arXiv:2304.01597v1 [cs.CL] 4 Apr 2023

语言模型中事实知识的无监督提升

Nafis Sadeq[*1], **Byungkyu Kang**[2], **Prarit Lamba**[2], **Julian McAuley**[1]
Intuit[2] 和加州大学圣迭戈分校[1]{nsadeq,jmcauley}@ucsd.edu{Jay_Kang,Prarit_Lamba}@intuit.com

摘要

掩码语言建模（MLM）在大语言模型的预训练中发挥着关键作用。然而，掩码语言模型目标往往由高频词主导，而这并不利于学习事实知识。在这项工作中，我们提出了一种影响掩码语言模型预训练的方法，从而提升语言模型在各种知识密集型任务上的表现。我们以完全无监督的方式，促使语言模型优先关注信息量较高的词。实验表明，所提出的方法能够显著提升预训练语言模型在事实回忆、问答、情感分析以及闭卷设置下自然语言推理等任务上的表现。

1 引言

预训练语言模型（PLM），例如 BERT（Devlin 等，2019）、RoBERTa（Liu 等，2019）、BART（Lewis 等，2020）和 T5（Raffel 等，2020），在预训练期间采用掩码语言建模（MLM）目标。然而，传统的掩码语言模型目标可能并不适用于知识密集型任务（Peters 等，2019）。研究表明，在训练目标中以实体嵌入（Peters 等，2019；Zhang 等，2019）、知识检索器（Guu 等，2020）、知识嵌入（Wang 等，2021；Sun 等，2020）或基于知识图谱构建的扩充预训练语料（Agarwal 等，2021）的形式融入知识，可以使语言模型受益。尽管这些方法效果显著，但它们依赖现有知识库和实体嵌入，才能在训练目标中融入知识。这些资源构建成本高昂，而且可能并非适用于所有语言和领域（Huang 等，2022）。

本文提出一种预训练方法，无需使用任何现有知识库，即可在知识密集型任务上取得更好的表现。我们结合两种关键策略来影响掩码语言模型目标。首先，信息相关性较高的词元应更频繁地被屏蔽（Sadeq 等人，2022年）。其次，应对信息丰富的词元上的错误施以更严厉的惩罚。通过对预训练语料进行一次遍历，即可高效计算词元的信息相关性。实验表明，所提出的训练策略有助于语言模型在事实知识回忆基准 LAMA（Petroni 等人，2019年）、抽取式问答（QA）基准 SQuAD（Rajpurkar 等人，2016年，2018年）、基于提示的情感分析和自然语言推理（NLI）任务上取得更好的表现（AutoPrompt）（Shin etal.，2020年）。

本研究的主要贡献是提出了一种完全无监督、独立的掩码语言模型预训练目标，可显著提升语言模型在知识密集型任务上的表现。与该领域以往的研究不同，我们的方法无需依赖现有知识库即可在预训练期间融入知识。我们已公开代码。[1]

2 相关工作

将预训练语言模型用作知识库已有研究表明，BERT 等大规模预训练语言模型可用作知识库（Petroni 等人，2019年，2020年）。以往研究主要关注生成式预训练语言模型中的事实知识（Liu 等人，2021年）、多语言设置（Jiang 等人，2020年a）、实体和查询类型（HeinzerlingandInui，2021年）、事实核查（Lee 等人，2020年）。

*本研究是在第一作者于 Intuit 实习期间完成的。

[1]代码可在此处获取 https://github.com/intuit/wMLM.git

设计更好的提示Jiang 等人（2020年b）提出了基于挖掘和基于释义的方法，用于自动生成提示，以提升事实回忆表现。类似的方法也见于Zhong 等人（2021年）；Ha- viv 等人（2021年）；Qin 和 Eisner（2021年）。Shinet al.（2020年）提出了一种方法，可自动创建掩码语言模型提示，用于情感分析、自然语言推理、关系抽取等多种任务。

预训练期间的知识整合Pe- ters 等人（2019年）使用现有知识库中的实体嵌入，并将实体链接损失与掩码语言模型损失联合使用，以提升 BERT 的事实回忆性能。同样，Zhang 等人（2019年）；Wang 等人（2021年）；Févry 等人（2020年）；Sun 等人（2020年）；Liu 等人（2020年）使用现有知识库中的实体表示或知识表示，将知识整合到预训练语言模型中。Guu 等人（2020年）将知识检索器与语言建模目标联合预训练，以实现知识整合。Agarwal 等人（2021年）从现有知识库中合成文本语料库，并在预训练期间使用该语料库。Sun 等人（2019年）在训练期间使用实体级和短语级知识掩码。

预训练后的知识修改De Cao 等人（2021年）；Zhu 等人（2020年）使用约束优化，在对其余事实知识影响最小的情况下，编辑预训练语言模型中的现有世界知识。同样，Verga 等人（2021年）开发了一种事实注入语言模型架构，无需额外预训练即可轻松将现有知识库整合到预训练语言模型中。

3 方法论

我们使用掩码语言模型目标进行预训练，与以往工作类似（Devlin 等人（2019）；Liu 等人（2019）；Lewis 等人（2020）；Raffel 等人（2020））。给定一个词元序列 Z，会随机采样一部分词元 $X \subset Z$进行替换（在 Devlin 等人（2019）中为 $|X|/|Z| \approx 0.15$）。对于 X中的替换候选词元，其中 80% 的情况下会使用特殊词元 [MASK]，10进行替换，[MASK]，10% 的词元会被替换为随机词元，另外 10% 的候选词元保持不变（Devlin 等人（2019）；Liu 等人（2019）；Joshi 等人（2020））。模型在预训练期间的任务是根据修改后的词元预测原始词元，从而完成被修改词元的恢复。

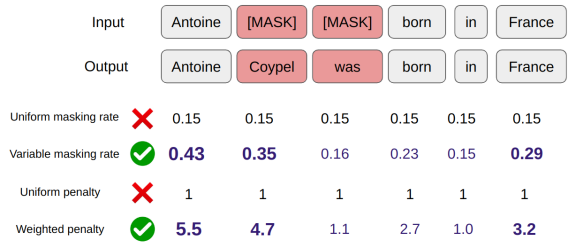


图1：可变掩码率和加权惩罚的简化示意图

定输入序列。对于一组被替换的词元$X(x_1, x_2, ..., x_N)$及其对应的输出词元 $Y(y_1, y_2, ..., y_N)$，损失 $\mathcal{L}_{MLM}$按如下方式计算：

$$\mathcal{L}_{MLM} = -\sum_{i=1}^{N} \log \frac{e^{x_{i,y_i}}}{\sum_{v \in V} e^{x_{i,v}}} \quad (1)$$

这里，$x_{i,\ j}$是给定输入 j时输出候选项 x_i的对数几率，而 V是词汇集合。在传统的掩码语言模型损失计算中，词汇表中的所有词元都受到相同的惩罚项。在我们的工作中，我们试图在预训练期间影响掩码语言模型目标，以融入更多事实知识。我们与传统的掩码语言模型预训练有两点不同：**(a)** 我们不以相同概率掩码所有词元，而是允许信息相关性较高的某些词元被更频繁地掩码；**(b)** 我们使用加权交叉熵损失，对信息相关性较高的某些词元的错误施加更严厉的惩罚。这两个概念的简单示意图如图1所示。我们按如下方式计算损失：

$$\mathcal{L}_{MLM} = -\sum_{i=1}^{N} w_{y_i} \log \frac{e^{x_{i,y_i}}}{\sum_{v \in V} e^{x_{i,v}}} \quad (2)$$

w_{y_i}是针对特定输出词元的惩罚权重 y_i。权重的大小根据词元的信息相关性确定。图1展示了这种加权方式。语言模型词汇表中的每个词元都有与其对应的唯一掩码率和惩罚权重。这些值可以在预训练之前通过一次遍历计算得出。

在此语境中，词元的信息相关性表示某个特定词元就事实知识而言的重要程度。对于事实知识更为重要的词元（例如命名实体）预计会具有更高的

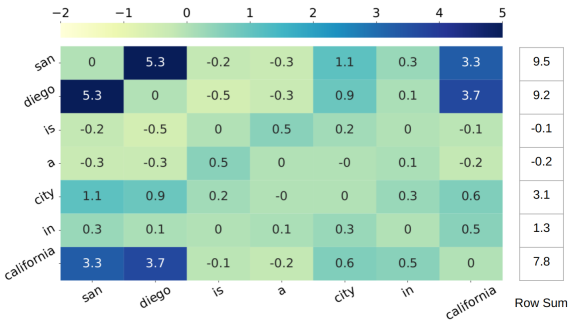


图2：通过对所有词元对的点互信息矩阵按行求和，计算特定文档中词元的信息相关性

信息相关性。我们使用点互信息（PMI（Fano，1961））以无监督的方式计算信息相关性。我们假设，与相邻词具有较高 PMI 的词往往具有较高的信息相关性。首先，我们在跳字窗口内计算预训练语料库的词共现统计信息。其次，计算词汇表中所有词对之间的 PMI。第三，我们以矩阵的形式考虑特定文档中所有词之间的成对 PMI（如图2所示），使该矩阵中的逐行求和反映该文档中各个词元的信息相关性。然后，在整个语料库中对词元的信息相关性取平均值。最后，对计算得到的值进行归一化，并将其转换为词元专属的掩码率和词元专属的惩罚权重。这些掩码率用于创建掩码输入，而惩罚权重则在计算掩码语言模型损失时纳入其中，如公式2所示。

4 实验

4.1 预训练设置

我们使用 Hugging Face（Lhoest 等人，2021年）提供的维基百科语料库进行预训练，使用词汇表大小为 100k 的 WordPiece分词器。词汇表大小经过选择，以确保包含大多数实体。使用大小为 10 的跳元窗口计算词语共现统计信息。存储词语之间点互信息的矩阵大小为 $100k \times 100k$。涉及词元信息相关性的单遍计算大约需要两小时，并需要 11 GB 的内存。单个词元的掩码率根据其信息相关性在 15%-50% 之间变化。所有词元的平均掩码率为 19%。词元的惩罚权重在 $[1, 5]$范围内进行归一化。训练使用 Hugging Face Transformers（Wolf 等人，2020年），在一台配备 4 个 Nvidia V100 GPU 的 AWS p3.8xlarge 机器上完成。我们的模型架构与 BERT-base（Devlin 等人，2019年）相似，包含 12 层，隐藏维度为 768。总批次大小为 128，学习率为 5e-5，并使用 AdamW优化器（Loshchilov 和 Hutter，2019年）。训练进行 10 个训练轮数，最大文档长度为 128。与 BERT（Devlin 等人，2019年）不同，我们在预训练期间不使用下一句预测目标。此外，提高后的掩码率和惩罚权重仅适用于整词词元。对于子词词元，我们使用最低掩码率 15% 和惩罚权重 1。

4.2 评测基准

我们使用 LAMA 知识探测集（Petroni 等人，2019年）评估模型的事实回忆能力。LAMA 包含约 7 万个样本，涵盖 46 种事实关系。为了评估抽取式问答的性能，我们使用 SQuAD v1 和 v2（Rajpurkar 等人，2016年，2018年）。为了评估闭卷问答的零样本性能，我们使用 LAMA 中的 SQuAD 部分（Petroni 等人，2019年）。对于闭卷情感分析和自然语言推理，我们使用 AutoPrompt 中的 SST2 和自然语言推理探测集（Shinet al.，2020年）。我们还报告模型在 GLUE 上的表现（Wang 等人，2018年）。

4.3 基线

我们使用第 4.1 节中提到的相同语料库、分词器和超参数设置，训练了四个模型：**(a)**BERT_{uu}：与 Devlin 等人(2019年）类似，它对所有词元使用统一的掩码率和统一的惩罚项。这是我们的基线。**(b)** BERT_{uw}：使用统一的掩码率和加权惩罚。**(c)** BERT_{vu} (Sadeq 等人，2022年）：对不同词元使用可变掩码率和统一的惩罚项。**(d)** BERT_{vw}：这是我们提出的方法，结合了针对不同词元的可变掩码率和加权惩罚。

4.4 结果与讨论

事实回忆与零样本问答 采用所提出的预训练方法的模型（BERT_{vw}）在 LAMA 的事实回忆任务上显著优于基线（BERT_{uu}）（见表1）。平均倒数排名（MRR）相较于基线提升了 17.5%，

模型	LAMA（Petroni 等人，2019年）				AutoPrompt（Shin et al.，2020年）		
	ConceptNet	GoogleRE	SQuAD	TREx	SST2	自然语言推理（三分类）	自然语言推理（二分类）
$BERT_{uu}$	0.114	0.281	0.156	0.578	0.651	0.397	0.620
$BERT_{uw}$	0.120	0.289	0.169	0.592	0.655	0.439	0.676
$BERT_{vu}$	0.129	0.292	0.175	0.616	0.700	0.457	0.697
$BERT_{vw}$	**0.134**	**0.298**	**0.187**	**0.625**	**0.704**	**0.481**	**0.711**

表1：LAMA 上的事实回忆表现，以及 AutoPrompt 上的情感分析和自然语言推理表现。LAMA 和 AutoPrompt 分别采用平均倒数排名（MRR）和准确率作为评估指标。

模型	SQuAD		GLUE（Wang 等人，2018年）								
	v1（2016年）	v2（2018年）	CoLA	SST2	MNLI	QNLI	QQP	STSB	RTE	WNLI	MRPC
$BERT_{uu}$	69.96	83.22	**31.06**	88.30	79.42	87.72	89.77	85.41	**66.43**	42.25	87.78
$BERT_{uw}$	71.17	84.17	28.55	89.11	79.82	87.15	89.59	85.70	58.84	49.30	87.93
$BERT_{vu}$	71.17	85.07	29.11	89.79	80.02	88.21	**90.10**	85.60	61.37	54.93	88.29
$BERT_{vw}$	**72.61**	**85.28**	28.93	**89.91**	**80.25**	**88.49**	89.82	**85.82**	59.93	**56.34**	**88.32**

表2：在 SQuAD 和 GLUE 开发集上的性能。对于 SQuAD，我们报告 F1分数。对于 CoLA，我们报告马修斯相关系数；对于 STSB，我们报告皮尔逊相关系数；对于其他 GLUE 任务，我们报告准确率。SQuAD 和 GLUE 的微调参数见附录 B。

对于 ConceptNet、GoogleRE 和 TREx，分别为 6% 和 8.1%。LAMA基准中的 SQuAD 部分是一组采用闭卷模板改编的零样本问答样本。在这项任务中，我们相较于基线取得了 19.9% 的相对提升。

关于事实回忆的案例研究见表 3。这些案例研究中有两个关键观察结果。首先，在知识探测期间，所提出的模型（$BERT_{vw}$）更有可能将真实标签排在更高位置。这有助于模型取得更好的总体平均倒数排名。其次，当基线模型只能生成通用词语时，所提出的模型更有可能在给定特定上下文的情况下生成具体词语。例如，当我们使用提示语“在第50届超级碗期间，这家 [MASK] 游戏公司首次发布了其广告”时，基线模型给出的前三个候选词是相对常见的词语，例如“电脑”“电子”和“美国人”。但所提出的模型能够生成与三家游戏公司相关的更具体词语（“任天堂”“Walt”和“雅达利”），其中包括正确答案“任天堂”。在探测语句“运营测量了落在亚马逊上的尘埃的卫星的组织是 [MASK]”中也可以观察到类似现象：所提出的模型能够根据给定上下文作出具体预测，例如“美国国家航空航天局”“巴西”和“亚马逊”。但基线模型只能生成通用词语例如“unknown”、“the”和“unclear”。

闭卷情感分析和自然语言推理 我们使用 AutoPrompt（Shin et al.，2020年）评估系统在闭卷情感分析和自然语言推理方面的表现。AutoPrompt 提供了一种将特定自然语言处理任务转换为基于模板的探测格式的方法。这种提示方式的优势在于，我们可以利用语言模型中的事实知识，而不受微调的限制（Wallat 等人，2020年）。提示包含输入、答案占位符以及一段触发词（提示模板见附录 C）。触发词通过训练数据集进行调优，随后用于评估。与基线相比，所提出的系统在情感分析、三分类自然语言推理和二分类自然语言推理任务上的准确率分别相对提升了 8.1%、21.1% 和 14.7%（表1）。

微调与提示调优 在抽取式问答基准 SQuAD（Rajpurkar 等人，2016年，2018年）和文本分类基准 GLUE（Wang 等人，2018年）上进行微调时，我们提出的模型比基线表现更好。在 SQuAD v1 和 v2 任务以及九项 GLUE 任务中的七项上，它的表现都优于基线（见表 2）。然而，微调带来的相对性能提升不如事实回忆、零样本问答，

输入	真实标签	BERTuu (Devlin 等人，2019年)		BERTvw (提出)	
		预测	得分	预测	得分
为了强调超级碗 50周年，使用了 [MASK] 颜色。	gold	黄色 red 绿色	0.17 0.13 0.12	**gold** 彩虹 橙色	0.09 0.06 0.06
在第50届超级碗期间， [MASK] 游戏公司 首次推出了他们的广告。	任天堂	电脑 电子 美国人	0.06 0.05 0.03	**任天堂** walt 雅达利	0.05 0.04 0.04
教师最有可能在 [MASK]授课。	学校	大学 **学校** 大学	0.61 0.26 0.03	**学校** 大学 神学院	0.40 0.34 0.09
光合作用会将 [MASK] 释放到地球的 大气中。	氧气	阳光 光合作用 光	0.13 0.09 0.09	**氧气** 碳 阳光	0.21 0.12 0.06
运行这颗卫星的组织，而这颗卫星测量了 落在亚马逊上的尘埃是 [MASK] 。	nasa	未知 the 不明确	0.11 0.03 0.03	**nasa** 巴西 亚马逊	0.06 0.05 0.02
美国的收入不平等开始加剧，时间是在 [MASK]。	20世纪 70年代	20世纪60年代 20世纪80年代 **20世纪 70年代**	0.21 0.18 0.17	**20世纪70年代** 20世纪60年代 20世纪80年代	0.14 0.13 0.12
他在 16 岁时搬到 [MASK] 完成他的高中教育 学校学习并获得了他的日本国籍 于1995年。	日本	东京 **日本** 横滨	0.42 0.21 0.03	**日本** 东京 夏威夷	0.19 0.18 0.06
《1914年犯罪法》是一项联邦 立法 [MASK]。	澳大利亚	加拿大 **澳大利亚** 英格兰	0.39 0.07 0.03	**澳大利亚** 田纳西州 加拿大	0.12 0.09 0.09
她也是赫尔辛基市议会议员，并且 是当地党组织的主席，任职于 [MASK]。	赫尔辛基	芬兰 **赫尔辛基** 埃斯波	0.52 0.38 0.01	**赫尔辛基** 芬兰 埃斯波	0.76 0.18 0.03
马克·施瓦恩（生于1966年7月5日）是一位美国人 [MASK], 导演和制片人。	编剧	演员 **编剧** 作家	0.66 0.14 0.13	**编剧** 作家 演员	0.53 0.21 0.16

表3：来自 LAMA（Petroni 等人，2019年）的事实回忆样本案例研究

或提示调优场景。其原因或许可以由 Wallat 等人(2020年) 的研究发现来解释。我们的方法的主要优势在于能够在预训练期间存储更多事实知识。然而，Wallat 等人(2020年) 表明，预训练期间学到的事实知识可能会在微调期间丢失，从而限制我们提出的系统所具备的优势。另一方面，关系探测、零样本问答以及基于提示调优的自然语言处理任务能够更有效地利用我们模型的额外知识，从而取得显著更好的表现。

消融研究 我们研究在掩码语言模型预训练期间，相较于加权惩罚，可变掩码率带来了多少性能提升。通过比较 BERTuw 与 BERTvu （表1和2），可以得出这一点。在大多数情况下，我们发现可变掩码率的表现略优于加权惩罚。

5 结论

在这项工作中，我们提出了一种预训练策略，可有效地将事实知识存储在语言模型中。额外的知识有助于模型在多种知识密集型自然语言处理任务上超越以往方法，例如事实回忆、零样本问答、闭卷情感分析和自然语言推理。我们的模型在针对 SQuAD 和 GLUE 任务进行微调时，也取得了更好的表现。未来，我们计划将这项工作扩展到 T5 等文本到文本预训练模型（Raffel 等人，2020年）。

致谢

本研究部分得到了 Intuit UniversityCollaborationProgramgrant 的资助。我们感谢

匿名审稿人为本研究提供的宝贵反馈。

局限性

所提出系统的一项局限是，在某些微调任务中，其表现不如基线，例如 CoLA（表 2）。所提出的训练目标降低了停用词在预训练目标中的重要性。这可能会对句法至关重要的任务产生负面影响。还需要进一步研究，以了解并缓解这一问题。

伦理声明

所提出系统的一个潜在问题是，如果相关关键词在训练目标中得到优先考虑，这种训练策略可能会放大语言模型已有的有害行为或偏见。减少所提出模型的有害行为或偏见，可以成为未来研究的一个有趣方向。

参考文献

Oshin Agarwal、Heming Ge、Siamak Shakeri 和 Rami Al-Rfou。2021年。基于知识图谱的合成语料库生成，用于知识增强语言模型预训练。载于*2021*年计算语言学协会北美分会会议论文集：人类语言技术，第 3554–3565 页，线上。计算语言学协会。

Nicola De Cao、Wilker Aziz 和 Ivan Titov。2021年。编辑语言模型中的事实知识。载于*2021*年自然语言处理实证方法会议论文集，第 6491–6506 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向变换器预训练。载于计算语言学协会北美分会*2019*年会议论文集：人类语言技术，*NAACL-HLT 2019*，明尼阿波利斯，明尼苏达州，美国，*2019*年*6*月*2*日至*7*日，第*1*卷（长篇与短篇论文），第4171–4186页。计算语言学协会。

Robert M Fano. 1961. 信息传输：通信统计理论。美国人物理学杂志，29(11):793–794。

Thibault Févry、Livio Baldini Soares、Nicholas FitzGer- ald、Eunsol Choi 和 Tom Kwiatkowski。2020年。作为专家的实体：通过实体实现稀疏内存访问监督。载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第 4937–4951 页，线上。计算语言学协会。

Kelvin Guu、Kenton Lee、Zora Tung、Panupong Pa- supat 和 Mingwei Chang。2020年。检索增强的语言模型预训练。载于国际机器学习会议，第 3929–3938 页。PMLR。

Adi Haviv、Jonathan Berant 和 Amir Globerson。2021年。BERTese：学习与 BERT 对话。载于第*16*届计算语言学协会欧洲分会会议论文集：主论文集，第3618–3623页，线上。计算语言学协会。

Benjamin Heinzerling 和 Kentaro Inui。2021年。 将语言模型视为知识库：关于实体表示、存储容量和释义查询。载于 第*16*届欧洲计算语言学协会会议论文集：主卷，第1772–1791页，线上。计算语言学协会。

Zijie Huang、Zheng Li、Haoming Jiang、Tianyu Cao、Hanqing Lu、Bing Yin、Karthik Subbian、Yizhou Sun 和 Wei Wang。2022年。使用自监督自适应图对齐的多语言知识图谱补全。载于 第*60*届计算语言学协会年会论文集（第*1*卷：长篇论文），第474–485页，爱尔兰都柏林。计算语言学协会。

Zhengbao Jiang、Antonios Anastasopoulos、Jun Araki、Haibo Ding 和 Graham Neubig。2020年a。 X- FACTR：从预训练语言模型中进行多语言事实知识检索。载于 *2020*年自然语言处理实证方法会议论文集（*EMNLP*），第 5943–5959 页，线上。计算语言学协会。

Zhengbao Jiang、Frank F. Xu、Jun Araki 和 Graham Neubig。2020年b。 我们如何知道语言模型知道什么？ 《计算语言学协会汇刊》，8:423–438。

Mandar Joshi、Danqi Chen、Yinhan Liu、Daniel S Weld、Luke Zettlemoyer 和 Omer Levy。2020年。SpanBERT：通过表示和预测文本片段改进预训练。计算语言学协会学报，8：64–77。

Nayeon Lee、Belinda Z. Li、Sinong Wang、Wen-tau Yih、Hao Ma 和 Madian Khabsa。2020年。语言模型能否充当事实核查器？ 载于第三届事实抽取与验证研讨会论文集（*FEVER*），第36–41页，线上。计算语言学协会。

Mike Lewis、Yinhan Liu、Naman Goyal、Mar-jan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于计算语言学协会第*58*届年会论文集，*ACL 2020*，线上，*2020*年*7*月*5*日至*10*日，第7871–7880页。计算语言学协会。

Quentin Lhoest、Albert Villanova del Moral、Yacine Jernite、Abhishek Thakur、Patrick von Platen、Suraj Patil、Julien Chaumond、Mariama Drame、Julien Plu、Lewis Tunstall 等。2021年。Datasets：一个用于自然语言处理的社- 区库。载于*2021*年自然语言处理实证方法会议论文集：系统演示，第 175–184 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Weijie Liu、Peng Zhou、Zhe Zhao、Zhiruo Wang、Qi Ju、Haotang Deng 和 Ping Wang。2020年。K-BERT：利用知识图谱实现语言表示。载于第三十四届*AAAI*人工智能会议，*AAAI 2020*，第三十二届人工智能创新应用会议，*IAAI 2020*，第十届*AAAI*人工智能教育进展研讨会，*EAAI 2020*，美国纽约，*2020*年*2*月*7*日至*12*日，第2901–2908页。AAAI出版社。

Xiao Liu、Yanan Zheng、Zhengxiao Du、Ming Ding、Yujie Qian、Zhilin Yang 和 Jie Tang。2021年。 GPT 也能理解。 *CoRR*，abs/2103.10385。

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。 *CoRR*，abs/1907.11692。

Ilya Loshchilov 和 Frank Hutter。2019年。 解耦权重衰减正则化。载于 第七届国际学习表示会议，*ICLR 2019*，美国路易斯安那州新奥尔良，*2019*年*5*月*6*日至*9*日。OpenReview.net。

Matthew E. Peters、Mark Neumann、Robert L. Logan IV、Roy Schwartz、Vidur Joshi、Sameer Singh 和 Noah A. Smith。2019年。 知识增强的上下文词表示。载于*2019*年自然语言处理经验方法会议暨第九届自然语言处理国际联合会议论文集，*EMNLP-IJCNLP 2019*，中国香港，*2019*年*11*月*3*日至*7*日，第43–54页。计算语言学协会。

Fabio Petroni、Patrick S. H. Lewis、Aleksandra Piktus、Tim Rocktäschel、Yuxiang Wu、Alexander H. Miller 和 Sebastian Riedel。2020年。上下文如何影响语言模型的事实预测。载于自动知识库构建会议，*AKBC 2020*，线上举行，*2020*年*6*月*22*日至*24*日。

Fabio Petroni、Tim Rocktäschel、Sebastian Riedel、Patrick Lewis、Anton Bakhtin、Yuxiang Wu 和 AlexanderMiller。2019年。语言模型能否作为知识库？载于*2019*年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第2463–2473页，中国香港。计算语言学协会。

Guanghui Qin 和 Jason Eisner。2021年。学习如何提问：使用软提示混合查询语言模型。载于*2021*年美国计算语言学协会北美分会会议论文集：人类语言技术，*NAACL-HLT 2021*，线上，*2021*年*6*月*6*日至*11*日，第5203–5212页。计算语言学协会。

Colin Raffel, Noam Shazeer, Adam Roberts, Kather- ine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li 和 Peter J. Liu。2020年。探索统一文本到文本转换器迁移学习的极限。《机器学习研究杂志》，21(140):1–67。

Pranav Rajpurkar、Robin Jia 和 Percy Liang。2018年。了解自己不知道什么：SQuAD 中无法回答的问题。载于 第 *56* 届计算语言学协会年会论文集（第 *2* 卷：短篇论文），第 784–789 页，澳大利亚墨尔本。计算语言学协会。

Pranav Rajpurkar、Jian Zhang、KonstantinLopyrev 和 Percy Liang。2016年。SQuAD：100 个000+ 用于文本机器阅读理解的问题。载于*2016*年自然语言处理实证方法会议论文集，第 2383–2392 页，得克萨斯州奥斯汀。计算语言学协会。

Nafis Sadeq、Canwen Xu 和 Julian McAuley。2022年。InforMask：用于语言模型预训练的无监督信息掩码。载于*2022*年自然语言处理实证方法会议论文集，第 5866–5878 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Taylor Shin、Yasaman Razeghi、Robert L. Logan IV、Eric Wallace 和 Sameer Singh。2020年。AutoPrompt：通过自动生成的提示从语言模型中提取知识。载于 *2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第 4222–4235 页，线上。计算语言学协会。

Tianxiang Sun、Yunfan Shao、Xipeng Qiu、Qipeng Guo、Yaru Hu、Xuanjing Huang 和 Zheng Zhang。2020年。Colake：上下文化语言与知识嵌入。载于 第*28*届国际计算语言学会议论文集，*COLING 2020*，西班牙巴塞罗那（线上），*2020*年*12*月*8*日至*13*日，第3660–3670页。国际计算语言学委员会。

Yu Sun、Shuohuan Wang、Yu-Kun Li、Shikun Feng、Xuyi Chen、Han Zhang、Xin Tian、Danxiang Zhu、Hao Tian 和 Hua Wu。2019年。ERNIE：通过知识整合增强表示。*CoRR*，abs/1904.09223。

Pat Verga、Haitian Sun、Livio Baldini Soares 和 William Cohen。2021年。可适配且可解释的神经记忆优于符号知识。载于*2021*年计算语言学协会北美分会会议论文集：人类语言技术，第3678–3691页，线上。计算语言学协会。

Jonas Wallat、Jaspreet Singh 和 Avishek Anand。2020年。BERTnesia：探究 BERT 中知识的捕获与遗忘。载于第三届 *BlackboxNLP*研讨会论文集：自然语言处理神经网络的分析与解读，第174–183页，线上。计算语言学协会。

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

Xiaozhi Wang、Tianyu Gao、Zhaocheng Zhu、Zhengyan Zhang、Zhiyuan Liu、Juanzi Li 和 Jian Tang。2021年。Kepler：一种用于知识嵌入和预训练语言表示的统一模型。计算语言学协会汇刊，9：176–194。

Thomas Wolf、Lysandre Debut、Victor Sanh、J ulien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Rémi Louf、Morgan Funtow- icz、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 M. Rush。2020年。Transformers：自然语言处理领域的最先进技术。载于 *2020*年自然语言处理实证方法会议论文集：系统演示，*EMNLP 2020* - 演示，线上，*2020*年*11*月*16*日至*20*日，第 38–45 页。计算语言学协会。

Zhengyan Zhang、Xu Han、Zhiyuan Liu、Xin Jiang、Maosong Sun 和 Qun Liu。2019年。ERNIE：通过信息丰富的实体增强语言表示。载于计算语言学协会第*57*届年会论文集（*ACL 2019*），意大利佛罗伦萨，*2019*年*7*月*28*日至*8*月*2*日，第*1*卷：长篇论文，第1441–1451页。计算语言学协会。

Zexuan Zhong、Dan Friedman 和 Danqi Chen。2021年。事实探测是 [MASK]：学习与学习回忆。在 *2021*年会议论文集的计算语言学协会北美分会：人类语言技术，第5017–5033页，线上。计算语言学协会。

ChenZhu、AnkitSinghRawat、ManzilZaheer、Srinadh Bhojanapalli、Daliang Li、Felix X. Yu 和 Sanjiv Kumar。2020年。修改Transformer模型中的记忆。*CoRR*，abs/2012.00363。

按关系划分的 PerformanceonLAMA

领域	数据集	$BERT_{uu}$	$BERT_{uw}$	$BERT_{vu}$	$BERT_{vw}$
ConceptNet	test	0.114	0.120	0.129	**0.134**
GoogleRE	出生日期	0.099	0.109	0.111	**0.113**
GoogleRE	出生地	0.456	0.459	0.461	**0.465**
GoogleRE	死亡地	0.288	0.300	0.305	**0.315**
Squad	test	0.156	0.169	0.175	**0.187**
TREx	P1001	0.779	0.770	0.793	**0.798**
TREx	P101	0.442	0.468	0.501	**0.514**
TREx	P103	0.822	0.834	**0.838**	0.836
TREx	P106	0.642	0.653	**0.675**	0.664
TREx	P108	0.491	0.526	0.538	**0.556**
TREx	P127	0.586	0.615	0.620	**0.636**
TREx	P1303	0.380	0.427	0.433	**0.472**
TREx	P131	0.690	0.702	0.741	**0.750**
TREx	P136	0.595	0.629	0.651	**0.675**
TREx	P1376	0.747	0.761	0.783	**0.792**
TREx	P138	0.633	0.640	0.656	**0.680**
TREx	P140	0.569	0.574	**0.608**	0.602
TREx	P1412	0.764	0.773	**0.785**	0.781
TREx	P159	0.535	0.551	0.573	**0.576**
TREx	P17	0.870	0.863	0.884	**0.887**
TREx	P176	0.647	0.673	0.699	**0.720**
TREx	P178	0.569	0.592	0.631	**0.639**
TREx	P19	0.477	0.478	0.509	**0.519**
TREx	P190	0.279	0.276	0.296	**0.297**
TREx	P20	0.511	0.533	0.559	**0.565**
TREx	P264	0.247	0.280	0.291	**0.313**
TREx	P27	0.745	0.756	0.767	**0.773**
TREx	P276	0.625	0.623	0.652	**0.663**
TREx	P279	0.512	0.544	0.562	**0.580**
TREx	P30	0.802	0.813	0.835	**0.842**
TREx	P31	0.616	0.627	0.635	**0.635**
TREx	P36	0.569	0.578	**0.618**	0.615
TREx	P361	0.530	0.538	0.567	**0.574**
TREx	P364	0.703	0.715	0.729	**0.742**
TREx	P37	0.701	0.688	**0.728**	0.715
TREx	P39	0.572	0.607	0.613	**0.630**
TREx	P407	0.638	0.630	0.647	**0.666**
TREx	P413	0.422	0.453	0.483	**0.507**
TREx	P449	0.416	0.444	0.454	**0.495**
TREx	P463	0.646	0.674	0.697	**0.713**
TREx	P47	0.492	0.508	0.564	**0.565**
TREx	P495	0.685	0.662	**0.699**	0.681
TREx	P527	0.423	0.452	0.521	**0.527**
TREx	P530	0.379	0.373	0.400	**0.416**
TREx	P740	0.407	0.414	0.438	**0.438**
TREx	P937	0.528	0.541	0.569	**0.569**

表 4：LAMA 上按关系划分的性能比较（Petroni 等人，2019年）

B GLUE、SQuAD 微调超参数

超参数	GLUE	SQuAD
批次大小	32	12
学习率	2e-5	3e-5
训练轮数	3	2
权重衰减	0.01	0.01

表 5：GLUE 和 SQuAD 的微调超参数

C AutoPrompt 的超参数

超参数	SST2	NLI
# 触发词元	3	4
# 候选项	100	10
批次大小	24	32
# 迭代次数	180	100

表 6：AutoPrompt 的提示调优超参数（Shin et al.，2020年）

Task	模板	提示示例	标签
SST2	{sentence} [T] . . . [T] [P]	导演罗布·马歇尔带着十足干劲，力图制作一部太棒了 电影导演 电影院 [MASK]	**词性：** 合作关系，好的 **负面：** 更差、糟糕
NLI	{prem}[P] [T] . . . [T] {hyp}	没有穿黑色夹克的人在做特技，身旁有一辆自行车。摩托车 [MASK] 奇怪的工作场所 一个人坐在黑色夹克正在摩托车上玩特技。	**反方：** 没有人，没有人，也没有 **实体：** 发现、方式、机构 **神经：** ##ponents, ##lary, ##已评估

表 7：提　示模板，用于 AutoPrompt 中的情感分析和自然语言推理任务（Shin 等人，1.，2020年）