Multi-agent Communication meets Natural Language:
Synergies between Functional and Structural Language Learning

Angeliki Lazaridou[*], **Anna Potapenko**[*], **Olivier Tieleman**[*]
DeepMind, London, UK
`{angeliki,apotapenko,tieleman}@google.com`

Abstract

We present a method for combining multi-agent communication and traditional data-driven approaches to natural language learning, with an end goal of teaching agents to communicate with humans in natural language. Our starting point is a language model that has been trained on generic, not task-specific language data. We then place this model in a multi-agent self-play environment that generates task-specific rewards used to adapt or modulate the model, turning it into a task-conditional language model. We introduce a new way for combining the two types of learning based on the idea of reranking language model samples, and show that this method outperforms others in communicating with humans in a visual referential communication task. Finally, we present a taxonomy of different types of language drift that can occur alongside a set of measures to detect them.

1 Introduction

In this work, we aim at making agents communicate with humans in natural language. Our starting point is a language model that has been trained on generic, not task-specific language data. We then place this model in a multi-agent communication environment that generates task-specific rewards, which are used to adapt or modulate the model, making it task-conditional. We thus propose to decompose the problem of learning language use into two components: learning "what" to say based on a given situation, and learning "how" to say it. The "what" is the essence of communication that underlies our intentions and is chosen by maximizing any given utility, making it a *functional*, utility-driven process. On the other hand, the "how" is a surface realization of our intentions, i.e., the words we use

———————————
[*]All authors contributed equally.

to communicate this "what" successfully. This factorization into *content planning* (here, "what") and *surface realization* (here, "how") moves us away from end-to-end neural generation systems and is in line with traditional methods of natural language generation (Reiter and Dale, 1997). More importantly, it enables us to bring together two different strands of research: traditional data-driven natural language learning and multi-agent communication.

Traditional approaches to natural language learning (Kneser and Ney, 1995; Mikolov et al., 2010; Sutskever et al., 2014; Vinyals and Le, 2015; Radford et al., 2019) are based on inferring *structural* properties of language from text corpora, often in a passive regime, dissociated from communication. While this type of learning is great for learning general statistical associations between symbols (e.g., adjectives come before nouns) and even inferring semantic relations, it ignores the functional aspects of communication, i.e., the fact that people use words to coordinate with others and make things happen in the world (Wittgenstein, 1953; Austin, 1975; Clark, 1996).

On the other hand, multi-agent communication research (Foerster et al., 2016; Lazaridou et al., 2017; Havrylov and Titov, 2017; Evtimova et al., 2017; Lee et al., 2019) puts communication at the heart of agents' (language) learning. Implemented within a multi-agent reinforcement learning setup, agents start *tabula rasa* and form communication protocols that maximize task rewards. While this purely utilitarian framework results in agents that successfully learn to solve the task by creating a communication protocol, these emergent communication protocols do not bear core properties of natural language. Chaabouni et al. (2019) show that protocols found through emergent communication, unlike natural language, do not conform to Zipf's Law of Abbreviation; Kottur et al. (2017) find that communication protocols do not follow composi-

tionality patterns of natural language, and Lazaridou et al. (2018) find emerged protocols to be sensitive to different experimental conditions. This growing set of alarming results on emergent communication raises doubts about the use of this type of *functional* learning as a viable alternative to language learning.

Concluding that neither approach on its own is adequate for learning language use, we propose a method for combining the best of both worlds. Generic language data can be used effectively as a good prior model of language, encapsulating its intrinsic *structural* properties, i.e., are only used for the "how" in the form of generic language models. Conversely, multi-agent interactions, that provide rewards specific to the task of interest, now only need to be used for the *functional* learning of language use, i.e., learning the "what".[1]

The contributions of this paper are as follows. First, we propose a general research program of language learning that combines two learning signals coming from multi-agent communication and traditional data-driven natural language learning techniques. We present a concrete study in the context of a referential communication game (see Section 2) between a speaker and a listener, where the traditional data-driven language learning takes the form of image captioning, and the functional learning takes the form of agent self-play (see Section 3). We then present a new approach for combining the two learning signals, i.e., reward-learned rerankers (see Section 4), and compare this to existing approaches using a human study (see Section 5). We discuss shortcomings of this program with respect to different types of *language drift* that can occur, and introduce a number of automatic measures to detect them (see Section 6). Finally, we show how such a program under oracle rewards can be a viable approach moving towards learning language use from human rewards (see Section 7).

---

[1]About the terminology: by 'traditional data-driven natural language learning', we mean language modelling of the next-word-prediction variety. This type of learning does not involve any *use* of the language or other context, and as such only focuses on word statistics. Since the structure of the language is a large part of those statistics, and the role of the generic language models in our proposed combined systems is to provide structural knowledge of language, we also use the term 'structural learning'. We contrast this with the purely usage-driven, reward-based learning of the type seen in emergent communication research. Since the function, rather than the structure or statistics, is the only thing that matters for such a learner, we also use the term 'functional learning'.

2 Research framing

Our research can be framed in the following scenario. An agent needs to perform a *functional communication task* in a natural language (in this work, English). However, examples of linguistic communication about this functional task are not available - the only natural language data that can be used consist of examples of *generic natural language*, which are not grounded in the functional task. Recasting the task as a multi-agent language game provides a way to obtain a *reward* that judges whether an utterance elicited the correct behaviour by a listener.

2.1 Experimental setup

In this work, we instantiate the research in the following way: the *functional task* is a visual referential communication game for a target image in the context of a distractor, the *reward* is based on success in referential communication where a listener is tasked to pick the correct image within distractors guided by the speaker's description, and the *generic natural language data* are captioning data.

Visual referential communication game. There are two players, the speaker and the listener. The speaker sees a target object and needs to communicate an utterance about it in the context of distractors; both target and distractors are represented as images. The listener is presented with the same set of images, but without the knowledge of which is the target, and needs to identify the target image relying on the utterance being communicated by the speaker. The utterance takes the form of sequences of word-like units. If the listener's choice is correct they both receive a positive reward, else they receive the same negative reward.[2]

Dataset and referential splits. For playing the visual referential communication game, we use a multi-modal dataset, the Abstract Scenes (Zitnick and Parikh, 2013) which contains 10k synthetic images accompanied with descriptive captions (on average 6 per image) (see Figure 1).[3] The cap-

---

[2]The task we consider is essentially *discriminative image captioning* (Vedantam et al., 2017; Dai and Lin, 2017; Andreas and Klein, 2016). Here we are using it as a placeholder of a communication task to illustrate our general framework. Thus, we are not incorporating any explicit bias in the model about this particular task. The only task-specific information we use is communicated via the reward.

[3]Other multi-modal datasets like MSCOCO (Lin et al., 2014) or Flickr (Thomee et al., 2016), while providing com-

Jenny is scared of the bear
Mike is scared of the bear
Jenny and Mike sit by a fire
Jenny and Mike are sitting
A bear is scaring mike and jenny

Figure 1: Example image and ground-truth captions from the Abstract Scenes dataset used in this study.

speaker (human)	easy	difficult
random	0.92	0.81
discriminative	1.0	0.97

Table 1: Accuracy performance of a human listener with a human speaker producing either random or discriminative caption on the easy and difficult splits.

tions typically refer to diverse aspects of the scene (characters and actions), providing a rich and challenging environment for an agent to evolve the captioning skills for successful communication. In our experiments, we split the dataset into 80/10/10 for train/validation/test sets. We use the test images to create two referential splits, i.e., *easy* and *difficult*, as a function of the similarity between the target and distractor images. Each split contains 1000 pairs of a target and a distractor.

Human performance and setup validation. In order to assess the difficulty of the task in the presence of the particular data (images and captions) we perform a human study in the reference game with a human speaker and a human listener, where the human speaker can only communicate one of the existing captions of the target image. We perform the human study under two conditions. In the first condition, the human speaker has only access to the ground-truth captions and does not have access to the distractor image, thus has to pick a random caption. This corresponds to the *perfect* structural knowledge of English but *no* knowledge of the functional task and it is the human upper-bound of a captioning system performance on this task. In the second condition, the speaker has access to both the ground-truth captions and the distractor image, thus is able to pick a discriminative caption to communicate. For each condition, we collect 50 rounds

---

plex naturalistic images, often have a repetitive set of captions, highlighting one particular aspect of the scene and suffer from a human reporting bias (Misra et al., 2016). By using Abstract Scenes, we have left certain visual challenges out of the scope of the work, obtained cleaner multi-modal associations between words and objects, and focused on the language use for referential communication.

of games and present results in Table 1. We see that the task-specific condition outperforms the first condition, indicating that in our current setup there is enough space to improve upon models based on structural-only learning (i.e., captioning models). Moreover, the good performance of discriminative caption speaker demonstrates that (in principle) the captioning data can be used in a successful communication with a human for this task.

3 Multi-agent communication setup

3.1 Speaker

The speaker is the primary learner in this research, aiming at creating a model that is able to use natural language in a communicative scenario, and consists of standard visual and language modules. To convert images to embeddings u, we use a pretrained ResNet (He et al., 2016) (parametrized by θ^{resnet}) and feed its last layer output into a one-layer MLP (parametrized by θ_S^{MLP}). To generate a message m, we use a one-layer LSTM (Hochreiter and Schmidhuber, 1997) (parametrized by θ_S^{LSTM}), adding embeddings u at each time step as additional context. Section 4 presents different speaker models consisting of these modules.

We also design two oracle speakers (with no weights) that have direct access to ground-truth captions of images at test time. The random caption speaker outputs one of the ground-truth captions for the target image at random. Since this speaker is not aware of the functional goal, their performance will indicate whether having only good grounded language skills is enough for communication success in our setup. We also build an oracle speaker that is task-aware; discriminative caption speaker uses a simple word-overlap heuristic to pick the target's caption that has the least word overlap with any of the distractor's captions (the score is normalized by the captions' length excluding stop-words).

3.2 Listener

Throughout the experiments, we need a way to estimate performance on the functional communication task, either for evaluation or to provide rewards during training acting as a scaffolding to learn the speaker model. Ideally, this performance signal should be provided by a human who is interacting online with the speaker agent. While we do so for evaluation reasons, for training we approximate this quantity with a learned component, an agent listener.

To convert images to embeddings u, we use the same pre-trained ResNet as for the speaker and feed its last layer output into a one-layer MLP (parametrized by θ_L^{MLP}). Following that, the listener uses an LSTM (parametrized by θ_L^{LSTM}) to embed the utterance m received by speaker, creating embedding v. Finally, the listener picks the image with the highest dot-product similarity between the embedded message v and the embeddings u_t and u_d for target and distractor. Since we know which image candidate is the intended referent, we cast this problem as supervised learning and update the listener's weights $\theta_L = \{\theta_L^{MLP}, \theta_L^{LSTM}\}$ optimizing cross-entropy. Finally, the listener assigns reward 1 to the speaker if they identified the correct image, else reward -1.

We consider two different setups: a joint listener, which is trained together with the speaker, as commonly done in the emergent communication literature, and a fixed listener that is pre-trained to perform best-response to the oracle discriminative caption speaker and stays fixed throughout the learning of the speakers with the sole use of providing them rewards. We expect the latter setup to be less prone to language drift issues due to the grounding of the discriminative caption speaker to language data. thus potentially resulting in better communication with human listeners. We also use the fixed listener for evaluation of all speakers.

4   Methods for learning language use

We describe ways to estimate the speaker's generative model $p_{\theta_S}(m|u,t)$ for message m, conditioned on target and distractor embeddings $u = [u_t, u_d]$ and target image index $t \in \{0, 1\}$.

4.1   Functional-only learning

This type of learning language use is identical to experiments commonly conducted in the literature of emergent communication (Lazaridou et al., 2017; Havrylov and Titov, 2017; Bouchacourt and Baroni, 2018; Evtimova et al., 2017; Graesser et al., 2019), i.e., the speaker learns to emit communication utterances m in order to maximize the communication task reward (see Section 3.2 for a discussion on how this reward is computed). Concretely, the weights $\theta_S = \{\theta_S^{MLP}, \theta_S^{LSTM}\}$ of the speaker policy $\pi_{\theta_S}(m|u,t)$ are updated via the REINFORCE update rule (Williams, 1992) using rewards r^L provided by the listener, i.e., we optimize $L^{functional} =$

$$-r^L(m,u,t)\sum_{i=1}^{I} \log p_{\theta_S^{LSTM}}(m^i|m^{<i},u), \quad \text{where}$$

$u = [u_t; u_d]$, $m^i \in V$, vocabulary size $|V| = 100$, and message length $I = 10$.[4] Note, that while this type of learning results in a language that is maximally functionally correct for the given task reward, this language is not natural language, i.e., the symbols are not grounded to natural language.

4.2   Structural-only learning

This type of learning ignores the functional aspect of communication and communicates utterances that reflect intrinsic structural properties of language, i.e., that are fluent, grammatical and related to the target. Here, we used paired data in the form $\langle u, c \rangle$, where u is a visual embedding and c is the associated caption, and learn an image captioning model. The speaker's parameters $\theta_S = \{\theta_S^{MLP}, \theta_S^{LSTM}\}$ are optimized using cross-entropy, i.e.,

$$L^{structural} = -\sum_{i=1}^{I} \log p_{\theta_S^{LSTM}}(c^i|c^{<i},u), \quad \text{where}$$

$u = u_t$, $c^i \in V$, $|V| = 2685$ (the vocabulary size) and $I = 25$, i.e., the longest caption in the dataset. We approximate the speaker model $p_{\theta_S}(m|u,t)$ with the captioning one, which ignores distractor, thus the communication task. We construct two speakers with different decoding schemes: greedy uses greedy decoding, while sample picks the highest probability message among $k = 20$ stochastic samples (temperature $\tau = 2.0$).

4.3   Structural and functional learning

We now describe several ways in which both types of learning are used to learn language use. In all cases, we equip the speaker with a base image captioning model similar to the one presented in Section 4.2, which is used to calculate $p_{\theta_S^{LSTM}}(c^i|c^{<i},u_t)$. The functional part is learned via the REINFORCE update rule optimizing the task reward (i.e., listener's accuracy in the referential task). However, speakers differ in how they parametrize $p_{\theta_S}(m|u,t)$ and whether the task reward is used to update the weights $\{\theta_S^{MLP}, \theta_S^{LSTM}\}$ of the base captioning model.

4.3.1   Reward finetuning

The simplest approach is to *first* use existing pre-trained components for which we have available corpora in order to learn the statistical properties

---

[4]In all experiments using REINFORCE we add an entropy regularization term to the loss.

of language, and *then* steer the language use to be functionally appropriate using reward finetuning for the given task. We use paired data in the form $\langle u, c \rangle$ to learn the weights $\theta_S = \{\theta_S^{MLP}, \theta_S^{LSTM}\}$ of a base image captioning model following Section 4.2, and then we perform functional learning by using the listener's reward to optimize the weights θ_S as in Section 4.1. While this method is conceptually simple, it becomes challenging when the task requires extending the conditioning part of the base model. Here, we need to change the conditioning of the base captioning model from $u = u_t$ to $u = [u_t; u_d]$, to allow conditioning on the distractor. Since this is not trivial (the base image captioning model has been learned by conditioning only on one image embedding), we keep the conditioning $u = u_t$ also during finetuning with REINFORCE. Thus, similar to the image captioning model, we approximate $p_{\theta_S}(m|u,t)$ with $p_{\theta_S^{LSTM}}(m|u_t)$. However, unlike image captioning, the information about distractors flows into the policy, since the weights θ_S are optimized using the listener's reward which considers distractors.

Since the gradients from optimizing the functional task are sent all the way into the base captioning model, this causes catastrophic forgetting of the *core knowledge* of language, leading to *language drift*. Thus, we use a language regularizer term in the form of Kullback-Leibler divergence between pre-trained and fine-tuned language modeling distributions (Havrylov and Titov, 2017).

4.3.2 Multi-task learning

An alternative is to conduct both types of learning (i.e., image captioning and functional learning) at the same time (Lazaridou et al., 2017; Lee et al., 2019). This takes the form of multi-task learning optimizing $\lambda_f L^{functional} + \lambda_s L^{structural}$, where $\lambda_f = 1$. Like in reward finetuning, the gradients of the reward learning flow into the weights of a base captioning model, leaving us with questions about a trade-off between task success and quality of language. Therefore, we introduce two variants of this model depending on the importance of the language component, i.e., one variant with $\lambda_s = 0.1$ and a language-regularized one with $\lambda_s = 1$.

4.3.3 Reward-learned rerankers

Finally, we introduce a new way of learning language use in the multi-agent communication setup. As before, we train the core language capabilities of a speaker using the image captioning task objec-

tive, but after this pre-training phase, the weights of this model are frozen. The functional part is then viewed as *learning to use* this general knowledge of language grounded in images. This is operationalized as learning to *rerank* samples obtained from the captioning model optimizing the listener's reward. The action space of this speaker are sentences, as opposed to words used commonly in the literature of emergent communication. We emphasize that by leveraging the idea of reranking, we are able to take a *task-unconditional* model, i.e., a captioning model that only conditions on the target, and extend its conditioning turning it into a *task-conditional* model, i.e., a discriminative captioning model that conditions also on the distractor.

Below we consider two concrete reranker models. In both cases, the message generation happens in two steps. First, we sample $|S| = 20$ candidates from the pre-trained and fixed image captioning model $p_{\theta_S^{LSTM}}(m|u_t)$. Then, we pick the best sample s using a task-conditional reranking score $p(s|u,t)$. The reranking score can be viewed as a new policy $\pi_{\theta_S}(s|u,t)$ that operates in the space of samples S drawn from the task-unconditional model. This policy introduces an additional set of trainable parameters θ_S^{rerank} that are learned with REINFORCE. Thus, the full set of weights for this speaker is $\theta_S = \{\theta_S^{MLP}, \theta_S^{LSTM}, \theta_S^{rerank}\}$. Crucially, the two learning signals, i.e., *structural* and *functional*, affect different set of weights, i.e., $\{\theta_S^{MLP}, \theta_S^{LSTM}\}$ and θ_S^{rerank} respectively, allowing submodules to specialize.

Product of experts reranker. In this model we parametrize the policy as a product of experts (PoE): $\pi_{\theta_S}(s|u,t) \propto p(s|u,t)^{\lambda_f} p(s|u_t)^{\lambda_s}$, where $u = [u_t; u_d]$ and $\lambda_f = 1$. The second term is the image captioning message probability, *renormalized* over the samples space, thus bringing general language knowledge grounded in images. The first term adjusts for the task specifics. To model that, we re-embed the samples using transformed bag-of-words, thus the trainable parameters of the reranker θ_S^{rerank} are word embeddings and additional MLP weights. We combine target and distractor embeddings into a single vector and compute the dot-product similarity between this vector and each of the bag-of-words representations of samples. Finally, these scores are passed through a softmax layer to obtain $p(s|u,t)$. We introduce two variants of the model, one with $\lambda_s = 0$ and a language-regularized one with $\lambda_s = 1$.

Noisy channel reranker. Following Bayes rule, we factorize the speaker's policy as follows: $\pi_{\theta_S}(s|u,t) \propto p(t|s,u)p(s|u)$, where $u = [u_t; u_d]$. We omit the distractor vector u_d in the conditioning of the prior, arriving to $p(s|u_t)$ from the PoE reranker above. The crucial difference is that the first term now represents the speaker's approximation of the listener's behaviour. As before, we represent samples with the transformed bag-of-words, but then compute their dot-product similarities with each image separately and normalize with softmax across the images to obtain the probability of the target $p(t|s,u)$. This reranker model is closely related to pragmatic speakers in Rational Speech Act (RSA) framework (Andreas and Klein, 2016; Monroe and Potts, 2015; Cohn-Gordon et al., 2018; Fried et al., 2018). However, while the RSA model assumes a given and fixed listener model, here we are learning the model of the listener that the speaker is using by optimizing end-to-end the listener's reward. Thus, when doing multi-agent communication using the noisy channel model, there exist two components that produce probability distributions of the same type $p(t|s,u)$; one belongs to the listener, thus the speaker has no access to it (e.g., this listener in the future could be a human sending rewards), while the other belongs to the speaker corresponding to their model of the listener.

5 Speakers trained jointly with listeners

Table 2 presents referential success when speakers are *trained* with rewards from a joint listener, i.e., a listener being learned jointly with the speaker.

We conduct three different evaluations: at *test* time we play against the fixed listener, *human* listeners and the joint listener the speaker was trained with. While fixed listener is the same for all speakers, the joint listener is speaker-specific. We report results on two splits: for the *easy* and *difficult* split we report referential success of the joint listener, and for the latter split, we also report results of the fixed and human listener.

To compute referential success using human listeners, we collect 400 annotations for each speaker model. To avoid annotators adapting to model-specific strategies, we group predictions of similar models and collect annotations in three sessions (one for each group), during which we present annotators with predictions from a model sampled from that group.[5]

Group 1: image captioning (greedy/sample), noisy chan-

	Easy split	Difficult split		
	joint	joint	fixed	human
Functional-only learning				
emergent (§4.1)	0.99	0.98	-	0.5
Structural-only learning				
image captioning (§4.2)				
sample	0.92	0.78	0.77	0.77
greedy	0.91	0.77	0.73	0.78
Structural & functional learning				
Gradients from reward affect base captioning model				
reward finetuning (§4.3.1)				
no KL	0.95	0.82	0.63	0.62
with KL	0.93	0.79	0.77	0.69
multi-task learning (§4.3.2)				
$\lambda_s = 0.1$	0.98	0.94	0.71	0.71
$\lambda_s = 1$	0.96	0.90	0.69	0.69
Reranking (§4.3.3), base captioning model unchanged				
PoE, $\lambda_s = 0$	0.99	0.92	0.81	0.81
PoE, $\lambda_s = 1$	0.98	0.91	0.83	0.78
noisy channel	0.96	0.83	0.84	0.86*
Utilizing ground-truth captions from the dataset				
Oracle speakers, no weights learned (§3.1)				
random	0.87	0.74	0.72	0.81
discriminative	0.87	0.73	0.82	0.87*
Reranking (§4.3.3) ground-truth captions				
PoE (§4.3.3)	0.95	0.88	0.85	0.93*
noisy channel (§4.3.3)	0.95	0.78	0.83	0.88*

Table 2: Referential success of speakers (by rows) *trained* with joint listener and then *tested* with joint, fixed and human listener (by columns).* indicates significance over the image captioning (greedy) when *tested* with humans ($p < 0.005$, bootstraping test).

5.1 Referential success of joint listeners

All models perform quite similarly in the *easy* split, whereas we observe larger gaps in the difficult split. In terms of joint accuracy results in the difficult split, reward finetuning has the lowest performance among models that are optimizing rewards, perhaps due to its large action space (i.e., the vocabulary size $|V| = 2685$), making it a hard RL exploration problem. multi-task, despite having the same action space performs better, probably due to the captioning objective being optimized concurrently facilitating the learning dynamics. Finally, the best results in both splits are obtained by the emergent communication model, that achieves near perfect performance. We believe this is the case since this speaker is the least constrained of all, since we can think of all other speakers (i.e., the ones that combine both types of learning) as being regularized towards producing natural language.

---

nel, PoE). Group 2: multi-task, reward finetuning. Group 3: random, discriminative, PoE and noisy channel with ground-truth captions.

5.2 Referential success of human listeners

Somewhat alarmingly, we observe the joint performance is not predictive of the human's one across the board, hinting to issues regarding *pragmatic* drift (we will further discuss this in Section 6). In the most extreme case, while the emergent communication speaker achieved the highest results when playing against a listener jointly learned with the speaker, this comes in the expense of human performance: functional learning alone results in maximally uninterpretable protocols, and as such humans are at random when playing against such a model.

Speakers that combine both types of learning achieve good human performance, with reward-learned reranker models, i.e., noisy channel and PoE being the best. In their case, they outperform the image captioning baselines, even approaching the discriminative oracle speaker based on ground-truth captions. This indicates their effectiveness in extending the conditioning of the underlying image captioning to the distractor image with the reward coming from the listener, turning like this the base image-captioning model into a task-specific referential captioning model. Moreover, when giving the rerankers a perfect captioning model in the form of ground-truth captions of target images, performance of noisy channel and PoE surpass the oracles' (see last two columns of Table 2); as the community improves the base language models, we should expect this to also result in net improvement in the reranker models.
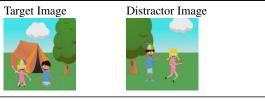
Finally, we also observe that the fixed grounded listener is significantly predictive of the human performance ($p < 0.005$, t-test).[6] This is encouraging, since as we will show in Section 7, we can use this listener as a fixed model that provides rewards to the speaker model.

6 Language drift and how to detect it

We show that the multi-agent communication framework is prone to language drift (Lee et al., 2019), i.e., when protocols diverge from human language. We present a taxonomy of different types that occur in this framework, alongside a set of automatic measures to detect it.

6.1 Structural drift: Definition and measures

The most basic type of drift that manifests in the emergent communication setup relates to the core

---

Target Image	Distractor Image

Structural-only learning
image captioning (§4.2)
| sample | **jenny** is wearing a hat |
| greedy | <u>mike</u> is wearing a hat |

Structural and functional learning
Gradients from reward affect base captioning model
reward finetuning (§4.3.1)
| no KL | it is camping **camping** [...] camping |
| with KL | mike is sitting <u>on</u> the tent |
multi-task learning (§4.3.2)
| $\lambda_s = 0.1$ | mike is jenny on <u>the the</u> tent |
| $\lambda_s = 1$ | mike is sitting on the ground |

- -
Reranking (§4.3.3), base captioning model unchanged
PoE, $\lambda_s = 0$	the tent is in the <u>tree</u>
PoE, $\lambda_s = 1$	mike and jenny are sitting **on the ground**
noisy channel	jenny is wearing a **funny hat**

Table 3: Examples of generated messages. We underline <u>wrong</u> and bold **correct** cases of language usage.

---

structural properties of the generated language, i.e., its fluency and grammaticality with respect to *natural language* (this is also referred to by Lee et al. (2019) as "syntactic"). Looking at Table 3, a clear example of this type of drift happens when models update the base captioning model. reward finetuning (no KL) does not produce at all grammatical sentences, while multi-task ($\lambda_s = 0.1$) appears to suffer less, only occasionally producing slightly ungrammatical sentences by repeating consecutive words. We term this *structural drift* and we quantify it as the log probability of the generated message under a pre-trained unconditional language model (column log p(m) in Table 4).

6.2 Semantic drift: Definition and measures

The second type of drift is the *semantic* drift. This relates to whether the generated message is grounded with regards to the target object, i.e., its adequacy with respect to the literal semantics of the target (this is also referenced by Lee et al. (2019) as "semantic"). We have qualitatively observed instances of this type of drift in the PoE, which occasionally shifts the semantics of words, e.g., using the word *tree* to refer to *ground* as seen in Table 3. To measure it, we use a pre-trained image-conditional language model and compute the target-image conditional log probability of the generated

---

[6] All t-tests are conducted between two distributions of scores dichotomized on human performance.

	log p(m)	log p(m|i)	1-gram	3-gram
Structural-only learning				
image captioning (§4.2)				
sample	-8.71	-7.77	0.81	0.37
greedy	-8.63	-7.72	0.73*	0.30*
Structural & functional learning				
Gradients from reward affect base captioning model				
reward finetuning (§4.3.1)				
no KL	-442.00	-279.55	0.33	0.00
with KL	-11.75	-10.78	0.70*	0.22*
multi-task learning (§4.3.2)				
$\lambda_s = 0.1$	-18.08*	-19.67	0.78*	0.18*
$\lambda_s = 1$	-10.68*	-10.64	0.63*	0.18
Reranking (§4.3.3), base captioning model unchanged				
PoE, $\lambda_s = 0$	-10.18*	-8.79	0.71*	0.24*
PoE, $\lambda_s = 1$	-8.95	-7.94	0.78*	0.30*
noisy channel	-10.02	-8.59	0.76	0.30

Table 4: Language drift measures, lower scores mean higher drift. *indicates that the measure was significantly predictive of the human listener performance on the referential task (p<0.005, t-test).

message (column log p(m|i) in Table 4).

These two log probability-based measures do not assume access to language data for the target objects, and as such can be computed from general unconditional and domain-specific conditional language models. In this particular case though, since we also have access to language data for the target images (i.e., captions in English), and assuming that these data describe everything that is true about the target, we can use simple n-gram statistics as proxies of semantic drift (i.e., in this case 1-gram word overlap ignoring stop word and 3-gram word overlap between the ground-truth captions and the speaker-generated message). Moreover, all these measures do not take into account the specific communication task the speaker has to perform, i.e., our measures do not consider any information about the distractor object, making them easily adaptable to other tasks.

6.3 Structural and semantic drift results

In Table 4 we report performance of different models under these automatic measures. The structural score log p(m) reflects the qualitative observations made from Table 3, i.e., multi-task and reward finetuning, have the highest structural drift, with the latter performing significantly worse than all the models. In contrast, the reranker models that do not update the base captioning model, i.e., PoE and noisy channel, perform the best on the semantic score by construction; both models directly incorporate in their

models a component associated with the semantic score (i.e., the samples taken from the image-conditional model alongside the associated probabilities). Moreover, they also perform well on all other measures, indicating their robustness against language drift. Finally, all the model-specific language regularizers (KL for reward finetining, $\lambda_s = 1$ for multi-task and $\lambda_s = 1$ for PoE) we introduced were effective in limiting both types of language drift (as also seen in Table 3).

6.4 Pragmatic drift

Finally, we identify a novel type of drift, i.e., *pragmatic* drift, which relates to the divergence between a human's interpretation of the message from the interpretation a speaker will assume. Unfortunately, this type of drift is perhaps the most difficult to capture in an automatic way as it is *task specific* and requires access to the exact interpretation that the human would ascribe to the message. As a proxy of pragmatic drift, we use the difference between the agent- and human-listener referential success; if the joint referential success is higher than the human's one, then the speaker assumes an interpretation of the message that is different from the human's one, resulting in lower human performance. An extreme example of this drift manifests when the joint listener achieves almost perfect referential success whereas a human listener is at random, as in the case of emergent communication. However, in this case the messages are maximally uninterpretable with the lowest possible performance in both structural and semantic scores.

Hence, a natural question to ask is *to what degree (if at all possible) pragmatic drift can manifest in the absence of the other two types of language drift.* Or, put differently, does the emergent communication for learning language use hide any other pathological behaviour for models that do not suffer a lot from structural and semantic drift, as in the case of PoE and noisy channel? To study this, we create a setup where PoE is guaranteed to have a perfect knowledge of (grounded) language. Namely, it uses the reward to rerank *ground-truth captions* associated with the target image (note, our dataset provides up to five captions per image). Moreover, we perform several ablations where we allow the updating of different parameters in the speaker's and listener's model by unfreezing components.

Table 5 presents the results of the joint and human referential success. The main finding is that

Weights learned with RL	joint	human	Δ
reranker	0.88	0.92	-0.04
reranker + speaker ResNet	0.92	0.90	+0.02
reranker + both agent ResNets	0.96	0.88	+0.08

Table 5: Referential success for PoE with gold captions when updating different components during training with joint listener.

Model	fixed	human
reward finetuning, with KL	0.81	0.75
multi-task learning, $\lambda_s = 0.1$	0.80	0.68
PoE, $\lambda_s = 0$	0.93	0.86
noisy channel	0.88	0.87

Table 6: Referential success when training speakers with the fixed listener.

by increasing the number of components that get updated using the joint reward, the margin between the referential success of the two types of listeners increases. Despite the fact that the speaker is using human language that is perfectly fluent and accurate with respect to the target image (since the reranker operates on captions associated with the target image), while the joint listener is able to communicate with the agent speaker, the human listener achieves significantly lower performance.

In one test example, the speaker said *Mike has a hat*, which was equally true for both images making the human pick at random. So, how could the listener pick correctly? The speaker had reached a pact with the listener that the interpretation of this message will be something beyond what the phrase means (e.g., *Mike has a yellow hat* or the intensity of the pixels in the target image is lower). Since speaker and listener learn together, they co-adapt, forming conventions (or conceptual pacts (Brennan and Clark, 1996)) that differ from humans', even in the presence of *fluent* and *grounded* language.

7 Speakers trained using fixed listener

In the previous section we showed that learning a speaker using a learned reward module as a scaffolding (i.e., the joint listener) can lead to pragmatic drift. In this section, we use a grounded reward as scaffolding. In the absence of a human listener to provide rewards for learning, we use the oracle fixed listener, which was found in Section 5 to be predictive of human referential success. It is pre-trained, stays fixed and just provides rewards for training the speaker. As speakers, we use the models that scored the highest in Table 2 and retrain them against fixed. Table 6 presents the results of referential success against fixed and human listeners. Using a grounded reward results in better performance for the weaker models. The small gap between the rerankers in the two experimental setups points that using a learned reward module (joint) holds promise, despite the different types of language drift. Moreover, we show that our mod-

els for learning language can be used against fixed reward models, potentially learning directly from human rewards (Ziegler et al., 2019).

8 Discussion and Limitations

We presented a method for teaching agents to communicate with humans in natural language, by **combining two learning signals coming from multi-agent communication and traditional data-driven natural language learning techniques**, which adds on recent efforts of blending emergent communication with natural language (Lowe et al., 2020; Lu et al., 2020).

Self-play between speakers and listeners can result in **language drift, the most severe of which being pragmatic drift**. Since speakers and listeners are learning concurrently, they can co-adapt to pair-specific policies that deviate from the policies that humans learn. This pathological behaviour of self-play is not specific to language and extends to other policies (Carroll et al., 2019).

Finally, we introduced the **reward-learned reranker approach which alleviates language drift and achieves the highest human performance**, by constraining the functional learning to happen on the level of utterances generated by a pre-trained language model. However, since the functional signal is not currently influencing the sampling from the language model, this will lead to poor performance when using more general language models with weaker conditioning (e.g. GPT-2 (Radford et al., 2019)) whose samples potentially do not fit the functional context. Moving towards integrating our findings into more realistic applications of self-play, e.g., user simulation in dialogue (Schatzmann et al., 2006; Shah et al., 2008), these shortcomings need to be addressed.

Acknowledgements

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

John Langshaw Austin. 1975. *How to do things with words*. Oxford university press.

Diane Bouchacourt and Marco Baroni. 2018. How agents see things: On visual representations in an emergent language game. *arXiv preprint arXiv:1808.10696*.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. In *Advances in Neural Information Processing Systems*, pages 5175–5186.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019. Anti-efficient encoding in emergent communication. *arXiv preprint arXiv:1905.12561*.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 439–443.

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907.

Katrina Evtimova, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho. 2017. Emergent communication in a multi-modal, multi-step referential game. *arXiv preprint arXiv:1705.10369*.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2137–2145. Curran Associates, Inc.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pages 3314–3325.

Laura Graesser, Kyunghyun Cho, and Douwe Kiela. 2019. Emergent linguistic phenomena in multi-agent communication games. *arXiv preprint arXiv:1901.08706*.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in neural information processing systems*, pages 2149–2159.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184 vol.1.

Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge'naturally'in multi-agent dialog. *arXiv preprint arXiv:1706.08502*.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. *ICLR*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. *ICLR*.

Jason Lee, Kyunghyun Cho, and Douwe Kiela. 2019. Countering language drift via visual grounding. 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)*, pages 4376–4386.

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

Ryan Lowe, Abhinav Gupta, Jakob N. Foerster, Douwe Kiela, and Joelle Pineau. 2020. On the interaction between supervision and self-play in emergent communication. *ArXiv*, abs/2002.01093.

Yuchen Lu, Soumye Singhal, Florian Strub, Olivier Pietquin, and Aaron C. Courville. 2020. Countering language drift with seeded iterated learning. *ArXiv*, abs/2003.12694.

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

Ishan Misra, C. Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. pages 2930–2939.

Will Monroe and Christopher Potts. 2015. Learning in the rational speech acts model. *arXiv preprint arXiv:1510.06807*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2008. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59(2):64–73.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

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

Ludwig Wittgenstein. 1953. *Philosophical investigations*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

C Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016.

A Hyperparameters

The following tables represent our choice of hyper-parameters in the speaker and listener agents. Hyperparameters in Table 7 where chosen in the image captioning task using the validation set. Hyperparameters in Table 8 where chosen in the referential task using the validation set.

Agent	Hyperparameter	Value
listener	LSTM hidden size	512
speaker	LSTM hidden size	512
listener	visual embeddings size	512
speaker	visual embeddings size	1024

Table 7: Settings shared across all experiments.

Model	Hyperparameter	Value
fine-tuning++	KL regulatization	0.1
multi-task	structural weight	0.1
PoE	structural weight	0
noisy channel / PoE	number of samples	20
noisy channel / PoE	message embedding size	1024
noisy channel / PoE	entropy regularization	0.1

Table 8: Settings for particular speakers.

B ResNet module

We use ResNet-50 (He et al., 2016) pre-trained on ImageNet. For image captioning and also for models that use the pre-trained captioning model (i.e. reward finetuning, PoE and noisy channel) we back-propagate gradients into the ResNet module. However, in all rerankers we freeze the ResNet during reward optimization. Moreover, we also keep the ResNet fixed in the jointly learned listener to prevent additional drift, however we back-propagate when we pre-train the fixed listener, grounded though the discriminative caption speaker.