Characterization Capacity of Agents and Compositionality from Naturally Emergent Communication

**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 this 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.

**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.

废话太多，简单强调发生语言很重要即可，可引用乔姆斯基XXX。

people try to make the emergent language similar to human natural language. XXX用了XXX方法做了XXX致力于让机器语言接近人类语言。（引用其他语言学相关工作，尤其是AAAI上的）

一个重要的引用：Barbara Partee 2004 提出的compostionality。

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.

这段废话也多，只需要强调High compositionality的语言的两个特征：

1. Syntax
2. 一个symbol对应一个concept

但是low compostionality的语言没有这两个特征

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compostionality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. 废话太多，应一句写完，长度不超过下面的however后面。

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 \emph{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.)这句话感觉不太符合表意，而且复杂度并不是指互信息的。理论证明的是“互信息矩阵与单位矩阵的相似度越低，表示单个symbol 越倾向于分散传递更多concepts的信息 (i.e. compostionality / bilaterality越低)”。后面理论分析部分得到结论的过程是“互信息矩阵与单位矩阵的相似度越低，表示单个symbol 越倾向于分散传递更多concepts的信息 (i.e. compostionality越低)，(从这里开始就不严谨了) 从而单个symbol携带的语义信息的复杂度越高，最终导致agents表征单个symbol中的语义信息需要的capacity越大。”

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 \emph{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.

**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影响因素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 compostional 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](file:///C:/Users/haoyi/AppData/Local/youdao/dict/Application/8.9.4.0/resultui/html/index.html#/javascript:;) ’ 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](file:///C:/Users/haoyi/AppData/Local/youdao/dict/Application/8.9.4.0/resultui/html/index.html#/javascript:;) measure characterization capacity’s impact on compositionality of emergent language。

Chen start

**Experimental Framework**

*#In this section, we introduce a referential game platform and our listener-speaker model.*

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。

*#subsection1: 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 t^；

c). 当t = t^时，认为agents在本次游戏成功，S和L分别获得reward R(t, ) = 1；否则，agents失败，并分别获得reward R(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 >= |M\_0|\*|M\_0|，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 \* |V|的vector，s由该vector表示。

predict result 由一个长度为|M\_0|\*|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 = (, )。具体地，[i \* |M1| + j] = 1 correspond to [i] = 1 & [j] = 1 (i = 0, ..., |M0| - 1; j = 0, ..., |M1| - 1)。

我们定义的 = t是指t和分别对应的object相同，i.e.对应的(c0, c1) = (, )。

*#subsection2: Agent architecture*

Agents以各自强化学习的策略进行上述referential game。将speaker agent S和listener agent L的policy分别记为pi\_S和pi\_L。pi\_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。pi\_L表示给定输入symbol sequence s = (s\_0, s\_1)，listener输出predict result 的条件概率P( | s\_0, s\_1)。listener L根据概率分布P( | s\_0, s\_1)随机采样输出。Agents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figure1所示。

Figure1. 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 \* 2。Splitting h^S equally得到两个长度为h\_size的neural vectors h^S\_0和h^S\_1。h^S\_i (i=0,1)各自依次经过一个全连接层和一个softmax得到output layer o^S\_i。o^S\_i (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(si| 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^L\_i，h^L\_i的神经元节点数也是h\_size。Concatenating h^L\_0和h^L\_1得到长度为h\_size \* 2的neural vectors 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|M0|\*|M1|的vector，其每个分量表示，给定输入symbol sequence s = (s0, s1)，的每个取值的概率，即P( |s0, s1)。

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。

*#subsection3: 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theta^S和theta^L分别表示speaker和listener的policy pi^S和pi^L的全部参数。训练speaker时，固定policy pi^L的参数theta^L，训练目标是调整参数theta^S，使其基于策略pi^S获得的期望奖励J(theta^S, theta^L) = E\_pi^S\_pi^L[R(t, t^)]最大。同理，训练listener时，固定policy pi^S的参数theta^S，最大化期望奖励J(theta^S, theta^L)。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：

（这里的m换成s\_0, s\_1）

（这里的m,c换成s\_0, s\_1）

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figure2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用于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。

Figure2. agents的Training Algorithm伪代码图

*#subsection4: 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compostionality的关系。当Listener agent L的正确率收敛到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compostionality。

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compostionality，据我们所知，目前并没有一个统一的度量标准。Topographic similarity (Brighton and Kirby, 2006)是一个广为接收的compostionality的度量(e.g., Lazaridou et al., 2018; Li and Bowling, 2019). 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 more similar symbol sequences in context. *Compositionality and Generalization in Emergent Languages*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。Posdis captures the intuition that each symbol should only be informative about a single concept. 但是posdis和topo一样，都只以speaker的policy为基础计算compotionality，并不能处理speaker和listener对symbol和concept的对应不完全相同的情况。如Figure2.5所示，在该language中，speaker XXXX。我们基于mutual information theory提出一种解决上述问题的compostionality的bilateral metric MID，将在后续的theoretical analysis中作详细介绍。

Figure 2.5 一个language反例

**Compositionality and Capacity in Artificial Language teaching**

*#We get a counter-intuitive observation on the relationship between the compositionality and the number of nodes in the hidden layer.*

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

实验配置如下：

object t =concept sequence (c\_0, c\_1)，concepts size |M\_0| = |M\_1| = 3，shape c\_0 = {triangle, circle, square}，color c\_1 = {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Figure3所示。Figure3(a)代表一种perfect compositional language LA with maximum compostionality，symbol sequence s = (s\_0, s\_1)中，s\_0代表shape, s\_1代表color。Figure3(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)。Figure3(c)表示一种non-compositional language LC with minimum compostionality，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 adjust h\_size，得到accuracy随训练iteration的变化曲线如Figure4所示。

结果显示，在h\_size=1时，agent的capcity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 >= X时，agent可以掌握LA和LB，但依然无法掌握LC。h\_size >= 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。

Figure3. (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

Figure4. 不同capacity (h\_size)，不同的artificial language LA,LB and LC，agent的correctness收敛曲线

Chen end

**Theoretical Analysis**

*#In this section, we analyze the observation above mentioned theoretically on the basis of mutual information theory.*

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 compostionality based on the MSC model

*#subsection1: 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可看作是一个子信道，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  = (, )。整体模型结构如Figure5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( = (, )| s\_0, s\_1)。

Figure5. MSC结构示意图

*#subsection2: 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。

（; 换成 , p换成大写P, x,y换成X,Y X,Y换成花体X和花体Y（一般用来表示矩阵的那种））

其中 P(X,Y) 是 X 和 Y 的[联合概率分布函数](https://zh.wikipedia.org/wiki/%E8%81%94%E5%90%88%E5%88%86%E5%B8%83)，而P(X)和P(Y)分别是 X 和 Y 的[边缘概率](https://zh.wikipedia.org/wiki/%E8%BE%B9%E7%BC%98%E6%A6%82%E7%8E%87)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。



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

RI(X, Y) = I(X, Y) / H(X)

RI(X, Y)的取值在[0,1]中。

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ci取Mi中每个值的概率都等于 1/|Mi|,且c\_0和c\_1独立。



可将上述全部信息比整理成speaker的传递信息比矩阵MRI^S:



同理，listener在MSC中以 (, )提取(s\_0, s\_1)中的信息。根据listener的policy，即概率分布P( = (, ) | 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。





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[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 compostional language，like LA in Figure3(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symbols s\_j (j = 0,1,...,M-1)对应多个concepts c\_i (i = 0,1,...,N-1)，NXM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. compostionality越低)，从而symbol s\_j携带信息复杂度越高，最终导致agents表征symbol s\_j中的信息需要的capacity越大。示意图如Figure6所示。这一分析结果与上个section中提到的observation一致。

Figure6. 分析示意图

*#subsection3: ‘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compostionality的metric MIS:

（最前面加上MIS = ）

In our case, M = N = 2.

（最前面加上MIS = ）

MIS captures the view that a single symbol of emergent languages with higher composionality 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Figure2.5中的语言，能更好的判断compostionality。

**Emergent Language Experiments**

*#In this section, we verify the hypothesis above mentioned by experimental results*

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：

1. Concept size |M0| = 3, |M1| = 3，vocabulary size |V| = 4；
2. Concept size |M0| = 3, |M1| = 3，vocabulary size |V| = 6；
3. Concept size |M0| = 3, |M1| = 3，vocabulary size |V| = 10；
4. Concept size |M0| = 3, |M1| = 6，vocabulary size |V| = 10；
5. Concept size |M0| = 8, |M1| = 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,...,100}。

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

得到配置(a)的实验结果如Figure7所示。Figure7(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，标准差的区别对不同h\_size不明显。Figure7(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compostionality。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Figure8中，图中结果表明，对不同的concept size和vocabulary size，capacity对compostionality的影响趋势一致，都符合上述猜想。

Figure7. 2张图分别画配置(a)下composionality - h\_size 均值-标准差 曲线图 和 散点图

Figure8. 5张图分别画不同concept size和vocabulary size配置下composionality - h\_size 均值-标准差曲线图

**Conclusion**

**Reference**