NewsQuote：用于事实核查中专家推荐的引语提取与归因数据集

Wenjia Zhang, [1] **Lin Gui,** [2] **Rob Procter,** [1,3] **Yulan He** [1,2,3]

[1] 英国华威大学计算机科学系 [2] 英国伦敦国王学院信息学系[3] 英国艾伦·图灵研究所{Wenjia.Zhang,Rob.Procter}@warwick.ac.uk, {lin.1.gui,yulan.he}@kcl.ac.uk

arXiv:2305.04825v1 [cs.IR] 5 May 2023

摘要

为提升从新闻文章中寻找可信证据的能力，我们提出了一项新的专家推荐任务，旨在识别特定新闻主题下值得信赖的专家。为实现这一目标，我们介绍了一个新的 NewsQuote 数据集的构建过程。该数据集包含 24,031 个出现在 COVID-19 新闻语料库中的引语-说话者对。我们展示了一套基于 BERT 问答模型的自动化流程，用于提取说话者和引语。随后，我们将专家推荐建模为文档检索任务：首先检索相关引语，作为识别专家的中间步骤；此外，还通过直接根据查询在给定候选专家条件下的概率检索来源，来实现专家检索。NewsQuote 数据集上的实验结果表明，与专家检索相比，文档检索在针对给定新闻主题识别相关专家方面更有效。[1]

1 引言

近年来，虚假信息迅速增长，受到学术界、记者、政治分析人士和事实核查组织的广泛关注，并推动了关于基于自然语言处理的技术和工具的研究，以支持事实核查工作和证据验证（Lazarski、Al-Khassaweneh 和 Howard，2021；Zeng、Abu-mansour 和 Zubiaga，2021；Guo、Schlichtkrull 和 Vlachos，2022）。这项研究工作大多基于一种以文档为中心的 事实核查工作模式，其最终目标是向记者或事实核查人员提供一份（自动生成的）与待核查主张相关的文档排序列表，供他们用作证据，以判断该主张可能的真实性（例如，Zhao 等人，2023）。

我们最近的研究表明，一些事实核查人员采用以专家为中心的模式，即寻找愿意被引用的可信且值得信赖的专家（Procter 等人（2023））。寻找这样的专家是一项重大挑战，记者和事实核查人员通常会设法采访多位专家，因为仅依赖单一来源可能不被认为具有足够的可信度。对于存在争议的说法，他们还可能需要确保其报道保持平衡（Procter 等人（2023））。

因此，迫切需要开发一种工具，帮助记者和事实核查人员依据专家过去被新闻媒体、事实核查机构及其他可信机构引用的记录来搜索专家。为实现这一目标，我们首先需要从新闻文章中自动提取引语及其来源，然后返回与查询相关的专家排名列表，供记者或事实核查人员评估。这可以表述为两项任务：（1）引语提取与归因；（2）专家推荐。

对于引语提取与归因这一首要任务，大多数数据集基于文学叙事构建，且由于依赖人工标注，规模有限（Zhang、Black 和 Sproat，2003；Elson 和 McKeown，2010；Fernandes、Motta 和 Milidi´u，2011；Lee 和 Yeung，2016）。但与小说相比，新闻通讯社中的独白和对话要少得多（O'Keefe 等，2012）。早期研究依赖规则驱动的框架和人工定义的语言模式，因此主要关注直接引语（Lee 和 Yeung，2016；Zhang 和 Liu，2021；Vaucher 等，2021）。与戏剧剧本或小说不同，新闻媒体报道中被引用的人并不局限于固定人物名单。此外，新闻文章报道的事件不断变化，新闻媒体机构采用的书写风格也各不相同，因此很难依靠正则表达式来识别专家和提取引语。

对于专家推荐的第二项任务，学术研究（Sun 等，2015；Silva，2014；Wang 等，2017）、在线社区（Yuan 等，2020）以及企业领域（Paul，2016；Askari、Verberne 和 Pasi，2022）都已开展了大量专家发现工作。然而，据我们所知，尚无研究根据专家在新闻文章中被引用的记录来寻找专家。本文提出一种半自动方法，基于 AYLIEN 冠状病毒数据集[2],构建新闻引文数据集，称为 NEWSQUOTE。该数据集包含来自约 440 个全球来源的超过 150 万篇英文新闻文章。我们利用新闻文章中句子的语义角色标注结果，提取引语触发动词、主体（即来源）和宾语（即引文），并根据相应的 DBpedia[2],本体类标签识别来源。

版权所有 © 2022，人工智能促进协会（www.aaai.org）。保留所有权利。[1]我们的源代码可通过以下地址访问：https://github.com/ WenjiaZh/NewsQuote

[2]这些数据由 AYLIEN 使用 AYLIEN 的新闻智能平台进行汇总、分析和丰富。https://aylien.com/resources/datasets/coronavirus-dataset,https: //aylien.com/blog/free-coronavirus-news-dataset

语料库	#引语	间接引语%	实体	数据来源
StylisticsCorpus	16,533	16	✗	小说、报纸、传记
PARC3	19,712	72	✗	《华尔街日报》
QuoteBank	1.78亿	-	✓	新闻文章
DirectQuote	10,279	0	✓	新闻文章
NEWSQUOTE	24,031	81	✓	新闻文章

Ta表 1：大规模（超过 10,000 条）源自新闻的英语引语语料库摘要。

我们利用新闻文章中句子的语义角色标注结果，提取引语触发动词、主体（即来源）和宾语（即引文），并根据相应的 DBpedia[3] 本体类标签识别来源。最终的数据集包含直接引语和间接引语，也包含混合引语，即只有部分引文置于引号内。我们提出了从新闻报道中寻找证据来源这一任务，并介绍了一系列方法，用于：（1）从文本中识别引文及其来源；（2）针对给定的新闻主题推荐潜在专家。实验结果表明，利用我们构建的 NEWSQUOTE 数据集开发自动化工具，以便记者和事实核查人员搜索并排序领域专家，是可行的。

2 相关工作

引语抽取与归属 引语抽取与归属最初是文学作品研究的一个方向（Zhang、Black 和 Sproat，2003），如今通常涵盖三个子任务：识别来源、抽取引文，以及将引文归属于其来源。表 1 汇总了若干基于新闻文章构建的大规模英语引文数据集。**StylisticsCorpus** （Semino 和 Short，2004）旨在研究英国叙事作品中书面语篇的呈现方式。由于硬新闻（例如事故、冲突和犯罪）（Bell，1991）具有传播广泛、叙事性强、真实性高且文化影响显著等特点，他们将其选作数据来源的一部分。在全部数据中，有 5407 个实例来自报刊。他们将这些样本分为言语、书写和思想三类，再将每一类细分为多种呈现类别，例如间接引语、自由间接引语、直接引语和自由直接引语。**PARC3** （Pareti，2016）项目旨在填补归属关系（AR）研究的空白。其标注方案对归属关系的三个构成要素进行了标注：来源、提示语和内容。他们根据引号的使用情况，将引语状态标为直接、间接或混合，并根据嵌套层级考察归属深度。**Quote-Bank** （Vaucher 等，2021）的构想源于当代新闻流错综复杂的特性。Vaucher 等（2021）利用不同媒体上的重复报道，学习引语-来源对的模式。他们聚焦于直接引语的归属，提出了一个名为 Quobert 的端到端弱监督框架，用于抽取并归属引语。他们使用 Quobert 从 Spinn3r 数据集（Burton 等，2011）生成 QuoteBank，并将来源实体链接到维基数据知识库。**DirectQuote** （Zhang 和 Liu，2021）包含从在线新闻媒体中人工标注的直接引语。与 QuoteBank 一样，其中每个来源都可以链接到维基数据中的命名实体，从而服务于各种下游任务。

在现有的新闻引语数据集中，Stylistic-Corpus 和 PARC3 同时包含直接引语和间接引语，但并非源自多平台新闻报道，也不提供指向维基数据的来源实体链接。另外两个数据集 QuoteBank 和 DirectQuote 均将其来源链接到维基数据中的命名实体，但只关注直接引语。相比之下，我们的 NEWSQUOTE 包含多种引语类型，包括直接引语、间接引语和混合引语；混合引语中，只有部分引文位于引号内。此外，所有来源都具有其 DBpedia 实体链接。

专家发现 专家发现的核心任务是针对给定查询，识别具备所需专业知识的候选项（Yuan 等，2020）。因此，相关解决方案着重于匹配搜索者的需求与相关专家的经验。在实践中，这一问题已扩展到不同情境，并需要考虑多种因素。学术领域占专家发现研究的比例高达 65%（Husain 等，2019）。寻找学术专家时，研究者会关注主题相关性、专家质量、研究关联性（Sun 等，2015；Silva，2014；Wang 等，2017），以及能力限制（Neshati、Beigy 和 Hiemstra，2014）。与此同时，许多专家发现系统应用于在线平台，例如社区问答、社交网络和论坛（Yuan 等，2020；Faisal、Daud 和 Akram，2017）。在企业领域，专家的可访问性和生产力被认为能带来显著的经济效益（Silva 等，2013；Paul，2016）。在医学领域，为特定患者寻找最合适的医生时，患者的潜在病情至关重要（Tekin、Atan 和 Van Der Schaar，2014）。在律师发现中，用户可能更倾向于选择同一州或城市的候选项，因此，地理位置尤为重要（Askari、Verberne 和 Pasi，2022）。

3 NewsQuote：数据集构建

本节将介绍数据集的构建过程，包括数据来源和预处理步骤的详细信息

[3]https://www.dbpedia.org/

以及测试集标注。最后将展示数据条目示例和数据集统计信息。

数据收集

我们基于 AYLIEN 冠状病毒数据集构建了 NEWSQUOTE 数据集，该数据集收录了 2019年11月至 2020年8月期间发布的内容。我们使用 AYLIEN 新闻 API[4] 检索新闻文章。除文本外，每篇文章还附带元数据，例如作者、关键词、摘要、来源、发布时间、依据互动广告局（IAB）分类体系[5] 和 IPTC 新闻代码[6], 编码的主题类别，以及从 DBpedia 提取的识别出的实体和实体链接。

预处理

数据去重由于同一篇新闻报道可能由多个来源发布，而且原始数据集中存在完全重复的内容，因此我们移除了与已发布新闻文章相似的新闻文章。首先按照时间顺序对新闻文章进行排序。随后，使用通过半监督学习、基于标题-正文对训练的 RoBERTa 分类器[7]检测重复新闻（R¨uckl´e、Moosavi 和 Gurevych，2019 年）。为提高处理效率，我们将数据集拆分为 16 个大小相同的子集。对于每个子集，按顺序将按时间排序的新闻文章的标题和新闻摘要的首句输入 RoBERTa 分类器。所有重复内容均被移除。数据去重后，剩余 158,325 篇新闻文章。新闻来源平台总数为 258 个，如图 A1b 所示，排名前 5 的新闻来源平台为：《每日邮报》、雅虎、Seeking Alpha、Business Insider 和路透社。

引语触发词筛选 对于每篇选定的文章，我们将正文划分为句子，然后使用基于预训练 BERT 的语义角色标注模型（Shi 和 Lin，2019年）提取动词（或谓词）、主体和宾语。我们根据动词的出现频率对其进行排序，得到一个候选动词列表。在人工检查出现次数超过 100 次的候选动词后，我们确定了 352 个更有可能表示直接引语或间接引语的引语触发词。动词列表详见我们的源代码仓库 [8]。其中一些动词明显表示引语，例如‘说’，而另一些动词则可能不属于传统意义上的引语相关动词，例如‘发推文’。确定引语触发词后，我们仅保留至少包含一个触发词、一个主体和一个宾语的句子。主体被视为潜在来源，而宾语则被视为潜在引语。为确保引文内容具有信息价值，我们还要求宾语的长度超过三个单词。

来源与引语筛选 我们要求候选句子的主语必须是人物或组织，因此通过数据集中附带的 DBpedia 本体标签 [9] 识别潜在来源实体。我们所选用的本体类见我们的源代码仓库 [10]。由于每个实体可能具有多个本体类，我们进一步删除了来源被标记为 位置、地点 和 国家的样本。由于同一主语可能被多次提及，我们使用 DBpedia 实体链接进行实体解析和规范化。此外，我们要求命名实体至少在数据集中出现两次。最后，为避免句子拆分错误，我们要求包含直接引语和混合引语的句子中的引号必须成对出现。

测试集标注

在实际应用中，给定一个主题，我们只能根据专家此前发表在新闻文章中的引语来识别专家，因此我们按照新闻文章的发布时间戳将数据集划分为训练集、验证集和测试集，并确保验证集和测试集中的引语-来源对出现时间晚于训练集中的对应内容。图 2 展示了引语-来源对根据其相关新闻文章发布日期的分布情况。[11]

为确保数据质量，一名标注员对测试集中的样本进行了人工筛查。表 A1 列出了原始测试集中出现的五类噪声及相应示例。属于这些噪声类别之一的数据均从测试集中删除。

数据集统计

我们的数据涵盖三类引语，如图1所示。简而言之，直接引语置于引号内，而间接引语不置于引号内；混合引语则只有部分引文置于引号内。我们根据引号的数量和位置，粗略估计了数据集中各类引语的占比：间接引语占81%，直接引语占10%，混合引语占9%。在测试集中，共有 1,867条（84%）间接引语、215条（10%）混合引语和 143条直接引语（6%）。表2展示了最终 NEWSQUOTE 数据集的统计信息。总的来说，我们共有 24,031对英语来源-引语对，涉及来自258个全球来源的 3,246个来源。更多相关统计数据和图表见附录A。

4 任务定义

在我们的数据集中，每个样本 S_i 由上下文 c_i、引文—来源对 $(q,\ e)_i$、关键词列表 k_i 以及元数据 m_i组成。上下文包含 3 个句子：来源和引文出现的主句、其前一句和后一句。关键词和元数据均在文档级别定义，并从

[4]https://aylien.com/product/news-api [5]https://www.iab.com [6]https://iptc.org/standards/newscodes/ [7]https://huggingface.co/vslaykovsky/roberta-news-duplicates [8]https://github.com/WenjiaZh/NewsQuote/blob/main/SelectedTriggerVerbs.csv

[9]http://mappings.dbpedia.org/server/ontology/classes/ [10]https://github.com/WenjiaZh/NewsQuote/blob/main/SelectedOntologyClasses.txt[11]原始数据集在2020年5月31日至2020年6月21日期间没有数据。

引文类型	句子
直接引语	“我希望这些地区的人们知道，我们不会落下任何一个人。” 福特说：“我们将在每周一提供各地区重新开放情况的最新消息。
间接引语	当 many 员工 已返回他们的 家乡 之后 the 宣布—— ment of 封锁，隔离，禁闭 丰田 says 在任何给定时点，只有 40% 至 45% 生产劳动力中的人员一直在岗工作，以维持各种形式的社交距离。
混合引文	好市多 已确认 to USA 《今日》 即“可能会有例外”，在某些 俱乐部，因为常规营业时间可能有所不同。

图 1：我们数据集中的三种引语类型。来源以蓝色突出显示，触发动词以红色突出显示，引语以黄色突出显示。

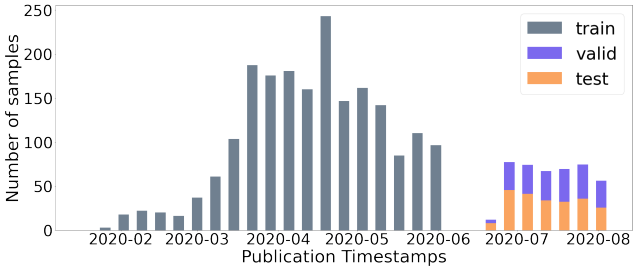


图 2：引语-来源对的分布。训练集包含 2020-01-19 至 2020-05- 31 发布的样本，验证集 / 测试集包含 2020 年6月21日 至 2020-08-02 发布的样本。

	Test	验证集	训练集
样本数	2,236	2,082	19,713
文章数	1,937	1,766	14,526
来源实体数 1,016	765	2,963	**平均引文长度** 28.38 29.16 28.99
新闻来源数 180	178	252	
新闻类别数 470	440	629	
每篇文章的平均关键词数	43.23	44.28	42.17

表2：NEWSQUOTE 数据集统计。

AYLIEN 冠状病毒数据集。我们在此 NEWSQUOTE 数据集上提出以下两个任务：

来源与引语抽取定义为从给定上下文c_i中自动提取来源-引语对（q，e）$_i$。

专家推荐 是指根据专家过去所说的内容，针对给定查询推荐一份按排名排序的专家列表。

5 种方法

我们介绍了来源与引语抽取和专家推荐的方法。图3展示了这些方法的概览。

来源与引语抽取

我们采用三种方法解决引语-来源对抽取问题：基于规则的方法、序列标注和问答。

方法 1：基于规则的引语标注器 可以使用类似正则表达式的规则来提取直接引语。我们在测试样本句子上运行来自斯坦福 CoreNLP（Manning 等人，2014 年）的引语标注器 [12]。它只能提取由引号界定的直接引语。

方法 2：序列标注 我们可以使用 5 类 BIO 标注方案为数据集中的每个样本添加标签。来源使用“B-S”和“I-S”标注，分别表示相应词元是否为来源提及的开头，或位于来源提及内部。同样，引语使用“B-Q”和“I-Q”标注，其他所有词元均标记为“O”。随后，我们对基于 BERT 的词元分类器（Devlin等人，2018年）进行微调，以从上下文中识别来源和引语。

方法3：问答（QA）流水线 我们使用问答流水线，通过依次提出两个问题来提取来源和引文：

问题一：来源是谁？问题二：［来源］说了什么？

在训练过程中，问题二中的［来源］是问题一的标准答案。在推理过程中，它是从问题一中抽取出的答案。输入上下文由一个问题、一条左侧句子、 l、一条主句、 s 和一条右侧句子r组成。为从上下文中抽取答案，我们对基于预训练BERT的抽取式问答模型（Devlin等人，2018年）进行了微调，其输入形式如下：

[CLS] 问题 [SEP] l [SEP] s [SEP]r [SEP]

专家推荐

我们可以将专家推荐表述为一个检索问题：给定一个查询，我们希望检索出能够评论查询中所讨论主题的来源，并按其与该查询的相关性排序。有两种可能的方法，一种是将来源过去的引语作为

[12]https://stanfordnlp.github.io/CoreNLP/quote.html

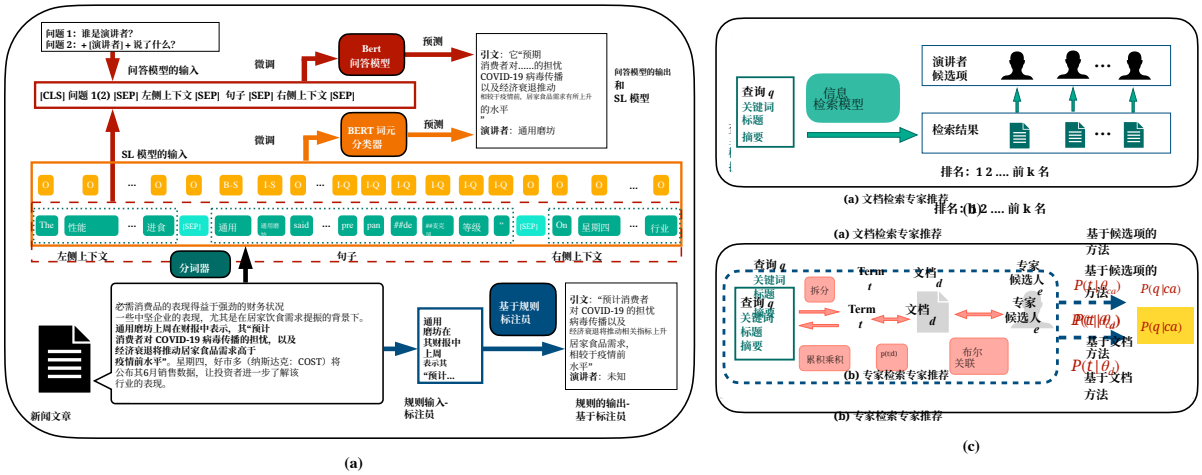


图 3：第 5 节中介绍的 5 种方法示意图。图 (a) 描述了问答流水线、序列标注以及用于引文来源抽取的基于规则的配额标注员。图 (b) 介绍了用于专家推荐的文档检索方法，图 (c) 展示了用于专家推荐的专家检索方法。

文档，执行文档检索，然后将检索到的引语的来源作为结果返回；另一种是直接执行专家检索。

方法 1：文档检索 文档检索旨在首先根据查询检索相关文档（即引文出现的上下文），然后从文档中提取来源作为结果。对于文档索引，我们通过 Pyserini[13]，分别使用稀疏词袋 Lucene 索引和四种 Transformer 编码的 Faiss 索引进行实验。随后，在稀疏索引上采用 BM25 排名方法，并在密集索引上进行最近邻搜索，以返回给定查询最相关的前 10 个文档。接着，将前 10 个检索结果文档中的来源确定为推荐的专家。

方法 2：专家检索 专家检索 根据给定候选来源 $P(q|e)$ 条件下查询的概率，直接引入来源。遵循 Balog、Azzopardi 和 de Rijke（2009）提出的框架，我们实现了基于候选项的方法和基于文档的专家发现方法。

基于候选项的专家检索 假设查询中的每个词项均以相同且相互独立的方式进行采样，同时假设文档与专家来源候选项在条件上相互独立，基于候选项的方法通过以下方式估计 $P(q|e)$:

$$P(q|e) = \prod_{t \in q} \{(1-\lambda)(\sum_{d \in D} p(t|d)p(d|e) + \lambda p(t)\}^{n(t,q)},$$

$$\lambda = \frac{\beta}{\beta + n(e)}, \quad \beta = \frac{\sum_E |\{d : n(e,d) > 0\}| \cdot |d|}{|E|},$$

其中，λ 是平滑参数，$p(t|d)$、$p(d|e)$和 $p(t)$分别是词项 t 在文档 d中的条件概率，

[13]https://github.com/castorini/pyserini

文档 d 给定来源 e的条件概率，以及词项 t的概率。 $p(t|d)$和$p(t)$均通过最大似然法估计。概率$p(d|e)$由布尔模型设定，稍后将对此进行讨论。 $|d|$ 是平均文档长度，$n(t,q)$是词项 t 在查询 q中出现的次数， $n(e,d)$是专家 e 在文档 d中出现的频率，而 $n(e)$是与来源 e相关联的文档中出现的总次数。

基于文档的专家检索 基于文档的专家检索方法通过相关文档集合搜索来源。该方法假设查询与候选项之间具有条件独立性，并估计词项 t 在每个文档中的概率:

$$P(q|e) = \sum_{d \in D} \{\prod_{t \in q} ((1-\lambda)p(t|d) + \lambda p(t))^{n(t,q)}\} p(d|e),$$

$$\lambda = \frac{\beta}{\beta + n(d)}, \quad \beta = |d|,$$

其中， $n(d)$ 表示文档 d 的长度。

在基于候选项和基于文档的专家发现方法中，文档与候选项之间的关联 $p(d|e)$ 通过简单的布尔模型进行估计：当 $n(e, d) > 0$ 时设为 1，否则设为 0。

6 实验

实验设置

对于基于规则的方法，我们将原始句子直接输入引语标注器。为了构建词元分类器，我们将输入文本分割为长度为512的词元序列，并以2e-7的初始学习率对模型进行100个训练集周期的微调。对于抽取式问答模型，在询问来源时，将抽取答案的最大长度设为30；在询问引语时，设为512。对于关于来源的问题，我们以2e-6的初始学习率对模型进行50个训练集周期的训练。对于关于引语的问题，我们

训练模型 100 个训练周期，初始学习率为 2e-5。

在专家推荐任务中，我们考虑两种类型的文档：来源/引文出现的主句，或主句及其周围上下文（即前后句）。作为专家检索的查询，我们使用新闻文章的标题、关键词或摘要的第一句。为进一步消除干扰，如果输入查询中包含来源名称，我们会将其删除。对于专家检索方法，我们仅取新闻文章标题（关键词列表或新闻摘要的第一句）的前 w 个词作为输入查询，以缩短运行时间。在验证 w 介于 1 到 10 之间后，我们最终设定 $w = 5$。

评估指标

为了衡量模型在引语提取和归因任务上的表现，我们采用 SQuAD（Rajpurkar 等，2016）中定义的两项指标：精确匹配和宏平均 F1 分数。**精确匹配**在预测结果与真实值完全一致时等于 1，而**（宏平均）F1**用于衡量预测答案与真实答案在词元级的平均重合度。

对于专家推荐，我们采用信息检索中常用的两种指标：平均准确率（MAP）和归一化折损累计增益（NDCG）。**平均准确率** 是检索到相关文档的位置上的准确率的平均值。**K 位的归一化折损累计增益**首先将第 i 位的增益按 $1\ \log_2(i)$ 进行折损，然后累加至第 k 位的转换后增益，最后根据理想排序进行归一化。此外，我们提出了**宽松指标**：如果检索到的专家与真实来源属于同一聚类，则将其视为相关。在构建宽松指标时，我们选取出现频率最高的 100 个来源 DBpedia 类别，并使用二进制向量嵌入来源[14]。随后，我们对来源嵌入进行 k-均值聚类。我们根据聚类一致性和分离度分数，通过经验设定 $k = 40$。

实验结果

我们首先介绍第 5 节中描述的三种引语提取与归因方法的结果，随后介绍第 5 节中提出的两种专家推荐方法的评估结果。

引语提取与归因 表 3 展示了基于规则的标注器、序列标注和问答流水线在测试集上的性能。基于规则的引语标注器表现最差，这并不令人意外，因为它只能使用类似正则表达式的规则提取直接引语。在我们的测试集中，2225 个样本中只有 337 个被识别为包含引语。在这个子集中，基于规则的标注器取得了较高的 49.65 精确匹配分数。与引文相比，来源抽取的表现更好。但在宏平均 F1 方面，直接引语抽取的表现远优于来源抽取。对于另外两类引语，即间接引语和混合引语，基于规则的标注器基本无法产生任何合理的结果。与基于规则的标注器相比，序列标注取得了好得多的结果。我们注意到，从精确匹配的角度来看，引语提取的表现似乎比来源抽取低近 10%，这表明模型在提取较长答案时遇到了困难。在引语的三种类别中，模型对直接引语的提取效果最好，其次是间接引语，而对混合引语的提取效果最差。这是意料之中的，因为与另外两类引语相比，混合引语的处理复杂得多。问答流水线在识别来源和提取引文这两项任务中均取得了最佳表现。在测试问答流水线的引语提取能力时，我们通过以下三种方式进行了实验：在询问引文的问题中使用真实来源名称、使用 QA_{source} 的结果中预测出的来源，或者使用代词“他们”替换来源，以完全从问题中移除来源信息。由于我们的问答模型在来源识别方面的准确率已经很高，因此，在引语提取问题中使用真实来源或预测来源并不会带来太大差异。然而，如果来源信息缺失，引语提取的表现就会下降近 2 个百分点的宏平均 F1，精确匹配则下降超过 4 个百分点。

专家推荐 表4展示了以新闻文章中的关键词作为查询、以引文上下文（来源和引文出现的主句，以及其前后句子）作为文档时的专家推荐结果。可以观察到，文档检索（**文档检索**）方法通常优于专家检索（**专家检索**）方法。在各种文档索引策略中，使用 Lucene 稀疏词袋索引（**文档检索**$_{sparse}$）的效果优于其他密集型 Transformer 编码的 Faiss 索引。正如预期的那样，采用宽松指标时，如果检索到的来源与真实来源位于同一簇中，就将其视为相关来源；与严格指标相比，我们获得了更好的结果。[15]

7 挑战与未来方向

我们介绍了 NewsQuote 数据集，并提出了一项新任务，即新闻与事实核查领域中的专家推荐。我们的实验证实，可以利用问答流程提取引语-来源对，也可以通过文档检索和专家检索找到专家来源。在此，我们概述了若干潜在的未来方向。

首先，在构建我们的数据集时，我们从出现频率最高的一组动词中人工挑选了引语触发动词。一方面，识别出的动词列表并未涵盖所有可能表示引文的动词，例如那些出现频率较低或与新冠肺炎主题关联不密切的动词；

[14]在我们的数据集中，每个来源平均被分配到 4 至 5 个 DBpedia 类别。

[15]采用其他文档检索或专家检索方法，并基于文档和查询的不同组合方式得出的结果见附录 C。

	总体		直接引语		间接引语		混合引语	
	宏平均 F1	精确匹配	宏平均 F1	精确匹配	宏平均 F1	精确匹配	宏平均 F1	精确匹配
Rule$_{source}$	5.76	5.62	50.58	49.65	0.214	0.214	24.11	23.26
Rule$_{quote}$	7.72	1.93	82.33	30.07	0.145	0.00	23.84	0.00
SL$_{source}$	98.06	95.37	98.63	95.80	97.99	95.34	98.23	95.35
SL$_{quote}$	95.65	85.17	**97.17**	89.51	95.61	85.11	95.05	82.79
问答$_{speakr}$	**98.86**	**98.61**	**99.30**	**99.30**	**98.77**	**98.50**	**99.38**	**99.07**
问答$_{quote}$								
w/ true source	**95.96**	**90.74**	95.83	**93.01**	**95.96**	**90.31**	**96.06**	**93.02**
w/ pred. source	95.61	89.93	95.78	93.01	95.55	89.34	96.06	93.02
w/ source mask	93.92	85.84	96.53	90.21	93.56	85.11	95.28	89.30

表 3：测试集上的来源和引语抽取结果。**规则** – 基于规则的标注器，**序列标注** – 序列标注，**问答** – 问答流程。下标表示模型的目标，即用于 *source* 抽取或用于 *quote*抽取。在问答*quote*中，“*w/ true source*”表示提问“说了什么？ + *[* 来源*]* +”时使用真实来源名称；“”使用问答结果中的预测来源，而“”使用泛称“他们”。

	严格指标			宽松指标		
	MAP	NDCG $_5$	NDCG$_{10}$	MAP	NDCG $_5$	NDCG$_{10}$
DR$_{sparse}$	**0.2903**	**0.2807**	**0.3590**	**0.4162**	**0.3925**	**0.5183**
文档检索$_{flat}$	0.1481	0.1440	0.1939	0.2886	0.2714	0.3887
文档检索$_{hnswpq}$	0.1509	0.1473	0.1926	0.2966	0.2805	0.3956
文档检索$_{hnsw}$	0.1446	0.1406	0.1889	0.2865	0.2686	0.3850
DR$_{pq}$	0.1395	0.1363	0.1838	0.2739	0.2583	0.3734
ER$_{can}$	0.1021	0.1106	0.1252	0.2306	0.2294	0.3135
专家检索$_{doc}$	0.1205	0.1281	0.1418	0.2465	0.2412	0.3285

表4：使用引文上下文作为文档、新闻文章关键词作为查询的专家推荐结果。在前五行中，**文档检索** 表示文档检索方法，下标代表第 5 节方法 1 中提到的 5 种检索索引：Lucene 稀疏词袋索引、Faiss 平面索引、Faiss HNSW PQ 索引、Faiss HNSW 索引和 Faiss PQ 索引。**专家检索**$_{can}$ 表示基于候选人的专家发现方法，**专家检索**$_{doc}$ 表示基于文档的专家发现方法。在基于文档的专家发现方法中，输入查询长度设为 5 个关键词。

另一方面，有些动词存在歧义，需要结合上下文判断它们是否确实是触发词。尽管我们在检查测试集时剔除了明确不属于此类的情况，但要对如此大规模的数据进行人工筛选并不现实。未来的工作可以探索利用其他大规模引语语料库训练模型，以检测引语触发词的可能性。此外，我们的数据集是根据有关冠状病毒的新闻文章构建的。未来可以将其扩展到更多主题，例如商业、技术、教育和政治。

其次，共指消解对于增加引语来源归属数据至关重要，因为在新闻文章中，常使用代词指代先前提到的来源。我们关于共指消解的初步实验得出了嘈杂的引语来源归属结果。未来的工作可以利用代词关联引文与候选来源引文之间的内容相似度和／或连贯性，来改进共指消解结果。

第三，将 DBpedia 链接视为我们数据集中来源的标识，即可引入外部知识作为证据，以提升专家推荐的效果。

第四，我们的框架使新闻编辑部能够建立引文—来源库，以协助评估信息的真实性。该库可向记者和事实核查人员提供每个来源所发表言论的摘要，包括哪些人曾引用该来源、引用时间以及引用涉及哪项真实性核查，从而减少重复劳动并促进协作。

最后，记者和事实核查人员务必注意，不要过度依赖我们在此介绍的工具（即避免陷入所谓的“自动化偏见”）。因此，必须谨慎解读结果，且应始终由记者或事实核查人员最终决定联系哪些专家。因此，此类模型提供的推荐依据必须能够供用户评估其可信度和相关性（Procter 等人（2023））。

8 结论

我们介绍了如何构建一个新颖的大规模数据集，其中包含从新闻文章中检索得到的引语-来源对。我们的 NEWSQUOTE 数据集由直接引语、间接引语及其组合构成。引语类型的多样性将促进更先进方法的发展，以应对间接引语和混合引语提取这一具有挑战性的任务。基于 NEWSQUOTE 数据集，我们证明了问答流水线在来源抽取任务中的精确匹配率可超过 98%，在引语提取任务中的精确匹配率接近 90%。此外，我们还引入了专家推荐任务，并表明，与其他稠密检索方法相比，采用稀疏索引的文档检索方法能够取得最佳效果。

伦理声明

我们使用的所有数据均来自公开来源。我们已获得 Aylien 的书面同意，可以下载其数据。根据数据所有者的要求，我们不会直接分享下载的数据，而是会分享下载脚本和所有预处理脚本，以便他人能够从 Aylien 的网站获取与本文所使用的相同数据集。

致谢

本研究部分得到了 EPSRC（资助编号：EP/V048597/1）的支持。YHis 获得了由 UKRI 资助的图灵人工智能奖学金（资助编号：EP/V020579/2）。

参考文献

Askari, A.；Verberne, S.；Pasi, G. 2022。《法律社区问答中的专家发现》。*arXiv* 预印本 *arXiv:2201.07667*。Balog, K.；Azzopardi, L.；de Rijke, M. 2009。用于专家发现的语言建模框架。《信息处理与管理》，45(1)：1–19。Bell, A. 1991。《新闻媒体的语言》。牛津 Blackwell 出版社。Burton, K.；Java, A.；Soboroff, I.；等。2011。ICWSM 2011 Spinn3r 数据集。载于第五届年度博客与社交媒体会议论文集（*ICWSM 2011*）。Devlin, J.；Chang, M.-W.；Lee, K.；Toutanova, K. 2018。BERT：用于语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*。Elson, D. K.；McKeown, K. R. 2010。文学叙事中引语言语的自动归属。载于第二十四届 *AAAI* 人工智能大会。Faisal, M.；Daud, A.；Akram, A. 2017。利用共存用户的声誉和答案质量进行专家排名。《阿拉伯国际信息技术期刊》（*IAJIT*），14(1)。Fernandes, W. P. D.；Motta, E.；Milidi´u, R. L. 2011。葡萄牙语引语抽取。载于第八届巴西信息与人类语言技术研讨会论文集。

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. 自动事实核查综述。计算语言学协会汇刊，10：178–206。Husain, O.; Salim, N.; Alias, R. A.; Abdelsalam, S.; and Hassan, A. 2019. 专家发现系统：系统性综述。应用科学，9(20)：4250。Lazarski, E.; Al-Khassaweneh, M.; and Howard, C. 2021. 使用自然语言处理进行事实核查：综述。*Designs*，5(3)：42。Lee, J.; and Yeung, C. Y. 2016. 直接引语标注语料库。在第十届国际语言资源与评估会议（*LREC'16*）论文集，1059–1063。斯洛文尼亚波尔托罗日：欧洲语言资源协会（ELRA）。Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; and McClosky, D. 2014. 斯坦福 CoreNLP 自然语言处理工具包。在计算语言学协会第*52*届年会论文集：系统演示，55–60。马里兰州巴尔的摩：计算语言学协会。Neshati, M.; Beigy, H.; and Hiemstra, D. 2014. 使用设施选址分析进行专家组组建。信息处理与管理，50(2)：361–383。O'Keefe, T.; Pareti, S.; Curran, J. R.; Koprinska, I.; and Honnibal, M. 2012. 引语归属的序列标注方法。在*2012*年自然语言处理实证方法会议与计算语言学习联合会议论文集，790–799。Pareti, S. 2016. PARC 3.0：归因关系语料库。在第十届国际语言资源与评估会议（*LREC'16*）论文集，3914–3920。Paul, S. A. 2016. 寻找专家：为正式帮助提供设计专家选择界面。在*2016*年计算机系统中的人为因素会议论文集，3038–3048。Procter, R.; Arana-catania, M.; He, Y.; Liakata, M.; Zubiaga, A.; Kochkina, E.; and Zhao, R. 2023. 关于事实核查工作的一些观察及其对计算支持的启示。在国际 *AAAI* 网络与社交媒体会议论文集。AAAI 出版社。Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD：100,000+个机器阅读理解问题。在*2016*年自然语言处理实证方法会议论文集，2383–2392。得克萨斯州奥斯汀：计算语言学协会。R¨uckl´e, A.; Moosavi, N. S.; and Gurevych, I. 2019. 无需标注训练数据的神经重复问题检测。在*2019*年自然语言处理实证方法会议与第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，1607–1617。中国香港：计算语言学协会。Semino, E.; and Short, M. 2004. 语料库文体学：英语书写语料库中的言语、书写与思想呈现。Routledge。

Shi, P.; and Lin, J. 2019. 用于关系抽取和语义角色标注的简单 BERT 模型。*arXiv* 预印本 *arXiv:1904.05255*。Silva, A. T. P. 2014. 面向研究社交网络中专家推荐的研究分析框架。博士论文，香港城市大学。Silva, T.; Guo, Z.; Ma, J.; Jiang, H.; and Chen, H. 2013. 社交网络赋能的项目选择研究分析框架。决策支持系统，55(4)：957–968。Sun, J.; Xu, W.; Ma, J.; and Sun, J. 2015. 利用 RAF 在研究社交网络服务中发现领域专家：一种基于 MapReduce 框架的大数据分析方法。国际生产经济学杂志，165：185–193。Tekin, C.; Atan, O.; and Van Der Schaar, M. 2014. 发现专家：面向医疗诊断的上下文自适应专家选择。*IEEE* 新兴计算主题汇刊，3(2)：220–234。Vaucher, T.; Spitz, A.; Catasta, M.; and West, R. 2021. QuoteBank：十年新闻引文语料库。在第*14*届 *ACM* 网络搜索与数据挖掘国际会议论文集，328–336。Wang, Q.; Ma, J.; Liao, X.; and Du, W. 2017. 面向高校与产业研发项目协作的上下文感知研究人员推荐系统。决策支持系统，103：46–57。Yuan, S.; Zhang, Y.; Tang, J.; Hall, W.; and Cabot`a, J. B. 2020. 社区问答中的专家发现：综述。人工智能评论，53(2)：843–874。Zeng, X.; Abumansour, A. S.; and Zubiaga, A. 2021. 自动事实核查：综述。语言与语言学指南，15(10)：e12438。Zhang, J. Y.; Black, A. W.; and Sproat, R. 2003. 识别儿童故事中的发言者以进行语音合成。在第八届欧洲语音通信与技术会议。Zhang, Y.; and Liu, Y. 2021. DirectQuote：新闻文章中的直接引语抽取与归属数据集。*arXiv* 预印本 *arXiv:2110.07827*。Zhao, R.; Arana-Catania, M.; Zhu, L.; Kochkina, E.; Gui, L.; Zubiaga, A.; Procter, R.; Liakata, M.; and He, Y. 2023. PANACEA：面向 COVID-19 的自动错误信息检测系统。*arXiv* 预印本 *arXiv:2303.01241*。

A 数据统计

图 A1a 展示了 25 个最常见的新闻文章类别。平均而言，每篇文章有 3.92 个类别标签。排名前五的类别是：“个人理财”、“股票”、“商业”、“法律、政府与政治”和“旅游”。

图 A1b 展示了 25 个最常见的新闻来源。在我们的数据集中，新闻来源平台总数为 258 个。排名前五的新闻来源平台是：《每日邮报》、雅虎、Seeking Alpha、Business Insider、路透社。

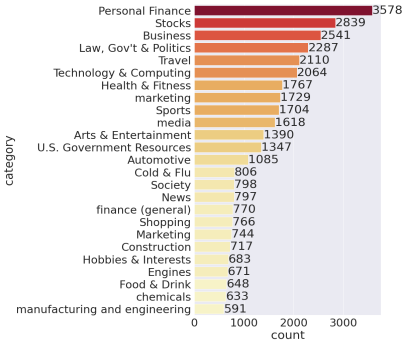
图 A1c 列出了 25 个最常见的来源。我们总共有 3246 个来源。排名前 5 的来源是：Amazon、苹果、Google、Microsoft、路透社。

B 噪声样本

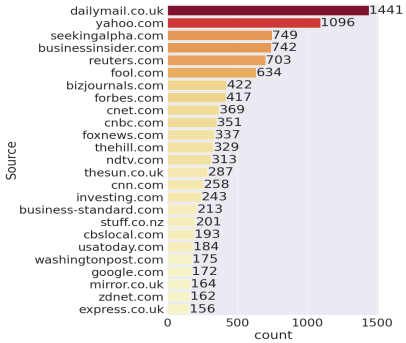
原始测试集中出现的五种噪声及其对应示例列于表 A1。

C 结果：专家推荐

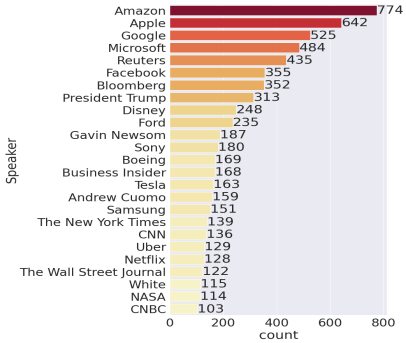
表 A2 展示了专家推荐的实验结果。在全部 6 种文档-查询组合中，文档检索（**文档检索**）方法的表现优于专家检索（**专家检索**）方法。在各种文档索引策略中，Lucene 稀疏词袋索引（**文档检索**$_{sparse}$）的效果优于其他密集型 Transformer 编码的 Faiss 索引。总体而言，使用上下文作为文档时，实验效果更好。我们认为，相邻上下文也可能包含有关来源及其引文的信息。



(a) 频率最高的新闻文章类别。计数表示相应类别中的文章数量。



(b) 热门新闻来源。计数表示相应新闻来源发布的文章数量。



(c) 引文-来源配对数据集中出现频率最高的来源。计数表示来自相应来源的引文数量。

图 A1

噪声类型	**示例文本**
乱序文本	-=-=-= +++导入文本 最后 八月， 苹果宣布，它 将 [分发给特别人士 iPhone](https://www.
引文标注错误 （仅加粗的文字被标记为引文）	此外，CBS 援引**了解情况的现任和前任外交官**报道称， 自2019年5月获提名以来，Gunter 营造了一个日益“难以为继”的工作环境， 他动辄“勃然大怒”，并随意更换使团副团长。
不恰当的触发词	”她说：“这些警告从一开始就不可避免，这也正是为什么移民及海关执法局（ICE） 本来就应该、今后也应该**释放**这些人，尤其是那些在医疗方面易受伤害的人， 以预防人道主义灾难。”
不恰当的来源	我们与多家公司保持专业的企业关系，这有助于我们挖掘 **市场数据**帮助我们生成准确的研究数据表，并**确认其中的最高准确性** **我们的市场预测**。
这不是肯定性陈述。	苹果**未回应**置评请求。

表 A1：从测试集中移除的噪声样本类型。

文档查询		方法	严格指标 MAP	NDCG$_5$	NDCG$_{10}$	宽松指标 MAP	NDCG$_5$	NDCG$_{10}$
上下文	关键词	**DR**$_{sparse}$	**0.2903**	**0.2807**	**0.3590**	**0.4162**	**0.3925**	**0.5183**
		文档检索$_{flat}$	0.1481	0.1440	0.1939	0.2886	0.2714	0.3887
		文档检索$_{hnswpq}$	0.1509	0.1473	0.1926	0.2966	0.2805	0.3956
		文档检索$_{hnsw}$	0.1446	0.1406	0.1889	0.2865	0.2686	0.3850
		DR$_{pq}$	0.1395	0.1363	0.1838	0.2739	0.2583	0.3734
		ER$_{can}$	0.1021	0.1106	0.1252	0.2306	0.2294	0.3135
		专家检索$_{doc}$	0.1205	0.1281	0.1418	0.2465	0.2412	0.3285
上下文	标题	**DR**$_{sparse}$	0.1357	0.1367	0.1719	0.3112	0.2980	0.4048
		文档检索$_{flat}$	0.1321	0.1304	0.1670	0.2979	0.2800	0.3917
		文档检索$_{hnswpq}$	0.1206	0.1177	0.1538	0.2830	0.2659	0.3790
		文档检索$_{hnsw}$	0.1311	0.1299	0.1659	0.2963	0.2787	0.3902
		DR$_{pq}$	0.1231	0.1223	0.1570	0.2931	0.2797	0.3873
		ER$_{can}$	0.1062	0.1116	0.1289	0.2516	0.2485	0.3366
		专家检索$_{doc}$	0.1111	0.1183	0.1320	0.2532	0.2518	0.3387
上下文	摘要	**DR**$_{sparse}$	0.1498	0.1484	0.1884	0.3285	0.3115	0.4297
		文档检索$_{flat}$	0.1435	0.1389	0.1818	0.3125	0.2945	0.4089
		文档检索$_{hnswpq}$	0.1285	0.1254	0.1648	0.2969	0.2761	0.3953
		文档检索$_{hnsw}$	0.1419	0.1373	0.1803	0.3097	0.2907	0.4056
		DR$_{pq}$	0.1402	0.1418	0.1773	0.3103	0.2953	0.4098
		ER$_{can}$	0.0918	0.0966	0.1157	0.2461	0.2419	0.3337
		专家检索$_{doc}$	0.1014	0.1069	0.1235	0.2570	0.2542	0.3479
引文	关键词	**DR**$_{sparse}$	0.1356	0.1366	0.1730	0.3059	0.2938	0.4074
		文档检索$_{flat}$	0.0721	0.0724	0.0999	0.2279	0.2127	0.3238
		文档检索$_{hnswpq}$	0.0756	0.0746	0.1027	0.2501	0.2410	0.3458
		文档检索$_{hnsw}$	0.0713	0.0719	0.0984	0.2266	0.2124	0.3214
		DR$_{pq}$	0.0692	0.0684	0.0931	0.2161	0.2036	0.3047
		ER$_{can}$	0.0685	0.0720	0.0815	0.1902	0.1902	0.2633
		专家检索$_{doc}$	0.0552	0.0588	0.0682	0.1685	0.1647	0.2404
引文	标题	**DR**$_{sparse}$	0.0955	0.0956	0.1204	0.2713	0.2614	0.3600
		文档检索$_{flat}$	0.0953	0.0958	0.1227	0.2744	0.2611	0.3729
		文档检索$_{hnswpq}$	0.0861	0.0865	0.1113	0.2683	0.2526	0.3654
		文档检索$_{hnsw}$	0.0943	0.0946	0.1209	0.2751	0.2620	0.3724
		DR$_{pq}$	0.0934	0.0947	0.1188	0.2674	0.2565	0.3615
		ER$_{can}$	0.0644	0.0669	0.0785	0.1951	0.1935	0.2725
		专家检索$_{doc}$	0.0558	0.0582	0.0678	0.1851	0.1832	0.2598
引文	摘要	**DR**$_{sparse}$	0.1021	0.1037	0.1306	0.2837	0.2739	0.3784
		文档检索$_{flat}$	0.1058	0.1074	0.1338	0.2902	0.2771	0.3895
		文档检索$_{hnswpq}$	0.0941	0.0911	0.1258	0.2912	0.2718	0.3933
		文档检索$_{hnsw}$	0.1038	0.1056	0.1309	0.2882	0.2756	0.3866
		DR$_{pq}$	0.1023	0.1013	0.1293	0.2882	0.2750	0.3854
		ER$_{can}$	0.0568	0.0585	0.0717	0.2002	0.1958	0.2778
		专家检索$_{doc}$	0.0457	0.0473	0.0569	0.1850	0.1799	0.2627

表 A2：**文档检索** 表示文档检索方法，下标表示第 5 节方法 1 中提到的 5 种检索索引：Lucene 稀疏词袋索引、Faiss 平面索引、Faiss HNSW PQ 索引、Faiss HNSW 索引和 Faiss PQ 索引。**专家检索**$_{can}$ 表示基于候选人的专家发现方法，**专家检索**$_{doc}$ 表示基于文档的专家发现方法。两种专家发现方法的输入查询长度均设为 5。