PIE-QG：面向小型语料库的无监督问题生成的释义信息抽取

DineshNagumothu、Bahadorreza Ofoghi、GuangyanHuang、Peter W. Eklund迪肯大学信息技术学院，澳大利亚维多利亚州伯伍德 Burwood Hwy 221号，邮编3125

{dnagumot,b.ofoghi,guangyan.huang,peter.eklund}@deakin.edu.au

arXiv:2301.01064v1 [cs.CL] 3 Jan 2023

摘要

监督式问答系统（问答系统）依赖由人工标注的领域专属数据进行训练。无监督问答系统会自行生成问题—答案训练对，通常借助辅助知识源来实现。我们的方法（称为 PIE-QG）使用开放信息抽取（OpenIE），从释义后的段落中生成合成训练问题，并将问答对用作基于 BERT 的最先进问答系统所用语言模型的训练数据。从每个段落中提取形式为 <主语、谓语、宾语> 的三元组，并以主语（或宾语）和谓语构造问题，同时将宾语（或主语）视为答案。在五个抽取式问答数据集上的实验表明，我们的方法性能可与现有最先进的问答系统相媲美，而且所需训练文档数量少一个数量级，也完全无需借助外部参考数据源。

1 引言

问答系统（QA 系统）能够针对以自然语言提出的输入问题提供答案。可以使用机器阅读理解（MRC）从非结构化文本中回答问题。给定一段文本、若干句子或一个段落，以及提出的问题，问答系统会生成最合适的答案。抽取式问答系统（EQA 系统）是问答系统的一个子集，涉及一项 MRC 任务，其预测答案是从段落中抽取的一段文字。借助预训练语言模型（Radford 等人，2018年），EQA 系统取得了出色的成果，甚至超越了人类表现。预训练语言模型，例如 BERT（Devlin 等人，2019）和 GPT（Radford 等人），可以通过微调来执行问答等下游任务。然而，要用大量数据训练这些模型使其适用于特定领域，因此这项任务十分耗费人力，需要投入大量精力来收集合适的领域特定训练数据。

EQA系统数据集中的每个训练实例都需要一个问题、一段段落和一个答案。可以借助先进的信息检索工具收集与领域相关的文档，再将文档拆分为若干相关句子或一个段落，从而形成段落。然而，如何根据给定段落生成问题和答案对，为问答系统提供训练集，被认为是最具挑战性的难题，这种方法称为无监督问答（Cui 等人，2004）。

现有的无监督问答系统技术，如（Lewis 等人，2019）和（Lyu 等人，2021），使用域外数据集生成问题；也就是说，除了目标语料库和预训练通用模型所能提供的内容外，还需要额外的训练来源。另一方面，基于规则的问答系统方法仅限于从语料库本身生成问答对，因此存在生成与段落词汇重叠度过高的问题，从而可能迫使模型学习词语匹配模式。（Fabbri 等人，2020）和（Li 等人，2020）采用基于信息检索的方法，例如弹性搜索和引文导航，从目标数据集中未出现的段落中生成问题。然而，这些方法可能无法生成足够的训练问题，尤其是在语料库规模较小且没有引文或文档间链接结构时。

本文聚焦于解决 EQA 系统的局限性，提出一种新颖的无监督问题生成方法——用于问题生成的释义信息抽取（PIE-QG）。该方法通过从给定语料库中抽取 <主语、关系、宾语> 三元组，生成合成训练数据。我们利用原始段落生成问题—

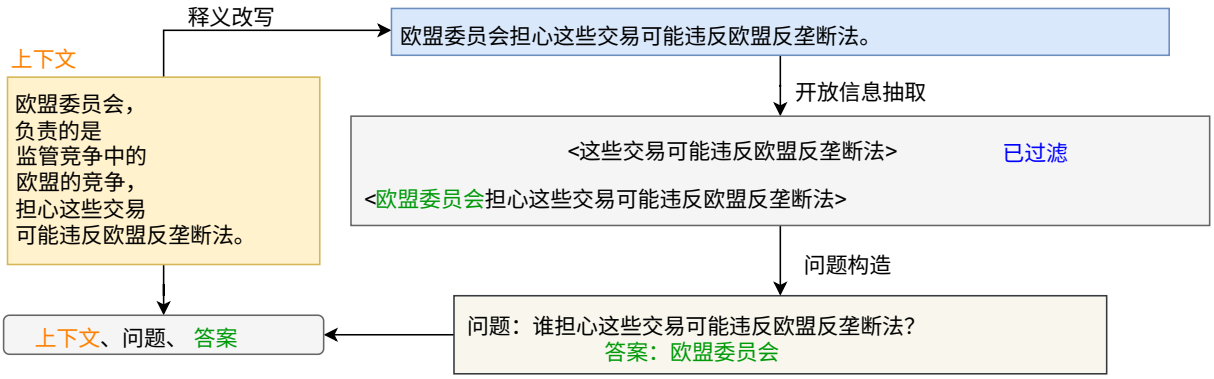

图 1：通过释义改写，随后使用 OpenIE 进行信息抽取，从上下文（左）生成问题。注：绿色文本表示选定的答案。

答案训练对：通过生成原始段落的释义版本，避免段落与问答之间出现词汇重叠。我们采用开放信息抽取（Kolluru 等人，2020年）从释义改写后的段落中的每个句子抽取 <主语、关系、宾语>三元组。这些三元组语义丰富，代表原始事实；因此，根据三元组生成问题—答案对，可以得到格式良好且有效的训练数据。此外，段落中的许多句子都有助于生成有意义的抽取结果，从而能够基于同一段落以不同方式提出问题。图1展示了我们提出的问题生成流程示例（称为 PIE-QG，即释义改写、信息抽取与问题生成）。本文的贡献如下：

1. 我们介绍 PIE-QG 方法：利用原始语料库中的释义段落生成问答对，无需依赖外部参考数据源，例如基于检索或跨文档链接导航的方法。对段落进行释义可以减轻段落与问题之间词汇重叠的影响。

2. 我们采用开放信息抽取从释义段落中提取事实，从而根据单个释义段落生成多个问题，增加从语料库中提取的问答对数量。

我们在四个抽取式问答数据集上进行了实验，结果表明，所提出的 PIE-QG 方法在完全匹配（精确匹配）和 F1 分数方面达到了相当的性能，同时所需的段落数量显著更少。

本文其余部分安排如下：第2节介绍相关工作。第3节描述所提出的 PIE-QG 方法。第4 节讨论实验设置。第5节评估我们方法的性能。第6 节介绍所提方法的局限性，第7 节给出总结性评述。

2 相关工作

预训练语言模型，例如 BERT（De-vlin etal.，2019），可针对下游任务进行微调，例如抽取式问答系统（EQA 系统）。完整的自然语言（NL）段落（可能由若干句子或一段自然语言文本组成）被视为模型寻找答案片段的上下文。输入问题和上下文被表示为一个序列，输入预训练模型，再通过计算答案片段首尾词元的概率来预测答案。预训练语言模型，如 BERT（De-vlin etal.，2019）、T5 Transformer（Raffelet al.，2020年）和 XLNet（Yang et al.，2019），在 EQA 系统中表现出色，但代价是依赖大量由人工标注的监督数据集。斯坦福问答数据集（SQuAD）（Rajpurkaret al.，2016年）是 EQA 系统广泛使用的数据集。

Lewis 等人(2019), Fabbri 等人(2020年), Li 等人(2020年),以及 Lyu 等人 (2021)从维基百科中随机抽取段落，并将命名实体或名词短语识别为答案，因为这些内容往往有助于问答。随后，根据段落和选定的答案短语，用自然语言生成问题。

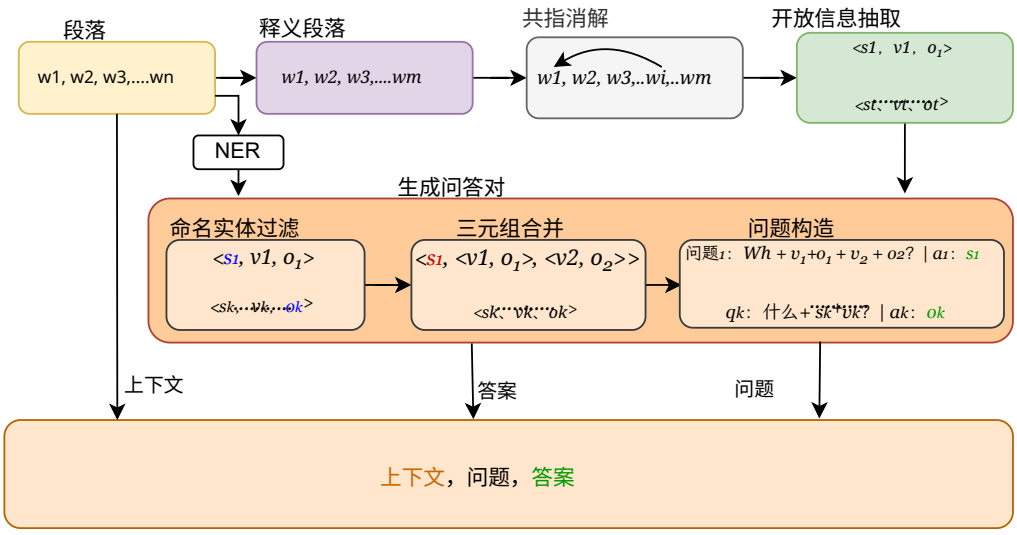

图2：使用释义改写和开放信息抽取进行问题生成的 PIE-QG 总体流程。注：蓝色表示命名实体，红色表示具有共同主语的合并三元组，绿色表示选定的答案。

无监督阅读理解问答通过完形翻译方法实现（Lewis 等人，2019）：从给定的目标语料库中构造段落、问答三元组。段落中的答案会被屏蔽，从而形成“填空式”问题，即所谓的完形填空问题。作者使用神经机器翻译（NMT）模型翻译自然语言问题，该模型通过包含完形填空问题和自然问题对的不同语料库进行训练。

直接根据段落生成的问题只能通过匹配段落中的文本来回答简单的完形填空问题；对于措辞不同的问题，这种方法无法给出正确答案。为了扩展用于训练 EQA 系统的问题，Fabbri 等人（2020年）使用取自不同段落的相似句子生成问题。实际段落被视为查询，并使用弹性搜索检索句子。最相似的句子必须包含答案但不包含原始查询段落，并且相似度低于 95%，以避免出现抄袭句子；该句子随后用于构成问题—答案对。这些句子中的答案会被遮盖，并按照“Wh+B+A？”规则生成问题，其中“wh”（what、when 或 who 三者之一）根据答案实体类型进行选择（“B”是答案遮盖部分之后的句子片段，“A”是答案遮盖部分之前的片段）。

Li 等人 (2020年) 使用引文生成段落的摘要。被引用的段落被视为上下文，而引文所在的句子则用于问题生成，以避免词汇重叠。问题生成过程包括用完形填空掩码遮蔽答案，该掩码仅指出答案实体的类型。句子的依存树经过调整，使完形填空掩码移至句首。随后，根据答案实体的类型，用恰当的“wh”疑问词替换完形填空掩码，从而生成问题。

Lyu 等人（2021年）通过基于文本摘要创建问题生成模型来实现无监督问答。该模型利用从摘要中提取的依存树和语义角色标注来生成问题。随后训练一个神经编码器—解码器模型，将文章转换为受摘要启发的问题。训练好的模型应用于实际段落以生成问题。然而，由于该方法依赖文本摘要数据集，我们将其视为迁移学习任务，而非无监督问题生成。我们的方法与Fabbri 等人（2020年）和Li 等人（2020年）的方法进行了比较，避免了基于句子和引文的检索，并尽可能降低了生成问答对所需的大型语料库规模。

3 释义信息抽取与问题生成

为克服对包含大量段落的外部参考数据源的依赖，我们采用了

开放信息抽取和释义段落来进行无监督合成问题生成。首先将实际段落改写为释义形式，再从释义段落中抽取 <主语、谓语、宾语>三元组。结合某些启发式规则，这些三元组构成问答对，随后与原始段落一起作为上下文，用于微调问答模型。

我们提出的 EQA 问题生成流程如图2所示。该流程的具体步骤如下。

(i) 释义改写：直接根据段落生成的问答对会导致问答系统的性能较差，因为由此训练出的模型几乎不具备泛化能力（Fabbri 等人，2020年）。因此，我们采用释义改写，在不改变段落实际含义的前提下对其进行改写。其背后的思路是，根据语义相似、但用词与原段落不同的段落来生成问题。已有研究表明，改写问答对本身会造成语义漂移（Pan 等人，2021）。相比之下，我们的方法改写的是段落，而非问答对。这提升了模型的性能。第5节将讨论释义改写的影响。

(ii) 共指消解：由于我们的目标是利用段落中的每个句子来生成问题，因此有些句子会因代词的存在而无法有效使用（Ma 等人，2021）。我们通过实施共指消解来解决这一问题，即将释义段落中的代词替换为其所指名词的专有名称。

(iii) 信息抽取：对释义段落应用开放信息抽取，从自然语言文本中以论元和关系的形式生成抽取结果（Mausam，2016年）。给定段落中的一个句子 w_i，$\{w_1, w_2, w_3, ..w_N\}$，开放信息抽取会生成抽取结果 $\{T_1, T_2, T_3, ...T_M\}$，其中每个抽取结果的形式为 <主语、谓语、宾语>，即三元组。研究证明，开放信息抽取是下游任务（如复杂问答）的有效解决方案（Khot 等人，2017）。

(v) 问题构造：从段落中生成的开放信息抽取结果用于构造问题，作为问答系统微调的合成训练集。

（***vi***）命名实体过滤：由于从段落中提取的三元组类型各不相同

算法 1： PIE-QG：从段落生成问题。

输入： 给定语料库中的一个段落 P
输出： 问答对列表

P' = 改写(P)
CP = 共指_消解(P')
T = 开放_信息抽取(CP)
命名_实体 = 命名实体识别(CP)
T_{ne} = 命名实体_过滤器(T, 命名_实体)
T_{IF} =相同三元组_过滤器(T_{ne})
T_M = 合并_三元组(T_{IF})
T_{IF} = 移除_已合并的_三元组(T_{IF}, T_M)
$QA_Pairs \leftarrow newlist$
对于 t_n 中的 T_M **执行**
  A = 选择_答案(t_n)
  Q = 疑问词
  对于 <主语、关系、宾语> 中的 t_n **执行**
    $Q = Q$ + 关系 + 宾语/主语
    $QA_Pairs \leftarrow append(\langle Q, A, P\rangle)$
  end
end
针对 <主语、关系、宾语> 在 T_{IF} **执行**
  Q = 疑问词 + 关系 + 宾语?
  A = 主语
  问答_对 $\leftarrow$ 追加($\langle Q, A, P\rangle$)
  Q = Wh + 主语 + 关系?
  A = 宾语
  QA_对 $\leftarrow$ append($\langle Q, A, P\rangle$)
end
返回 QA_Pairs

我们会选择包含答案中的命名实体的三元组。换句话说，只有当主语（或宾语）是命名实体时，才会将其选作答案。

(vii) 消除重复三元组：开放信息抽取的一个缺点是存在重复或语义冗余三元组。从相似或重复三元组中分别生成问题会导致 EQA 系统模型的性能下降，因此会对冗余三元组进行排序，并从排序结果中选取最长的三元组，作为最终问题生成的唯一来源。

(viii) 三元组合并：使用上述方法从三元组生成的问题简单易答。为了进行鲁棒模型训练，我们将主语或宾语相同的三元组分组，从多个三元组生成更复杂的问题。例如，如果有两个形式为 $\{\langle s_1, r_1, o_1\rangle, \langle s_2, r_2, o_2\rangle\}$ 的三元组

和 $s_1 = s_2$，我们会构造一个问答对，以“Wh + r_1 + o_1, r_2 + o_2?”作为问题，以 s_1 （或s_2）作为答案。

从释义段落中提取的每个三元组都可以生成两个问题，答案分别为主语或宾语。选择主语作为答案时，问题的形式为“Wh +关系 + 宾语？”。反之，选择宾语作为答案时，生成的问题形式为“Wh + 主语 +关系？”。“Wh”是这些句式中的疑问词，具体形式根据答案实体类型，从前文所述的列表中选取。

4 实验平台

数据集 我们使用现有的 EQA 数据集，通过精确匹配（EM）和 F1 分数评估问题生成方法的性能。具体包括 SQuAD v1.1（Rajpurkar 等人，2016年）开发集，以及 NewsQA（Trischler 等人，2016年）、BioASQ（Tsatsaronis 等人，2015年）和 DuoRC（Saha 等人，2018年）测试集。SQuAD 1.1 版本取自官方网站[1]；而 NewsQA、BioASQ 和 DuoRC 使用的是 Fisch 等人（2019）发布的测试集版本。较新的 SQuAD v2.0（Rajpurkar 等人，2018年）不适用于我们的实验，因为合成训练集中不包含无法回答的问题。

问题生成 我们从（Li 等人，2020年）抽取的维基百科数据集中选取了相对较小的子集，其中包含 30,000 个段落，用于问题生成和模型训练。所提出的问题生成技术的伪代码见算法 1。

这一过程中生成的部分问题可能存在语法错误。我们希望模型在推理过程中接收的问题使用自然语言且语法正确，因此尝试在流程中加入语法纠错模块，以生成句法准确的问题；但后来由于其影响（见第5节），我们移除了该模块。

使用基于 PEGASUS Transformer 的预训练模型，将从维基百科获取的段落转换为释义段落[2] （Zhang 等人，2020年）。释义段落中的代词会被替换为它们所指代的名词。为此，我们使用 neuralcoref[3]，这是 spaCy 对基于强化学习的预训练共指消解方法的实现（Clark 和 Manning，2016年）。我们使用 OpenIE6 从 <主语、谓语、宾语> 三元组中抽取信息，这些三元组来自将代词替换后的释义段落。OpenIE6 使用迭代网格标注，并基于 BERT。我们使用基于 spaCy 的命名实体识别（NER）模块（Honnibal 等人，2020年）从段落中生成命名实体列表。命名实体识别（NER）对于筛选三元组以及确定答案实体类型、从而恰当选择“疑问词”尤其有帮助。根据Fabbri 等人（2020年），针对特定命名实体选择“疑问词”的最简单方法是采用这一做法。使用基于 RoBERTa 的语法纠正模块“GECToR”（Liu 等人，2019）对这一流程生成的问题进行语法纠正（Omelianchuk 等人，2020年）。所有模型均直接采用上述来源提供的版本，不进行特定领域的微调。

问答微调 我们使用 Devlin 等人 (2019) 提出的预训练 BERT 模型作为基线，并使用生成的训练数据对模型进行微调，以完成下游问答系统任务。生成的问题及其上下文（即包含问题及其答案的实际自然语言段落）表示为一个序列，并通过不同的片段掩码和“[SEP]”词元分隔。通过计算每个词元的对数似然，训练模型的最终线性层，以识别答案的起始和结束片段。所有实验均使用未区分大小写的 BERT-base 模型进行，学习率为 3e-5，最大序列长度为 384，批量大小为 12，文档步幅为 128，训练 2 个轮次，并每 500 步保存一次检查点。通过在从合成训练数据中随机抽取的 5000 个问答对上进行验证，选出最佳检查点。我们使用 Huggingface[4] 实现进行输入分词、模型初始化和训练。为了与最先进的 EQA 模型进行比较，我们还使用相同的训练数据，对采用全词掩码的 BERT-large 版本进行了实验。所有模型均在单张 NVIDIA Tesla A100 GPU 上训练和验证。

[1] https://rajpurkar.github.io/SQuAD-explorer/ [2] https://huggingface.co/tuner007/pegasus_改写

[3] https://spacy.io/universe/project/neuralcoref [4] https://huggingface.co

	SQuAD1.1		NewsQA		BioASQ		DuoRC	
PIE-QG 启发式方法	**EM**	**F-1**	**EM**	**F-1**	**EM**	**F-1**	**EM**	**F-1**
开放式信息抽取	22.8	36.5	13.0	23.9	16.6	24.3	22.0	28.5
+ 释义改写	37.7	53.6	19.9	32.3	20.3	31.6	32.6	40.7
+ 共指消解	44.2	53.4	21.1	31.8	26.5	34.7	34.8	40.4
+ 命名实体过滤器	46.6	56.5	21.7	32.5	**30.3**	36.9	**36.9**	**42.4**
+ 相同三元组过滤	47.5	57.8	21.8	32.2	30.1	37.1	35.1	41.1
+ 三元组合并	**48.6**	**58.7**	21.8	**32.8**	29.6	**37.5**	34.3	40.1
+ 语法纠正	47.1	56.8	**21.9**	32.3	29.1	36.5	35.2	40.7

表1：对 PIE-QG 中使用的不同技术进行消融研究，并分析微调 BERT-base 模型后这些技术对精确匹配和 F1 分数的影响。注：每一步都代表在前一步的基础上逐步改进问题生成。

5 结果与讨论

通过训练 BERT-base 模型，并在现有的 EQA 开发集和测试集上进行评估，来衡量使用 PIE-QG 方法生成的问答数据的有效性。我们选用完全匹配（精确匹配）和 F1 分数作为指标，以评估问答模型各个组件的有效性。初始问题集使用开放信息抽取（OpenIE）创建，直接利用段落构建三元组并生成问题，具体方法如第 3 节所述。使用 OpenIE 的初衷是从单个段落中生成多个问题。然而，如前所述，这种简单直接的方法会导致答案为代词、问题不合语法，以及段落与问题之间的词汇相似度过高，因此大多数抽取出的三元组仅适用于词匹配。

释义改写的影响 使用经过释义改写的段落来生成问题，可以避免与段落产生词汇重叠，并提升模型性能。我们使用 PEGASUS（Zhang 等人，2020年）释义改写生成模型，为段落中的每个句子生成十种不同的释义改写。针对每个改写后的句子，计算其与原句之间的詹森-香农散度（JSD）。JSD 根据两个句子的词分布计算散度分数：JSD 值越高，表示句子差异越大；JSD 值越低，则表示词汇重叠越高。在我们的 PIE-QG 流水线中，我们选择 JSD 值最高的句子来生成问题，使问题在句法上有所不同。释义改写对模型有显著的积极影响，在所有评估集上，精确匹配和 F1 分数分别至少提高了 4% 和 7%。

共指消解的影响 段落中出现代词会导致问答对毫无意义。例如，“瓦索·塞帕什维利（生于*1969*年*12*月*17*日）是一名退役的格鲁吉亚职业足球运动员。他于*1990*年代表阿克纠宾斯克阿克纠宾涅茨足球俱乐部，在苏联乙级联赛*B*组完成职业生涯首秀。”就是该段落。这会生成一个三元组“<他，完成了，在*1990*年代表阿克纠宾斯克阿克纠宾涅茨足球俱乐部于苏联乙级联赛*B*组完成职业生涯首秀>”。关系和宾语共同构成问题：“谁在1990年代表阿克纠宾斯克阿克纠宾涅茨足球俱乐部于苏联乙级联赛B组完成职业生涯首秀？”，其中主语“他”被选作答案。该问题的最佳答案可以在前一句中通过共指找到，其中代词“他”指的是“瓦索·塞帕什维利”。为解决这一问题，我们通过共指消解修改段落，将所有代词替换为其所指的专有名词。上述句子经过修改后，提取出的三元组变为“<瓦索·塞帕什维利，在*1990*年代表阿克纠宾斯克阿克纠宾涅茨足球俱乐部于苏联乙级联赛*B*组完成了职业生涯首秀>”，从而选出理想答案。使用此方法将代词替换为其所指名词，以生成有意义的问题；同时保留原始段落，用于训练问答模型。这样，共指消解对模型性能有积极影响，使所有数据集上的精确匹配提高了 2%–6%。

命名实体过滤由于三元组是训练问题的直接来源，三元组的质量会影响 PIE-QG 模型生成训练问题的质量。一般而言，OpenIE6 会从一个句子中提取所有可能的三元组，但为了生成更好的问答对，选择合适的三元组至关重要。为帮助确定最佳的三元组集合，我们会过滤掉不包含命名实体的三元组。我们使用命名实体识别（NER）从段落中提取所有命名实体。

	SQuAD1.1		NewsQA		BioASQ		DuoRC		
微调模型	**EM**	**F-1**	**EM**	**F-1**	**EM**	**F-1**	**EM**	**F-1**	**#训练上下文**
BERT-base									
句子检索 (Fabbri 等人，2020年)	46.1†	56.8†	20.1	31.1	29.4	**38.1**	28.8	35.0	45K
PIE-QG（本文方法）	**48.6**	**58.7**	**21.8**	**32.5**	**29.6**	37.5	**34.3**	**40.1**	**20-28K**
BERT-large									
完形填空翻译 (Lewis 等人, 2019) †	45.4	55.6	19.6	28.5	18.9	27.0	26.0	32.6	782K
RefQA (Li 等人,2020年)	57.1 †	66.8 †	27.6	41.0	42.0	54.9	41.6	49.7	178K
+ 迭代式数据精炼	**62.5** †	**72.6** †	**32.1**	**45.1**	**44.1**	**57.4**	**45.7**	**54.2**	240K
PIE-QG（我们的方法）	61.2	**72.6**	29.7	44.1	43.6	55.1	44.6	52.9	**20-28K**

表 2：PIE-QG 与最先进的无监督问答模型的比较。注：迭代式数据精炼通过借助引文和文档内链接对语料库进行结构分析，实现了最佳性能；该模型所需的上下文数量是我们提出的 PIE-QG 模型的 $\times 8$ 倍。“†”表示取自现有文献的结果，其他所有数值均基于已发布的合成训练数据（或）预训练模型进行评估。“训练上下文数量”根据各自已发布的合成数据集进行测量。每个模型都使用来自维基百科的相同合成训练数据进行微调，并在标准 EQA 数据集上进行评估。

要成为问题生成流程中可供选择的候选项，三元组中的主语或宾语必须至少包含一个命名实体。这种 NER 过滤方法有助于模型从训练集中剔除许多不切实际的问答对，并将整体完全匹配（精确匹配）和 F1 分数提高 2%，但 NewsQA 除外。

相同三元组过滤的效果 使用 OpenIE6 可以构造出词汇重叠度很高的语义相似三元组。根据这些三元组生成问题会导致问题重复，可能降低模型性能，甚至造成过拟合。为了过滤相似或重复的三元组，我们会将每个三元组与从段落中抽取的其他三元组进行核对，以发现它们之间的词汇重叠。如果一个以句子形式构成的三元组是另一个三元组的子字符串，则会从训练集中删除较短的三元组，以避免生成冗余问题。从图 1 中可以看到，诸如 <这些交易，可能违反，欧盟反垄断法> 和 <欧盟委员会，担心，这些交易可能违反欧盟反垄断法>这样的三元组表达了相同的含义，且词汇重叠度很高，因此删除前者。以这种方式过滤相同三元组 对模型有小幅但积极的影响，如表 1 中的消融实验总结所示。

三元组合并的影响段落中的主语（或宾语）可能与多个宾语（或主语）存在关系。将主语相同的三元组合并，形成复杂问题，使问答模型能够理解复杂关系。三元组合并对模型性能有小幅但积极的提升，如表1所示，精确匹配提高了 1.2%，F1 分数提高了 0.9%。

语法纠正的影响上述过程中生成的问题通常存在语法错误，这可能会对模型性能产生负面影响。我们试用了“GECToR”[5], 这一语法纠错模块，它会标记并纠正存在语法错误的输入问题。例如，问题“担心这些交易可能违反欧盟反垄断法的是什么？”就是一个例子。动词“是”重复出现，显然是个错误。语法纠错模块会修改该问题，最终将其纠正为“担心这些交易可能违反欧盟反垄断法的是什么？”。根据表格1中呈现的启发式方法，直到“三元组合并”为止的所有渐进式改进都提升了模型性能，但语法纠正并未带来提升，因此将其从流程中移除。

训练数据规模的影响我们进行了实验，测量不同合成数据规模下的精确匹配和 F 分数，以确定最佳训练问题数量。图3 展示了这些实验的结果，表明使用 BERT-large 模型时，PIE-QG 在 3万至5万个训练问题之间达到最佳性能，超过该数量后便开始过拟合。在（Fabbri 等人，2020年）中也观察到了相同的效应。确定最佳训练问题数量的方法是将生成的问答对按每块 1 万个划分，然后将这些数据划分为训练集和验证集。每经过固定的 500 个训练步数，就会使用验证集

[5]https://github.com/grammarly/gector

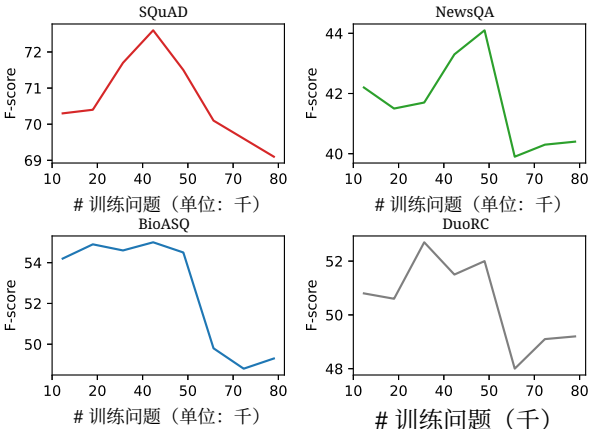

图3：使用 BERT-large 模型时，PIE-QG 模型在不同数据集上的 F 分数随训练集问题数量的变化情况；每个数据集的最佳数量范围为 30–50K。

对问答模型进行评估。通过这种渐进式评估，可以确定模型针对用于训练的问答对数量达到最佳效果的时机。图3显示，每个数据集的这一时机都出现在 3万至5万个问答对的范围内。超过一定数量后，增加模板式训练问题会降低评估性能，这是因为模型记住了合成数据模式。

与最先进水平的比较 Fabbri 等人 (2020年) 在实验中使用 BERT-base 模型作为骨干模型，而Lewis 等人 (2019) 和 Li 等人 (2020年) 则采用了 BERT-large 整词掩码预训练模型。PIE-QG 模型生成的问题表现优于Fabbri 等人 (2020年) 提出的基于信息检索的方法，在 SQuAD 1.1 开发集上的精确匹配提升了 2.5%，F1 分数提升了 1.9%。在比较 BERT-large 模型时，PIE-QG 模型优于基于引文检索的 RefQA；该方法涉及依存树重建。不过，采用了改进技术的 RefQA 性能最佳，其 F1 分数比 PIE-QG 高出 1%–2.5%，但代价是使用 8× 更多段落和 10×更多训练问题。此外，RefQA 的改进是在训练数据上通过 SQuAD 1.1 开发集的迭代交叉验证完成的，而 PIE-QG 模型不涉及此类过程。各方法使用的段落和问题数量详见表3。PIE-QG 优于基于检索的

系统	#上下文	#问题
Fabbri 等人（2020年）	45K	50K
RefQA Li 等人（2020年）	178K	300K
+ IDR	240K	480K
PIE-QG	20-28K	30-50K

表 3：现有无监督问题生成方法与 PIE-QG 所生成合成训练数据的统计对比。

问题生成方法在每个数据集上的表现，并且在使用8× 更少段落的情况下，取得了与 RefQA 相当的效果。

总而言之，实验结果展示了 PIE-QG 方法的优势：

1. 对原始段落进行释义改写，无需借助外部知识源即可避免词汇重叠；

2. 使用我们提出的方法，通过开放信息抽取生成多个问题，降低了对大型语料库的需求，同时不牺牲性能。

6 局限性

PIE-QG 无监督问题生成流程的缺点在于依赖释义改写、开放信息抽取和命名实体识别等外部模块，而这些模块可能并不存在于英语以外的语言中。因此，用于训练问答模型的问答对质量取决于这些模块在所选语料库上的性能。不过，我们预计 PIE-QG 在任何英语语料库上都能取得类似的良好表现。未来的工作是将这些模块应用于 PIE-QG 流程中的其他语言，并在能够获得可比的特定语言模型时分析其性能表现。

7 结论

PIE-QG 模型无需依赖任何外部参考语料库，而是利用释义改写和开放信息抽取（OpenIE），为基于 BERT 的问答系统中的语言模型生成合成训练问题，以进行微调。从释义段落中抽取形式为 <主语、谓语、宾语> 的三元组，并以主语（或宾语）作为答案来构造问题。消解代词共指项，并在可能的情况下合并三元组，同时删除重复及高度相似的三元组。此外，不包含命名实体的三元组

P	佐治亚理工学院的本科项目一直表现出色，我很高兴我们能够长期保持这一卓越水平，”代理校长兼教务长 Gary Schuster 说。
Q	谁表示，他很高兴佐治亚理工学院的本科项目继续表现出色?
A	**Gary Schuster**
P	“我们非常难过，也非常愤怒，”35 岁的美国籍律师 Raphael Felli 说道，他是被处决者的儿子阿奇扬庞政府时期的外交部长罗杰·费利上校。
Q	这位常驻美国的律师是谁？他是被处决的罗杰·费利上校的儿子。
A	**Raphael Felli**
P	《自由与暴政》销量达百万。Politics Radio 主持人 Mark R. Levin 的畅销书《自由与暴政：保守派宣言》已售出一百万册，据出版商 Threshold Editions……《自由》于2009年3月24日出版，首次登上《纽约时报》榜首《纽约时报》畅销书榜 .
Q	什么销量达到一百万，并登上了《纽约时报》畅销书榜?
A	**自由**

表4：使用 PIE-QG 生成的示例合成问答对。注：**P** 表示从文档中提取的段落。**Q**和**A** 分别表示生成的问题和从段落中选取的答案。

会被剔除。PIE-QG 流程生成高质量的问答训练集，用于训练问答模型。使用 PIE-QG 流程得到的问答模型，只需使用少得多的段落，便能达到与最先进水平相当的性能。它仅略逊于 RefQA；RefQA 采用迭代式数据精炼方法，因此依赖语料库的引文结构以及 $\times 10$ 更多训练问题。

参考文献

Kevin Clark 和 Christopher D. Manning。2016年。用于提及排序共指模型的深度强化学习。载于 *2016* 年自然语言处理实证方法会议论文集，第2256–2262页，得克萨斯州奥斯汀。计算语言学协会。

Hang Cui、Min-Yen Kan 和 Tat-Seng Chua。2004年。用于定义型问答的软模式无监督学习。载于第十三届万维网大会（*WWW 2004*）论文集，第90–99页。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深层双向变换器预训练。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，明尼阿波利斯，明尼苏达州。计算语言学协会。

Alexander Fabbri、Patrick Ng、Zhiguo Wang、Ramesh Nallapati 和 Bing Xiang。2020年。基于模板的从检索句子生成问题，以改进无监督问答。载于计算语言学协会第*58*届年会论文集，第4508–4513页，线上。计算语言学协会。

Adam Fisch、Alon Talmor、Robin Jia、Minjoon Seo、Eun-sol Choi 和 Danqi Chen。2019。MRQA 2019 共享任务：评估阅读理解中的泛化能力。载于*EMNLP* 第二届机器阅读理解（*MRQA*）研讨会论文集。

Matthew Honnibal、Ines Montani、Sofie Van Landeghem 和 Adriane Boyd。2020年。spaCy：基于 Python 的工业级自然语言处理。

Tushar Khot、Ashish Sabharwal 和 Peter Clark。2017年。利用开放信息抽取回答复杂问题。载于计算语言学协会第*55*届年会论文集（第*2*卷：短论文），第311–316页，加拿大温哥华。计算语言学协会。

Keshav Kolluru、Vaibhav Adlakha、Samarth Aggarwal、Mausam 和 Soumen Chakrabarti。2020年。OpenIE6：用于开放信息抽取的迭代网格标注与并列结构分析。载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第 3748–3761 页，线上。计算语言学协会。

Patrick Lewis、Ludovic Denoyer 和 Sebastian Riedel。2019 年。通过完形填空进行无监督问答

翻译。载于第 *57* 届计算语言学协会年会论文集，第 4896–4910 页，意大利佛罗伦萨。计算语言学协会。

Zhongli Li、Wenhui Wang、Li Dong、Furu Wei 和 Ke Xu。2020年。为无监督问答采集并精炼问题—答案对。载于第*58*届计算语言学协会年会论文集，第6719–6728页，线上。计算语言学协会。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019。RoBERTa：一种经过鲁棒优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Chenyang Lyu、Lifeng Shang、Yvette Graham、Jennifer Foster、Xin Jiang 和 Qun Liu。2021年。通过摘要信息辅助的问题生成改进无监督问题问答。载于*2021*年自然语言处理实证方法会议论文集，第4134–4148页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Kaixin Ma、Filip Ilievski、Jonathan Francis、Yonatan Bisk、Eric Nyberg 和 Alessandro Oltramari。2021年。面向常识问答零样本评估的知识驱动型数据构建。载于*AAAI* 人工智能大会论文集，第35卷，第13507–13515页。

Mausam Mausam。2016年。开放信息抽取系统及其下游应用。载于第二十五届国际人工智能联合会议论文集，第4074–4077页。

Kostiantyn Omelianchuk、Vitaliy Atrasevych、Artem Chernodub 和 Oleksandr Skurzhanskyi。2020年。GECToR——语法错误纠正：标注，而非改写。载于创新运用自然语言处理构建教育应用研讨会第十五届论文集，第163–170页，美国华盛顿州西雅图 →线上。计算语言学协会。

Liangming Pan、Wenhu Chen、Wenhan Xiong、Min- Yen Kan 和 William Yang Wang。2021年。通过问题生成进行无监督多跳问题解答。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5866–5880页，线上。计算语言学协会。

Alec Radford、KarthikNarasimhan、Tim Salimans 和 Ilya Sutskever。2018年。通过生成式预训练提升语言理解能力。

Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。《语言模型是无监督的多任务学习者》。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020年。探索迁移学习的极限：使用统一的文本到文本转换器。机器学习研究杂志，21(140):1–67。

Pranav Rajpurkar、Robin Jia 和 Percy Liang。2018年。了解自己所不知道的：SQuAD 中无法回答的问题。载于计算语言学协会第*56*届年会论文集（第*2*卷：短论文），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*年自然语言处理经验方法会议论文集，第2383–2392 页，得克萨斯州奥斯汀。计算语言学协会。

Amrita Saha、Rahul Aralikatte、Mitesh M. Khapra 和 Karthik Sankaranarayanan。2018年。DuoRC：通过释义阅读理解迈向复杂语言理解。载于计算语言学协会第 *56* 届年会论文集（第 *1* 卷：长篇论文），第 1683–1693 页，澳大利亚墨尔本。计算语言学协会。

Adam Trischler、Tong Wang、Xingdi (Eric) Yuan、Justin D. Harris、Alessandro Sordoni、Philip Bachman 和 Kaheer Suleman。2016 年。Newsqa：一个机器阅读理解数据集。

George Tsatsaronis、Georgios Balikas、Prodromos Malakasiotis、Ioannis Partalas、Matthias Zschunke、Michael R Alvers、Dirk Weissenborn、Anastasia Krithara、Sergios Petridis、Dimitris Polychronopou- los 等。2015年。BioASQ 大规模生物医学语义索引与问答竞赛概述。《*BMC*生物信息学》，16(1):1–28。

Zhilin Yang、Zihang Dai、Yiming Yang、Jaime Car- bonell、Russ RSalakhutdinov 和 Quoc V Le。2019。XLNet：面向语言理解的广义自回归预训练。神经信息处理系统进展，32。

Jingqing Zhang、Yao Zhao、Mohammad Saleh 和 Pe- ter Liu。2020年。PEGASUS：利用提取的间隔句进行抽象式摘要预训练。载于国际机器学习会议，第11328–11339页。PMLR。