Structural and functional learning for learning language use

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

Abstract

One of the most ambitious goals of AI is to develop agents that are able to communicate with humans. While many existing systems are already capable of producing human-like utterances, they often focus on learning *structural* properties of language and miss the utilitarian and *functional* aspects of communication, i.e., that humans use words to coordinate with others and make things happen in the world. In this work, we investigate if and how we could use the *multi-agent interactions* (between an agent and a user simulator) as a building component for learning natural language use, and how to harness the structural knowledge of language, that is easily extractable from large collections of texts using *language models*.

One of the most ambitious goals of AI is to develop intelligent agents that are able to communicate with humans. Thus, communication and interaction should be at the core of the language learning process of these agents. However, traditional machine learning approaches to language learning [14, 17, 18] are dissociated from communication but are based on static, passive, and mainly supervised (or self-supervised) regimes, focusing on learning from corpora about the *structural* properties of language. While this is a great way to learn general statistical single-modality associations between symbols (e.g., the fact that adjectives come before nouns and after determiners) or even multi-modal associations between symbols and things in the world (e.g, the fact that the word cat refers to the furry animal with the four legs) it misses the functional aspects of communication, i.e., that humans use words to coordinate with others and make things happen in the real world [1, 3, 20].

One way to add communication in the core learning of agents is to cast functional language learning (i.e., learning to communicate grounded in a goal) as a supervised learning task and collect language data grounded to a particular goal. However this would require us to collect data of all potential language usages that we would want our agent to be able to communicate about. Motivated by this, previous research [12, 11] has focused on ways to emerge a communication protocol in a completely utilitarian framework implemented within a multi-agent setup where agents learn to communicate in order to maximize a task reward. While this purely utilitarian framework results in agents that successfully learn to solve the task by creating a communication protocol, these emergent communication protocols bear (at best) very little resemblance to natural language and pose doubts to the use of this type of functional learning as a viable alternative to language learning.

Thus, it becomes clear that neither framework on its own is completely adequate for learning language use. Instead, in this work we 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, for us at least, the essence of communication that underlies our intentions. The "what" is chosen by maximizing a given utility, which can be anything, 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 to communicate this "what" successfully. Since our goal is to communicate with humans, there are particular constraints that govern the form of "how" so that it is understandable by humans, i.e.,

[*] Shared first co-authorship.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

structural properties of natural language that relate, among others, to grammaticality and fluency. This factorization into *content planning* (here, “what”) and surface realization (here, “how”), which can lead to meaning representations which are amenable to reinforcement learning, moves away from end-to-end neural generation system and is inline with more traditional views of natural language generation [16].

Under this factorization, generic language data do not have to be used as gold-standard of functional language learning (which, as we explained above is problematic) but can be used effectively as a good prior model of language, encapsulating all the intrinsic *structural* knowledge of language. In other words, language data are only used for the “how”. On the other hand, multi-agent interactions that provide task-rewards for the task of interest, can now be used only for the *functional* learning of the language use. This combination of functional and structural learning guarantees that, in theory, the emergent communication of agents arising from multi-agent interactions will be grounded in natural language semantics, bringing us closer to learning *natural language*.

In this work, we present preliminary results of implementing this factorization of language use into “what” and ”how” and effective ways to combine functional (i.e., learning in the context of communicating with another agent so as to achieve a particular goal) and structural (i.e., traditional supervised learning of language) language learning.

1 Research framing

Our research can be framed in the following general scenario: an agent needs to perform a **functional communication task** in natural language (in this work, we are considering only English). However, we do not have examples of linguistic communication in natural language about this functional task. Framing the task into a multi-agent language game gives a way to obtain a **reward** that judges whether an utterance elicited the correct behaviour by a listener. We also have examples of **generic natural language**, that however are not grounded in the aforementioned functional task.

1.1 Experimental setup

In the first set of experiments, we looked into the following instantiation of the research. **Functional task**: visual referential communication game for a target image in the context of a distractor image. **Reward**: success in referential communication where a listener needs to identify the correct image within a set of distractors guided by the speaker’s description. **Generic natural language**: captioning data.

Visual referential communication game. There are two players, the speaker and the listener. The speaker is presented with two objects represented as images, a target and a distractor. The listener is presented with the same objects, however without knowledge of which object is the intended target. The listener needs to identify the target image from the distractors relying on an utterance being communicated by the speaker. The utterance takes the form of sequences of word-like units. If the listener is correct in identifying the target, they both receive a positive reward, else they receive the same negative reward.

Datasets. In our experiments we use two visual datasets, MSCOCO [13] (real images, Figure 1a) and Abstract Scenes [21] (synthetic images, Figure 1b). Both datasets are accompanied with captioning data that describe the images. Moreover, for evaluation purposes only, we introduce two different splits (i.e., *easy* and *hard*) that control for the difficulty of the discrimination task as a function of the semantic similarity of target and distractor.

1.2 Methods for learning language use

1.2.1 Speaker

The speaker model is the primary learner in this research, which aims at creating a model that is able to use natural language in a communicative scenario. The speaker is constructed with standard modules. For visual processing, we use a pre-trained ResNet [9] which extracts features from images using the last layer. For generating a message, we initialize a one-layer LSTM [10] with the ResNet-extracted features of the target image.

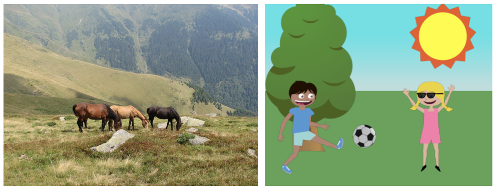

Figure 1: Images from the two visual datasets used in this study. **a**: MSCOCO. **b**: Abstract Scenes.

We now discuss several ways of learning language use and updating the weights of the language component of the speaker (i.e., its LSTM), including **functional-only learning**, **structural-only learning** as well as ways to combine the best of both worlds (**structural + functional learning**).

Functional-only learning. As we do not have language instances of this communication task, the speaker learns to emit communication utterances in order to maximize the communication task reward end-to-end (see Section 1.2.2 for a discussion on how this reward is computed). This type of learning of language use is identical to experiments commonly conducted in the literature of **emergent communication** [12, 8, 2, 5, 7]. Concretely, the weights of the speaker's LSTM are being updated via the REINFORCE [19] update rule (we assume the actions of the speaker are the words they emit). Note, that while this type of learning will result in a language that is maximally functionally correct for the given task reward, this is not natural language, i.e., the symbols are not grounded to natural language and have emergent semantics.

Structural-only learning. An alternative is to learn language use by altogether ignoring the functional aspect of communication and just learning to communicate utterances that reflect intrinsic structural properties of language, i.e., utterances that are fluent and grammatical. In our task, this type of structural learning takes the form of **image captioning**, thus the speaker's LSTM weights are updated in order to minimize cross-entropy on the captioning data. While a system trained on this supervised task can learn to describe images in a fluent and grammatical way, it is not clear that this system will also be able to correctly use these language skills in another language situation governed by a different functionality, in this case the visual referential communication game. Moreover, we also design a speaker that has access to **gold captions** of images at test time and uses them directly. Performance of these speakers will indicate to what degree having good language skills is adequate for functional communication task.

Structural + functional learning. Here, we describe ways in which both types of learning are used to learn language use. The simplest perhaps, is to *first* learn about the statistical properties of language from canned corpora (in our case, pairs of images and captions). While this knowledge of language is dissociated from the communicative function of the task, we can *then* do fine-tuning using the task reward to steer the language use to be functionally appropriate. We will refer to this speaker in the Results section as **image captioning + reward fine-tuning**.

Another alternative is to conduct both types of learning at the same time, i.e., to use **multi-task**. Here, image captioning will be teaching the speaker about statistical properties of language and associations of symbols, while functional learning will be optimizing for reward. Crucially, these two objectives are optimized simultaneously with a weighted loss.

We note that in both of these types of combined learning, the functional learning is interacting with the structural learning, i.e., the gradients from optimizing the functional task are back-propagated all the way into the LSTM language model of the speaker. This might have a negative impact on the *core knowledge* of language and its properties, leading to *language drift*.

Motivated by this, we introduce a third way of learning language use. We start by training the core language capabilities of a speaker, i.e., the image-conditional language model, on structural language learning, i.e., the image captioning task. This gives the speaker general knowledge of language grounded in images. The functional task learning is instead viewed as *learning to use* the existing knowledge. Concretely, structural learning is performed first to learn an image-conditional language model by optimizing cross-entropy. Following that, the weights of the LSTM are frozen. The functional learning task is implemented as learning to rerank samples obtained from the image-conditional language model. The weights of the reranker are being updated using the REINFORCE rule in order to optimize the task reward (i.e., communication success in the reference game). The reranker may be given an additional loss term proportional to the log probability assigned to the sampled utterances by language model, scaled by a weight λ.

Unlike the existing emergent communication setups, in this setup we assume the actions of the speaker are instead whole utterances and reinforcement learning is conducted on the utterance level rather than on the word level. This means that learning to use the language in the functional task is not going to affect the core language learning capabilities, e.g., by back-propagating through the core language component. We will refer to this speaker in the Results section as **image captioning + reranking**.

1.2.2 Listener

Throughout all the experiments, we need a way to estimate performance on the functional communication task, either for training or evaluation purposes. Ideally, this performance signal should be provided by a human who is interacting with our speaker agent online. However, we start by approximating this quantity with a *learned* component, an agent listener. Since, we always know which of the image candidates is the intended referent, we can treat this problem as an instance of supervised learning. The listener, similarly to the speakers, uses a pre-trained ResNet which converts features from all images (i.e., the target and the distractor) using the last layer. Following that, the listener uses an LSTM to embed the utterance from the speaker. Finally, the listener picks the image with the highest dot-product similarity between the embedded message and the features of the images. The weights of the modules are trained to map a communication utterance to the correct image target by optimizing cross-entropy. The listener assigns reward 1 to the speaker if they identified the correct image, else the listener assigns reward -1. Alternatively, the listener can also provide its negative cross-entropy loss as a reward: the higher the probability assigned to the correct image by the listener, the better the speaker has done. In the experiments reported here, the listener is trained jointly with the speaker.

2 Results and Discussion

Speaker type	Learning		MSCOCO		Abstract Scenes		Natural
	Functional	Structural	easy	hard	easy	hard	Language
gold captions	-	+	0.97	0.59	0.84	0.72	Yes
image captioning	-	+	0.99	0.75	0.91	0.84	Yes
emergent communication	+	-	**0.99**	**0.98**	**0.99**	**0.98**	**No**
image captioning + reward fine-tuning	+	+	0.95	0.75	0.89	0.78	Drifted
multi-task	+	+	0.99	0.79	0.92	0.82	Drifted
image captioning + reranking ($\lambda = 1$)	+	+	0.98	0.78	0.93	0.88	Yes
image captioning + reranking ($\lambda = 0.9$)	+	+	0.99	0.80	0.96	0.92	Yes
image captioning + reranking ($\lambda = 0$)	+	+	**0.99**	**0.87**	**0.97**	**0.95**	**Yes**

Table 1: Ratio of successful communications on held-out data for games with 2 distractors.

Results. In Table 1 we present preliminary results of this research. First, we observe that using **gold captions** verbatim for referential functional communication is sub-optimal, confirming the hypothesis that for successful language use we need to be aware of the particular functional goal and *adapt* for it. With **image-captioning** approach, the listener is trained on stochastic samples from the conditional language model, as opposed to the fixed gold captions, which allows it to perform at higher, but still sub-optimal, accuracy scores. As expected, the best results for functional communication are obtained when optimized for it using **emergent communication**. However, this type of learning results in a speaker who is not communicating in natural language, as indicated by the last column of Table 1, i.e., the speaker's communication utterances are incomprehensible.

In hybrid **multi-task** and **reward fine-tuning** scenarios, communication is kept close to natural language by the language model loss, however, both of them suffer from *language drift*, happening through the back-propagation of gradients into the core language component. We observe *structural drift* resulting in less fluent utterances, and *semantic drift*, where concepts can obtain different names, allowing the listener to bias communication channel to the game needs. Finally, we observe that the **reranking** methods that use a language model as a proposal model and learn to rerank its samples have English-looking utterances and achieve good performance. When setting the weighting coefficient $\lambda = 0$, the re-ranking is guided purely by the listener reward, resulting in the increase in the game accuracy scores. Overall, the proposed ways of combined functional and structural learning outperform the pure structural ones, indicating that goal-oriented language learning is beneficial for learning language use.

Discussion. We believe that combining structural language learning in the form of language modeling and functional learning in the form of multi-agent interactions is an exciting new avenue for (semi-supervised) learning of language use. The reward reranking model seems to be a method that combines many desirable properties, i.e., communicating sampled directly from a pre-trained language model, introducing rich conditioning indirectly in the language model, and using the reward to search for the more appropriate sample. While we do not observe *structural* drift (i.e., the output of the reranker is english-looking) it is challenging to prevent *semantic* drift (i.e., the output of the reranker can have low adequacy by referring to, say, cats as dogs, especially by setting low λ values and taking large number of samples). Quantifying and controlling these types of drift would allow for further improvements in the proposed approaches of learning language use.

Finally, in the near future we would like to consolidate ideas from *pragmatics*, a field of research that, just like us, puts the listener's behaviour at the heart of communication and has attracted attention both uni-modal [15] and multi-modal NLP [6, 4].

Acknowledgements

We would like to thanks the anonymous reviewers for the comments, as well as Kris Cao and Chris Dyer for valuable feedback.

References

[1] John Langshaw Austin. *How to do things with words*. Oxford university press, 1975.

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

[3] Herbert H Clark. *Using language*. Cambridge university press, 1996.

[4] Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 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, 2018.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[20] Ludwig Wittgenstein. *Philosophical investigations*. John Wiley & Sons, 2009.

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