

ChatGPT 等大模型的语言处理机制 及其理论蕴涵

袁毓林^{1,2}

(1. 澳门大学 中国语言文学系, 澳门 519000; 2. 北京大学 中文系, 北京 100080)

摘 要: 本文简单解释 ChatGPT 等现代大型语言模型的语言处理机制和数学原理及其理论蕴涵。首先, 通过 ChatGPT 对代词指称歧义句的理解, 来说明语言大模型在语义理解和常识推理方面的性能; 接着, 介绍转换器这种语言模型的新颖构架, 特别是其注意力机制及其功能; 还介绍基于分布语义学的词的向量化嵌入表示, 以及词向量在语言处理和类比推理方面的作用; 然后详细介绍转换器模型怎样通过注意力机制和前馈网络, 来追踪和传递词语之间的句法语义关系信息, 从而成功地预测下一个词语并生成合适的文本; 最后简介语言大模型的训练方式, 并说明大模型怎样用“再造语言”的方式, 帮助我们重新认识人类自然语言的有关特点(分布性和预测性), 启发我们反思既有的句法学和语义学理论。

关键词: ChatGPT; 语言大模型; 转换器; 注意力机制; 前馈网络; 词向量

ChatGPT and Other Large Language Models: Their Language Processing Mechanisms and Their Theoretical Implications

YUAN Yulin^{1,2}

(1. Department of Chinese Language and Literature, University of Macau, Macau 519000, China;
2. Department of Chinese Language and Literature, Peking University, Beijing 100080, China)

Abstract: This paper briefly explains the language processing mechanisms, mathematical foundations and theoretical implications of ChatGPT and other modern large language models. Firstly, it demonstrates the performance of large language models in semantic understanding and common sense reasoning by testing ChatGPT's understanding of ambiguous sentences. Secondly, it introduces the transformer, which is equipped with what is referred to as multi-headed attention (MHA) and functions as a novel module of these large language models. Additionally, it presents word embedding and real-valued vector representations based on distribution at semantics, as well as the role of word vectors in language processing and analogical reasoning. Thirdly, it details how transformers successfully predict the next word and generate appropriate texts by tracing and passing on syntactic and semantic relationship information between words through multi-headed attention (MHA) and feed forward network (FFN). Finally, it provides an overview of the training methods of large language models and shows how they use the method of “recreating a language” to help us re-assure relevant design features (including: distributivity and predictability) of human natural languages and to inspire us to re-examine various syntactic, semantic theories that have been developed and formulated so far.

Key words: ChatGPT; large language models; transformer; multi-headed attention (MHA); feed forward network (FFN); word vectors

1. 引言: 了解大模型的语言运用能力的来源

大家可能都知道, ChatGPT 等现代大型语言模型 (large language models, LLMs) 在语言理解与生成方面有强大的功能; 但是, 未必了解它们的工作方式、内部机制和数学原理。那么, ChatGPT 等语言大模型到底是怎么达到这种境界的? 能不能用一种语言学家也能够理解的方式来简单地解释一下? 大模型的这种语言处理机制在语言学理论方面有什么含义? 这也许是一些语言学工作者 (包括研究生和教师) 甚至一般的网友所普遍关心的问题。

本文尝试根据笔者有限的文献阅读和粗浅的专业知识, 来介绍和回答这些问题。为了方便, 我们将主要集中在两个方面: (一) 基于分布语义学的词向量表示及其类比推理功能, (二) 基于转换器的语言大模型的注意力机制和前馈网络的工作机理。

2. ChatGPT 的语义理解与常识推理水平例示

任何一个语言学家, 只要亲自去测试一下, 就会留下如此深刻的印象: ChatGPT 在语义理解和常识推理方面, 简直有接近人类水平的表现。比如, 代词和先行语的回指 (anaphora) 规律是什么? 这种回指规则是纯粹句法的、语义的、语用的、认知的, 还是兼而有之的? 这在当代语言学理论上长期存在争论, 可以说是意见纷纭、莫衷一是。(详见徐烈炯 1995: 227-253) 但是, ChatGPT 却可以在并不理会诸如此类的理论假设的情况下, 像一个实干家一样, 能够比较准确地识别有关代词的先行语, 得出代词的正确指称意义解读; 并且, 还能够说明作出这种判断的理由。(详见袁毓林 2024a, b) 比如, 下面的 (1)-(3) 是我们用维诺格拉德模式 (Winograd Schema, WS) 句子, 对它进行测试的部分例子:^①

(1) a. Pete envies Martin because he is very successful.

Q: Does he refer to Pete?

Q: Does he refer to Martin?

ChatGPT: In this sentence, “he” most likely refers to Martin. The phrase “Pete envies Martin because he is very successful” indicates that the reason for Pete’s envy is the other person’s success, suggesting that “he” is referring to Martin.

So: Q: Does “he” refer to Pete? No.

Q: Does “he” refer to Martin? Yes.

b. Pete envies Martin because he is very unsuccessful.

Q: Does he refer to Pete?

Q: Does he refer to Martin?

ChatGPT: In this sentence, “he” most likely refers to Pete. The phrase “Pete envies Martin because he is very unsuccessful” suggests that despite his lack of success, Pete is envious of Martin, which could be for various reasons not stated in the sentence.

So: Q: Does “he” refer to Pete? Yes.

Q: Does “he” refer to Martin? No.

^① 我们参考了 Levesque et al. (2012) 和 Bernard and Han (2020) 等文献对于维诺格拉德模式句对的设计。

可见, ChatGPT 能够利用“envy”的语义(希望自己在才能、荣誉、地位、境遇等方面跟某人一样好,但是没有达到,所以内心怨恨该人)及相关的常识:更加可能的情况是不成功的人对成功的人心怀怨恨,来确定到底“(un) successful”的是嫉人者 Pete 还是被嫉者 Martin; 从而成功地消解代词“he”的指称歧义,并且还能够说明为什么这样做的理由。

下面,我们来看一下 ChatGPT 对相关的汉语句子的理解能力:

(2)a. 陈亮坐在邓忠后面,根本看不到黑板上的字,因为他太高了。

问题:他指谁? / 谁太高了? / 是陈亮太高,还是邓忠太高?

ChatGPT: 在句子 a 中,“他”应该指的是邓忠。邓忠太高了。句子的意思是因为邓忠太高,坐在他后面的陈亮根本看不到黑板上的字。所以,是邓忠太高,而不是陈亮。

b. 陈亮坐在邓忠后面,根本看不到黑板上的字,因为他太矮了。

问题:他指谁? / 谁太矮了? / 是陈亮太矮,还是邓忠太矮?

ChatGPT: 在句子 b 中,“他”应该指的是陈亮。陈亮太矮了。句子的意思是因为陈亮太矮,坐在邓忠后面,根本看不到黑板上的字。所以,是陈亮太矮,而不是邓忠。

可见, ChatGPT 不仅能够理解方位词“后面”的意义,理解反义形容词“(太)高”和“(太)矮”的意义,理解否定表达“看不到”的意义;而且,还拥有关于人体高矮与位置前后能否形成遮挡的常识。从而,得以综合利用这些知识,来推断是邓忠太高和陈亮太矮,最终识别这个句对中两个代词“他”的不同的指称意义。为了检验它是不是真的具有上述关于否定表达的语言知识,我们把例(2)改编成下面的肯定句,来测试 ChatGPT:

(3)a. 陈亮坐在邓忠后面,也能看到黑板上的字,因为他太高了。

问题:他指谁? / 谁太高了? / 是陈亮太高,还是邓忠太高?

ChatGPT: 在句子 a 中,“他”应该指的是陈亮。陈亮太高了。句子的意思是即使陈亮坐在邓忠的后面,他也能看到黑板上的字,因为他太高了。所以,是陈亮太高,而不是邓忠。

b. 陈亮坐在邓忠后面,也能看到黑板上的字,因为他太矮了。

问题:他指谁? / 谁太矮了? / 是陈亮太矮,还是邓忠太矮?

ChatGPT: 在句子 b 中,“他”应该指的是邓忠。邓忠太矮了。句子的意思是即使陈亮坐在邓忠的后面,他也能看到黑板上的字,因为邓忠太矮了。所以,是邓忠太矮,而不是陈亮。

可见,对于方位词、反义形容词和动补短语的否定表达, ChatGPT 都能进行正确的语义解读;并且,它还能够利用方位和物体高矮能否形成遮挡的百科知识,来消解代词可能的指称歧义。

既然 ChatGPT 等大模型有如此强大的语言运用能力:可以比较准确地理解人类的自然语言,并且生成比较流畅和自然的话语和文本;那么,接下来的问题自然是:大型语言模型是如何做到这一点的呢?下面,我们拟先从语言学的角度介绍语言大模型的工作机制。

3. 基于概率的语言模型和转换器的构架

从技术上讲, ChatGPT 等大模型是一种语言模型(language model, LM)。所谓语言模型,指为人类的自然语言所建立的数学模型。通俗地说,就是用数学公式来表示自然语言。比如,早期的生成语法把语言定义成一个由该语言的句子组成的集合,再用重写规则(rewriting rules)来定义句子。例如(详见 Chomsky 1957: 26-48):

(4) Σ : Sentence:

(i) $Sentence \rightarrow NP + VP$

(ii) $NP \rightarrow T + N$

(iii) $VP \rightarrow Verb + NP$

(iv) $T \rightarrow the$

(v) $N \rightarrow man, ball, etc$

(vi) $Verb \rightarrow hit, took, etc$

这可以看作是用基于产生式(production)规则的短语结构语法(phrase structure grammar)来给人类自然语言建模(modeling)。但是,这种基于规则的定义方式是离散式的,机器无法理解;所以,是不可计算的,无法在数字计算机上用程序来实现。

与此不同,计算机科学家采用基于统计的概率模型,来估算自然语言中每一个句子出现的可能性,即概率分布(probability distribution)。具体方法是把对句子的概率的计算,分解为组成句子的各个词的出现概率的乘积;即由组成句子的一串词中的每一个词的出现概率(条件概率)相乘,就得到句子的概率。经典的公式为:

$$(5) P(W) = P(w_1, w_2, \dots, w_m) = P(w_1) P(w_2 | w_1) P(w_3 | w_1 w_2) \dots P(w_m | w_1 w_2 \dots w_{m-1})$$

虽然这种概率模型不能判断该句子是否符合语法,但是可以用来在给定一个句子的前序词串(previous words)时,计算和预测出下一个位置出现某一个词的概率。比如,上面公式中的 $P(w_3 | w_1 w_2)$,即表示通过对一定规模的真实文本的统计,就可以由前面出现在 $w_1 w_2$ 位置上的某两个词,来算出不同的词在 w_3 位置上出现的概率。^②显然,统计大型语料库中各个词的出现概率和条件概率,这种计量方法是容易实现的。相对于(4)这种基于短语结构语法的语言模型来说,(5)应该算是基于词语法(word grammar)或 n-元语法(n-tuple grammar, n-grams)的语言模型。

下面先简单说明 ChatGPT 这个名称的意思。前缀 Chat,说明这是一款专注于聊天的人工智能软件;词干 GPT 是一个简称,全称是 Generative Pre-trained Transformer(生成式预训练转换器)。中心语“转换器”(Transformer)指一种具有良好的并行计算能力的多层的神经网络,可以处理序列数据(如文本、语音、图像等),是神经网络语言模型的基础组件。转换器一般由编码器(encoder)和解码器(decoder)两个部分组成,前者用以把输入的序列数据转换为一种维度相同的特征表示(feature representation);即一种叫做隐藏状态(hidden state)的向量表示,它可以捕捉数据中的特征和信息;后者用以根据编码器的输出和自身的状态,生成输出的数据并完成有关任务。

以翻译任务为例,编码器的输入是源语言的句子。虽然编码器可以一次性地看到整个句子,但是解码器只能一个一个地生成。因为,自回归(auto-regressive)模型只能根据自己(前面已经预测出来的词语)预测自己(后面即将被预测出来的词语)。解码器以此前时刻的一些输出作为此时此刻的输入,一个词例(token)一个词例地向右转移(shifted right)。因此,在翻译任务中,向右转移输出的是训练数据中源语言对应的目标语言的一个一个向右转移的词例语序。这种工作顺序,在训练阶段是并行的,在预测阶段是串行的。输入是通过“输入嵌入”(input embedding)把词例编码成分布式表示的向量,然后通过“位置编码”(positional encoding)来编码每一个词例的位置信息;因为输入没有位置信息,结果“狗咬人”和“人咬狗”就没有区别。然后,通过多头注意力机制(multi-headed attention, MHA)提取词例与词例之间的相关性。接下来,通过“求和归一”(Add & Norm)进行“残差连接”(residual connection)和“归一化”(normalization)。前者把前几层的输入连接到当前层,防止“梯度消失”(gradient vanishing);后者通过平均来减少数据的差异,以保证网络的有效性。然后,通过“前馈网络”(feed forward network, FFN,一种把数据向前传递进行线性变换和连接的网络),增加可学习的参数的数量。编码以后的结果输入到解码器。解码器跟编码器的结构类似,只

^② 关于语言模型及其相关的 n 元语法(n-grams),详见李沐等(2018),第 45-50 页。

是多了带掩码的多头注意力机制。^③

转换器模型的主要创新点是引入注意力机制 MHA。(详见 Vaswani et al. 2017)。多头注意力机制是转换器的核心,它可以让语言模型关注输入或输出数据中最重要或最相关部分,并且捕捉数据之间的长距离依赖关系,从而提高模型的理解和生成能力,以及处理复杂问题和目标的能力。每一个注意力头都包含相同的结构——带缩放的点积注意力机制(scaled dot-product attention),但是有着独立的参数。

然而,ChatGPT 只有解码器,没有编码器。这样,它只能根据之前生成的数据来预测下一个数据,而不能同时考虑整个数据序列。因此,这是一种自回归(autoregressive)的语言模型。修饰语“生成式的”(Generative)指一种跟“判别式的”(discriminative)相对的机器学习(machine learning)模型,后者通过寻找一个决策边界来将样本划分到对应类别,而前者通过学习每个类别的边界,体现更多数据本身的分布信息,可以用来生成样本,所以普适性更广(详见 Microstrong 2020)。生成式语言模型的基本思路是根据前面已知的词语,根据分布概率来预测接下来最可能出现的词语。修饰语“预训练的”(pre-trained)中的“预训练”(pre-train),指基于大型语料库对语言模型进行统一训练,并不考虑具体场景和下游任务(如:翻译、造句、摘要、文案创作等),完成预训练后,再根据应用场景和下游任务的具体情况,对语言模型进行微调(Fine-Tuning, FT)。

文献上通常称类似 GPT 这样的语言模型为“(语言)大模型”“大型语言模型”,或者“现代大型语言模型”,因为这种语言模型的结构相对复杂,主要是由多层神经元组成的巨大的神经网络,模型参数众多。比如,作为 ChatGPT 基础的 GPT-3 有 1750 亿个参数,那意味着构成转换器的神经网络至少有 1750 亿个神经节点之间的连接及其权重(weight)。但是,从数学上看,不管语言模型多么复杂,它们都只是一个函数,可以抽象地表示为“ $f(x) = y$ ”这样的函数式。比如,假如 x 的取值是英语句子, y 的取值是法语句子;那么, f 这个函数就是一个把英语翻译为法语的语言模型。当然,大型语言模型包含的变量数目是非常多的。

由于词向量是语言模型表示自然语言的基础,因而下面一节先介绍词向量。

4. 分布语义学和词语的向量化嵌入表示

受语言符号任意性(arbitrariness)原理的制约,词的形式并不能反映词的意义,也就不能反映不同的词在意义上的同异和远近。比如,我们知道“猫”“狗”的语义关系比“鱼”“虾”近,“草”“树”的语义关系比“雨”“雾”近。问题是,我们是怎么获得这种语义直觉的?有没有一种可操作的衡量词语之间语义相似性的办法?如果采用指称论(referential theory)的立场,那么词语的意义就是它所指称的事物;或者曲折一点说,词语的意义就是它跟其所指称的事物之间的关系。^④根据这种观点,人们可以凭借词语所代表的事物之间的相似性,来获得有关词语之间相似性的知识。但是,这种指谓语义学(denotational semantics)之下的认识,不容易形成可操作的衡量词语之间语义相似性的办法。相反,如果采用 Wittgenstein (1958)的用法论(use theory)的立场,那么“词的意义是它在语言中的用法”;^⑤或者像 Firth (1957)所说的:“You should know a word by the company it keeps!”(观其邻,知其意)。^⑥这种观点换成描写语言学的说法就是“words that used and occur in the same contexts tend to purport similar meanings”(被用于和出现在相同上下文中的词语往往指谓相似的意思)

③ 对于转换器的具体说明,主要根据黄子峰(2023)。

④ 详见徐烈炯(1995)第1章:指称论,第5-16页。

⑤ 详见徐烈炯(1995)第6章:用法论,第83-85页。

⑥ 这种表达是仿拟谚语:“a man is known by the company he keeps”(观其邻,知其人)。

义),^⑦这后来成为分布语义学(distributional semantics)的核心思想。

显然,这种分布语义学假设可以成为统计语义学(statistical semantics)的基础,也比较容易形成可操作的衡量词语之间语义相似性的办法:通过调查词语在使用的文本的主题领域(topical regions,如新闻、公文、小说、诗歌、博客,等等)、各种组合关系(搭配关系、具体的构式框架)和聚合关系(替换关系和上下位等语义层级关系,等等)等具体维度上的相似性,设定不同类别的词语在语义距离上的远近;并且,投射到一个多维度(或高维度)的意义空间(meaning space),或者语言学的特征空间(linguistic feature space)。

有了意义空间和语义距离的概念,我们就可以用类似建立坐标系(coordinate system)的办法,来确定不同的词语在某种意义空间中的位置及其数字表示。可见,所谓词嵌入(word embedding)就是根据词语在某种意义空间中的位置而指派一个数字(一般是数组),确保意义与用法(分布)上相似的词语在数值上也相似。这就是词语的嵌入性向量表示的要旨。^⑧由于词语的意义和用法(分布)太复杂,无法仅用二维来表示;因而现代语言模型往往用几百甚至几千维度的向量空间。于是,每个词的向量就是一长串数字(即数组)。比如,2013年Google发布了word2vec语言模型是300多维。Google研究团队分析了从Google新闻中收集的数百万篇文档,找出哪些单词倾向于出现在相似的句子中。经过多轮次训练,让神经网络学会把类别或意义相似的单词(如“猫”、“狗”和“宠物”等)放置在向量空间中相邻位置。比如,word2vec把英语单词“cat”(猫)表示为在300个维度上取值的300个数字的数组:^⑨[0.0074, 0.0030, -0.0105, 0.0742, 0.0765, -0.0011, 0.0265, 0.0106, ..., 0.0002]。

值得一提的是,随着现代语言模型规模的扩大,其所用的词向量的维度也越来越多,以便表达更加丰富而细微的语义信息。比如,OpenAI在2018年发布的GPT-1,使用了768维的词向量;其转换器有12层神经网络,神经节点之间的连接权重参数有1.17亿个。几个月以后发布的GPT-2,其最大版本拥有1,600维的词向量;其转换器有48层神经网络,神经节点之间的连接权重参数有15亿个。2020年发布的GPT-3具有12,288维的词向量;其转换器有96层神经网络,神经节点之间的连接权重参数有1,750亿个。这样,语言模型不仅能够区分同音词(如“bank”)的不同意义(银行vs.河岸),而且能够区别多义词(如“magazine”)的不同意义(杂志vs.杂志社)。^⑩

基于分布语义学的词语的语义空间,不仅能够反映词语在概念意义方面的亲疏远近,而且,还能反映不同词类的词语在二维的特征空间中的位置关系;特别是语义空间中有关词语之间的某些平行的传播关系(parallel transport),可以降维到二维平面上来明显而直观地揭示。图略(详见Wolfram 2023)。

这样,反映词语在高维的语义空间中的坐标点位置的词向量,就不仅可以捕捉到许多平行或对立的语义关系,而且还可以进行向量的加减运算和词语的意义关系的推理。比如,Google研究团队用“biggest”(最大的)的向量减去“big”(大的)的向量,再加上“small”(小的)的向量,得到的向量最接近于“smallest”(最小的)的向量。换句话说,“biggest”的向量减去“big”的向量,得到的是近于最高级“-est”的向量;最高级“-est”的向量加上“small”的向量,结果自然就是最近于“smallest”的向量。同样,根据“man”与“woman”类似于“king”和“queen”,进行向量运算,不仅可以学会性别

^⑦ 详见 Harris (1954)。关于分布语义学,详见 Distributional semantics, Wikipedia, https://en.wikipedia.org/wiki/Distributional_semantics。

^⑧ 详见 Word embedding, Wikipedia, https://en.wikipedia.org/w/index.php?title=Word_embedding&oldid=1170647139。

^⑨ 详见 Lee and Trott (2023)及其中文编译本。

^⑩ 详见 Lee and Trott (2023)及其中文编译本。

角色,而且可以学会“王室君主”这一概念(“king”的向量减去“man”的向量 \approx “queen”的向量减去“woman”的向量 \approx “royal”的向量)。推而广之,可以从“印度人”与“印度”类似于“美国人”和“美国”上,学会“国籍”这一概念;从“巴黎”与“法国”类似于“柏林”和“德国”上,学会“首都”这一概念;从“不道德的”与“道德的”类似于“不可能的”和“可能的”上,学会“反义关系”这一概念;从“mice”与“mouse”类似于“dollars”和“dollar”上,学会“复数形式”这一概念。总之,词向量编码了词语之间微妙而重要的关系,成为现代语言模型的一个必不可少的基础。^①

5. 转换器语言模型是如何进行后词预测的?

如上文所说,ChatGPT的前身GPT-3模型的转换器有96层神经网络;其中每一层都是一个转换器,都接受一系列跟输入文本相对应的向量作为输入;并且,每一层都根据上下文等信息来添加信息,帮助澄清有关歧义或模糊词语的意义,以便更好地预测下一个可能出现的词语。比如,假定输入文本是“John wants his bank to cash the”(约翰想让他银行兑现)。这些单词首先被表示为word2vec风格的向量,并传送到第一个转换器;这个转换器可能根据上下文信息,确定“wants”和“cash”都是动词;然后模型会通过修改词向量的方式来存储这一信息。这些新的向量被称为“隐藏状态”(hidden states),并传送到第二个转换器。这第二个转换器可能添加了另外两个词的上下文信息,来帮助澄清“bank”指银行,而不是指河岸;并且,“his”回指先行词“John”。于是,它产生了另一组隐藏状态向量。这种向量反映的是该模型之前学习到的所有信息。有研究指出,GPT-3模型的前几层专注于理解句子的语法,并解决上文所示的歧义问题;后几层则致力于理解整个段落的高层次信息,比如文本中人物的性别、年龄、跟其他人的角色关系、当前及以前的位置、个性、目标等等。其实,迄今为止,研究人员也并不完全了解语言大模型是怎样跟踪这些信息的,只是从逻辑上推断:模型在各层神经网络之间传递信息,必须通过修改隐藏状态向量来实现。

大家可以设想一下,GPT-3的最强版本使用具有12,288个维度的词向量;也就是说,每一个词都由一个包含12,288个数字的数字列表来表示。于是,每一个词的上下文的各种有关信息都得到了记录,并且可以被后面的网络层读取和修改。这样,不仅加深了模型对于输入文本(包括通过后词预测而新产生的文本)的意义的理解,而且便于网络的最后一层输出一个包含所有必要信息的隐藏状态,来准确地预测下一个单词。^②

下面,我们一起来看看每个转换器内部的工作情况。转换器在更新输入段落的每一个单词的隐藏状态时,主要有两个处理过程:1)在注意力步骤中,神经网络会在每一个词的周围进行观察,查找具有相关背景并彼此共享信息的其他词语。2)在前馈步骤中,神经网络会“思考”每一个词在前面注意力步骤中收集到的信息,并且尝试预测下一个单词。为了直观,大家可以把注意力机制看作是单词之间的一种撮合服务(matchmaking service)。神经网络为每一个词都设置三个向量(相当于三张检查表):一个是查询向量(query vector),描述该词所寻找的词语的特征;一个是键向量(key vector),描述该词自身的特征;还有一个是值向量(value vector),反映注意力加权后的词向量。神经网络通过把每个键向量跟每个查询向量进行比对(通过计算点积),来找到最佳匹配的单词。一旦找到匹配的项目,神经网络就把产生键向量的单词的有关信息,传递到产生查询向量的单词。比如,在处理前文提到的句子“John wants his bank to cash the”时,转换器内部的加工过程很可

^① 这一节内容,详见Wolfram (2023)和Lee and Trott (2023)及其中文编译本。另外,关于word2vec和词向量的运算,详见袁毓林(2022)§2。关于向量运算(vector calculation),详见Mikolov, et al. (2013)。

^② 这一节内容,主要根据Lee and Trott (2023)及其中文编译本。

能是这样的：“his”的查询向量可能会表示为‘寻找一个描述男性的名词’，而“John”的键向量可能会表示为‘这是一个描述男性的名词’。网络检测到了这两个向量匹配，于是把“John”的向量信息转移给“his”的向量。

正如前面所说的，每一个注意力层都有多个“注意力头”。这使得上述信息交换过程可以在每一层上会多次并行地进行，并且每一个注意力头各专注于不同的任务：有的把名词跟代词匹配，以建立回指关系；有的解析同音词或多义词，消除词汇歧义；有的把诸如“Joe”和“Biden”等几个单词连接在一起，识别出有关的姓名或命名实体(named entity)。就这样，一个注意力层上的操作结果，成为下一层中一个注意力头的输入。事实上，上面所提及的任务往往都需要动用多个注意力头。大家可以想象一下，GPT-3的最强版本有96层，每一层有96个注意力头；因此，每一次预测下一个词时，GPT-3将执行9,216个注意力操作。

根据研究机构 Redwood Research 的科学家 Wang, et al. (2022)介绍，他们研究了 ChatGPT 的前身 GPT-2 怎样预测下一个单词的过程。他们发现，对于“When Mary and John went to the store, John gave a drink to”（玛丽和约翰进入商店以后，约翰把一杯饮料给了）这样的段落，有三种类型的注意力头为预测出这下一个单词“Mary”作出了贡献：1)他们称之为“名称移动头”(Name Move Heads)的三个注意力头，把信息从“Mary”的向量复制到最后的输入词“to”的向量；2)他们称之为“主语抑制头”(Subject Inhibition Heads)的四个注意力头，它们标记了第二个“John”的向量，阻止名称移动头复制主语的“John”这个名字；3)他们称之为“复制词例头”(Duplicate Token Heads)的两个注意力头，它们把第二个“John”的向量标记为第一个“John”的向量的复制副本，帮助主语抑制头决定不应该复制“John”。这后面两种注意力头的功能，他们是通过 GPT-2 的计算过程进行逐步逆向推导而发现的。这三种注意力头的协同作用，使得 GPT-2 的神经网络能够决定“Mary”是正确的下一个词。也就是说，这九个注意力头使得 GPT-2 能够理解“John gave a drink to Mary”是有意义的，而“John gave a drink to John”是没有意义的。为此，他们发表了一篇长达 25 页的论文，来解释他们是如何识别和验证这些注意力头的。这从侧面反映：要完全解释 GPT 等现代大型语言模型的工作机制，将是一个无比巨大的工程，人类不太可能在短时间内完成。^⑬

正如上文所说的，在注意力头与词向量之间传输信息以后，前馈网络会“思考”每个词向量，并且尝试预测下一个单词。在这个阶段，虽然单词之间没有交换信息，但是前馈层会独立分析每个单词，并且可以访问之前由注意力头复制的任何信息。前馈层的神经元是计算其输入加权值的数学函数。前馈层之所以强大，是因为它有大量的连接。比如，输出层有 12,288 个神经元（对应模型的 12,288 维词向量），隐藏层有 49,152 个神经元，它的每个神经元有 12,288 个输入值（因此每个神经元有 12,288 个权重参数）。这就意味着，每个前馈层有 $(49,152 * 12,288 + 12,288 * 49,152) = 12$ 亿个权重参数。这样，96 个前馈层就共有 $(12 \text{ 亿} * 96) = 1160$ 亿个参数。这相当于作为 ChatGPT 基础的 GPT-3.5 的参数量（1,750 亿）的三分之二。

以色列特拉维夫大学的研究人员发现，前馈层通过模式匹配进行工作；隐藏层中的每个神经元都能匹配输入文本中的特定模式，比如，分别是以“substitutes”结尾的词语序列，跟军事有关并且以“base”或“bases”结尾的词语序列，以时间范围结尾的词语序列（如“在下午 3 点到 7 点之间”），跟电视节目相关的词语序列（如“原始的 NBC 日间版本，已经存档”），等等。基本倾向是前面的层匹配特定的单词，后面的层匹配更加广泛的语义类别的短语。当一个神经元跟其中的一个模式匹配时，它会向词向量中添加信息。虽然这些信息目前还不太容易解释，但是在许多情况下，好像都可

^⑬ 这一节内容，主要根据 Lee and Trott (2023) 及其中文编译本。

以看作是针对下一个词的临时性预测。而布朗大学的研究人员发现,前馈层有时使用如上文所示的向量运算来进行类比推理,从而准确地预测下一个单词。例略。

可见,前述的注意力机制和前馈层有一定的分工:前者专注于从提示等较早生成的部分检索信息,而后者使语言模型能够“记住”从训练数据中学到的信息。其中,靠前的前馈层更可能编码跟特定单词相关的事实,比如“特朗普”经常在“唐纳德”之后出现;靠后的前馈层则编码更复杂的关系,比如,添加某个向量来使一个国家的名称变成其首都的名称。^⑭

6. 大型语言模型的训练方式

上面介绍了大型语言模型的注意力机制和前馈层神经网络的工作原理,现在简单介绍一下这种拥有上千亿个参数的大模型是如何训练的。

抽象地说,在自然语言处理中,专家们通常把一个文本看作一个词例(token)序列。给定一个语料库,其中的词例的概率分布记作 P_{data} ,语言模型希望通过训练,使得模型输出的概率分布 P_{θ} 拟合(fitting)数据的概率分布 P_{data} 。而自回归语言模型把联合概率建模成条件概率的乘积,公式为:

$$P_{\theta}(w_1, w_2, \dots, w_n) = \prod_t P_{\theta}(w_t | w_{<t})$$

其优化目标为极大似然(Maximum Likelihood)。公式为:

$$\operatorname{argmax}_{\theta} E_w \sim P_{data} \left[\sum \log P_{\theta}(w_t | w_{<t}) \right]$$

在理想情况下,模型收敛到最优,即 $P_{\theta} = P_{data}$ (参考秦兵 2023)。也就是说,完成了对语言模型的训练。

但是,实际训练起来却困难重重。好在大型语言模型有一个成功的创新,那就是不需要用人工事先标注好的数据来训练,而是通过预测文本中的下一个词或者中间被掩码的(masked)词语,来学习人类自然语言。这具体的后一个词到底预测得对不对,这是可以在语料库中核实的。这样,质量较高的未经标注的语料库,俨然成了提供标准答案的“有标注”的训练数据库。这种语料是容易获得并且数量庞大的,从而有力地推动了大模型的训练规模和质量。比如,大模型接受输入“I like my coffee with cream and”(我喜欢在咖啡里加奶油和),并且试图预测接下来出现的单词是“sugar”(糖)。也就是说,像小学生通过在句子末尾或中间填空而学会组词造句,从而学会一种语言的语法一样。更进一步看,这种后词预测训练等于是在做信息压缩工作(后面的信息被有效地打包在前面的词例序列中);^⑮并且,从本质上说,也是在做知识压缩工作(因此,适当的提示就可以唤醒和召回语料库中可能存在的信息)。当然,一个初始化的模型在这方面会表现得很糟糕,因为它的每一个权重参数最初都是从一个随机数字开始的;后来,随着模型“看到”更多的例子(动辄上千亿个单词的文本数据集),它的权重参数逐渐调整,慢慢接近或达到正确的预测。这种情况,有点儿像你住旅馆的第一天洗澡前,尝试调整水龙头的把手的方向,并触摸水流:如果水温太烫或太冷,你会向相反的方向转动把手;越是接近你需要的水温时,你转动把手的幅度越来越小。

现在,我们利用上面这个类比进行想象:首先,有5万多个水龙头,每一个对应着由5万多个词构成的词汇总表中的一个特定的单词;你的目标是,只让跟给定词语序列中下一个单词相对应的水

^⑭ 这一节内容,主要根据 Lee and Trott (2023) 及其中文编译本。与此不同的一种观点认为,转换器的前馈层是一种存储大量具体知识的 Key-Value 存储器。详见 Geva, et al. (2021) 和 Dai, et al. (2021)。

^⑮ 笔者猜想,这可能就是 GPT-4 的重要缔造者、前 OpenAI 首席科学家 Ilya Sutskever 所谓的“通过[数据]压缩来进行无监督学习”。详见王庆法(2023)。

龙头出水(即接续正确的单词)。其次,水龙头后面有一大堆互相联通的管道,管道的上面有一大堆阀门。所以,如果水从错误的管道里出来(即接续不正确的单词),那么你不能只调整水龙头的旋钮。怎么办?如果你是魔法师,或者你会使用魔法;那么你可以派遣一个精灵,或者一支“松鼠部队”,去追踪每一条管道,并且沿途调整它们找到的每一个阀门。由于同一条管道通常供应多个水龙头,因而精灵或松鼠们需要仔细思考如何确定拧紧或松开哪些阀门,以及松紧的程度有多大。当然,这个比喻太疯狂了,因为 GPT-3 的最强版本有 1750 亿个参数,人们不可能去建造一个拥有 1750 亿个阀门的管道网络。但是,拜摩尔定律(Moore's Law)所赐,现代计算机的确可以并且确实以这种规模运行。

实际上,大型语言模型的注意力头和前馈层的神经元都被实现为一系列简单的数学函数(主要是矩阵乘法),其行为由可调整的权重参数来确定。就像上面想象的故事中的精灵或松鼠部队通过松紧阀门来控制水流一样,训练算法通过增加或减少语言模型的权重参数来控制信息在神经网络中的流动。

训练过程分成两个步骤:首先进行“前向传递”(forward pass),相当于上面故事中的打开水源,并且检查水是否从正确的水龙头流出。然后关闭水源,进行“后向传递”(backwards pass),精灵或松鼠们沿着每根管道快速移动,并且拧紧或松开阀门。在数字神经网络中,精灵或松鼠的角色由一个称为“反向传播”(backward propagation)的算法来扮演。该算法“逆行”(walks backwards)通过网络,使用微积分来估计需要改变每个权重参数的程度。完成这样一个过程(即对一个示例先进行前向传递,再进行后向传递,来提高神经网络在这个示例上的性能),需要进行数百亿次数学运算。因此,所谓语言模型的训练,主要是调整网络神经元之间的连接权重参数,简称“调参”,或者戏称“炼丹”。训练像 GPT-3 这样的大型语言模型,需要重复这个过程数十亿次,因为对每个训练数据的每一个词都要进行训练。据 OpenAI 估计,训练 GPT-3 需要超过 3,000 亿万次浮点计算(对每个权重参数进行近 2 万亿次浮点计算),而这需要几十个高端计算机芯片运行几个月。^⑩

对于 ChatGPT 来说,它的最长路径大约有 400 个核心层,并且有数百万个神经元,总共大约有 1750 亿个连接,所以也就有 1750 亿个权重参数。需要注意的是,ChatGPT 每次生成一个新的词例时,都必须进行涉及每个权重参数的计算。在实现上,这些计算可以被组织成高度并行的数组操作,可以方便地在 GPU 上完成。但是,对于产生的每个词例,仍然需要进行 1750 亿次计算(最后还要多一点)。因此,当 ChatGPT 生成一段长文本时,往往需要较长一段时间也就可想而知了。^⑪

7. 语言大模型的成功帮助我们重新认识语言

物理学家费曼(Richard Phillips Feynman, 1918—1988)说过:“我不能做出来的事物,我就不理解。”(What I cannot create, I do not understand.)我把它演绎为:“理解一个事物的最好的办法就是把它做出来。”(The best way to understand something is to create it.)照此说来,理解人类自然语言的结构机制和工作原理的最好的办法,莫过于造出一种能够生成和理解自然语言的机器,来反观和检验语言学家关于语言的结构和功能的有关理论。不知道是幸运还是不幸,软件工程师们经过 70 多年不懈的努力,已经为我们造出了 ChatGPT 等语言运用能力十分强大的现代语言模型。此情此景之下,我们的语言学家与其坐在扶手椅上对 ChatGPT 评头品足,挑剔指责其“虚假的承诺”(false promise,即提供虚妄的前景),^⑫不如想一下:ChatGPT 等现代语言模型的成功,能够给当代语言学带

^⑩ 这节的大部分内容,主要根据 Lee and Trott (2023)及其中文编译本改写而来。

^⑪ 详见 Wolfram (2023)及其中文编译本。

^⑫ 详见 Chomsky, et al. The False Promise of ChatGPT, *New York Times*, Mar. 8, 2023。

来什么样的启示？大模型通过“再造语言”(recreate a language)的方式提醒我们，人类自然语言有哪些值得注意的特点？相应地，我们的句法学和语义学的理论建设和研究方向，应该进行哪些反思？

总结有关学者和相关文献的观点，我们可以发现，这种启示至少有以下五点：

1) 基于分布语义学的词嵌入向量表示，对于捕捉和表示语言的意义十分有效。正是认识到词语的具体意义取决于上下文，利用上文信息来预测下一个单词的方法才得以实现；并且，使得语言模型能够通过找出最合适的下一个词的方式，来学习人类语言的运作方式。当然，遗憾还是有的，那就是：最终得到的这种人工智能系统，其内部的运作方式，还不能被人类专家所完全理解；也就是说，现代大型语言模型还没有达到“可解释的人工智能”这种机理透明的“表里俱澄澈”的科学境界。

2) 分布语义学的基本假设：语义上越相似的两个词语在分布上也越相似，结果它们越倾向于出现在相似的上下文语境中。这个假设不仅帮助软件工程师解决计算建模(computational modeling)时遇到的“数据稀疏问题”(the data-sparsity problem)(详见 Wishart et al. 2017)，得以借助同义词或近义词集合中数据丰富的词语的分布来为数据稀疏的其他词语建模；而且，帮助认知科学家回答儿童习得语言时的“刺激贫乏问题”(the problem of the poverty of stimulus，即儿童是如何在输入相对贫乏的情况下快速地学会一种语言的？)(详见 Yarlett 2008)。也就是说，我们有理由假定：儿童可能利用同义词或近义词进行类推和泛化(generalization)，来从有限的输入语料上学会一种语言。^①我们知道，乔姆斯基正是基于“刺激贫乏问题”而坚信先天的普遍语法的必要性的。推而广之，人类的这种类推和泛化能力，也可用以回答“柏拉图问题”(Plato's problem)：在可借鉴的事物极端贫乏的情况下，人类是如何获得如此丰富的知识的？

3) 语义的分布假设既然可以解释语言学习中的泛化现象，那么自然可以解释：为什么人类可以通过少量的例子来学会完成各种语言任务？比如，从陈述句上变换出疑问句，从主动句上变换出被动句，等等。值得一提的是，2020 年 OpenAI 在推出 GPT-3 时，发表了一篇由 Tom Brown 领衔，包括 Ilya Sutskever 在内的 31 位作者，长达 75 页的论文《语言模型是少样本学习者》(Language Models are Few-Shot Learners)，说明只需要给语言模型看几个示范的例子(即少样本)，它就可以完成诸如把英语词语翻译成相应的法语词语之类的任务，这就是所谓的“在上下文中学习”(in-context learning)的提示策略。其背后朴实的理念就是“人类不需要很多有监督数据就可以学会大多数的语言任务”(humans do not require large supervised datasets to learn most language tasks)(详见 Brown et al. 2020)。换句话说，这种类推推理和泛化是基于人类一般的认知能力。因此，先天遗传的专门化的语言能力和普遍语法假设并不是必需的。

4) 语言本身是可预测的，语言的结构规律比我们通常设想的要简单得多。因为，语言的规律性往往跟物质世界的规律性相联系；所以，当语言模型学习了单词之间的语义关系时，通常也在隐性地学习外部世界中事物之间存在的关系。否则，如果人类自然语言真像当代形式句法学和形式语义学所描绘的层次繁复的结构图景，那么人工神经网络语言模型网络恐怕是无法处理的。于是，随之而来的问题便是：当代形式句法学对句子结构的叠床架屋式的树形图构想与分析，有没有掐准自然语言的命门？有没有语言事实根据和理论或技术上的必要？同样，当代形式语义学对句子的语义的多重隐性算子约束的逻辑结构，有没有意义表达上的根据和逻辑技术上的必要？(详见袁毓林 2019)

5) 语言运用本身是一种预测的过程，因为人脑是一台“预测机器”，对环境的良好预测和表示是生物能够适应环境的前提。因此，预测应该是生物智能和人工智能的基础。^②在语言信息的处理

^① 参考 Distributional semantics, Wikipedia, https://en.wikipedia.org/wiki/Distributional_semantics。

^② 这一节内容，详见 Wolfram (2023)、Lee and Trott (2023) 及其中文编译本。

过程中,对后续文本的预测也是必不可少的。这方面的研究,在心理语言学领域比较多,但是在理论语言学领域很少见到。我们希望,像预期、反预期、惊异等新兴的语言学理论范畴的探索,能够吸收人工智能和心理语言学上关于语言预测的相关研究,积极为人工智能时代的语言学研究开拓新的学术生长点。

6)最后,但并非最不重要的是,语言能力可能并不是专属于人类的(language maybe not unique to human beings)。像 ChatGPT 等语言大模型的语言运用能力,已经直逼人类的语言运用水平(详见袁毓林 2024b)。这可能使某些人感到不舒服,觉得人类的自尊心受到伤害;因为它挑战了“语言独异说”,使得语言好像不再是唯一地属于人类的。但是,我们认为,这是我们人类自己设计出来的机械能够“再造语言”,我们的“人性”(human nature)并没有受到冒犯。

参考文献:

- [1] Bernard, T. & T. Han. Mandarinograd: A Chinese collection of Winograd schemas[C]// European Language Resources Association (ELRA). *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Marseille: European Language Resources Association, 2020.
- [2] Brown, T. B., Mann, B. & N. Ryder. et al. Language models are few-shot learners[J]. *Advances in Neural Information Processing Systems*, 2020, (33): 1877–1901.
- [3] Chomsky, N. *Syntactic Structure*[M]. The Hague: Mouton. 邢公畹等,译. 北京: 中国社会科学出版社, 1979.
- [4] Chomsky, N. *Syntactic Structure*[M]. The Hague: Mouton. 陈满华,译. 北京: 商务印书馆, 2022.
- [5] Dai, D., Dong, L. & Y. Hao. et al. Knowledge neurons in pretrained transformers[J]. *arXiv preprint arXiv*, 2021, 2104. 08696.
- [6] Firth, J. R. A synopsis of linguistic theory 1930–1955[C]// Firth, J. R. & F. R. Palmer. *Selected Papers of J. R. Firth 1952–1959*. London: Longman, 1968.
- [7] Geva, M., Schuster, R., Berant, J., & O. Levy. Transformer feed-forward layers are key-value memories[J]. *arXiv preprint arXiv*, 2020, 2012.14913.
- [8] Harris, Z. S. Distributional structure[J]. *Word*, 1954, 10(2–3): 146–162.
- [9] Lee, T. B. & S. Trott. Large language models, explained with a minimum of math and jargon[J/OL]. *Understanding AI*, 2023, (27). <https://www.understandingai.org/p/large-language-models-explained-with>.
- [10] Levesque, H., Davis, E. & L. Morgenstern. The winograd schema challenge[C]// Levesque, H., Davis, E. & L. Morgenstern. *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Rome: AAAI Press, 2012.
- [11] Mikolov, T., Sutskever, I. & K. Chen. et al. Distributed representations of words and phrases and their compositionality[J]. *Advances in neural information processing systems*, 2013, 26: 3111–3119.
- [12] Vaswani, A., Shazeer, N. & N. Parmar. et al. Attention is all you need[J]. *Advances in neural information processing systems*, 2017, 30: 5998–6008.
- [13] Wang, K., Variengien, A. & A. Conmy. et al. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small[J]. *arXiv preprint arXiv*, 2022, 2211. 00593.
- [14] Wishart, R. & P. Prokopidis. Topic Modelling Experiments on Hellenistic Corpora[C]// *Proceedings of the Workshop on Corpora in the Digital Humanities*. Bloomington, 2017.
- [15] Wittgenstein, L. *Philosophical Investigations*[M]. Trans. Anscombe, G. E. M. Oxford: Basil Blackwell, 1958.
- [16] Wolfram, S. *What Is ChatGPT Doing and Why does It work?*[M]. WOLFRAM 传媒汉化小组,译. 北京: 人民邮电出版社, 2023.
- [17] Yarlett, D. & M. Ramscar. Language learning through similarity-based generalization [D]. Palo Alto: Stanford University, 2008.

- [18] 黄子峰. 详解 ChatGPT 的智能从何而来?[EB/OL]. (2023-02-20) [2023-02-19]. <https://www.acfun.cn/a/ac40704856>.
- [19] 李沐, 刘树杰, 张冬冬, 周明. 机器翻译 [M]. 北京: 高等教育出版社, 2018.
- [20] 秦兵. 大语言模型之人类价值观对齐 [EB/OL]. (2023-08-03) [2023-08-10]. <https://mp.weixin.qq.com/s/888XZ43VP8nefVXgGww9Bw>.
- [21] 王庆法. OpenAI 首席科学家透露 GPT4 技术原理 [EB/OL]. (2023-03-17) [2023-03-30]. <https://mp.weixin.qq.com/s/H8vNSn-0Ho2Ho4I0n7YDfQ>.
- [22] 徐烈炯. 语义学 (修订本)[M]. 北京: 语文出版社, 1995.
- [23] 袁毓林. 为什么要给语言建造一座宫殿?——从符号系统的转喻本质看语言学的过度附魅[J]. 语言战略研究, 2019, 4(4): 60-73.
- [24] 袁毓林. 在人类生境约束下思考语言的设计原理和运作机制[J]. 语言战略研究, 2022, 7(6): 85-96.
- [25] 袁毓林. 如何测试 ChatGPT 的语义理解和常识推理水平?——兼谈大语言模型时代语言学的挑战与机会[J]. 语言战略研究, 2024, 9(1): 49-63.
- [26] 袁毓林. 语义理解与常识推理的机器表现和人类基线之比较——怎样评估 ChatGPT 等大型语言模型的语言运用能力?[J]. 汉语学报, 2024, 待刊。

基金项目: 澳门大学讲座教授研究与发展基金(CPG2024-00005-FAH)和启动研究基金(SRG2022-00011-FAH)

收稿日期: 2024-03-01

作者简介: 袁毓林, 博士, 教授, 博士生导师。研究方向: 理论语言学和汉语语言学, 特别是句法学、语义学、语用学、中文信息处理和计算语言学。

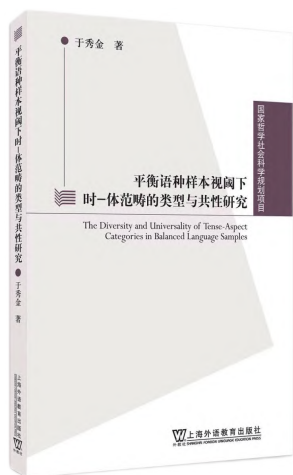
《平衡语种样本视阈下时一体范畴的类型与共性研究》

作者: 于秀金 著

出版时间: 2024 年 6 月

定价: 52.00 元

出版社: 上海外语教育出版社



该书基于所选取的 155 种语种样本, 考察了世界语言中的时范畴、体范畴及时一体与情态之间的关系, 探讨了时一体类型差异的跨语言表现和差异中隐含的共性, 并重点关注汉语和英语时一体的类型与共性, 在此基础上, 构建了跨语言时一体—情态的范畴化层级, 以解释包括汉语和英语在内的世界语言中时一体—情态的复杂形—义对应关系。该书力图用语言表征的形式特征及其原型义来评估不同语言中时一体的范畴化程度和显赫性差异, 揭示跨语言时一体范畴地位不同的原因, 为时一体的类型学研究打开了一个新视角。