使用自动生成提示从语言模型中引出知识

Taylor Shin[*◇] **Yasaman Razeghi**[*◇] **Robert L. Logan IV**[*◇]
Eric Wallace[♠] **Sameer Singh**[◇]
◇加利福尼亚大学尔湾分校 ♠加利福尼亚大学伯克利分校
{tshin1, yrazeghi, rlogan, sameer}@uci.edu
ericwallace@berkeley.edu

摘要

预训练语言模型的显著成功激发了对这些模型在预训练期间学习了哪些知识的研究。 将任务重新表述为填空问题（例如，完形填空测试）是评估这种知识的一种自然方法，但是，由于编写合适提示所需的手动工作和猜测，其使用受到限制。 为了解决这个问题，我们开发了一个名为AUTOPROMPT的自动化方法，基于梯度引导搜索来创建各种任务的提示。 使用AUTO-PROMPT，我们展示了掩码语言模型（MLM）在没有额外参数或微调的情况下执行情感分析和自然语言推理的内在能力，有时性能与最新的监督模型相当。 我们还展示了我们的提示在LAMA基准测试中从MLMs中引出更准确的事实知识，以及MLMs可以比监督关系提取模型更有效地用作关系提取器。 这些结果表明，自动生成的提示是现有探测方法的一种可行的无参数替代方案，并且随着预训练语言模型变得更加复杂和强大，有可能成为微调的替代方案。

1 引言

预训练的语言模型（LMs）通过微调（Peters等，2018年；Devlin等，2019年）在下游任务中取得了非凡的成功。 尽管明确地表明预训练提高了准确性，但很难确定微调的LMs所包含的知识是在预训练期间学习的还是在微调期间学习的。

过程。 我们如何直接评估预训练的LMs中存在的语言、事实、常识或任务特定的知识？

已经提出了许多技术来通过分析预训练的LMs的内部表示来引出这种知识。 一种常见的策略是使用探测分类器，即使用LMs的表示作为特征来预测某些属性（Conneau等，2018年；Liu等，2019年）。 然而，探测分类器需要额外的学习参数，因此容易产生误报；高探测准确性并不足以得出LM包含某个特定知识的结论（Hewitt和Liang，2019年；Voita和Titov，2020年）。 另一种常见的技术是注意力可视化，但它也存在类似的失败模式：注意力分数可能与目标知识相关，但并不是由底层目标知识引起的，这导致对它们的使用提出了批评（Jain和Wallace，2019年；Wiegreffe和Pinter，2019年）。 探测和注意力可视化都难以评估无法表示为简单的令牌或序列级分类任务的知识。

一种更直接的方法是通过提示来引出这些模型的知识，因为它们毕竟是语言模型。 例如，Radford等人（2019）通过在文章末尾添加“TL;DR:”来将摘要建模为语言建模任务，然后从语言模型中生成。类似地，Petroni等人（2019）通过手动重构知识库完成任务，将其转化为填空测试（即，填写空白问题）。 与现有的模型分析方法相比，提示是非侵入性的：它不会引入大量的额外参数，也不需要直接检查模型的表示。因此，提示提供了一种有效的方法来从语言模型中获取知识。

[*] 前三位作者贡献相同。

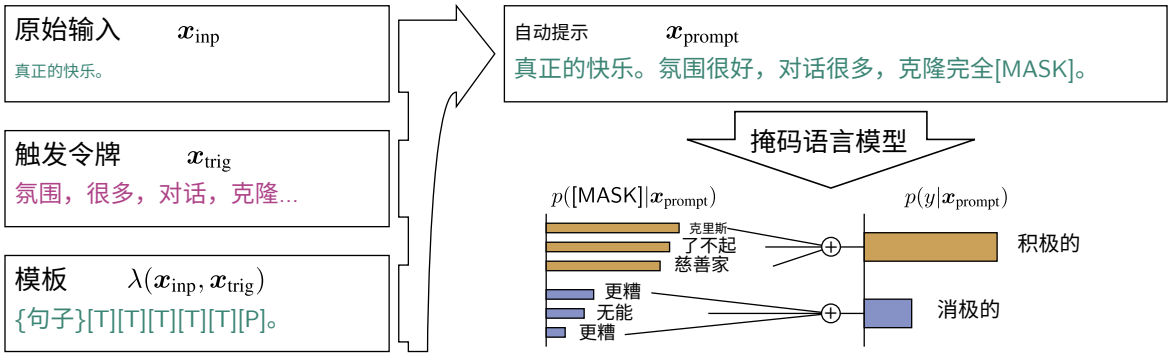


图1：应用AUTOPROMPT来探测掩码语言模型（MLM）进行情感分析的示意图。 每个输入， x_{inp}，都放置在一个自然语言提示， x_{prompt}，中，其中包含一个[MASK]令牌。 使用模板， λ，将原始输入与一组触发令牌， x_{trig}，结合创建提示。 触发令牌在所有输入之间共享，并使用基于梯度的搜索（第2.2节）确定。 然后，通过对自动检测到的标签令牌集合（第2.3节）进行边际化MLM预测，获得每个类别标签， y，的概率。

提供了模型“知道”的下限，因此是一个更有用的分析工具。 然而，提示不幸地需要手动制作上下文来输入模型。 这不仅耗时且对许多任务（例如文本蕴含）来说不直观，更重要的是，模型对这个上下文非常敏感：构造不当的上下文会导致人为地低性能（Jiang等，2020年）。 克服手动指定提示的需求将使提示成为一个更广泛有用的分析工具。

在本文中，我们介绍了AUTOPROMPT——一种用于生成任何任务提示的自动化方法，如图1所示。 给定一个任务，例如情感分析，AUTOPROMPT通过将原始任务输入（例如评论）与一组触发令牌结合起来，根据一个模板创建一个提示。 对于所有输入都使用相同的触发令牌，并使用Wallace等人（2019年）提出的一种梯度搜索策略的变体进行学习。 通过对与标签令牌相关联的一组标签进行边际化，将提示的LM预测转换为类别概率，这些标签可以是学习或预先指定的，使得可以像评估任何其他分类器一样评估LM。

我们在多个实验中验证了AUTOPROMPT的有效性。 首先，我们使用AUTO-PROMPT构建了用于测试预训练掩码语言模型（MLM）在情感分析和自然语言推理（NLI）上的提示。 我们的测试结果显示，在没有任何微调的情况下，MLMs在这两个任务上表现良好-

经过适当提示的RoBERTa在SST-2上达到了91%的准确率（优于经过微调的ELMo模型（Peters等人，2018）），在平衡的SICK-E数据集变体上达到了69%的准确率（Marelli等人，2014）。 接下来，我们将AUTOPROMPT应用于LAMA（Petroni等人，2019）的事实检索任务中，我们能够构建出比使用手动和语料库挖掘方法生成的提示更有效地引出MLM的事实知识的提示。 具体而言，我们实现了43.3%的1-精确度，而当前最佳单提示结果为34.1%（Jiang等人，2020）。 我们还引入了这个任务的一个变体，类似于关系抽取（RE），测试MLMs是否能够从给定的文本中提取知识。 我们展示了当提供具有真实事实的上下文句子时，MLMs实际上可以胜过现有的关系抽取模型，然而，当上下文句子被人为篡改时，它们会遇到困难。

最后，尽管AUTOPROMPT的目标是分析模型，但我们发现它在实践中比微调具有一定的优势。 首先，在低数据情况下，AU-TOPROMPT的平均和最差准确性都比微调更高。

此外，与微调不同，提示语言模型不需要大量的磁盘空间来存储模型检查点；一旦找到一个提示，它可以在现成的预训练语言模型上使用。这在为多个任务提供模型服务时非常有益。

AUTOPROMPT概述

从预训练语言模型中引出知识的一种自然方法是将任务构建为填空问题。

然而，编写提示不仅耗时，而且不清楚相同的措辞是否对每个模型都有效，也不清楚什么标准决定了特定措辞对于引出所需信息最好。 鉴于此，我们引入了AUTOPROMPT，一种为特定任务和感兴趣的MLM构建定制提示的方法，以使MLM产生所需的知识。[1]AUTOPROMPT的示例如图1所示。 提示是通过将原始任务输入（一个或多个令牌序列，例如图1中的评论）映射到一个令牌序列来构建的，使用了一个模板。 在接下来的几节中，我们将介绍AUTOPROMPT如何使用带标签的训练数据来构建提示，以及如何将MLM的输出作为任务的预测。

2.1 背景和符号

为了构建提示，我们区分原始任务输入 $\boldsymbol{x}_{\text{inp}}$（例如，图1中的评论“a real joy.”）和提示 $\boldsymbol{x}_{\text{prompt}}$（例如，“a real joy. atmosphere alot dialogue Clonetotally [MASK].”）被馈送到MLM中。从 $\boldsymbol{x}_{\text{inp}}$到 $\boldsymbol{x}_{\text{prompt}}$的映射使用模板$\lambda$来执行。 该模板定义了每个输入序列在提示中的位置，以及任何其他标记的放置位置。 特别地，它还必须定义MLM填充的特殊 [MASK]标记的位置（在模板中用[P]表示以区别于可能出现的其他 [MASK]标记）。 将提示输入MLM会产生一个概率分布 p（[MASK]$|\boldsymbol{x}_{\text{prompt}}$），描述哪些标记最有可能填充空白处。

如果类别标签自然对应于词汇中的标记（例如，知识库完成任务中的实体名称），则该分布可以被解释为类别标签的分布。 然而，对于情感分析等任务，可能存在一组与特定标签y对应的标记 $\mathcal{V}\,y$。 例如，在图1中，“Cris”、“marvelous”和“philanthrop”都表示积极情感。 在这种情况下，类别概率通过对标记进行边际化得到。

[1]虽然我们在这项工作中只关注自回归语言模型，但我们的方法可以轻松扩展到自动回归语言模型。唯一的调整是预测标记必须出现在提示的末尾。

标记集合：

$$p(y|\boldsymbol{x}_{\text{prompt}}) = \sum_{w \in \mathcal{V}_y} p([\text{MASK}] = w|\boldsymbol{x}_{\text{prompt}}) \tag{1}$$

2.2 基于梯度的提示搜索

到目前为止，我们已经展示了如何使用提示将分类任务重新定义为语言建模任务。 在这里，我们提出了一种基于Wallace等人（2019年）的方法来自动构建提示。
这个想法是添加一些“触发”标记，这些标记在所有提示中都是共享的（在图1的示例模板中用[T]表示）。 这些标记被初始化为[MASK]标记，然后在批量示例上迭代地更新，以最大化标签似然（方程（1））。

形式上，在每一步，我们计算出通过交换第j个触发标记x（j）而产生的对数似然变化的一阶近似。 然后，我们确定一个候选集$\mathcal{V}$cand，这些候选集是估计会导

$$\mathcal{V}_{\text{cand}} = \underset{w \in \mathcal{V}}{\text{top-}k} \left[\boldsymbol{w}_{\text{在}}^{T} \nabla \log p(y|\boldsymbol{x}_{\text{prompt}})_{\text{中}} \right] \tag{2}$$

其中 $\boldsymbol{w}_{\text{in}}$是 w的输入嵌入，梯度是相对于输入嵌入 $x^{(j)}$的

触发. 请注意，计算这个候选集的开销与模型的单次前向传播和反向传播相当（点积需要与计算LM输出投影相同数量的乘法）。 对于这个集合中的每个候选项，我们在更新后的提示上重新评估方程（1），并保留下一步中概率最高的提示-这需要模型的 k次前向传播。 这种方法生成的情感分析任务的示例提示如图1所示。

2.3 自动化标签令牌选择

虽然在某些情况下，标签令牌的选择是显而易见的（例如，当类标签直接对应词汇中的单词时），但对于涉及更抽象类标签的问题（例如，NLI），什么样的标签令牌是合适的则不太清楚。在本节中，我们开发了一种通用的两步方法来自动选择标签令牌集合$\mathcal{V}_y$。 在第一步中，我们使用上下文化嵌入的 [MASK]令牌作为输入，训练一个逻辑回归分类器来预测类标签：

$$\boldsymbol{h} = \text{Transformer}_{\text{enc}}(\boldsymbol{x}) \tag{3}$$

我们将这个分类器的输出写为：

$$p(y|\boldsymbol{h}^{(i)}) \propto \exp(\boldsymbol{h}^{(i)} \cdot \boldsymbol{y} + \beta_y) \tag{4}$$

其中 $\boldsymbol{y}$和 β_y是学习到的权重和偏置项， i表示[MASK]标记的索引。

在第二步中，我们用MLM的输出词嵌入 $\boldsymbol{w}_{\text{out}}$替换 $\boldsymbol{h}^{(i)}$以获得得分 $s(y, w) = p(y|\boldsymbol{w}_{\text{out}})$。直观地说，因为对于与特定上下文相关的单词和标签， $\boldsymbol{w}_{out} \cdot \boldsymbol{h}$和 $\boldsymbol{y} \cdot \boldsymbol{h}$都很大，所以对于通常与给定标签相关联的单词，$s_w \propto \exp(\boldsymbol{w}_{\text{out}} \cdot \boldsymbol{y} + \beta_y)$应该很大。 然后，从得分最高的$k$个单词中构建标签标记的集合：

$$\mathcal{V}_y = \underset{w \in \mathcal{V}}{\text{top-}k} \left[s(y, w) \right] \tag{5}$$

2.4 与其他提示方法的关系

我们的工作属于通过提示来探索语言模型知识的一系列工作。 以前的研究使用手动定义的提示来研究语言模型在常识推理（Trinh和Le，2018；Kwon等，2019；Shwartz等，2020）、问答（Lewis等，2019）、事实回忆（Petroni等，2019；Jiang等，2020；Bouraoui等，2019）、摘要（Radford等，2019）和其他监督任务（Brown等，2020）方面的能力。 Schick和Schutze（2020）在少样本学习中使用手动构建的提示与半监督学习相结合。 相反，我们自动创建任何任务的提示，这提高了准确性并开启了新的现象分析。

2.5 评估设置

在接下来的几节中，我们使用自动生成的提示来探测BERTBASE2（110M参数）和RoBERTa$_{\text{LARGE}}$（355M参数）对以下任务的知识：情感分析、自然语言推理（NLI）、事实检索和关系抽取。 我们使用由transformersPython库（Wolf等，2019）提供的PyTorch实现和预训练权重。

对于情感分析和NLI，我们使用基于逻辑回归的启发式方法在第2.3节中描述的方式找到标签标记。对于事实检索和关系抽取，我们跳过这一步骤，因为标签（实体）直接对应词汇表中的标记。 对于所有任务，我们执行多次迭代的搜索。

我们使用第2.2节中描述的搜索方法进行多次迭代。 在每次迭代中，我们使用一批训练数据来确定替换触发标记的候选集 $\mathcal{V}$cand。 然后，我们在单独的一批数据上评估更新后的提示的标签似然，并在搜索的下一次迭代中保留最佳触发标记。 在每次迭代结束时，我们在保留的开发数据上测量标签似然，并将整个搜索过程中找到的最佳提示作为最终输出。 性能评估使用相应的任务特定度量标准，例如情感分析和NLI的准确性，以及事实检索的精确度@k。在单独的保留测试集上进行评估。我们的AUTOPROMPT实现公开可用，网址为http://ucinlp.github.io/autoprompt，并支持在HuggingFacetransformers库（Wolf等，2019）上对预训练模型进行提示生成。

情感分析3

情感分析是自然语言处理中的一个基本任务，既用于自然语言理解研究，也用于实际应用。 在没有微调的情况下，很难判断MLM对情感的理解程度。

设置 我们将我们的方法应用于将实例从二进制斯坦福情感树库（Socher等人，2013年，SST-2）转换为提示，使用标准的训练/测试分割。 我们使用基于Ta-ble [3]中的模板的提示找到标签令牌。 对于我们基于梯度的提示搜索，我们在以下超参数上进行网格搜索： $|\mathcal{V}_{cand}| \in \{10, 100\}$, $|\mathcal{V}_y| \in \{1, 3, 5\}$,$|\boldsymbol{x}_{\text{trig}}| \in [3, 6]$.[3] 所有提示都使用相同的模板进行初始化，该模板用于找到标签集。我们还手动构建了一个提示（在生成自动提示之前，以避免偏见），基于SST-2由电影评论组成的直觉。 我们使用“{sentence}这部电影[P]。"作为模板，并使用“糟糕”和“棒极了”作为负面和正面标签令牌。”。

结果 我们在表1中展示了结果，同时还有来自GLUE（Wang等人，2019）SST-2排行榜的参考分数，以及通过对LM标记表示的逐元素平均进行训练的线性探测器的分数。 由AUTOPROMPT生成的提示显示BERT和

为了简洁起见，我们将在模型名称中省略下标。

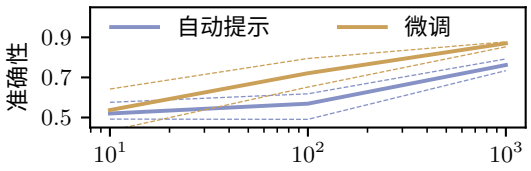
每次运行在一个NVIDIA 2080Ti GPU运行需要2天时间。

模型	开发集	测试集
双向LSTM	-	82.8†
双向LSTM + ELMo	-	89.3†
BERT（线性探测）	85.2	83.4
BERT（微调）	-	93.5†
RoBERTa（线性探测）	87.9	88.8
RoBERTa（微调）	-	96.7†
BERT（手动）	63.2	63.2
BERT (AUTOPROMPT)	80.9	82.3
RoBERTa（手动）	85.3	85.2
RoBERTa (AUTOPROMPT)	91.2	91.4

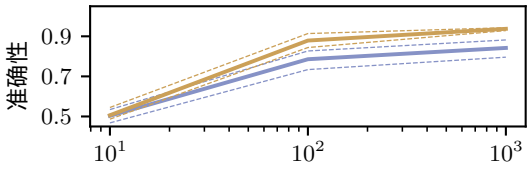
表1：情感分析在**SST**-2测试集上的性能，包括有监督分类器（顶部）和填空MLMs（底部）。 得分标有 †来自GLUE排行榜：http://gluebenchmark.com/leaderboard。

BERT和RoBERTa在情感分析方面具有很强的知识：在没有任何微调的情况下，BERT的性能与有监督的双向LSTM相当，而RoBERTa的准确率与微调的BERT和ELMo模型相当。 此外，我们观察到我们自动生成的提示比手动提示更有效，并且使用人类直觉很难构建这些提示：对于RoBERTa来说，最好的模板是“{sentence} at-mosphere alot dialogue Clone totally [P].”我们在附录A中包含了关于AUTOPROMPT超参数的影响结果。

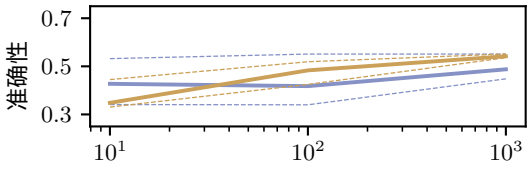
在低数据环境中的准确性尽管**A**UTOPROMPT的目标是探测模型的知识，但我们还发现它可能是低数据情况下微调的一个可行替代方案。 为了证明这一点，我们测量了使用训练数据的10、100和1000个随机子集时**A**U-TOPROMPT提示的开发集准确率。 我们使用 $|\boldsymbol{x}_{\text{trig}}|=10$,$|\mathcal{V}_y|=3$和 $|\mathcal{V}_{\text{cand}}|=10$来运行我们的提示搜索。 我们将其与在相同数据上微调的BERT和**RoBERTa**的性能进行比较。 为了公平比较**A**U-TOPROMPT和微调，我们使用**Mosbach**等人(2020)在小数据集上推荐的微调参数：训练20个**epochs**，使用**Adam W**(**Loshchilov**和**Hutter**, 2018)进行偏差校正，并且学习率在前10%的迭代中线性增加到 2×10^{-5}，之后线性减少到 0。 实验在数据的随机子集上重复10次(以及微调模型的种子)。 最佳情况、最差情况和平均性能如图2所示。 请注意，EMNLP版本中的结果存在一个已经修复的错误。



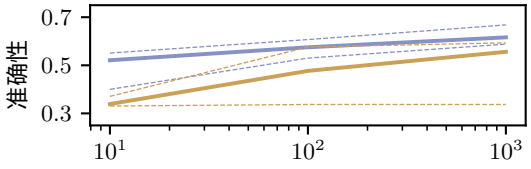
(a) SST-2上的BERT



(b) SST-2上的RoBERTa



(c) SICK-E上的BERT



(d) SICK-E上的RoBERTa

图2：使用训练数据对情感分析和NLI的AUTOPROMPT与微调的效果。 X轴表示训练过程中使用的数据点数量。 误差条显示了在10次独立运行中观察到的最大和最小准确性。(自从**E***M*NL*P*版本修订以来)。

我们观察到，尽管在情感分析上微调的表现优于AUTOPROMPT，但在NLI上，AUTOPROMPT可以比微调表现更好。值得注意的是，仅使用10个训练示例，AUTOPROMPT从BERT和RoBERTa中提取出更好的平均性能。 此外，RoBERTa的结果在所有样本大小上更加稳定，而微调可能会导致“失败的运行”（与Dodge等人2020年的研究一致）。 这种在低数据情况下的行为是一个有趣的现象，并且表明当将MLM转换为微调的分类器时，存在一些障碍，而在以掩码语言建模的形式呈现任务时则不会遇到这些障碍。

4 自然语言推理

为了评估预训练语言模型的语义理解能力，我们在自然语言推理上进行了实验

模型	SICK-E 数据集		
	标准的	3路	2路
多数	56.7	33.3	50.0
BERT (微调)	86.7	84.0	95.6
BERT (线性探测)	68.0	49.5	91.9
RoBERTa (线性探测)	72.6	49.4	91.1
BERT (自动生成提示)	62.3	55.4	85.7
RoBERTa (自动生成提示)	65.0	69.3	87.3

表2：SICK-E测试集及其变体的自然语言推理性能（顶部）基准分类器（底部）填空式MLM

（NLI） NLI在阅读理解和常识推理等许多任务中至关重要（Bowman等，2015），它被用作语言理解的常见基准。

设置我们使用SICK数据集（Marelli等，2014年，SICK-E）中的蕴涵任务，该数据集包含约10,000对人工注释的句子，标记为蕴涵、矛盾和中性。标准数据集对中性类别有偏见，占实例的56.7%。

我们还进行了一个无偏差的变体实验，将矛盾与蕴涵的2路分类（*2-way*）和一个无偏差的3路分类变体（*3-way*）进行了实验。 用于AUTOPROMPT的模板在表3中提供。 我们在以下参数上进行搜索： $|\mathcal{V}_{cand}| \in \{10, 50\}$，$|\mathcal{V}_y| \in \{1, 3, 5, 10\}$， $|\boldsymbol{x}_{\text{trig}}| \in [1, 5]$，并根据开发集准确性选择最佳提示。

结果表2显示，AUTOPROMPT在所有实验中明显优于大多数基准。 例如，在2-way SICK-Dataset上，AUTOPROMPT与经过监督微调的BERT相当。我们还测试了线性探测器——在冻结的MLM表示上进行训练的线性分类器，并发现AUTOPROMPT具有相当或更高的准确性，尽管线性探测器容易产生误报。 总的来说，这些结果表明BERT和RoBERTa对自然语言推理具有一定的内在知识。

我们还检查了AUTOPROMPT在低数据情况下（使用与SST-2相同的过程）对无偏3路SICK-E数据的有效性。 图2的结果显示，AUTOPROMPT在低数据设置中表现与微调的BERT相当，并且明显优于微调的RoBERTa。

MLMs在矛盾方面表现出色我们发现标签标记对于矛盾更易解释。

与蕴含或中性相比，我们发现标签标记对于矛盾更易解释（表3中的示例）。 我们调查了这是否会对蕴含和中性类别的模型性能造成影响。 我们测量了3路平衡SICK-E数据集中每个标签的精确度。BERT分别获得了74.9%、54.4%和36.8%的矛盾、蕴含和中性情况的精确度，而RoBERTa分别获得了84.9%、65.1%和57.3%的精确度。 这些结果表明，AUTOPROMPT对于可以使用自然标签标记轻松表达的概念可能更准确。

5个事实检索

一个重要的问题是预训练的MLM是否了解现实世界实体的事实。 LAMA数据集（Petroni等，2019）使用填空测试来评估这一点，该测试由(sub, rel, obj)三元组组成，例如(奥巴马,出生于,夏威夷)，以及带有缺失对象的手动创建的提示，例如“奥巴马出生在[MASK]。”LPAQA（Jiang等，2020）通过挖掘维基百科、改写和众包的方式系统地创建提示来扩展这个想法。 在本节中，我们使用相同的填空式设置，但自动生成提示以更好地评估MLM的事实知识。我们将我们的方法与专门设计用于事实检索任务的LAMA和LPAQA进行比较。

设置我们通过使用模板将(sub,rel,obj)三元组映射到提示中来重新定义事实检索，其中触发令牌是特定于关系 rel的，正确的对象obj是标签令牌。”，其中触发令牌是特定于关系rel的，正确的对象obj是标签令牌。 我们使用来自LAMA（Petroni等，2019）的原始测试集，以下简称为*Orig-inal*。 为了收集AUTOPROMPT的训练数据，我们从T-REx数据集（ElSahar等，2018）中收集了LAMA中的41个关系的最多1000个事实。 对于仍然具有少于1000个样本的关系，我们直接从Wikidata收集额外的事实。我们确保T-REx三元组中没有出现在测试集中，并将数据按80-20的比例分为训练集和开发集。 此外，由于收集的T-REx数据与LAMA测试集的分布略有不同，我们还考虑了单独的评估，将T-REx三元组分为60-20-20的训练/开发/测试集，并在测试集上进行评估。当训练和测试数据来自相同分布时，我们使用此T-REx数据集来衡量我们的提示性能。

任务	提示模板	由AUTOPROMPT发现的提示	标签标记
情感分析	{句子}[T]. . . [T] [P].	毫不动摇的黑暗和绝望 写作学术界的海外 将出现[MASK]。	**pos**: 合作伙伴，非凡，##bla **neg**: 更糟，持续，违宪
NLI	{前提}[P][T]. . . [T]{假设}	两只狗在摔跤和 拥抱[MASK]的混凝土路径 工作场所没有狗 摔跤和拥抱	**con**: 没有人，没有人，也没有 **ent**: ##found，##ways，机构 **neu**: ##ponents, ##lary, ##uated
事实检索	*X*演奏*Y*的音乐 {子 }[T]. . . [T][P].	Hall Overton的壁炉制成的古董 儿子的高音[MASK]。	
关系抽取	*X*是一个职业为*Y*的人 {sent}{sub}[T]. . . [T][P].	莱纳德·伍德（Leonard Wood） （1942年2月4日出生）是一 位前加拿大政治家。 莱纳德·伍德体育馆兄弟自 称另一个[MASK]。	

表3：每个任务的示例提示由AUTOPROMPT生成。 在左边，我们展示了提示模板，它结合了输入、一些触发词 [T]，和一个预测词 [**P**]。 对于分类任务（情感分析和自然语言推理），我们通过对一些自动选择的标签词的模型概率求和来进行预测。 对于事实检索和关系抽取，我们选择模型预测的最可能的标记。

我们使用包含5或7个标记的AUTOPROMPT，并使用T-REx开发集来选择搜索参数。我们防止将专有名词和出现在训练数据中作为黄金对象的标记选择为触发词。 这样做是为了防止AUTOPROMPT通过在提示中嵌入常见答案来“作弊”。 为了评估，我们观察MLM中真实对象在标签词分布中的排名，并使用标准排名指标：平均倒数排名（MRR）、精确度@1（P@1）和精确度@10（P@10）。

结果表4展示了使用不同提示方法的MLMs的性能，我们在表3和附录C中展示了定性示例。使用AUTOPROMPT生成的提示可以更有效地从BERT中提取事实知识：我们将P@1提高了多达12个百分点。 此外，尽管AUTOPROMPT每个关系只使用一个提示，但它仍然比LPAQA的集成方法（平均预测多达30个提示）表现出约4个百分点的优势。 使用7个触发令牌比5个触发令牌获得稍高的分数，尽管差异不大。这表明我们的方法对于触发长度的选择是稳定的，这与我们的情感分析结果一致。 总的来说，这些结果表明AUTOPROMPT可以比过去的提示方法更有效地检索事实，从而证明BERT包含的事实知识比以前估计的要多。

关系细分我们还在附录C的表7中提供了**Petroni**等人(2019)和AUTOPROMPT找到的提示以及它们的相关准确性的详细细分。 手动提示在指定简单的提示时具有竞争力，例如“出生于”这样的提示用于出生地关系。 另一方面，AUTOPROMPT在难以用自然语言提示来指定的关系中表现特别好。 例如，**Petroni**等人(2019)对于球队上的位置关系的提示是“{sub}打球在[MASK]位置”，这不如所需的关系具体。 尽管来自AUTOMPT的提示不符合语法（“*{sub} ediatricstriker ice baseman defensive {obj}*”），但它包含与体育直接相关的标记。

BERT优于RoBERTa 我们最终直接比较了**BERT**和**RoBERTa**。为此，我们对LAMA测试集进行了子采样，使其包含**BERT**和**RoBERTa**的单个令牌的示例（原始-**RoBERTa**）。实际上，**BERT**稍微优于**RoBERTa**，并且我们发现为**RoBERTa**生成的提示往往包含更多无关的词语（请参见附录C，表7）。 例如，**RoBERTa**为PLAYS INSTRUMENT关系生成的提示包含诸如“**Trump**”之类的词语和诸如“,” (),”之类的符号，而对于POSITION PLAYED ON TEAM关系。 令人惊讶的是，**RoBERTa**并没有

[4]原始数据集由BERT的单个令牌示例组成。

提示类型	原始			T-REx		
	MRR	P@10	P@1	MRR	P@10	P@1
LAMA	40.27	59.49	31.10	35.79	54.29	26.38
LPAQA (Top1)	43.57	62.03	34.10	39.86	57.27	31.16
自动生成提示5个标记	53.06	72.17	42.94	54.42	70.80	45.40
自动生成提示7个标记	53.89	73.93	43.34	54.89	72.02	45.57

模型	MRR	P@10	P@1
BERT	55.22	74.01	45.23
RoBERTa	49.90	68.34	40.01

表4：事实检索：在左侧，我们使用来自Petroni等人（2019）的原始LAMA数据集评估BERT的事实检索。对于三个指标（平均倒数排名，平均精确度@10（P@10）和平均精确度@1（P@1）），自动生成提示明显优于过去的提示方法。我们还报告了使用T-REx数据的结果（详见文本）。在右侧，我们使用5个标记的自动生成提示比较BERT和RoBERTa在LAMA数据的子集上的表现。

在未来的工作中，它比BERT表现更好，值得进一步研究。此外，要记住提示是模型知识的下限：相对较低的性能并不意味着模型实际上知道得更少。

6 关系抽取

除了评估MLM是否知道事实外，评估它们是否能够从文本中提取知识也很重要。在本节中，我们使用关系抽取（RE）任务来识别给定句子中的实体之间的关系，这是信息提取中的重要任务。我们以与事实检索类似的方式创建关系抽取提示：对于给定的三元组（主体，关系，客体）和表达此关系的句子，我们构建一个提示为"{sent}{sub}[T]...[T][P]。"其中触发词是特定于关系的，标签词是正确的客体obj（示例见表3）。

设置 我们使用T-Rex数据集进行关系抽取，因为每个T-Rex事实都附带有提及主语和宾语表面形式的上下文句子。我们将AUTOPROMPT与LAMA和LPAQA进行比较（它们的提示在这里仍然有用），以及最近的监督关系抽取模型（Sorokin和Gurevych，2017），该模型也被Petroni等人（2019）使用。为了对监督关系抽取模型进行公平评估，我们修改了标准的关系抽取评估方法。只要模型不为主语和宾语预测不同的关系，即我们忽略"无关系"预测和其他所有关系，我们就给予模型信用。对于模型的命名实体提取器无法识别主语和宾语作为实体的所有句子，我们也将其从评估中删除。有关详细信息，请参见附录B。对于所有系统的评估，如果预测结果是对象的规范版本（例如，"美国"）或

对于给定的三元组中的任何上下文句子，都可以生成表面形式（例如，"美国"）的表达。

结果 表5展示了BERT和RoBERTa的结果。MLMs可以比监督式RE模型更有效地提取关系信息，当使用AUTOPROMPT时，任务性能提升了33%。RoBERTa也优于监督式RE模型，尽管比BERT差（可能是因为我们在第5节中概述的类似原因）。对于BERT和RoBERTa，我们注意到触发词与相应关系相关的词汇（完整列表请参见附录D，表8），例如RoBERTa选择"违反商标的同名制造商"作为关系。

制造商/产品制造商.

扰动句子评估MLM在关系提取设置中取得强大结果的一个可能解释是它们可能已经了解许多关系。因此，它们可以直接预测对象而不是提取它们。为了分离这种影响，我们通过将测试数据中的每个对象替换为其他随机对象并对提示进行相同的更改来合成扰动的关系提取数据集。例如，"Ryo Kase（出生于*1974*年*11*月*9*日，横滨→约克郡）是一位日本演员"，其中Ryo Kase是主语，横滨是原始对象，约克郡是新对象。我们使用扰动版本~~的数据重~~新生成提示。

RE模型在扰动数据上的准确率没有显著变化（表5），然而，MLMs的准确率显著下降。这表明，MLM的准确率的一个显著部分来自背景信息而不是关系抽取。然而，我们为BERT生成的提示优于它们的LAMA和LPAQA对应物，这进一步证明了AUTOPROMPT生成更好的探测器。

模型	原始	扰动
监督式关系抽取 LSTM	57.95	58.81
BERT (LAMA)	69.06	28.02
BERT (LPAQA)	76.55	30.79
BERT (AUTOPROMPT)	90.73	56.43
RoBERTa (AUTOPROMPT)	60.33	28.95

表5：关系抽取：我们使用提示来测试预训练的MLMs在关系抽取上的性能。 与2017年的最先进LSTM模型相比，MLMs具有更高的平均准确率（P@1），尤其是在使用AUTOPROMPT的提示时。 我们还测试了模型在包含错误事实的句子上的表现。 MLMs在这些句子上的准确率显著下降，表明它们的高性能源于它们的事实知识。

7 讨论

提示作为微调的替代方法
提示语言模型的目标是探索模型从预训练中获得的知识。 然而，与解决实际任务的微调相比，提示具有一些实际优势。 首先，如第3节所示，使用AUTOPROMPT生成的提示在低数据情况下可以实现更高的准确性。

此外，与微调相比，提示在解决许多不同任务时具有优势（例如，OpenAI GPT-3 API的许多用户（Brown等人，2020））。 特别是，微调需要为每个单独的任务存储大型语言模型检查点，并且因为需要同时部署许多不同的模型，它会大大增加系统成本和复杂性。 提示解决了这两个问题。 每个单独的任务只存储提示，而所有任务都使用相同的预训练模型。

提示的限制有一些现象很难通过提示从预训练的语言模型中引出。 在我们对数据集（如QQP（Iyer等，2017）和RTE（Dagan等，2005））进行初步评估时，手动生成的提示和使用自动生成的提示并没有比随机猜测表现得更好。 然而，我们不能通过这些结果得出BERT不了解释义或蕴含的结论。 一般来说，不同的探测方法适用于不同的任务和现象：自动生成提示使得基于提示的探测更具普适性，但它仍然只是可解释性研究者工具箱中的一种工具。

自动生成提示的限制自动生成提示的一个缺点是它需要有标记的训练数据。 尽管其他探测技术（如线性探测分类器）也需要这个，但手动提示依赖于领域/语言的见解而不是标记数据。 与人工设计的提示相比，自动生成的提示缺乏可解释性，这与其他探测技术（如线性探测分类器）类似。 自动生成提示的另一个限制是在训练数据高度不平衡时可能会遇到困难。 例如，在第4节和第5节中，我们展示了提示通常只会增加多数标签的可能性。 重新平衡训练数据可以帮助缓解这个问题。 最后，由于在大量离散短语空间上进行贪婪搜索，自动生成提示有时会很脆弱；我们将更有效的构建技术留给未来的研究方向。

8 结论

在本文中，我们介绍了一种名为自动生成提示（AUTO PROMPT）的方法，该方法通过自动生成的提示从预训练的语言模型中提取知识，适用于各种任务。 我们证明了这些提示在需要较少人力的情况下优于手动提示。 此外，情感分析和文本蕴涵的结果表明，在某些数据稀缺的情况下，提示语言模型可能比对其进行微调更有效。 尽管本文仅关注掩码语言模型，但我们的方法可以轻松扩展到标准语言模型，因此可能对构建类似GPT-3（Brown等，2020）的模型的输入很有用。 本文中重现结果所需的源代码和数据集可在http://ucinlp.github.io/autoprompt获取。

致谢

我们要感谢LAMA和LPAQA团队对我们问题的回答。 我们还要感谢UCI NLP的成员Matt Gardner、Sebastian Riedel和Antoine Bosselut对我们的宝贵反馈。 本材料基于DARPA MCS计划的工作，根据合同编号N660011924033与美国海军研究办公室签订。

参考文献

Zied Bouraoui, Jose Camacho-Collados, and StevenSchockaert. 2019. 从BERT中引导关系知识。在 *AAAI*中。

Samuel R Bowman, Gabor Angeli, Christopher Potts,and Christopher D Manning. 2015. 用于学习自然语言推理的大型注释语料库。在 *EMNLP*中。

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, ArvindNeelakantan, Pranav Shyam, Girish Sastry, AmandaAskell, et al. 2020. 语言模型是少样本学习器。ar*Xiv*预印本 ar*Xiv:*200*5.14165*。

Alexis Conneau, Germ´an Kruszewski, Guillaume Lam-ple, Lo¨ıc Barrault, and Marco Baroni. 2018. 你可以塞进一个向量的东西：探究句子嵌入的语言属性。在 *ACL*中。

Ido Dagan，Oren Glickman和Bernardo Magnini。2005年。 PASCAL识别文本蕴含挑战。 在机器学习挑战工-作坊。

Jacob Devlin，Ming-Wei Chang，Kenton Lee和Kristina Toutanova。2019年。 BERT：深度双向变压器的预训练语言理解。在 *NAACL*。

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi，Hannaneh Hajishirzi和Noah Smith。2020年。 微调预训练语言模型：权重初始化，数据顺序和早停-止。*arXiv*预印本*arXiv:2002.06305*。

Hady ElSahar，Pavlos Vougiouklis，Arslen Remaci，Christophe Gravier，Jonathon S. Hare，Fr´ed´eriqueLaforest和Elena Simperl。2018年。 T-REx：自然语言与知识库三元组的大规模对齐。在 *LREC*。

John Hewitt和Percy Liang。2019年。 设计和解释带有控制任务的探针。在 *EMNLP*中。

Shankar Iyer，Nikhil Dandekar和Kornel Csernai。2017年。 Quora首个数据集发布：问题对。

Sarthak Jain和Byron C Wallace。2019年。 注意力不是解释。在 *NAACL*中。

Zhengbao Jiang，Frank F Xu，Jun Araki和GrahamNeubig。2020年。 我们如何知道语言模型知道什么？在 *TACL*中。

Sunjae Kwon，Cheongwoong Kang，Jiyeon Han和Jaesik Choi。2019年。 为什么遮蔽的神经语言模型仍然需要常识知识？ ar*Xiv*预印本ar*Xiv:*1911.0*3024*。

Patrick Lewis，Ludovic Denoyer和Sebastian Riedel。2019年。 无监督问答通过填空翻译。在 *ACL*中。

Nelson F Liu, Matt Gardner, Yonatan Belinkov,Matthew Peters, and Noah A Smith. 2019. 语言学知识和上下文表示的可迁移性。在 *NAACL*.

Ilya Loshchilov and Frank Hutter. 2018. 解耦的权重衰减正则化。 在国际学习表示会议.

Marco Marelli, Stefano Menini, Marco Baroni, LuisaBentivogli, Raffaella Bernardi, Roberto Zamparelli,et al. 2014. 用于组合分布语义模型评估的SICK疗法。在 *LREC*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. 关于BERT微调的稳定性：误解、解释和强基线。

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and LukeZettlemoyer. 2018. 深度上下文化的词表示。在 *NAACL*.

Fabio Petroni，Tim Rockt¨aschel，Patrick Lewis，AntonBakhtin，Yuxiang Wu，Alexander H Miller和Se-bastian Riedel。2019年。 语言模型作为知识库？在 *EMNLP*中。

Alec Radford，Jeffrey Wu，Rewon Child，David Luan，Dario Amodei和Ilya Sutskever。2019年。 语言模型是无监督的多任务学习者。技术报告。

Timo Schick和Hinrich Sch¨utze。2020年。 利用填空问题进行少样本文本分类和自然语言推理。 ar*Xiv*预印本*arXiv:2001.07676*。

Vered Shwartz，Peter West，Ronan Le Bras，Chandra Bhagavatula和Yejin Choi。2020年。 无监督的常识问答与自我对话。ar*Xiv*预印本ar*Xiv:*200*4.05483*。

Richard Socher, Alex Perelygin, Jean Wu, JasonChuang, Christopher D Manning, Andrew Ng, andChristopher Potts. 2013. 递归深度模型用于情感树的语义组合性。在 *EMNLP*中。

Daniil Sorokin和Iryna Gurevych。2017年。 上下文感知表示用于知识库关系提取。在 *EMNLP*中。

Trieu H Trinh和Quoc V Le。2018年。 一种简单的常识推理方法。ar*Xiv*预印本 *arXiv:1806.02847*。

Elena Voita和Ivan Titov。2020年。 最小描述长度的信息论探测。在 *EMNLP*中。

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner,and Sameer Singh. 2019. 用于攻击和分析NLP的通用对抗触发器。在 *EMNLP*中。

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill，Omer Levy和Samuel R Bowman。2019年。GLUE：自然语言理解的多任务基准和分析平台。在 *ICLR*中。

Sarah Wiegreffe和Yuval Pinter。2019年。 注意力不是解释。在 *EMNLP*中。

Thomas Wolf，Lysandre Debut，Victor Sanh，JulienChaumond，Clement Delangue，Anthony Moi，Pier-ric Cistac，Tim Rault，R´emi Louf，Morgan Fun-towicz，Joe Davison，Sam Shleifer，Patrick vonPlaten，Clara Ma，Yacine Jernite，Julien Plu，Can-wen Xu，Teven Le Scao，Sylvain Gugger，MariamaDrame，Quentin Lhoest和Alexander M. Rush。2019年。 HuggingFace的转换器：最先进的自然语言处理。 arXiv预印本*arXiv:1910.03771*。

超参数对情感分析的影响

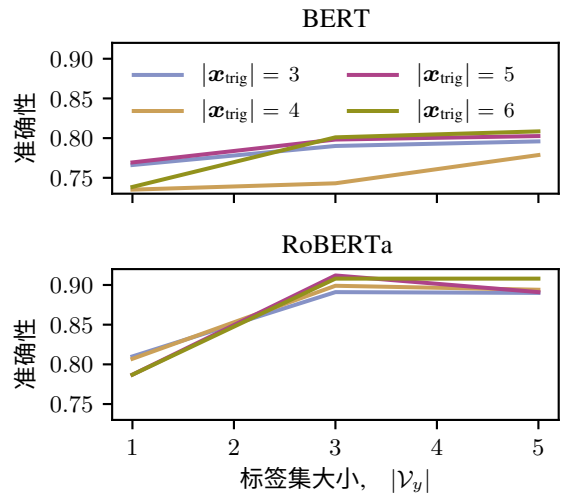


图3: 标签集大小和触发器集大小对情感分析的影响。 候选替换数量固定为 $|\mathcal{V}_{\text{cand}}|= 100$。 增加标签集大小可以提高性能，而改变触发器长度的影响不大。

为了衡量自动生成提示的搜索超参数的影响，我们在图3中绘制了验证准确率与标签集大小 $|\mathcal{V}_y|$和触发器令牌数量 $|\boldsymbol{x}_{\text{trig}}|$的关系。 我们将候选数量固定为 $|\mathcal{V}_{\text{cand}}|= 100$。 当 $|\mathcal{V}_{\text{cand}}|= 10$时，我们观察到类似的趋势。

改变触发令牌的数量通常没有太大影响。 另一方面，当标签集大小从1增加到3时，准确率显著提高（BERT约+5%，RoBERTa约+10%）。在分析标签集之后，我们发现我们的方法通常产生直观的结果-对于RoBERTa，“了不起”的和“仁慈”的与积极情感相关，而“更糟”的和“无能”的与消极情感相关。

B关系抽取细节

根据Petroni等人（2019）的方法，我们使用Sorokin和Gurevych（2017）的预训练RE模型作为我们的基线。 为了编码句子，该模型使用基于LSTM的关系编码器和注意机制的组合。 为了进行预测，模型构建了一个知识图谱，其边缘是提取的关系三元组。 标准的RE评估衡量模型在句子级别上预测实体对的关系类型的准确性。

由于我们的目标是提取关系三元组的对象，而不是关系本身，因此我们对标准的关系抽取评估进行了调整。 我们将关系抽取模型输入测试事实的句子，并查询结果图中包含给定主体和关系的所有边。 然后我们选择置信度最高的三元组，并将其对象与黄金对象进行比较。 我们对每个事实都进行这样的操作，并对所有关系取平均以得到总体精度。 关系抽取模型没有经过训练来预测原始的两个T-REx关系。 为了公平比较，我们在评估中排除了这两个关系。

C 附加事实检索结果

关系手动提示（LAMA）		#train	LAMA	LPAQA	AUTOPROMPT
P1001	[X]是[Y]中的一个法律术语	1000	70.47	72.75	82.45
P101	[X]在[Y]领域工作	864	9.91	5.32	12.79
P103	[X]的母语是[Y]	1000	72.16	72.16	82.09
P106	[X]是一位[Y]职业者	1000	0.63	0.0	14.72
P108	[X]为[Y]工作	376	6.79	5.74	8.62
P127	[X]归[Y]所有	548	34.79	32.46	35.95
P1303	[X]演奏[Y]	1000	7.59	18.02	15.38
P131	[X]位于[Y]	1000	23.27	22.81	37.46
P136	[X]演奏[Y]音乐	1000	0.75	16.76	55.42
P1376	[X]是[Y]的首都	310	73.93	59.83	40.17
P138	[X]以[Y]命名	856	61.55	59.69	66.05
P140	[X]与[Y]宗教有关	445	0.63	59.83	75.26
P1412	[X]曾用[Y]进行交流	1000	65.02	64.71	71.21
P159	[X]的总部位于[Y]	1000	32.37	35.57	35.47
P17	[X]位于[Y]	1000	31.29	35.48	52.15
P176	[X]由[Y]生产	1000	85.64	81.67	87.78
P178	[X]由[Y]开发	560	62.84	59.12	66.72
P19	[X]出生在[Y]	1000	21.08	20.87	19.92
P190	[X]和[Y]是双子城市	895	2.41	1.91	2.31
P20	[X]在[Y]去世	1000	27.91	27.91	31.16
P264	[X]由音乐厂牌[Y]代表	1000	9.56	10.26	43.82
P27	[X]是[Y]公民	1000	0.0	41.51	46.69
P276	[X]位于[Y]	1000	41.5	41.5	44.11
P279	[X]是[Y]的子类	1000	30.74	14.75	54.93
P30	[X]位于[Y]	1000	25.44	18.56	70.36
P31	[X]是[Y]	1000	36.66	36.66	51.95
P36	[X]的首都是[Y]	1000	62.16	62.16	60.6
P361	[X]是[Y]的一部分	1000	23.61	31.44	17.7
P364	[X]的原始语言是[Y]	1000	44.51	43.93	48.48
P37	[X]的官方语言是[Y]	311	54.55	56.83	62.63
P39	[X]担任[Y]的职位	1000	7.96	16.14	30.72
P407	[X]是用[Y]编写的	1000	59.18	65.22	68.42
P413	[X]在[Y]位置上比赛	1000	0.53	23.74	41.7
P449	[X]最初在[Y]上播出	1000	20.89	9.08	34.39
P463	[X]是[Y]的成员	679	67.11	57.33	54.22
P47	[X]与[Y]接壤	1000	13.67	13.34	19.52
P495	[X]创建于[Y]	1000	16.5	32.23	36.63
P527	[X]由[Y]组成	1000	11.07	10.55	25.61
P530	[X]与[Y]保持外交关系927		2.81	3.92	3.11
P740	[X]成立于[Y]	1000	7.59	13.68	13.89
P937	[X]曾在[Y]工作	1000	29.77	39.1	38.36

表6：Petroni等人（2019）的原始数据集上的事实检索的所有关系细分。 我们比较了LAMA、LPAQA和我们的方法生成的五个提示标记的P@1。

关系方法		提示	P@1
P101	手动	[X]在[Y]领域工作	11.52
	AUTOPROMPT BERT	[X]概率最早的名声总计研究[Y]	15.01
	AUTOPROMPT RoBERTa	[X] 1830年的论文应用mathsucci [Y]	0.17
P103	手动	[X]的母语是[Y]	74.54
	AUTOPROMPT BERT	[X]PA communerug说得很好[Y]	84.87
	AUTOPROMPT RoBERTa	[X]neau可选fluent!?'traditional [Y]	81.61
P106	手动	[X]是一位[Y]职业者	0.73
	AUTOPROMPT BERT	[X]支持者研究的政治家音乐家转变[Y]	15.83
	AUTOPROMPT RoBERTa	[X] (),天文学家商人·前[Y]	19.24
P127	手动	[X]归[Y]所有	36.67
	AUTOPROMPT BERT	[X]是后翅膀主线架构内[Y]	47.01
	AUTOPROMPT RoBERTa	[X] picThom不愿意正式统治[Y]	39.58
P1303	手动	[X]演奏[Y]	18.91
	AUTOPROMPT BERT	[X]演奏鼓协奏曲电动[Y]	42.69
	AUTOPROMPT RoBERTa	[X]特朗普学会了独奏基夫古典[Y]	44.44
P136	手动	[X]演奏[Y]音乐	0.7
	AUTOPROMPT BERT	[X]该死的类型管弦乐团虚构酸[Y]	59.95
	AUTOPROMPT RoBERTa	[X]融合战后人质剧情萨克斯[Y]	52.97
P1376	手动	[X]是[Y]的首都	81.11
	AUTOPROMPT BERT	[X]自豪地称之为传统领土[Y]	63.33
	AUTOPROMPT RoBERTa	[X]石灰岩沉积在自治市区内[Y]	28.33
P178	手动	[X]由[Y]开发	62.76
	AUTOPROMPT BERT	[X]是由[Y]品牌打造的记忆游乐场	64.45
	AUTOPROMPT RoBERTa	[X]1987年软盘模拟器用户起诉[Y]	69.56
P20	手动	[X]在[Y]去世	32.07
	AUTOPROMPT BERT	[X]在[Y]中的摄影工作室重组类型	33.53
	AUTOPROMPT RoBERTa	[X].. 神秘的二十世纪现在在[Y]附近	31.33
P27	手动	[X]是[Y]公民	0.0
	AUTOPROMPT BERT	[X]立方米的羽毛球器材在国际上代表[Y]46.13	
	AUTOPROMPT RoBERTa	[X]的fic组织森林法规西北[Y]	42.07
P276	手动	[X]位于[Y]	43.73
	AUTOPROMPT BERT	[X]由千克组成，以[Y]为中心的社区	44.64
	AUTOPROMPT RoBERTa	[X]的机动构建举报者山丘附近[Y]	37.47
P279	手动	[X]是[Y]的子类	31.04
	AUTOPROMPT BERT	[X]被称为[Y]的涂层	55.65
	AUTOPROMPT RoBERTa	[X]，以前的祈祷者不稳定的[Y]	52.55
P37	手动	[X]的官方语言是[Y]	56.89
	AUTOPROMPT BERT	[X]新恩方言与官方语言完全相似[Y]	54.44
	AUTOPROMPT RoBERTa	[X]恩族后裔主要说[Y]	53.67
P407	手动	[X]是用[Y]编写的	60.21
	AUTOPROMPT BERT	[X]每个方言都玩[Y]	69.31
	AUTOPROMPT RoBERTa	[X] scaven pronunciation.*Wikipedia speaks [Y]	72.0
P413	手动	[X]在[Y]位置上比赛	0.53
	AUTOPROMPT BERT	[X] played colors skier ↔ defensive [Y]	41.71
	AUTOPROMPT RoBERTa	[X]," (), ex-,Liverpool [Y]	23.21

表7：手动提示的示例（第一行，显示BERT的P@1）和通过AUTOPROMPT生成的提示用于事实检索。

D 附加关系抽取结果

关系	模型	上下文和提示	预测
P103（母语）	BERT	亚历山德拉·拉米（出生于1971年10月14日）是一位法国女演员。 亚历山德拉·拉米说空军滴水百分之[MASK]。	法语
P36（首都）	RoBERTa	柯克出生于俄亥俄州克林顿县，他在俄亥俄州威尔明顿市服役 。克林顿县以动物园而闻名，影响了[MASK]。	威尔明顿
P530（外交关系）	BERT	黑海形成了一个东西向椭圆形的洼地，位于保加利亚、格鲁吉亚、罗马尼亚、俄罗斯、土耳其和乌克兰之间。乌克兰实际上有一些移民进入了[MASK]。	俄罗斯
P106（职业）	RoBERTa	斯宾塞·特里特·克拉克（出生于1987年9月24日）是一位美国演员，曾出演过多部电影，包括《角斗士》、《神秘河》和《不可思议的》。斯宾塞·特里特·克拉克以其英俊的外貌而闻名，[MASK]。	绿巨人
P276（位置）	BERT	《不朽的对局》是由阿道夫·安德森和莱昂内尔·基瑟里茨基于1851年6月21日在伦敦期间的第一届国际锦标赛休息期间进行的一场国际象棋对局。 《不朽的对局》位于[MASK] 。	首尔
P176（制造商）	BERT	罗伯塔本田Civic del Sol是一款由本田在1990年代制造的2座前置前驱敞篷车。 本田Civic del Sol违反了同名制造商的商标[MASK]。	丰田
P279（子类）	BERT	Mizeria是一道波兰沙拉三明治，由薄切或刨碎的黄瓜制成，通常加入酸奶油，尽管有些情况下也会加入油。Mizeria被称为直接的高度[MASK]。	食物
P463（成员）	BERT	罗伯塔Rush是一个加拿大摇滚乐队，由Geddy Lee（贝斯，主唱，键盘手），Alex Lifeson（吉他手）和Neil Peart（鼓手，打击乐器，词曲创作）组成。 Alex Lifeson与国际上的缩写[MASK]有关 。	吻

表8：使用自动生成的提示生成的关系抽取示例。 下划线表示黄金对象。 表的下半部分显示了我们增强评估的示例，其中原始对象（用划掉的词表示）被新对象替换。