Compositionality Exploration

Ofir Shechter
ofir1847@gmail.com

Shahaf Goren
shahafgoren@mail.tau.ac.il

1 Introduction

Natural language allows us to refer to composite concepts by combining expressions denoting their parts according to systematic rules, a property known as compositionality. Linguists studying this trait use their knowledge of the word meanings and composition rules to study compositionality, these assumptions do not hold when analyzing emergent languages and therefore other methods must be used.

Assessment of compositionality in languages has received significant attention in linguistics and adjacent fields. Languages are often studied using computer science techniques however the compositionality aspect sometimes stays behind. One way to use computer science to study natural languages is to use emergent languages that arise in communications between simulated agents. It is important to analyze how compositional is this formed language to conclude meaningful realizations about natural languages since they are compositional. An interesting aspect is also the compositionality of embedding vectors that are commonly used in machine learning algorithms, but in this article, we will mainly focus on emerging languages consisting of bit-vector messages. Another interesting information compositionality gives us is the ability to understand if the success of the cooperative paradigm comes from strategies akin to human-language compositionality.

2 Related Work

2.1 Emergent languages for language research

Certain aspects of linguistics relating to the evolution of languages are extremely hard to properly investigate with naturally evolved languages. This is especially true when desiring an experimental approach. We can look at the study of pidgin languages as an example. Pidgin languages are a grammatically simplified means of communication that develop in the contact between two or more groups that do not have a language in common. Pidgin language research is complicated and mostly based on observational work for long periods due to the slow nature of language development. In (Master, Schumann, and Sokolik 1989) an attempt to experimentally create a pidgin was done but this is a very hard and expensive process. A modern approach to study the evolution of languages involve computational modeling of communications between agents or communities of agents cooperating to get some sort of reward from the environment. This allows for rapid and careful experimentation with the caveat of dealing with languages that are mostly confined to the task at hand and that are to be analyzed only with analytical tools since they are not humanly understandable. This method was used in many works such as (Hurford 1989) and (Lewis 2008) and recently saw even more research with the advances in deep learning (Lazaridou, Peysakhovich, and Baroni 2016); (Havrylov and Titov 2017). In the work of (Graesser, Cho, and Kiela 2019), such a framework was used to study the creation of pidgins by simulating interactions between communities of different sizes speaking different languages. We elaborate on their work by using their framework to study the linguistic aspect of Compositionality

2.2 Compositionality Measurement

The principle of compositionality is normally taken to quantify over expressions of some particular language L: For every complex expression $e \in L$, the meaning of e (in the language L) is determined by the structure of e and the meanings of the constituents of e in the language. Gottlob Frege is widely credited for the first modern formulation of compositionality, and this principle is also called Frege's principle. His primacy claim

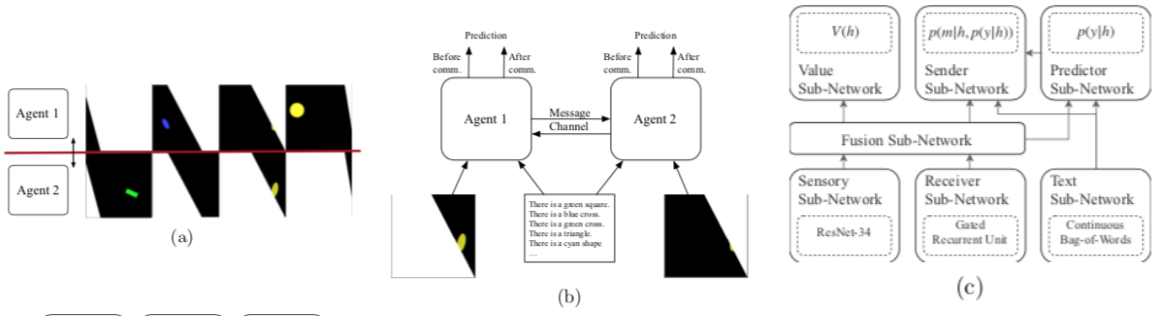

Figure 1: (a) Example training data. Only a random half of each image (dark background) is presented to one agent, necessitating communication in order to solve the game. (b) A graphical illustration of the proposed game. Each of two agents observes a partition of an input image and decides which of ten textual captions best describes the entire image before and after exchanging messages with the other agent. (c) The internal structure of an agent. The structure is modular in that each sub-network could be replaced by an alternative without requiring any change in other parts of the proposed framework.

was formulated to a determination claim, similar to the above (The meaning of an expression is determined by the meanings of all complex expressions in which it occurs as a constituent) (Szabó 2004).

Jacob Andreas developed a procedure called Tree Rconstruction Error (TRE) (Andreas 2019) to evaluate compositionality by measuring how well the true representation can be approximated by a model that explicitly composes a collection of inferred representational primitives - an empirical characterization of compositional structure. We were curious about the effect of word-space on the compositionality of the emerging language.

3 Method

3.1 Language Emergence Framework

We used the framework and modified code from (Graesser, Cho, and Kiela 2019). In the following section, we will describe it in detail.

At each step t a partially observable cooperative game is played between two agents a_1^t, a_2^t sampled from an agent pool. The agents receive different parts I_1^t, I_2^t of an image I^t (Figure 1) containing a colored shape (7 colors, 8 shapes) together with 10 sentences of scene descriptions, e.g. "There is a red square". The sentences are represented by their GloVe embedding. Only one sentence y^* is a correct scene description and the goal of the agents is to select the correct one. Since the image might be split in a way that one agent sees the whole shape and the other sees nothing, they are allowed to exchange a message before the prediction stage.

The agents are represented as Deep neural networks composed of several modules (Figure 1. All agents share the same architecture but learn different weights while playing the games. At the beginning of the game, each agent makes an initial prediction $y^t_{1,before}, y^t_{2,before}$ followed by an exchange of messages (1 per agent) and a final prediction $y^t_{1,after}, y^t_{2,after}$. Then both receive a reward equal to the gains from communicating (same for both). The gain of agent 1 from communication at step t is defined as

$$r^t_{1,comm} = \mathbb{1}_{y^*=y^t_{1,after}} - \mathbb{1}_{y^*=y^t_{1,before}}$$

And the final reward given is

$$r^t = r^t_{1,comm} + r^t_{2,comm}$$

This reward is designed to encourage cooperation between the agents and was empirically validated in the original paper.

The agents are trained in a mix of supervised and reinforcement learning. The supervised part uses the fact that we know y^* and therefore can back-propagate the log-loss over the agents' prediction before and after communication $y^t_{i,before}$ and $y^t_{i,after}$ through the net. Since the message generation process is discrete we cannot back-propagate the loss through it and so we use the REINFORCE algorithm (Williams 1992) to maximize $E_m[r^t]$ where m is the generated message. A precise description of all loss functions used can be found in the original paper.

3.2 Tree Reconstruction Error (TRE)

Representations: A representation learning problem is defined by a data-set I of observations $I_t \in I$ (the images the agents sees) a space Θ of representations $m \in \Theta$ (8-bit binary vectors) and a model $f : I \rightarrow \Theta$ (i.e the representation).

Derivations: We assume that inputs can be labeled with tree-structured derivations d, defined by a finite set D_0 of primitives and a binary bracketing $\langle \cdot, \cdot \rangle$, such that if d_i and d_j are derivations, $\langle d_i, d_j \rangle$ is a derivation. Derivations are produced by a derivation oracle $D : I \rightarrow D$. In our case these are the tuples that describe the Images (e.g $\langle$ BLUE, SQUARE $\rangle$).

Compositionality: A model f is compositional if it is a homomorphism from derivations to representations. We require that for any $I_t, I_{t,a}, I_{t,b} \in I$ with $D(I_t) = \langle D(I_{t,a}), D(I_{t,b}) \rangle$ the following holds: $f(I_t) = f(I_{t,a}) * f(I_{t,b})$. $I_{t,a}$ is the image which is described by $D(I_{t,a})$

TRE framework:

- Define distance function δ as cosine distance - $\delta(\theta, \theta') = 1 - \theta^T \theta' / (\|\theta\| \|\theta'\|)$
- Define a composition function $* : \Theta \times \Theta \rightarrow \Theta$ as the addition function.
- Use RMSprop optimizer and a simple neural network in order to learn $\hat{f}$, a compositional approximation of f.
- Compute the TRE value for each sample: $TRE(x) = \delta(f(x), \hat{f}(D(x)))$
- Approximate the expectation - $TRE(X) = \frac{1}{n} \sum_i TRE(x_i)$

For convenience we will mark the return value as tre. Note that $tre \in [0, 1]$ while lower value implies higher compositionality.

3.3 compositionality via tree classifier

TRE was validated with long floating point vectors of learned word embeddings. This is quite different than the context we would like to use it in - short binary vectors. Therefor we thought of a way to validate and further investigate the compositionality of the emergent language. to do so we performed some manual exploration and came up with another compositionality test suitable for binary languages.

We try to predict the shape and color of the image from a given message. We predict only for messages that led to a successful round and to messages sent by agents that saw some part of the shape (had an informative part of the image). By restricting the hypothesis class we can learn about the underlying language. For example, if we train a tree classifier for property A (for example shape) and limit the number of leaves to the number of values n_A in the category then we force the prediction to learn only on the most robust bit combination describing each class or ignore a whole class in favor of representing a class in more than one leaf. Looking at the feature importance's can teach us which bits are important for the prediction of the whole category. If the language is completely non-compositional then with the proposed limitation we can expect poor performance as we do not allow for enough complexity to describe every shape and color combination and a we expect the feature importance to be roughly uniformly distributed over bits. On the other hand, if it is perfectly compositional we can expect good performance and the importance should be concentrated over a subset of bits describing this category only.

4 Results

4.1 Compositionality testing

We started by testing the language of a pool of agents communicating by words of length 8. The pool was trained for 800K steps. The language corpus was generated by evaluating all agents as first speakers and as second speakers. Evaluation was preformed over the whole data set so some shape and color combinations were new. Average success on the unseen combinations was 75%. The successful generalization hints towards the compositionality of the language but it is not conclusive.

Running the TRE procedure resulted in a low TRE value of 0.145. In the original paper Andreas 2019 the TRE values were computed over English word embeddings and found to be similar to this value (between 0.1-0.2). The authors conclusion was that the embeddings are indeed compositional, so we also conclude that the emergent language is compositional.

We performed a further manual analysis of the corpus as described above. We fit a tree classifier for the shape and color property without any limitation and found that it is performing quite well with a macro f1 score (averaged over classes) of 0.71. When inspecting the feature importance's we see that most of the weight is placed over 4 bits (3,5,6,7 take 87% of the importance). When limiting the classifier to only 8 leaves (one per value)

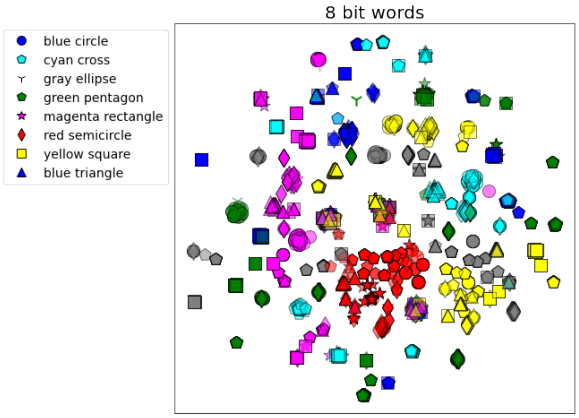

Figure 2: scatter plot of all messages after TSNE (Van der Maaten and Hinton 2008) dimensionality reduction

we get the same macro f1 score but now the same bits take all the weight. When repeating this analysis with the color property we get the same results (macro f1 of 0.85 for both) but with the 4 unused bits (0,1,2,4). These results suggest that the language is highly compositional, yet not perfectly as the messages used were informative and led to successful rounds but still we did not get perfect predictions.

We performed some extra analysis to better understand the language structure. Two aspects were of interest to us: (1) can we predict if an agent saw an informative image or not? (2) how is the message effected when the correct description has no specific shape (e.g. "there is a yellow shape"). Predicting both labels failed with Area Under Curve (AUC) scores of 0.55 for both. From that, we conclude that the agents are not signaling differently to each other when they need more assistance and that the message stays the same when the correct description has no specific shape in it. Another thing we notice is that visually close shapes and colors were closer in their learned representation as can be seen in figure 3.

4.2 word length effect over compositionality

The absolute minimum required for this task is 3 bits for each attribute. Therefore it is natural to ask what language would emerge if we use 6-bit messages. we hypothesized that the emergent language will be more compositional out of necessity to compress. To check this claim we trained a pool of agents with message lengths of 6 and 10 for 400k and 300k steps respectively and repeated the previous tests. we did not reach convergence

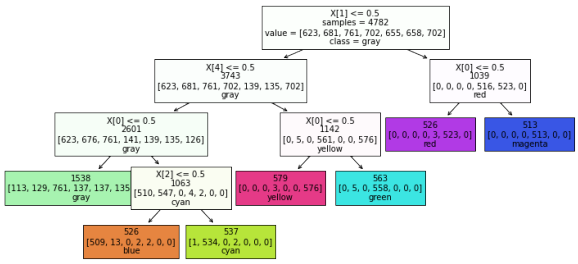

Figure 3: Classification tree for colors. X[i] signifies split over bit i close colors happen to be close together also in the learned representation

(75% success on new shape color combinations) due to the long training time required but we can still draw some preliminary conclusions.

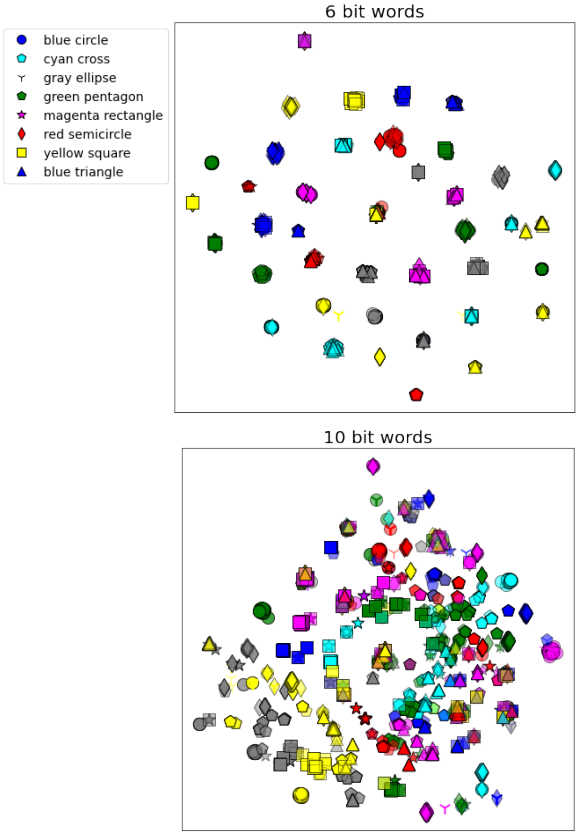

Figure 4: scatter plots of all messages of length 6 and 10 after TSNE dimensionality reduction

TRE values results for these languages were surprising. For a language with 6-bit messages we got $tre_6 = 0.1375$ and for 10-bit messages $tre_{10} = 0.1129$. The results were different from our hypothesis. To validate that the final TRE score did not come from some extreme distance distribution we plotted the histograms of the distances between the learned $\hat{f}(D(I_t))$ which is

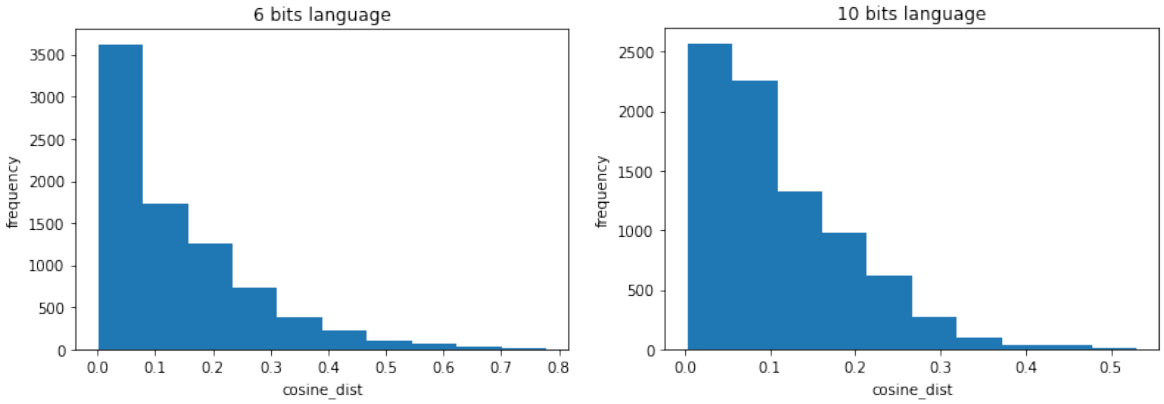

Figure 5: Histograms of the cosine distance between the learned $\hat{f}$ (100% compositional) in TRE procedure to the real f (the messages)

	6 bit		10 bit	
	shape	color	shape	color
unrestricted classifier	0.41	0.85	0.65	0.71
restricted classifier	0.3	0.85	0.55	0.72

Table 1: macro f1 scores for tree classifiers with restricted number of leaves and without restrictions for each classification task and language

100% compositional to the real $f(I_t)$ (the agents probabilities over bits given an image) and indeed we see that the distances for the 10 bit language are much smaller. Meaning the learned language is very close to compositional As can be seen in the histograms (Figure 5). we find the effect of higher dimension exciting and non-intuitive.

Similarly to the analysis of the 8 bit language we fitted two tree classifiers to predict each attribute (shape and color). For a compositional language we expect that each value of an attribute (e.g. color) would have a stable combination of bits describing it independently from the other attribute (e.g. shape). That way a limited classifier with one leaf per category should perform as good as a classifier with no limitations. We found that the relative difference in f1 scores was smaller for the 10 bit language than the 6 bit language (+18% vs +33%) (see Table 1) indicating that indeed the 10 bit language is more compositional, like we saw with the TRE scores.

5 Discussion

Compositionality understanding is important, it can help us learn about the similarities between simulated evolved languages and naturally evolved languages and understand the way machines learn to communicate (similar or different to human-like communication).

We used an advanced RL-task to create an emergent language with different dimension space, a state of the art compositionality assessment tool (TRE), and some other machine-learning tools for further investigation and got some exciting results.

Our work also validates further the TRE method since the TRE values correlated with more straightforward analysis that implied compositionality as well. We also noticed some relevant insights that cannot be seen in the single value the TRE procedure returns such as the relationship between visually similar shapes/colors and close learned representations of the attributes. So for further compositionality research and computation, we highly recommend performing more specific analysis to avoid missing some interesting results.

The most interesting result we got was the counter intuitive finding that higher dimension leads to higher compositionality. Humans' natural languages have very high dimensionality and yet these languages are characterized as compositional languages. More research is required in order to verify our results, potential directions could be to check if our findings hold after the languages converge and to see if the higher compositionality effect is further increased by taking the message dimension to extreme. Though Further research is required in order understand the compositionality of emergent languages and why it evolves we feel that we made a decent contribution towards understanding it.

6 Acknowledgments

we had a great time working on the research project. We had to face academic papers from the frontier of current knowledge, use a complicated codebase and change it according to the requirements of our research and get a good understanding of new fields (RL tasks and compositionality). We hope you enjoy reading about our work as much as we enjoyed making it.

*all relevant code can be found in:
`https://github.com/OfirShechter/NLPMultimodalGame`
`https://github.com/OfirShechter/tre`

References

[And19] Jacob Andreas. "Measuring compositionality in representation learning". In: *arXiv preprint arXiv:1902.07181* (2019).

[GCK19] Laura Graesser, Kyunghyun Cho, and Douwe Kiela. "Emergent linguistic phenomena in multi-agent communication games". In: *arXiv preprint arXiv:1901.08706* (2019).

[HT17] Serhii Havrylov and Ivan Titov. "Emergence of language with multi-agent games: Learning to communicate with sequences of symbols". In: *arXiv preprint arXiv:1705.11192* (2017).

[Hur89] James R Hurford. "Biological evolution of the Saussurean sign as a component of the language acquisition device". In: *Lingua* 77.2 (1989), pp. 187–222.

[Lew08] David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 2008.

[LPB16] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. "Multi-agent cooperation and the emergence of (natural) language". In: *arXiv preprint arXiv:1612.07182* (2016).

[MSS89] Peter Master, John H Schumann, and Margaret E Sokolik. "The experimental creation of a pidgin language". In: *Journal of Pidgin and Creole Languages* 4.1 (1989), pp. 37–63.

[Sza04] Zoltán Gendler Szabó. "Compositionality". In: (2004).

[VH08] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[Wil92] Ronald J Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine learning* 8.3-4 (1992), pp. 229–256.