PreCog：探索预训练语言模型中记忆与性能之间的关系

Leonardo Ranaldi（•）、Elena Sofia Ruzzetti 和 Fabio Massimo Zanzotto罗马第二大学企业工程系 ART 研究组，意大利罗马 Viale del Politecnico 1号，邮编00133；（•）{leonardo.ranaldi}@uniroma2.it

arXiv:2305.04673v2 [cs.CL] 9 May 2023

摘要

BERT 等预训练语言模型是令人印象深刻的机器，能够记忆可能经过泛化的学习示例。本文针对记忆与 BERT 在下游任务中的性能之间的相互作用，提出一项小规模且聚焦的分析贡献。我们提出*PreCog*，一种用于评估预训练所产生记忆的度量方法，并分析其与 BERT 性能之间的相关性。实验表明，记忆程度较高的示例更容易被正确分类，这说明记忆是 BERT 成功的关键因素。

1 引言

预训练语言模型（PTLMs）（Peters 等人，2018；Devlin 等人，2019；Liu 等人，2019）是引人关注的机器，凭借在合成神经元中记忆文本泛化规律的能力，主导着自然语言处理任务领域。经过大量无标注数据上的长期预训练后，预训练语言模型已展现出能够利用有限的标注数据有效学习下游任务（Howard 和 Ruder，2018），并泛化到分布外示例（Hendrycks 等人，2020）的能力。大量研究表明，这些预训练语言模型往往会模仿传统的语言学句法模型（Jawahar 等人，2019）和传统的自然语言处理流程（Tenney 等人，2019）。因此，一个关键问题是阐明为何预训练语言模型比传统自然语言处理模块更能利用预训练，而传统模块利用标注语料的效果不及前者。

理解预训练语言模型的学习过程，有助于理解其在下游任务中的表现，并在其表现不佳的场景中改进其语言表征（Kumar 等人，2020）。事实上，与传统流水线中的通用自然语言处理模块不同，预训练语言模型需要针对特定任务进行微调（Devlin 等人，2019），并最终针对新语料库中的特定语言进行领域适配（Jin 等人，2022）。此外，与许多其他机器学习模型一样，经过微调的如果随后针对另一个任务进行微调，预训练语言模型就会失去解决某个任务的能力（Xu 等人，2020），尽管它们的语言模型显然并未改变（Merchant 等人，2020）。这种现象称为灾难性遗忘 (Kirkpatrick 等人，2017），是机器学习中的一种现象。那么，这些模型如何利用预训练和训练样例，目前仍不清楚。

PTLMs（如 BERT（Devlin 等人，2019年）展现出令人印象深刻的能力，能够记忆并可能泛化学习样例。这种能力受到了广泛研究，因为它可能造成极大的危害。事实上，这些预训练语言模型可能会泄露在预训练过程中获取的敏感信息。例如，生成式预训练 Transformer（GPT）的记忆（Radford 和 Narasimhan，2018）曾遭到攻击，并泄露了电话号码和用户名（Carlini 等人，2021年；Thakkar 等人，2021年）。然而，这种简单的记忆能力可能在预训练语言模型的下游任务表现中发挥关键作用。

本文聚焦于探讨记忆在 BERT 执行下游任务时所发挥的作用，提出了一项范围较小但具有针对性的贡献。我们提出*PreCog*，一种极为简单的覆盖度量，用于评估预训练对建模给定示例所需信息的覆盖程度；或者更准确地说，评估 BERT 是否已经部分见过该示例——也就是它是否预先认知了该示例。我们的目标是评估 PreCog 能否预先认知哪些示例会让适配了下游任务的 BERT 做出更好的推理。我们在 GLUE 任务（Wang 等人，2018）上使用 BERT，对 PreCog 进行了广泛实验，观察到 PreCog 能够预测任务适配后的 BERT 在哪些示例上表现更好。除作为预测指标外，PreCog 还表明，对示例的记忆是 BERT 成功的关键因素。

2 相关工作

语言神经网络模型记忆事实的能力毋庸置疑。由于这一能力涉及隐私问题，人们对此进行了深入研究。事实上，LSTM 语言模型对事实的记忆十分牢固，以至于在推理过程中可以检索出个别事实（Carlini 等人，2019年）。这些事实可能会泄露敏感的个人信息，例如与特定人士相关的姓名和地址。此外，Kanerva于1988年提出的稀疏分布式记忆理念重新受到关注，Petroni 等人（2019年）据此提出假设：大语言模型或许可以作为一种巧妙且廉价的方式，轻松构建知识库。即使在图像分类等其他领域，似乎大型神经网络也可能记住整个数据集，因为即便目标标签是随机生成的，这些网络在数据集上的错误率仍然极低（Zhang 等人，2017）。然而，这种记忆事实的能力究竟在多大程度上有助于神经网络执行下游任务，仍不清楚。

一个关键的研究问题是理解大型预训练神经网络如何对已记忆的示例进行泛化。预训练似乎是提升泛化能力的有效策略。事实上，与未经过预训练的方法相比，预训练模型在分布外数据上的泛化能力更强，也更善于检测此类数据（Hendrycks 等人，2020）。然而，这些模型需要大量训练实例，才能在下游任务中发挥这种泛化能力（Tänzer 等人，2022）。因此，鉴于在特定数据集上微调似乎与示例的灾难性遗忘相关（Xu 等人，2020），泛化与记忆可能存在紧密关联。

为了探索记忆与下游任务性能之间的相关性，我们提出了一种分析句子覆盖情况的机制。具体而言，我们采用扰动掩蔽方法，研究基于 Transformer 的预训练语言模型（ PLM ）在预训练阶段见过多少句子。这些方法使我们能够观察预训练对下游任务性能的影响。之所以需要这种新颖的度量方法，是因为当前用于理解覆盖情况的度量方法，例如“遗忘事件”（Toneva 等人，2019年）和反事实记忆（Zhang 等人，2021年），混淆了模型性能与实际记忆。

3 方法与数据

本节介绍了 PreCog，这是我们用于评估预训练在多大程度上涵盖了对给定示例进行建模所需信息的指标（第 3.1节）、两种比较指标 $Lenght$ 和 $LexCov$（第 3.2节），以及实验设置（第 3.3节）。

3.1 *PreCog*：一种用于评估预训练覆盖度的指标

BERT（Devlin 等人，2019年）使用掩码语言模型（MLM）作为两项主要学习任务之一，在数十亿文本词元上进行预训练。具体而言，在预训练期间，MLM 会随机选择任意给定序列中 15% 的词元并将其屏蔽。这 15% 的词元中，有些会以 80% 的概率被替换为特殊词元 [MASK]，，有些会以 10% 的概率被随机词元替换，还有些则会以 10% 的概率保持不变。随后，BERT 学习预测这些被屏蔽的词元。这项任务会持续训练，直到模型接近过拟合。BERT 的主要能力之一便是还原被屏蔽的词元。

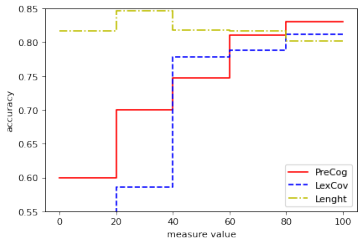
我们旨在衡量 Transformer（如 BERT）的预训练在多大程度上覆盖了一个词元序列。为此，我们基于 BERT 的核心能力，即预测掩码词元。因此，如果 BERT 能够预测给定词元序列中的掩码词元，它可能就掌握了更好地处理该序列所需的知识。我们的直觉是，基于预测掩码词元构建的度量能够描述 BERT 对序列的“先验”知识。

给定由词元列表 $x = [x_1, ..., x_T]$ 表示的句子或文本摘录，我们的函数 $PreCog(x)$ 定义如下。首先，我们逐个遮蔽 x 中的每个词元，得到 T 个不同的序列 $\hat{x}_i = [x_1, ..., x_{i-1}, [MASK], x_{i+1}.., x_T]$。然后，该度量直接定义为：

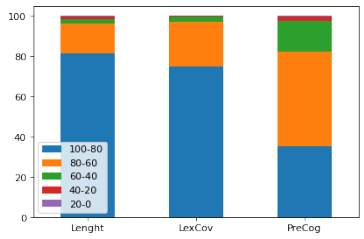
$$PreCog_l(x) = \frac{\sum_{i=0}^{T} \delta(x_i \in BERT_{MLM}(\hat{x}_i))}{T} \quad (1)$$

其中，$BERT_{MLM}(\hat{x}_i)$ 是 BERT 针对位置 i 预测的前 100 个词元的集合，而 $\delta(x_i \in X)$ 若 $x_i \in X$ 则为 1，否则为 0。

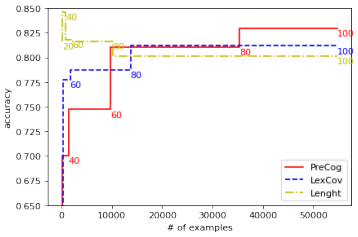
PreCog 是一种非常简单的度量方法。然而，它可能揭示 BERT 如何在下游任务中使用预训练文本的重要事实。一个非常重要的问题是，要了解 PreCog 是否与 BERT 在这些任务中的表现相关。稳定的正相关将为理解预训练的作用提供重要线索



(a) $BERT_{FT}$的准确率直方图，按 20 个点分箱，并绘制其与各项度量值的关系。



(b) 数据集的覆盖率在各项度量值区间内。



(c) $BERT_{FT}$的准确率直方图，按 20 个点分箱，并绘制其与三项度量下各分箱测试集覆盖率的关系。曲线上标出了与相应度量相关的各分箱两端的值。

图 1： $BERT_{FT}$的准确率图，采用 GLUE 各任务准确率的加权和。

理解预训练的作用。

3.2 替代覆盖率度量

为了对 $PreCog$ 进行比较评估，我们使用两个度量：长度和 LexCov。长度旨在衡量 BERT 的准确率与样本长度之间的相关性，而 LexCov 则衡量 BERT 词典的覆盖率。随后，这些度量定义如下：

- $Length(x) = \frac{T - min_D}{max_D - min_D}$其中，T 是 x 的长度， min_D 和 max_D 分别是数据集 D中样本长度的最小值和最大值；
- $LexCov(x) = \frac{T - |OOV(x)|}{T}$其中 $OOV(x)$ 是该示例中相对于 BERT 词汇表的词汇表外词集合 x

3.3 实验设置

为了在多种任务上进行实验，我们使用 GLUE 基准测试（Wang et al., 2018），其中包含以下任务：（1）自然语言推理，即多体裁自然语言推理（MNLI）（Williams et al., 2018）、问题自然语言推理（QNLI）（Wang et al., 2018）、文本蕴涵识别（RTE）（Bentivogli et al., 2009）和 Winograd 自然语言推理（WNLI）（Levesque et al., 2012）；（2）语义相似度，即微软研究释义语料库（MRPC）（Dolan and Brockett, 2005）、语义文本相似度基准测试（STS-B）（Cer et al., 2017）和 Quora 问题对（QQP）（Sharma et al., 2019）；情感分类——斯坦福情感树库（SST-2）（Socher et al., 2013）；以及语言可接受性语料库（CoLA）（Warstadt et al., 2019）。SST-2 和 CoLA 都是单句任务。

我们使用了两个版本的 BERT（Devlin 等人，2019年）： $BERT_{FT}$经过微调的版本和 $BERT_{DA}$经过领域自适应的版本。这两个版本均基于 BERT 序列分类的预训练版本（参见（Wolf 等人，2020））。微调过程采用传统的 BERT 方法。对于每项下游任务，我们均选择 Adam 优化器（Kingma 和 Ba，2015），批量大小设为 16，并遵循原始论文（Devlin 等人，2019年），对 BERT 进行 4 个轮次的微调。在超参数调优方面，每项任务的最佳学习率各不相同，所有原始作者均从以下范围中选择一个值：1×10^{-5} 和 5×10^{-5}。

我们使用配备 CUDA v11.3 的 NVIDIA RTX A6000 GPU 进行实验。我们使用 Transformer 库（Wolf 等人，2020）中的模型，并使用 PyTorch v1.12.0 运行。

为了研究 BERT 的表现与三种指标之一（PreCog、长度或 LexCov）之间的相关性，我们根据指标值将序列分为 5 个区间 x，并绘制了 BERT 准确率相对于这三种指标的直方图（图 1），还计算了指标与准确率之间的皮尔逊相关系数（表 2）。

4 实验结果与讨论

图 1a和图 1c 以及表 2中报告的准确率，是各项 GLUE 任务准确率的加权和。这确保每个 20 分区间都有足够的样本来计算稳定的准确率。

PreCog 与$BERT_{FT}$的准确率的相关性优于 Length 和 LexCov（见图1a和表 2）。不同分箱中 PreCog 的准确率比另外两种度量的准确率下降得更加均匀（图 1a中的红色实线）。此外，PreCog 值与 $BERT_{FT}$ 准确率之间的皮尔逊相关系数为 0.9737，

Task	全局 $BERT_{FT}$	$BERT_{DA}$	区间	# 样本	长度 $BERT_{FT}$	$BERT_{DA}$	# 样本	LexCov $BERT_{FT}$	$BERT_{DA}$	# 样本	PreCog $BERT_{FT}$	$BERT_{DA}$
CoLa	0.920	0.935	(80,100]	499	0.906	0.918	857	0.926	0.940	577	0.951	0.972
			[0,80]	446	0.935	0.955	88	0.852	0.886	368	0.870	0.878
mnli	0.716	0.721	(80,100]	7782	0.717	0.721	6512	0.739	0.745	3508	0.759	0.770
			[0,80]	1361	0.716	0.718	2631	0.660	0.660	5635	0.690	0.690
mrpc	0.806	0.861	(80,100]	59	0.780	0.831	924	0.818	0.877	376	0.867	0.880
			[0,80]	1590	0.806	0.861	725	0.789	0.839	1273	0.787	0.854
qnli	0.808	0.829	(80,100]	3245	0.802	0.832	3123	0.809	0.831	1769	0.832	0.846
			[0,80]	1970	0.817	0.825	2092	0.807	0.827	3446	0.796	0.821
qqp	0.822	0.845	(80,100]	32728	0.820	0.845	28862	0.823	0.843	12810	0.840	0.860
			[0,80]	3990	0.834	0.842	7856	0.816	0.850	23908	0.812	0.837
rte	0.646	0.653	(80,100]	146	0.671	0.678	155	0.716	0.723	46	0.652	0.674
			[0,80]	122	0.615	0.623	113	0.549	0.558	222	0.644	0.649
sst2	0.939	0.924	(80,100]	151	0.907	0.887	607	0.951	0.946	333	0.970	0.970
			[0,80]	655	0.947	0.933	199	0.905	0.859	473	0.918	0.892
wnli	0.565	0.594	(80,100]	31	0.452	0.484	61	0.590	0.623	39	0.590	0.615
			[0,80]	38	0.658	0.684	8	0.375	0.375	30	0.533	0.567

表 1：准确率：根据三项指标的取值对数据集进行分组后，计算得到的 GLUE 任务准确率 s - PreCog, Le xCov 和长度——针对 $BERT_{FT}$ 和 $BERT_{DA}$。

指标	相关性	p 值
长度	-0.5922	0.292
LexCov	0.9014	0.037
PreCog	0.9737	0.005

表 2：在合并的 GLUE 任务中，$BERT_{FT}$ 的各项度量指标与准确率分箱之间的皮尔逊相关系数。

p 值为 0.005，高于 LexCov 的 0.9014（p 值为 0.037）；而 Length 不存在相关性（见表 2）。

PreCog 能更好地区分测试集中的样本。乍一看，LexCov 似乎是区分准确率预期较高和较低的样本的更好模型。LexCov 值低于 40 的样本准确率较低（见图 1a）。然而，LexCov 值介于 0 和 40 之间的样本非常少见（图 1b）。通过绘制按照覆盖率重新缩放后的各区间准确率，可以得到更好的观察结果（图 1c）。事实上，PreCog 比 LexCov 更善于区分样本（图 1c中的红色实线与蓝色虚线）：对于 PreCog，18,000 到 55,000 之间的样本分布在两个区间中，而对于 LexCov，这些样本仅分布在一个区间中。因此，PreCog 比 LexCov 具有更强的判别能力。

在任务层面，结果得到了充分验证：PreCog 能更好地预测任务准确率，也能更好地区分不同类别的样本（见表 1）。对于 $BERT_{FT}$，PreCog 值处于区间 $[80, 100]$ 的样本，其准确率通常高于另外两种指标处于相同区间的样本。$LexCov$ 仅在 RTE 任务中，对 $[80, 100]$ 区间内的样本具有更高的准确率。此外，PreCog 和 LexCov 的 $[80, 100]$ 区间内样本的准确率始终高于 $[0, 80]$ 区间内的样本。然而，PreCog 对样本的划分更均衡，且各区间准确率之间的差异更大。通常，区间 $[80, 100]$和 $[0, 80]$ 之间的准确率差异更高。

此外，领域适应并未改变上述发现。对于 $BERT_{DA}$ 除 SST2 和 WNLI 之外的所有任务，领域适应后的准确率通常都高于未进行领域适应时的准确率（表 2）。此外，聚焦于 PreCog，CoLA、MNLI 和 RTE 中准确率的整体提升源于区间 $[80, 100]$内样本数量的增加。这一事实表明，$BERT_{DA}$针对这些样本获得了更好的模型。

最后需要指出的是，BERT 似乎在预训练期间至少部分见过的句子上表现更好。事实上，PreCog 是一种衡量句子被预训练任务掩码语言模型（MLM）覆盖程度的指标。通常，BERT 会在预训练期间对 MLM 过拟合。因此，PreCog 可以用来判断句子是否已经被部分见过。相比之下，LexCov 描述了句子中有多少词语被 BERT 的词汇表覆盖。由于 PreCog 与 LexCov 在任务预测准确率方面存在很大差异，我们可以得出结论：当目标句子的常识知识已经在预训练期间获得时，BERT 的表现会更好。

5 结论

预训练样例的记忆在 BERT 的性能中发挥着非常重要的作用。事实上，我们提出的 PreCog 衡量了被记忆的预训练知识覆盖目标样例的程度，并且与 BERT 在推理中的性能高度相关。因此，PreCog 还可以作为衡量基于 BERT 的模型在下游任务中所作决策置信度的指标。

鉴于 BERT 的成功部分归功于对样例的简单记忆，

而且 ChatGPT 已广泛普及，未来研究的一个方向应是更深入地理解实际训练样例与推理之间的关系，从而向知识生产者致谢。

局限性

本文为理解记忆与预训练语言模型（PTLMs）性能之间的关系做出了小而集中的贡献。然而，对于这一更长期的目标，仍有一些问题尚未解决。事实上，我们仅针对一种特定的 PTLM——BERT，以及一种特定的预训练任务——掩码语言模型（MLM）探索了这一想法。未来的分析应探讨我们的发现是否适用于其他基于 MLM 的 PTLM。此外，我们尚未探究 PTLM 所使用的预训练语料库在多大程度上真正涵盖了任务样例。应进一步研究 PreCog 与实际训练样例之间的相关性。最后，PreCog 不适用于基于非 MLM 预训练任务的 PTLM。在这些情况下，应定义其他覆盖度指标。

参考文献

Luisa Bentivogli、Bernardo Magnini、Ido Dagan、Hoa Trang Dang 和 Danilo Giampiccolo。2009。第五届 PASCAL 文本蕴含识别挑战赛。载于第二届文本分析会议论文集，*TAC 2009*，盖瑟斯堡，马里兰州，美国，*2009*年*11*月*16*日至*17*日。美国国家标准与技术研究院。

Nicholas Carlini、Chang Liu、Úlfar Erlingsson、Jernej Kos 和 Dawn Xiaodong Song。2019年。《秘密共享者：评估和测试神经网络中的非预期记忆》。载于*USENIX*安全研讨会。

Nicholas Carlini、Florian Tramer、Eric Wallace、Matthew Jagielski、Ariel Herbert-Voss、Katherine Lee、Adam Roberts、Tom Brown、Dawn Song、Ul- far Erlingsson、Alina Oprea 和 Colin Raffel。2021年。从大语言模型中提取训练数据。

Daniel Cer、Mona Diab、Eneko Agirre、Iñigo Lopez- Gazpio 和 Lucia Specia。2017。SemEval-2017任务1：语义文本相似度多语言与跨语言重点评测。载于第*11*届语义评测国际研讨会论文集（*SemEval-2017*），第1–14页，温哥华，加拿大。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 KristinaToutanova。2019年。BERT：用于语言理解的深度双向 Transformer。载于*2019*年计算语言学协会北美分会人类语言技术会议论文集：第*1*卷（长篇与短篇论文），第4171–4186页，美国明尼苏达州明尼阿波利斯。计算语言学协会。

William B. Dolan 和 Chris Brockett。2005年。自动构建句子释义语料库。载于第三届国际释义研讨会（*IWP2005*）论文集。

Dan Hendrycks、Xiaoyuan Liu、Eric Wallace、Adam Dziedzic、Rishabh Krishnan 和 Dawn Song。2020。预训练变换器提升分布外鲁棒性。载于计算语言学协会第*58*届年会论文集，第2744–2751页，线上。计算语言学协会。

Jeremy Howard 和 Sebastian Ruder。2018。面向文本分类的通用语言模型微调。载于计算语言学协会年会。

Ganesh Jawahar、Benoît Sagot 和 Djamé Seddah。2019年。BERT 学到了哪些语言结构？载于计算语言学协会第*57*届年会论文集，第3651–3657页，意大利佛罗伦萨。计算语言学协会。

Xisen Jin、Dejiao Zhang、Henghui Zhu、Wei Xiao、Shang-Wen Li、Xiaokai Wei、Andrew Arnold 和 Xiang Ren。2022。终身预训练：持续适应新兴语料库中的语言模型。载于*2022*年北美计算语言学协会分会年会论文集：人类语言技术，第4764–4780页，美国西雅图。计算语言学协会。

Pentti Kanerva。1988。稀疏分布式记忆。麻省理工学院出版社。

Diederik P. Kingma 和 Jimmy Ba。2015。Adam：一种随机优化方法。*CoRR*，abs/1412.6980。

James Kirkpatrick、Razvan Pascanu、Neil Rabinowitz、JoelVeness、Guillaume Desjardins、AndreiA. Rusu、Kieran Milan、John Quan、Tiago Ramalho、Agnieszka Grabska-Barwinska、Demis Hassabis、Claudia Clopath、Dharshan Kumaran 和 Raia Hadsell。2017。克服神经网络中的灾难性遗忘。《美国国家科学院院刊》，114(13)：3521–3526。

Ankit Kumar、Piyush Makhija 和 Anuj Gupta。2020 年。噪声文本数据：BERT的阿喀琉斯之踵。载于第六届噪声用户生成文本研讨会论文集（*W-NUT 2020*），第 16–21 页，线上。计算语言学协会。

Hector J. Levesque、Ernest Davis 和 Leora Morgenstern。2012 年。Winograd 图式挑战。载于第十三届知识表示与推理原理国际会议论文集，KR'12，第 552–561 页。AAAI 出版社。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。Roberta：一种经过稳健优化的 BERT 预训练方法。*ArXiv*，abs/1907.11692。

Amil Merchant、Elahe Rahimtoroghi、Ellie Pavlick 和 Ian Tenney。2020。 BERT 嵌入在微调过程中发生了什么？载于 第三届关于分析和解释自然语言处理神经网络的 *BlackboxNLP* 研讨会论文集，第 33–44 页，线上。计算语言学协会。

Matthew E. Peters、Mark Neumann、Mohit Iyyer、Matt Gardner、Christopher Clark、Kenton Lee 和 Luke Zettlemoyer。2018。 深度上下文化词表示。载于 *2018* 年北美计算语言学协会分会会议论文集：人类语言技术，第 *1* 卷（长篇论文），第 2227–2237 页，新奥尔良，路易斯安那州。计算语言学协会。

Fabio Petroni、Tim Rocktäschel、Patrick Lewis、Anton Bakhtin、Yuxiang Wu、Alexander H. Miller 和 Sebastian Riedel。2019年。语言模型能否作为知识库？*ArXiv*，abs/1909.01066。

Alec Radford 和 Karthik Narasimhan。2018 年。通过生成式预训练改进语言理解。

LakshaySharma、Laura Graesser、Nikita Nangia 和 Utku Evci。2019年。基于 Quora 问题对数据集的自然语言理解。*ArXiv*，abs/1907.01041。

Richard Socher、Alex Perelygin、Jean Wu、Jason Chuang、Christopher D. Manning、Andrew Ng 和 Christopher Potts。2013 年。基于情感树库的递归深度模型：探索语义组合性载于*2013* 年自然语言处理实证方法会议论文集，第 1631–1642 页，美国华盛顿州西雅图。计算语言学协会。

Michael Tänzer、Sebastian Ruder 和 Marek Rei。2022年。预训练语言模型中的记忆与泛化。载于第*60*届计算语言学协会年会论文集（第一卷：长篇论文），第7564–7578页，爱尔兰都柏林。计算语言学协会。

Ian Tenney、Dipanjan Das 和 Ellie Pavlick。2019年。BERT 重新发现经典的自然语言处理流程。载于第*57*届计算语言学协会年会论文集，第4593–4601页，意大利佛罗伦萨。计算语言学协会。

Om Thakkar、Swaroop Ramaswamy、Rajiv Mathews 和 Françoise Beaufays。2021年。理解联邦学习中的非预期记忆。载于计算语言学协会北美分会*2021*年年会（*NAACL 2021*）第三届自然语言处理隐私研讨会（*PrivateNLP 2021*）。

Mariya Toneva、Alessandro Sordoni、Remi Tachet des Combes、Adam Trischler、Yoshua Bengio 和 Geoffrey J Gordon。2019年。深度神经网络学习过程中的样例遗忘实证研究。载于*ICLR*。

Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel Bowman。2018年。GLUE：自然语言理解多任务基准与分析平台。载于*2018*年*EMNLP*研讨会 *BlackboxNLP*：自然语言处理神经网络的分析与解释论文集，第353–355页，比利时布鲁塞尔。计算语言学协会。

Alex Warstadt、Amanpreet Singh 和 Samuel R. Bowman。2019年。 神经网络可接受性判断。计算语言学协会汇刊，7：625–641。

Adina Williams、Nikita Nangia 和 Samuel Bowman。2018年。 用于通过推理理解句子的广覆盖挑战语料库。载于*2018*年北美计算语言学协会分会会议论文集：人类语言技术，第*1*卷（长篇论文），第1112–1122页，新奥尔良，路易斯安那州。计算语言学协会。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Remi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020。 Transformers：自然语言处理的最先进技术。载于*2020*年自然语言处理经验方法会议论文集：系统演示，第38–45页，线上。计算语言学协会。

Ying Xu、Xu Zhong、Antonio Jose Jimeno Yepes 和 Jey Han Lau。2020 年。别忘了我：减少阅读理解领域适应中的灾难性遗忘。载于*2020*年国际联合神经网络会议（*IJCNN*），第 1–8 页。

Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht 和 Oriol Vinyals。2017年。理解深度学习需要重新思考泛化。载于 第五届国际学习表征会议（*ICLR 2017*）论文集，法国土伦，*2017*年*4*月*24*日至*26*日，会议论文集。OpenReview.net。

Chiyuan Zhang、Daphne Ippolito、Katherine Lee、Matthew Jagielski、Florian Tramèr 和 Nicholas Carlini。2021年。神经语言模型中的反事实记忆。*ArXiv*, abs/2112.12938。