The Relationship Between Compositional Generalization and Misinformation in Emergent Communication

Heeyoung Lee
Sungkyunkwan University
Suwon, South Korea
hy18284@g.skku.edu

Abstract

Social interactions involve adversarial nature as well as cooperative aspects. In communications, a sender's message may undergo adversarially motivated actors before reaching the intended receiver. These actors may modify the message with malign intents while preserving the overall characteristics of the message. This form of misinformation is prevalent in real-world communications. This work investigates how misinformation affects language emergence. We find that risks of malign misrepresentation induce more generalizable languages in simulations of communicative agents[1].

1 Introduction

The field of emergent communication studies how languages evolve over time. The recent developments in artificial neural nets allow simulations of communicative agents, and the field has produced numerous findings on the dynamics of human language evolution through these simulations (Lazaridou et al., 2017). Of these, emergence of compositionality is one of the most studied aspects in the field. Compositional languages allow expression of novel concepts by combinations of already familiar attributes with utilization of a rule structure. Numerous studies explore the core environmental factors in emergence of compositionality (Li and Bowling, 2019; Chaabouni et al., 2020).

This work investigates how misinformation affects compositionality of languages. Social interactions take complex forms. A sender's message may be relayed by another agent before arriving at the intended recipient. The integrity of the message may be compromised by this relay agent. This agent could modify the message with malign intents while preserving the overall appearance of the message. Misinformation by malign actors are commonly observed in real-world communications, yet its effects on language emergence remain to be explored.

We introduce an adversarially motivated message relay agent to the communication game. We find that the added risk of malign misrepresentation induces more compositional and generalizable languages when compared to undirected random noisy channels (Kuciński et al., 2021).

2 Communication game under misinformation

We investigate the effects of misinformation in language emergence with agents playing a variant of Lewis reconstruction game (Lewis, 1969). The sender π_ϕ observes an object $x \in \mathcal{X}^N$. An object is characterized by N attributes and each attribute can take one of $|\mathcal{X}|$ values. The sender outputs a message $m \sim \pi_\phi(\cdot \mid x)$ describing the object x and sends the message to the receiver π_θ. A message is consisted of a fixed length of L symbols, and each symbol belongs to the vocabulary $\mathcal{V}$. With probability p, the receiver π_θ receives the original message and outputs its prediction for the object x as $\hat{x} \sim \pi_\theta(\cdot \mid m)$. With probability $1-p$, the message is sent to the adversarial relay agent π_ψ. The adversarial relay agent observes the message m and the object x and modifies $k \leq L$ symbols within the message with the intent of inducing an incorrect prediction from the receiver π_θ as $m' \sim \pi_\psi(\cdot \mid m, x)$. The modified message is relayed to the receiver and the receiver outputs its prediction for the object x from the modified message m' as $\hat{x} \sim \pi_\theta(\cdot \mid m')$. The sender and receiver are rewarded if the receiver outputs the correct prediction for the object x either from m or m'. The adversarial relay agent is rewarded when the receiver outputs an incorrect prediction of the object x from the modified message m'.

[1] A preliminary version of this work is to be presented at EDAI 2024 workshop in October. This version incorporates new results on generalization ability.

3 Experimental setup

Game setup We set the probability that an original message is directly passed to the receiver, p, to 0.5. We vary the value of k, the number of symbols modified by the adversarial relay agent, from 1 to 4 and observe its effect on the language structure. The number of modified symbols are kept to exactly k by reverting back (or modifying) randomly chosen modified (or unmodified) symbols if the number of modified symbols do not match k in the modified message. We compare directed misinformation of adversarial relay agent with undirected random noise, where randomly chosen k symbols are modified to take different symbols each with probability $\frac{1}{|\mathcal{V}|-1}$. Clean communication channel setup is also compared, where no modifications are made to the messages. The message length, T, and the vocabulary size, $|\mathcal{V}|$, are both set to 10. All agents are single-layer GRUs (Cho et al., 2014) of size 128. See Appendix A for full description.

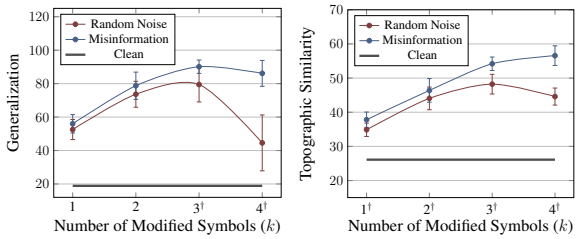
Dataset We utilize simple attribute-value dataset where each attribute is represented with one-hot encoding. The number of attributes, N, is set to 5, and the number of values an attribute can take, $|\mathcal{X}|$, is set to 10. We set aside 95% of the attribute-value combinations as the test set and use the rest as the train set.

Optimization The receiver is optimized with average cross entropy loss from the prediction distributions of attributes. This is written as $-\frac{1}{N}\sum_{n=1}^{N}\log\pi_\theta(x_n \mid m)$, where x_n denotes the ground truth value of the n-th attribute. This loss is directly backpropagated to the receiver. The sender and adversarial relay agent are optimized with the REINFORCE algorithm (Williams, 1992). The above cross entropy loss from the receiver is used as the reward for the adversarial relay agent and negative of the cross entropy loss is used as the reward for the sender. We run 10 experiments with different random seeds and report the average results. See Appendix B for full description.

4 Evaluation metrics

Compositional generalization We measure languages' generalization ability with accuracy on the test set containing unseen attribute combinations.

Topographic similarity (TopSim) Topographic similarity (Brighton and Kirby, 2006) is Spearman's rank correlation of D_{obj} and D_{msg} over the joint uniform object, message distribution, where D_{obj}, D_{msg} refer to distance measures in the object and message spaces, respectively. In other words, $D_{\text{obj}} : \mathcal{X}^N \times \mathcal{X}^N \to \mathbb{R}^+$ and $D_{\text{msg}} : \mathcal{V}^L \times \mathcal{V}^L \to \mathbb{R}^+$. High topographic similarity indicates that messages that are similar to each other refer to objects are also similar to each other. We use cosine distance for D_{obj} and Levenshtein distance (Levenshtein, 1965) for D_{msg}.



(a) (b)

Figure 1: Comparison of language properties under varying degrees message corruption. Statistically significant differences from random noise are denoted with † (one-tailed t-test with $p < 0.05$). Error bars represent 95% confidence interval.

5 Experimental results

In Figure 1a, we observe that directed misinformation tends to induce more generalizable languages when compared to random channel noise. The differences get more pronounced as the number of modified symbols is increased. We also observe a similar trend in TopSim scores from Figure 1b suggesting that languages developed under risks of malign misrepresentation are more compositional. We hypothesize that adversarial relay agent's ability to more effectively exploit vulnerabilities in a language forces the language to be more compositional. We discuss the performance of the clean communication channel setup in Appendix C.

6 Conclusion

This work explores the effects of misinformation in emergent languages. We introduce an adversarial message relay agent that promotes misinformation. We find that directed attempts at misrepresentation with malign intents force the agents to develop more compositional languages compared to random symbol corruptions of messages. We provide a hypothesis that this is due to the adversarial relay agent's ability to exploit more obscure vulnerabilities in the language. This work presents a connection between the prevalent act of misrepresentation in social interactions and language emergence.

Acknowledgements

We thank the anonymous reviewers. Their comments helped improve the presentation and provided a deeper insight into the research.

References

Henry Brighton and Simon Kirby. 2006. Understanding Linguistic Evolution by Visualizing the Emergence of Topographic Mappings. *Artificial Life*, 12(2):229–242.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Online. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Ł ukasz Kuciński, Tomasz Korbak, PawełKoł odziej, and Piotr Mił oś. 2021. Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication. In *Advances in Neural Information Processing Systems*, volume 34, pages 23075–23088. Curran Associates, Inc.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

David Kellogg Lewis. 1969. *Convention: A Philosophical Study*. Wiley-Blackwell, Cambridge, MA, USA.

Fushan Li and Michael Bowling. 2019. Ease-of-teaching and language structure from emergent communication. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

A Agent architectures

Sender architecture The sender takes the object x and encodes it with a linear layer. This initializes a single-layer GRU (Cho et al., 2014) of size 128. The GRU layer recursively processes the object for a total of $L = 10$ steps, each time outputting the next symbol of the message from a probability distribution over the vocabulary $\mathcal{V}$ of size 10.

Receiver architecture The receiver recursively processes the message with a single-layer GRU of size 128. The last state is then passed through N linear layers, each of which corresponds to one of the N attributes and produces activations of size $|\mathcal{X}|$. Softmax activation function induces N distributions over the $|\mathcal{X}|$ values from these activations.

Adversarial relay agent architecture Two separate linear layers each process the object and the corresponding message. The two activations are summed in element-wise manner. This initializes a single-layer GRU of size 128 and it produces a modified message in the same manner as in the sender.

B Optimization details

Average negative entropy of the symbol distributions is utilized as an additional loss to encourage exploration for the sender and adversarial relay agent. The entropy term is multiplied by scaling hyperparameters of 0.5 and 2.0 for the sender and adversarial relay agent, respectively. The adversarial sender agent takes an additional loss corresponding to the deviations from the number of symbols to be modified, k. This is written as $|k - k'|$, where k' denotes the number of symbols modified by the adversarial relay agent. All agent's are optimized with ADAM (Kingma and Ba, 2017) optimizer with learning rate of 0.001 and $\beta_1 = 0.9$, $\beta_2 = 0.999$. We train the agents for 5,000 epochs with a batch size of 4096. We exclude a few runs that do not reach training accuracy of 95%.

C Generalization ability in the clean communication channel setup

In the clean communication channel setup, the agents achieve a moderate generalization ability in early stages of training. However, we observe that they strongly overfit to the train set as the training progresses. Decreasing tendencies in generalization ability are also observed in misinformation

and random noise setups, but their degree is not as pronounced, resulting in relatively high final test set accuracies.