

Characterization Capacity of Agents and Compositionality
from Naturally Emergent Communication

Hao Yifan

Contents

1	Abstract	2
2	Introduction	2
3	Related Work	3
4	Experimental Framework	4
4.1	Set up	4
4.2	Agent architecture	5
4.3	Training Algorithm	6
4.4	Evaluation	6
4.5	Compositionality and Capacity in Artificial Language teaching	7
5	Theoretical Analysis	9
5.1	the MSC model	9
5.2	analysis on the basis of mutual information theory	9
5.3	‘bilateral’ metrics for communication	11
6	Emergent Language Experiments	12
	Bibliography	15

1 Abstract

Recent advance on symbolic language in neural network based multi-agent systems have shown great progress in compositionality, which is taken as a distinguished feature of human language different from animal language. However, these efforts only explored environmental pressures, without realizing the importance of characterization capacity of agents.

In the work, we explore the relationship between the characterization capacity of agents and the compositionality of symbolic languages. By both proving with mutual information theory and verifying with extensive experiments, we made the counter-intuitive conclusion that lower characterization capacity facilitates the emergence of symbolic language with higher compositionality.

2 Introduction

The emergence and evolution of human language has always been an important and controversial issue. The problem covers many fields, including artificial intelligence in computer science. Computer scientists induce the emergence and evolution of languages in multi-agent systems by setting up pure communication scenarios, such as referential games and communication-action policies.

Researchers have confirmed that agents can master a symbolic language to complete appointed tasks. Such symbolic language is a communication protocol using symbols or characters to represent concepts.

people try to make the emergent language similar to human natural language.

Compositionality is a widely accepted metric used to measure the hierarchical complexity of language structure, and it is also a key feature to distinguish human language from animal language. Syntactic languages with high compositionality, such as human natural language, are able to express complex concepts through the combination of symbols and to produce certain syntax. In contrast, non-syntactic languages with low compositionality, such as animal languages, are almost impossible to extract specific concepts (i.e. attributes of objects) from a single symbol.

Researchers have found that various environmental pressures would affect compositionality. e.g. small vocabulary sizes, memoryless, carefully constructed distractors, ease-of-teaching.

但是, 他们都是研究环境对 compotionality 的影响. 我们发现了模型本身也对 compotionality 有影响.

Besides environmental pressures, we suggest that the impact of internal factors from agents themselves on compositionality is equally significant.

Many people believe that the cranial capacity of animals is not big enough to master languages with high compositionality. In neuron network based multi-agent systems, this hypothesis corresponds to a point of view that it' s difficult for agents with insufficient characterization capacity (i.e. number of neural nodes) to master languages with high compositionality.

However, we found that lower characterization capacity facilitates the emergence of symbolic language with higher compositionality, within the range afforded by the need for successful communication. We prove the point with mutual information theory and experiments.

From theoretical analysis, we define *bilaterality* as the quantitative metrics for compositionality. The bilaterality is the similarity between an identity matrix and the mutual information matrix of concepts and symbols (after normalization). We use the MSC (Markov Series Channel) to model the language transmission process and use the probability distribution of symbols and concepts to model policies of agents. Combining the MSC model with mutual information theory, we prove that (the complexity of the mutual information between original concepts received by the speaker and predicted concepts outputted by the listener is anti-correlated with the compositionality of the emergent language, which can be characterized by the definition of bilaterality.)

Then with experiments we show that a low-bilateral (i.e. low-compositionality) language needs higher capacity of the model to emerge. We build a listener-speaker referential game as experimental framework, and train agents with the correctness of forecast output from the listener as the only criterion. (The criterion does not imply any environmental pressures on the agents) 这句的描述还是不太准确, 体现 “不包含任何 environmental pressures” 的不是 correctness 这个 criterion 本身, 而是我们仅仅使用 correctness 训练 agents. Therefore, we can study the impact of capacity on the compositionality without any environmental pressures’ affection (因为前面还在提 environmental pressures, 这里突然没了有点突兀, 所以感觉是不是加个后缀算是给 environmental pressures 收个尾). Moreover, to study the impact of capacity on the compositionality under a more ‘natural’ environment, the speaker and listener are individual agents, i.e. disconnected models without sharing parameters (想给 individual 的定义加上一个 “模型不相连”, 这一点还是比较重要的, 如果允许相连就是个 auto-encoder 了, auto-encoder 里的编码不能称作 emergent language). The conclusion suggests that by restricting the number of neurons in a model the emerging languages attend to have higher bilaterality, thus higher compositionality.

To sum up, our contributions are as follows:

- a) We propose a novel metric, namely *bilaterality*, to quantitatively measure the compositionality of the emerging language.
- b) With experiments we found that the capacity of model is anti-correlated with the bilaterality, showing that restricting the number of neurons in a model attends to emerging a language with higher compositionality.

3 Related Work

一些工作是基于某个启发式的猜想, 提出某个 environmental pressure 对 compositionality 的影响. XXX 提出了 small vocabulary sizes; XXX 提出了 memoryless; XXX 提出了 carefully

constructed distractors; XXX 提出了 ease-of-teaching. 他们都忽略了一个源于模型本身的重要因素 characterization capacity.

不仅如此, 'naturally' emergent communication 也是一个值得关注的问题. 部分工作中使用了精心构造的 scenarios, models, reward, loss function. XXX 使用了 XXX……这些做法实质上等同于对 agents 施加了额外的人为诱导, 不仅削弱了 'naturally emergent compositional language' 的相关结论, 而且也分散/模糊/稀释了单个因素对 compositionality 的影响.

此外, 关于 metrics to measure communication 的争议也从未停止. 许多工作都提出了关于度量 compositionality and the degree of alignment between symbols and concepts 的 metrics. On the Pitfalls of Measuring Emergent Communication 这篇文章整理了近年来出现的 widely accepted metrics, 并将它们分为两类: those that measure positive signaling, 这类 metrics 是站在 speaker 的视角, 用于衡量 speaker 说出的 symbols 和接收的 concepts 之间关系, 例如 XXX; and those that measure positive listening, 这类 metrics 是站在 listener 的视角, 用于衡量 listener 收到的 symbols 和预测的 concepts 之间的关系, 例如 XXX. 总的来说, 这些 metrics 全都是 'unilateral' metrics, 但它们都缺少一个非常重要的 'bilateral' 特征: speaker 和 listener 的相互理解程度, i.e. 在 concepts 和 symbols 的对应上的一致性.

综上, 这些工作都无法回答这样一个问题: 在 'natural' 环境中, 模型的 characterization capacity 对 compositionality of emergent language 有怎样的影响? 这也是这篇文章要解决的问题. 我们结合理论分析以及实验结果, 并建立一种更合理的 'bilateral' metrics, so that we can quantificationally measure characterization capacity's impact on compositionality of emergent language.

4 Experimental Framework

我们在 referential game 中搭建实验框架. referential game 是一种 speaker 和 listener 通过交流达成合作的场景. 许多工作, 例如 XXX, 都使用 referential game 研究 emergent language. 下面, 我们分别介绍实验的 set up, agent 的模型结构, 训练算法和评估方法.

4.1 Set up

在我们使用的 referential game 中, 每次游戏都遵守如下基础规则:

- a) speaker agent S 根据 input object t 输出 symbol sequence s
- b) listener agent L 根据 symbol sequence s 输出 predict result \hat{t}
- c) 当 $t = \hat{t}$ 时, 认为 agents 在本次游戏成功, S 和 L 分别获得 reward $R(t, \hat{t}) = 1$. 否则 agents 失败, 并分别获得 reward $R(t, \hat{t}) = -1$

object t 由固定长度的 concept sequence (c_0, c_1) 组成, 记为 $t = (c_0, c_1)$. 其中 concept c_0 (shape) 和 c_1 (color) 分别有自己的取值集合 M_0 和 M_1 . 实验中, we let $|M_i| (i = 0, 1)$ range from 3 to 8. 我们用长度为 $|M_0|$ 的 one-hot vector 表示 shape c_0 , 用长度为 $|M_1|$ 的 one-hot vector 表示 color c_1 . 这两个 one-hot vector concatenate 成一个长度为 $|M_0| + |M_1|$ 的 vector, t 由该 vector 表示.

s 是固定长度的 symbol sequence (s_0, s_1) . 其中每个 symbol $s_i (i = 0, 1)$ 的取值都属于 vocabulary set V . 实验中, we let $|V|$ range from 3 to 10. 并且保证 $|v|^2 \geq |M_0| \times |M_1|$, 即保证 symbol sequence (s_0, s_1) 足够分别描述所有情况的 object t . 我们用两个长度为 $|V|$ 的 one-hot vector 分别表示 s_0 和 s_1 . 这两个 one-hot vector concatenate 成一个长度为 $2 \times |V|$ 的 vector, s 由该 vector 表示.

predict result \hat{t} 由一个长度为 $|M_0| \times |M_1|$ 的 one-hot vector 表示. 该 one-hot vector 中的每个 bit 对应一个 object, 即一个 shape 和 color 的组合, 记为 $\hat{t} = (\hat{c}_0, \hat{c}_1)$. 具体地, $\hat{t}[i \times |M_1| + j] = 1$ correspond to $\hat{c}_0[i] = 1$ and $\hat{c}_1[j] = 1 (i = 0, \dots, |M_0| - 1; j = 0, \dots, |M_1| - 1)$.

我们定义的 $\hat{t} = t$ 是指 t 和 \hat{t} 分别对应的 object 相同, i.e. 对应的 $(c_0, c_1) = (\hat{c}_0, \hat{c}_1)$.

4.2 Agent architecture

Agents 以各自强化学习的策略进行上述的 referential game. 将 speaker agent S 和 listener agent L 的 policy 分别记为 π_S 和 π_L . π_S 表示给定输入 object t , speaker 输出 symbol s_0 和 s_1 的条件概率 $P(s_0|t)$ 和 $P(s_1|t)$. speaker S 分别根据概率分布 $P(s_0|t)$ 和 $P(s_1|t)$ 随机采样输出 s_0 和 s_1 . π_L 表示给定输入 symbol sequence $s = (s_0, s_1)$, listener 输出 predict result \hat{t} 的条件概率 $P(\hat{t}|s_0, s_1)$. listener L 根据条件概率分布 $P(\hat{t}|s_0, s_1)$ 的随机采样输出 \hat{t} . Agent 分别用一个神经网络连接各自的 policy 的输入和输出. 模型的 architecture 如图 1 所示.

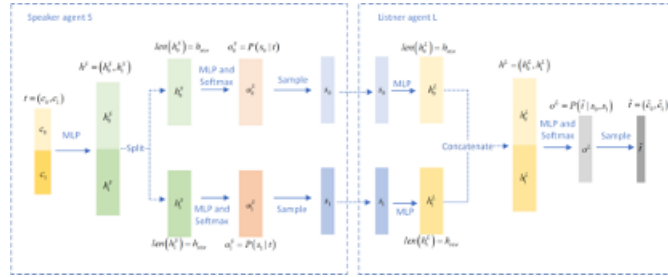


Figure 1: 模型的 architecture 示意图

对于 speaker 的神经网络模型, 输入 t 经过一个全连接层并激活得到 hidden layer h^S , h^S 的神经元节点数为 $h_{size} \times 2$. Splitting h^S equally 得到两个长度为 h_{size} 的 neural vectors h_0^S 和 $h_1^S (i = 0, 1)$ 各自依次经过一个全连接层和一个 softmax 得到 output layer o_i^S . $o_i^S (i = 0, 1)$ 是一个长度为 $|V|$ 的 vector, 其每个分量表示, 给定输入 t 时, S_i 的每个取值的概率, 即 $P(s_i|t)$.

对于 listener 的神经网络模型, 输入的 symbol sequence $s = (s_0, s_1)$ 中, $s_i (i = 0, 1)$ 各自经过一个全连接层并激活得到 hidden layer h_i^L , h_i^L 的神经元节点数也是 h_{size} . Concatenating h_0^L 和 h_1^L 得到长度为 $h_{size} \times 2$ 的 neural vector h^L . h^L 依次经过一个全连接层和一个 softmax 得到 output layer o^L . o^L 是一个长度为 $|M_0| \times |M_1|$ 的 vector, 其每个分量表示, 给定输入 symbol sequence $s = (s_0, s_1)$, \hat{t} 的每个取值的概率, 即 $P(\hat{t}|s_0, s_1)$.

在实验中, h_{size} 取一组离散的取值, 用于定量地表示 agents 模型的 capacity.

4.3 Training Algorithm

在我们的实验中, 我们使用 Stochastic Policy Gradient Methodology 单独训练 speaker agent S 和 listener agent L . 我们用 θ^S 和 θ^L 分别表示 speaker 和 listener 的 policy π^S 和 π^L 的全部参数. 训练 speaker 时, 固定 policy π^L 的参数 θ^L , 训练目标是调整参数 θ^S , 使其基于策略 π^S 获得的期望奖励 $J(\theta^S, \theta^L) = E_{\pi^S, \pi^L} [R(t, \hat{t})]$ 最大. 同时, 为了排除其他因素, 以及最小化人为诱导对 emergent language 的影响, 我们仅使用 listener 预测结果是否正确作为奖励, 分别对 listener agent L 和 speaker agent S 计算训练目标 $J(\theta^S, \theta^L)$ 的 gradients.

$$\nabla_{\theta^S} J = \mathbb{E}_{\pi^S, \pi^L} [R(\hat{t}, t) \cdot \nabla_{\theta^S} \log \pi^S(s_0, s_1 | t)] \quad (1)$$

$$\nabla_{\theta^L} J = \mathbb{E}_{\pi^S, \pi^L} [R(\hat{t}, t) \cdot \nabla_{\theta^L} \log \pi^L(\hat{t} | s_0, s_1)] \quad (2)$$

agents 的模型相互独立, 不共享任何模型参数也没有结构上的直接相连, 模型之间的联系仅为相互传递 symbol sequence $s = (s_0, s_1)$. 训练过程如图 2 所示. 训练过程中, 两个 agents 模型交替更新; 并且使用一个平行的神经网络保存 old parameters, 该网络定期将参数与实际输出的网络的参数同步, 从而限制 policy 的更新幅度, 使训练过程更加稳定.



Figure 2: agents 的 Training Algorithm 伪代码图

4.4 Evaluation

我们的目的是在保证模型收敛的前提下, 研究模型的 capacity 和 emergent language 的 compositionality 的关系. 当 Listener agent L 的正确性收敛到 100% 时, 我们认为模型收敛, 此时结束训练. 所以, 在完成一次训练后, 我们从 2 个方面对模型进行评估: 模型的 capacity; emergent language 的 compositionality.

Agents 的 capacity 可以由神经网络模型的隐层节点数 (i.e. h_{size}) 量化衡量. 对于 compositionality, 据我们所知, 目前并没有一个统一的度量标准. Topographic similarity¹ 是一个广泛接受的 compositionality 的度量^{3;4}. Topographic similarity 计算的是 symbol sequence 的 minimum edit distance 和 object 的差异度之间的 spearman correlation. In our case, symbol sequence $s = (s_0, s_1)$, object $t = (c_0, c_1)$, higher topographic similarity means similar objects have

		Speaker				
$c_1 \backslash c_0$		circle	square	Listener		
	red	(a, c)	(a, a)	$s_1 \backslash s_0$	a	b
blue		(b, c)	(b, b)	a	(square, red)	(circle, blue)
				b	(circle, red)	(square, blue)
				c	(circle, red)	(circle, blue)

(a)

(b)

Figure 3: 一个 language 反例

more similar symbol sequences in context. Compositionality and Generalization in Emergent Languages² 这篇文章指出 topographic similarity is agnostic about the type of similarity as long as it is captured by minimum edit distance. 并且提出了一个 metric posdis. 但是 posdis 和 topo 一样, 都只以 speaker 的 policy 为基础计算 compositionality, 并不能处理 speaker 和 listener 对 symbol 和 concept 的对应不完全相同的情况. 如图 3 所示, 在该 language 中, speaker **TODO: XXXX**. 我们基于 mutual information theory 提出一种解决上述问题的 compositionality 的 bilateral metric MID, 将在后续的 theoretical analysis 中作详细介绍.

4.5 Compositionality and Capacity in Artificial Language teaching

Ease-of-Teaching and Language Structure from Emergent Communication⁴ 这篇文章指出, languages with higher compositionality are easier-to-teach. 这一结论的一个隐含前约束是: agents have same capacity. 为了观察 capacity 对 compositionality 的影响, 我们取消了这一约束 and teach artificial languages with different compositionality to agents with different capacity.

实验配置如下:

object t = concept sequence(c_0, c_1), concept size $|M_0| = |M_1| = 3$, shape $c_0 = \{\text{triangle, circle, square}\}$, color $c_1 = \{\text{red, blue, green}\}$;

symbol sequence $s = (s_0, s_1)$, vocabulary size $|V| = 9$, $s_i (i = 0, 1) = \{a, b, c, d, e, f, g, h, i\}$;

Count of neural nodes in the hidden layer $h_{size} = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

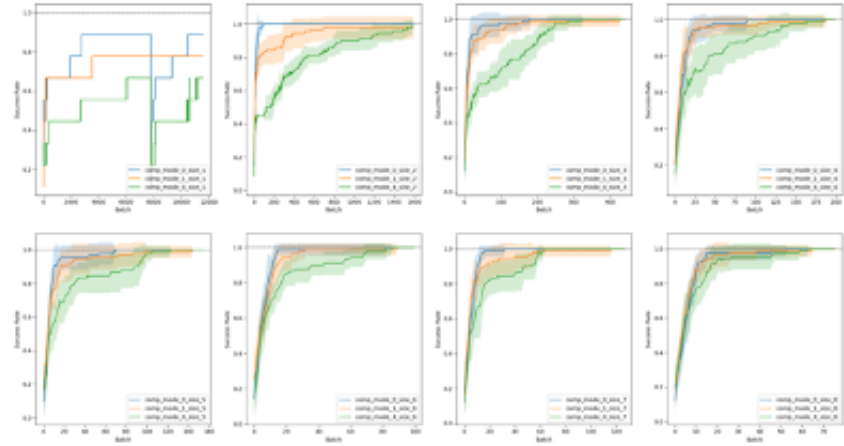
Specifically, we generate 3 different languages, 分别如图 4 所示. 图 4(a) 代表一种 perfect compositional language LA with maximum compositionality, symbol sequence $s = (s_0, s_1)$ 中, s_0 代表 shape, s_1 代表 color. 图 4(b) LB 是一种随机生成的语言, s_0 和 s_1 单独不能代表任何 concept (shape or color). 图 4(c) 表示一种 non-compositional language LC with minimum compositionality, s_0 独自表示 shape 和 color 的组合. We teach LA, LB and LC respectively to a Listener agent and change its capacity by adjusting h_{size} , 得到 accuracy 随训练 iteration 的变化曲线如图 5 所示.

结果显示, 在 $h_{size} = 1$ 时, agent 的 capacity 太小, LA, LB 和 LC 都无法掌握. $h_{size} = 2$ 时, agent 可以掌握 LA, 但无法掌握 LB 和 LC. $h_{size} \geq X$ 时, agent 可以掌握 LA 和 LB, 但依然无法掌握 LC. $h_{size} \geq Y$ 时, agent 的 capacity 足够掌握这三种语言. 综上, we get an observation that languages with higher compositionality require lower capacity of agents. 下面, 我们通过理论分析解释观察到该现象的原因.

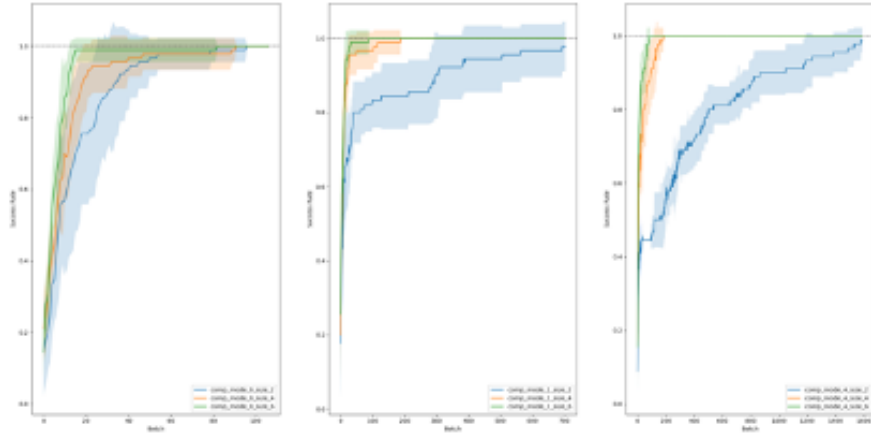
LA	circle	square	triangle	LB	circle	square	triangle	LC	circle	square	triangle
red	(a, a)	(a, b)	(a, c)	red	(a, a)	(a, b)	(a, c)	red	(a, a)	(b, a)	(c, a)
blue	(b, a)	(b, b)	(b, c)	blue	(a, d)	(a, e)	(a, f)	blue	(d, a)	(e, a)	(f, a)
green	(c, a)	(c, b)	(c, c)	green	(b, a)	(b, b)	(b, c)	green	(g, a)	(h, a)	(i, a)

(a) 完美组合的语言 LA (b) 随机生成的语言 LB (c) 完全不组合的语言 LC

Figure 4: 这里应该有一个标题....



(a)



(b)

Figure 5: 不同 capacity (h_{size}), 不同的 artificial language LA, LB and LC, agent 的 correctness 收敛曲线

5 Theoretical Analysis

理论分析过程主要分 3 步:

- a) 用 MSC model 对 listener-speaker 的语言传递过程建模
- b) analyze why languages with higher compositionality require lower capacity of agents
- c) propose a metric MIS to measure compositionality based on the MSC model

5.1 the MSC model

We use the MSC (Markov Series Channel) to model a speaker-listener combination. MSC 由多个子信道串联形成, 并且信息在其中的传递具有马尔可夫性质, 即信道中某节点的值只与前一节点的值有关. In our case, speaker agent S 可看做是一个子信道, 其输入为 concept sequence (c_0, c_1) , 输出 symbol sequence (s_0, s_1) ; listener agent L 也同样作为另一个子信道, 其输入为 symbol sequence (s_0, s_1) , 输出为 predict result $\hat{t} = (\hat{c}_0, \hat{c}_1)$. 整体模型结构如图 6 所示. Speaker 的 policy 可以表示为概率分布 $P(s_0|t = (c_0, c_1))$ 和 $P(s_1|t = (c_0, c_1))$; Listener 的 policy 可表示为概率分布 $P(\hat{t} = (\hat{c}_0, \hat{c}_1)|s_0, s_1)$.

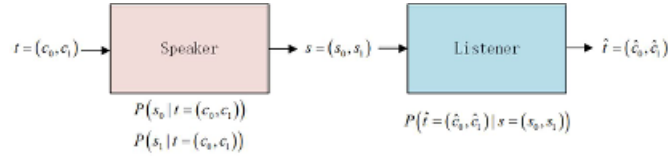


Figure 6: MSC 结构示意图

5.2 analysis on the basis of mutual information theory

结合 MSC 模型和 mutual information theory, 我们接下来对信息传递的过程进行分析.

在 Mutual information theory 中, 互信息 $I(X, Y)$ 表示确定 Y 的取值前后关于信息源 X 的不确定度减少的量, 即从 Y 获得的关于信息源 X 的信息量.

$$I(X, Y) = \sum_{Y \in \mathbb{Y}} \sum_{X \in \mathbb{X}} p(X, Y) \log \left(\frac{p(X, Y)}{p(X)p(Y)} \right)$$

其中 $P(X, Y)$ 是 X 和 Y 的联合概率分布函数, 而 $P(X)$ 和 $P(Y)$ 分别是 X 和 Y 的边缘概率分布函数. 信息源 X 的总信息量为信息熵 $H(X)$, 可以由 X 的边缘概率分布 $P(X)$ 直接求得.

$$H(x) = - \sum_{x \in X} P(X) \cdot \log P(x)$$

由不等式 $I(X, Y) = H(X) - H(X|Y) \leq H(X)$, 我们定义 Y 传递 X 的信息比,

$$RI(X, Y) = \frac{I(X, Y)}{H(X)}, RI(X, Y) \in [0, 1]$$

In our case, speaker 在 MSC 中用 (s_0, s_1) 传递 (c_0, c_1) 的信息给 listener. 根据 speaker 的 policy, 即概率分布 $P(s_0|t = (c_0, c_1))$ 和 $P(s_1|t = (c_0, c_1))$, 可以分别计算出 symbol $s_j (j = 0, 1)$ 传递 concept $c_i (i = 0, 1)$ 的信息比 $RI^{S(c_i, s_j)}$. 其中 $c_i (i = 0, 1)$ 的边缘分布 $P(c_i)$ 为离散均匀分布, 即 c_i 取 M_i 中每个值的概率都等于 $\frac{1}{|M_i|}$, 且 c_0 和 c_1 独立.

$$RI^{(c_i, s_j)} = \frac{I(c_i, s_j)}{H(c_i)} \quad (3)$$

$$I(c_i, s_j) = \sum_{c_i \in M_i} \sum_{s_j \in V} P(c_i, s_j) \cdot \log \left(\frac{P(c_i, s_j)}{P(c_i)P(s_j)} \right) \quad (4)$$

$$H(c_i) = - \sum_{c_i \in M_i} P(c_i) \cdot \log P(c_i) = \log |M_i| \quad (5)$$

$$P(c_i, s_j) = \sum_{c_{1-i} \in M_{1-i}} P(s_j | (c_i, c_{1-i})) \cdot P(c_i)P(c_{1-i}) \quad (6)$$

$$P(s_j) = \sum_{c_i \in M_i} P(c_i, s_j) \quad (7)$$

$$P(c_i) = \frac{1}{|M_i|} \quad (8)$$

可将上述全部信息比整理成 speaker 的传递信息比矩阵 MRI^S :

$$MRI^S = \begin{pmatrix} RI^S(c_0, s_0) & RI^S(c_0, s_1) \\ RI^S(c_1, s_0) & RI^S(c_1, s_1) \end{pmatrix}$$

同理, listener 在 MSC 中以 (\hat{c}_0, \hat{c}_1) 提取 (s_0, s_1) 中的信息. 根据 listener 的 policy, 即概率分布 $P(\hat{t} = (\hat{c}_0, \hat{c}_1) | s_0, s_1)$, 可以计算出 listener 的传递信息比矩阵 MRI^L . 其中 $s_j (j = 0, 1)$ 的边缘分布 $P(s_j)$ 已经在计算 MRI^S 时根据 speaker 的 policy 求得.

$$RI^{(s_j, \hat{c}_i)} = \frac{I(s_j, \hat{c}_i)}{H(s_j)} \quad (9)$$

$$I(s_j, \hat{c}_i) = \sum_{\hat{c}_i \in M_i} \sum_{s_j \in V} P(s_j, \hat{c}_i) \cdot \log \left(\frac{P(s_j, \hat{c}_i)}{P(s_j)P(\hat{c}_i)} \right) \quad (10)$$

$$H(c_i) = - \sum_{s_j \in V} P(s_j) \cdot \log P(s_j) \quad (11)$$

$$P(s_j, \hat{c}_i) = \sum_{\hat{c}_{1-i} \in M_{1-i}} \sum_{s_{1-j} \in V} P(\hat{c}_i, \hat{c}_{1-i} | s_j, s_{1-j}) \cdot P(s_j)P(s_{1-i}) \quad (12)$$

$$P(\hat{c}_i) = \sum_{s_j \in V} P(s_j, \hat{c}_i) \quad (13)$$

$$MRI^L = \begin{pmatrix} RI^S(s_0, \hat{c}_0) & RI^S(s_1, \hat{c}_0) \\ RI^S(s_0, \hat{c}_1) & RI^S(s_1, \hat{c}_1) \end{pmatrix}$$

将 MRI^S 和 MRI^L 作 element-wise 相乘, 我们可以得到信息源 (c_0, c_1) 中的信息在经过由

speaker 和 listener 组成的 MSC 后, 传递的信息比矩阵 MRI^B :

$$MRI^B = MRI^S \odot MRI^L$$

$MRI^B[i, j](i = 0, 1; j = 0, 1)$ 表示在 speaker 和 listener 之间, 由 symbol s_j 传递的 concept c_i 中包含的信息的比例. For a perfect compositional language, like LA in 图 4(a), 一个 symbol 仅传递一个 concept 的信息, 并且传递比例为 1, 即 $MRI^B[i, j]$ 的每一列都是一个 one-hot vector.

推广到一般情况, 即多个 symbol $s_j(j = 0, 1, \dots, M - 1)$ 对应多个 concept $c_i(i = 0, 1, \dots, N - 1)$, $N \times M$ 维的矩阵 MRI^B 的第 j 列向量表示 symbol s_j 传递所有 M 个 concepts 的信息比. 该列向量与 one-hot vector 的相似度越低, 表示 symbol s_j 越倾向于分散传递更多 concepts 的信息 (i.e. compositionality 越低), 从而 symbol s_j 携带的信息复杂度越高, 最终导致 agents 表征 symbol s_j 中的信息需要的 capacity 越大. 示意图如图 7 所示. 这一分析与节 4 中提到的 observation 一致.

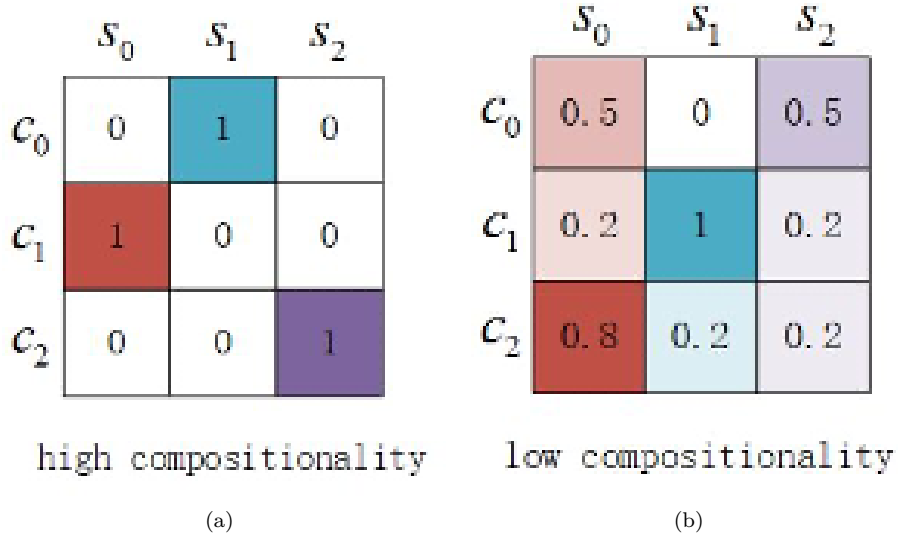


Figure 7: 分析示意图

5.3 ‘bilateral’ metrics for communication

此外, 我们用欧氏距离计算 MRI^B 的列向量与 one-hot vector 的相似度, 归一化之后得到一个 compositionality 的 metric MIS:

$$MIS = 1 - \frac{1}{\sqrt{M(N-1)}} \sqrt{K}$$

$$K = \sum_{j=0}^{M-1} \left[\left(1 - \max_{i=0, \dots, N-1} RI^B[i, j] \right)^2 + \sum_{i=0}^{N-1} (RI^B[i, j])^2 - \left(\max_{i=0, \dots, N-1} RI^B[i, j] \right)^2 \right]$$

In our case, $M = N = 2$.

$$MIS = 1 - \frac{1}{\sqrt{2}} \sqrt{\sum_{j=0}^1 \left[\left(1 - \max_{i=0,1} RI^B[i, j] \right)^2 + \left(\min_{i=0,1} RI^B[i, j] \right)^2 \right]}$$

MIS captures the view that a single symbol of emergent languages with higher compositionality should be used to ground or transmit a certain concept ‘bilaterally’ and more exclusively between listener and speaker. 与其他的 metric 不同, 例如 topo 和 posdis, MIS 同时考虑了 listener 和 speaker 的语义一致性, 对于 speaker 和 listener 语义不完全一致的情况, 例如图 3 中的语言, 能更好的判断 compositionality.

6 Emergent Language Experiments

We get an observation that teach languages with higher compositionality to a listener agent require lower capacity of model, 并且在上一个 section 中通过理论分析解释了其合理性. 进一步的, 我们提出一个猜想: lower capacity facilitates the emergence of language with higher compositionality. 下面我们通过实验验证这一猜想.

实验不预先指定 artificial languages, 而是通过交替训练 speaker 和 listener 自然产生语言. Agent 的模型结构, 训练算法, 和评估方法与 Experimental Framework 中所述一致. 实验选取 5 组 concept size 和 vocabulary size 的配置如下:

- (a) Concept size $|M_0| = 3$, $|M_1| = 3$, vocabulary size $|V| = 4$;
- (b) Concept size $|M_0| = 3$, $|M_1| = 3$, vocabulary size $|V| = 6$;
- (c) Concept size $|M_0| = 3$, $|M_1| = 3$, vocabulary size $|V| = 10$;
- (d) Concept size $|M_0| = 3$, $|M_1| = 6$, vocabulary size $|V| = 10$;
- (e) Concept size $|M_0| = 8$, $|M_1| = 4$, vocabulary size $|V| = 10$;

在每组配置中, 改变模型的 capacity (i.e. h_{size}), 并对每个 h_{size} 的 agents 训练多次至收敛, 即分别产生多个语言. h_{size} 的取值如下:

$$h_{size} = 2, 4, 6, 8, 10, 15, 20, 30, 40, \dots, 100$$

分别统计产生语言的 compositionality (measured by MID) 的平均值和标准差.

得到配置 (a) 的实验结果如图 8 所示. 图 8(a) 是不同 h_{size} 下, 产生语言的 MID 值的均值-标准差曲线图. 从图中可以看出, 随着 h_{size} 降低, MID 的均值呈明显的上升趋势, 标准差的区间对不同 h_{size} 不明显. 图 8(b) 是 MID- h_{size} 的散点图, 每个点代表一个 h_{size} 下产生的一种语言的 compositionality. 从图中可以看出, 对于 h_{size} 较大的情况, 例如 $h_{size} = 100$, MID 也偶尔可以接近 1, 但多数在 XX 到 XX 之间浮动; 对于 h_{size} 较小的情况, 例如 $h_{size} = 2$, MID 没有低于 XX. 这表示 agents with lower capacity 将被迫掌握更高组合性的语言, 因为这些 agents 无

法表征低组合度语言的 symbols 中所包含的高复杂度的信息. 这验证了我们之前的猜想: lower capacity facilitates the emergence of language with higher compositionality.

所有配置 (a)(b)(c)(d)(e) 的实验结果汇总在图 9 中, 图中结果表明, 对不同的 concept size 和 vocabulary size, capacity 对 compositionality 的影响趋势一致, 都符合上述猜想.

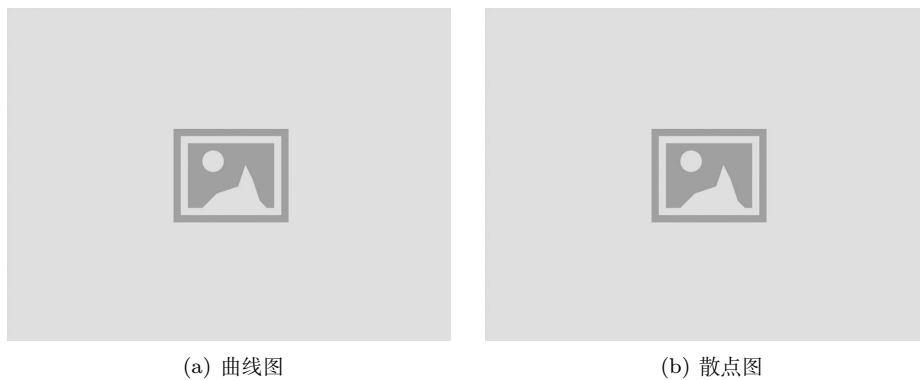


Figure 8: 2 张图分别画配置 (a) 下 compositionality - h_{size} 均值-标准差曲线图和散点图

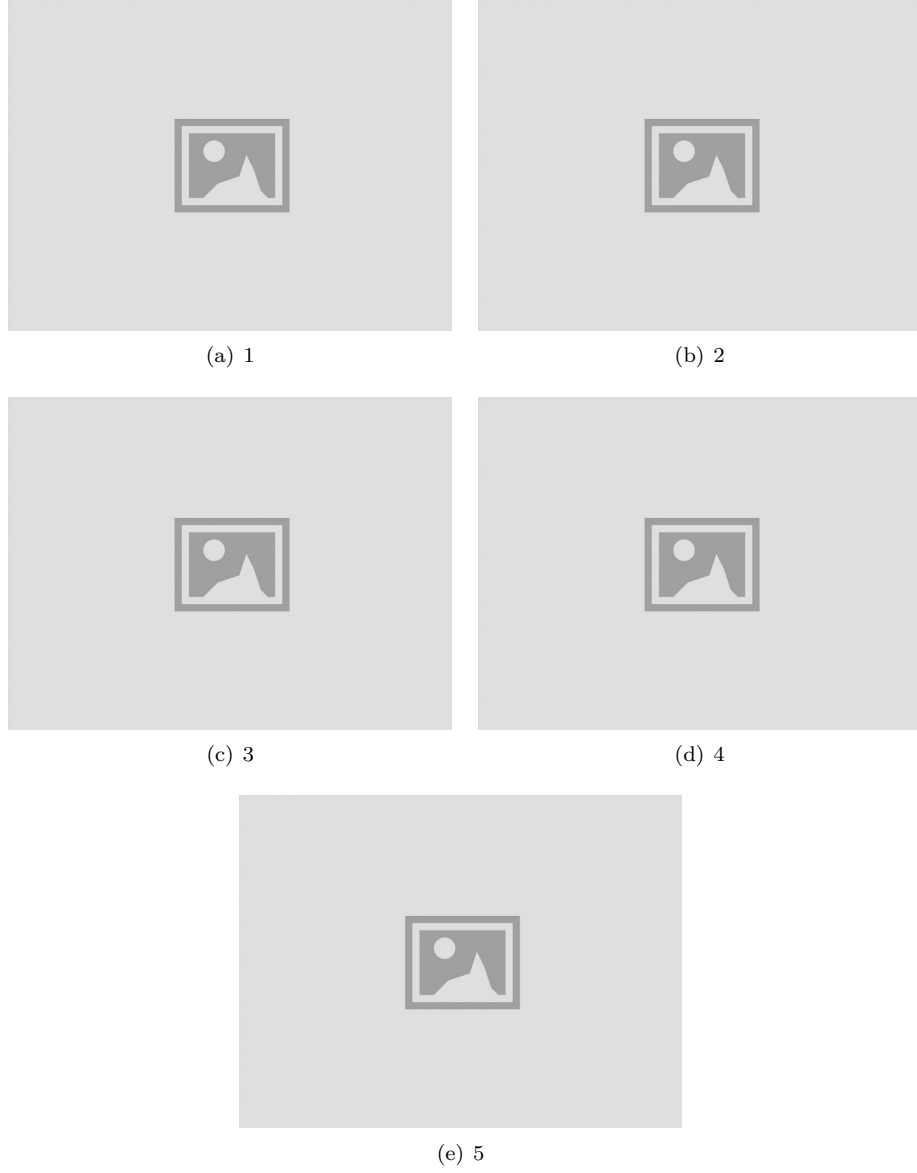


Figure 9: 5 张图分别画不同 concept size 和 vocabulary size 配置下 compositionality - h_{size} 均值-标准差曲线图

Bibliography

- [1] Henry Brighton and Simon Kirby. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial life*, 12(2):229–242, 2006.
- [2] Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. *arXiv preprint arXiv:2004.09124*, 2020.
- [3] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv preprint arXiv:1804.03984*, 2018.
- [4] Fushan Li and Michael Bowling. Ease-of-teaching and language structure from emergent communication. In *Advances in Neural Information Processing Systems*, pages 15851–15861, 2019.