



大模型关联度预测的形式化和语义解释研究

陈小平

引用本文:

陈小平. 大模型关联度预测的形式化和语义解释研究[J]. *智能系统学报*, 2023, 18(4): 894–900.

CHEN Xiaoping. Research on formalization and semantic interpretations of correlation degree prediction in large language models[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(4): 894–900.

在线阅读 View online: <https://dx.doi.org/10.11992/tis.202306045>

您可能感兴趣的其他文章

融合多层次特征的中文语义角色标注

Chinese semantic role labeling with multi-level linguistic features

智能系统学报. 2020, 15(1): 107–113 <https://dx.doi.org/10.11992/tis.201910012>

可拓聚类的科教人际网络节点重要性动态分析方法

Dynamic analysis method of importance of science and education interpersonal network nodes based on extension clustering

智能系统学报. 2019, 14(5): 915–921 <https://dx.doi.org/10.11992/tis.201811012>

因素空间理论——机制主义人工智能理论的数学基础

Factor space-mathematical basis of mechanism based artificial intelligence theory

智能系统学报. 2018, 13(1): 37–54 <https://dx.doi.org/10.11992/tis.201711034>

机制主义人工智能理论——一种通用的人工智能理论

Mechanism-based artificial intelligence theory: a universal theory of artificial intelligence

智能系统学报. 2018, 13(1): 2–18 <https://dx.doi.org/10.11992/tis.201711032>

基于面向对象时间Petri网的智能电商物流服务系统建模

Modeling of intelligent E-commerce logistics service system based on object-oriented time petri nets

智能系统学报. 2018, 13(2): 303–313 <https://dx.doi.org/10.11992/tis.201612031>

利用智能引导和KDML增强可拓模型人机建模能力研究

Research on enhancing the human-machine modeling ability for an extension model using the intelligent guide and KDML

智能系统学报. 2017, 12(3): 348–354 <https://dx.doi.org/10.11992/tis.201610017>

DOI: 10.11992/tis.202306045

大模型关联度预测的形式化和语义解释研究

陈小平

(中国科学技术大学 计算机科学与技术学院, 安徽 合肥 230026)

摘要: 本文探讨“大型语言模型是什么”的问题。为此对大模型的评判标准展开实验观察, 对大模型的基础设施关联度预测进行直观分析, 构建关联度预测的一种形式化 L_C , 进而研究关联度预测的语义解释。在此基础上讨论大模型的真实挑战、共识挑战、内容属性挑战和而非封闭性挑战。主要发现包括: 语义关联度是体现人类语言习惯的可自动提取的语言痕迹; 关联度预测具有语境相关的统计性质; L_C 具有弱共识性实质语义; L_C 是一个非概念化公理系统。这些特点颠覆了科学理论、形式化方法和软件的传统理念在人工智能领域的主导地位, 是大模型输出既出人预料、又符合语言习惯的深层原因。

关键词: 大模型; 形式化; 语义; 概念化; 弱共识

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673-4785(2023)04-0894-07

中文引用格式: 陈小平. 大模型关联度预测的形式化和语义解释研究 [J]. 智能系统学报, 2023, 18(4): 894-900.

英文引用格式: CHEN Xiaoping. Research on formalization and semantic interpretations of correlation degree prediction in large language models[J]. CAAI transactions on intelligent systems, 2023, 18(4): 894-900.

Research on formalization and semantic interpretations of correlation degree prediction in large language models

CHEN Xiaoping

(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: To explore the problem of what a large language model is, we conduct experimental observation on the evaluation criteria for large language models, intuitively analyze the infrastructure of large language models—correlation degree prediction, of which a formalization L_C is constructed and semantic interpretations are explored. On top of these, four challenges of truthfulness, consensus, content attribute, and non-closeness for large language models are discussed. The main findings include: the correlation degrees between tokens are automatically extractable language traces that reflect human language habits; correlation degree prediction has the context-sensitive statistical property; L_C has a substantive semantics of weak consensus; L_C is a non-conceptualized axiomatic system. These radically differ from the traditional notions of scientific theory, formal methods, artificial intelligence(AI) and software, and are the deep reasons why large language models can behave unexpectedly yet consistent with human language habits.

Keywords: large language models; formalization; semantics; conceptualization; weak consensus

当前大型语言模型(简称大模型)基础研究的首要问题是: 大模型是什么? 这一问题引起了普遍困惑和关注, 不仅波及面广, 而且涉及基础研究中的多项颠覆。为此, 本文以相应的非常规思

路展开探讨。首先从普通用户对大模型的评判标准入手, 尝试理解大模型引发普遍反应的真实原因。其次, 针对大模型的基础设施——关联度预测, 进行直观分析。在此基础上探索大模型的理论研究, 构造关联度预测的一种形式化, 包括形式公理系统 L_C 及其语义结构分析, 以澄清关联度预测的基础特性——语境相关的统计性质、弱共

收稿日期: 2023-06-20.

基金项目: 国家自然科学基金项目(92048301, U1613216); 国家重点研发计划项目(2020YFB1313602).

通信作者: 陈小平. E-mail: xpchen@ustc.edu.cn.

识性实质语义和非概念化公理系统。弱共识性和非概念化等特性颠覆了科学理论、形式化方法和软件的基本理念在人工智能领域的传统支配地位,是大模型得以产生既出人预料、又符合人类语言习惯的效果的深层原因。依据这些特性,进一步讨论大模型带来的4项新挑战——真实性挑战、共识挑战、内容属性挑战和非封闭性挑战。

1 实验观察:语言运用的习惯性

自然语言人机交互是大模型的核心功能。因此,普通用户对大模型的反应是大模型研究的重要内容之一。目前全球几亿人对大模型做了测试,人提出问题,大模型回答,人评判回答是否正确。普通用户正是根据这种评判,来决定是否接受大模型。所以,普通用户的评判标准具有决定性作用,有必要对其加以研究。

人类对推理进行了2000多年持续性研究,形成了推理的各种定义,这些定义可以作为评判大模型推理能力的专业性标准。以此为基础,本文对大模型进行了实验观察。首先反复向ChatGPT提出一个有关推理的问题,收集它的所有回答,然后分析回答所依据的评判标准。所有回答中最好的一个如下。

问:在王之涣的《登鹳雀楼》中,出现了哪些推理?请回答得尽可能全面。

答:在这首古诗中,包含了一些简单的逻辑推理和推断。具体如下:

1)“白日依山尽”和“黄河入海流”这2句诗句都是对自然现象的描述,暗示了时间的流逝和河流的长远,这可以看作是对现象的归纳推理;

2)“欲穷千里目”这句话表达了诗人想要看到更远的景象,这可以看作是一种归纳推理,即从局部的现象中推断出更广阔的事物;

3)“更上一层楼”这句话则是根据前面的推理得出的结论,即要看到更远的景象,就需要站在更高的地方,这可以看作是一种演绎推理。

综上所述,这首古诗中出现的推理包括归纳推理和演绎推理。

归纳推理的一种定义是:归纳是从个别到一般的推理。回答1)和回答2)是符合这个定义的,而且网上可以找到有关归纳推理的大量类似的说法。但是,归纳推理还应符合推理的一般定义——推理是前提与结论之间关系,而前提和结论都是用语言表达出来的。回答1)和回答2)显然不符合这个一般定义,而这一定义在网上很少出现。

演绎推理的一种定义是:在给定条件下,如果

前提成立,则结论一定成立。回答3)认为第3句与第4句之间是演绎推理关系,这也符合网上的流行说法。然而,实际上第3句是表达愿望的,第4句可理解为表达愿望或表达行动,但后一理解与全诗的意趣不符,舍去。于是,第3句和第4句作为愿望的表达,它们之间并非演绎推理关系,因为在给定条件下,从第3句表达的愿望,不仅可以诱发“更上一层楼”的愿望,也可引发其他愿望如“变作一只鸟”。这表明,确定适当的评判标准,有时并不容易。一般情况下,普通用户不会仔细推敲评判标准,而是依据自己的经验(大脑的第一反应)做出评判。

与推理相比较,理解、学习、决策、规划等功能的评判甚至更加复杂,所以普通用户更不会仔细推敲对应的评判标准。还有一些评判,比如对语言流畅性的评判,也难以依据严格的标准进行,凭感觉反而更准确。总之,普通用户对大模型的评判所依据的,主要是个人的语言经验而非科学标准,本文称之为“习惯性评判”,而相应的科学标准往往尚未建立起来。

值得注意的是,人对大模型语言运用的评判与人对自己语言运用的评判,采用的评判标准是相同的,不可能另设一套标准。既然人对自身的语言运用是否正确采取习惯性评判,那么人的语言运用也一定是依据习惯的,即根据自己的语言运用习惯回答问题,这样做具有现实可行性。

实验得出的主要观察是:人的语言运用通常是习惯性的,并根据自己的语言习惯评判大模型的行为。这一观察为理解大部分普通用户对大模型的反应、认识大模型人机交互的应用特点,提供了重要参考。

2 大模型的基础设施:关联度预测

根据Andrej Karpathy在微软Build 2023开发者大会专题演讲中的介绍,基础模型预训练占大模型整个训练计算时间的99%,这一关键事实揭示了基础模型预训练的重大意义。基础模型预训练的目标是利用语元关联度预测下一个语元(简称关联度预测)。因此,关联度预测是整个大模型技术体系的基础设施,在很大程度上决定了大模型的行为和性能。

在人工智能(artificial intelligence, AI)领域, Bengio等^[1]提出了关联度预测的最初想法。支持该想法的一项关键技术是利用词向量间的距离表达文本语义关系的距离^[2]。Transformer^[3]可以有

效提取不同上下文中单词和句子的文本语义,成为目前大模型的基础架构。另一项关键技术是在海量无标注语料上进行预训练的自监督训练方法^[4]。其他一些技术也为大模型的工程实现发挥了作用。

大模型技术中,语元(token)指的是字、词和标点符号,或者由字词和标点符号组成的字符串(实际训练中往往使用其他类型的符号串,如字符编码中的字节)。下面用一个极度简化的例子,帮助说明什么是关联度预测。考虑只有 2 个不同句子的语料,2 个句子在语料中重复出现,出现的概率分别为 0.6 和 0.4,如表 1 所示。

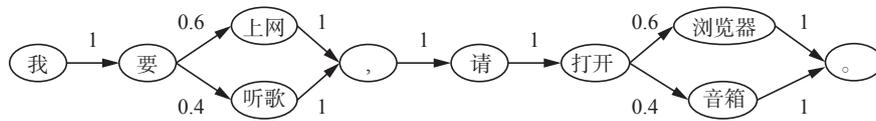


图 1 2 个句子的关联概率

Fig. 1 Correlation probability graph of the two sentences

任意给定一个语元,可以根据关联概率预测下一个语元是什么、出现的概率多大。比如给定语元“我”,根据图 1 可预测下一个语元是“要”的概率为 1;给定“打开”,下一个语元是“浏览器”或“音箱”的概率分别为 0.6 或 0.4。当下一个语元有多种可能时,最简单的办法是选关联概率最高的语元为预测结果,如选“浏览器”为“打开”之后的下一个语元。这个预测的正确率为 60%,错误率达 40%。

如果能够根据已经出现的多个语元进行预测,则可大幅降低预测错误率。例如,如果已经出现了“听歌”和“打开”,则可做出预测:下一个语元是“音箱”的概率为 1,这个预测的错误率降到 0。不过,“听歌”与“音箱”相距很远,它们之间的关联概率没有包含在图 1 中。事实上,图 1 只包含相邻语元之间的关联概率,而降低预测错误率需要综合利用相邻和不相邻(尤其远距离)语元之间的关联概率。

大模型以人类规模的原始语料作为训练样本。有人估计,ChatGPT 的训练样本达到互联网文本总量的 1/3~2/3。从这么多语料中提取远距离语元之间的关联概率所需的计算量远远超过目前的计算能力。另一方面,按现有概率、统计方法进行预测,效果未必好。为此,人工智能引入了精巧的启发式方法^[3],对不同距离的语元自动提取关联概率的某种近似。本文称这种近似为语元关联的强度,简称关联度,远距离语元之间的关联度简称远程关联度。

表 1 2 个句子及其概率
Table 1 Two sentences and their probabilities

句子	概率
我要上网,请打开浏览器。	0.6
我要听歌,请打开音箱。	0.4

针对表 1 中的语料,可以生成一个关联概率图,如图 1 所示。图 1 中的每一个结点(椭圆)代表语料中的一个语元,每一个箭头代表 2 个语元之间的关联(又称依赖),箭头上的数字是关联的概率。如语元“要”与语元“上网”的关联概率为 0.6,与“听歌”的关联概率为 0.4。

大量远程关联度的使用为提升大模型的性能发挥了关键作用。对于用户提出的一个问题,大模型对大量关联度进行综合,根据综合结果预测该问题的下一个语元、下下个语元...,直到产生一个完整的回答。这是一种启发式统计预测,是本文的主要研究对象。

为了达到实用要求,还需对预训练模型做后续优化,主要包括模型细调(fine-tuning,又译精调、微调)^[5-6]和提示(prompting)工程^[7],这些内容不在本文形式化研究的范围内。

3 关联度预测的形式化

关联度预测是现有数学模型不能刻画的一种新机制。本节为关联度预测构建一种新的形式公理系统 L_C ,下节考察 L_C 的语义解释。

L_C 的形式语言包括 3 类语法元素: 1) 语元符号 a, b, c, a_1, a_2, \dots , 代表语元; 2) 序列项 $\langle a_i, \dots, a_j \rangle (1 \leq i \leq j \leq n)$, 代表语元序列, 简记为 a_i^j , 其中 $a_k (i \leq k \leq j)$ 为语元符号; 3) 函数符号 C , 用于组成关联度表达式 $C(a_i^j, b|a_i^j) (1 \leq i \leq j \leq n)$, 代表语元序列 a_i^j 与语元 b 在语境 a_i^j 下的关联度, $C(a_i^j, b|a_i^j)$ 简写为 $C(a_i^j, b)$, a_i^j 简写为 a_i 。

L_C 有如下 2 条公理, 其中 a_i, b, c 是任意语元符号, $a_i^n = \langle a_1, a_2, \dots, a_n \rangle (1 \leq n)$ 是任意序列项。

公理 1 $0 \leq C(a_i, b|a_i^n) \leq 1$;

公理 2 $\bigwedge_{i=1}^n (C(a_i, b|a_i^n) \leq C(a_i, c|a_i^n)) \supset C(a_i^n, b) \leq C(a_i^n, c)$ 。

公理1称为语境关联度公理,它假定任何语境 a_i^n 下,语元 a_i 与另一语元 b 之间都有值在 $[0, 1]$ 中的关联度。在实际应用中,大模型训练中没有实际获得的所有关联度的值可虚设为0,从而在形式上满足公理1的要求,但可能对实际效果产生严重影响,引发非封闭性挑战(见第5节)。

公理2称为综合单调性公理,它假定语元序列 a_i^n 与语元 b 的关联度 $C(a_i^n, b)$ 由序列中每个语元 a_i ($i=1, 2, \dots, n$)与语元 b 的关联度 $c(a_i, b|a_i^n)$ 综合得出,并且这种综合具有单调性,即如果对每个 a_i 有 $C(a_i, b|a_i^n) \leq C(a_i, c|a_i^n)$,则 $C(a_i^n, b) \leq C(a_i^n, c)$ 。因此,综合单调性隐含着关联度的可综合性。

公理2还有一种稍显复杂的形式,针对的是预测在语元序列 a_i^n 的第 k ($1 \leq k \leq n$)个位置上出现的下一个语元,而不是预测整个序列之后出现的下一个语元,表达为

公理2' $\bigwedge_{i=1}^{k-1} (C(a_i, b|a_i^n) \leq C(a_i, c|a_i^n)) \wedge \bigwedge_{i=k+1}^n (C(a_i, b|a_i^n) \leq C(a_i, c|a_i^n)) \supset C(a_i^n, b) \leq C(a_i^n, c)$ 。

显然,公理2是公理2'当 $k=n+1$ 时的特例。然而就本文目的而言,只需针对公理2展开讨论就够了。

注意 L_C 不包含有关预测规则的公理。仅当需要考虑预测规则时,在 L_C 上加以补充相应的公理,如刻画预测的最大化规则的公理3,从而得到类 L_C 公理系统。

公理3 $\operatorname{argmax}_b c(a_i^n, b)$ 。

公理3称为预测选择公理,它规定选择使 $c(a_i^n, b)$ 取最大值的语元 b 作为预测结果。选用不同的预测规则,预测结果可能不同,但整个预测过程同样遵循 L_C 的公理1和公理2。大模型实验表明,不同的预测规则各有长短,适合于不同场合,不宜强求统一。所以,本文在 L_C 中不预置预测选择公理,而是需要时在 L_C 上进行补充,形成相应的类 L_C 。

在类 L_C 中,关联度预测包含2个主要步骤——关联度综合和依据给定预测规则的选择。假设采用最大化规则,对应的预测过程如下:1)任给语元序列 a_i^n ,利用公理1中的语元关联度实例 $c(a_i, b_j|a_i^n)$,其中 $i=1, 2, \dots, n, j=1, 2, \dots, m$,分别进行 m 次关联度综合,产生 m 个综合结果 $c(a_i^n, b_j), j=1, 2, \dots, m$;2)依据最大化规则,从这 m 个综合结果中选择使 $c(a_i^n, b)$ 的值最大的语元 $b = b_{j^*}$,以 b_{j^*} 作为对 a_i^n 下一个语元的预测结果。

类 L_C 为关联度预测规定了如下基本约束:1)对于语元,只要求它们是可区分的、相互之间存在关联度,对其他的一切(包括语元的直观含义)

都没有规定。2)对于语元关联度,要求其度量是归一化的和可综合的,且综合是单调的。3)类 L_C 中的推理(预测)模式是从给定语境下的一组关联度实例 $c(a_i, b|a_i^n)$ 出发,经过关联度综合和依据预测规则的选择,产生推理(预测)结果。

作为一种新型形式公理系统, L_C 有2个鲜明的特点:1) L_C 的形式语言和公理都不含变元,所以 L_C 具有实例性,而以往的公理系统都是用少量含有变元的规则概括大量实例。尤其值得注意的是, L_C 只含一条推理规则即公理2,同时拥有大量推理前提即公理1,它们在 L_C 的推理中发挥了巨大的、主要的作用。对比经典逻辑系统和传统AI推理系统,发挥主要作用的是推理规则,推理前提的作用相对较小。2) L_C 中的推理(预测)具有语境相关性,即关联度及关联度预测总是相对于语境的。因此,出现在语境 a_i^n 中的语元 a_i ($i=1, 2, \dots, n$)发挥着关键作用——激活公理1中的关联度实例 $C(a_i, b|a_i^n)$,于是 b 进入预测范围。假设 b 被选为预测结果,则在继续预测下一个语元 c 时, b 是 c 的预测语境 $a_i^{n+1} = \langle a_1, a_2, \dots, a_n, b \rangle$ 中的一个语元,并激活关联度实例 $C(b, c|a_i^{n+1})$ 。语境相关性是大模型激发(prompt)机制的形式化,该机制对大模型具有极其重要的意义。

实际应用中预测一个语元时,大模型可处理的语境至少包含几千个语元,甚至可达10万个语元。通过回看这些语元,激活大量远程关联度,根据对它们的综合对比来决定预测结果。公理1中的语元关联度是从人类规模的语料中提取的,所以公理1和公理2以实例化形式承载着人类的某些语言运用经验,从而使大模型得以直接利用这些经验。相对而言,人类的每一位成员通常只掌握人类语言运用经验的极小部分,而其余大部分则不在个人的掌握之中,导致很多问题无法回答。所以,大模型可以产生大量既出人预料、又符合人类语言运用经验的输出。由此可见, L_C 刻画了一种新型推理(预测)模式,是人的习惯性语言运用方式(见第1节)的形式化表达,是大模型产生出人预料效果的主要机制。

4 L_C 的语义结构分析

一个公理系统如果不是纯逻辑系统,其语义通常包含形式语义和实质语义2个部分。 L_C 的语义也包含形式语义和实质语义2个部分,这就是 L_C 的语义结构。

L_C 的形式语义是一个四元组 $\mathcal{S} = \langle \mathcal{D}, \llbracket a \rrbracket, \llbracket a_i^n \rrbracket, \llbracket C \rrbracket \rangle$,其中 \mathcal{D} 是一个给定的训练语料集,对

于 L_C 形式语言的每一类语法元素, 都有一个定义在 \mathcal{D} 上的解释函数, 包括:

1) 对任何语元符号 a , 其语义解释是函数 $\llbracket a \rrbracket$, 将 a 映射为 \mathcal{D} 的一个元素 (语元)。不同的 \mathcal{D} 和 $\llbracket a \rrbracket$ 产生 L_C 的不同语义解释。

2) 对任何序列项 a_i^j , 它的语义解释是函数 $\llbracket a_i^j \rrbracket$, 将语元符号序列 $\langle a_i, \dots, a_j \rangle$ 映射为 \mathcal{D} 上的语元序列 $\langle \llbracket a_i \rrbracket, \dots, \llbracket a_j \rrbracket \rangle$, 其中 $\llbracket a_k \rrbracket (i \leq k \leq j)$ 是 \mathcal{D} 中用来解释语元符号 a_k 的语元。

3) 对函数符号 C , 解释为一个函数 $\llbracket C \rrbracket$, 其定义域为 \mathcal{D} , 值域为实数区间 $[0, 1]$, 并且 $\llbracket C \rrbracket$ 与 $\llbracket a \rrbracket$ 、 $\llbracket a_i^j \rrbracket$ 的复合形成关联度表达式 $C(a_i^j, b|a_i^j)$ 的语义解释, 使得:

$$\llbracket C(a_i^j, b|a_i^j) \rrbracket = \llbracket C \rrbracket (\llbracket a_i^j \rrbracket, \llbracket b \rrbracket | \llbracket a_i^j \rrbracket), 1 \leq i \leq j \leq n \quad (1)$$

要求 $\mathcal{D} = \langle \mathcal{D}, \llbracket a \rrbracket, \llbracket a_i^j \rrbracket, \llbracket C \rrbracket \rangle$ 满足以下条件:

① \mathcal{D} 具有组合性, 即 $\llbracket C \rrbracket$ 与 $\llbracket a \rrbracket$ 和 $\llbracket a_i^j \rrbracket$ 的复合遵守式 (1);

② \mathcal{D} 满足公理 1 的要求, 即 $0 \leq \llbracket C \rrbracket (\llbracket a_i \rrbracket, \llbracket b \rrbracket | \llbracket a_i^j \rrbracket) \leq 1$;

③ \mathcal{D} 满足公理 2 的要求——如果对所有 $i=1, 2, \dots, n$, 有 $\llbracket C \rrbracket (\llbracket a_i \rrbracket, \llbracket b \rrbracket | \llbracket a_i^j \rrbracket) \leq \llbracket C \rrbracket (\llbracket a_i \rrbracket, \llbracket c \rrbracket | \llbracket a_i^j \rrbracket)$, 则 $\llbracket C \rrbracket (\llbracket a_i^j \rrbracket, \llbracket b \rrbracket) \leq \llbracket C \rrbracket (\llbracket a_i^j \rrbracket, \llbracket c \rrbracket)$ 。

在这 3 个条件下, $C(a_i^j, b|a_i^j)$ 的语义解释 $\llbracket C(a_i^j, b|a_i^j) \rrbracket$ 就是“语境 $\llbracket a_i^j \rrbracket$ 下语元序列 $\llbracket a_i^j \rrbracket$ 与语元 $\llbracket b \rrbracket$ 的关联度”的某种数学刻画。不难看出, 第②、③条件 (即公理 1、2) 规定了语元关联度的基本特性——语境相关的统计性质。

符合上述要求的函数 $\llbracket C \rrbracket$ (以及 $\llbracket a \rrbracket$ 和 $\llbracket a_i^j \rrbracket$) 不是唯一的, 如简单平均函数 $(\llbracket C \rrbracket (\llbracket a_i^j \rrbracket, \llbracket b \rrbracket) = \sum_{i=1}^n \llbracket C \rrbracket (\llbracket a_i \rrbracket, \llbracket b \rrbracket | \llbracket a_i^j \rrbracket) / n$ 、加权平均函数和其他很多综合函数都符合要求, 但它们用于预测的效率和性能差别很大。如从图 1 可知, 在预测语元序列“我要上网, 请打开”的下一个语元时, 语元“上网”的作用比序列中其他语元大得多, 所以简单平均用作 $\llbracket C \rrbracket$ 函数肯定效果不好。 L_C 的语义框架能否支持更好地分析不同 $\llbracket C \rrbracket$ 函数的效率和性能, 留给后续工作。

下面讨论 L_C 的实质语义。一般而言, 只要一个公理系统不是纯逻辑的, 它的形式语义就不能完全规定其含义, 不得不辅之以非形式的、直观的解释, 即实质语义, 以弥补形式语义的不足。如牛顿力学的公理系统由 4 条非逻辑公理组成, 即牛顿运动三定律和万有引力定律。牛顿本人对牛顿力学的理解, 绝不限于这 4 条公理及其形式

化解释, 即形式语义, 还包含对形式语义的直观理解, 即实质语义。

在 L_C 的形式语义中, 只考虑语元之间的关联度 (一种语境相关的统计性质), 忽略语元的任何其他性质, 包括语元在自然语言中携带的所有其他信息, 从而将语元归结为无直观含义的符号。另一方面, 在大模型与人交互的过程中, 大多数用户不可能将语元当作无直观含义的符号, 而是会不自觉地“脑补”——对大模型说出的任何语元及语元序列, 根据个人的语言习惯去理解语元的直观含义, 并默认个人理解的语元含义正是大模型理解和传递的语元含义。本文第 1 节的实验观察为这个判断提供了证据。

综合上述两方面的情况可知, 关联度预测的实质语义是一种弱共识性实质语义。也就是说, L_C 的实质语义仅限于语元之间的关联度, 不规定语元本身的直观含义, 允许不同用户对同一个语元做不同的理解。作为对比, 以往的公理系统的实质语义都是强共识性的, 即追求统一的直观理解。如对牛顿力学基本概念“力”、“质量”、“加速度”等的理解, 科学共同体不仅有共识, 而且这种共识具有强制性, 违反共识的个人理解是不被接受的。因此, 大模型以 L_C 的弱共识性实质语义颠覆了科学传统中的强共识性实质语义。

非概念化是 L_C 语义的另一个颠覆。在 2 500 多年的科学传统中, 任何学科的知识体系都是概念化 (conceptualization) 的。概念化是对一个学科领域或现实世界某个局部的抽象, 用一套专门化的概念系统描述该领域的现象, 进而做出逻辑判断——断定什么是真的、什么是假的、真假的依据是什么, 以形成该领域的一种学术主张。如牛顿力学是对宏观物理世界力学现象的概念化, 描述了宏观物理世界的力学规律, 符合规律的是真的, 不符合的是假的。以此为典范, 现代科学的几乎所有分支都试图仿照牛顿力学, 建立概念化的知识体系。

然而, L_C 只是利用语元相关性的统计度量回答用户提出的问题, 并不对问题和回答进行逻辑判断。对用户提出的问题 Q , L_C 给出回答 A , 是因为 L_C 认为 A 与 Q 是某种度量下最相关的, L_C 并不判断 Q 和 A 的真假, 而且 L_C 也没有逻辑判断的机制。由于放弃了概念化的核心功能——逻辑判断, L_C 不是概念化公理系统, 而是一个非概念化公理系统, 也是历史上第 1 个具有重大影响的非概念化公理系统。

5 大模型带来的新挑战

上文分析表明,语境相关的统计性质、弱共识性实质语义和非概念化公理系统是关联度预测的基础特性。明确这组基础特性有助于探讨大模型带来的前所未有的新挑战。

真实性挑战 众所周知,大模型的输出有时不是真实的,这对大模型的很多应用构成重大挑战。这个挑战的根源,不在于大模型判断真假的能力不足,而在于根本没有判断机制,因为关联度预测是非概念化的。

除非为大模型构建配套的逻辑判断机制,否则它不可能具备真正的逻辑判断能力。大模型的细调和其他后续优化对其真实性问题进行了各种改进,取得了一定进展。但现有证据显示,迄今尚未形成完整的逻辑判断机制。

下面的分析表明,真实性挑战还与共识挑战和内容属性挑战有关,所以不能孤立地研究真实性挑战、孤立地探索其解决方案。

共识挑战 通常意义下的共识是强共识,即面向所有交互者、关于所有重要内容的强制性共识。科学理论和大模型之前的人工智能追求强共识,前者成功了,后者没有成功。从人机交互的观点看,强共识不成功的原因在于用户之间没有强共识,所以人工智能与用户之间也没有强共识,既没有依据也没有权力让用户接受一种强共识。

大模型走了另一条路——弱共识,即以某种人类语言习惯为人机交互的共识,并由 L_C 的弱共识性实质语义体现这种人类语言习惯。目前看来,这种弱共识取得了一定程度的成功,这有3方面的原因。1)从人类规模的文本语料中,可以提取人类的语言痕迹——语元关联度,作为一种人类语言习惯的数据载体,用于基础模型的预训练。语元的直观含义具有更大的不确定性,目前不能从语料中提取对应的语言痕迹。2)利用这种语言痕迹,借助于关联度预测,可以自动生成问题的回答,这种回答在多数情况下符合人类的语言习惯。3)普通用户对自己和机器的语言运用采取习惯性评判(第1节),而大模型的多数回答恰恰符合这一评判标准。本文认为,这是大模型被普通用户广泛接受的真正原因。作为对比,传统软件都是基于强共识的,强制用户遵从软件定义的共识,所以用户必须经过专门的培训才能学会软件的使用。大模型一举颠覆了这一传统。

在人机交互中,有无可能制定比上述弱共识更强的共识?只要把大模型用作人机交互的产

品,把“用户是上帝”的信条用于这种产品,就立即排除了更强共识的可能性——每一个用户都有权坚持自己对语元直观含义的个人理解,而机器无权改变用户的理解。所以在上述用户观念下, L_C 的弱共识性实质语义是一种最佳选择。

由上述分析得到的一个直接推论是相对于现实世界的真实性和相对于语料的真实性2种真实性概念,在人机交互的一些场景下是相互矛盾的;也就是说,遵从用户理解与符合客观真实有时是冲突的。所以真实性挑战不是一个单纯的科学问题,人工智能也不是一个纯粹客观的研究领域。事实上,智能机器^[8]是非人非物的第3种存在物^[9]。

内容属性挑战 大模型输出的内容是什么属性,是信息、知识还是别的什么?目前没有定论。一个明显的观察是大模型输出的内容十分丰富,大大超出了信息的范围。不过,大模型的输出也不符合通常对知识的定义。人工智能界接受英国哲学家罗素的观点,以真实作为知识的必要条件(即“知识公理”^[10])。由于大模型没有判断真假的机制,导致大模型无法保证自己的输出是知识。

目前看来,将大模型的输出视为“重构资讯”,更接近现实情况。因为大模型是以训练语料中提取的语言痕迹为基础的,所以输出的内容是语料中资讯的重新组织和表述,有可能改变了语料中的原始表述及其含义,但总体上是符合人类语言习惯的。

重构资讯有2种真实性概念:1)相对于现实世界的真实性,类似于牛顿力学等科学理论的真实性。2)相对于训练语料的真实性,这种真实性保证大模型的输出与它的训练语料是一致的,不保证输出内容在现实世界中是真实的。大模型的不同应用需要不同的真实性,这给真实性挑战带来更大的复杂性。

大模型输出的内容属性对于大模型的应用方式形成刚性制约。既然输出内容不是知识,就不宜简单地将大模型用于问答等知识型应用。除了聊天,在大多数应用场景中,大模型的输出需要有资格的人进行内容审核,以控制可能的风险。

非封闭性挑战 人工智能研究涉及的各种数学理论隐含着各自的封闭性假设,而人工智能的应用场景通常不符合任何封闭性假设,这一矛盾称为人工智能的非封闭性挑战^[11]。如Kolmogoroff概率公理要求一个样本空间中所有样本点上的概率之和为1。然而,如果一个应用场景是非封闭的,则只能确定部分样本点,并依据上述

公理设置这些样本点上的概率之和为 1。但由于这个样本空间有遗漏,所以其中样本点的概率之和实际上不为 1,即部分样本点的概率是不真实的,由此推出的结论一定存在错误,而且难以判断错在哪里。这意味着,非封闭性破坏了 Kolmogoroff 概率论的应用条件。这种情况对其他数学理论同样存在,是人工智能基础研究中的一个深层挑战。

大模型同样面临着非封闭性挑战。 L_C 的公理 1 要求对所有语元 a_i 和 b 、语元序列 a_i^n , 语境 a_i^n 下 a_i 和 b 之间存在关联度 $C(a_i, b|a_i^n)$ 。而在实际应用中,即使将人类现有全部语言记录用作训练语料,仍远远不足以满足公理 1 对训练数据的要求。所以,一定有大量关联度 $C(a_i, b|a_i^n)$ 的值被虚设为 0,以便形式上满足公理 1 的要求。于是,这些虚设关联度就如同上文中遗漏的样本点一样,破坏了 L_C 的应用条件,导致 L_C 的推理(预测)无法避免产生不真实的结果。这意味着,大模型真实性不是一项单纯的工程挑战,同时也是基础理论挑战。

致谢 感谢马长山、殷实为本文实验及资料收集提供的帮助。

参考文献:

- [1] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model[J]. NIPS, 2000: 932–938.
- [2] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[EB/OL]. (2013–10–16)[2023–06–20]. <https://arxiv.org/abs/1310.4546>.
- [3] VASWANI A, SHAZEER N M, PARMAR N, et al. Attention is all you need[EB/OL]. (2017–06–12)[2023–06–20]. <https://arxiv.org/abs/1706.03762>.
- [4] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. NAACL-HLT (1), 2019: 4171–4176.
- [5] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[EB/OL]. (2020–05–28)[2023–06–20]. <https://arxiv.org/abs/2005.14165>.
- [6] OUYANG Long, WU J, JIANG Xu, et al. Training language models to follow instructions with human feedback [EB/OL]. (2020–03–04)[2023–06–20]. <https://arxiv.org/abs/2203.02155>.
- [7] PETRONI F, ROCKTASCHEL T, LEWIS P, et al. Language models as knowledge bases? [EB/OL]. (2019–09–03)[2023–06–20]. <https://arxiv.org/abs/1909.01066>.
- [8] TURING A. Intelligent machinery[M]. The Turing digital archive, 1948.
- [9] 陈小平. 人工智能伦理建设的目标、任务与路径: 六个议题及其依据 [J]. 哲学研究, 2020(9): 79–87, 107, 129.
CHEN Xiaoping. The target, tasks, and implementation of artificial intelligence ethics: six issues and the rationale behind them[J]. Philosophical research, 2020(9): 79–87, 107, 129.
- [10] NILSSON N J. Artificial intelligence: A new synthesis[M]. Beijing: China Machine Press, 1999.
- [11] 陈小平. 人工智能中的封闭性和强封闭性: 现有成果的能力边界、应用条件和伦理风险 [J]. 智能系统学报, 2020, 15(1): 114–120.
CHEN Xiaoping. Criteria of closeness and strong closeness in artificial intelligence—limits, application conditions and ethical risks of existing technologies[J]. CAAI transactions on intelligent systems, 2020, 15(1): 114–120.

作者简介:



陈小平, 教授、机器人实验室主任, 广东省科学院人工智能首席科学家, 中国人工智能学会人工智能伦理与治理专委会主任, 中国管理科学学会大数据管理专委会副主任。曾任 2015 世界人工智能联合大会 (IJCAI2015) 机器人领域主席、2008 和 2015 机器人世界杯及学术大会 (RoboCup2008, 2015) 主席、Journal of Artificial Intelligence Research 和 Knowledge Engineering Review 编委。

[责任编辑: 李雪莲]