

从神经编解码到语言认知计算

宗成庆

中国科学院自动化研究所

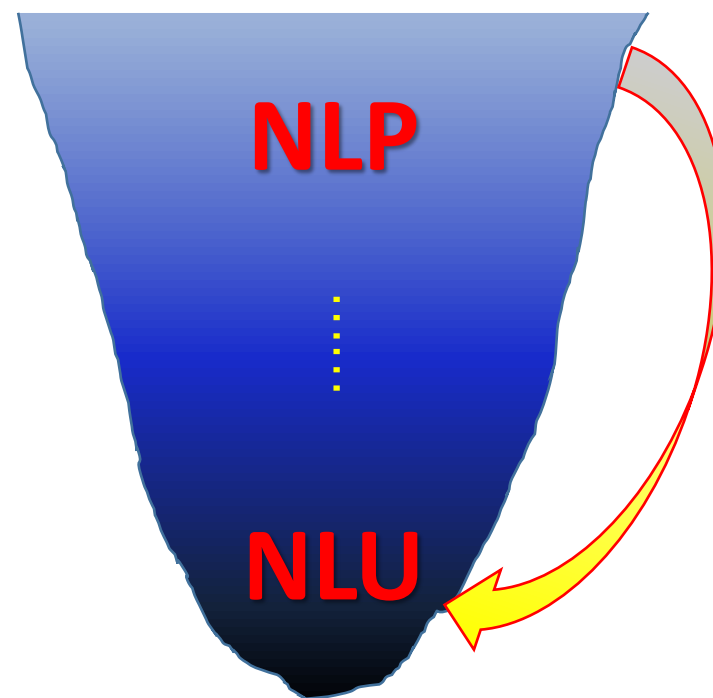
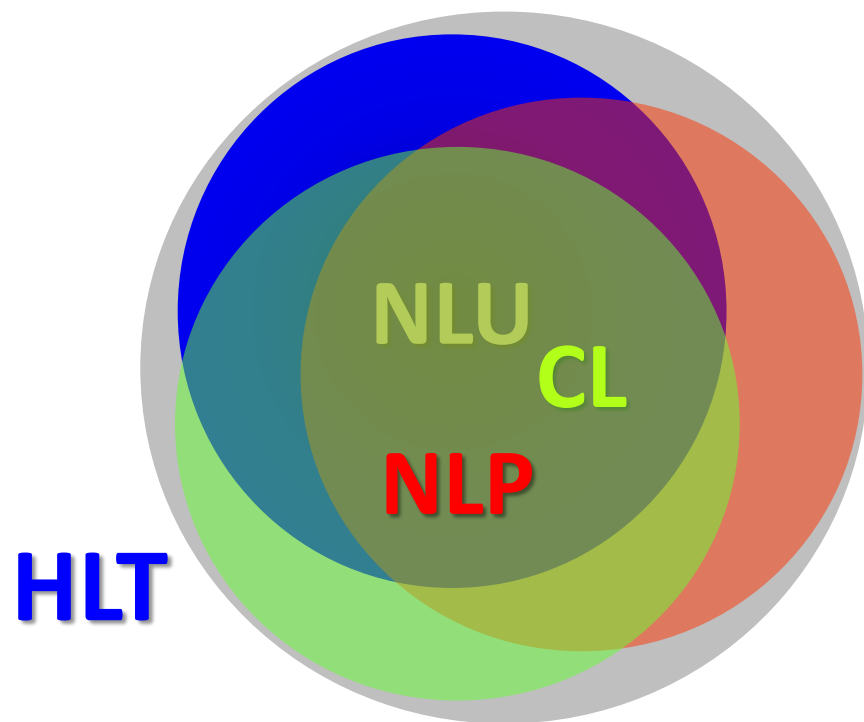
cqzong@nlpr.ia.ac.cn

提 纲

- ➡ 1. 问题的提出
- 2. 语言的表征与解码
- 3. 结束语

1. 问题思考

- 自然语言理解 (natural language understanding, NLU) (1956)
- 计算语言学 (computational linguistics, CL) (1960s)
- 自然语言处理 (natural language processing, NLP) (1970~80s)



- 人类语言技术 (human language technology, HLT) (1980s)

1. 问题思考

◆研究范式

- 大数据、大模型、大算力
- 多模态信息“一视同仁”
- 建立多任务统一框架

◆神经网络方法一统天下

- 分布式表示
- 自注意力机制
- 端到端的序列生成模型 (transformer)
- 预训练语言模型

A. Vaswani et al. [Attention Is All You Need](#).
Proc. of NIPS'2017, pp 6000–6010

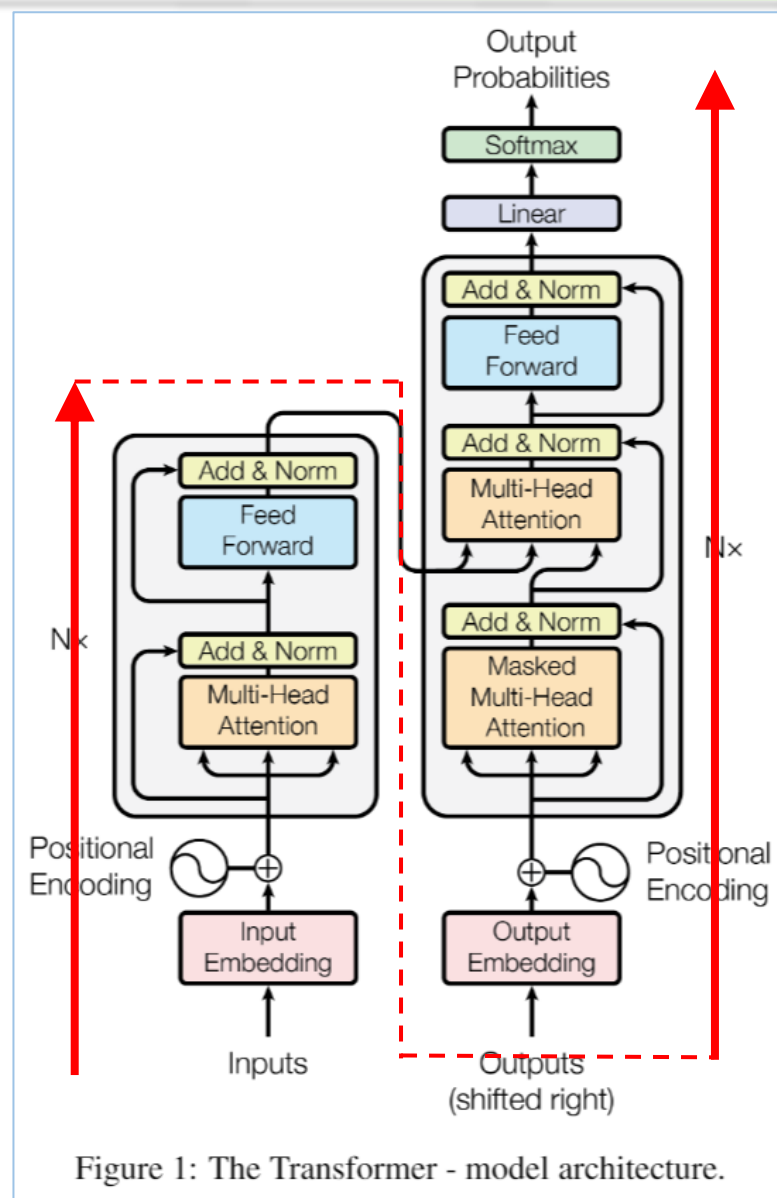


Figure 1: The Transformer - model architecture.

1. 问题思考

◆研究范式

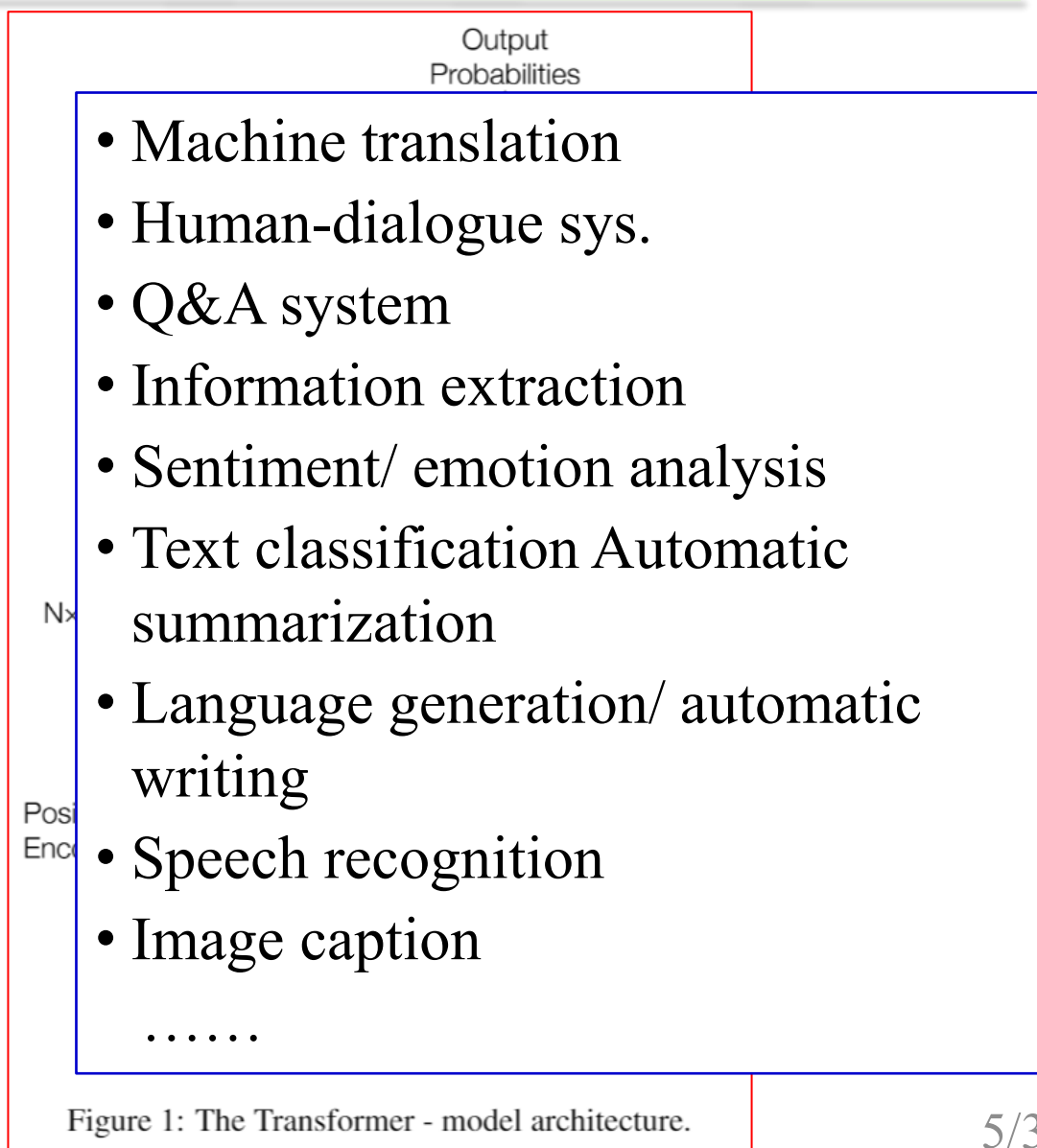
- 大数据、大模型、大算力
- 多模态信息“一视同仁”
- 建立多任务统一框架

◆神经网络方法一统天下

- 分布式表示
- 自注意力机制
- 端到端的序列生成模型 (transformer)
- 预训练语言模型

A. Vaswani et al. [Attention Is All You Need](#).

Proc. of NIPS'2017, pp 6000–6010



1. 问题思考

◆思考之一：大规模预训练语言模型是NLP的终结者吗？

➤ 模型名称：GPT-3

发布时间：2020, OpenAI

参数规模：1750 亿

测试地址：<https://beta.openai.com/examples>

➤ 模型名称：盘古α

发布时间：2021.4, 鹏城实验室、华为、北大

参数规模：2000 亿

测试地址：<https://pangu-alpha.openi.org.cn/>

➤ 模型名称：悟道2.0

发布时间：2021.6

参数规模：1980 亿

测试地址：<https://wudao.aminer.cn/os/>

➤ 模型名称：文心(ERNIE 3.0)

发布时间：2021.7, 鹏城实验室、百度

参数规模：2600 亿

测试地址：<https://wenxin.baidu.com/wenxin/ernie>

➤ 模型名称：源1.0

发布时间：2021.9, 浪潮

参数规模：2457 亿

测试地址：未开放

➤ 模型名称：紫东•太初

发布时间：2021.9

参数规模：千亿级

测试地址：<http://172.18.29.61:8888/#/main>

1. 问题思考

◆思考之二：NLP未来发展的方向在哪里？

往哪里走？



●**企业界**：拥有市场，赚钱是硬道理。

●**学术界**：是紧随企业的脚步亦步亦趋，还是另辟蹊径？

Green NLP !

1. 问题思考

◆思考之三：人脑是如何实现语言表征和理解的？

- 大脑是如何表征句法和语义信息的？
- 分布式表示模型编码的何种信息与人脑语言理解语言相关联？
- 分布式表示模型能帮助我们揭示不同脑区编码的语义差别吗？
- 不同类型的分布式表示模型与其对应文本的人脑神经表征的关系是什么？
- 不同的表示方法在人脑解码中的性能表现如何？
- 如何解释基于不同模态信息获得的语义表示？

.....

$$\begin{matrix} \text{山} \\ \begin{bmatrix} 0.15 \\ \cdot \\ \cdot \\ \cdot \\ 0.21 \end{bmatrix} \end{matrix} \quad ? \quad \begin{matrix} \text{药} \\ \begin{bmatrix} 0.23 \\ \cdot \\ \cdot \\ \cdot \\ 0.19 \end{bmatrix} \end{matrix} = \begin{matrix} \text{山药} \\ \begin{bmatrix} 0.18 \\ \cdot \\ \cdot \\ \cdot \\ 0.06 \end{bmatrix} \end{matrix}$$

天 * 狗 = 天狗？

- 100 – 47%
- 2500 – 97%+
- 4000 – 99%+

提 纲

1. 问题的提出

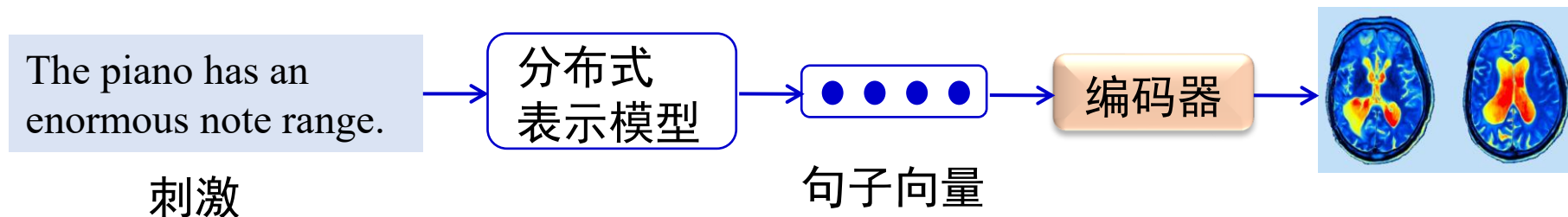
 2. 语言的表征与解码

3. 结束语

2. 语言的表征与解码

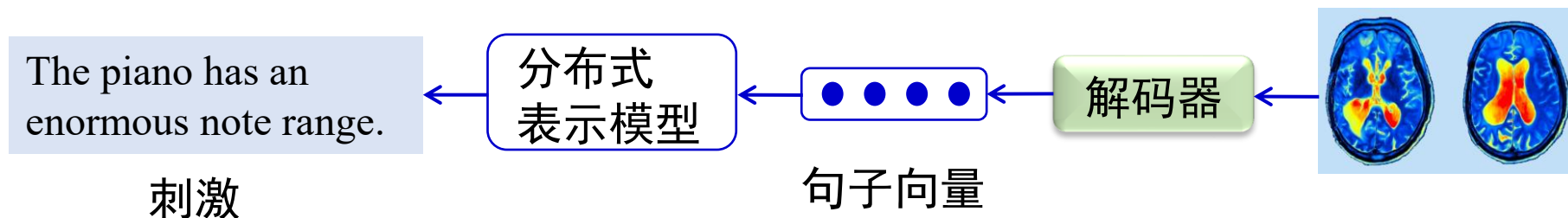
❖ 神经编码

目标：通过语言刺激的变化来预测对应各脑区、体素点神经活动，进而帮助解析人脑的语言表征机制。



❖ 神经解码

目标：从脑神经活动中重建出人所感知到的刺激，是脑机接口的计算基础之一。



2. 语言的表征与解码

◆工作1：分布式表示与人脑之间的关系

- 不同类型的分布式表示模型与其对应文本的人脑神经表征的关系是什么？
- 分布式表示模型编码的何种信息与人脑语言理解语言相关联？
- 分布式表示模型可以帮助我们揭示不同脑区编码的语义差别？

2. 语言的表征与解码

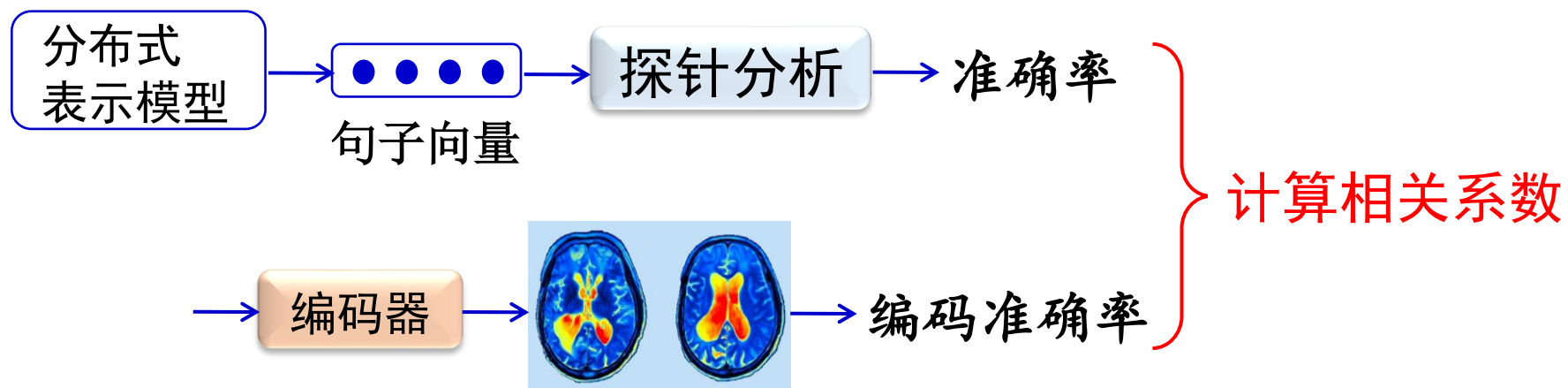
● 常用的分布式文本表示方法

模型分类		名称	结构基础
非结构化模型	朴素池化	① 平均池化(AVG)	平均池化
		② 最大池化(MAX)	最大池化
		③ 联合池化(AVMA)	拼接平均池化和最大池化
	复杂池化	④ SIF	加权平均池化
结构化模型	无监督模型	⑤ Skip-thought	LSTM
		⑥ FairSeq	CNN
		⑦ ELMo	CNN-BiLSTM
		⑧ BERT	Transformer
		⑨ GPT2	
		⑩ RoBerTa	
	有监督模型	⑪ GenSen	BiLSTM
		⑫ InferSent	CNN-BiLSTM

2. 语言的表征与解码

- 基于探针任务和消融测试的编解码分析算法

- 应用探针任务



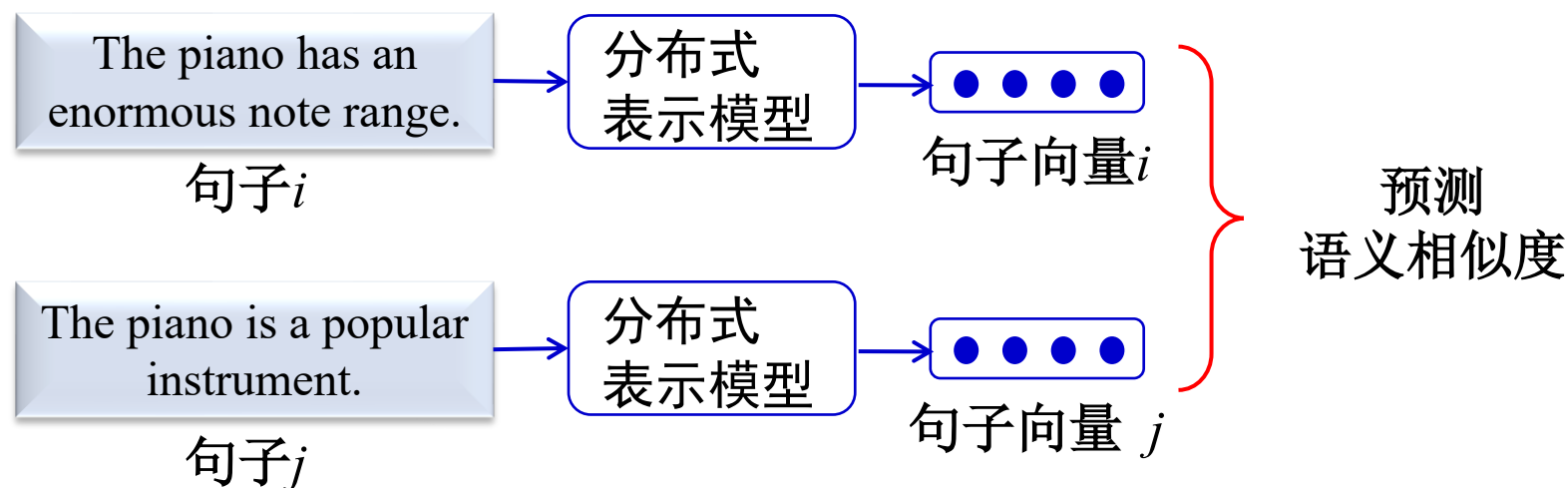
- 探针任务可针对语义、语法、形态学等句子特征设计；
- 通过探针任务，能够测试预训练分布式表示编码了何种特征，从而导致了在神经编解码中有对应的表现。

2. 语言的表征与解码

➤ 探针任务

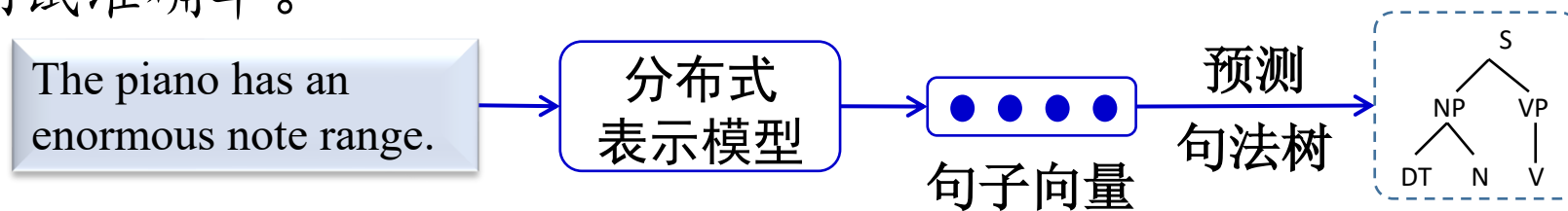
- 目的：测试分布式句子表示模型编码各种语言学特征的能力；
- 种类：我们使用了三种探针任务：

①语义探针任务：基于句子向量预测句子之间的语义相似度，与人工标注相似度对比作为测试准确率。

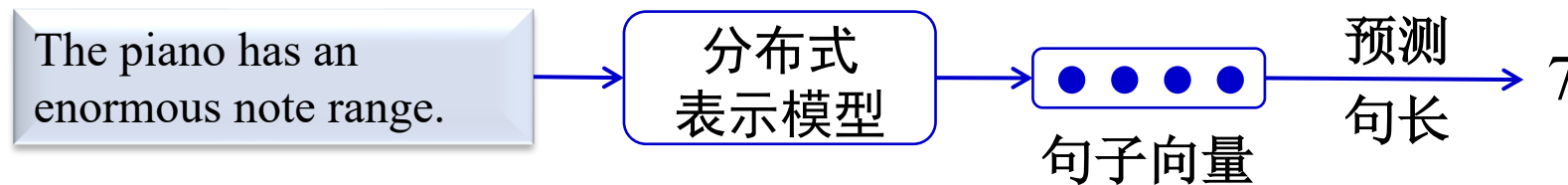


2. 语言的表征与解码

②句法探针任务：基于句子向量生成原句子的句法树，与**正确句法树的相似度对比**作为测试准确率。



③形态学探针任务：基于句子向量预测原句子的长度，与**正确长度对比**作为测试准确率。

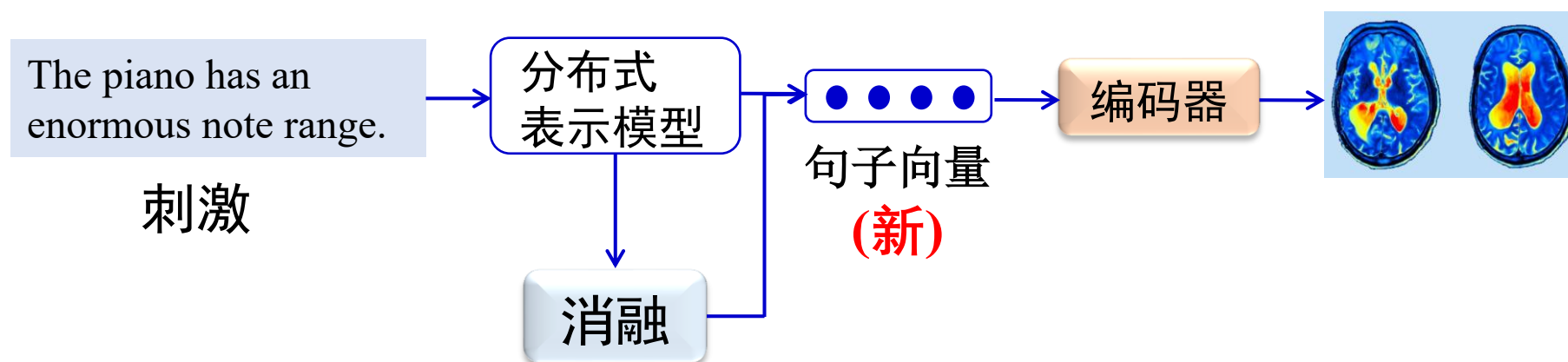


通过探针任务的准确率和表示模型编解码的准确率计算相关系数，从而探测出某种分布式表示编码了哪种语言特征，导致了在神经编解码中产生相应的表现。

2. 语言的表征与解码

➤ 消融任务

消融测试用于削弱模型编码某些特征的能力，以此验证探针任务的发现。

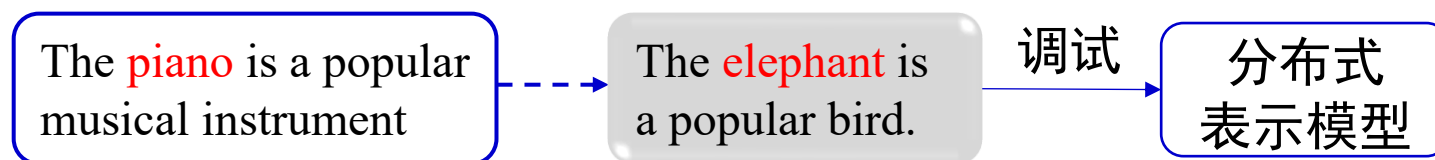


- **目的：**削弱分布式文本表示模型编码某个特定语言学特征的能力，将削弱后的表示模型再应用于神经编解码，观察神经编解码准确率的变化，即可验证探针任务的发现。

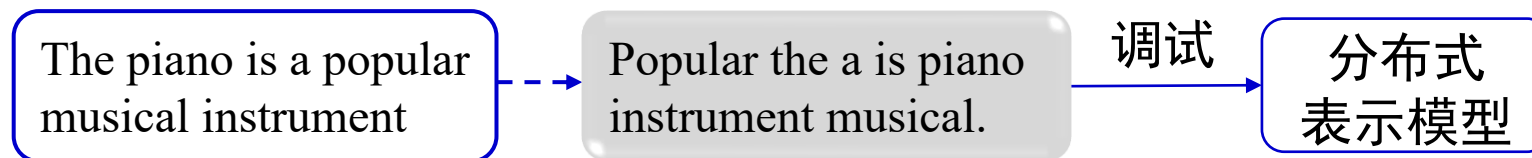
2. 语言的表征与解码

- 两种消融测试：

❶ 语义消融测试：在不出现句法错误的前提下**随机替换句子中的名词或动词**，并用替换后的语料调试(fine-tune)分布式表示模型；



❷ 句法消融测试：不更改句子中的词但是**打乱词序**，使得句子失去正确的句法结构。使用打乱后的语料fine-tune分布式表示模型。



2. 语言的表征与解码

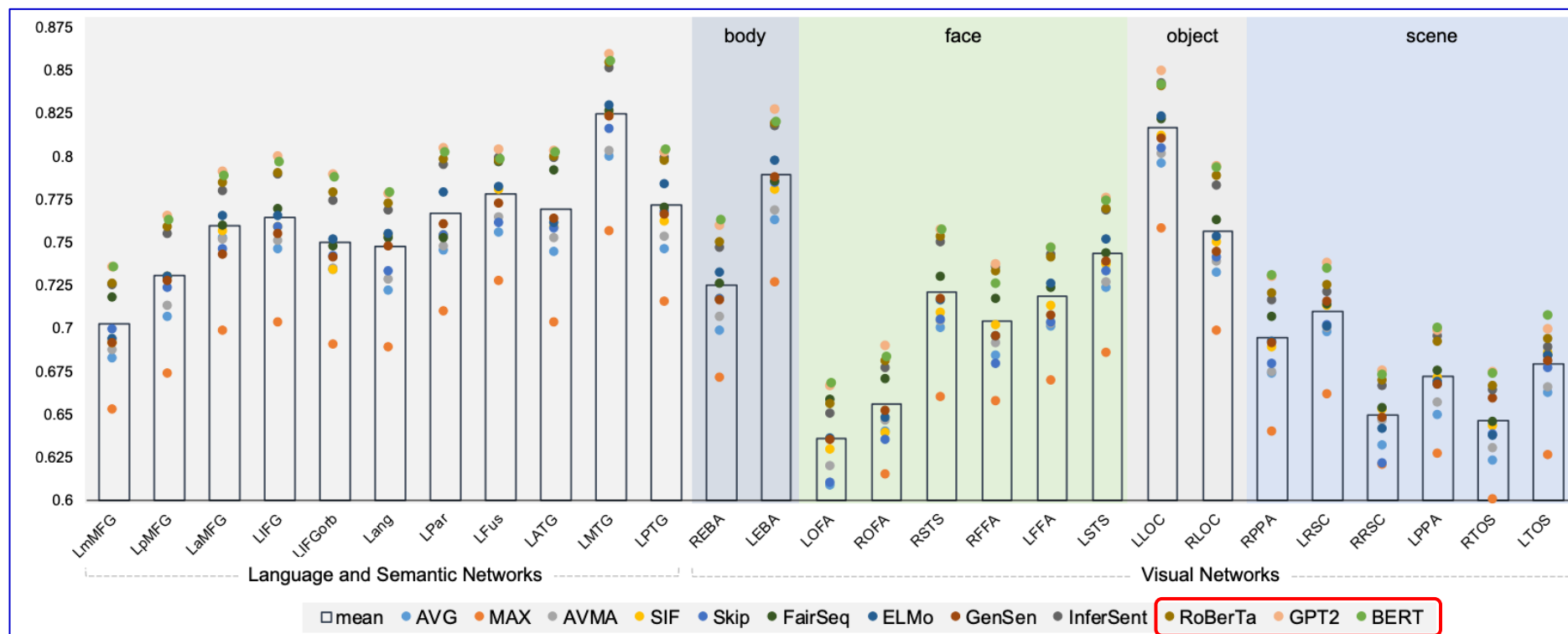
➤ 句子刺激结构及样例

Topic	Passages	Sentences
Musical Instruments	Piano	1. The piano is a popular musical instrument...
		2. Pressing a piano key causes a felt-tipped hammer...
		3. The piano has an enormous note range.
	Accordion	1. A clarinet is a woodwind musical instrument...
		2. It is a long black tube with a flare at the bottom
		3. The player chooses notes by pressing keys and holes.
	Clarinet	1. An accordion is a portable musical instrument.
		2. One keyboard is used for individual notes and the other for chords.
		3. Accordions produce sound with bellow that blow air through reeds
		

48个主题，每个主题3~5个段落，每个段落3~4个句子，每个句子长10~20个词；共637个句子。

2. 语言的表征与解码

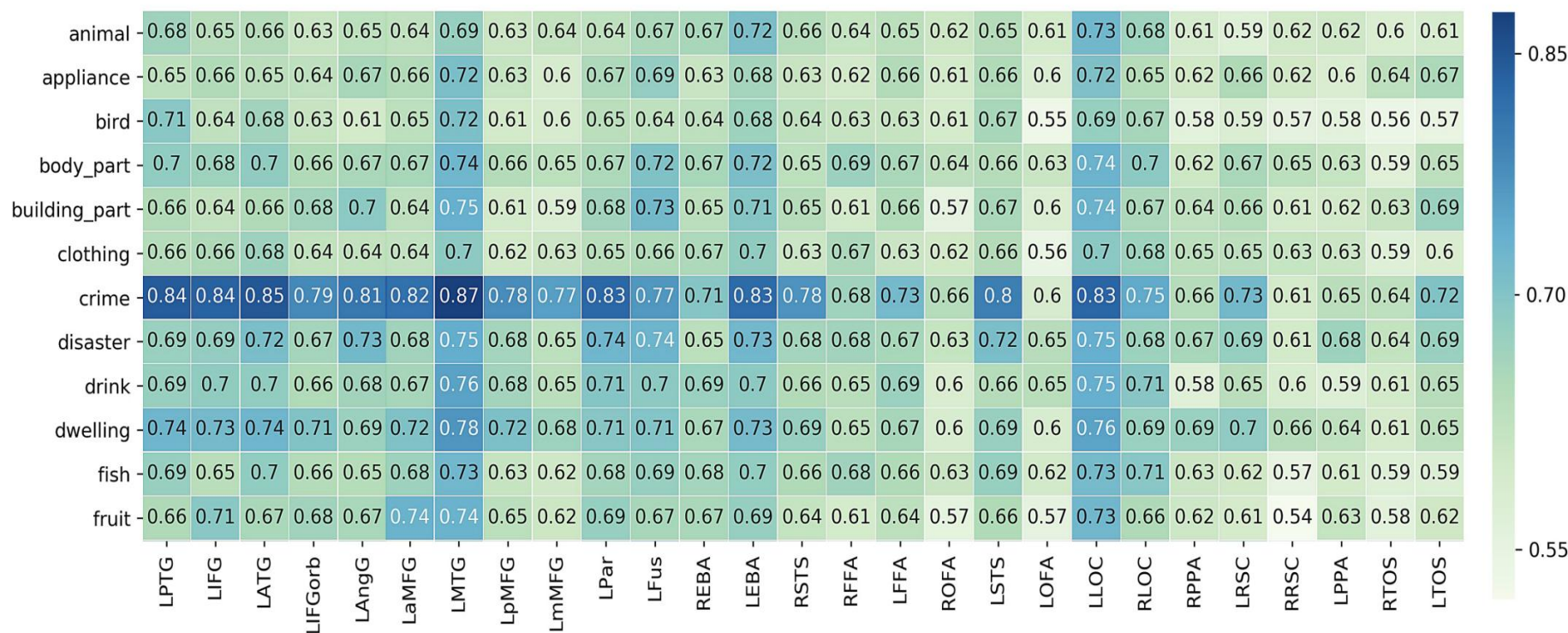
➤ 分布式表示在脑语言及视觉网络上的编码准确率



- 候选模型在大脑左颞叶中回(LMTG)和左枕叶复合体(LLOC)上编码表现突出;
- 基于Transformer 的模型在语言网络上的编码准确率显著优于其他候选模型;
- 应用最广泛的平均方法, 其编码表现仅排在所有候选模型的中游。

2. 语言的表征与解码

❖ 分布式表示模型的在不同脑区、不同主题上的编码准确率

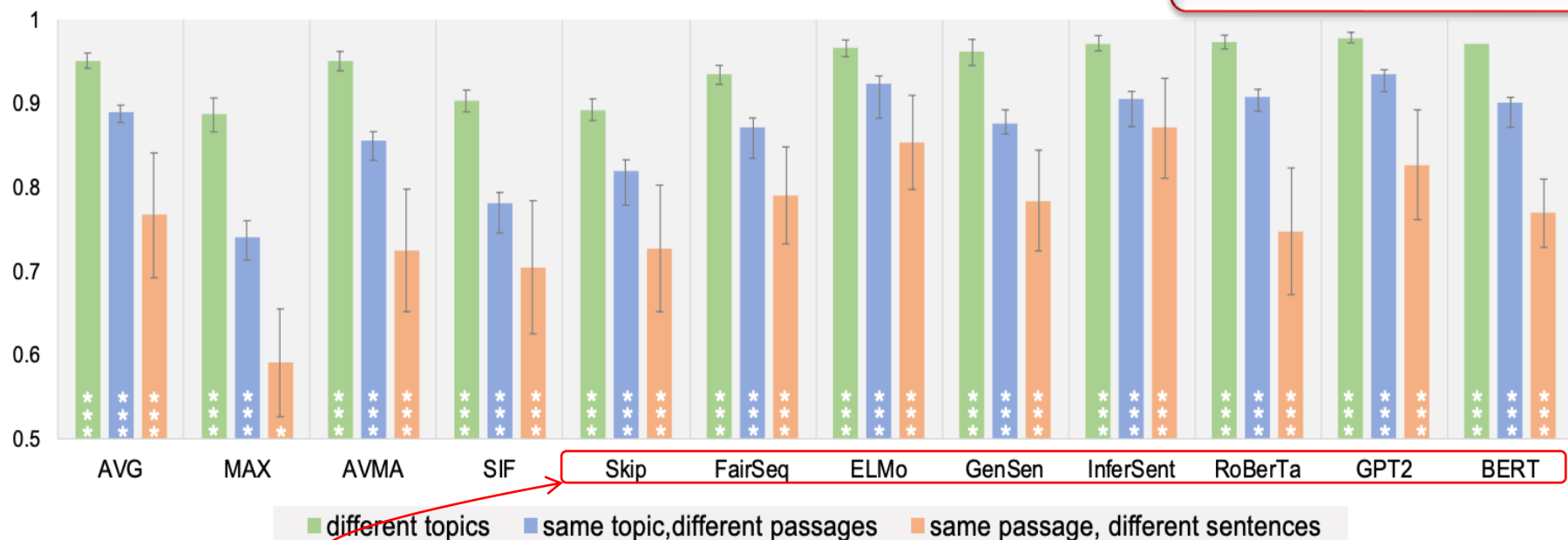


- 在主题层次上，左颞叶中回(LMTG)比其它语言和语义网络中的兴趣区(ROI)的预测准确率更高；左枕叶复合体(LLOC)比其他视觉网络中的ROI的预测准确率更高。
- 一个主题可以在多个ROI得到准确的预测，同时在一个特定ROI上可以预测出多个主题。这说明在人脑中，某个主题并不仅仅被一个或者几个脑区表示，而是广泛分布于多个脑区之中。

2. 语言的表征与解码

❖ 分布式表示模型的神经解码准确率

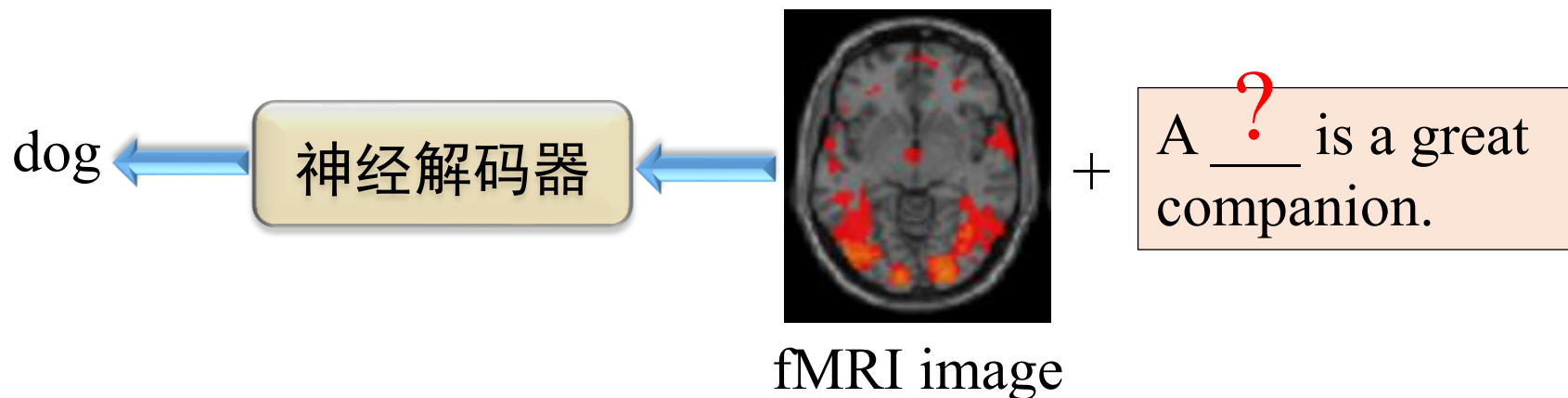
IEEE TNNLS 2020



- 在神经解码准确率上，结构化模型并不始终显著地优于非结构化模型；
- 结构化模型在处理细粒度的语义差别上效果较好，而非结构化模型在处理粗粒度的语义差别上可以取得较为满意的表现；
- 值得注意的是，有监督的结构化模型在细粒度的语义解码任务上表现突出；监督任务引入的额外语义信息如何帮助了解码任务，值得进一步探讨。

2. 语言的表征与解码

◆工作2：神经解码的可行性

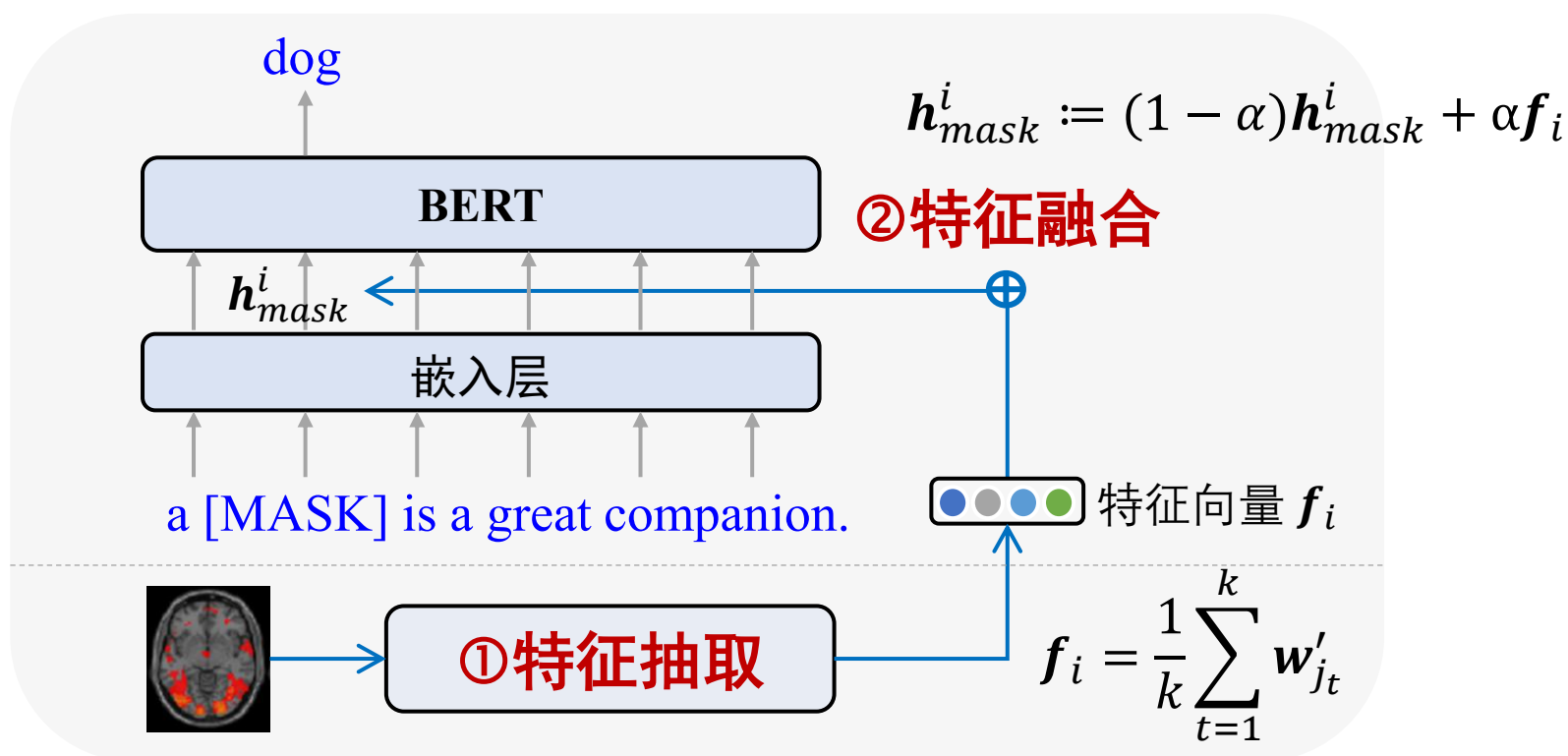


- 已有的工作大都将其看作成对分类任务
- **我们的思路：**引入上下文作为辅助信息，对神经影像进行直接词语预测，将其形式化为跨模态完形填空任务：**完形填空任务 + 直接预测任务**
 - 零样本学习：能够预测未出现在训练集中的词语；
 - 单样本学习：一个词语只有一张神经影像与之对应。

2. 语言的表征与解码

● 神经解码方法

将神经影像转换为特征向量，输入到预训练语言模型BERT中进行目标词语预测。



2. 语言的表征与解码

● 实验数据

➤ 开源fMRI数据集

fMRI60

- 9个被试
- 60个英文词语

carrot, dog, hammer, igloo, skirt



dog

fMRI180

- 15个被试
- 180个英文词语

ability, big, damage, experiment, seafood

A **dog** is a great companion.

gerbil cat
dog
 chow
animal pet

fMRI60: Tom M. Mitchell et al. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, pages 1191–1195

fMRI180: Francisco Pereira et al. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963–963

2. 语言的表征与解码

➤ 构造跨模态完形填空数据集 ({神经影像, 上下文}, 词语)

- 上下文构造: 为每个词语构造6段上下文 (4~13个词, 平均7个词)

	词语	上下文
fMRI60_CMC	carrot	the [MASK] is his favorite vegetable.
	hammer	she puts the [MASK] down on the ground.
fMRI180_CMC	ability	he has the [MASK] to cultivate creativity.
	damage	the accident left some serious [MASK].

- 数据集统计:

	被试	样本量	词语	上下文	神经影像
fMRI60_CMC	9	360	60	360	60
fMRI180_CMC	15	1080	180	1080	180

2. 语言的表征与解码

● 实验设置

- 数据划分: 10折交叉验证: 8折训练、1折验证、1折测试
- 体素选择: 对fMRI影像选择信息度最高的5000个体素 → 5000维fMRI向量

● 模型设置

- 基础模型架构为BERT-base, 不微调, 词表为语料中的词语;
- 跨模态映射使用岭回归, 中间词向量使用BERT深层词向量;
- 对所有被试, 通过验证集调整超参数为 $k = 5$, $\alpha = 0.7$ 。

2. 语言的表征与解码

● 主要实验结果

(%)	fMRI60_CMC		fMRI180_CMC	
	Top-1 acc	Top-5 acc	Top-1 acc	Top-5 acc
BERT	27.50	45.56	26.02	50.93
BERT-Direct Fusion	27.78	49.54	26.51	51.78
BERT-Retri Fusion (random)	24.81	44.01	25.91	50.83
BERT-Retri Fusion	31.08(+3.58)	55.99(+10.43)	27.60(+1.59)	53.11(+2.19)

● 结论：

- fMRI影像中抽取的特征向量携带了语义信息，可指导BERT预测目标词语；
- 我们的方法可以作为跨模态完形填空任务的一个强基线模型；
- 在给定上下文的情况下，将fMRI影像预测到大词表中的词语是可行的。

2. 语言的表征与解码

◆工作3：大脑语义编码的网络结构

- 大脑理解语言时是多脑区同时被激活的，如何通过影像数据解析网络结构与语义编码之间的关系，构建自然语言的语义图谱？
- 如何构建汉语的语义图谱？
- 在语义图谱的基础上，如何发现语义组合、概念形成及演变的规律？

2. 语言的表征与解码

◆ 数据采集

(1) 磁共振成像仪



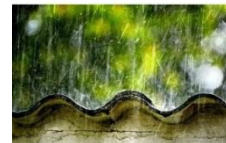
每个词看3s,间隔2s(十字空屏)。每个人每次看2小时, 672个词全部看完。

➤ 实验刺激

选取《同义词词林》中每个类别的代表性词汇, 包括名词、动词、形容词, 共计672个, 每个词汇呈现6遍。已采集11个被试8h数据和9个被试2h数据。



下雨



下雨



下雨



下雨



下雨

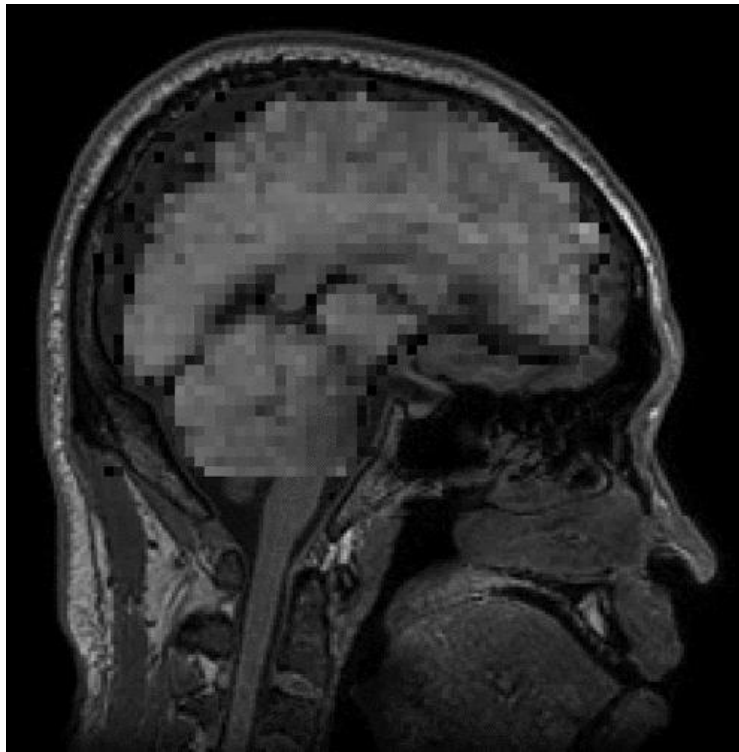


下雨

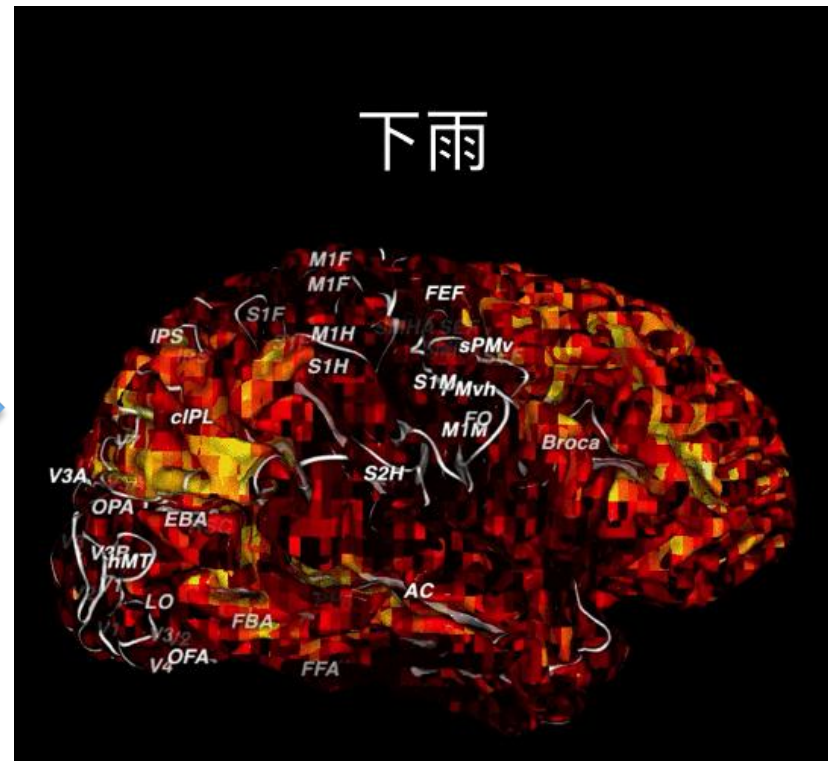
2. 语言的表征与解码

● 观察

➤ 血流动力学响应

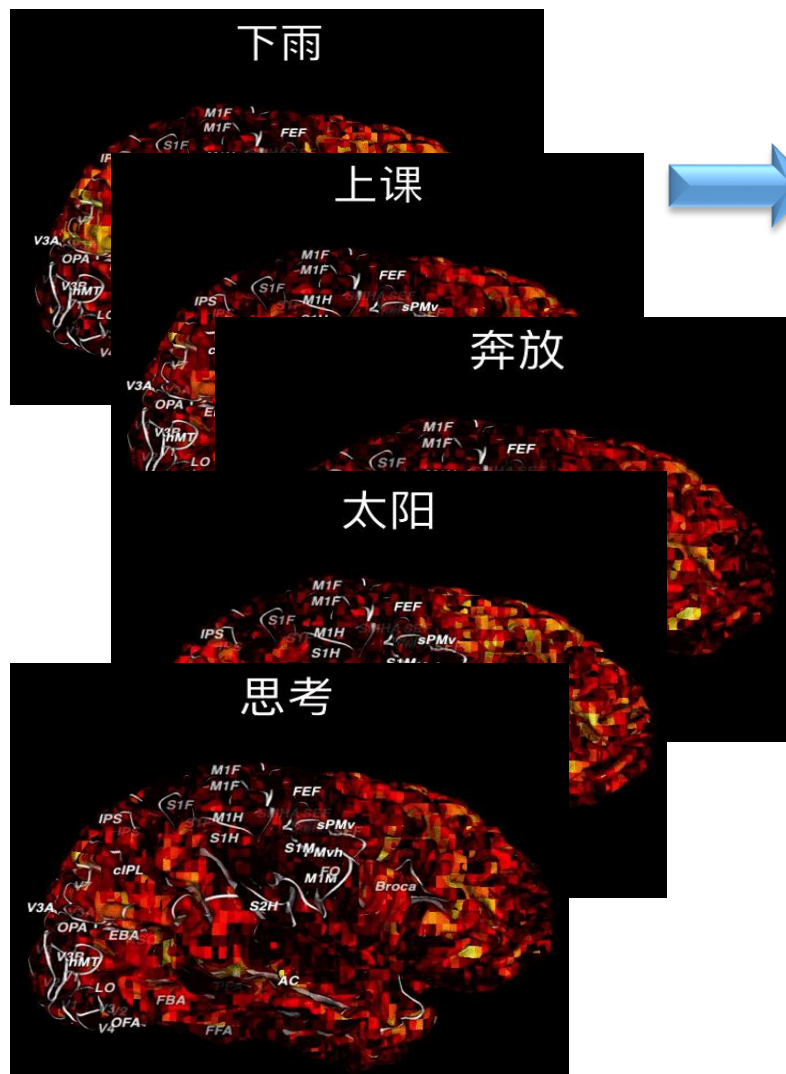


➤ 词汇概念响应

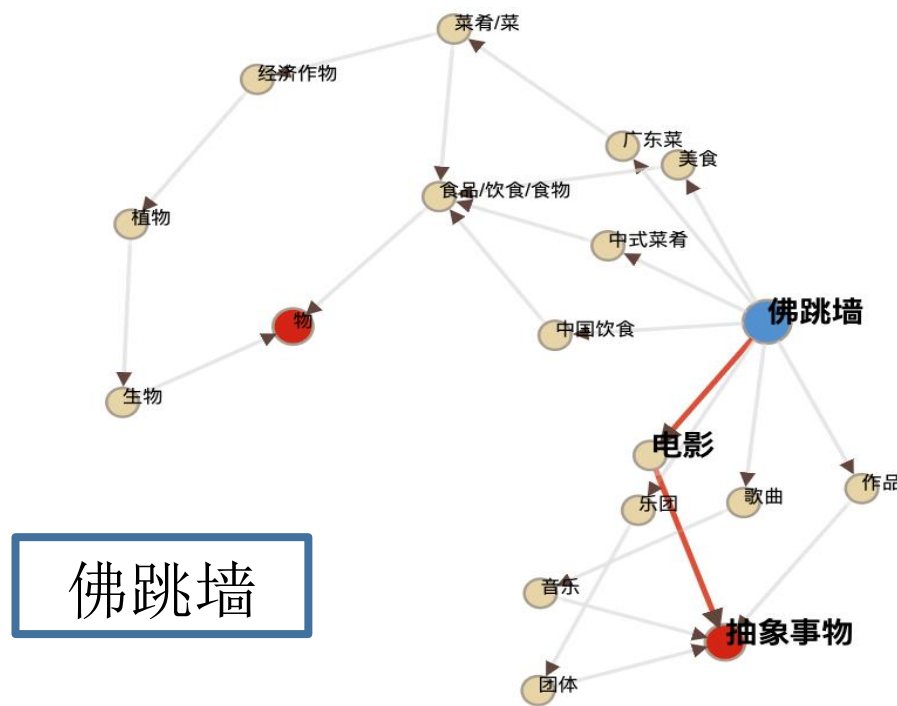


2. 语言的表征与解码

• 分析



- 大脑每个体素编码了什么语义信息？
- 不同(类型)词汇的大脑表征有何异同？
- 词汇关系是如何在大脑中编码的？
- 大脑语义编码与语言学中概念表征的异同在哪里？



佛跳墙

2. 语言的表征与解码

(2) 脑磁图仪

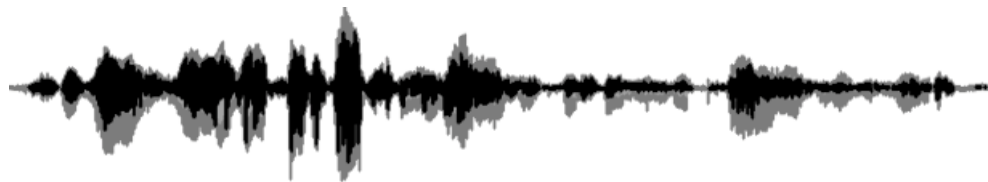


+ 磁共振成像仪



➤ 实验刺激

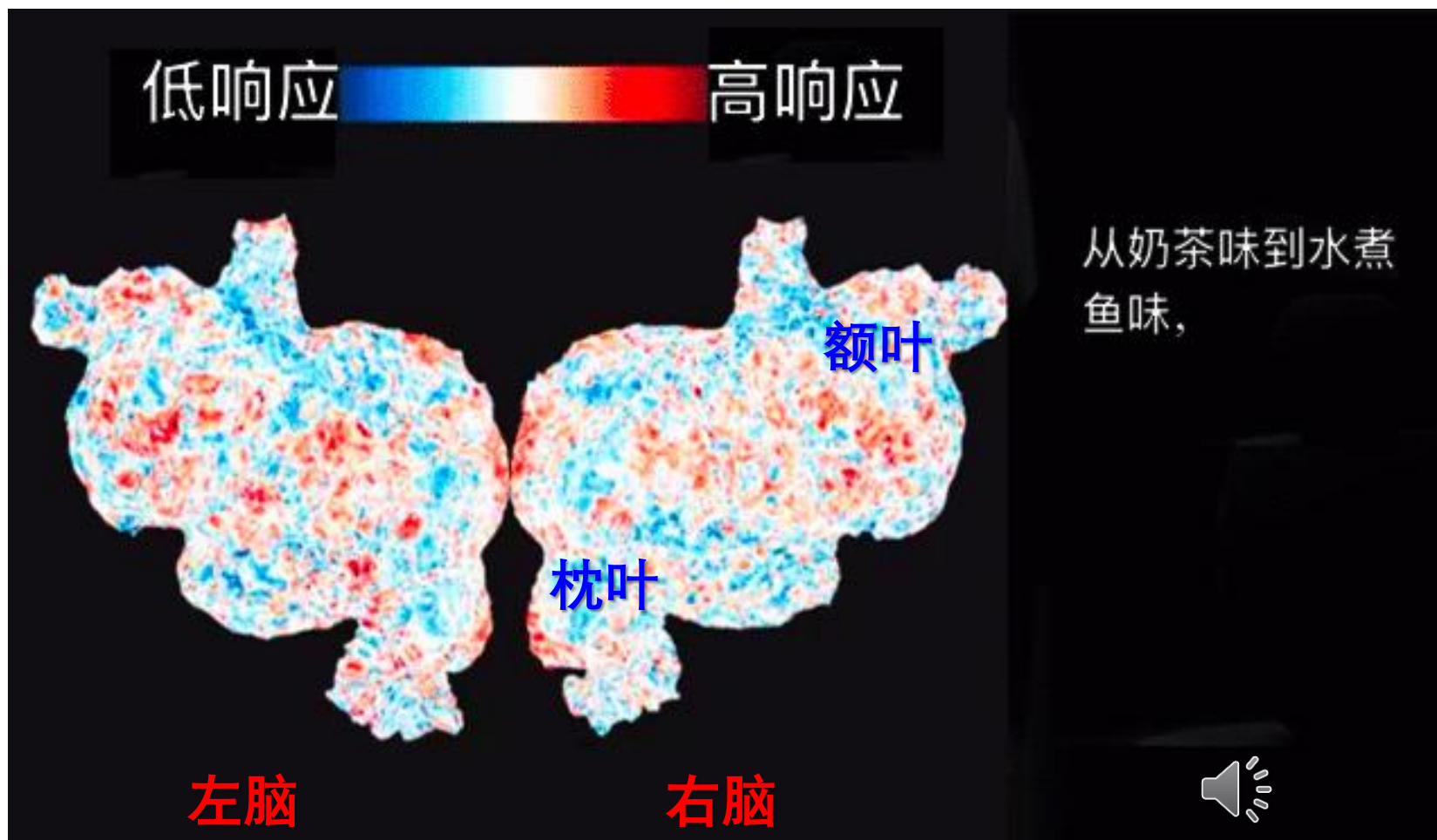
- 60个中文故事，长度在600~900个词之间，总计40145个中文单词，词汇量为10152，每个词出现的频次范围为1-210；
- 音频摘自《人民日报评论》；
- 已采集12个被试的9个小时fMRI和6个小时MEG数据。



“从奶茶味到水煮鱼味，从折耳根味到钵钵鸡味，刷新味蕾的不仅只有月饼、酸奶，薯片也爆款频出，甚至一袋难求...”

2. 语言的表征与解码

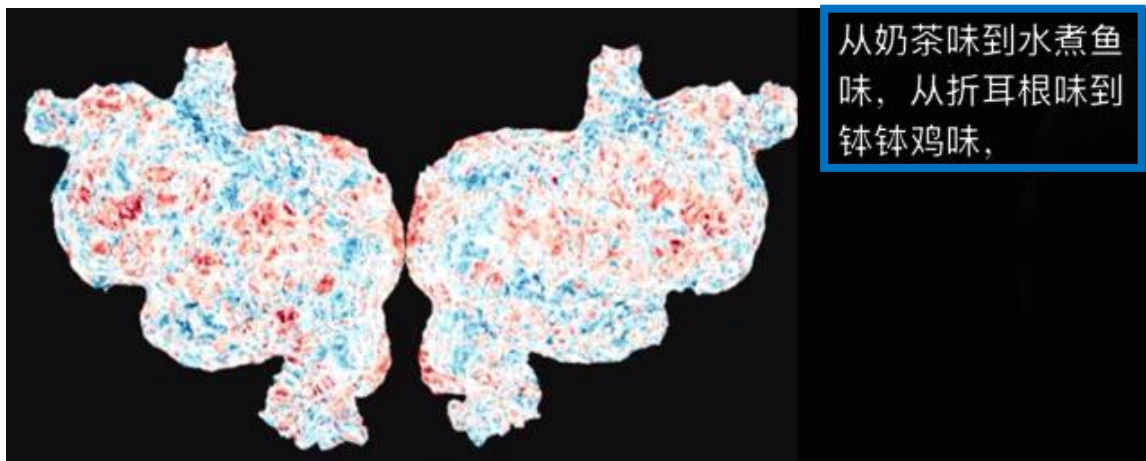
- 故事理解响应



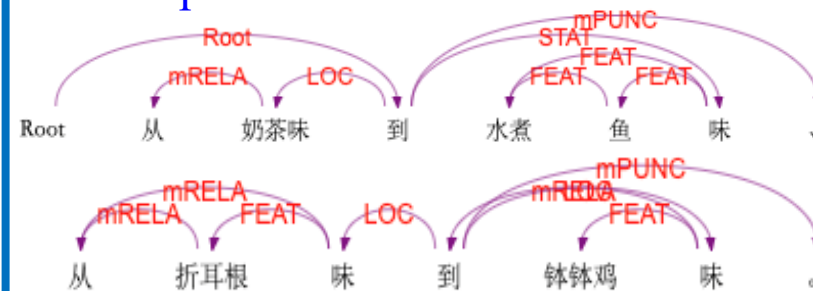
大脑皮层展开后示意图

2. 语言的表征与解码

● 分析



从/p 奶茶味/n 到/p 水煮/n 鱼/n 味/n 、
/wp 从/p 折耳根/n 味/n 到/p 钵钵鸡/n 味/n 。 /wp



- 不同语言特征（词义、词法、句法、篇章关系等）是如何在大脑中编码的？
- 不同语言特征的时间响应关系是什么？
- 大脑故事表征过程与通用语言模型之间有什么联系？

[1]Shaonan Wang et al. A synchronized multimodal neuroimaging dataset for studying brain language processing. *Nature- Scientific Data*, (2022) 9: 590. [<https://doi.org/10.1038/s41597-022-01708-5>]

[2]Shaonan Wang et al. An fMRI Dataset for Concept Representation with Semantic Feature Annotations. *To appear in Nature- Scientific Data*. [<https://doi.org/10.1038/s41597-022-01840-2>]

提 纲

1. 问题的提出
2. 语言的表征与解码
- ➡ 3. 结束语

3. 结束语

- 数据驱动的NLP方法和大规模预训练语言模型将在实际应用中发挥巨大作用。
- 依靠数据和算力的自然语言处理能力存在性能上限，在某些应用任务上或许可以解决80~90%的问题，但剩余10~20%的攻坚问题，如果没有类脑智能的参与，是不可想象的。
- 以节约资源和能源，建立低碳环保的绿色NLP技术需要得到应有的关注和重视，学术界应成为该方向的先锋。
- 探索人脑理解语言的认知机理、研发类脑语言智能技术必将成为数据驱动技术未来发展的重要方向。
- 实现真正的语言智能，从感知到认知，从计算到推理，从数据浅层处理到深层语义理解，建立可解释、能推理、不需要超大规模数据支撑的语言智能技术是未来发展的必然趋势。

谢谢!
Thanks!



个人主页: <http://www.nlpr.ia.ac.cn/cip/cqzong.htm>