自然语言处理中的探针可解释方法综述

鞠天杰[1)] 刘功申[1)] 张倬胜[1)] 张 茹[2)]

[1)](上海交通大学网络空间安全学院 上海 200240)

[2)](北京邮电大学网络空间安全学院 北京 100876)

摘 要 随着大规模预训练模型的广泛应用,自然语言处理的多个领域(如文本分类和机器翻译)取得了长足的发展.然而,受限于预训练模型的"黑盒"特性,其内部的决策模式以及编码的知识信息被认为是不透明的.以OpenAI发布的ChatGPT和GPT-4为代表的先进预训练模型为例,它们在多个领域取得重大性能突破的同时,由于无法获知其内部是否真正编码了人们期望的知识或语言属性,以及是否潜藏一些不期望的歧视或偏见,因此仍然无法将其应用于重视安全性和公平性的领域.近年来,一种新颖的可解释性方法"探针任务"有望提升人们对预训练模型各层编码的语言属性的理解.探针任务通过在模型的某一区域训练辅助语言任务,来检验该区域是否编码了感兴趣的语言属性.例如,现有研究通过冻结模型参数并在不同层训练探针任务,已经证明预训练模型在低层编码了更多词性属性而在高层编码了更多语义属性,但由于预训练数据的毒性,很有可能在参数中编码了大量有害内容.该文首先介绍了探针任务的基本框架,包括任务的定义和基本流程;然后对自然语言处理中现有的探针任务方法进行了系统性的归纳与总结,包括最常用的诊断分类器以及由此衍生出的其他探针方法,为读者提供设计合理探针任务的思路;接着从对比和控制的角度介绍如何解释探针任务的实验结果,以说明探测位置编码感兴趣属性的程度;最后对探针任务的主要应用和未来的关键研究方向进行展望,并讨论了当前探针任务亟待解决的问题与挑战.

关键词 探针任务;可解释;自然语言处理;预训练模型;深度学习;人工智能安全

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2024.00733

A Review of Probe Interpretable Methods in Natural Language Processing

JU Tian-Jie[1)] LIU Gong-Shen[1)] ZHANG Zhuo-Sheng[1)] ZHANG Ru[2)]

[1)](*School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai* 200240)

[2)](*School of Cyber Science and Engineering, Beijing University of Posts and Telecommunications, Beijing* 100876)

Abstract The widespread adoption of large-scale pre-trained models in multiple fields, particularly in natural language processing such as text classification and machine translation, has paved the way for remarkable advancements. Nonetheless, due to the "black box" nature of pre-trained language models, the internal decision patterns and encoded knowledge information are considered to be opaque. While advanced pre-trained language models such as ChatGPT and GPT-4 released by OpenAI have achieved significant performance breakthroughs in various domains, they may not be appropriate for fields that place high importance on security and fairness. This is attributed to the difficulty in verifying if these models inherently encode the desired knowledge and language properties without entailing any internal biases or discrimination. In the pursuit of better understandability and transparency of pre-trained models, a new interpretable scheme known as

收稿日期:2023-04-03;在线发布日期:2023-12-16. 本课题得到社会治理与智慧社会科技支撑重点专项(2023YFC3303805)、国家自然科学基金联合重点项目(U21B2020)、科技创新2030——新一代人工智能重大项目(2022ZD0120304)、上海市科技计划项目(22511104400)资助.**鞠天杰**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、深度学习可解释性. E-mail:jometeorie@sjtu.edu.cn. **刘功申**(**通信作者**),博士,教授,中国计算机学会(CCF)会员,主要研究领域为人工智能安全、自然语言处理. E-mail:lgshen@sjtu.edu.cn. **张倬胜**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为深度学习、自然语言处理、预训练模型. **张 茹**,博士,教授,主要研究领域为数字内容安全.

the "probing task" has emerged in recent years. This task promises to enhance our understanding of the linguistic properties encoded in each layer of pre-trained models. It assimilates model outputs from arbitrary positions as input, employing a probing model for training auxiliary linguistic tasks (e.g., part-of-speech tagging, dependency parsing), and subsequently gauges the degree to which specific linguistic properties are encoded within the layer under analysis based on the auxiliary model's performance on the test set. For example, existing studies have demonstrated that pre-trained models encode more lexical properties at lower layers and more semantic properties at higher layers by freezing the model parameters and training the probing task at different layers. However, due to the toxicity within pre-training data, there is a significant possibility that the parameters encode a substantial amount of harmful content. Our review begins with an introduction to the basic framework of the probing task, where we delve into the definition of probing tasks and outline the basic workflow of carrying out such a task. Then we systematically summarize existing schemes for probing tasks in natural language processing, including the most commonly used diagnostic classifiers and other probing methods derived from them (structural probing, intervention-based probing and prompt-based probing) to provide readers with ideas for designing reasonable probing tasks. For diagnostic classifiers, we also focus on the selection of probing model complexity and probing datasets to guide the design of more reliable probe experiments. After that, we describe how to interpret the experimental results of probing tasks from the perspective of comparisons and controls to illustrate the extent to which the probing position encodes properties of interest. Finally, as we come to the end of the review, we take stock of the main applications and discuss potential key research directions to be pursued. We further ruminate on the current issues and challenges that the field of probing tasks faces and needs to address. Undeniably, as a relatively novel area of research, extant probing methods remain insufficiently mature, encompassing both theoretical shortcomings inherent in the design of probing tasks themselves and an inadequacy in exploring more intricate linguistic properties. This paper aspires to furnish readers with a comprehensive "diagnostic report" concerning ongoing probing task research, while advocating for increased scholarly investment in pertinent domains.

Keywords probing task; interpretability; natural language processing; pre-trained model; deep learning; artificial intelligence security

1 引 言

自然语言处理(Natural Language Processing, NLP)是人工智能(Artificial Intelligence, AI)领域的重要研究方向,其研究核心包括语言建模、词法分析、句法分析和语义分析[1].近年来,受益于深度学习强大的拟合能力,NLP 各领域的性能普遍取得显著提升.OpenAI 在 2023 年最新发布的 GPT-4[2]模型更是在多类任务中达到甚至超过了人类水平,实现了迈向通用人工智能的第一步,并有望兴起新一轮的 NLP 研究热潮.

然而,深度学习在各领域取得高性能的同时,往往需要海量的模型参数.例如,谷歌在 2018 年发布的两个 BERT[3]预训练模型参数量已分别达到 1.1 亿和 3.4 亿,OpenAI 在 2020 年发布的 GPT-3[4]预训练模型参数量达到 1750 亿,而尽管 GPT-4 的参数量并未公开,其在各方面出色的表现使得人们普遍对其抱有更高的参数量预期.这些参数使得模型变得更加不透明,并引起人们的担忧.从安全角度考虑,由于无法确定模型内部的工作机理,人们对模型输出的信任度有限[5],使得模型难以被部署于医疗、金融、军事等对安全有着极高要求的领域;从性能的角度考虑,模型的大小限制了人们进行细粒度消融实验的能力,从而无法进一步优化结构[6].

为了解决复杂模型的不透明性引起的一系列问题,越来越多的研究致力于模型可解释性的探索[5-10],即对复杂的模型行为作出人们所能理解的

解释,如特征重要度解释[11-12]、生成式解释[13-14]和反事实解释[15-16].这些通用的可解释方法展示了模型输出某一结果的原因,却没有分析模型内部究竟编码了什么信息.而相比于图像、视频等其他领域,自然语言天生地具有各种语言属性和现象(如词性、时态、句法成分、复杂语义等),被人们归纳总结后用于教学和日常生活中.因此,对于 NLP 任务,人们迫切想要了解一个端到端训练的语言模型是否在对原始输入编码后的表征中包含了期望的词法、句法、语义属性,以此验证模型的可信度和鲁棒性;以及确定模型是否利用数据集中潜在的偏置嵌入了歧视或有害的信息,对社会秩序造成负面影响.以"探针任务"为代表的可解释方法有望同时满足这些需求.

探针任务接受模型任意位置的输出作为输入,利用插入的辅助模型训练辅助语言任务(如词性标注、依存句法树分析、时态分析),并根据辅助模型在测试集上的性能得出待分析层中编码某种语言属性或现象的程度.以图 1 所示的探针任务为例,可以将隐藏层 1 的编码结果作为辅助模型(例如线性分类器)的输入,以训练额外的词性标注任务,从而探测该层编码词性相关语言属性的程度.这种实用而有效的可解释性方法近年来在 NLP 领域引起了广泛的关注.

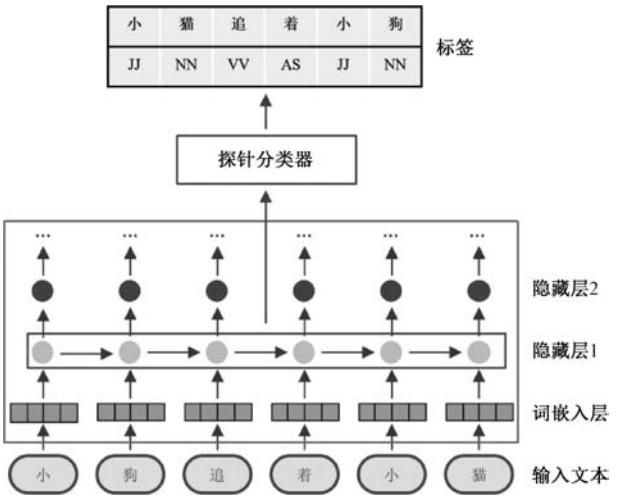

图 1　探针任务示例

在本文中,我们首先对探针任务的框架进行明确的定义;接着根据不同探针模型的结构进行了分析和归类,包括最常用的诊断分类器以及其他 AI 领域的技术在探针任务上的应用(如表 1 所示);然后介绍对探针任务性能的解释,以回答探测区域究竟编码了多少相关的语言属性信息;最后从多个角度对未来的关键研究方向进行展望,并对文章内容进行总结.

表 1　常见探针方法分类

方　法	特　点
诊断分类器	易于实现,解释性强,但受分类器复杂度和数据集等干扰因素影响.
结构探针	用于对具有树结构或图结构的语言属性的探测,具有一定局限性.
基于干预的探针方法	可以测试细粒度语言属性作用于模型决策的因果效应,但往往具有较高复杂度.
基于提示的探针方法	无需中间层表征,适用于未公布参数的大模型探针,但检测的因果效应有待验证.
无参数探针方法	消除了额外引入的参数对探针结果的影响,但可信的探针结果依赖于无参数模型方法本身的设计,目前缺乏统一标准.

2　探针任务框架

2.1　术语和标记

本节对探针任务中所使用的术语进行基本定义,并给出在后文中使用的对应标记.

(1)原始模型 M :待探测语言属性的模型;

(2)语言属性 Z :待探测的属性,例如词性、时态等;

(3)探针任务 T :在模型 M 上所训练的辅助语言任务,例如词性标注、语义推理等;

(4)原始数据集 $D=\{x,y\}$:用于在 M 上训练原始任务的数据集,x 和 y 分别表示一组训练数据的输入和对应标签;

(5)中间表征 R_ℓ :输入数据在 M 第 ℓ 层上的表征,根据探针任务的不同可以分为词向量和句向量;

(6)探针模型 M_p :用于训练探针任务以检验某种语言属性的编码程度的模型;

(7)探针数据集 $D_p=\{x,z\}$:用于训练辅助语言任务的数据集,z 表示 x 对应的某种语言属性标签;

(8)探针输出 R_p :输入数据在探针模型中的输出;

(9)探针性能 $P(M,M_p,D,D_p)$:探针模型在探针数据集上训练后的性能.

2.2　任务定义

探针可解释方法采用探针任务进行解释,通过在预训练模型的某一组件中插入并训练辅助模型来检测特定语言属性.探针任务在形式上与一般 NLP 任务类似,它利用 M_p 训练输入的中间层表征到语言属性的映射 $R_\ell \to Z$,最终根据探针性能 $P(M,M_p,D,D_p)$ 推测语言属性 Z 的编码程度,这一过程中 M 的参数始终保持冻结.如果在测试集上具有较高的探针性能,则证明表征 R_ℓ 与语言属性 Z 之

间存在良好的映射关系，即 M 在端到端的预训练或微调中编码了相关语言属性；反之则说明 R_ℓ 与 Z 之间具有较低的相关性，即缺少对相关语言属性的编码.

随着该领域研究的深入，Pimentel 等人[17]提出可以从信息论角度定义探针任务的过程，即量化 R_l 和 Z 之间的互信息 $I(R_\ell;Z)$. Zhu 等人[18]在交叉熵损失作为探针任务优化目标的前提下，进一步对互信息量进行拆解，以解释用交叉熵损失近似互信息的误差来源. 具体来说，属性与任务间的互信息可以拆解为

$$I(R_\ell;Z)=H(Z)-H(Z\mid R_\ell) \tag{1}$$

而条件熵 $H(Z\mid R_\ell)$ 又可以继续拆解为

$$\begin{aligned}H(Z\mid R_\ell)&=-\mathbb{E}_{p(Z|R_\ell)}\log p(Z\mid R_\ell)\\&=-\mathbb{E}_{p(Z|R_\ell)}\log\frac{p(Z\mid R_\ell)q_\theta(Z\mid R_\ell)}{q_\theta(Z\mid R_\ell)}\\&=H(p,q_\theta)-\mathrm{KL}(p\,||\,q_\theta)\end{aligned} \tag{2}$$

因此

$$H(p,q_\theta)=H(Z)-I(R_\ell;Z)+\mathrm{KL}(p\,||\,q_\theta) \tag{3}$$

其中 $p=p(Z\mid R_\ell)$ 表示真实的未知概率分布，$q_\theta=q_\theta(Z\mid R_\ell)$ 是探针模型对真实概率分布的近似.

也就是说，探针模型的交叉熵 $H(p,q_\theta)$ 可以分解为独立于表征的常数项 $H(Z)$、表征和语言属性之间的互信息 $I(R_\ell;Z)$（希望探测的内容）和探针模型本身捕获的任务内容 $\mathrm{KL}(p\,||\,q_\theta)$（误差来源）. 这些研究为系统性地解释与分析探针任务结果提供了理论依据.

2.3 基本流程

根据定义，要想设计一个良好的探针任务，需要遵循如下步骤：

(1)设计合理的问题

实验开始之前，首先要确定感兴趣的任务内容，包括探测的位置和语言属性. 探测的位置可以是词嵌入[19-20]、句嵌入[21-22]、模型的中间层输出[23-29]、跨层的标量混合表征[30-31]等. 根据探测位置可能包含的信息，再设计感兴趣的语言属性. 语言属性又可以简单分为表面属性、词法属性、句法属性和语义属性. 例如，可以研究词嵌入中是否编码了词性或时态相关的词法属性[20]、句嵌入中是否包含了实体间关系的语义属性[32]. 实际中，我们往往同时对多个位置和语言属性感兴趣，但一次训练周期一般只能选定一个位置和语言属性进行探测.

(2)选择可靠的探针数据集

数据集作为探针任务的近似替代，需要经过细致分析和选择. 受限于数据集大小[29]和可能存在的偏置[33]，Ravichander 等人[34]指出数据集难以真正做到探针任务的替代. Choudhury 等人[33]在问答任务的探针实验中进一步验证探针数据集中偏置的存在，Kuznetsov 等人[35]则在角色语义任务中证实了不同探针数据集中的语言形式主义对探针结果有显著影响. 且由于语言属性的复杂多元性，难以定义统一的标杆数据集. 因此，如何选择可靠的数据集作为探针任务的近似是实验设计的难点.

(3)构建合适的探针方法

一般的探针任务采用诊断分类器[26]进行训练，即训练一个接受输入表征并预测某些语言属性的分类器. 如果分类器表现良好，则说明输入表征学习到了特定属性[36]. 这类方法的研究重点在于如何选择具有合适复杂度的分类器，使其能够揭示探测区域编码特定属性的程度的同时，不会由于自身的学习能力干扰到实验结果[34]. 近年来，一些研究开始尝试越过诊断分类器的范式进行探索，例如提示学习[37]、因果干预[38]等，使得探针任务获得了新的发展.

(4)提供客观的解释

探针任务训练结束后，往往会从测试集上得到探针性能，例如词性标注的准确率. 然而，探针性能本身并不能作为很好的对探针结果的解释，且因为数据集的难易程度不同，仅凭 70%、80%等准确率数值作为结果不具备说服力. 因此，需要设计合格的上下界作为对比，或通过控制任务[39]排除可能存在的干扰项对结果做出解释.

目前，探针任务方法的构建和对结果的解释是现有研究的热点，将在下文中重点介绍. 而数据集的设计与选择，以及其与探针任务本身的关联是该领域的痛点与难点，需要未来投入更多的研究.

3 探针方法

探针方法是探针任务的核心，包括辅助模型的构建与数据集的选择. 本节总结了现有针对自然语言处理的探针方法的相关研究和进展，并根据不同方法的特点进行了归纳与分析. 首先重点介绍了探针任务的主流方法诊断分类器，包括对近年来应用诊断分类器进行探测的文章的总结，对模型复杂度权衡的讨论以及数据集的选择；接着依次介绍了近年来涌现的全新探针任务方法，包括结构探针、基于干预的探针、基于提示的探针以及无参数探针.

3.1　诊断分类器

3.1.1　探测方法

诊断分类器是探针任务中最为经典和常用的方法,它通过在冻结参数后的探测位置插入分类器并训练辅助语言任务实现对感兴趣的语言属性的检测.辅助语言任务一般以与任务高度相关的有标签数据集作为替代.例如图 1 中,用标注了每个字对应的词性标签的数据集作为辅助任务的近似,来探测原始模型的隐藏层中编码词性属性的程度.直觉上,分类器在测试集上的准确率越高,意味着分类器的输入,也就是探测位置所编码的相关语言属性越多;如果分类器的准确率很低,甚至接近于随机预测等基线,则意味着探测位置几乎没有编码该语言属性.

早期研究可以追溯到 2015 年.Gupta 等人[19]对 Skip-Gram[40]训练出的词向量进行探针实验,通过训练一个逻辑回归分类器检测词向量中编码的一般概念知识;Köhn 等人[20]采用 L_2 正则线性分类器分析了 6 种多语言静态嵌入编码的词法和句法属性.作为探针任务的试点实验,这些工作为诊断分类器的可行性奠定了基础.

随着研究的深入,研究人员尝试将诊断分类器应用于 NLP 的各种任务中.Shi 等人[24]首先研究基于 LSTM[41]的机器翻译模型,探究了模型编码器学习词语层面和句子层面各种语法信息的程度;Belinkov 等人[42]针对 LSTM 翻译模型进一步探测,发现较低层表征编码了更多词性属性,而较高层编码了更多语义属性;Xu 等人[43]则通过诊断分类器推翻了翻译任务中编码器层捕获源信息,解码器层负责传输的假设.Van 等人[44]聚焦于在问答任务上微调的 BERT 模型,采用多层感知机(Multilayer Perceptron,MLP)分析了隐藏层中的语义属性;Choudhury 等人[33]进一步在问答任务上探测,发现微调后编码的语义信息不会增多,并假设探针模型可能错误地利用了数据集中的偏置.对话任务中,Wu 等人[45]详细评估了 12 个模型在 4 个面向任务的核心对话任务的探针结果,证实了其中编码的语义属性.关系提取任务中,Alt 等人[46]介绍了 14 个针对关系提取相关语言属性的探针任务,并使用它们研究了 40 多个不同编码器的语言属性.

除了以上针对单一领域的分析,一些研究提供了全方位的分析工具,并作为诊断分类器的标杆被用于后续工作.例如,Adi 等人[22]围绕句子结构的孤立方面(句子长度、单词内容和单词顺序)定义探针任务,并训练 MLP 测试了 CBOW[40]、LSTM 和 Skip-Thought(ST)[47]向量各属性的编码程度,成为早期研究的标杆之一;Conneau 等人[23]设计了表面属性、句法属性和语义属性在内的 11 种探针任务,同样被许多研究采纳;Liu 等人[48]研究了 ELMo[49]、GPT[50]和 BERT 表征在 16 个探针任务上的结果;Tenney 等人[32]提出边缘探针(如图 2 所示),通过接受一个或两个文本片段的跨度表征作为输入训练探针模型,探究词法、句法和语义共计 11 种语言属性.目前,边缘探针已成为诊断分类器首选的基线之一.以上研究多数集中于 2015 年至 2019 年,总结于表 2 中,包括采用的探针模型结构(在下一小节中重点讨论),待探测的语言模型、位置、属性以及利用诊断分类器发现的基本语言属性以外的内容.

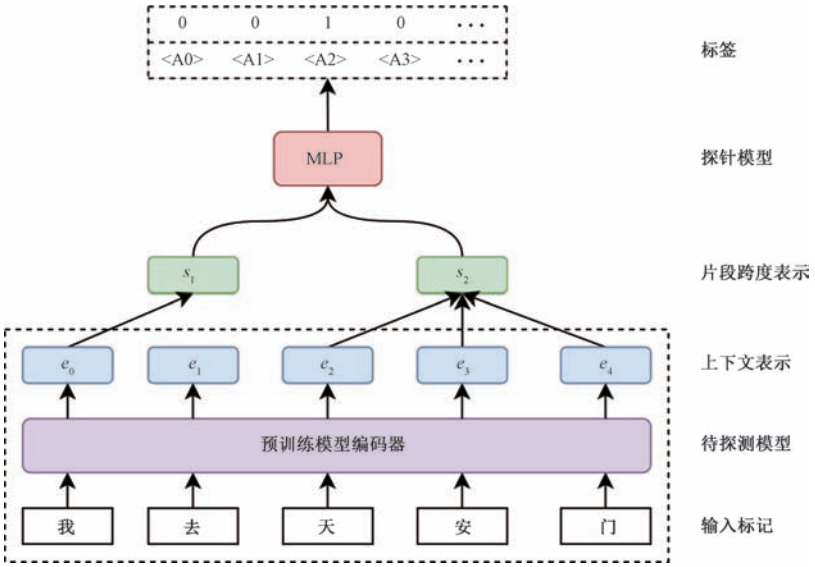

图 2　边缘探针模型结构

最后，一些研究尝试利用诊断分类器检测语言属性以外的内容，并发现了一些有趣的现象. 例如，Dalvi 等人[51]采用诊断分类器分析序列标记和序列分类任务，发现 BERT 和 XLNet[52]每一层都编码了关于任务的足够知识，意味着存在特定于任务的冗余. Michael 等人[53]提出正则化后的潜在子类学习对诊断分类器的输入进行潜在分类，以发掘模型涌现出的细粒度结构. 例如在以命名实体标注作弱监督信号的前提下，ELMo 和 BERT 分别涌现出人格概念和时间概念. Mohebbi 等人[54]将基于梯度的归因分析法[55]应用于诊断分类器，发现一些表征在特定语言任务中起到的关键作用（例如“[SEP]”和“.”对于句子长度预测）. Ousidhoum 等人[56]使用搜集的仇恨和攻击性言语作为探针数据集，用于量化语言模型潜藏的有害内容. Aghazadeh 等人[57]构造包含隐喻和非隐喻的探针数据集，用于检测预训练模型是否学会了对人类语言中的隐喻知识进行编码. Gurnee 等人[58]首次对大语言模型编码连续事实的能力进行了探索，证明了大语言模型在中间层就已经编码了时间和空间等结构化知识. Azaria 等人[59]通过对大语言模型中间层表征进行探测，发现可以很容易区分陈述的真实性，即模型可能知道自己是否在“撒谎”. 这些研究启发人们利用诊断分类器对语言模型背后的机制进行更广泛的探索.

表 2 诊断分类器方法总结（2015 年至 2019 年）

方法	探针模型	待探测语言模型	待探测位置	待探测属性				其他发现
				表面	词法	句法	语义	
Gupta, et al.[19]	LR	Skip-Gram[40]	词嵌入	×	×	×	√	—
Köhn, et al.[20]	LR	CBOW, Skip-Gram, dep[60], GloVe[61], cca[62], brown[63]	词嵌入	×	√	√	×	—
Ettinger,et al.[21]	LR	GloVe, Paragram[64], ST	句嵌入	×	×	×	√	—
Shi, et al.[24]	LR	LSTM	中间层输出	×	√	√	×	—
Qian, et al.[65]	MLP	CBOW, Skip-Gram, LSTM	词嵌入	×	√	√	√	探针模型结构影响
Adi, et al.[22]	MLP	CBOW, LSTM, ST	句嵌入	√	×	×	×	—
Belinkov,et al.[25]	MLP	LSTM, CharCNN[66]	中间层输出	×	√	×	×	探针模型结构影响
Hupkes,et al[26]	LR	GRU[67]	中间层输出	×	×	×	×	模型运行策略
Blevins, et al.[27]	MLP	LSTM	中间层输出	×	√	√	×	—
Peters, et al.[28]	LR	LSTM, Transformer[68], Gated CNN[69]	中间层输出	×	√	√	×	—
Conneau, et al.[23]	MLP	LSTM, Gated CNN	句嵌入	√	×	√	√	—
Zhang, et al.[29]	MLP	LSTM	中间层输出	×	√	√	×	数据量对探针任务的影响
Belinkov, et al.[42]	MLP	LSTM	中间层输出	×	√	×	√	—
Tenney,et al.[32]	MLP	CoVe[70], ELMo, GPT, BERT	词嵌入	×	√	√	√	—
Liu, et al.[48]	LR, MLP	ELMo, GPT, BERT	中间层输出	√	√	√	√	模型/层的迁移性、探针模型结构影响
Jawahar, et al.[71]	MLP	BERT	中间层输出	√	×	√	√	—
Tenney, et al.[30]	MLP	BERT	跨层标量混合表征	×	√	√	√	各层重要度、各层对任务的增益
Yaghoobzadeh, et al.[72]	LR, MLP, KNN	Skip-Gram, Structured Skip-Gram[73]	词嵌入	×	×	×	√	探针模型结构影响
Chen, et al.[74]	LR	BERT, ELMo	中间层输出	×	×	√	√	—
Van, et al.[44]	MLP	BERT	中间层输出	×	×	×	√	—
Clark, et al.[75]	LR	BERT	中间层输出	×	×	√	×	注意力机制分析
Warstadt[76]	LR	LSTM, GloVe, ELMo	句嵌入	×	×	√	√	—

注：LR＝Logistic Regression（逻辑回归），MLP＝Multi-Layer Perception（多层感知机），√＝包含，×＝不包含，—＝不存在.

由于诊断分类器易于实现且能够得到相对直观的解释，因此已被广泛应用于各种 NLP 模型语言属性的探测. 这类工作往往将重心和创新点聚焦于探针范式的第一步，即设计合理的问题. 不同工作设计的实验主要区别于三个维度：待探测模型、待探测位置、待探测语言属性和其他语言现象. 表 3 总结了 2020 年至今的前沿工作研究. 对比表 2，相关研究逐步由对传统小模型（如 CBOW）的探测转向对大规模预训练模型（如 BERT）的探测，这与 NLP 领域的发展路径相吻合. 另一方面，随着基本语言属性被逐步研究透彻，近年来更多工作尝试探测编码在模型中的其他语言现象（如涌现性、公平性）. 遗憾的是，基于诊断分类器的探针方法需要获取模型的中间层表征，因此对于尚未公开参数的 ChatGPT、GPT-4 等大模型无法采用这种方法.

表 3　诊断分类器方法总结(2020 年至今)

方法	探针模型	待探测语言模型	待探测位置	待探测属性				其他发现
				表面	词法	句法	语义	
Merchant, et al.[31]	MLP	BERT	跨层标量混合表征	×	√	√	√	—
Dalvi,et al.[51]	LR	BERT, XLNet	中间层输出	×	√	×	√	特定于任务的冗余
Chiang, et al.[77]	LR	ALBERT[78]	中间层输出	×	√	√	√	—
Wu, et al.[45]	LR	BERT, ALBERT, DistilBERT[79], RoBERTa[80], GPT-2[81], ELECTRA[82], ConveRT[83], DialoGPT[84], TOD-BERT[85], TOD-GPT	句嵌入	×	×	×	√	—
Xu, et al.[43]	Transformer	Transformer	中间层输出	×	×	×	×	翻译任务中编码器和解码器的分工
Klafka, et al.[86]	MLP	BERT, ELMo, GPT	词嵌入	×	×	√	√	词嵌入对周围单词信息的编码程度、探针模型结构影响
Edmiston, et al.[87]	LR, MLP	BERT	中间层输出	×	√	×	×	探针模型结构影响
Alt,et al.[46]	LR	CNN[88−89], LSTM, GCN[90], Transformer	句嵌入	√	×	√	×	架构和语言特征对探针任务的偏见
Şahin, et al.[91]	MLP	Word2Vec[39], FastText[92], GloVe, MUSE-supervised[93], ELMo	词嵌入	√	√	√	√	—
Kuznetsov, et al.[35]	LR	BERT	词嵌入、句嵌入	×	×	×	√	语言形式主义对探针结果的影响
Michael, et al.[53]	LR	BERT, ELMo	词嵌入	×	×	√	√	涌现的细粒度结构
Zhang, et al.[94]	MLP	RoBERTa	词嵌入	√	√	√	√	预训练数据量对语言属性的影响
Lyu, et al.[95]	LR	BERT, RoBERTa, BART[96], GPT-3	词嵌入	×	×	×	√	—
Mohebbi,et al.[54]	MLP	BERT, RoBERTa	句嵌入	√	×	√	√	特定标记在一些任务中起重要作用
Liu, et al.[97]	LR	RoBERTa	词嵌入	×	√	√	×	—
Choudhury, etal.[33]	MLP	BERT	词嵌入	×	√	×	√	探针模型可能错误利用了数据集偏置
Ousidhoum, et al.[56]	LR	BERT, RoBERTa, GPT-2	句嵌入	×	×	×	√	模型潜藏有害内容
Shapiro, et al.[98]	LR	BERT	词嵌入	×	√	√	×	跨语言编码的形态句法特征
Aghazadeh,et al.[57]	MLP	BERT, RoBERTa, ELECTRA	词嵌入	×	×	×	√	预训练模型掌握隐喻知识的程度
Conia, et al.[99]	LR, MLP	BERT, RoBERTa,	跨层标量混合表征	×	×	×	√	—
Gurnee, et al.[58]	LR, MLP	Llama2[100], Pythia[101]	中间层输出	×	×	×	√	预训练模型掌握结构化知识的程度
Azaria, et al.[59]	MLP	BERT, OPT[102], Llama 2	中间层输出	×	×	×	√	模型可能知道自己是否在“撒谎”

注:LR=Logistic Regression(逻辑回归),MLP=Multi-Layer Perception(多层感知机),√=包含,×=不包含,—=不存在.

3.1.2　复杂度权衡

尽管诊断分类器在语言属性探针任务中被广泛应用,但是关于该使用简单的单层线性分类器还是复杂的多层非线性分类器,学术界一直有较大的争议.

早期研究往往忽略分类器复杂度对探针结果的影响,而是从经验性角度选择分类器结构[103]. Qian 等人[65]作为较早探索分类器结构对结果影响的论文,发现复杂的多层非线性分类器的检测结果往往优于简单的线性分类器,但两者的大致趋势基本一致,类似的实验现象也被 Belinkov 等人[25]发现.

然而,一些研究在实验过程中却得出相悖的结

论. Liu 等人[48]发现在命名实体识别等相对复杂的探针任务上将线性分类器替换为非线性分类器会产生明显的性能提升,并得出线性分类器能力不足的结论;Yaghoobzadeh 等人[72]在各项任务上的结果均显示 MLP 始终优于 LR 和 KNN,同样提出应使用比线性分类器更强大的模型. 与之相反,Edmiston 等人[87]则在探针任务中发现两者具有相似的性能.

近年来,部分工作将重点投入到诊断分类器复杂度对探针结果影响的探索中,并得出许多具有指导意义的结论.

Hewitt 等人[39]认为复杂的分类器本身就具有较强的对探针任务的拟合能力,因此其较高的探针性能并不能反映待探测位置编码的语言属性程度. 为了排除探针模型本身的拟合能力对结果的影响,文章提出控制任务(如图 3 所示),通过为每个输入随机分配语言属性标签构造控制数据集,并将分类器在探针数据集和控制数据集上的性能差记为选择性. 结果表明,复杂度较低的分类器往往具有更高的选择性,即在探针结果上具有更高的可靠度.

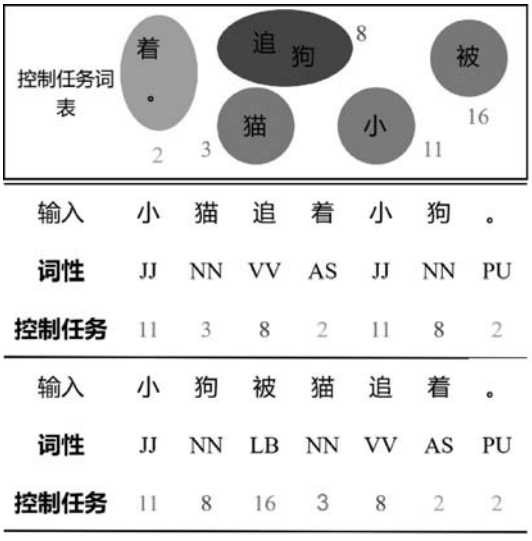

图 3 词性标注控制任务

另一部分研究则主张采用相对复杂的分类器. Saphra 等人[104]从线性代数的角度批判了简单分类器的滥用,因为大多数神经表征都不是以使信息线性可分离为明确目标来估计的. Pimentel 等人[17]则从信息论的角度重新审视探针任务,即计算表征和任务之间的互信息:

$$I(Z;R_\ell):=H(Z)-H(Z\mid R_\ell) \tag{4}$$

而由式(2)可知

$$I(Z;R_\ell)\geqslant H(Z)-H(p,q_\theta) \tag{5}$$

其中 $H(p,q_\theta)$ 是由分类器估计的交叉熵. 因此,人们应该始终选择性能最好的分类器,即使它更复杂,因为它将导致更严格的对互信息的估计,从而揭示更多表征中固有的信息. 作为对控制任务的反驳,文章提出控制函数,认为应将表征随机化而不是标签随机化作为探针任务的性能下界,并将控制后的交叉熵与探针任务交叉熵的差作为信息增益.

Zhu 等人[18]对控制任务和控制函数的本质进行研究,在 Pimentel 等人[17]信息论的框架下写出了 Hewitt 等人[39]对探针任务的分解方式(式 1),并进一步证明控制任务和控制函数本质上是等价的,具有高度相关性,都可以作为分类器复杂度的选择指标.

Voita 等人[105]提出用最小描述长度(Minimum Description Length, MDL)的信息理论探针作为标准探针的替代,即通过估计传输已知语言属性所需的最小代码长度来探测分类器的性能和复杂度. 其中变分代码作为 MDL 的一种实现,可以用于检查诱导分类器的复杂度. 实验发现为语言任务学习的分类器复杂度要远小于为控制任务学习的分类器复杂度.

Pimentel 等人[106]认为准确性和复杂度应该被视为双目标优化问题,简单地根据单个目标选择分类器都会导致边缘退化,因此文章主张一种反映分类器性能和复杂性之间的基本权衡的探针度量:帕累托超体积. 以图 4 为例,横坐标为打乱后的结构化输入映射到随机标签的能力,即独立于领域的复杂度,纵坐标是语言任务上的探针性能,图例表示对不同表征的探针结果. 该图能够清晰地显示不同探针任务在不同复杂度下的性能变化趋势和曲线下的面积,以指导选择具有高选择性和准确率的分类器.

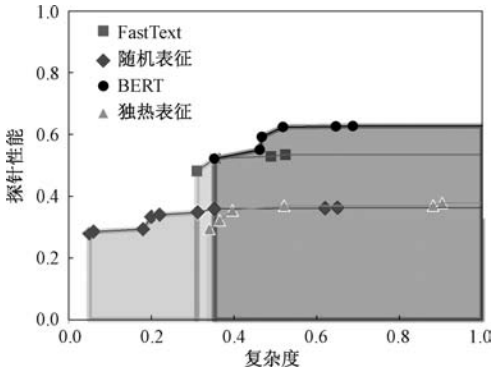

图 4 不同探针任务的帕累托曲线

延续自己的研究,Pimentel 等人[107]近期提出仅将重点放在模型复杂度选择上的反对意见. 他们

认为,受限于数据量的有限性,人们在实际中无法获取真实概率分布,因此,采用互信息作为探针任务的目标是不合理的.文章提出贝叶斯互信息框架以克服传统信息论框架下数据无法增熵、数据处理不增加信息等局限,根据给定关于任务本身有限的知识(即观察到的数据)的情况下,检查贝叶斯代理从表征中提取的信息量,并绘制帕累托曲线.该研究否认了根据探针模型复杂度检测数据的易提取性,而应根据任务的复杂度进行检测.

Cao 等人[108]提出一种基于减法修剪的探针,通过对分类头的监督梯度下降和对掩码的松弛来搜索子网,并在不同的粒度级别(包括修剪权重、神经元和层)进行掩码,实现低复杂度的分类器,相对于 MLP 在帕累托曲线中始终占据主导地位,即在给定任意探针复杂度预算下,都能取得较高精度.

Immer 等人[109]提出从贝叶斯的观点比较不同表征好坏的同时,选择合适复杂度的诊断分类器.文章将表征和诊断分类器的组合视为一个模型,并使用贝叶斯模型证据率确定该模型的归纳偏差程度.文章认为探针模型 M_p 和原始模型 M 的输出共同指定了表征-探针组成的先验对,只需要指定对应的似然函数,而两者组合后的先验联合分布由下式给出:

$$p(z \mid R_p) \times p(R_p \mid R_\ell, M_p)(R_\ell \mid x, M) \quad (6)$$

其中乘号左右两侧分别为似然函数和先验,通过积分可以计算表征-探针组的归纳偏差度:

$$\begin{aligned} & p(z \mid x, M, M_p) \\ &= \iint p(z, R_p, R_\ell \mid x, M, M_p) \mathrm{d}R_p \mathrm{d}R_\ell \\ &= \iint p(z, R_p \mid R_\ell, M_p) \mathrm{d}R_p \delta(M(x) - R_\ell) \mathrm{d}R_\ell \end{aligned} \quad (7)$$

其中 $M_\ell(x)$ 表示输入 x 在 M 中的输出,$\delta(\cdot)$ 表示狄拉克分布,这是由于表征空间一般是狭窄且离散的,即每一个表征函数仅对应一个离散的输出向量.通过最大化上式,可以用于比较不同表征优劣的同时,选取给定任务下合适的探针模型复杂度.

总结来说,关于诊断分类器复杂度是否重要,以及如何选择合适复杂度的诊断分类器,学术界仍存在一定的争议.但在以信息论和概率论为基础的研究框架下,该领域的发展日益完善,未来有望给出一套完整的诊断分类器设计方案.

3.1.3　数据集选择

第 2 章在介绍探针任务基本流程时提到,探针性能还会受到探针数据集 D_p 的影响.基于诊断分类器的探针方法往往需要构建与任务高度相关的探针数据集用于训练和测试,它的质量和大小共同决定了探针结果的可靠性.

遗憾的是,多数研究忽略了数据集的精心选择与设计,且由于需要探测的语言属性和语言现象的多样化,现有任务缺乏统一的标杆数据集.现阶段各研究获取探针数据集的方法大致可以分为两类.第一类研究直接使用已被广泛认可的 NLP 各领域数据集.例如标杆工作边缘探针[32]采用公开标注数据集 OntoNotes 5.0、Winograd 等用于词性、句子成分、句法依赖、命名实体、语义角色、关系分类、共指标记等一系列任务,作为词法、句法和语义任务的替代,用于全面检测模型各语言属性的能力.第二类研究则通过各类分析工具对语料标注得到数据集.例如标杆工作 Conneau 等人[23]随机抽取 Toronto Book Corpus 中的语料并使用 Stanford 解析器等工具进行解析标注,用于表面、句法和语义任务.表 4 总结了目前的主流的探针数据集及其适用任务.

从表中不难发现,当前探针数据集的使用过于多样化,且数据集之间缺少横向对比.例如,同为机器阅读理解问答的 SQuAD、SQuAD 2.0、HotpotQA 和 bAbI 数据集是否在不同程度上反映了模型编码相关语言属性的能力,一个在 SQuAD 上性能良好,但在 HotpotQA 上性能一般的模型是否能够反映出其某方面能力的不足?这些问题的存在增加了探针任务标准化的难度[35],业界仍缺少对特定任务的统一衡量标准.

另一方面,探针任务本质上是用探针模型在某一数据集上的表现作为编码特定属性能力的近似,因此数据集与任务间的偏置程度会影响探针结果的可靠性.Choudhury 等人[33]发现边缘探针模型很容易利用探针数据中的伪相关性,而这些伪相关性正是源于数据集与任务本身的偏置,势必会对结果造成偏差.Ravichander 等人[34]发现即使构造输入输出毫无关联的控制数据集,探针模型仍然具备一定的拟合能力,说明在数据集上的探针性能可能部分来源于噪声,而非任务本身.

3.2　其他方法

3.2.1　结构探针方法

结构探针由 Hewitt 等人[110]首次提出,用于检测深度模型是否在其单词表征中编码了完整的句法依赖树.文章认为,如果词向量间的距离表现出与句法依赖树相似的距离模式,就可以说明模型的单词表征学习到了句法树结构(如图 5).

表 4 主流探针数据集总结

数据集	采用论文	适用任务
Countries and Cities	文献[19]	对国家、城市等常识知识的掌握程度
Köhn	文献[20]	词性标注、依赖标注、单词性别、单词大小写、单词数量、单词时态
Adi, et al.	文献[22]	句子长度预测、单词内容预测、单词顺序预测
GLUE	文献[28,31,51,76,77]	语言可接受性、语义相似性、情感分析、语言推理、文本蕴含
OntoNotes 5.0	文献[28,30,32-33,53,75,94]	词性标注、句法依赖、命名实体识别、共指消解、语义角色标注
Conneau, et al.	文献[23,54,71]	句子长度预测、单词内容预测、语序预测、句法树深度预测、句法成分预测、时态预测、主宾语数量预测、语义奇偶预测、从句顺序预测
Winograd	文献[32]	共指消解
WIKI-PSE	文献[72]	词嵌入语义探测
DiscoEval	文献[74]	句嵌入是否包括关于句子在其话语语境中的作用的信息
SQuAD	文献[31,33,44]	机器阅读理解问答
HotpotQA	文献[33,44]	多跳机器阅读理解问答
bAbI	文献[44]	多跳机器阅读理解问答
SQuAD 2.0	文献[77]	机器阅读理解问答(包含无法回答的问题)
MWOZ	文献[45]	跨领域任务型对话
OOS	文献[45]	任务驱动的对话意图检测
Klafka, et al.	文献[86]	句中不同属性在上下文嵌入中的分布
LINSPECTOR	文献[91,97]	词性分析、句法依赖、语义角色标记、命名实体识别、自然语言推理
RNPC	文献[95]	递归名词短语理解
Ousidhoum, et al.	文献[56]	多语言潜在有害内容量化
Gurnee, et al.	文献[58]	对空间(世界、美国、纽约市)和时间(历史人物、艺术品、新闻)知识的掌握程度
Azaria, et al.	文献[59]	真实和虚假信息的辨识度

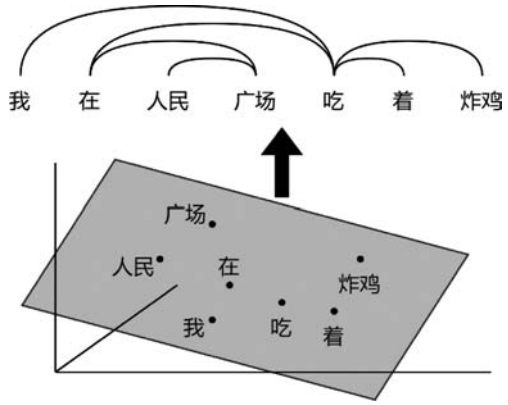

图 5 结构探针方法

具体来说,单词 w^i 和单词 w^j 在第 ℓ 层上的表征 R_ℓ^i 和 R_ℓ^j 的距离 d_B 可由一个半正定的测度矩阵 $A=B^{\mathrm{T}}B$ 表示:

$$\begin{aligned} d_B(R_\ell^i,R_\ell^j)^2&=(R_\ell^i-R_\ell^j)^{\mathrm{T}}A(R_\ell^i-R_\ell^j)\\ &=[B(R_\ell^i-R_\ell^j)]^{\mathrm{T}}B(R_\ell^i-R_\ell^j) \end{aligned} \tag{8}$$

再通过训练神经网络学习矩阵 B 的参数,使得词向量之间的距离 $d_B(R_i^\ell,R_j^\ell)^2$ 与解析树中单词间的距离 $d_\ell^{\mathrm{tree}}(w^i,w^j)$ 的差尽可能小:

$$\min_B \sum_l \frac{1}{|s^\ell|^2}\sum_{i,j}\left|d_\ell^{\mathrm{tree}}(w^i,w^j)-d_B(R_\ell^i,R_\ell^j)^2\right| \tag{9}$$

其中 $|s^\ell|$ 表示句子长度.

利用上述过程,可以学习到语法空间中单词对之间的距离.将距离函数 $d_B(h_i^\ell,h_j^\ell)$ 替换为向量的平方范数 $\|h_i\|_B^2$,可以用于探测单词在解析树中的深度.近年来,结构探针已被广泛应用于句法解析树检测等与树结构相关的探针任务中[31,103,111-114].

Limisiewicz 等人[115]对结构探针的矩阵 B 进行奇异值分解,将线性投影分解为同构旋转和线性缩放.文章为每个任务使用共享的同构旋转和不同的缩放向量,从而适用于多探针任务的联合训练,并具有良好的探针性能和选择性.在词法和句法上的多任务联合训练后的结果表明,BERT 等模型的词法和句法信息在表征中是趋于分离的.此外,Limisiewicz 等人[116]在后续工作中利用该探针继续在多语言嵌入中实验,实现了在小样本甚至零样本的场景下检测语言嵌入是否可以在多语言共享的空间中对齐.

White 等人[117]对线性映射的假设提出质疑,认为部分句法知识是通过非线性编码的,将结构探针框架重新定义为度量学习,并引入非线性的核变换.在 BERT 中的实验发现,正则化后的高斯核具有明显的探针性能优势,并认为这可能与自注意力机制引入的非线性相关.

Chen 等人[118]对词嵌入位于欧式空间的假设提

出质疑,提出庞加莱探针,通过训练两个权重矩阵将嵌入投影到具有明确定义的层次结构的双曲空间中,再采用类似于结构探针中的回归任务,使得嵌入之间的平方庞加莱距离和到原点的平方庞加莱距离分别近似于树中对应节点的距离和深度.在句法依赖树和词汇情感中的探针实验表明,庞加莱探针能够更忠实地恢复模型中的句法树结构,并揭示单词的上下文嵌入在情感子空间中的极性分布.

Kulmizev 等人[110]将结构探针扩展到有向依赖图,利用最大生成树算法[119]探测出的距离和深度信息导出有向依赖图. Müller-Eberstein 等人[120-121]进一步提出一种能够同时提取依赖树各边方向和对应关系标记的探针,他们认为具有多个依赖关系分类的任务可以简化为使用线性变换的标记任务,并结合有向结构探针,迭代地计算依赖树中的各有向边及其对应的关系类型,实现轻量级参数下的多任务覆盖.

Hou 等人[122]采用与先前研究不一样的思路探究语言中的结构化信息.他们从信息论的角度出发,估计语言图和上下文嵌入之间的互信息(如图 6 所示).基于互信息的不变性,文章采用 DeepWalk 算法[123]将离散的语言图映射为连续图嵌入.此外,文章还提出基于扰动分析的方法对局部语言结构进行分析.实验表明,基于信息论方法的语言图结构探针相比于诊断分类器具有更高的可靠性,并证实了 BERT 模型对句法和语义图结构的编码能力.

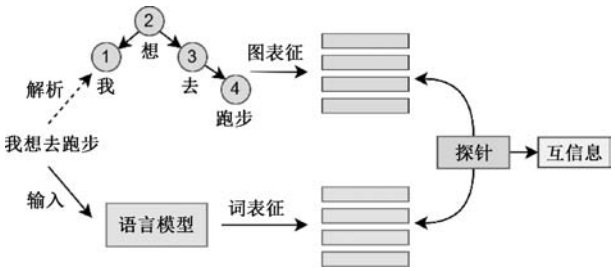

图 6　基于信息论方法的语言图结构探针

除了对词嵌入句法树结构的探测,Nikolaev 等人[124]提出语义结构探针将句嵌入映射到较低维度的子空间用于计算句子间任务特定的成对距离.这些距离可以根据任务属性进行解释,并且可以通过改变投影空间的维度衡量嵌入包含的信息量.文章发现不同模型族在处理语义相似度和文本推理等任务的内在行为存在很大差异.这一研究扩展了结构探针的应用场景.

3.2.2　基于干预的探针方法

为了测试和发现待检测区域的鲁棒性和因果效应,一些研究将干预手段应用于探针任务,以提升探针性能或深入探索模型的决策依据.

Giulianelli 等人[125]作为早期在探针任务中应用干预方法探索的研究,在主谓一致的探针任务中使用探针模型中的梯度修改原模型中的表征,使其预测结果稍微接近于真实标签.实验发现,虽然干预后的表征对一般语言建模的性能影响很小,但可以显著影响探针性能,这意味着探针模型可以正确识别原始模型实际使用的特征.

Tamkin 等人[126]将探针性能视为表征从一个任务到另一个任务的可迁移性.为了量化 BERT 等预训练模型各层的迁移性,他们对预训练模型的参数进行了简单的干预,即将第 k 层后的参数随机化,再进行下游任务的微调.通过比较干预前后的参数经过微调后在探针任务上的结果,文章发现 BERT 底层参数对微调任务表现出显著的迁移性.

Tucker 等人[127]将干预上升至因果分析的角度,利用探针模型梯度生成反事实嵌入.他们认为第 ℓ 层的探针模型获取输出所利用的 R_ℓ 中的特征应该与原始模型 M 利用的 R_ℓ 中的特征存在因果关联,如果改变 R_ℓ 产生新的探针预测,则模型输出应产生一致的变化(如图 7 所示).为此,作者根据梯度下降方向更新 R_ℓ 以获取改变探针输出后的反事实嵌入,并观察原始模型的输出是否同样产生了改变,以验证原始模型是否学会充分利用探针任务检测出的语言属性. Tucker 等人[128]延续基于梯度的方法,通过引入丢弃层避免模型中信息冗余导致生成的反事实嵌入只简单地利用部分信息.

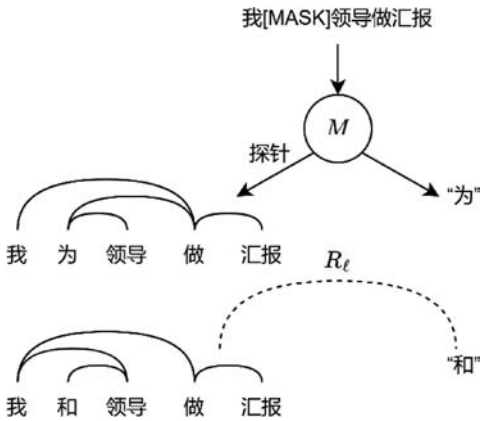

图 7　干预探针结果的探针方法

另一部分工作利用干预方法从嵌入中删除部分信息. Elazar 等人[38]基于遗忘探针算法[129]从探测的上下文嵌入中找到属性 Z 并将它删除,观察对结果的影响.例如对于输入文本"电影很棒",可以通过

干预删除词嵌入“棒”中的形容词词性，观察模型输出结果的变化，以验证模型决策过程中是否利用了“棒”的词性特征(如图 8 所示).具体来说，为了删除语言属性 Z，文章迭代地训练一个遗忘探针模型用于根据当前词向量 R_ℓ 预测 Z，再将 R_ℓ 投影至零空间，则相当于从中删除了与 Z 关联度最大的分支.通过多次迭代，可以从原表征中完全删除属性 Z，从而得到干预后的反事实嵌入用于模型分析.Lasri 等人[130]采用类似的方法着重研究 BERT 如何对语法数进行编码，以及如何利用这些编码解决相关语法任务.Ravfogel 等人[131]利用遗忘探针算法将单词的表征空间划分为属于关系从句和不属于关系从句两部分，生成两个具有相反语言属性的反事实嵌入，再将反事实嵌入的预测结果与原始嵌入的进行比较，得出有关语言特征对模型的因果效应.与之类似，Feder 等人[132]通过对抗性训练从 R_ℓ 中删除属性，并引入额外的辅助任务控制不应删除的语言属性，从而精确估计 Z 对下游任务的影响.

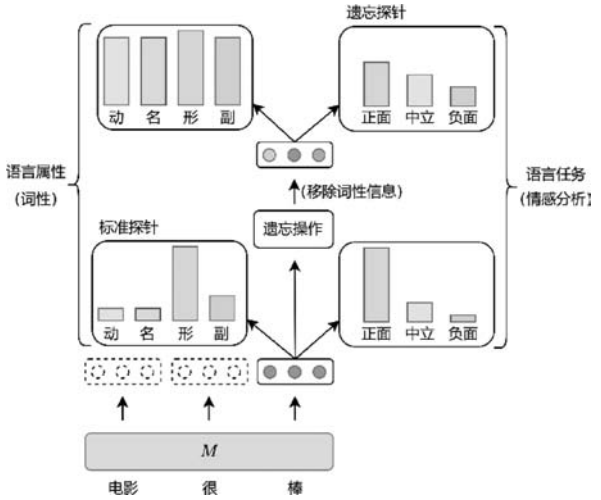

图 8 干预嵌入属性的探针方法

除了对文本嵌入的干预方法，一些研究也尝试进行输入级别的干预，即通过因果中介分析理论控制输入语句的最小变化，确保语句仅在单个分析属性上有所不同，再计算该属性变化所产生的因果效应[133-135].这种干预方法非常适用于模型偏见性检测.例如在高度受控的环境下将文本中的男性属性变为女性，并计算产生的总效应，可以有效评估模型的性别偏见程度.

3.2.3 基于提示的探针方法

提示学习①通过人工设计或机器训练一套伴随输入的模板用于指导模型生成期望的输出.例如，对于电影影评情感分析任务，可以人工构造模板“[X]总体来说，这是一部[Z]的电影”来指导具有丰富世界知识的预训练模型输出与任务相关的结果.其中[X]为待填入的输入文本，[Z]为预训练模型需要预测的内容.

受启发于提示学习的思路，许多研究通过人工设计大量提示模板，让预训练模型预测遮掩位置的内容，来检验模型是否掌握了世界知识[136-138]、常识推理[139-140]或特定领域的知识[141-142].其中，Petroni 等人[136]提出的 LAMA 探针任务为最具影响力的工作之一，他们提供了一组由主体－关系－客体三元组事实组成的知识源，每一组都会被转换为完形填空式的问答.例如给定一个三元组(刘翔，出生在，上海)，可以构造完形填空“刘翔出生在[Z]”以评估模型是否编码了相关知识.Talmor 等人[143]采用问答和填空的形式构造了 8 个推理任务挑战数据集，用于评估模型在宾语比较、连词分类等任务上表现的复现能力.

上述方法主要依赖人工设计的领域模板，从某种意义上，可以将这些模板看作隐式的探针模型，用于检测原始模型的语言属性.在这种观点下，用于探针任务的提示模板不仅可以是人工设计的，还可以是由机器训练得到的.Shin 等人[144]通过在提示模板中加入一定数量触发词，并迭代地根据预测的似然值选择触发词的替代，以此来选择最优模板，用于对模型编码的语言属性的下界做出更准确的评估.Zhong 等人[145]去除了在离散空间中优化的限制，他们认为提示模板不一定由真实的自然语言组成，而利用梯度下降方法直接在连续空间中优化提示模板，从而大幅提升了预训练模型在 LAMA 探针任务上的性能.Chen 等人[146]采用多提示的集成和提示搜索方法找到不同提示的最佳组合，用于探测预训练模型蕴含的明喻知识.Li 等人[37]采用前缀提示[147]方法，利用如图 9 所示的模板形式，将输入文本和待探测内容分别置于特殊符号“SEP”和“EOS”前，并训练前缀提示代替边缘探针任务，实验表明前缀提示在提取信息方面相比诊断分类器更优，并具有更好的选择性.Li 等人[148]对基于前缀提示的探针方法进行改进，通过引入多源注意力机制来整合来自多个来源的局部特征，并引入后期融合模块来捕捉全局特征，从而更好地探测开放领域对话模型的对话理解能力.

① 大多数提示学习算法的主要目的是提升模型少样本场景性能或减少模型待训练参数数量，并未显示地对探针任务进行研究；且这些引入的提示模板参数可能编码了额外的知识，不能直接用于原始模型语言属性的检测，因此相关论文将不在本文中介绍.

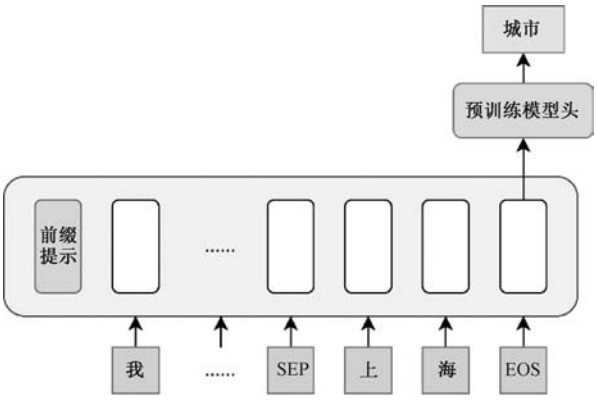

图 9　基于前缀提示的探针方法

随着 ChatGPT 和 GPT-4 等大语言模型的火热,基于提示的探针方法有望成为探测它们语言属性和语言现象的有效工具.一方面,由于这些模型的参数不对外公开,诊断分类器、结构探针等一系列需要中间层表征的探针方法无法进行有效探测;另一方面,GPT 系列模型均采用自回归的预训练模式,并在 ChatGPT 后引入指令学习[149],非常符合提示学习的流程.因此,现阶段相关探针方法以构造人工设计的提示模板为主.Bang 等人[150]作为早期对 ChatGPT 的探针工作,构造了 23 个模板数据集对 ChatGPT 等模型进行了广泛的技术评估,发现 ChatGPT 更擅长演绎推理而不是归纳推理,其逻辑能力依旧不足.Huang 等人[151]和 Zhong 等人[152]构造涵盖各种学科领域的不同难度问题首次对大模型的中文复杂问题推理的能力进行了全面评估.Deshpande 等人[153]聚焦于大语言模型的偏见问题,发现通过给 ChatGPT 分配一个特定角色(例如某个拳击手)可能会显著增加生成内容的有害性.Ortega-Martín 等人[154]通过询问 ChatGPT 句子是否存在歧义,来检测其对语言歧义的理解程度.Yin 等人[155]设计了 2858 个无法被回答的问题作为提示模板,检测大模型自我认知的能力,发现相比于人类,现有的大语言模型认知能力依旧不足,普遍存在"一本正经地胡说八道"的现象.总体来说,随着更多基于提示的探针方法被提出,人们正逐步构建对大语言模型语言输出能力的全方位基准评估,揭开 ChatGPT 等一系列大模型的神秘面纱[156].

然而,最近的一些研究提出了对基于提示的探针任务范式可靠性的担忧.Cao 等人[157]通过全面研究基于提示的探针行为,发现预训练模型通常会生成与提示本身相关但与其内部编码知识无关的不可靠预测,这意味着将提示应用于探针任务会导致预测性能的不准确.继续先前的研究,Cao 等人[158]利用结构因果模型[159]从因果角度证实预训练模型、探针数据和提示之间存在复杂的隐式联系,导致探针结果产生包括提示偏好、实例词汇和样本差异在内的系统性偏差(即从模型 M 到性能 $P(M,M_p,D,D_p)$ 存在的三条后门路径),并再次建议采用因果干预手段消除偏差.

除了基于提示的探针任务框架本身可能存在的缺陷,一些研究发现随着大模型训练数据来源的多样化,用于评估模型性能的探针数据集可能在预训练阶段被提前泄露,造成测试性能的虚高.Zhou 等人[160]通过模拟现实中极端数据泄露的情况,来测试大模型产生的影响,发现当训练集中包含了某一探测数据集后,会显著提升其测试性能,但在其他任务中的表现却有所下降.

总结来说,基于提示的探针方法的可靠性还有待进一步研究,如何将提示方法与干预手段相结合以及避免构建的探针数据集在训练阶段被提前泄露有望成为未来研究的热门方向.

3.2.4　无参数探针方法

许多工作致力于无参数的探针任务研究,根据模型在精心设计的探针数据集中的测试性能直接得出编码相关语言属性的程度.

第一类工作侧重于受控测试集的构建,通过对错误样例进行分析,以反向推演模型编码的属性.相比于基于提示的探针方法,受控测试集构造的范围更广,不止局限于完形填空式的提示模板.Marvin 等人[161]通过自动构建大量符合语法和不符合语法的句子对创建标杆数据集用于评估模型的语法属性;Salazar 等人[162]介绍了一组关于语言最小对的行文探针数据集,隔离了词法、语法、语义学中的特定现象.Zhou 等人[163]引入常识探针数据集,用于衡量原始模型是否能给符合常识的句子更高的平均对数概率.Steinborn 等人[164]通过构建男女对比的句子对研究多语言模型可能存在的性别偏见现象.Zhou 等人[165]为对话响应任务自动生成常识因果解释,再通过众包进行人工验证,用于探测对话响应生成模型是否捕捉到常识解释和响应间的逻辑关系.例如考虑图 10 中的对话,回复者难过的原因并没有明示,而常识告诉我们他是因为复习了很久还没有通过考试才难过的.文章对对话系统中常识性的解释进行了探针实验.

第二类工作侧重于直接对待探测区域进行分析,它们往往通过设计某种重要性得分作为推断语言属性的信号[36].Ethayarajh 等人[166]提出自相似

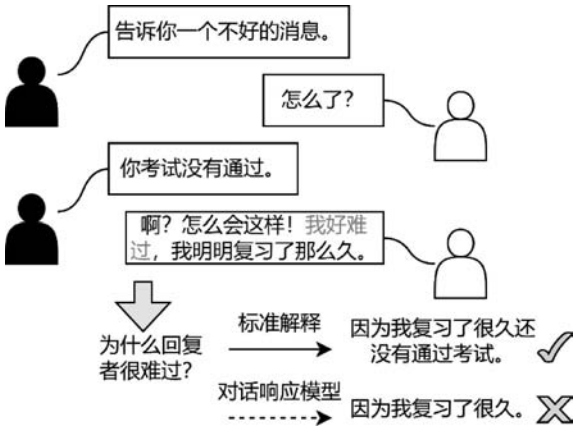

图 10 常识性解释探针方法

度、句内相似度和最大可解释性方差用于直接测量预训练模型上下文嵌入的语境化程度. Lepori 等人[167]使用表征相似性得分析(Representational Similarity Analysis, RSA)研究了动词嵌入对动词主语、代词嵌入对代词前置词以及全句表征对句首词的编码程度. Chrupała 等人[168]采用结构化 RSA 和回归式 RSA 直接量化模型编码的信息与语法树等符号结构所代表的信息的对应程度. Evanson 等人[169]采用可解释性判断[76]比较语法句子和对应的非语法句子的估计概率,用于检测 GPT－2 模型在不同训练阶段的语言能力,发现其学习顺序类似于人类儿童由浅入深的过程. 尽管无法获取 ChatGPT 等大模型训练过程中的断点模型,但这启发我们可以通过对与大模型结构、训练方式相近的小模型深入探测,来归纳出大模型可能存在的类似语言属性和现象.

第三类工作则侧重于无参数模型本身的方法设计. Htut 等人[170]通过将最高注意力权重对应的单词识别为父词或计算注意力矩阵上最大生成树提取 Transformer 识别的依赖关系,以检测注意力机制中是否编码了相关的句法知识. Wu 等人[171]采用扰动掩码技术,依次对单词 i 和 j 进行遮掩并得到前后向量差,作为单词 j 对单词 i 的影响程度,重复该过程可以得到单词间的影响程度矩阵,进而可以推导出句法依赖树等全局语言属性的编码程度. Zhou 等人[172]开发了一种基于层次聚类的启发式方法从几何角度对嵌入进行分析,确保簇与簇间不重叠的情况下,用尽可能少的簇数使簇内元素具有相同的标签. 根据聚类后簇的数量、簇间距离和空间相似性,可以容易地得出语言属性表征的线性程度、内部结构以及空间相似度,该方法也被 Mehta 等人[173]用于检测低资源语言的注释质量.

3.3 探针方法讨论

探针方法由于其直观和易于实现的特性,已被广泛应用于语言模型的各种属性和现象的检测. 由于侧重点不同,未来探针方法的研究将围绕以下两个不同的方向发展.

(1)以诊断分类器、结构探针和基于干预的探针为主的方法将继续围绕模型内部进行探索,分析模型不同结构编码的语言属性程度,从而使黑盒模型更加透明. 由于这些方法往往涉及额外的探针模型,因此如何控制探针模型复杂度,缓解额外参数对探针结果的影响也将成为研究重点.

(2)提示型探针和无参数探针则主要以对模型在不同情形下的最终输出的分析为主. 这类方法缺乏对模型内部细粒度属性的探测能力,但非常适合对 ChatGPT 等未公开参数的完全黑盒的大模型的能力探测. 随着大模型时代的到来,这类方法将具有更大的潜力和研究动机.

4 探针结果解释

基于现有的探针方法对目标位置的语言属性进行检测,往往可以得到测试集上的探针性能 $P(M, M_p, D, D_p)$,但该性能本身仍然是不可解释的. 例如,假设探针准确率达到 65%,是否意味着原始模型的检测区域编码了足够的相关语言属性? 另外,探针性能又与探针模型 M_p、原始数据集 D 和探针数据集 D_p 这些干扰项相关,它们又在多大程度上反应了人们真正关心的内容? 针对以上问题,本章将从对比方法和控制方法分别介绍现有研究是如何对探针结果做出合理的解释,并排除干扰因素.

4.1 对比方法

对比方法通过设计一些对比实验来逼近探针性能的上界或下界,再根据探针性能与近似上界或下界的性能差给出合理的解释. 常用的下界包括多数类、随机输入、静态嵌入和其他受控下界,常用的上界包括人工结果和其他受控上界. 各对比方法对应的相关研究详细列表见 4.3 节末的表 3.

多数类下界将全部样本预测为训练集中出现次数最多的类别[23,37,46,91,113,130,137,144-145]或者将具有多种可能类别的输入预测为最有可能的一类[27,38,42,130]. 例如,对于句子长度检测任务,可以遍历全体训练集找出句子长度的众数,作为所有测试集的预测结果[23];而对于词性标注任务,由于单词在不同语境下可能有不同的词性,因此可以根据训练集将每个

单词映射为最有可能的词性类别[27]. 多数类下界表明模型在完全不利用词嵌入或句嵌入等特征训练的情况下所能达到的最高准确率，经过训练后的探针性能与它的差值越大，说明探测位置编码了越多相关语言属性.

随机输入下界将输入打乱或者随机化以消除输入和输出间的关联性，这种情况下探针性能将完全无法从任务本身获取，只能通过随机性带来的“巧合”得到较低的性能. 例如，为了检测单词顺序对编码句子的重要程度，可以将单词顺序打乱进行探针实验[22]；对于依存句法分析等常见的探针任务，可以将嵌入随机化后交给探针模型训练[31]. 由于该方法的实用性和易实现性，大量文献采用随机输入作为探针下界之一.

在对 BERT 等上下文嵌入的语境化能力进行探针实验时，一些研究将 Word2Vec 等静态嵌入的探针结果作为近似下界. 例如，对于关系识别任务，需要得知两个实体在上下文语境中的信息，而静态嵌入则不具备任何语境化能力[48]. 因此，与静态嵌入的探针性能的差值可以作为模型语境化能力的解释.

除了以上三类常见的近似下界，一些研究尝试自行设计具有约束的基线作为探针结果的对比解释. Conneau 等人[23]将句子长度和单词的 TF-IDF 值分别作为唯一特征交由探针模型训练. Peters 等人[28]在代词共指解析任务中选择距离代词最接近的名词作为预测结果. Tenney 等人[32]利用字符级 CNN 层限制编码器利用上下文的长度，以检测其多大程度上捕获了长期依赖关系. Chen 等人[74]删除或修改预训练模型的损失函数以测试不同损失函数对结果的影响. Petroni 等人[136]采用关系提取模型[174]从基于提示的探针任务中计算结果. 对于句法解析任务，Clark 等人[75]在句法解析任务中总是选择距离单词固定位置的另一个单词作为头部，并汇报最佳结果，Hewitt 等人[110]将句中单词从左到右形成链作为预测的句法树，Manning 等人[111]采用全字符串匹配、首词匹配等基于规则的共指系统预测句法树. 最后，Hewitt 等人[175]提出一种全新的对比方法测量包含在表征中但不在随机基线中的信息. 相比于直接比较表征 R_ℓ 和随机基线 B 的探针性能差，他们提出比较随机基线与表征的拼接 $[B;R_\ell]$ 和随机基线与零向量的拼接 $[B;0]$ 的探针性能差，测量 R_ℓ 在预测 Z 方面超出 B 的贡献，并通过 γ —信息论[176]证明了该对比方法的优势.

对于近似上界，除了常用的人工评估手段，一些研究也尝试设计其他受控上界. Gupta 等人[19]在利用国家和城市分布向量预测其数字和分类 FreeBase 属性值的任务中，直接将 FreeBase 属性值作为探针模型的输入进行训练. Köhn 等人[20]对静态嵌入进行分类时，考虑到嵌入无法利用上下文信息，因此将在训练集中出现频率最高的类别作为单词的预测结果. Belinkov 等人[42]将序列标注探针任务直接用于编码器－解码器的下游训练，确保模型一定编码了相关语言属性. Michael 等人[53]将直接在真实标签上训练的诊断分类器结果作为提出的弱监督潜在子类学习方法的上界.

4.2　控制方法

对比方法大多基于主观直觉设计易于实现的上界或下界，并未深入考虑混杂因素对结果的影响. 为了克服这一问题，一些研究采用控制的思想设计基线，对探针结果做出更可靠的解释.

正如 3.1.2 节所提到的，Hewitt 等人[39]首先考虑到探针模型对探针性能的影响，通过为输入分配随机标签构造控制任务，将控制任务和探针任务的性能差记为选择性 SEL，作为探针结果的解释：

$$\mathrm{SEL}=P(M,M_p,D,D_p)-P(M,M_p,D,D_{\mathrm{rand}}) \tag{10}$$

其中 D_{rand} 为随机分配标签后的探针数据集.

类似的思路被 Ma 等人[177]进行改进，通过提出一种随机单词替换和随机标签匹配的控制任务应用于句法探针任务，从而显著提升句法探针任务的结果一致性.

进一步，Pimentel 等人[17]从信息论角度提出质疑，并设计控制函数用于将表征 R_ℓ 映射为随机向量，并计算映射前后属性 Z 和表征间互信息的差记为信息增益 G，作为探针结果的解释：

$$\begin{aligned} G&=I(Z;R_\ell)-I(Z;c(R_\ell))\\ &=H(Z\mid c(R_\ell))-H(Z\mid R_\ell)\geqslant 0 \end{aligned} \tag{11}$$

其中 $c(\cdot)$ 为控制函数.

为了近似计算增益函数，Pimentel 等人[17]首先以如下方式近似增益：

$$G\approx H_{q_{\theta 2}}(Z;c(R_\ell))-H_{q_{\theta 1}}(Z\mid R_\ell) \tag{12}$$

其中右式为利用模型 θ_1 和 θ_2 预估的增益函数 G_{q_θ}.

通过证明两个模型 KL 散度的边界关系，可以得出增益函数的变分边界，从而证明预估增益函数能够保证误差在可以接受的范围内：

$$G_{q_\theta}-\mathrm{KL}_{q_{\theta_2}}(Z;c(R_\ell))\leqslant G\leqslant G_{q_\theta}+\mathrm{KL}_{q_{\theta_1}}(Z;R_\ell) \tag{13}$$

考虑到完全消除探针模型干扰的困难性，Pimentel 等人[106]提出帕累托曲线作为控制方案，通过绘制不同探针模型复杂度下的探针性能给出解释，让人们更清晰地看出探针模型对探针结果的影响程度，以及探测区域究竟编码了多少相关语言属性.

最后，Ravichander 等人[34]设计控制数据集对探针结果进行控制(如图 11 所示)，他们修改原始数据集，使得所有输出都具有相同的属性值. 这样，从理论上讲，在控制数据集上训练的模型将不会包含任何与语言属性相关的内容，因为它对训练任务没有任何帮助. 如果在原始数据集上训练的模型和在控制数据集上训练的模型具有相似的探针性能，则无法说明探测位置编码了足够的语言属性.

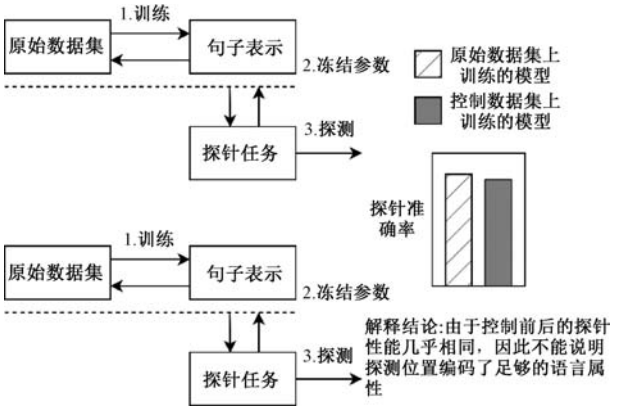

图 11 基于控制数据集的控制方法

4.3 探针结果解释总结

本节归纳总结了现有探针任务研究对结果的解释方法，详细结果如表 5 所示. 可以看出，多数研究仍采取对比方法，即设计一个简单基线作为结果的解释. 本文呼吁未来的研究能够更多地采用控制方法，以从控制变量的角度给出更严谨的解释.

表 5 探针结果解释方法

解释策略	方法	相关文献
对比方法	多数类	[19-20]，[23-24]，[27]，[33]，[37-38]，[42]，[46]，[76]，[91]，[95]，[113]，[130]，[136-137]，[142]，[144-145]，[167]
	随机输入	[22]，[29]，[31-33]，[37]，[58]，[72]，[75]，[87]，[94]，[97]，[99]，[104]，[110]，[122]，[151]，[163]，[165]，[167-168]，[170]
	静态嵌入	[28]，[30-32]，[42-43]，[48]，[53]，[71]，[76]，[86]，[97]，[99]，[110]，[113]，[118]，[143]，[153]，[166]，[168]
	受控下界	[23]，[28]，[32]，[42]，[46]，[72]，[74-75]，[59]，[99]，[110-111]，[113-114]，[118]，[126]，[136]，[153]，[171]，[174-175]
	人工结果	[23]，[74]，[76]，[94-95]，[138]，[140]，[155]，[161]，[163]，[165]
	受控上界	[19-20]，[42]，[53]，[122]
控制方法	控制任务	[18]，[38-39]，[98]，[105]，[143]，[177]
	控制函数	[17-18]
	控制数据集	[34]
	帕累托曲线	[106-108]

5 探针任务应用

随着大规模预训练模型的涌现，探针任务在自然语言处理领域中正扮演着愈发关键的角色. 本节概括了探针任务的关键应用，包括对模型编码语言能力更深入的理解、对模型工作机理的理解和对偏见歧视等不公平现象的检测.

(1)理解模型局部或整体的语言能力

作为一种高效且易于实现的全局可解释性方法，探针任务为理解模型特定区域编码的语言属性和语言现象的程度提供了标准化流程，这也是探针任务最主要的应用场景.

此外，基于提示的探针方法和无参数探针方法也成为诊断大语言模型整体语言能力的主要工具. 通过设计一系列与常识推理[150]、世界知识[151]、自我认知[155]等任务相关的提示模板，让模型像人类一样进行答题或问卷调查，从而准确评估模型在特定领域的能力.

(2)理解模型工作机理

探针任务对理解模型系统性的工作流程具有重要意义. 通过对模型特定组件进行分析，可以理解各组件负责的功能，以及信息如何在模型组件间传递. 例如，对注意力层进行探针可以发现特定注意力头部有助于主语－动词一致性，间接说明了多头注意力机制的重要性[87]；对 BERT 如何编码语法数字任务的基于干预的探针实验中可以发现 BERT 依赖于语法数字的线性编码来产生正确的行为输出，且对名词和动词使用了单独的语法数字编码[130]；而采用基于提示的探针任务对预训练模型进行知识检

测可以发现模型通常会生成与提示本身相关但与其内部编码知识无关的不可靠预测[157].这些系统性机理的发现不仅可以帮助人们更好地理解大模型能够有效应用于下游任务的原理,还可以帮助人们发现潜藏在模型中的缺陷和捷径,有助于指导下一阶段的训练和优化.

(3)检测模型偏见或歧视程度

探针任务在帮助检测模型偏见或歧视方面同样发挥着关键作用.通过设计针对性的探针任务,可以有效检测模型在处理不同种族、性别、宗教或社会群体时是否存在不公平的差异或偏向.对于黑盒模型,可以采用基于提示的探针[153,178-180]检测模型的输出文本是否包含对特定群体的偏见.对于白盒模型,可以采用搜集的有害数据集训练诊断分类器进行偏见性评估[56],也可以采用基于干预的探针方法,通过测量针对性的干预措施产生的直接或间接效应,来精确定位各种偏见歧视是如何在模型不同组件之间传播的[133-135].随着国家对生成式大模型安全规范的进一步重视,探针任务有望成为监管大模型公平性和安全性的重要工具.

6　未来研究方向

尽管近年来探针任务技术已取得广泛的应用和长足的发展,但回到 2.3 节介绍的设计一个良好的探针任务需要考虑的四个基本步骤,可以发现仍有许多亟待解决的问题与挑战.本节针对现有探针任务的不足之处进行讨论,并对未来的关键研究方向做出展望.

(1)设计更广泛的语言属性探测

现有的研究大多将精力集中于词嵌入、句嵌入和中间层输出,缺少对模型其他部件的分析,例如不同激活层或丢弃层对编码语言属性的影响,提示学习中的提示位置的语言属性编码程度等.对于感兴趣的语言属性,现有研究大多集中于词法、句法和语义的不同粒度语言属性的探索,缺少利用探针实验对其他可能存在的语言现象的分析,例如模型内部可能存在的系统性偏见、捷径学习、后门攻击等内容.因此,未来研究可以朝着更全面的模型部件与语言现象分析发展.

另外,目前的探针任务大多针对英文数据,而缺少对中文数据全面和独立的分析,仅有部分研究在讨论翻译任务或多语言预训练模型时涉及对中文的分析[42,56,98].然而,中文包含的语言属性往往与英文大相径庭.因此,一份全面的针对中文(或其他非英文语言)预训练模型的探针实验报告显得尤为重要

(2)探究探针数据集与探针任务的关联.

探针实验采用探针数据集作为任务的近似替代,不可避免地会存在一些偏置与误差[34].一方面,过少的数据集难以反映探针任务的真实概率分布;另一方面,过多的数据集不仅会增大搜集难度和训练开销,还会导致探针模型发挥“记忆”的特性在探针任务中过拟合.

目前,已有许多研究发现这一问题并呼吁针对探针数据集的选择展开更深入的研究[29,33-34],但是相关研究仍然较少.Zhu 等人[181]提出从泛化边界的角度决定数据集的大小,并通过功效分析评估该数据建议下的可靠性.然而,尽管找到了可靠的下界,对于是否存在上界、数据集与探针任务和探针性能的关联仍没有得出清晰的结论.本文认为,对探针数据集与探针任务关系的进一步探索,是补全探针领域理论基础的重要研究方向.

(3)深入研究探针任务的本质

尽管主流观点仍然认为探针任务是用于探测模型中间层编码特定语言属性的程度,但近年来一些论文对这一结论提出质疑,并提出全新的观点解释探针任务的本质,例如计算表征与任务之间的互信息量的操作[17],量化表征从一个任务到另一个任务的可转移性[126],量化表征的归纳偏置程度[182]等.另一方面,Kunz 等人[183]指出目前人们对表征以及它们与探针模型的学习能力的相互作用缺乏深入了解,因而无法从结果中推断出某些特定特征的存在.因此,未来任务可以集中从理论角度分析探针任务究竟学习了什么,以及人们能够从探针结果中得出什么可靠性结论.

(4)构建更有效的探针模型

探针模型作为探针任务中研究最多的部分,已经衍生出包括诊断分类器、结构探针、基于干预的探针、基于提示的探针和无参数探针等多种模型分支.然而,这些方法均存在一定的应用局限和理论不足.未来,设计更有效且可靠的探针模型将仍然是探针社区的主要研究方向.

(5)提供更合理的对比和控制实验

现有的对比方法通过设计主观可以理解的近似界限来解释检测区域编码的语言属性的程度,然而越来越多的研究提出反对意见,例如未排除干扰因素[34,39,107],无法检测表征相比基线更有用的内容[175]等.尽管后续基于控制的方法从理论上排除

了探针模型或探针数据集对结果的干扰，但如何对各种干扰因素同时进行控制，并根据对比和控制结果给出更可靠的结论（而不止是说相比于基线编码了更多相关语言属性），同样是未来的重点研究方向之一.

7 结 论

大规模预训练模型在 NLP 各项领域取得性能突破的同时，也引起了人们对模型内部可解释性与安全性的重视. 本文聚焦于 NLP 领域中常见的可解释性方法——探针任务，并从任务介绍、探针方法和探针结果解释等角度对现有的优秀研究进行了归纳与总结. 作为一个相对新颖的研究领域，现有的探针方法仍然不够成熟，包括探针任务本身设计的理论缺陷以及对更复杂语言属性的探索匮乏等. 本文希望为读者提供一份关于现有探针任务研究的全面“诊断书”，并鼓励更多的研究投入到相关领域中.

参 考 文 献

[1] Otter D W, Medina J R, Kalita J K. A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(2): 604-624

[2] Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712, 2023

[3] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 4171-4186

[4] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners//Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, 2020: 1877-1901

[5] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable ai: A review of machine learning interpretability methods. Entropy, 2020, 23(1): 18

[6] Rogers A, Kovaleva O, Rumshisky A. A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 2020, 8(2): 842-866

[7] Danilevsky M, Qian K, Aharonov R, et al. A survey of the state of explainable AI for natural language processing//Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. Suzhou, China, 2020: 447-459

[8] Samek W, Montavon G, Lapuschkin S, et al. Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 2021, 109(3): 247-278

[9] Li X, Xiong H, Li X, et al. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. Knowledge and Information Systems, 2022, 64(12): 3197-3234

[10] Hua Y Y, Zhang D C, et al. Research progress in the interpretability of deep learning models. Journal of Cyber Security, 2020, 5(3): 1-12 (in Chinese)
(化盈盈，张岱墀，葛仕明. 深度学习模型可解释性的研究进展. 信息安全学报，2020，5(3)：1-12.)

[11] Ribeiro M T, Singh S, Guestrin C. “Why should I trust you?” Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1135-1144

[12] Lundberg S M, Lee S I. A unified approach to interpreting model predictions//Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017, 30: 4765-4774

[13] Park D H, Hendricks L A, Akata Z, et al. Multimodal explanations: Justifying decisions and pointing to the evidence//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8779-8788

[14] Camburu O M, Rocktäschel T, Lukasiewicz T, et al. E-snli: Natural language inference with natural language explanations//Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2018, 31: 9560-9572

[15] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology. 2017, 31: 841

[16] Mothilal R K, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain. 2020: 607-617

[17] Pimentel T, Valvoda J, Maudslay R H, et al. Information-theoretic probing for linguistic structure//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 4609-4622

[18] Zhu Z, Rudzicz F. An information theoretic view on selecting linguistic probes//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 9251-9262

[19] Gupta A, Boleda G, Baroni M, et al. Distributional vectors

encode referential attributes//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Lisbon, Portugal, 2015: 12-21

[20] Köhn A. What's in an embedding? Analyzing word embeddings through multilingual evaluation//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Lisbon, Portugal, 2015: 2067-2073

[21] Ettinger A, Elgohary A, Resnik P. Probing for semantic evidence of composition by means of simple classification tasks//Proceedings of the 1st Workshop on Evaluating Vector-space Representations for NLP. Berlin, Germany, 2016: 134-139

[22] Adi Y, Kermany E, Belinkov Y, et al. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks//Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017: 1-13

[23] Conneau A, Kruszewski G, Lample G, et al. What you can cram into a single vector: Probing sentence embeddings for linguistic properties//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018: 2126-2136

[24] Shi X, Padhi I, Knight K. Does string-based neural MT learn source syntax? //Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP). Austin, USA, 2016: 1526-1534

[25] Belinkov Y, Durrani N, Dalvi F, et al. What do neural machine translation models learn about morphology? //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017: 861-872

[26] Hupkes D, Veldhoen S, Zuidema W. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. Journal of Artificial Intelligence Research, 2018, 61: 907-926

[27] Blevins T, Levy O, Zettlemoyer L. Deep RNNs encode soft hierarchical syntax//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia, 2018: 14-19

[28] Peters M E, Neumann M, Zettlemoyer L, et al. Dissecting contextual word embeddings: Architecture and representation//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium, 2018: 1499-1509

[29] Zhang K W, Bowman S R. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis//Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium, 2018: 359-361

[30] Tenney I, Das D, Pavlick E. BERT rediscovers the classical NLP pipeline//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 4593-4601

[31] Merchant A, Rahimtoroghi E, Pavlick E, et al. What happens to bert embeddings during fine-tuning? //Proceedings of the 3rd Blackbox NLP Workshop on Analyzing and Interpreting Neural Networks for NLP. 2020: 33-44

[32] Tenney I, Xia P, Chen B, et al. What do you learn from context? probing for sentence structure in contextualized word representations//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA, 2019: 1-17

[33] Choudhury S R, Bhutani N, Augenstein I. Can edge probing tasks reveal linguistic knowledge in QA models? //Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea, 2022: 1620-1635

[34] Ravichander A, Belinkov Y, Hovy E. Probing the probing paradigm: Does probing accuracy entail task relevance? //Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 3363-3377

[35] Kuznetsov I, Gurevych I. A matter of framing: The impact of linguistic formalism on probing results//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 171-182

[36] Belinkov Y. Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 2022, 48(1): 207-219

[37] Li J, Cotterell R, Sachan M. Probing via prompting//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA. 2022: 1144-1157

[38] Elazar Y, Ravfogel S, Jacovi A, et al. Amnesic probing: Behavioral explanation with amnesic counterfactuals. Transactions of the Association for Computational Linguistics, 2021, 9: 160-175

[39] Hewitt J, Liang P. Designing and interpreting probes with control tasks//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019: 2733-2743

[40] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, USA, 2013: 746-751

[41] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[42] Belinkov Y, Màrquez L, Sajjad H, et al. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks//Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Taipei, China, 2017: 1-10

[43] Xu H, van Genabith J, Liu Q, et al. Probing word translations in the transformer and trading decoder for encoder lay-

ers//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 74-85

[44] Van Aken B, Winter B, Löser A, et al. How does bert answer questions? A layer-wise analysis of transformer representations//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China, 2019: 1823-1832

[45] Wu C S, Xiong C. Probing task-oriented dialogue representation from language models//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 5036-5051

[46] Alt C, Gabryszak A, Hennig L. Probing linguistic features of sentence-level representations in neural relation extraction//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 1534-1545

[47] Kiros R, Zhu Y, Salakhutdinov R R, et al. Skip-thought vectors//Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2015, 28: 3294-3302

[48] Liu N F, Gardner M, Belinkov Y, et al. Linguistic knowledge and transferability of contextual representations//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 1073-1094

[49] Peters M , Neumann M , Iyyer M , et al. Deep contextualized word representations//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, USA, 2018: 2227-2237

[50] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. OpenAI, USA. Technical report, 0112018

[51] Dalvi F, Sajjad H, Durrani N, et al. Analyzing redundancy in pretrained transformer models//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4908-4926

[52] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding//Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019, 32: 5754-5764

[53] Michael J, Botha J A, Tenney I. Asking without telling: Exploring latent ontologies in contextual representations//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6792-6812

[54] Mohebbi H, Modarressi A, Pilehvar M T. Exploring the role of BERT token representations to explain sentence probing results//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic, 2021: 792-806

[55] Yuan H, Chen Y, Hu X, et al. Interpreting deep models for text analysis via optimization and regularization methods//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33(01): 5717-5724

[56] Ousidhoum N, Zhao X, Fang T, et al. Probing toxic content in large pre-trained language models//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 4262-4274

[57] Aghazadeh E, Fayyaz M, Yaghoobzadeh Y. Metaphors in pre-trained language models: Probing and generalization across datasets and languages//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 2037-2050

[58] Gurnee W, Tegmark M. Language models represent space and time. arXiv preprint arXiv:2310.02207, 2023

[59] Azaria A, Mitchell T. The internal state of an llm knows when its lying//Proceedings of the Association for Computational Linguistics: EMNLP 2023. Singapore, 2023: 967-976

[60] Levy O, Goldberg Y. Dependency-based word embeddings//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Baltimore, USA. 2014: 302-308

[61] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, 2014: 1532-1543

[62] Stratos K, Collins M, Hsu D. Model-based word embeddings from decompositions of count matrices//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long papers). Beijing, China, 2015: 1282-1291

[63] Brown P F, Della Pietra V J, Desouza P V, et al. Class-based n-gram models of natural language. Computational Linguistics, 1992, 18(4): 467-480

[64] Wieting J, Bansal M, Gimpel K, et al. Towards universal paraphrastic sentence embeddings//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016: 1-19

[65] Qian P, Qiu X, Huang X J. Investigating language universal and specific properties in word embeddings//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). San Juan, Puerto Rico, 2016: 1478-1488

[66] Kim Y, Jernite Y, Sontag D, et al. Character-aware neural language models//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, Arizona, 2016: 2741-2749

[67] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling//Proceedings of the NIPS 2014 Workshop on Deep Learning. Montreal, Canada, 2014: 1-9

[68] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017, 30: 5998-6008

[69] Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated convolutional networks//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 933-941

[70] McCann B, Bradbury J, Xiong C, et al. Learned in translation: Contextualized word vectors//Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017, 30: 6294-6305

[71] Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language? //Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 3651-3657

[72] Yaghoobzadeh Y, Kann K, Hazen T J, et al. Probing for semantic classes: Diagnosing the meaning content of word embeddings//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 5740-5753

[73] Ling W, Dyer C, Black A W, et al. Two/too simple adaptations of word2vec for syntax problems//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, USA, 2015: 1299-1304

[74] Chen M, Chu Z, Gimpel K. Evaluation benchmarks and learning criteria for discourse-aware sentence representations//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019: 649-662

[75] Clark K, Khandelwal U, Levy O, et al. What does bert look at? an analysis of BERT's attention//Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Florence,Italy, 2019: 276-286

[76] Warstadt A, Singh A, Bowman S R. Neural network acceptability judgments. Transactions of the Association for Computational Linguistics, 2019, 7: 625-641

[77] Chiang C H, Huang S F, Lee H. Pretrained language model embryology: The birth of ALBERT//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6813-6828

[78] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-17

[79] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter//Proceedings of the NIPS 2019 Workshop on Energy Efficient Machine Learning and Cognitive Computing. New York, USA, 2019: 1-5

[80] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach//Proceedings of the 20th Chinese National Conference on Computational Linguistics. Huhhot, China, 2021: 1218-1227

[81] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9

[82] Clark K, Luong M T, Le Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-18

[83] Henderson M, Casanueva I, Mrkšïc N, et al. ConveRT: Efficient and accurate conversational representations from transformers//Proceedings of the Association for Computational Linguistics: EMNLP 2020. 2020: 2161-2174

[84] Zhang Y, Sun S, Galley M, et al. Dialogpt: Large-scale generative pre-training for conversational response generation//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020: 270-278

[85] Wu C S, Hoi S C H, Socher R, et al. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 917-929

[86] Klafka J, Ettinger A. Spying on Your Neighbors: Fine-grained probing of contextual embeddings for information about surrounding words//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 4801-4811

[87] Edmiston D. A systematic analysis of morphological content in BERT models for multiple languages. arXiv preprint arXiv:2004.03032, 2020

[88] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland, 2014: 2335-2344

[89] Nguyen T H, Grishman R. Relation extraction: Perspective from convolutional neural networks//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Denver, USA, 2015: 39-48

[90] Zhang Y, Qi P, Manning C D. Graph convolution over pruned dependency trees improves relation extraction//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium, 2018: 2205-2215

[91] Ahin G G, Vania C, Kuznetsov I, et al. LINSPECTOR: Multilingual probing tasks for word representations. Computational Linguistics, 2020, 46(2): 335-385

[92] Joulin A, Grave É, Bojanowski P, et al. Bag of tricks for efficient text classification//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Valencia,Spain, 2017: 427-431

[93] Lample G, Conneau A, Ranzato M A, et al. Word translation without parallel data//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-14

[94] Zhang Y, Warstadt A, Li X, et al. When do you need billions of words of pretraining data? //Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 1112-1125

[95] Lyu Q, Hua Z, Li D, et al. Is "my favorite new movie" my favorite movie? Probing the understanding of recursive noun phrases//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA, 2022: 5286-5302

[96] Lewis M, Liu Y, Goyal N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7871-7880

[97] Liu Z, Wang Y, Kasai J, et al. Probing across time: What does RoBERTa know and when? //Proceedings of the Association for Computational Linguistics: EMNLP. Punta Cana, Dominican Republic, 2021: 820-842

[98] Shapiro N, Paullada A, Steinert-Threlkeld S. A multilabel approach to morphosyntactic probing//Proceedings of the Association for Computational Linguistics: EMNLP. Punta Cana, Dominican Republic, 2021: 4486-4524

[99] Conia S, Navigli R. Probing for predicate argument structures in pretrained language models//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 4622-4632

[100] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023

[101] Biderman S, Schoelkopf H, Anthony Q G, et al. Pythia: A suite for analyzing large language models across training and scaling//International Conference on Machine Learning. Honolulu, USA, 2023: 2397-2430

[102] Zhang S, Roller S, Goyal N, et al. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068, 2022

[103] Maudslay R H, Valvoda J, Pimentel T, et al. A Tale of a probe and a parser//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7389-7395

[104] Saphra N, Lopez A. Understanding learning dynamics of language models with SVCCA//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 3257-3267

[105] Voita E, Titov I. Information-theoretic probing with minimum description length//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 183-196

[106] Pimentel T, Saphra N, Williams A, et al. Pareto probing: Trading off accuracy for complexity//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 3138-3153

[107] Pimentel T, Cotterell R. A bayesian framework for information-theoretic probing//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic. 2021: 2869-2887

[108] Cao S, Sanh V, Rush A M. Low-complexity probing via finding subnetworks//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 960-966

[109] Immer A, Hennigen L T, Fortuin V, et al. Probing as quantifying inductive bias//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 1839-1851

[110] Hewitt J, Manning C D. A structural probe for finding syntax in word representations//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 4129-4138

[111] Manning C D, Clark K, Hewitt J, et al. Emergent linguistic structure in artificial neural networks trained by self-supervision//Proceedings of the National Academy of Sciences. USA, 2020, 117(48): 30046-30054

[112] Kulmizev A, Ravishankar V, Abdou M, et al. Do neural language models show preferences for syntactic formalisms? //Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 4077-4091

[113] Maudslay R H, Cotterell R. Do syntactic probes probe syntax? Experiments with jabberwocky probing//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 124-131

[114] Chi E A, Hewitt J, Manning C D. Finding universal grammatical relations in multilingual BERT//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5564-5577

[115] Limisiewicz T, Mare? ek D. Introducing orthogonal constraint in structural probes//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 428-442

[116] Limisiewicz T, Mare? ek D. Examining cross-lingual contextual embeddings with orthogonal structural probes//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic, 2021: 4589-4598

[117] White J C, Pimentel T, Saphra N, et al. A non-linear structural probe//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 132-138

[118] Chen B, Fu Y, Xu G, et al. Probing BERT in hyperbolic spaces//Proceedings of the 9th International Conference on Learning Representations. Austria, 2021: 1-24

[119] Chu Y J. On the shortest arborescence of a directed graph. Scientia Sinica, 1965, 14: 1396-1400

[120] Müller-Eberstein M, Van Der Goot R, Plank B. Probing for labeled dependency trees//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 7711-7726

[121] Müller-Eberstein M, Goot R, Plank B. Sort by structure: Language model ranking as dependency probing//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA, 2022: 1296-1307

[122] Hou Y, Sachan M. Bird's Eye: Probing for linguistic graph structures with a simple information-theoretic approach//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 1844-1859

[123] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2014: 701-710

[124] Nikolaev D, Padó S. Investigating semantic subspaces of transformer sentence embeddings through linear structural probing//Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP. Singapore, 2023: 142-154

[125] Giulianelli M, Harding J, Mohnert F, et al. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information//Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium, 2018: 240-248

[126] Tamkin A, Singh T, Giovanardi D, et al. Investigating transferability in pretrained language models//Proceedings of the Association for Computational Linguistics: EMNLP 2020. 2020: 1393-1401

[127] Tucker M, Qian P, Levy R. What if this modified that? Syntactic interventions with counterfactual embeddings//Proceedings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 862-875

[128] Tucker M, Eisape T, Qian P, et al. When does syntax mediate neural language model performance? Evidence from dropout probes//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA, 2022: 5393-5408

[129] Ravfogel S, Elazar Y, Gonen H, et al. Null it out: Guarding protected attributes by iterative nullspace projection//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7237-7256

[130] Lasri K, Pimentel T, Lenci A, et al. Probing for the usage of grammatical number//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 8818-8831

[131] Ravfogel S, Prasad G, Linzen T, et al. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction//Proceedings of the 25th Conference on Computational Natural Language Learning. 2021: 194-209

[132] Feder A, Oved N, Shalit U, et al. Causalm: Causal model explanation through counterfactual language models. Computational Linguistics, 2021, 47(2): 333-386

[133] Vig J, Gehrmann S, Belinkov Y, et al. Investigating gender bias in language models using causal mediation analysis//Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems. 2020, 33: 12388-12401

[134] Finlayson M, Mueller A, Gehrmann S, et al. Causal analysis of syntactic agreement mechanisms in neural language models//Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021. Association for Computational Linguistics (ACL), 2021: 1828-1843

[135] Amini A, Pimentel T, Meister C, et al. Naturalistic causal probing for morpho-syntax. Transactions of the Association for Computational Linguistics, 2023, 11(3): 384-403

[136] Petroni F, Rocktäschel T, Riedel S, et al. Language models as knowledge bases? //Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019: 2463-2473

[137] Jiang Z, Xu F F, Araki J, et al. How can we know what language models know? Transactions of the Association for Computational Linguistics, 2020, 8: 423-438

[138] Beloucif M, Biemann C. Probing pre-trained language models for semantic attributes and their values//Proceedings of the Association for Computational Linguistics: EMNLP. Punta Cana, Dominican Republic, 2021: 2554-2559

[139] Kassner N, Schütze H. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7811-7818

[140] Lin B Y, Lee S, Khanna R, et al. Birds have four legs NumerSense: Probing numerical commonsense knowledge of pre-trained language models//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6862-6868

[141] Meng Z, Liu F, Shareghi E, et al. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 4798-4810

[142] Sung M, Lee J, Yi S, et al. Can language models be biomedical knowledge bases? //Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic, 2021: 4723-4734

[143] Talmor A, Elazar Y, Goldberg Y, et al. oLMpics-on what language model pre-training captures. Transactions of the Association for Computational Linguistics, 2020, 8: 743-758

[144] Shin T, Razeghi Y, Logan IV R L, et al. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4222-4235

[145] Zhong Z, Friedman D, Chen D. Factual probing is [MASK]: Learning vs. learning to recall//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 5017-5033

[146] Chen W, Chang Y, Zhang R, et al. Probing simile knowledge from pre-trained language models//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 5875-5887

[147] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 4582-4597

[148] Li Y, Zhou H, Zhou J, et al. Multi-source probing for open-domain conversational understanding//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore, 2023: 12491-12505

[149] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback//Proceedings of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems. 2022, 35: 27730-27744

[150] Bang Y, Cahyawijaya S, Lee N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv e-prints, 2023: arXiv: 2302.04023

[151] Huang Y, Bai Y, Zhu Z, et al. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models//Proceedings of the Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems. 2023

[152] Zhong W, Cui R, Guo Y, et al. Agieval: A human-centric benchmark for evaluating foundation models. arXiv preprint arXiv:2304.06364, 2023

[153] Deshpande A, Murahari V, Rajpurohit T, et al. Toxicity in ChatGPT: Analyzing persona-assigned language models//Proceedings of the Association for Computational Linguistics: EMNLP. Singapore, 2023: 1236-1270

[154] Ortega-Martín M, García-Sierra Ó, Ardoiz A, et al. Linguistic ambiguity analysis in ChatGPT. arXiv e-prints, 2023: arXiv: 2302.06426

[155] Yin Z, Sun Q, Guo Q, et al. Do large language models know what they don't know? //Proceedings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada, 2022: 8653-8665

[156] Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. arXiv preprint arXiv: 2307.03109, 2023

[157] Cao B, Lin H, Han X, et al. Knowledgeable or educated guess? Revisiting language models as knowledge bases//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 1860-1874

[158] Cao B, Lin H, Han X, et al. Can prompt probe pretrained language models? Understanding the invisible risks from a causal view//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 5796-5808

[159] Pearl J. Models, reasoning and inference. Cambridge, UK: Cambridge University Press, 2000, 19(2):24-32

[160] Zhou K, Zhu Y, Chen Z, et al. Don't make your LLM an evaluation benchmark cheater. arXiv preprint arXiv: 2311.01964, 2023

[161] Marvin R, Linzen T. Targeted syntactic evaluation of language models//Proceedings of the Society for Computation in Linguistics (SCiL), 2019: 373-374

[162] Salazar J, Liang D, Nguyen T Q, et al. Masked language model scoring//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2699-2712

[163] Zhou X, Zhang Y, Cui L, et al. Evaluating commonsense in pre-trained language models//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(05): 9733-9740

[164] Steinborn V, Dufter P, Jabbar H, et al. An information-theoretic approach and dataset for probing gender stereotypes in multilingual masked language models//Proceedings of the Association for Computational Linguistics: NAACL. Seattle, USA, 2022: 921-932

[165] Zhou P, Jandaghi P, Cho H, et al. Probing commonsense explanation in dialogue response generation//Proceedings of the Association for Computational Linguistics: EMNL. Punta Cana, Dominican Republic. 2021: 4132-4146

[166] Ethayarajh K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT－2 embeddings//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019: 55-65

[167] Lepori M, McCoy R T. Picking BERT's brain: Probing for linguistic dependencies in contextualized embeddings using representational similarity analysis//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain, 2020: 3637-3651

[168] Chrupała G, Alishahi A. Correlating neural and symbolic representations of language//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. 2019: 2952-2962

[169] Evanson L, Lakretz Y, King J R. Language acquisition: Do children and language models follow similar learning stages? //Proceedings of the Association for Computational Linguistics: AC. Toronto, Canada, 2023: 12205-12218

[170] Htut P M, Phang J, Bordia S, et al. Do attention heads in BERT track syntactic dependencies? arXiv preprint arXiv: 1911.12246, 2019

[171] Wu Z, Chen Y, Kao B, et al. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 4166-4176

[172] Zhou Y, Srikumar V. DirectProbe: Studying representations without classifiers//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 5070-5083

[173] Mehta M, Srikumar V. Verifying annotation agreement without multiple experts: A case study with gujarati SNACS//Proceedings of the Association for Computational Linguistics: ACL. Toronto, Canada, 2023: 10941-10958

[174] Sorokin D, Gurevych I. Context-aware representations for knowledge base relation extraction//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, Denmark, 2017: 1784-1789

[175] Hewitt J, Ethayarajh K, Liang P, et al. Conditional probing: measuring usable information beyond a baseline//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic, 2021: 1626-1639

[176] Xu Y, Zhao S, Song J, et al. A theory of usable information under computational constraints//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-24

[177] Ma W, Wang B, Zhang H, et al. Improving syntactic probing correctness and robustness with control tasks//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Toronto, Canada, 2023: 402-415

[178] Cao Y, Zhou L, Lee S, et al. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study//Proceedings of the 1st Workshop on Cross-Cultural Considerations in NLP (C3NLP). Dubrovnik, Croatia. 2023: 53-67

[179] Rutinowski J, Franke S, Endendyk J, et al. The self-perception and political biases of ChatGPT. arXiv preprint arXiv:2304.07333, 2023

[180] Zhuo T Y, Huang Y, Chen C, et al. Exploring ai ethics of chatgpt: A diagnostic analysis//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 2023: 13538-13556

[181] Zhu Z, Wang J, Li B, et al. On the data requirements of probing//Proceedings of the Association for Computational Linguistics: ACL. Dublin, Ireland. 2022: 4132-4147

[182] Lovering C, Jha R, Linzen T, et al. Predicting inductive biases of pre-trained models//Proceedings of the 9th International Conference on Learning Representations. 2021: 1-27

[183] Kunz J, Kuhlmann M. Classifier probes may just learn from linear context features//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain, 2020: 5136-5146

JU Tian-Jie, Ph. D. candidate. His current research interests include natural language processing and interpretability of deep learning.

LIU Gong-Shen, Ph. D., professor. His current research interests include artificial intelligence security and natural language processing.

ZHANG Zhuo-Sheng, Ph. D. candidate. His current research interests include deep learning, natural language processing and pre-trained models.

ZHANG Ru, Ph. D., professor. Her current research interest is digital content security.

Background

This paper mainly focuses on the probing task in NLP, which belongs to AI Interpretability area. As a practical and easy-to-understand diagnostic tool, the probing task has emerged as one of the most popular methods in the NLP field for interpreting linguistic properties encoded by large-scale pre-trained language models. However, recent studies have analyzed the limitations and unreliability of the commonly used probing model, the diagnostic classifier, from both theoretical and experimental perspectives, confirming that the probing results may not accurately reflect the degree of linguistic properties encoded in models, which has attracted wider research interest.

In recent years, several works have conducted general surveys in related areas. These works either take a more detailed perspective (e. g., a discussion of diagnostic classifiers) or a broader one (e. g., a discussion of analytical methods for NLP), lacking a systematic analysis of the latest probing tasks. After investigating a large variety of studies on the application, improvement and discussion of the probing task, this paper summarizes and concludes the existing studies, including basic paradigms, probing methods and interpretation of probing results. For applications, we provide detailed descriptions of commonly used probing methods, especially diagnostic classifiers. For discussing, we summarize the main concerns of the NLP community about the probing task, including the complexity trade-off, interference from confounding factors, and the essence of the probing task. Finally, considering the priorities and difficulties of existing studies, we explore future research directions of designing a more reliable probing task.

This work is partly supported by the Joint Funds of the National Natural Science Foundation of China under No. U21B2020, the National Key R&D Program of China under No. 2023YFC3303805, the National Key R&D Program of China under No. 2022ZD0120304 and the Shanghai Science and Technology Plan under No. 22511104400.