文章编号：1003-0077（2017）00-0000-00

# 分布式文本表示与大脑语言表征相似性的可解释性分析

张肖寒[1,2] 王少楠[1,2] 宗成庆[1,2]

（1.中国科学院自动化研究所 模式识别国家重点实验室，北京市 100000；
2.中国科学院大学 人工智能学院，北京市 100000）

**摘要：**近年来，研究发现分布式文本表示与大脑的语言表征存在一定相似性，但不同的文本表示模型与大脑语言表征的相似程度存在差异。该文针对这一问题，使用可解释的语义和句法信息作为中介，分析导致差异的原因。该文选择了 Word2Vec、GloVe、MacBERT、GPT2 四种文本表示模型，分别计算每种文本表示预测语义特征、句法特征和脑活动的准确率，并在此基础上进一步分析模型预测脑活动的表现与模型编码各语言特征表现之间的关系。实验结果表明，文本表示编码语义和句法的表现均与其预测脑活动的表现存在显著相关，而其中句法特征的相关性相比语义特征更高。这一结果表明，在使用分布式文本表示解释大脑语言机制，尤其是大脑的句法机制时，所选择的文本表示模型可能会对结果有较大影响。

**关键词：**分布式文本表示；脑活动；神经编码

**中图分类号：**TP391 **文献标识码：**A

# Explaining the Similarity Between Distributed Text Representations and the Brain with Linguistic Factors

Xiaohan Zhang[1,2], Shaonan Wang[1,2], and Chengqing Zong[1,2]

(1. National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing 100000, China;
2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100000, China)

**Abstract:** Recent studies have shown that distributed text representations are brain-like at various degrees. In this paper, we investigated the explainable linguistic factors behind this phenomenon. We chose four types of text representations, i.e., Word2Vec, GloVe, MacBERT, and GPT2, and computed their accuracies in predicting explainable semantic and syntactic features, as well as brain activation. Then, we analyzed the relationship between the prediction performance of linguistic features and brain activation. Results show that the ability to encode both syntax and semantics significantly correlates with the prediction performance of brain activation, and syntax has a higher correlation than semantics. This finding outlines the importance of choosing text representation models when explaining the neural basis of language, especially syntax in the brain.

**Key words:** distributed text representation; brain activation; neural encoding

## 0 引言

语言是人类特有的认知功能，大脑的语言处理机制是神经科学领域的研究重点之一[1]。近年来，自然语言处理技术快速发展，深度语言模型在包括机器翻译、自动摘要等许多曾被认为只有人类才能完成的任务上取得了很好的性能。因此，语

**收稿收稿日期**：20XX- XX-XX；**定稿日期**：20XX- XX-XX
**基金项目**：国家自然科学基金（61906189）；国家自然科学基金（62036001）

言模型是否学到了一般性的语言特征，其表征语言的方式和人脑是否有相似之处，以及能否借助深度语言模型来研究大脑的语言处理机制，成为近年来广受关注的研究问题[2,3]。

语言模型能够将文本转换成高维空间中的分布式向量，已有的研究表明这些向量中编码了一般性的语言学特征[4,5]，包括语义特征和句法特征。而人的大脑在接收到语言信号之后，会激活一系列脑区，进行一系列操作和计算，整个过程也包含对语义信息和句法信息的表征[6]。因此，很多工作开始关注语言特征在分布式文本表示中的编码方式与其在大脑中的表征方式是否存在相似性，能否借助分布式文本表示来研究语义和句法对应的大脑表征机制。近年来的一些研究工作发现语言模型建立的文本表示和语言在大脑中的表征存在一定相似性，即文本表示可以在一定程度上预测大脑在处理语言时的活动。同时，一些工作开始使用分布式文本表示作为语言学特征的向量表征，通过建立文本表示与大脑活动间的关系，来研究这些语言学特征在大脑中的表征[7-14]。例如Caucheteux等[14]使用深度语言模型不同层的输出来表示句法和语义，通过神经编码模型（neural encoding）来分析语义和句法在大脑中的表征。

这些工作的重要前提之一是其所使用的文本表示中编码了目标特征。然而，分布式文本表示缺少可解释性。目前虽然有一些工作尝试解释分布式文本表示中所编码的语言特征[15-19]，但这些工作大多是通过文本表示在探针任务（probing task）上的表现来分析其中是否编码了完成探针任务所需的信息，缺少对语言特征的直接定义。目前对于不同类型的文本表示模型在编码不同语言信息能力上的差异缺少较为系统的分析。更重要的是，不同的文本表示编码不同语言特征的能力可能是不同的，而且预测大脑活动的准确率也是不同的。目前尚不清楚文本表示编码语言特征的能力会如何影响其预测大脑活动的能力。随着近年来越来越多的工作开始使用分布式文本表示研究大脑的语言表征机制，明确这一问题对于后续文本表示模型在大脑语言机制研究中进一步的应用至关重要。

在此背景下，本文旨在使用语言特征来分析不同文本表示模型中编码的语言信息类型和导致文本表示与大脑语言表征相似程度差异的原因。为此，我们选择了基于大脑功能划分的六种语义特征，包括：视觉，动作，社会，情绪，空间，时间。这六个语义维度包含感觉-运动和非感觉-运动特征，与经历和获取语义知识的神经系统相对应，是语言理解过程中影响神经活动最重要的语义特征[20]。此外，我们还选择了与大脑构建句法结构难度相关的五种句法特征，包括：自上而下、自下而上、左角方式构建短语结构时的句法分析步数，短语结构树的深度，依存结构中每个词与其中心词的距离。这五种句法特征来自句法分析中最主要的两种句法结构——短语结构和依存结构[21]，同时也是大脑句法机制研究中重要的特征[22-25]。本文使用这11种可解释的语言特征作为中介，分析了不同的文本表示模型中所编码的语言特征类型以及导致文本表示模型与大脑语言表征相似程度差异的原因。

实验结果表明，不同的文本表示模型编码不同信息的能力存在明显差异。整体来说，相比于上下文无关的文本表示，上下文相关的文本表示编码语义特征和句法特征的能力更好；与之前发现不同的是，我们的实验结果中深度语言模型的3-5层编码语义的能力最好，而7-10层编码句法特征的能力最好。此外，实验结果显示文本表示模型之间编码语义和句法特征能力的差异都与其预测脑活动准确率的差异存在较高的相关性，其中，编码句法的能力与预测脑活动的能力相关性要更高一些。该结果表明，在使用分布式文本表示研究大脑的句法机制时，应选择编码句法较好的文本表示模型，而深度语言模型的中间层是较好的选择。

综上所述，本文的主要贡献如下：

（1）采用根据心理学和认知神经科学发现的基于大脑功能划分的一组基础语义特征和根据大脑处理句法结构难度构建的一组基础句法特征分析了分布式文本表示中编码的语言特征；

（2）以语言学特征为中介，分析了影响文本表示与大脑语言表征相似性的原因，发现句法因素的作用要强于语义因素。

## 1 相关工作

### 1.1 分布式文本表示的可解释性

由于深度语言模型的优异表现，近年来很多工作探索了其生成的分布式文本表示是否编码了一般性的语言特征[4,5,15-19,26-28]。这类工作可以分为两类，一类是直接通过探针任务（probing tasks）来研究文本表示中是否编码了某些特征。例如Tennry[26]等人提出的边界探针（edge probing）方法，通过线性分类任务分析词向量中是否编码了词性等句法信息和共指消解等语义信息。他们的实验发现上下文相关词向量相比于上下文无关词向量的提升主要表现在句法任务上，在语义任务上提升较小。另一类是通过分析文本表示内部不同维度的特点来研究语言特征在文本表示中的组织形式[27,28]。例如 Hennigen[27]等提出了内部探针（intrinsic probing）的方法，通过寻找和形态句法最相关的一组维度子集，来分析形态句法信息为分布式或局部编码，实验结果发现 BERT 需要较多的维度来编码形态句法信息，而形态句法在 fastText 中的编码则较为局部，集中在少数维度。

除了上述对于文本表示编码特征的研究之外，一些工作对于深度语言模型的不同层所编码的特征进行了分析。例如，Jawahar 等[18]将语言特征分成表层（surface）特征、句法（syntactic）特征和语义（semantic）特征，通过为不同特征定义相应的探针任务，来分析 BERT 的不同层分别学到了哪些特征。他们的实验结果表明，BERT 的低层学到了表层特征，中间层学到了句法特征，而高层学到了语义信息。

这些工作虽然已经发现分布式文本表示中编码了丰富的句法和语义信息，但是，上述已有工作大多通过定义探针任务，根据模型在这些任务上的表现来间接判断其中是否编码了相关的语义和句法特征。然而，这些探针任务本身以及完成任务所需要的信息都比较复杂，缺少对语义和句法信息的直接定义。因此，本文采用了细粒度的可解释的语义和句法特征，分析不同的文本表示模型是否编码了这些特征，以及深度模型内不同层分别学到了哪些特征。而且，由于本文选择的特征是基于心理学和认知神经科学对大脑功能的划分，因此在此基础上分析的结果更适合研究分布式文本表示和大脑语言表征的相似性。

### 1.2 分布式文本表示与大脑语言表征的相似性

由于分布式文本表示在自然语言处理任务上取得的性能进步，一些研究开始关注其与大脑语言表征的相似性[7-14]。其中最主要的一类研究是借助文本表示，通过神经编码或神经解码的方式来探索大脑的语言理解机制。例如，Zhang 等[7]使用神经编码模型，通过词向量来预测大脑处理语言时的活动，来分析概念在大脑中的组织形式。Wang 等[12]使用分离的语义向量和句法向量，通过表征相似性分析（representational similarity analysis，RSA）来研究句法和语义在大脑中的表征方式。Caucheteux 等[14]通过对 GPT2 不同层的表示进行线性组合得到语义和句法向量，并使用神经编码模型分析语义和句法在大脑中的表征。这些研究工作中的一个重要前提是所使用的文本表征中编码了目标信息。若所使用的文本表征没有编码目标信息，或者编码某种信息的能力较弱，那么得到的结果很可能并不可靠。此外，如果不同的文本表示编码某些语言特征的能力不同，那么采用不同文本表示得到的实验结果也会有差异，同样也会降低结论的可靠性。

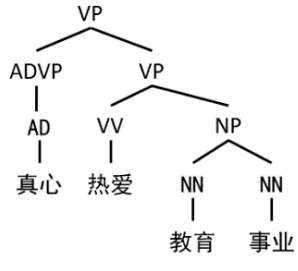
目前有一些工作对于可能影响文本表示与大脑活动相似性的因素进行了分析[13,29,30]，例如，Pasquiou 等[29]从文本表示模型的隐层维度、模型结构（例如 LSTM 或 Transformer）、损失函数、训练数据量等方面，对影响文本表示与脑语言表征相似性的因素进行了系统的分析，发现文本表示模型的结构和训练数据等对结果有较大的影响。而且，模型预测下一个词的表现与预测大脑活动的表现之间有较强的相关性。然而，这些工作仅从文本表示模型的结构和训练过程等方面进行了分析，缺少基于语言特征的可解释性分析。

综上，本文采用根据心理学和认知神经科学领域的发现基于大脑功能划分的一组基础语义特征，以及根据大脑处理句法结构难度构建的一组基础句法特征，来分析不同文本表示中所编码的特征类别以及影响文本表示与大脑相似性的语言特征。

## 2 实验材料

### 2.1 脑活动数据

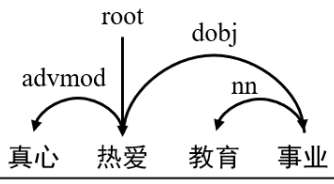
为了获得更接近真实语言场景下的大脑语言表征，我们采集了基于自然语言刺激的汉语神经影像数据。在自然语言刺激选择方面，我们在《人民日报评论》选择了 60 个故事，每个故事的音频时长 5 分钟左右，总时长约 5 小时，故事内容涵盖教育、科技等多个主题。每段故事都有对应的文本，每段故事的文本都对照音频进行了人工校正，并标注了故事中每个词出现和结束的时间，以便于和 fMRI 数据进行对齐。经统计，60 个故事共 52269 个词，去掉重复词后，文本的词表共包含 9153 个词。




| 词 | 真心 | 热爱 | 教育 | 事业 |
|---|---|---|---|---|
| **depth** | 4 | 4 | 5 | 5 |

| 短语结构树 | **nc_td** | | **nc_bu** | | **nc_lc** | |
|---|---|---|---|---|---|---|
| | 生成过程 | 特征值 | 生成过程 | 特征值 | 生成过程 | 特征值 |
| | VP → ADVP VP | | AD → 真心 | 2 | AD → 真心 | |
| | ADVP → AD | 3 | ADVP → AD | | ADVP → AD | 3 |
| | AD → 真心 | | VV → 热爱 | 1 | VP → ADVP VP | |
| | VP → VV NP | | NN → 教育 | 1 | VV → 热爱 | 2 |
| | VV → 热爱 | 2 | NN → 事业 | | VP → VV NP | |
| | NP → NN | 2 | NP → NN | 4 | NN → 教育 | 2 |
| | NN → 教育 | | VP → VV NP | | NP → NN | |
| | NN → 事业 | 1 | VP → ADVP VP | | NN → 事业 | 1 |

图 1 短语结构句法特征计算示例




| 依存结构树 | **depdist** | |
|---|---|---|
| | 真心 | 0 |
| | 热爱 | 0 |
| | 教育 | 0 |
| | 事业 | 1 |

图 2 依存结构句法计算示例

我们招募了 12 位汉语母语者被试，所有被试的年龄均在 20-30 岁之间，且均为在校大学生。我们采集了每位被试在听到每个汉语故事时的大脑核磁共振成像（functional Magnetic Resonance Imaging, fMRI）数据。fMRI 数据共分 6 次进行采集，每位被试每次实验需听 10 个故事。所采集的 fMRI 数据空间分辨率为 2 毫米，重复时间（Repetition Time, TR）为 0.71 秒。原始数据采集全部结束后，我们按照 HCP pipeline[31]对 fMRI 数据进行了预处理，并对数据进行了技术验证，结果表明了 fMRI 数据具有较高的质量。关于该 fMRI 数据集的更多细节可以参考论文[32]。

## 2.2 语义和句法特征标注

为了得到可解释的语义和句法特征，本研究选择了基于大脑功能划分的六种语义特征，分别为动作（action）、视觉（vision）、空间（space）、时间（time）、社会（social）、情绪（emotion），以及 5 种由依存结构和短语结构得到的句法特征，分别为短语结构树深度（depth），分别用自下而上（bototm-up, nc_bu）、自上而下（top-down, nc_td）、左角（left- corner, nc_lc）方法构建短语结构树时每个词的句法分析步数（node count），和依存距离（dependency distance, depdist），这些句法特征分别代表大脑处理不同句法结构时的难度。

对于语义特征，我们选择了 60 个故事的词表中除专有名词、功能词等之外的 7513 个词，标注了每个词在上述 6 个语义维度上的评分。其中，情绪维度的评分为[-6, 6]，6 表示正向，-6 表示负向，0 表示中立。其余五种语义标注评分为从 1 到 7，7 分表示非常高，1 分表示非常低。我们招募了 30 名汉语母语被试进行语义标注，并将 30 人的结果平均，作为每个词的最终语义评分。表 1 中展示了这 6 种语义特征的标注实例。

对于句法特征，故事中的每个句子都人工标注了短语结构树，并使用 Stanford Core NLP 将短语结构树转换成对应的依存结构树。本论文所使用的 5 种句法特征有 4 种来自于短语结构句法，分别是 depth，nc_lc，nc_bu，nc_td；另外一种来自依存句法，为 depdist。

表 1 语义特征标注示例

| 词 | 语义特征 | | | | | |
|---|---|---|---|---|---|---|
| | 动作 | 视觉 | 空间 | 时间 | 社会 | 情绪 |
| 真心 | 1.77 | 1.83 | 1.07 | 1.03 | 1.53 | 3.1 |
| 热爱 | 1.87 | 1.7 | 1.13 | 1.17 | 1.93 | 3.5 |
| 教育 | 2.6 | 2.17 | 1.53 | 1.93 | 5.57 | 1.47 |
| 事业 | 1.33 | 2.27 | 1.27 | 1.13 | 4.27 | 1.03 |

## 2.3 分布式文本表示模型

本文选择了 4 种文本表示模型：GloVe[33]、Word2Vec[34]、MacBERT[35]、GPT2[36]。模型详情如下：

- Word2Vec：300 维，训练语料为新华社新闻，语料大小在 12G 左右
- GloVe：300 维，模型训练与 Word2Vec 相同；
- MacBERT：12 层，隐层维度为 768，模型来自 https://huggingface.co/hfl/chinese-macbert-base；

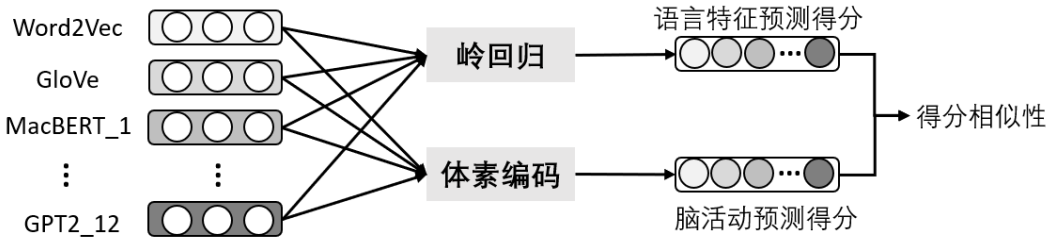



图 3 整体框架

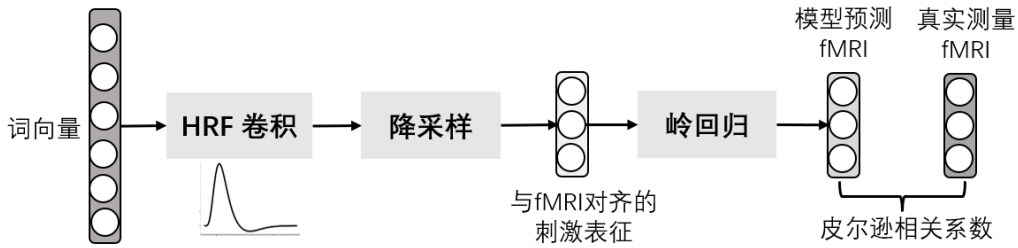



图 4 体素编码模型

- GPT2：12 层，隐层维度为 768，模型来自 https://huggingface.co/uer/gpt2-chinese-cluecorpussmall。

## 3 方法

### 3.1 整体框架

本文的整体框架，如图 3 所示，共分为两部分：对于每一种文本表示，1）为每一种语义和句法特征训练一个岭回归模型，根据词向量预测特征值；2）训练体素编码模型，预测大脑活动。最后，我们计算不同文本表示在预测语言特征和大脑活动上得分的相关性，以判断导致文本模型与大脑语言表征相似性差异的原因。

用 $w_i$ 表示单词，$\boldsymbol{x}_i$ 表示该单词的词向量，$y_{ij}$ 表示该单词在特征 $j$ 上的标注值，则预测语言特征的回归模型可以写成：

$$y_{ij} = \boldsymbol{W}_j \boldsymbol{x}_i$$

其中 $\boldsymbol{W}_j$ 为回归模型的权重。

体素编码模型的内部构成如图 4 所示。由于 fMRI 的特性，词向量无法直接预测 fMRI 数据，因为 fMRI 数据的时间分辨率比较低，每一个 TR 的 fMRI 数据都可能对应若干个词。此外，fMRI 并不是直接测量神经元的活动，而是神经元活动所引发的血氧水平依赖（blood oxygen level dependent, BOLD）信号。BOLD 信号的变化相对神经元活动较为缓慢，通常会在神经元放电之后 6 秒左右达到峰值，而后慢慢降低。整个过程一般用血液动力学响应函数（hemodynamic response function, HRF）表示。因此，体素编码需要首先将词向量与 HRF 函数卷积，并降采样到 fMRI 的采样频率，实现与 fMRI 的对齐；然后为大脑的每一个体素训练一个回归模型，根据对齐后的文本表示预测其活动。用 $[w_1, w_2, \ldots, w_i]$ 表示单词序列，$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i]$ 表示对应的的词向量序列，$\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k]$ 表示所采集的 fMRI 数据，则体素编码模型可以表示为：

$$\boldsymbol{R} = \boldsymbol{W} \times downsample(conv(HRF, \boldsymbol{X}))$$

词向量在语言特征和大脑活动上的预测得分都使用预测结果和实际值的皮尔逊相关系数（Pearson correlation）来表示。在得到所有文本表示模型的语言特征预测得分和大脑活动预测得分后，我们将所有文本表示模型在每一种语言特征上的得分看作一个向量，并计算其与大脑活动得分之间的相关性。其背后的逻辑为，如果文本表示模型编码某一语言特征的能力是影响其与大脑表征相似性的主要因素，那么文本表示模型预测这一语言特征的得分应该与预测大脑活动的得分变化模式相似，即在这一语言特征上得分高的文本表示在预测脑活动的得分上也应该较高。通过这种方式，我们可以分析导致文本表示与大脑语言表征相似性差异的语言特征。

### 3.2 实验设置

本文的实验在 Word2Vec、GloVe、MacBERT 和 GPT2 的全部 12 层在内共 26 种文本表示上进行，因此语言特征预测得分和脑活动预测得分均为 1×26 维的向量。本文采用嵌套交叉验证的方法来训练岭回归模型。嵌套交叉验证包含两层循环，其中内层循环将训练集分为训练子集和验证子集，选取最优超参，外层循环则在测试集上计

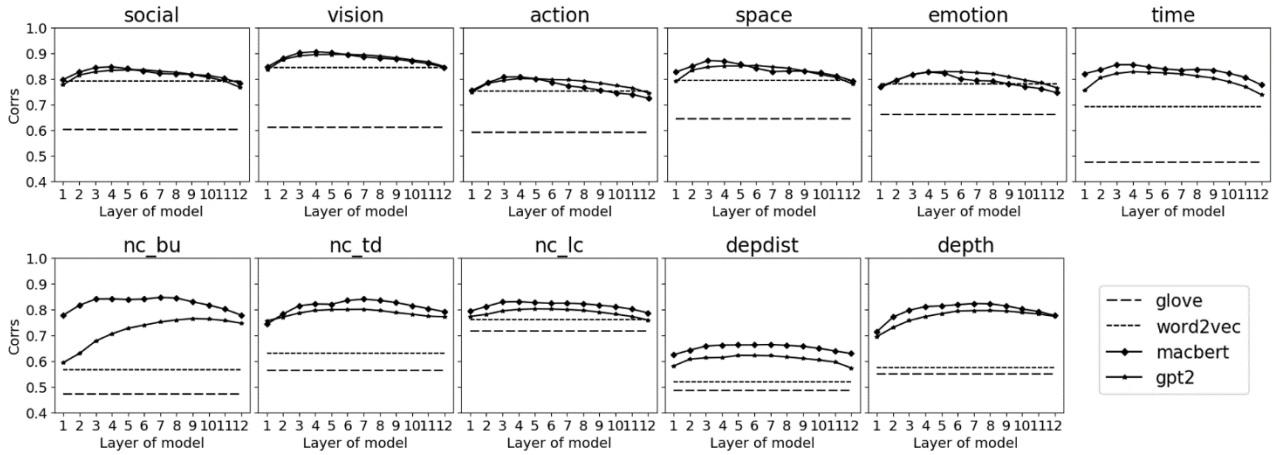



图 5 语义和句法特征的词向量预测结果，纵轴为预测值与实际值的相关系数，横轴为模型层数

算模型预测结果和真实结果之间的皮尔逊相关系数，并将外层循环每一折的相关系数平均，作为词向量在这一特征上的预测得分。

训练语言特征的岭回归模型内层和外层循环均为 10 折交叉验证。体素编码模型也使用嵌套交叉验证的方式训练。对每一个体素，我们计算外层的每一折测试集上模型预测结果与测量值之间的皮尔逊相关系数，将所有折的相关系数平均作为该体素的最终得分。由于大脑中包含上万个体素，我们将所有体素的得分平均，作为最终的脑活动预测得分。

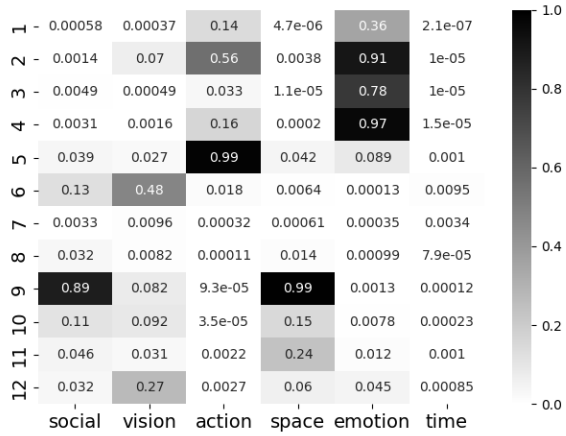



图 6 GPT2 与 MacBERT 相同层得分差异的显著性检验结果（p 值）

## 4 结果和分析

### 4.1 文本表示预测语义和句法特征

不同文本表示模型的语义特征预测得分和句法特征预测得分如图 5 所示，其中横轴代表文本表示模型的不同层，纵轴表示模型的预测得分。本文从上下文相关向量与上下文无关向量之间、上下文相关向量不同层之间的表现差异来进行分析。

从图中可以看出，MacBERT 和 GPT2，尤其是中间层，编码语义和句法信息的能力要好于 Word2Vec 和 GloVe，这说明上下文相关向量学到了更多的语义和句法信息。对于两种上下文无关的词向量，Word2Vec 在每一种语言特征上的得分都显著高于 GloVe（$p<10^{-7}$）。然而，这两种模型之间以及与上下文相关模型之间的差异模式并不相同。在语义特征上，Word2Vec 的表现仅略低于上下文相关词向量，而 GloVe 的表现则远低于其他三种模型；在句法特征上，Word2Vec 和 GloVe 之间表现差异较小，二者与上下文相关词向量差别较大。这说明 Word2Vec 相比 GloVe 能更好的编码语义信息，且能达到和上下文相关模型相近的效果。这一发现与 Tenney 等[26]的发现类似，即上下文模型在句法任务上相比非上下文模型提升较大，而在语义任务上仅有较小的提高。而本文中所发现的 Word2Vec 和 GloVe 的差异则表明，导致这一现象的原因可能是模型的训练方式。Word2Vec、MacBERT 和 GPT2 的训练方式都是最大化给定上文或上下文时单词的条件概率，而 GloVe 则是通过词共现矩阵分解的方式计算得到。前者的训练方式相比后者可能让模型更好的学到语义信息。相比于句法，语义特征在不同的上下文中是相对稳定的，而句法特征则高度依赖上下文，这可能导致上下文模型在句法特征上取得了远好于非上下文模型的表现。在 MacBERT 和 GPT2 这两种模型之间，相同层之间的语义特征得分差异较小。图 6 展示了两种模型相同层之间语义得分差异的显著性。可以看出，除了时间（time）这一语义特征，MacBERT 和 GPT2 在其

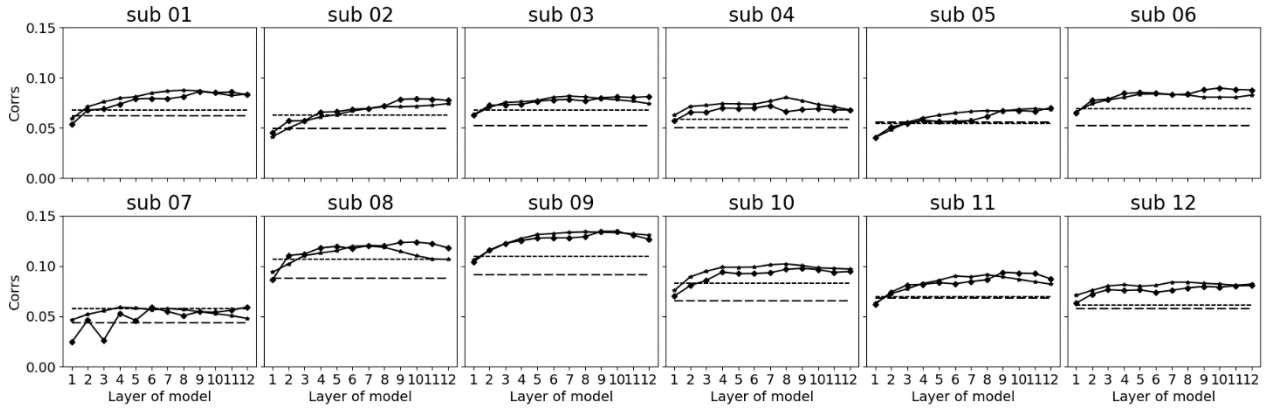



图 7 所有被试的脑活动预测得分，纵轴为预测值与实际值的相关系数，横轴为模型层数

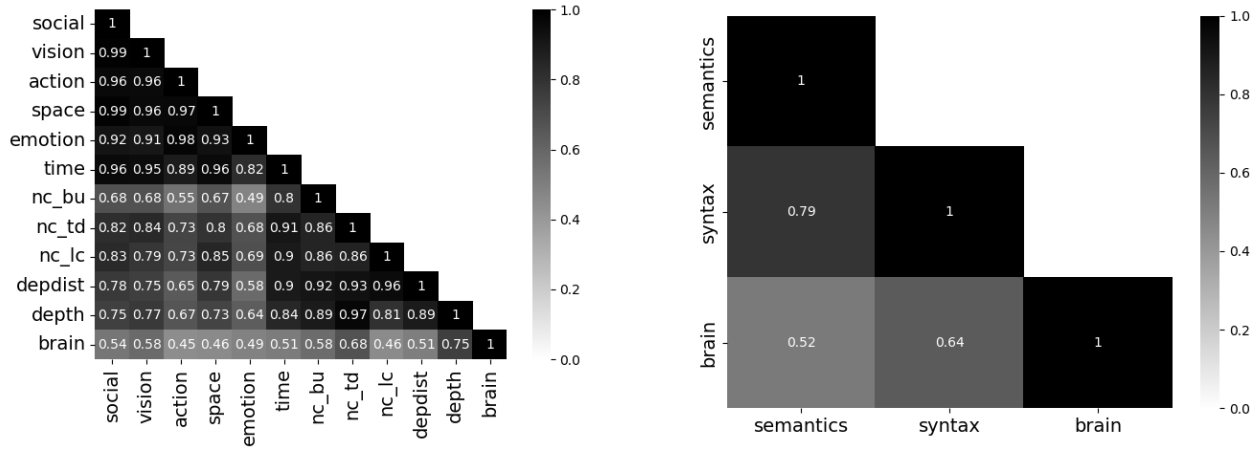



图 8 语言特征得分与脑活动得分的相关系数

他的语义特征上，相同层之间的差异大多数不显著的。例如在情感（emotion）这一特征上，MacBERT 和 GPT2 的前 5 层和最后两层并没有显著的差异（$p>0.01$），而且尤其是二到四层，二者几乎没有差异。由此推断，MacBERT 和 GPT2 这两种都是基于 transformer 结构的模型在编码语义的能力上并没有差异。而在句法特征上，除了二者的第 12 层在树的深度（depth）这一特征上的差异不显著之外，其他的层 MacBERT 得分都显著高于 GPT2（$p<0.001$）。这可能与两种模型的训练方式不同有关。GPT2 作为一种自回归语言模型，其训练过程为从左到右的单向训练，通过上文来预测下文。而 MacBERT 作为自编码语言模型，其训练过程中可以同时用到上下文的信息。这可能是导致两个模型之间差异的重要原因之一，即，根据上文足以预测下文的语义，而句法信息则需要更完整的上下文才能较好的学到。然而，这一观点仍需要在更多的自回归语言模型和自编码语言模型上进行验证。此外，本文所使用的 GPT2 模型为字符级别的模型，而 MacBERT 在训练过程中用到了 n-gram mask，这对于汉语这种以词汇作为最小可独立运用句法单位的语言，可能会对模型学习句法信息的能力有一定影响。

在 MacBERT 和 GPT2 模型内部，二者的变化趋势比较一致，语义得分最高的层一般为第三层或第四层，而句法得分最高的层大多为第七层或第八层。在句法上的这一发现与之前的一些工作结论比较一致[17,18]，然而在语义特征上，本文的结果与之前的工作存在差异。Jawahar 等[18]通过动词/名词随机替换敏感程度（sensitivity to random replacement of a noun/verb, SOMO）等需要语义信息的任务，发现 BERT 的高层相比低层能更好的编码语义信息。这与本文的发现正好相反。但是值得指出的是，Jawahar 等的研究中是通过提出需要语义信息的任务来间接分析文本表示中是否编码了语义信息，而本文是采用直接定义的语义信息。而且前者所探究的语义信息是在句子层面，而本文所关注的语义信息则是在词汇层面。这也可能是导致差异的原因。

综上，本文通过对多种文本表示模型所编码的语义和句法特征进行对比分析，发现上下文相关模型在语义和句法任务上的表现都要优于上下

文无关向量；同样作为上下文相关向量的MacBERT 编码语义信息的能力与 GPT2 没有显著差异，而编码句法信息的能力则显著优于GPT2；MacBERT 和 GPT2 的第三层和第四层编码语义信息的能力最好，而第七层和第八层编码句法信息的能力最好，说明模型不同层在编码不同特征的能力上出现了分化。

### 4.2 文本表示预测大脑活动

图 7 展示了所有文本表示在预测脑活动上得分的差异。尽管在不同被试上的表现存在差异，但是显著性检验结果表明，Word2Vec 的脑活动预测得分显著高于 GloVe（$p<0.001$），MacBERT 和 GPT2 的 3-12 层得分显著高于 Word2Vec 和 GloVe（$p<0.001$）。而在 MacBERT 和 GPT2 内部，高层的得分显著高于底层，MacBERT 中第 9 层和第 10 层的得分最高，GPT2 中第 7 层到第 10 层的得分显著高于其他层。这一结果也与之前工作中的发现相符合[2]，即模型的中间层与脑活动最相似。单从这一结果上来看，文本表示模型编码句法特征的能力与预测脑活动的能力呈现出较为相近的规律，即：编码句法信息较好的层，脑活动得分也越高。

为了进一步量化分析文本表示编码不同语言特征的能力与预测脑活动能力之间的关系，本文将所有被试的脑活动预测得分平均，计算了所有文本表示预测语言特征的得分和大脑活动平均得分之间的相关性，结果如图 8（左）所示。可以看出，句法特征预测得分与脑活动预测得分的相关性普遍较高一些，在所有特征中相关性最高的前两种特征都是句法特征，分别为树的深度（depth），自上而下句法分析步数（nc_td），其次是句法特征自下而上句法分析步数（nc_bu）和语义特征视觉性（vision）。值得注意的是，上述提到的三种句法特征都来自成分句法的短语结构树。而由依存句法提取的依存距离（depdist）这一特征的相关性较低。本文并不认为这说明大脑的句法处理机制为成分句法分析，因为已有研究已经多次表明受到大脑认知资源（如工作记忆）的限制和提高效率的要求，依存距离最小化是人类语言的一个重要特征[37]。事实上，本次实验中文本表示的依存距离预测得分比较低，因此可能是单词的文本表示中没有较好的编码这一特征导致其得分与脑活动得分相关性较低。此外，左角句法分析步数（nc_lc）得分与大脑得分的相关性也比较低。在图 4 中可以看出所有模型及内部的每一层在这一特征上的得分都比较高，而且比较接近，说明这些文本表示均较好的学到了这一特征，因此这一特征不是导致文本表示与脑活动相似性差异的重要原因。

其他的语义特征预测得分和脑活动预测得分的相关性虽然相对较低，但相关系数也都在 0.4 以上，说明文本表示中编码的语义特征也会影响其与脑活动表征的相似性。因此，本文将所有语义特征的得分平均，将所有句法特征的得分平均，计算语义和句法特征平均得分和脑活动得分的相关性，结果如图 8（右）所示。可以看出，整体来看，文本表示中编码的句法特征是影响其与脑活动相似程度的主要原因。

## 5 总结

随着分布式文本表示在大脑语言机制研究中的深入使用，了解影响分布式文本表示和大脑语言表征相似性的因素变得尤为重要。因此，本文选择了四种经典的文本表示模型，将基于大脑语言功能的可解释语义和句法特征作为中介，通过分析发现文本表示编码句法特征的能力是影响其与大脑语言表征相似性的主要原因。这一发现表明在使用文本表示研究大脑的语言机制时应该慎重选择文本表示模型，尤其在研究大脑句法处理机制时应选择编码句法特征较好的文本表示，否则可能会影响结论的准确性。

本文的实验存在一定的局限性。首先，本文的结论基于本文所选择的 6 种语义特征和 5 种句法特征得出，其中语义特征且均为人为标注，尽管包含了大脑的基本功能划分，但仍然不能涵盖全部的语义特征，句法特征也仅代表了构建句法结构的难度，因此结论能否推广到更多的语义和句法特征上仍需进一步实验验证；其次，本文的实验是在全脑水平上进行的分析，而大脑中存在不同的功能分区，将来的工作可在脑区级别上进行进一步的分析；最后，本文实验针对的语言为汉语，不同语言间是否存在类似的结论仍需要进一步的研究。

## 参考文献

[1] Gazzaniga, Michael S., et al. 认知神经科学——关于心智的生物学（第 3 版）[M]. 北京：中国轻工业出版社，2019

[2] Gauthier, Jon, and Roger Levy. Linking Artificial and Human Neural Representations of Language. [C]//Proceedings of the 2019 Conference on Empirical

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 529–539.

[3] Sun, Jingyuan, et al. Neural Encoding and Decoding with Distributed Sentence Representations [J]. IEEE Transactions on Neural Networks, 2021, vol. 32, no. 2, pp. 589–603.

[4] Manning, Christopher D et al. Emergent linguistic structure in artificial neural networks trained by self-supervision. [C]//Proceedings of the National Academy of Sciences of the United States of America vol. 117, no. 48, 2020, pp. 30046-30054.

[5] Pimentel, Tiago, et al. Information-Theoretic Probing for Linguistic Structure. [C]//Proceedings of the 58th Annual Meeting of the Association-for-Computational-Linguistics (ACL 2020) (Virtual), 2020, pp. 4609–4622.

[6] Pylkkänen, Liina. The Neural Basis of Combinatory Syntax and Semantics [J]. Science, 2019, vol. 366, no. 6461, pp. 62–66.

[7] Zhang, Yizhen, et al. Connecting Concepts in the Brain by Mapping Cortical Representations of Semantic Relations [J]. Nature Communications, 2020, vol. 11, no. 1, p. 1877.

[8] Shailee Jain, et al. Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). 2020, pp. 13738–13749.

[9] Wehbe, Leila, et al. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses [J]. PLOS ONE, 2014, vol. 9, no. 11.

[10] Sun, Jingyuan, et al. Towards Sentence-Level Brain Decoding with Distributed Representations. [C]//Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 1, 2019, pp. 7047–7054.

[11] Zhang, Xiaohan., et al. Probing Word Syntactic Representations in the Brain by a Feature Elimination Method. [C]//Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 10, 2022, pp. 11721-11729.

[12] Wang, Shaonan, et al. Probing Brain Activation Patterns by Dissociating Semantics and Syntax in Sentences. [C]//Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 5, 2020, pp. 9201–9208.

[13] Schrimpf, M., et al. Artificial Neural Networks Accurately Predict Language Processing in the Brain. bioRxiv.2020.

[14] Charlotte Caucheteux, et al. Disentangling syntax and semantics in the brain with deep networks. [C]//Proceedings of the 38th International Conference on Machine Learning, PMLR vol. 139, 2021, pp. 1336-1348.

[15] Linzen, Tal, et al. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies [J]. Transactions of the Association for Computational Linguistics, 2016, vol. 4, no. 1, pp. 521–535.

[16] Conneau, Alexis, and Douwe Kiela. SentEval: An Evaluation Toolkit for Universal Sentence Representations. [C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[17] Hewitt, John, and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2019, pp. 4129–4138.

[18] Jawahar, Ganesh, et al. What Does BERT Learn about the Structure of Language. [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3651–3657.

[19] Reif, Emily, et al. Visualizing and Measuring the Geometry of BERT. [C]//Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 8592–8600.

[20] Binder, J. R. et al. Toward a brain-based componential semantic representation [J]. Cogn. Neuropsychology, 2016, vol. 33, pp. 130–174.

[21] 宗成庆. 统计自然语言处理（第 2 版）[M]. 北京：清华大学出版社，2013.

[22] Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael, Malach, et al. Syntactic structure building in the anterior temporal lobe during natural story listening[J]. Brain and Language, 2012, vol. 120, pp. 163–173.

[23] Richard Futrell, Kyle Mahowald, and Edward Gibson. Large-scale evidence of dependency length minimization in 37 languages. [C]//Proceedings of the National Academy of Sciences, 2015, 112:10336 – 10341.

[24] John Hale, David Lutz, Wen-Ming Luh, et al. Modeling fmri time courses with linguistic structure at various grain sizes. [C]//Proceedings of CMCL@NAACL-HLT. 2015.

[25] Reddy, A. J.; and Wehbe, L. Syntactic representations in the human brain: beyond effort-based metrics. In bioRxiv, 2021.

[26] Tenney, Ian et al. What do you learn from context? Probing for sentence structure in contextualized word representations. ArXiv, 2019.

[27] Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. Intrinsic Probing through Dimension Selection. [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. pp. 197–216.

[28] Dalvi, F., et al. What Is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models. [C]//Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, 2019, pp. 6309-6317.

[29] Pasquiou, A., et al. Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps. [C]//Proceedings of the 39th International Conference on Machine Learning, 2022, pp.17499-17516.

[30] Charlotte Caucheteux, Jean-Rémi King. Brains and algorithms partially converge in natural language

processing [J]. Communications Biology, Nature Publishing Group, 2022

[31] Glasser, Matthew F., et al. The Minimal Preprocessing Pipelines for the Human Connectome Project [J]. NeuroImage, 2013, vol. 80, pp. 105–124.

[32] Wang, Shaonan., et al. A synchronized multimodal neuroimaging dataset for studying brain language processing[J]. Sci Data 9, 590, 2022.

[33] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543

[34] Mikolov, Tomas et al. Efficient Estimation of Word Representations in Vector Space. [C]//International Conference on Learning Representations, 2013.

[35] Yiming Cui, et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing. [C]//Findings of the Association for Computational Linguistics: EMNLP, 2020, pp. 657–668

[36] Radford, Alec et al. Language Models are Unsupervised Multitask Learners. 2019.

[37] Futrell, Richard et al. Large-scale evidence of dependency length minimization in 37 languages. [C]//Proceedings of the National Academy of Sciences of the United States of America vol. 112,33, 2015, pp. 10336-41.

张肖寒（1995—），博士研究生，主要研究领域为自然语言处理与语言认知。
E-mail：xiaohan.zhang@nlpr.ia.ac.cn

王少楠（1990—），博士，副研究员，主要研究领域为自然语言处理、语言认知。
E-mail：shaonan.wang@nlpr.ia.ac.cn

宗成庆（1963—），博士，研究员，博士生导师，主要研究领域为自然语言处理。
E-mail：cqzong@nlpr.ia.ac.cn