like natural language, these results have a technical significance: by using a fundamentally continuous communication protocol, with discreteness emerging, we can train end-to-end using standard backpropagation techniques, instead of reinforcement learning algorithms like REINFORCE and its refinements (Williams, 1992; Schulman et al., 2015; Mnih et al., 2016), which are difficult to use in practice.

2 Related Work

A related line of work attempts to avoid the difficulties of reinforcement learning—which introduces stochastic nodes into a computation graph—by reparameterization and/or non-stochastic estimators (Bengio et al., 2013; Schulman et al., 2015). In the emergent communication case, where the stochastic nodes are discrete (e.g. sampling a

message from a sender distribution), the Gumbel-Softmax estimator has become increasingly popular (Jang et al., 2017; Maddison et al., 2017).

This work enables standard backpropagation to be used for training by optimizing approximations to the true reinforcement learning signal. By contrast, using the auto-encoder analogy more literally, we do not approximate the discrete RL learning signal, but rather ask under what conditions discreteness will emerge. Probably can be worded better.

3 Function Games

We here introduce a general communication game setting, which we call Function Games. Our games contain three basic components: (i) a set of contexts C, (ii) a set of actions A, (iii) a family of functions F, from contexts to actions. One play of a Function Game game runs as follows:

1. Nature chooses $f \in F$ and a context $c \in C$.
2. Sender sees the context c and f.
3. Sender sends a message m to Receiver.
4. Receiver sees *a possibly different* context c' and the message m and chooses an action a'.
5. Both are 'rewarded' iff $a' = f(c')$.

Abstractly, the function f represents some piece of knowledge available primarily for Sender, and which determines what action is appropriate in any given context. Two concrete interpretations will help illustrate the variety of communication protocols and goals that this framework encompasses.

Generalized referential games. A reference game is one in which Sender tries to get Receiver to pick the correct object out of a given set (Skyrms, 2010; Lazaridou et al., 2017, 2018; Havrylov and Titov, 2017; Chaabouni et al., 2019). Here, contexts are sets of objects (i.e. an $m \times n$ matrix, with m objects represented by n features). Normally (though we will drop this assumption later), $c' = \texttt{shuffled}(c)$: Sender and Receiver see the same objects, but in a different arrangement. Actions are the objects, and the functions $f \in F$ are *choice functions*: $f(c) \in c$ for every context c.

Belief update games. Contexts can represent the (possibly different) belief states of the agents. 'Actions' can represent updated belief states ($A = C$), the different functions in F then representing how to update an agent's beliefs in the light of learning a particular piece of information (passed directly to Sender, and only through the message to Receiver).

What should we cite here? Something from dynamic semantics?

4 Experiment

Because we are interested in the simultaneous emergence both of discrete signals and of compositional messages, we use a Function Game called the Extremity Game designed to incentivize and test rich compositionality (Steinert-Threlkeld, 2019). In this game, it is natural to think of the n dimensions of the objects as gradable properties, e.g., size and darkness, so that a 2D object is determined by a given size and shade of gray. For the functions, we set $F = \{\arg\min_i, \arg\max_i : 0 \leq i < n\}$ and, hopefully, the emerging language should contain messages like '(pick the) MOST + BIG', 'LEAST + DARK', etc.

cite my forthcoming philosophy of science paper instead? pre-print is only on my website, so URL would be sort of de-anonymizing

4.1 Model

Our model resembles an encoder-decoder architecture, with the Sender encoding the context/target pair into a message, and the Receiver decoding the message (together with its context c') into an action. Both the encoder and decoder are multi-layer perceptrons with two hidden layers of size 64 and rectified linear activation (ReLU) (Nair and Hinton, 2010; Glorot et al., 2011). A smaller, intermediate layer without an activation function bridges the encoder and decoder and represents the transformation of the input information to messages. Figure 1 depicts this architecture.

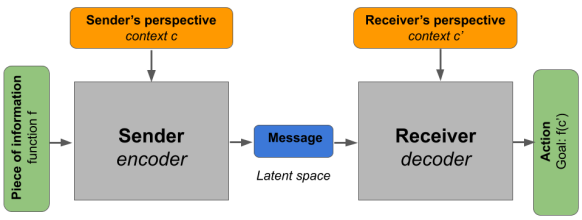

Figure 1: Our model architecture, mixing terminology from the autoencoder and signaling game traditions.

4.2 Game Parameters

We manipulated the following parameters:

- Context identity. In the *shared* setting, Receiver sees a shuffled version of Sender's context ($c' = \texttt{shuffled}(c)$). In the *non-shared* setting, Receiver's context c' is entirely distinct from Sender's. This forces displacement and may incentivize compositional messages, since Sender cannot rely on the raw properties of the target object in communication.

- Context strictness. In *strict* contexts, there is a one-to-one (and onto) correspondence between F and A.[1] In *non-strict* contexts, an object may be the $\arg\max$ or $\arg\min$ of several dimensions, or of no dimension.

In all experiments, the latent space (message) dimension was always 2,[2] and objects had 5 dimensions. Strict contexts therefore contained 10 objects, while non-strict contexts could contain 5, 10, or 15 objects.

4.3 Training Details

By using a continuous latent space, the entire model, including the communication channel, is differentiable and so can be trained end-to-end using backpropagation to compute gradients. We can drop the previous sentence if we need space. We used the Adam optimizer (Kingma and Ba, 2015) with learning rate 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The model was trained for 5,000 epochs by feeding the network with mini-batches of 64 contexts concatenated with one-hot function selectors. The network's loss is taken as the MSE between the target object $f(c')$ and the object generated by the Receiver. For each setting of the above parameters, we ran 20 trials with different random seeds.

Code and data will be made available once the paper can be de-anonymized. Save space by shortening this remark or integrating it better to a paragraph

5 Results

5.1 Communicative success

We measure the communicative success of the network by calculating the accuracy of recovering the correct object from c'. Receiver's prediction is considered correct if its output is closest to $f(c')$ than to all other objects in c'. Accuracy of the different settings is reported in Table 1. While the network handles well displacement (*non-shared contexts*), the model struggles with non-strict contexts. Note that although accuracy is not 100%, it is still well above chance, since e.g. for a context of 10 objects random guessing yields an expected accuracy of 10%. Accordingly, the recovery accuracy for networks before training are at the expected chance level (e.g., for a 10-object, shared, strict setting, accuracy is $9.92\% \pm 0.01$, with MSE 8 times larger before than after training).

	Shared	Non-shared
Strict		
10 objects	$63.78\% \pm 1.63$	$60.22\% \pm 1.56$
Non-strict		
5 objects	$49.37\% \pm 1.67$	$43.55\% \pm 1.69$
10 objects	$33.06\% \pm 1.47$	$31.89\% \pm 1.63$
15 objects	$27.58\% \pm 1.30$	$27.95\% \pm 1.24$

Table 1: Communicative success, as measured by object recovery accuracy.

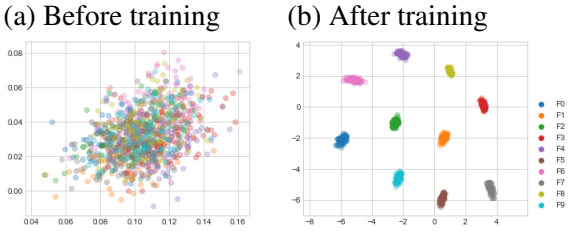

Figure 2: Messages sampled from latent space for objects of size 5 and contexts of 10 objects for (a) an untrained and (b) a trained network. Colors represent the $f_i \in F$ input part of the Sender.

As mentioned, our setting can use traditional backpropagation, and it reaches higher accuracy levels than settings using *a priori* discrete signals ((Steinert-Threlkeld, 2019) did not perform well with more than 4 objectsCHECK). Move this paragraph to discussion section? I think we should remove this little paragraph: 6 objects was the highest I did, and mean accuracy was either 73% or 69%. It's plausible that it would not be as good at 10+ objects, but we can't do a direct comparison.

5.2 Discrete signals

Figure 2 depicts message vectors sampled from the latent space layer, before and after training. It is apparent that discrete messages emerge from the imposed learning regime. We measured cluster tendency more quantitatively through two measures, one considering Sender's *production*, and the other Receiver's *perception*.

First, we sampled 100 contexts, and collected the output of the trained encoder for each of these contexts combined with each possible function f. We applied an unsupervized clustering algorithm to this set of produced messages (DBSCAN, Ester et al., 1996, with $\epsilon = 0.5$). A label was assigned to each cluster using the ground truth: the label of a cluster is the function f that was most often at the source of a point in this cluster. This allowed us to

[1] These are the contexts used in the original Extremity Game from (Steinert-Threlkeld, 2019). Drop if we need space? Or incorporate in text with a couple of words, like 'As in the original Extremity Game from (Steinert-Threlkeld, 2019)'?

[2] The model also performs well with messages of size 1, not reported here. Using messages of size 2 makes it easier to inspect the latent space using 2-D visualization.

	Shared	Non-shared
Strict		
10 objects	1.00 ± 0.00	0.90 ± 0.09
Non-strict		
5 objects	0.99 ± 0.02	0.54 ± 0.15
10 objects	1.00 ± 0.00	0.99 ± 0.01
15 objects	1.00 ± 0.00	1.00 ± 0.00

Table 2: Discreteness in production, as measured by F1 scores for automatically clusterized messages.

	Shared	Non-shared
Strict		
10 objects	$63.39\% \pm 1.45$	$55.37\% \pm 3.43$
Non-strict		
5 objects	$46.94\% \pm 1.70$	$29.40\% \pm 5.59$
10 objects	$32.63\% \pm 1.43$	$31.51\% \pm 1.62$
15 objects	$28.24\% \pm 1.11$	$27.94\% \pm 1.20$

Table 3: Discreteness in perception, as measured by object recovery accuracy from artificial messages.

compute F1-scores, which results are reported in Table 2. The model reached near-optimal clusterization measures for both shared and non-shared contexts, and for both strict and non-strict contexts.

The second approach is akin to studying perception. Given the clusterization of the message space, we are able to sample new messages from each cluster, and test Receiver's perception of these 'artificial' messages, messages which have never been produced by Sender. Optional: In other words, we ask whether we can speak the language developed by the players, and be understood by Receiver. As a sampling procedure for artificial messages, we take the average of 10 messages from a (now labelled) cluster. These artificial messages are fed to Receiver, the output object accuracy for these artificial messages is given in Table 3. The model achieves recovery accuracy similar to when messages are generated using actual messages.

In sum, we can identify discrete, abstract regions of the latent space corresponding to different functions in the input, just like words form clusters in acoustic space. This analysis in terms of production and perception can lead to more psycholinguistic-like queries about these emergent languages. For instance, one may ask whether classical 'Categorical Perception' effects obtain, whereby messages shifting continuously in the latent space may be interpreted as if they crossed a sharp boundary for interpretation purposes (see Damper and Harnad, 2000 for early discussions in the context of neural architectures). Improve that last paragraph above? Move it to the discussion section?

5.3 Compositionality

Our agents are capable of communicating in abstract situations, namely some in which their contexts are different in the first place. This generalizability is already a sign of 'compositionality', or at least of productivity. But how do the messages relate to the structure in the family of functions F?

First, in the pioneering Mikolov et al., 2013, compositionality is looked for at the level of word embeddings (WE) through addition, most classically asking whether WE(queen)=WE(king)-WE(man)+WE(woman). In the current Game, the analogous question is whether the messages are related as follows, for any dimensions i and j: $\text{M}(\arg\max_i)$=$\text{M}(\arg\max_j)$-$\text{M}(\arg\min_j)$+$\text{M}(\arg\min_i)$. We find that using the right-hand side of the equation above, that is the compositionally appropriate message, in principle, leads to important degradation of communicative success (a drop of at least 36 percentage points across parameter combinations, to around chance level).

Second, we note as others that the composition as addition assumption is however disputable, both in general and in the original application case (Linzen, 2016; Chen et al., 2017). To abstract away from this issue, we train a network at predicting $\text{M}(\arg\max_i)$ from $\text{M}(\arg\max_j)$, $\text{M}(\arg\min_j)$ and $\text{M}(\arg\min_i)$, therefore letting it discover any function for mixing values, and not involving addition *a priori*. We left out one dimension i_0 from training, and considered the message predicted by such a network from $\text{M}(\arg\max_j)$, $\text{M}(\arg\min_j)$ and $\text{M}(\arg\min_{i_0})$. If the language was compositional, this predicted message should behave like $\text{M}(\arg\max_{i_0})$, but we found that, as in the case of addition, communication accuracy dropped dramatically (again, at least 36 percentage points drop). cite Baroni's (or others) similar tests?

6 Conclusion

Signaling games have long been used to study which properties of languages are shaped by communicative pressures. This tradition has recently been merged with the machine learning literature, by using neural agents for the relevant simulations. Here we propose a general framework in which

more games can be simulated, fewer *a priori* assumptions are imposed on the conversational situations. We find in particular that under appropriate conditions, which are met by most[all] studies involving neural signaling games, messages become discrete without the analyst having to force this property into the language (and having to deal with non-derivability issues). Likely due to this improvement, we find that at equivalent difficulty levels, our training regime leads to better communication than systems forcing discreteness *a priori*, while, nonetheless, the resulting language spontaneously evolves effectively discrete signals.

References

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019. Anti-efficient encoding in emergent communication. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Dawn Chen, Joshua C. Peterson, and Thomas L. Griffiths. 2017. Evaluating vector-space models of analogy. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Edward Choi, Angeliki Lazaridou, and Nando de Freitas. 2018. Compositional Obverter Communication Learning from Raw Visual Input. In *International Conference of Learning Representations (ICLR 2018)*, pages 1–18.

R.I. Damper and S.R. Harnad. 2000. Neural network models of categorical perception.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Michael Franke. 2016. The Evolution of Compositionality in Signaling Games. *Journal of Logic, Language and Information*.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. In *14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 315–323.

Serhii Havrylov and Ivan Titov. 2017. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*.

Charles F Hockett. 1960. The Origin of Speech. *Scienctific American*, 203:88–111.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International Conference of Learning Representations (ICLR)*.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. EGG: a toolkit for research on Emergence of lanGuage in Games. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 55–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference of Learning Representations (ICLR)*.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. In *International Conference of Learning Representations (ICLR 2018)*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-Agent Cooperation and the Emergence of (Natural) Language. In *International Conference of Learning Representations (ICLR2017)*.

David Lewis. 1969. *Convention*. Blackwell.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

Chris J Maddison, Andriy Mnih, Yee Whye Teh, United Kingdom, and United Kingdom. 2017. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference of Learning Representations (ICLR)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Tim Harley, Timothy P Lillicrap, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous Methods for Deep Reinforcement Learning. In *International Conference on Machine Learning (ICML)*.

Igor Mordatch and Pieter Abbeel. 2018. Emergence of Grounded Compositional Language in Multi-Agent Populations. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. 2015. Gradient Estimation Using Stochastic Computation Graphs. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*.

Brian Skyrms. 2010. *Signals: Evolution, Learning, and Information*. Oxford University Press.

Shane Steinert-Threlkeld. 2016. Compositional Signaling in a Complex World. *Journal of Logic, Language and Information*, 25(3):379–397.

Shane Steinert-Threlkeld. 2019. Paying Attention to Function Words. In *Emergent Communication Workshop @ NeurIPS 2018*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256.