Seeded self-play for language learning

Abhinav Gupta[*]
MILA
abhinavg@nyu.edu

Ryan Lowe[*]
MILA
ryan.lowe@mail.mcgill.ca

Jakob Foerster
Facebook AI Research

Douwe Kiela
Facebook AI Research

Joelle Pineau
Facebook AI Research
MILA

Abstract

How can we teach artificial agents to use human language flexibly to solve problems in real-world environments? We have an example of this in nature: human babies eventually learn to use human language to solve problems, and they are taught with an adult human-in-the-loop. Unfortunately, current machine learning methods (e.g. from deep reinforcement learning) are too data inefficient to learn language in this way. An outstanding goal is finding an algorithm with a suitable 'language learning prior' that allows it to learn human language, while minimizing the number of on-policy human interactions. In this paper, we propose to learn such a prior in simulation using an approach we call, Learning to Learn to Communicate (L2C). Specifically, in L2C we train a meta-learning agent in simulation to interact with populations of pre-trained agents, each with their own distinct communication protocol. Once the meta-learning agent is able to quickly adapt to each population of agents, it can be deployed in new populations, including populations speaking human language. Our key insight is that such populations can be obtained via self-play, after pre-training agents with imitation learning on a small amount of off-policy human language data. We call this latter technique Seeded Self-Play (S2P). Our preliminary experiments show that agents trained with L2C and S2P need fewer on-policy samples to learn a compositional language in a Lewis signaling game.

1 Introduction

Language is one of the most important aspects of human intelligence; it allows humans to coordinate and share knowledge with each other. We will want artificial agents to understand language as it is a natural means for us to specify their goals.

So how can we train agents to understand language? We adopt the functional view of language (Wittgenstein, 1953) that has recently gained popularity (Lazaridou et al., 2016; Gauthier and Mordatch, 2016; Mordatch and Abbeel, 2017): agents understand language when they can use language to carry out tasks in the real world. One approach to training agents that can use language in their environment is via *emergent communication*, where researchers train randomly initialized agents to solve tasks requiring communication (Foerster et al., 2016; Lazaridou et al., 2018; Tieleman et al., 2018). An open question in emergent communication is how the resulting communication protocols can be transferred to learning human language (Baroni, 2019; Hupkes et al., 2019). Existing approaches attempt to do this using auxiliary tasks (Lee et al., 2018, 2017), for example having agents predict the label of an image in English while simultaneously playing an image-based referential game (Evtimova et al., 2017). While this works for learning the names of objects, it's unclear if simply using an auxiliary loss will scale to learning the English names of complex concepts, or learning to use English to interact in an grounded environment.

One approach that we know will work (eventually) for training language learning agents is using a human-in-the-loop, as this is how human babies acquire language. In other words, if we had a good enough model architecture and learning algorithm, the human-in-the-loop approach should work. However, recent work in this direction has concluded that current algorithms are too sample inefficient to effectively learn a language with compositional properties from humans (Chevalier-Boisvert et al., 2018). Human guidance is expensive, and thus we would want such an algorithm to be as sample efficient as possible. An open problem is thus to create an algorithm or training procedure that results in increased sample-efficiency for lan-

[*] Equal contribution.

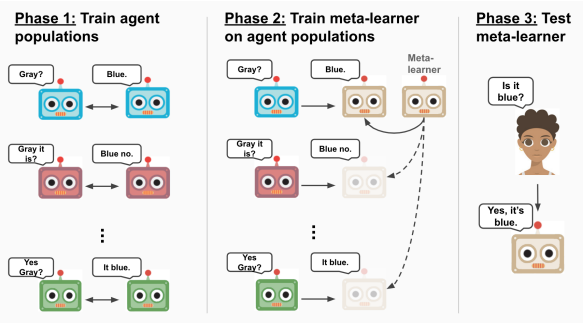

Figure 1: Schematic diagram of the L2C framework. An advantage of L2C is that agents can be trained in an external environment (which *grounds* the language), where agents interact with the environment via actions and language. Thus, (in theory) L2C could be scaled to learn complicated grounded tasks involving language.

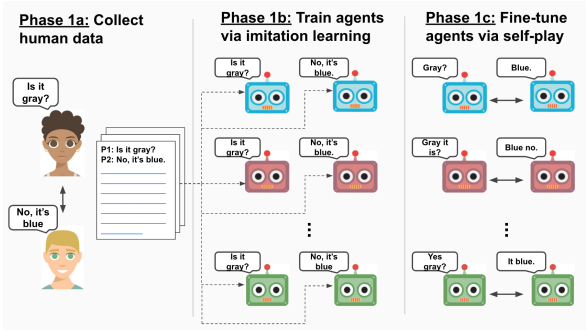

Figure 2: Schematic diagram of the S2P framework. Phase 1b and 1c are carried out in alternation or over some *schedule* to counter language drift while achieving high reward in the corresponding task. See Sec 3 for more details.

guage learning with a human-in-the-loop.

In this paper, we present the Learning to Learn to Communicate (L2C) framework, with the goal of training agents to quickly learn new (human) languages. The core idea behind L2C is to leverage the increasing amount of available compute for machine learning experiments (Amodei and Hernandez, 2018) to learn a 'language learning prior' by training agents via meta-learning in simulation. Specifically, we train a meta-learning agent in simulation to interact with populations of pre-trained agents, each with their own distinct communication protocol. Once the meta-learning agent is able to quickly adapt to each population of agents, it can be deployed in new populations unseen during training, including populations of humans. The L2C framework has two main advantages: (1) permits for agents to learn language that is *grounded* in an environment with which the agents can *interact* (i.e. it is not limited to referential games); and (2) in contrast with work from the instruction following literature (Bahdanau et al., 2018), agents can be trained via L2C to both *speak* (output language to help accomplish their goal) and *listen* (map from the language to a goal or sequence of actions).

To show the promise of the L2C framework, we provide some preliminary experiments in a Lewis signaling game (David, 1969). Specifically, we show that agents trained with L2C are able to learn a simple form of human language (represented by a hand-coded compositional language) in fewer iterations than randomly initialized agents. These preliminary results suggest that L2C is a promising framework for training agents to learn human language from few human interactions.

2 Learning to learn to communicate

L2C is a training procedure that is composed of three main phases: (1) **Training agent populations:** Training populations of agents to solve some task (or set of tasks) in an environment. (2) **Train meta-learner on agent populations:** We train a meta-learning agent to 'perform well' (i.e. achieve a high reward) on tasks in each of the training populations, after a small number of updates. (3) **Testing the meta-learner:** testing the meta-learning agent's ability to learn new languages, which could be both artificial (emerged languages unseen during training) or human.

A diagram giving an overview of the L2C framework is shown in Figure 1. Phase 1 can be achieved in any number of ways, either through supervised learning (using approximate backpropogation) or via reinforcement learning (RL). Phases 2 and 3 follow the typical meta-learning set-up: to conserve space, we do not replicate a formal description of the meta-learning framework, but we direct interested readers to Section 2.1 of (Finn et al., 2017). In our case, each 'task' involves a separate population of agents with its own emergent language. While meta-training can also be performed via supervised learning or RL, Phase 3 must be done using RL, as it involves interacting with humans which cannot be differentiated through.

3 Seeded self-play

Seeded self-play (S2P) is a straightforward technique for training agents in simulation to develop complex behaviours. The idea is to *'seed'* a population of agents and use that data to train other populations. Fig 2 gives a pictorial representation

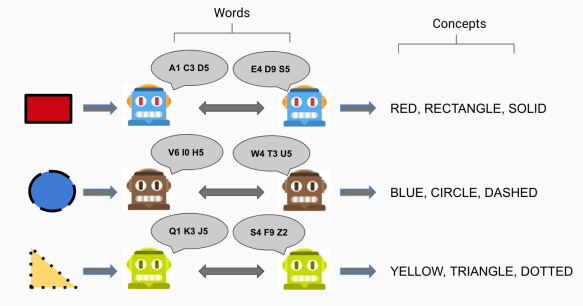

Figure 3: Lewis Signalling game with 2 agents, a Speaker and a Listener. For more details refer to Sec 4.

of S2P.

We collect some data which is sampled from a fixed seed population. This corresponds to the actual number of samples that we care about which is basically the number of human demonstrations. We first train each agent (a listener and a speaker) that performs well on these human samples. We call this step as the *imitation-learning* step. Then we take each of these trained agents (a pair of speaker and a listener) and deploy them against each other to solve the task via emergent communication. We call this step as the *fine-tuning* step. While these agents are exchanging messages in their emergent language, we make sure that the language does not diverge too much form the original language (i.e. the language of the fixed seed population). We enforce this by having a schedule over the *fine-tuning* and the *imitation-learning* steps such that both the agents are able to solve the task while also keeping a perfect accuracy over the seed data. We call this process of generating populations as *seeded self-play* (S2P).

4 Problem set-up

A speaker-listener game We construct a referential game similar to the Task & Talk game from (Kottur et al., 2017), except with a single turn. The game is cooperative and consists of 2 agents, a speaker and a listener. The speaker agent observes an object with a certain set of properties, and must describe the object to the listener using a sequence of words (represented by one-hot vectors). The listener then attempts to reconstruct the object. More specifically, the input space consists of p *properties* (e.g. shape, color) and t *types* per property (e.g. triangle, square). The speaker observes a symbolic representation of the input x, consisting of the concatenation of p one-hot vectors, each of length t. The number of possible inputs scales as t^p. We define the vocabulary size (length of each one-hot vector sent from the speaker) as $|V|$, and fix the number of words sent to be w.

Developing a compositional language To simulate a simplified form of human language on this task, we programatically generate a perfectly compositional language, by assigning each 'concept' (each type of each property) a unique symbol. In other words, to describe a blue shaded triangle, we create a language where the output description would be "blue, triangle, shaded", in some arbitrary order and without prepositions. By 'unique symbol', we mean that no two concepts are assigned the same word. We call these agents speaking compositional language as *Compositional Bots*. By generating this language programmatically, we avoid the need to have humans in the loop for testing, which allows us to iterate much more quickly. This is feasible because of the simplicity of our speaker-listener environment; we do not expect that generating these programmatic languages is practical when scaling to more complex environments.

Implementation details The speaker and listener are parameterized by recurrent policies, both using an embedding layer of size 200 followed by an LSTM of size 200. Both the speaker and the listener agents use the same number of parameters for encoding/decoding the message. The message produced by the speaker is a sequence of p categorical random variables which are discretized using Gumbel-Softmax relaxation (Jang et al., 2016; Maddison et al., 2016) with an initial temperature $\tau = 1$ which is annealed over a schedule with $\gamma = 0.7$ for every 10000 iterations. We set the vocabulary size to be equal to the total number of concepts $p \cdot t$. In our initial experiments, we train our agents using Cross Entropy which is summed over each property p. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 and a scheduler which anneals the learning rate with a $\gamma = 0.5$. We first demonstrate the results with a meta-learning listener (a *meta-listener*), that learns from the different speakers of each training population.

5 Experiments

5.1 Meta-learning improves sample efficiency

Here, we describe our initial results into the factors affecting the performance of L2C. Since our ultimate goal is to transfer to learning a human

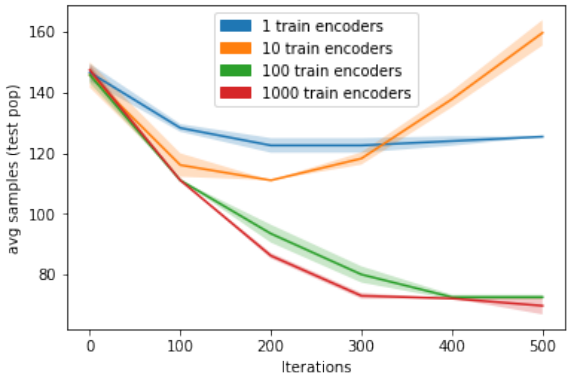

Figure 4: Performance of the meta-learner (in terms of number of training samples required to achieve >95% test accuracy on a test population) over the course of meta-training (horizontal axis), while varying the number of training encoders. Results averaged over 5 seeds, with standard error shown. Note the vertical axis is in log scale.

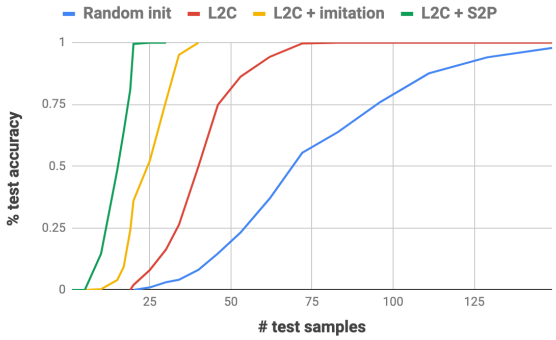

Figure 5: Varying performance across different number of test samples for all combinations of proposed frameworks. L2C+S2P performs the best, only needing 20 samples as compared to the 60 samples for L2C and 150 samples for randomly initialized agent.

language in as few human interactions as possible, we measure success based on the number of samples required for the meta-learner to reach a certain performance level (95%) on a held-out test population, and we permit ourselves as much computation during pre-training as necessary.

As can be inferred from Figure 4, having more training populations improves performance. Having too few training populations (eg: 5 train encoders) results in overfitting to the set of training populations and as the meta-learning progresses, the model performs worse on the test populations. For more than 10 training encoders, models trained with L2C require fewer samples to generalize to a held-out test population than a model not trained with L2C.

5.2 Self-play improves sample efficiency

We wanted to see if we can further reduce the number of samples required after L2C. So instead of doing L2C on a population of compositional bots, we train the population of agents using Seeded self-play (S2P). We collect some seed data from a single compositional bot which we call as *seed dataset*. Now we partition this data into train and test sets where the train set is used to train the agents via S2P. This set of trained populations is now used as the set of populations for meta-training (L2C). Fig 5 compares the results of populations trained via S2P and the compositional bots. It is evident that we need 40 fewer samples to generalize on the test set, when the populations are trained via S2P than using hard-coded bots.

6 Outstanding challenges

There are several immediate directions for future work: training the meta-agent via RL rather than supervised learning, and training the meta-agent as a joint speaker-listener (i.e. taking turns speaking and listening), as opposed to only listening. We also want to scale L2C training to more complicated tasks involving grounded language learning, such as the Talk the Walk dataset (de Vries et al., 2018), which involves two agents learning to navigate New York using language.

More broadly, there are still many challenges that remain for the L2C framework. In fact, there are unique problems that face each of the phases described in Section 2. In Phase 1, how do we know we can train agents to solve the tasks we want? Recent work has shown that learning emergent communication protocols is very difficult, even for simple tasks (Lowe et al., 2017). This is particularly true in the multi-agent reinforcement learning (RL) setting, where deciding on a communication protocol is difficult due to the non-stationary and high variance of the gradients (Lowe et al., 2017). This could be addressed in at least two ways: (1) by assuming the environment is differentiable, and backpropagating gradients using a stochastic differentiation procedure (Jang et al., 2016; Mordatch and Abbeel, 2017), or (2) by 'seeding' each population with a small number of human demonstrations. Point (1) is feasible because we are training in simulation, and we have control over how we build that simulation — in short, it doesn't really matter *how* we get our trained agent populations, so long

as they are useful for the meta-learner in Phase 2.

In Phase 2, the most pertinent question is: how can we be sure that a small number of updates is sufficient for a meta-agent to learn a language it has never seen before? The short answer is that it doesn't need to completely learn the language in only a few updates; rather it just needs to perform better on the language-task in the host population after a few updates, in order to provide a useful training signal to the meta-learner. For instance, it has been shown that the model agnostic meta-learning (MAML) algorithm can perform well when multiple gradient steps are taken at test time, even if it is only trained with a single inner gradient step. Another way to improve the meta-learner performance is to provide a *dataset* of agent interactions for each population. In other words, rather than needing to meta-learner perform well after interacting with a population a few times, we'd like it to perform well after doing some supervised learning on this dataset of language, and after a few interactions. After all, we do have lots of available datasets of human language, and not using this seems like a waste.

References

Amodei, D. and Hernandez, D. (2018). Ai and compute. *https://blog. openai. com/aiand-compute*.

Bahdanau, D., Hill, F., Leike, J., Hughes, E., Kohli, P., and Grefenstette, E. (2018). Learning to follow language instructions with adversarial reward induction. *arXiv preprint arXiv:1806.01946*.

Baroni, M. (2019). Linguistic generalization and compositionality in modern artificial neural networks.

Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T. H., and Bengio, Y. (2018). Babyai: First steps towards grounded language learning with a human in the loop. *arXiv preprint arXiv:1810.08272*.

David, L. (1969). Convention: a philosophical study.

de Vries, H., Shuster, K., Batra, D., Parikh, D., Weston, J., and Kiela, D. (2018). Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*.

Evtimova, K., Drozdov, A., Kiela, D., and Cho, K. (2017). Emergent Communication in a Multi-Modal, Multi-Step Referential Game. page 13.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*.

Foerster, J. N., Assael, Y. M., de Freitas, N., and Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. *CoRR*, abs/1605.06676.

Gauthier, J. and Mordatch, I. (2016). A paradigm for situated and goal-driven language learning. *arXiv preprint arXiv:1610.03585*.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. (2019). The compositionality of neural networks: integrating symbolism and connectionism. *arXiv:1908.08351 [cs, stat]*. arXiv: 1908.08351.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kottur, S., Moura, J. M., Lee, S., and Batra, D. (2017). Natural language does not emerge'naturally'in multi-agent dialog. *arXiv preprint arXiv:1706.08502*.

Lazaridou, A., Hermann, K. M., Tuyls, K., and Clark, S. (2018). Emergence Of Linguistic Communication From Referential Games With Symbolic And Pixel. page 13.

Lazaridou, A., Peysakhovich, A., and Baroni, M. (2016). Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*.

Lee, J., Cho, K., and Kiela, D. (2018). Countering language drift via grounding.

Lee, J., Cho, K., Weston, J., and Kiela, D. (2017). Emergent Translation in Multi-Agent Communication. *arXiv:1710.06922 [cs]*. arXiv: 1710.06922.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Mordatch, I. and Abbeel, P. (2017). Emergence of grounded compositional language in multi-agent populations. *arXiv preprint arXiv:1703.04908*.

Tieleman, O., Lazaridou, A., Mourad, S., Blundell, C., and Precup, D. (2018). Shaping representations through communication. *OpenReview*.

Wittgenstein, L. (1953). Philosophical investigations.