从网络搜索引擎自动生成证据收集查询

Nestor Prieto-Chavana
萨塞克斯大学
n.prieto-chavana
@sussex.ac.uk

Julie Weeds萨塞克斯大学 juliewe
@sussex.ac.uk

David Weir萨塞克斯大学 d.j.weir
@sussex.ac.uk

arXiv:2303.08652v1 [cs.CL] 15 Mar 2023

摘要

人们普遍认为，可以通过在互联网上搜索信息来核查所谓的事实。这个过程要求事实核查员根据事实拟定搜索查询，并将其提交给搜索引擎。随后，需要从搜索结果中找出相关且可信的段落，再做出判断。许多新闻和媒体机构的副编辑每天都会进行这一工作。在此，我们探讨能否将第一步——生成查询——自动化。我们能否根据事实陈述自动拟定搜索查询，使其与人类专家拟定的查询相似？在这里，我们从文本相似度以及搜索引擎返回相关文档的情况两个方面考察相似性。首先，我们介绍一个中等规模的证据收集数据集，其中包含390条事实陈述，以及与之对应的人工生成的搜索查询和搜索结果。随后，我们研究使用多种基于规则的方法，以及基于预训练大语言模型（LLMs）的自动文本生成方法来生成查询。我们展示了这些方法各有优劣，并提出一种在实际应用中表现更优的混合方法。

1 引言

事实核查，即核验被当作事实呈现的陈述是否属实，在数字时代已成为日常生活中不可或缺的一部分。事实核查可以由事实的陈述者开展，例如作为新闻机构内部副编辑流程的一部分；也可以由事实的接收者在外部进行。无论是哪种情况，收集证据都是事实核查流程中的关键步骤。人们需要事实信息，才能对主张是否属实作出明智判断。证据可以来自多种来源，其中许多来源都会在互联网上免费提供信息。

搜索引擎已成为人们访问互联网的默认方式。在一家英国大型报纸跟班观察副编辑履行事实核查职责时，我们发现，网络搜索往往是寻找核验证据的第一步。随后，副编辑会从搜索结果中选择页面阅读；如果没有找到相关页面，他们就会调整查询，添加内容或进行其他修改，然后再次搜索，直到找到合适的证据页面。

我们着手开发一个自动化系统的证据收集模块，以复现这一行为。该模块负责根据已识别出主张的句子生成搜索查询，并收集这些页面中的文本。筛选相关片段以及评估主张的真实性不在本文讨论范围之内。然而，拟定恰当的搜索查询并非易事。长句可能陈述多个潜在事实，用户实际要核验的事实也可能难以确定。因此，所采用的查询生成技术是证据收集模块的关键组成部分。

在这项工作中，我们分析了多种条件文本生成技术，并将其性能与多种基于规则的生成基线进行比较。我们从两个方面评估自动生成的查询与人工编写的搜索查询之间的相似度：文本相似度，以及在搜索引擎上执行后产生的搜索结果。

我们的贡献有三点。首先，我们引入了一个专为事实核查应用设计的证据收集数据集，以支持编辑工作。其次，我们展示了如何有效运用条件文本生成技术，单独或结合基于规则的方法，创建实用的证据收集搜索引擎

。第三，我们证明，与人工创建的搜索查询的相似度，是判断自动生成的查询能否有效检索到相同证据的有用替代指标。不过，也存在一些不容忽视的情况：看似不同的搜索查询可能会收集到相同的证据。

本文结构安排如下：第2节讨论事实核查中的证据检索相关工作，以及提示和上下文学习。第3 节介绍我们为查询生成任务标注的数据集。第4 节阐述我们用于条件生成的方法及评估方案。第5 节描述实验。第6 节展示结果，第7节总结全文并展望未来工作。

2 背景与相关工作

本节回顾事实核查中的证据收集（第2.1节）以及大语言模型提示/上下文学习（第2.2节）方面的文献。

2.1 事实核查工作中的证据收集

以整个网络为信息来源开展事实核查的研究相对较少，而本研究正是采用这种方式。在Popat 等人(2016年)中，作者对从 Snopes[1] 事实核查网站和维基百科收集的主张进行可信度评估，并将主张分为真实或虚假。为从网络收集证据文章，作者将完整的主张文本输入 Google 搜索引擎，并抓取每条主张搜索结果的前三页。作者根据网站的 AlexaRank 和 PageRank（Page 等人，1999）为证据页面计算可靠性评分，并使用一系列词典来构建这些文章语言风格的表示。他们发现，来源可靠性评分对结果影响甚微。

Karadzhov 等人(2017年)提出了一种方法，将网络搜索作为事实核查的证据来源，并使用必应和 Google 作为搜索引擎。搜索查询通过结合以下规则构建：使用词频-逆文档频率添加排名靠前的动词、名词和形容词，以及添加主张中出现的所有命名实体。查询长度限制为 5-10 个词元。他们在由Ma 等人(2016年)发布的谣言检测数据集上评估了该方法。他们报告称，使用必应或 Google 的性能在实际应用中没有差异。他们还发现，使用页面全文作为证据与仅使用搜索引擎提供的摘要片段，几乎没有区别。

FEVER 任务（Thorne 等人，2018年）根据维基百科中包含的证据，将主张分为支持、驳斥和信息不足。对系统的评估不仅看其能否给出正确分类，也看其能否检索到正确证据。FEVER 的参与者采用了多种证据检索方法，有时会使用网络搜索或类似的搜索 API。例如，Hanselowski 等人（2018年）从主张文本中提取名词短语，并将其作为搜索查询提交给维基百科 API；该 API 使用维基百科搜索引擎查找匹配的文章。这种方法在共享任务首届赛事的所有参与者中取得了最高的证据召回率，之后的届次中也有参与者沿用了该方法。在 FAKTA（Nadeem 等人，2019年）中，作者除了使用 2017年维基百科转储中的证据外，还使用 Google 自定义搜索 API 在特定网站上搜索证据；这些网站的可靠性由 Media Bias/Fact Check 网站进行分类[2]。查询仅根据主张中出现的动词、名词、形容词和命名实体生成。

在Fan 等人(2020年)中，作者介绍了一个旨在生成事实核查简报的数据集，以提供可协助事实核查人员的信息。其中包括问答简报，它们根据主张提取具体问题及答案。为生成这些内容，作者训练了一个文本到文本模型 BART（Lewis 等人，2020年），使其能够根据文本中的主张自动生成问题。随后，他们训练了一个问答模型，根据问题文本以及最初用于核查源主张的源文档，为这些问题生成答案。我们的工作与此不同：我们利用主张文本生成搜索查询，并将其直接用于搜索引擎，以便从网络上查找证据文档。

[1] https://www.snopes.com

[2] https://mediabiasfactcheck.com

2.2 提示与上下文学习

大语言模型（LLM）的使用近期发生了一种范式转变，这种方法被称为“预训练、提示与预测”Liu 等人，2022年）。其背后的思路是调整当前任务，使其更接近模型的预训练任务。由于大语言模型是通过语言建模任务进行预训练的，这些调整通常是改变输入，使其更接近自然语言。根据具体任务，可以通过使用模板来实现，例如将分类任务转换为完形填空任务（Petroni 等人，2019年）。也可以使用提示（Brown 等人，2020年），即添加到模型输入中的简短任务说明。专注于寻找有效提示的研究方向称为提示工程（Liu 等人，2022年）。事实上，T5 模型的主要动机（Raffel 等人，2020年）是将传统自然语言处理任务重新表述为文本到文本任务（文本作为输入，文本作为输出），从而能够通过提示方法尝试执行其他任务。

这种方法的一种扩展是，同时加入模型需要执行的任务的已解示例。这被称为上下文学习（Brown 等人，2020年），也称为少样本学习。研究表明，在翻译、问答和完形填空任务等任务中，使用上下文示例所取得的结果可以与微调方法相媲美。虽然性能尚无法超越微调，但上下文学习的优势在于所需的数据量和计算资源较少，因为模型只进行推理。

如何选择加入上下文的示例，是这种方法的关键要素之一，因为这会对模型的性能产生重大影响（Lu 等人，2022年）。这些示例应体现模型预期执行的任务。示例可以从训练数据集中随机抽取（Brown 等人，2020年），也可以从规模较小的留出开发集中抽取。然而，在数据真正有限的情况下，这种方法可能并不可行。另一种做法是从密切相关的任务中抽取示例。然而，Perez 等人（2021年）反对使用此类域外数据。

3 数据集

本节介绍我们在这项研究中编纂并标注的数据集。该数据集由390条包含可验证信息的主张组成。每条主张都附有一个目标网址、一个包含验证该主张所需信息的证据页面，以及一个用于查找目标网址的目标搜索查询。主张来源于《泰晤士报》和《星期日泰晤士报》发布的新闻文章档案，涵盖20个不同的版块。我们从该收集中随机选取了100篇发表于2021年五月至六月的文章进行标注。

在标注过程中，我们首先使用 spaCy[3] 库将文章拆分成句子。其次，标注人员逐句阅读文章，判断每个句子是否包含可核实信息。只有被标记为可核实的句子才会进入后续标注步骤。第三，标注人员寻找相关证据，以便核实主张。为此，他们使用必应搜索引擎查找相关网页。此流程旨在遵循事实核查员基于主张查找信息的常规做法，不对所用查询的格式施加任何人为限制。唯一的限制是，所选网页必须出现在搜索结果的第一页。标注人员可以根据主张句中的信息构建搜索查询，必要时也可以使用文章其他句子中的补充背景信息。如果初次查询未找到相关网页，他们可以修改查询并根据需要反复尝试。标注人员会查看搜索结果，并选择排名最高且包含核实主张所需信息的网页。找到证据网页后，标注人员会为主张句标注用于找到该网页的最终搜索查询。最后，他们还会标注所选证据网页的统一资源定位符。处理历史文章数据带来了一些挑战，其中之一是许多主张会被其他媒体转载，导致搜索引擎提供的搜索结果常被此类内容占据。在整个标注过程中，我们要求标注人员优先使用一手来源，并尽可能避免使用其他新闻／媒体来源。如果新闻／媒体来源是唯一明显可用于验证的来源，则优先选择发布日期早于待核查文章的来源。

[3] https://spacy.io/

在实验过程中，我们发现，即使使用相同的查询，必应搜索引擎返回的搜索结果也会有所不同。原始数据集包含的主张多于目前的390条。反复执行目标搜索查询会产生不同的结果，而且搜索结果中并非总能找到目标网址。这令人担忧，因为目标网址是否出现将是评估生成的查询质量所用的指标之一。为确保数据集只包含能够通过相应目标搜索查询找到目标网址的样本，我们在3个不同日期于必应上执行了所有目标搜索查询，每个查询共执行12次，并收集了搜索结果。只有在12次执行中至少50%的次数能在搜索结果中找到目标网址的样本，才会保留在数据集中。最终，数据集中保留了390个样本。所有微调和评估均基于这390个样本进行。表1 展示了一些主张示例，以及各自对应的目标搜索查询和目标网址。

4 方法

在本节中，我们首先定义查询生成任务 4.1。接着概述基于规则 4.2 和条件文本生成 4.3方法以及所采用的评估方法4.4。

4.1 查询生成任务

查询生成任务定义如下：每个样本由一条文本声明 c、一个人工创建的查询 q 和一个目标网址 url 组成，其中包含验证 c 所需的证据。给定主张文本 c，生成一条搜索查询 q'，使其在搜索引擎上执行后，能在结果的第一页中返回 url。我们通过衡量 q' 与 q 的相似度，以及在必应搜索引擎上执行该查询并在结果中查找 url，来评估其质量。

4.2 基于规则的查询生成

基于规则的查询生成方法会考虑主张 c 中词元的句法特征，以决定是否将其纳入查询 q。我们考虑以下 3 种方法：

逐字内容：此方法使用主张的完整文本作为查询。搜索引擎很可能会在 c中找到与该文本高度相似的释义。

命名实体：在句子中识别出的所有命名实体会被连接在一起。这种方法有望找到同时提及这些实体的结果，其中可能描述了它们之间相似的关系，但未必与源句子中的主张一致。

名词短语：将句子中识别出的所有名词短语连接在一起。这与上述方法类似；不过，由于名词短语的定义比命名实体更宽泛，因此这种方法会纳入主张中的其他内容。

4.3 条件文本生成

我们将查询生成任务建模为条件文本生成。也就是说，将主张的完整文本 c 以及任何额外的上下文信息输入文本生成模型，并输出查询 q'。我们尝试了三种条件文本生成方法：

零样本学习：使用提示模板为主张添加前缀 c，然后将其输入文本生成模型。我们尝试了不同的提示，详见第5.2节。

少样本学习：在少样本学习中，模型会先看到少量示例，然后再尝试执行任务。具体做法是将这些示例直接放在主要任务输入之前。

常规微调：最后，我们还对常规微调进行了实验。在这种设置下，T5 模型以主张句 c 作为输入，以人工创建的查询q 作为目标进行微调。由于数据集规模有限，我们采用 4 折交叉验证。对于每一折，模型使用 75% 的数据进行微调，并在剩余的 25% 数据上测试。我们对 4 个测试集上的评估指标取平均值。

4.4 评估

我们通过两种方式评估不同的查询生成方法：（1）生成的查询与目标查询的相似度；（2）生成的查询返回目标网址的能力。

查询相似度评估

我们采用这一评估方法的理由是：如果模型生成的查询与人工标注员给出的查询相同，那么该模型的质量就很高。进一步来说，相同的查询比不同的查询更有可能产生相同的搜索结果。即使考虑到前文所述的搜索结果不一致问题，这一点依然成立。

主张	目标搜索查询	目标网址
利兹·特拉斯，国际贸易大臣，秘书长表示，该组织的成员资格_久久爱全面与进步跨太平洋伙伴关系协定代表着“一个巨大的机遇英国的机会”。	Liz 特拉斯 国际贸易大臣	https://en.wikipedia.org/wiki/利兹·特拉斯_
经济合作与发展组织上调了其全球增长将预测值从 5.6% 上调至 5.75% 的每우우우우우从5.6%上调至5.75%，今年的预测值从4%上调至接近4.5%。	oecd 全球 gdp 增长预测五月 2021	https://www.oecd.org/economic-outlook/五月-2021年
伽利略，欧洲系统，运-行高度为23,000公里，但英国正被逐出伽利略系统因为英国脱欧 .	伽利略 运行高度	https://en.wikipedia.org/wiki/伽利略_（卫星_导航）

表1：数据集中的主张、目标搜索查询及相应目标网址示例。

Rouge：我们使用 Rouge（Lin，2004年）评估自动生成的查询与目标查询之间的文本相似度。这一指标广泛用于评估摘要生成和翻译等任务中的模型表现。我们使用 Google Research 提供的 Rouge Python 实现[4]。由于目标查询和自动生成的查询都很短，因此我们改用基于字符的 Rouge，即用按字符拆分句子的方式替代标准分词器。我们报告所有实验的 Rouge-1、Rouge-2 和 Rouge-L 指标。

我们还考虑过使用查询之间的莱文斯坦距离或编辑距离（Levenshtein 等人，1966年）。理论上，基于字符的 Rouge 分数受查询词语重排的影响应该较小，但我们在实践中发现，莱文斯坦距离分数与 Rouge-L 分数高度相关，因此未在此处列出。不过，我们使用莱文斯坦距离来选择少样本学习中作为上下文示例的最优样本（见第5.3节）。

搜索结果评估

我们通过 Bing 搜索 API[5]执行查询并收集排名前 10 的搜索结果，这些结果对应于浏览器中的第一页。我们评估目标网址是否出现在搜索结果中。

这项评估存在一些复杂之处。首先是搜索结果的一致性：在实验中，我们发现，使用相同的查询并不能保证从搜索引擎获得相同的结果。为应对这种差异，我们将每个查询执行 3 次，并收集三份不同的搜索结果排序列表。其次，这项评估会产生使用搜索服务提供商 API 的费用。在评估不同的方法和超参数配置时，无论是微调还是生成，费用都可能迅速累积。为提高结果的一致性而进行的重复执行又会进一步增加费用。为此，我们在调整模型超参数时使用查询相似度评估，而只对最终模型进行搜索结果评估。我们使用以下指标进行搜索结果评估：

全部命中百分比（FA%）：在所有三次执行中，目标网址均出现在搜索结果中的测试样本所占的比例。

多数命中百分比 (FM%)：目标网址在多数执行过程中（即至少两次）出现在搜索结果中的测试样本所占的比例。

至少一次命中百分比 (FO%)：至少在一次执行的结果中找到目标网址的样本所占的比例。

平均倒数排名：平均倒数排名（MRR@K）是信息检索任务评估中常用的一项指标（Voorhees 和 Tice，2000年）。该指标考虑检索到的文档中排名最高的相关答案。我们设定 $K = 10$，对应搜索引擎搜索结果的第一页。每个已执行的搜索查询所对应的搜索结果列表都单独计算，并计入总体 MRR 得分。

[4]https://github.com/google-research/google-research/tree/master/rouge [5] https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

5 实验

本节将介绍我们实验中所用各方法的具体实现。

5.1 基于规则的查询创建基线

我们使用 spaCy 库从主张文本中提取命名实体提及以及句法树。表 9 见附录A 展示了不同基于规则的方法的应用示例。

5.2 条件文本生成

我们使用 T5 语言模型（Raffel 等人，2020年），具体来说，使用 Lester 等人（2021年）提出的语言生成适配模型（t5-lm-adapt）。该版本针对语言模型任务额外训练了 10 万步，以克服原始 T5 模型的训练局限性；原始模型难以利用未损坏的文本。我们使用该模型的大型检查点，其参数量为 8 亿。我们使用 Hugging Face Transformers Python 库提供的实现[6]。文本生成时，我们设置了以下参数：采用 10 束的束搜索，限制模型重复生成二元组，并将生成提前停止设为 true。由于生成的查询不应过长，我们将新词元的最大数量设为 16。

在实验过程中，我们注意到 T5 模型倾向于在生成的文本中复现提示的部分内容。我们在少样本学习和零样本学习实验中尤其观察到了这一现象。为此，我们在生成后增加了一个处理步骤：如果模型输出中逐字包含提示文本（来自前缀或后缀），就将其删除。部分包含的内容不会从输出中删除。

正如 Petroni 等人 （2019年）；Brown 等人（2020年）；Schick 和 Schütze （2021年）所指出的，所使用的提示选择会显著影响模型性能。在 T5 论文原有任务中，我们认为摘要生成与查询提取任务最为接近，因此在此尝试使用“summarize”提示。我们还尝试了不同描述模型需执行任务的提示：生成搜索查询、对以下句子进行事实核查、验证以下句子和总结以下句子提示。我们还考虑了更短的提示，仅使用一两个词元来指示模型需要执行的操作：搜索查询、事实核查、验证和总结。我们尝试将较短的提示同时用作模型输入的前缀和后缀。实验中使用的提示详见附录A。

5.2.1 零样本学习

使用这种方法时，提示选择是唯一指导模型生成输出的因素。

5.3 少样本学习

我们使用 4 折交叉验证评估少样本学习方法。划分依据是主张的来源文章，也就是说，源自同一篇文章的主张只会出现在其中一个集合中，即训练集或测试集。各折训练集和测试集中的记录数量见附录10 A。在每一折中，模型使用 25% 的数据进行评估，其余 75% 可用作文本生成模型的上下文示例。为了选择最优的上下文示例，我们首先进行单样本学习评估，将训练集中的每个样本分别用作示例，并使用莱文斯坦距离指标，将生成的查询与训练集中的其余部分进行比较。表现最好的 3 个示例随后会在单样本、双样本和三样本设置下接受完整评估。

5.4 常规微调

我们再次采用4折交叉验证。每次使用75%的主张进行微调，并在剩余25%的主张上评估。与之前一样，这种划分依据来源文章进行。对于每一折，从训练数据中选取15%作为验证集。每个微调轮次结束时，使用验证集评估模型性能。评估依据是模型生成的验证集目标搜索查询的莱文斯坦相似度比率。微调结束时，保存模型表现最佳的检查点。每个模型都在单块 GeForce RTX 3090 GPU 上微调10轮。遵循 Raffel 等人(2020年)，我们使用 Adafactor 优化器，学习率设为1e-3，最大输入序列长度为512。生成参数与零样本学习和少样本学习所用参数相同。

6 结果与讨论

表2 展示了基于规则的基线方法，以及零样本、少样本和微调方法中表现最佳的提示在 FM% 指标上的结果。提示选择的影响

[6]https://github.com/huggingface/transformers

生成方法	R-1	R-2	R-L	FA%	FM%	FO%	MRR
逐字主张	0.343	0.294	0.314	15.1	17.2	18.2	0.114
名词短语	0.455	0.368	0.396	19.7	20.5	22	0.145
命名实体	*0.577*	*0.423*	*0.484*	*30.0*	*32.3*	*33.8*	*0.217*
零样本	0.533	0.388	0.438	25.1	26.2	29	0.181
单样本	0.546	0.339	0.418	19.0	20.5	23.1	0.15
双样本	0.582	0.382	0.46	25.6	26.4	28.2	0.191
三样本	0.593	0.402	0.474	30.3	31.5	32.8	0.225
微调	**0.646**	**0.472**	**0.56**	**41.0**	**41.8**	**42.6**	**0.321**

表2：各查询生成方法和基于规则的基线方法的性能指标。对于每种生成方法，均展示表现最佳的模板。

生成方法	FA%	FM%	FO%	MRR
目标查询	92.1	93.6	95.1	0.697

表3：使用目标查询进行网络搜索的评估结果。

	FA%	FM%	FO%	MRR
Rouge-1	0.842	0.831	0.817	0.848
Rouge-2	0.976	**0.98**	**0.984**	0.972
Rouge-L	**0.977**	0.972	0.966	**0.979**

表 4：基于相似度的指标与基于搜索的指标之间的相关性。

对结果的影响将在附录A中讨论。对结果的分析表明，FA%、FM% 和 FO% 相差不大。在所有实验中，三项指标的平均差异为 4%。换言之，只要目标结果至少在搜索结果中出现过一次，它很可能在大多数时候都会出现。然而，这些指标之间的差异也证实，网页搜索结果具有非确定性。它们也可能在短时间内发生变化：每个查询的所有必应搜索都是紧接着连续进行的，但仍经常得到不同的结果。

我们还观察到，网页似乎很快就会从结果中消失。请注意，在创建数据集时，我们对目标查询进行了验证，以确认它们能够返回目标网址。只有在大多数情况下都能返回目标网址的查询才被纳入数据集。之后，在实验期间，我们测试了使用目标查询进行收集的效果。如表3所示，只有93.6%的目标查询在结果中返回了目标网址，而95.1%的查询至少返回过一次。这意味着，从创建数据集到开展这些实验的这段时间（约两个月）内，近5%的样本所对应的目标网址已完全从结果中消失。这构成了任何系统所能达到性能的上限。

由于我们评估生成的查询与目标查询之间的文本相似度，并评估搜索结果的质量，因此我们评估这两种评估方法之间的相关性。结果如表4 所示，并证实与目标查询的相似度是衡量搜索结果潜在质量的一个很好的指标。Rouge-2 指标与 FA%、FO% 和 FM% 指标的相关性很高，而 Rouge-L 与平均倒数排名得分的相关性最高。

然而，相似度得分低并不一定意味着该查询无法返回目标网址。在分析结果时，我们发现许多情况下，Rouge 得分较低的查询仍然产生了预期的搜索结果。在这些情况中，大多数目标网址都是维基百科页面，这意味着在搜索查询中仅包含一个命名实体，就可能足以让搜索引擎返回相关页面。反过来，我们也发现了一些与目标查询相似度极高（基于 Rouge-2 指标）的生成查询，却仍未能返回目标网址。

回到表2中的结果，我们注意到，在基于规则的基线方法中，简单的命名实体基线与目标查询最为接近。这表明，人工创建的查询中有很大一部分包含命名实体。

在上下文内生成方法中，三样本学习的表现与

	命名实体匹配	命名实体失败	总计
微调匹配	96	67	163
微调失败	30	197	227
总计	126	264	390

表5：命名实体基线和微调生成方法在搜索结果中生成目标统一资源定位符的样本重叠情况。

命名实体基线不相上下，但未能超越它。其他上下文内方法的表现都远不及基线。与此同时，微调的表现超过了命名实体基线。这表明，虽然命名实体是目标查询的重要组成部分，但有效查询的构成还涉及其他因素。例如，当文本中出现多个命名实体时，需要选出正确的命名实体，并加入其他非命名实体词语。

为探究不同生成方法的性能差异，我们考察了排名前两位的方法各自能够正确解决的具体样本。表5 展示了命名实体基线与微调方法正确处理的样本重叠情况。尽管命名实体基线的表现不如微调方法，但仍有相当一部分样本（最多227个中的30个）是命名实体基线能够正确处理而微调方法无法处理的。这表明不同方法各有所长，最佳的使用方式是将这些优势结合起来。

为了进一步探索集成系统的潜力，我们采用改进的博尔达计数法来合并各查询生成方法的排名，同时优先考虑单独表现最好的方法。我们生成一份合并后的搜索结果前十排名，并使用与评估单项方法相同的指标对其评分。我们采用这种方法，合并表现最佳的前两种查询生成方法（命名实体和微调）以及表现最佳的前三种方法（命名实体、微调和三样本）。我们还尝试合并所有基于规则的基线方法。如表6所示，不同方法的集成表现可以优于其中任何一种单独方法。

6.1 误差分析

即使合并了表现最佳的生成方法，性能指标仍显示出很大的改进空间。为了找出最佳改进机会，我们分析了一部分微调模型输出，这些输出未能生成目标网址，并对其进行分类。详细分类见附录11中的表A。最常见的错误是模型未能在查询中包含关键术语，从而改变了所要查找信息的性质。第二常见的情况是，目标查询包含主张句之外的术语或上下文。可能的解决方法是在模型输入中加入文章标题或前文句子。第三常见的错误原因是，当主张文本中出现多个命名实体时，模型关注了错误的实体。我们还发现，幻觉出现的频率相同：模型输出没有考虑输入文本，而是生成了看似来自训练样本的文本。接下来一种失败情况是，模型仅将主张文本逐字复现，直至达到令牌上限。最后，我们发现有少数情况下，尽管查询与目标查询不同，但似乎涵盖了获取正确搜索结果所需的关键术语。这些情况可能是由于网站结构发生变化，或仅仅是网络搜索引擎结果本身存在固有差异所致。

7 结论

本文介绍了一个专门用于新闻副编辑情境下事实核查的证据收集数据集。我们证明，条件文本生成方法在生成正确搜索结果方面优于基于规则的基线方法。我们还探讨了集成多种此类方法的效果，并证明其表现优于任何单一方法。我们也分析了不同生成方法的不足之处，并发现了潜在的改进方向。未来可以尝试在自动化模型输入中加入更多上下文。也可以尝试更复杂的方式来组合不同方法的输入，例如，利用基于规则的基线方法的输出来增强自动化

生成方法	FA%	FM%	FO%	MRR
命名实体 + 微调	45.1	46.9	47.9	0.332
命名实体 + 三样本 + 微调	44.9	46.7	47.9	0.345
全部基于规则（命名实体 + 名词短语 + 逐字内容）	35.4	37.9	41.3	0.227
全部基于规则 + 微调	47.9	49.2	50.5	0.331
全部基于规则 + 三样本 +微调	47.4	49.7	51.3	0.356

表6：使用波达计数法组合不同方法的网络搜索评估结果。

生成方法的输入。最后，还可以采用进一步的数据标注或数据增强方法，以应对数据集规模相对有限所带来的挑战。

局限性

受实际条件限制，我们的数据集仅为每条主张标注了一个目标网址。不过，网络上可能还有其他同样有用的证据页面，这一点是合理的。因此，我们的结果可以视为对所展示方法实际表现的下限：对于失败样本，返回的内容也可能是有效证据，只不过不是标注人员选中的证据。另一项局限是证据具有时效性。对于数据集中的一些主张，证据仅在有限时间内有效。所选目标网址的相关性可能会随时间推移而降低，这些主张中的信息也可能随时间变化（例如，相关人员的职位可能发生变动，统计数据也可能有所变化）。时间久了之后，可能很难复现本文展示的结果。

致谢

我们感谢《泰晤士报》和《星期日泰晤士报》分享用于创建该数据集的数据，也感谢其副编辑团队分享他们对事实核查流程的见解。我们还要感谢匿名审稿人提供的反馈和意见。

参考文献

Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess, Jack Clark, Christopher Berner, Sam Mc- Candlish, Alec Radford, Ilya Sutskever 和 Dario Amodei。2020年。《语言模型是少样本学习者》。载于第*34*届神经信息处理系统国际会议论文集，NIPS'20，美国纽约州雷德胡克。Curran Associates Inc.

Angela Fan、Aleksandra Piktus、Fabio Petroni、Guillaume Wenzek、Marzieh Saeidi、Andreas Vlachos、Antoine Bordes 和 Sebastian Riedel。2020年。生成事实核查简报。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第7147–7161页，线上。计算语言学协会。

Andreas Hanselowski、Hao Zhang、Zile Li、Daniil Sorokin、Benjamin Schiller、Claudia Schulz 和 Iryna Gurevych。2018年。UKP-athene：用于声明验证的多句文本蕴含。载于首届事实提取与验证研讨会（*FEVER*）论文集，第103–108页，比利时布鲁塞尔。计算语言学协会。

Georgi Karadzhov、Preslav Nakov、Lluis Marquez、Alberto Barron-Cedeno 和 Ivan Koychev。2017年。使用外部来源进行全自动事实核查。载于*RANLP 2017*- 自然语言处理最新进展与深度学习，第 344–353 页。Incoma Ltd. 舒门，保加利亚。

Brian Lester、Rami Al-Rfou 和 Noah Constant。2021年。参数高效提示调优的规模效应。载于*2021*年自然语言处理实证方法会议论文集，第 3045–3059 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Vladimir I Levenshtein 等。1966年。能够纠正删除、插入和反转的二进制码。载于《苏联物理学报告》，第 10 卷，第 707–710 页。苏联。

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，第7871–7880页，线上。计算语言学协会。

Chin-Yew Lin。2004年。ROUGE：自动评估摘要的工具包。载于文本摘要领域的新进展，第74–81页，西班牙巴塞罗那。计算语言学协会。

Pengfei Liu、Weizhe Yuan、Jinlan Fu、Zhengbao Jiang、Hiroaki Hayashi 和 Graham Neubig。2022年。预训练、提示与预测：自然语言处理领域提示方法的系统性综述。*ACM*计算机调查已接受。

Yao Lu、Max Bartolo、Alastair Moore、Sebastian Riedel 和Pontus Stenetorp。2022年。奇妙排序的提示及其所在之处：克服少样本提示顺序敏感性。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8086–8098页，爱尔兰都柏林。计算语言学协会。

Jing Ma、Wei Gao、Prasenjit Mitra、Sejeong Kwon、Bernard J. Jansen、Kam-Fai Wong 和 Meeyoung Cha。2016年。使用循环神经网络检测微博中的谣言。载于第二十五届国际人工智能联合会议论文集，IJCAI'16，第3818–3824页。AAAI出版社。

Moin Nadeem、Wei Fang、Brian Xu、Mitra Mohtarami 和 James Glass。2019年。FAKTA：一种自动化端到端事实核查系统。载于《*2019*年北美计算语言学协会年会论文集（演示）》，第78–83页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Lawrence Page、Sergey Brin、Rajeev Motwani 和 Terry Winograd。1999年。PageRank 引文排名：为网络带来秩序。技术报告 1999-66，斯坦福信息实验室。

Ethan Perez、Douwe Kiela 和 Kyunghyun Cho。2021年。借助语言模型实现真正的少样本学习。载于*NeurIPS*。

Fabio Petroni、Tim Rocktäschel、Sebastian Riedel、Patrick Lewis、Anton Bakhtin、Yuxiang Wu 和 Alexander Miller。2019年。语言模型是知识库吗？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*）》，第2463–2473页，中国香港。计算语言学协会。

Kashyap Popat、Subhabrata Mukherjee、Jannik Ströt- gen 和 Gerhard Weikum。2016年。网络文本主张的可信度评估。载于第*25*届美国计算机协会国际信息与知识管理会议论文集，CIKM '16，第 2173–2178 页，纽约，美国。美国计算机协会。

Colin Raffel、NoamShazeer、Adam Roberts、Katherine 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(1)。

Timo Schick 和 Hinrich Schütze。2021年。重要的不只是规模：小型语言模型也是少样本学习器。载于*2021*年计算语言学协会北美分会人类语言技术会议论文集，第 2339–2352 页，线上。计算语言学协会。

James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018年。FEVER：用于事实抽取与验证的大规模数据集。载于 *2018*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（长篇论文），第 809–819 页，美国路易斯安那州新奥尔良。计算语言学协会。

Ellen M. Voorhees 和 Dawn M. Tice。2000年。TREC-8 问答赛道。载于第二届国际语言资源与评估会议（*LREC'00*）论文集，希腊雅典。欧洲语言资源协会（ELRA）。

A 附录

表7展示了在零样本和少样本生成方法中用于提示的不同模板。这些相同的模板也应用于微调过程中。表6所报告的分数是在以下设置下取得的：零样本学习使用 template-03，单样本学习使用 template-06，两样本和三样本学习使用 template-05，微调则不使用提示。总体而言，任务描述较长的提示对零样本学习更有帮助。与此同时，简短的后缀提示在少样本学习中取得了最佳结果。使用较短提示时，将其作为输入的后缀能带来更好的结果，而简短的前缀提示效果较差，尤其是在少样本设置下。对于微调，提示的选择几乎没有影响；完全不使用提示时，性能最佳。

各方法对提示选择的敏感程度差异很大。少样本设置表现出最强的敏感性：不使用提示或使用短前缀类型提示的实验，性能都非常低。这种影响在零样本学习中不那么明显，而在微调中，提示选择的影响可以忽略不计。表8展示了各方法在所有提示模板上的平均性能，以及标准

格式	模板 ID	前缀	后缀	零样本示例
无提示	无提示	-	-	Gerry Ford，63 岁时创立了这家企业成立于 1997 年。
长解释	模板-01	生成搜索查询：	搜索查询：	生成搜索查询：Gerry Ford，63 岁，创立于 the 企业 in 1997 年。
	模板-02	事实核查以下句子：	事实核查：	核查以下句子：Gerry Ford，63岁，创办了这家企业，时间是在 1997. 事实核查：
	模板-03	核实 the 以下句子：	核实：	验证以下句子：Gerry Ford，63，创立了企业，时间是 1997年。验证：
	模板-04	总结以下内容句子：	总结：	总结以下内容句子：Gerry Ford，63 岁，于.....创立了这家企业 1997 年。总结：
短后缀	模板-05	-	搜索查询：	Gerry Ford，63 岁，创立于 the 企业 in 1997 年。搜索查询：
	模板-06	-	事实核查：	Gerry Ford，63岁，创立了这家企业于1997年。事实核查- 核查：
	模板-07	-	验证：	Gerry Ford，63岁时创立了这家企业，时间是1997年。*Ver-ify:*
	模板-08	-	总结：	Gerry Ford，63岁时创立了这家企业，时间是1997年。*Sum-* 总结：
短前缀	模板-09	搜索查询：	-	搜索查询：Gerry Ford，63 岁时创立了企业于 1997 年。
	模板-10	事实核查：	-	事实核查：Gerry Ford，63，创立了企业于1997年。
	模板-11	验证：	-	验证：Gerry Ford，63，创立了这家企业 1997。
	模板-12	总结：	-	总结：Gerry Ford，63，创立了这家企业于1997年。

表 7：用于零样本、少样本和微调方法的提示模板。

错误。我们还展示了少样本设置下相同的指标，但未纳入不使用提示或使用短前缀式提示的结果。

表9 展示了将基于规则的基线方法应用于以下示例句子的输出：内-塔尼亚胡并非偶然成为以色列任职时间最长的总理。待核实的事实是，内塔尼亚胡是否确实是以色列任职时间最长的总理。

表10 展示了通过4折交叉验证划分得到的训练集、开发集和测试集的规模。划分依据是主张所属的来源文章。

表11 列出了对部分微调模型预测进行错误分析的结果和示例。

方法	R-2	FM%
零样本	0.340 (0.008)	19.7 (1.1)
微调	0.454 (0.004)	38.4 (0.5)
单样本	0.263 (0.014)	12.1 (1.7)
双样本	0.285 (0.026)	16.2 (2.9)
三样本	0.297 (0.027)	18.8 (3.2)
单样本*	0.299 (0.009)	16.5 (1.1)
双样本*	0.359 (0.004)	24.4 (0.8)
三样本*	0.372 (0.005)	27.9 (0.7)

表8：提示选择导致的结果差异。数值为使用不同提示模板进行实验所得指标的平均值。括号内的数值为标准误。带星号的方法不包括使用无提示或短前缀提示进行的实验。

生成方法	查询
逐字内容	内塔尼亚胡并未成为以色列任期最长的 误将总理。
命名实体	内塔尼亚胡，以色列
名词短语	内塔尼亚胡，以色列任职时间最长的总理， 错误

表9：不同基于规则的查询生成基线的输出。

Fold	训练集	Dev	Test
K0	245	45	100
K1	246	48	96
K2	243	48	99
K3	248	47	95

表10：微调数据划分中的样本数量。

原因	示例	百分比
查询缺少关键信息/术mation/term	主张文本：　　与此同时，苏格兰青年旅舍协会——英国青年旅舍协会在边境以北的对应组织今年迎来成立90周年——两天前开始分阶段重新开放其60多家旅舍。 目标查询：苏格兰青年旅舍协会历史 生成的查询：苏格兰青年旅舍协会	30%
查询所需的上下文主张之外	主张文本：开幕式将在东京国立竞技场举行，日期定于7月23日。 目标查询：东京奥运会开幕式仪式 生成的查询：东京国立竞技场东京	26.7%
提取到错误的命名实体	主张文本：普丽蒂·帕特尔正在考虑让英国退出《欧洲社会宪章》的一部分条款该条款规定，来自26个国家的公民可享受55英镑的折扣，适用于大多数工作签证的申请费。 目标查询：《欧洲社会宪章》费用豁免 生成的查询：普丽蒂·帕特尔	16.7%
幻觉/记忆内容来自训练数据	主张文本：对于这家网约车巨头，二月的那项判决可能是一场更广泛追责的开端，这将使其面临超过20亿英镑的增值税账单。 目标查询：uber 二月 £2billion 生成的查询：mup 苏格兰	16.7%
重构的主张文本	主张文本：科罗拉多州和俄勒冈州已经效仿，推出了类似的彩票，并且周四，加利福尼亚州宣布了一项规模更大的抽奖活动，设有十个奖项，价值150万美元。 目标查询：加利福尼亚州疫苗抽奖彩票娱乐数量 生成的查询：科罗拉多州和俄勒冈州也纷纷推出了类似的彩票。	6.7%
查询看起来不错	主张文本：Chris James，创始人来自影响力投资公司 Engine No. 1，上周在对决中大获全胜，击败了大型石油公司丛林中最凶猛的巨兽：埃克森美孚。 目标查询：chris james engine no 1 生成的查询：Chris James 创始人影响力投资公司 Engine No. 的	3.3%

表 11：微调模型输出错误分析。