arXiv:2305.03507v1 [cs.CL] 2 May 2023

再读两遍：通过重新审视证据实现忠实且可解释的事实核查

Xuming Hu Zhaochen Hong Zhijiang Guo 清华大学 清华大学 剑桥大学
hxm19@mails.tsinghua.edu.cn hongzc20@mails.tsinghua.edu.cn zg283@cam.ac.uk

Lijie Wen 清华大学
w
enlj@tsinghua.edu.cn

Philip S. Yu 伊利诺伊大学
芝加哥分校
psyu@cs.uic.edu

摘要

现实世界中的事实核查任务旨在通过从源文档中检索证据来验证主张的事实性。检索到的证据质量对主张验证至关重要。理想情况下，检索到的证据应当忠实 （反映模型在主张验证中的决策过程），并且合理 （能够令人信服），同时提升核查任务的准确性 。尽管现有方法利用主张与文档在语义或表层形式上的相似度来检索证据，但它们都依赖某些启发式方法，因此无法满足全部三项要求。为此，我们提出了一种名为 ReRead 的事实核查模型，用于检索证据并验证主张，具体做法是：(1) 训练证据检索器以获取可解释的证据（即满足忠实性和合理性标准）；(2) 训练主张验证器重新审视经优化的证据检索器所检索到的证据，以提高准确性。在不同设置下，该系统均显著优于此前报告的最佳模型。

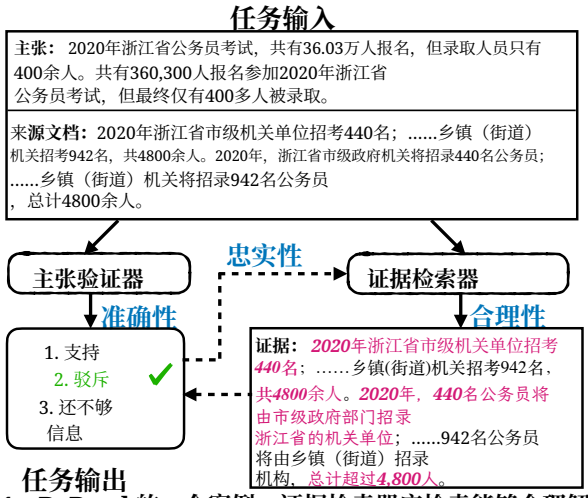

图1：ReRead 的一个案例。证据检索器应检索能够合理解释验证结果为何为“驳斥”，并反映验证器决策过程的证据。通过训练证据检索器，它可以为验证器提供更好的证据，以便重新审视并提高事实核查任务的准确性。

CCS 概念

• **信息系统** → **信息系统应用**。

关键词

自动事实核查，现实世界系统，潜变量模型，证据检索，主张验证

ACM 参考格式：
Xuming Hu、Zhaochen Hong、Zhijiang Guo、Lijie Wen 和 Philip S. Yu。2023年。《Read it Twice：通过重新审视证据实现忠实且可解释的事实核查》。载于 第*46*届 *ACM* 信息检索研究与开发国际会议论文集（*SIGIR '23*），*2023*年*7*月*23*日至*27*日，中国台湾台北。 ACM，纽约州纽约市，美国，5页。https://doi.org/10.1145/3539618.3592049

1 引言

错误信息的传播已成为当今社会的一大问题，尤其是在数字时代，信息可以轻易地通过各种平台传播和分享[3, 24, 33]。因此，事实核查已成为一项至关重要的任务

通过评估书面或口头语言中主张的事实性来应对这一问题 [1, 4, 7, 22, 32]。为实现这一目标，必须有恰当的证据来支持或反驳主张。因此，如何从大量源文档中检索合适的证据，是事实核查的关键环节。

如图1所示，我们通过谷歌搜索引擎检索了一个来自中国社交媒体的真实世界主张及其对应的源文档。我们需要从噪声文档中检索忠实的 （反映验证器在主张验证中的决策过程）和合理的 （解释主张事实性的原因）证据，以提高主张验证任务的准确性 [8, 36]。在本例中，需要检索诸如“超过4800人”这样的证据，以反驳“仅有400多人”的主张。尽管证据在事实核查中发挥着至关重要的作用，但早期的自动化事实核查尝试忽视了这一点，仅依赖主张的表层模式进行验证，而忽略了证据所提供的信息[25, 31]。因此，这些方法无法识别经过精心伪装的错误信息 [26]。近期，为解决这一问题，研究人员尝试邀请标注员通过修改维基百科文章中的句子来创建主张和证据 [2, 28]。然而，

本作品在知识共享署名国际 4.0 许可协议授权。*SIGIR '23*，*2023*年7月 [illegible] 湾台北。2023年版权归所有者/作者所有。ACM ISBN 978-1-4503-9408-6/23/07。https://doi.org/10.1145/3539618.3592049

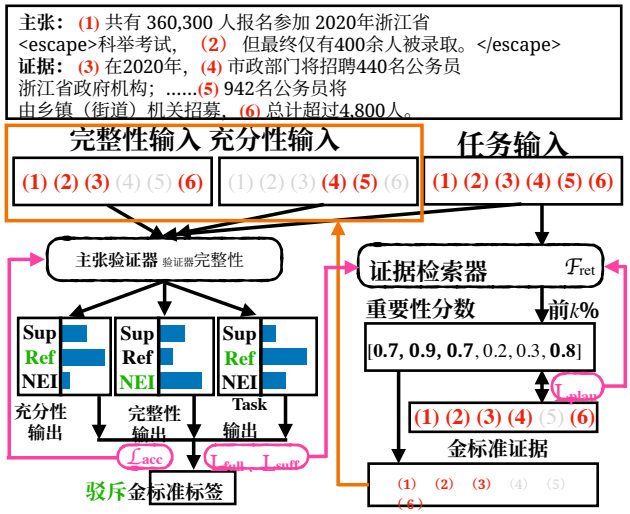

图2：ReRead 的架构。

这些基于维基百科生成的合成主张无法替代媒体生态系统中传播的现实世界主张。因此，其他研究转而从事实核查网站抓取主张，并利用搜索引擎查找支持性文档 [9, 10, 34]。然而，通过这种方式检索到的源文档往往包含大量噪声，从而影响核查任务的准确性。为了解决这一问题，Hu 等人 [10] 通过衡量主张与证据之间的语义相似度，从源文档中检索相关证据；Gupta 和 Srikumar [9] 则开发了一种基于注意力机制的证据聚合模型。然而，这些方法都依赖于特定的启发式方法，无法同时满足忠实性、合理性以及提高事实核查准确性这三项要求。

我们提出了新颖的真实世界事实核查模型 ReRead，通过以下方式满足三个关键要求：(1) 基于忠实性和合理性标准，训练用于获取可解释证据的证据检索器；(2) 训练主张验证器，对优化后的检索器获取的证据进行重新评估，从而提升准确性。如图1所示，ReRead 首先利用标注数据对验证器进行微调，然后借助验证器帮助检索器获取忠实证据。检索器还利用金标准证据来提升合理性。经过训练的检索器提供的改进证据，使验证器能够进一步优化准确性。我们的主要贡献包括：(1) 提出一种用于检索忠实且合理证据的新颖模型，从而提高验证准确性；(2) 实验表明，在真实世界数据集上，该模型的 F1 性能比当前最先进的基线提升了 4.31%，并通过广泛分析验证了 ReRead 的有效性。

2 训练目标分析

我们有三个训练目标：（1）检索到的证据需要具备**忠实性**，即证据在多大程度上准确反映了验证器预测验证标签 [14] 的真实推理过程。我们采用两个指标：**完整性**反映从源文档中移除证据后预测标签概率的变化。**充分性**反映仅使用证据预测标签时的概率变化；换句话说，如果证据确实具有影响力，标签的概率就不会发生显著变化。（2）检索到的证据需要具备**合理性**，以使验证器相信其预测结果 [6]。我们采用金标准证据来训练检索到的证据。（3）通过重新审视检索到的证据，需要提高任务的**准确性**。

3 模型架构

如图2所示，ReRead 首先利用标注数据，通过 L_{acc} 微调主张验证器。ReRead 利用金标准证据来评估检索证据的合理性（L^{plau}），并利用金标准标签来评估证据的忠实性（L^{full} 和 L^{suff}）。

3.1 句子编码器

我们采用 BERT 编码器 [5] 来获取主张和源文档中每个句子的语义嵌入。对于给定的主张 C 及其对应的源文档 D，我们通过在每个句子的开头添加特殊标记 [CLS]，并利用 [CLS] 位置嵌入来获取它们的句子嵌入。这将为主张和文档生成一个嵌入矩阵 $S_{emb} \in \mathbb{R}^{l \times d}$，其中 l 表示句子的总数，且 $d = 768$。

3.2 主张验证器

我们的主张验证器以 S_{emb} 为输入，并将主张分为三类：驳斥（Ref）、支持（Sup）和信息不足（NEI）。训练过程中，验证器根据主张和文档进行分类。

我们使用基于神经网络的分类器 F_{ver} 来实现这一目标。它以 S_{emb} 为输入，并输出概率预测向量 $F_{ver}(S_{emb}) = (p_{Sup}, p_{Ref}, p_{NEI})^\top$，其中 p_{Sup}、p_{Ref}和 p_{NEI}分别表示主张属于支持（Sup）、驳斥（Ref）或信息不足（NEI）的概率。我们将验证结果记为随机变量 v。

3.2.1 **准确性**。我们采用准确性这一准则来训练主张验证器执行主张验证。为评估其性能，我们使用交叉熵损失 $L_{CE}(F_{ver}(S_{emb}), y^*)$，该损失计算验证器的概率预测 $F_{ver}(S_{emb})$ 与真实标签 $y^* \in \{0, 1, 2\}$ 之间的差异，其中标签分别表示驳斥（Ref）、支持（Sup）和信息不足（NEI）。因此，我们将准确性损失函数定义如下：

$$\mathcal{L}_{acc} = \mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}), y^*), \quad (1)$$

用于训练主张验证器和句子编码器。

3.3 证据检索器

主张验证器训练完成后，将使用训练好的验证器和数据集中的黄金证据来训练证据检索器，以提高检索证据的忠实性并确保其合理性。经过优化的证据还能进一步提升验证性能。为此，我们使用基于神经网络的分类器 F_{ret}，并利用句子编码器的输出获取语义信息。具体而言，F_{ret} 以句子编码器的输出S_{emb} 为输入，并输出向量 $F_{ret}(S_{emb}) \in [0, 1]^l$，该向量用于量化文档中每个 l 句子对主张验证的重要性概率。我们用 1，0 分别表示句子是否被选中。将选出的证据输入句子编码器后得到的句子嵌入记为 E_{emb}。

为确保忠实性，我们采用完整性和充分性的标准。为了获得更合理的证据，我们采用合理性的标准，即

该标准促使检索器进行符合人类认知的证据选择。我们分别将完整性、充分性和合理性的损失函数记为 L^{full}、L^{suff} 和 L^{plau}。因此，我们可以使用 L 将这三个损失函数联合表示为证据检索器的目标函数：

$$\mathcal{L} = \alpha_{full}\mathcal{L}_{full} + \alpha_{suff}\mathcal{L}_{suff} + \alpha_{plau}\mathcal{L}_{plau}. \quad (2)$$

3.3.1 合理性。 我们引入合理性准则，用于衡量并提升证据对人类而言的合理程度。为选择对主张验证器最重要的句子，我们使用一种 Top k 算法，选择概率分数最高的句子。具体而言，我们根据概率分数选择文档中排名前 k% 的句子。所选证据表示为 E。

我们采用具有相应金标准证据的主张，并使用二元交叉熵损失衡量预测证据与金标准证据之间的差异。我们将 $\boldsymbol{g_i} \in \{0, 1\}^{|S|}$ 表示为金标准证据，其中 0 或 1 表示是否选择某个句子。合理性损失函数可定义为：

$$\mathcal{L}_{plau} = \mathcal{L}_{BCE}(\mathcal{F}_{ret}(S_{emb}), \boldsymbol{g_i}), \quad (3)$$

这可以促使检索器在训练期间选择更合理的证据句子。

3.3.2 忠实性—完整性。 如果从文档中删除某些句子会导致错误的验证结果，我们可以认为这些句子包含在验证结果中发挥关键作用的关键证据。为了选择最关键的证据，我们应当识别出这样的句子：如果将其删除，主张验证器的性能会显著下降。

我们使用交叉熵损失 $L_{CE}(F_{ver}(S_{emb})\ y^*)$ 来衡量验证性能，其中标签 y^* 表示三个类别中的一个。为了评估删除证据句子的影响，我们可以将 $S_{emb}\backslash E_{emb}$ 的性能与原始输入进行比较。具体而言，我们可以使用以下公式衡量删除证据句子的影响：

$$\mathcal{L}_{full} = \mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}), y^*) - \mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}\backslash E_{emb}, y^*). \quad (4)$$

损失函数 L^{full} 可以促使检索器选择所有对主张验证至关重要的句子。

理想情况下，证据检索器会选择在验证过程中发挥决定性作用的关键证据句子，使得 $L^{full}_0 < 0$。为了解决这一问题，我们可以首先将 L'^{full} 设置为在相应的 $L_{full} < -B_f$ 时，其中 $B_f > 0$ 是一个超参数。为了转换原始损失值的范围，使其始终大于或等于 0，当 $L_{full} > -B_f$ 时，我们可以将 L'^{full} 记为 $L_{full} + B_f$，从而使重新 formulated 的损失值 $L'_{full} \geq 0$。形式上，我们可以如下定义 L'^{full}：

$$\mathcal{L}'_{full} = \max(0, \mathcal{L}_{full} + B_f), \quad (5)$$

从而可将 L^{full} 的值限制在 $[0, +\infty$ 范围内）。

3.3.3 忠实性-充分性。 为确保所选证据能够在原始源文档所提供的信息之外进一步提升验证性能，我们采用充分性准则。该准则促使检索器选择能够最大程度提升主张验证性能的证据，相较于仅使用原始文档。

更具体地说，我们采用 $L_{CE}(F_{ver}(E_{emb})\ y^*)$ 来表示使用证据替代文档时的性能，而 $L_{CE}(F_{ver}(S_{emb})\ y^*)$ 则代表以主张和文档作为输入提供给主张验证器时的原始性能。因此，我们将充分性损失函数定义为：

$$\mathcal{L}_{suff} = \mathcal{L}_{CE}(\mathcal{F}_{ver}(E_{emb}), y^*) - \mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}), y^*), \quad (6)$$

这鼓励检索器选出主张验证过程中使用的所有重要句子。当检索器训练充分时，损失函数 L^{suff} 也可能为负。为避免损失函数为负，我们可以采用类似的度量方式，设置一个足够大的超参数 $B_s > 0$，并将取值范围转换为 $[0, +\infty)$。因此，我们可以将充分性损失函数定义为：

$$\mathcal{L}'_{suff} = \max(0, \mathcal{L}_{suff} + B_s) \quad (7)$$

经过优化的检索器将检索到更好的证据，从而通过重新审视第 3.2 节中的验证器来改善其结果。

4 实验与分析

4.1 设置与基线

设置： 请注意，只有 CHEF [10] 为现实世界主张标注了金标准证据。尽管 FEVER、 [29], FEVER 2.0 [30], 和 FEVEROUS [2]标注了从维基百科检索到的证据，但它们并不提供现实世界中的主张。因此，我们仅使用 CHEF。为衡量 ReRead 的效果，我们在训练集上调整参数，并报告 CHEF 开发集和测试集上的结果。CHEF 的训练集／开发集／测试集分别包含 8,002／999／999 个样本。CHEF 还提供 Google 摘要作为证据，这些摘要是谷歌 [9]提供的源文档内容概要。遵循先前的研究 [9, 10, 21],，我们采用微平均 F1 和宏平均 F1 作为评估指标。对于基础编码器，我们采用 BERT-Base-Chinese [5] 和 RoBERTa-Base-Chinese [20]。我们将 k设为源文档中全部句子数量的 5%。我们使用 BertAdam [15]，以 4e-5 的学习率和 0.07 的预热来优化交叉熵损失，并将批次大小设为 16。为简化起见，我们将 α_{full}、 α_{suff}和 α_{plau}分别设为 1。

基线：遵循以往的工作 [9, 10], ，我们采用两类基线：流水线系统和联合系统。流水线系统首先根据主张从文档中检索证据，然后使用检索到的证据验证主张。证据检索器和主张验证是两个独立的步骤。我们采用（1）Google 摘要 [9]。（2）表层排序器 [2]。（3）语义排序器 [21]。（4）混合排序器 [27]。联合系统将证据抽取视为潜变量，并通过主张验证损失联合优化证据抽取过程。我们采用（5）基于强化学习的方法 [16]。（6）基于多任务的方法 [35]。（7）基于潜变量的方法 [10]。此外，我们还提供（8）无证据和（9）金标准证据，以展示结果的下限和上限。

4.2 结果与分析

总体性能。表1展示了在 CHEF 的开发集和测试集上进行 5 次训练和测试所得结果的均值和标准差。我们观察到，使用真实世界证据可以提升主张验证的效果，而源文档带来的提升比 Google 摘要更大，这与源文档包含更多信息这一事实有关。

表1：ReRead 与基线模型在 CHEF 测试集和开发集上的 Micro F1 和宏平均 F1 结果。

系统 / 证据			测试集				开发集			
			基于 BERT 的模型		基于 RoBERTa 的模型		基于 BERT 的模型		基于 RoBERTa 的模型	
			微平均 F1	宏平均 F1	微平均 F1	宏平均 F1	微平均 F1	宏平均 F1	微平均 F1	宏平均 F1
流水线	无证据		54.46±2.89	52.49±2.44	55.34±2.68	53.22±2.59	54.76±2.35	52.97±2.12	55.73±2.06	53.61±2.17
	Google 摘要 [9]		62.07±2.55	60.61±2.96	62.53±2.13	61.55±2.69	62.31±1.97	60.87±2.07	62.96±2.17	61.93±2.42
	表层排序器 [2]		63.17±1.67	61.47±2.02	64.21±1.94	62.05±2.17	63.53±1.78	61.78±1.95	64.66±1.86	62.49±2.08
	语义排序器 [21]		63.47±1.71	61.94±1.66	64.35±1.76	62.24±1.52	63.73±1.68	62.42±1.49	64.71±1.45	62.59±1.38
	混合排序器 [27]		63.29±1.65	61.80±2.31	63.98±1.53	61.78±1.48	63.12±1.72	61.53±1.59	64.32±1.83	62.11±1.43
联合	强化学习 [16]	Google 摘要	63.76±1.52	61.74±1.88	64.46±1.82	62.42±1.67	63.54±1.38	61.48±1.63	64.81±1.69	62.80±1.72
		源文档	64.37±1.65	62.46±1.72	65.04±1.59	63.05±1.47	64.68±1.62	62.63±1.49	65.48±1.68	63.41±1.39
	多任务 [35]	Google 摘要	62.78±1.41	61.98±2.59	64.19±1.98	62.62±1.76	62.94±1.86	62.37±1.65	64.51±1.79	63.05±1.76
		源文档	65.02±1.46	63.12±1.78	65.87±1.68	63.79±1.84	65.41±1.80	63.38±1.62	66.19±1.63	64.12±1.55
	潜在变量 [10]	Google 摘要	64.45±1.68	62.52±2.23	65.11±1.86	63.14±1.82	64.71±1.69	62.80±1.48	65.08±1.62	63.50±1.77
		源文档	66.77±1.43	64.65±1.74	66.95±1.68	65.13±1.57	66.96±1.45	64.92±1.50	67.33±1.26	65.57±1.39
流水线	**ReRead**	源文档	**70.87±1.05**	**68.78±1.21**	**71.24±1.11**	**69.52±0.96**	**71.31±1.08**	**69.25±1.18**	**71.79±1.26**	**69.98±1.09**
	无 $\mathcal{L}^{plau}$	源文档	67.67±1.32	65.84±1.46	68.03±1.35	66.11±1.48	67.96±1.57	66.04±1.51	68.14±1.42	66.31±1.56
	w/o $\mathcal{L}^{full}$&$\mathcal{L}^{suff}$	源文档	68.24±1.42	66.15±1.39	68.58±1.50	66.39±1.44	68.53±1.32	66.31±1.53	68.70±1.44	66.59±1.37
流水线	金标准证据		**78.99±0.82**	**77.62±1.02**	**79.14±0.93**	**78.59±1.02**	**79.26±0.94**	**78.04±1.10**	**79.98±0.89**	**78.81±1.01**

表2：检索证据质量分析。

方法	测试集				开发集			
	BERT-Base		RoBERTa-Base		BERT-Base		RoBERTa-Base	
	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1
表层	0.43	85.3	0.46	86.6	0.42	84.6	0.44	85.5
语义	0.53	88.1	0.55	89.5	0.52	88.4	0.56	89.4
混合式	0.48	87.7	0.50	88.9	0.46	87.5	0.48	88.6
强化学习	0.63	89.6	0.66	90.4	0.62	89.3	0.64	90.3
多任务	0.66	90.4	0.67	91.5	0.64	90.3	0.65	90.8
潜在变量	0.68	90.8	0.69	91.4	0.67	90.5	0.69	91.2
ReRead	**0.84**	**95.3**	**0.86**	**95.4**	**0.85**	**95.1**	**0.87**	**95.7**

相应地，这些源文档也包含更多噪声内容，但 ReRead 仍然始终优于基线。更具体地说，与此前的 SOTA 模型相比：Latent [10], ReRead 在开发集和测试集上的微平均 F1 平均高出 4.30%，宏平均 F1 平均高出 4.32%。我们将 ReRead 的持续改进归因于 ReRead 从源文档中检索到的忠实且具有合理性的证据。在考虑标准差时，ReRead 比所有基线都更加稳健，因为证据检索器通过合理性监督获得了金标准证据，从而提供了质量更高的证据。

消融研究。我们在开发集和测试集上开展消融研究，以展示 ReRead 中不同损失的效果。ReRead 不含L^{plau} 表示移除了可信损失函数，因此证据检索器不再使用金标准证据来训练所选证据。ReRead 不含L^{full}和L^{suff}表示从主张验证器中移除了忠实损失函数，这会导致证据检索器获取的证据不再取决于主张验证结果。表1中的消融实验结果总体表明，所有损失函数都有助于提升性能。具体而言，移除 L^{plau} 后，所选证据会变得缺乏说服力，导致 F1 性能下降 3.33%。移除 L^{full}&L^{suff} 后，模型会选择与任务无关的证据，导致 F1 性能下降 2.90%。

检索证据质量分析。 我们通过将检索到的证据与开发集和测试集中的金标准证据进行比较，评估检索证据的质量。我们使用 BLEU [23] 衡量检索证据与金标准证据之间的相似度，BLEU 越高表示质量越好。此外，5 名博士生根据检索到的证据为 100 条主张标注核验标签，另有 2 名博士生验证这些数据。这有助于我们评估检索证据的**可解释性** 。表2展示了 BLEU 和微平均 F1 分数。与最先进的基线相比，ReRead 的 BLEU 提高了显著的 17%，这证明，在证据检索器训练中纳入合理性损失有助于 ReRead 获得质量更高的证据，从而使人工标注的 F1 核验准确性提高了 5.87%。

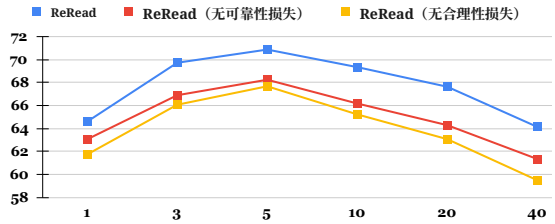

图 3：测试集上不同 k 的微平均 F1 结果。

选择比例的影响 k. 如图 3 所示，我们报告了 BERT-Base 编码器在测试集上采用不同 k 时的微平均 F1 分数。较低的 k 值可能会损害检索到的证据的信息充分性，进而影响验证结果。随着无关证据的加入，ReRead 的 F1 分数并未单调上升。当 k = 5时，模型性能最佳，这意味着选择 5% 的句子作为证据最为合适。如果移除忠实性和合理性损失，由于缺少金标准标签和证据的指导，ReRead 的 F1 性能平均会下降 3.24%。

5 结论

本文提出了一种新颖的事实核查框架 ReRead，该框架采用合理性、完整性和充分性标准，从真实世界的文档中检索合适的证据。检索到的证据能够反映主张的事实性，并让人信服。通过对证据检索器进行训练，它还能为主张验证器提供更好的证据，以供重新审视，从而提高核查任务的准确性。在真实世界数据集上的实验表明，ReRead 是有效的。未来，我们可以将忠实解释方面的研究扩展到知识图谱的构建 [11–13, 19, 37], 结构化知识的提取与问答 [17, 18]。

致谢

Lijie Wen 为通讯作者。Xuming Hu 和 Zhaochen Hong 对本研究的贡献相同。我们的代码可在以下地址获取：https://github.com/THU-BPM/ReRead。本研究得到中国国家重点研发计划（编号：2019YFB1704003）、国家自然科学基金（编号：62021002）、NSF 项目 III1909323、清华大学 BNrist 以及北京工业大数据系统与应用重点实验室的资助。

参考文献

[1] Bill Adair、Chengkai Li、Jun Yang 和 Cong Yu。2017年。迈向“圣杯”：持续探索实现自动事实核查。载于 *2017年计算+新闻学研讨会论文集*。[2] Rami Aly、Zhijiang Guo、Michael Sejr Schlichtkrull、James Thorne、Andreas Vlachos、Christos Christodoulopoulos、Oana Cocarascu 和 Arpit Mittal。2021年。FEVEROUS：在非结构化和结构化信息上进行事实提取与验证。载于 *第三十五届神经信息处理系统会议数据集与基准赛道（第一轮）*。[3] Bjarte Botnevik、Eirik Sakariassen 和 Vinay Setty。2020年。Brenda：用于虚假新闻检测的浏览器扩展。载于 *SIGIR*论文集。2117–2120。[4] Jiangui Chen、Ruqing Zhang、Jiafeng Guo、Yixing Fan 和 Xueqi Cheng。2022年。GERE：用于事实核查的生成式证据检索。载于 *SIGIR*论文集。2184–2189。[5] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：面向语言理解的深度双向Transformer预训练。载于 *NAACL-HLT*论文集。4171–4186。[6] Jay DeYoung、Sarthak Jain、Nazneen Fatema Rajani、Eric Lehman、Caiming Xiong、Richard Socher 和 Byron C Wallace。2020年。ERASER：评估合理化NLP模型的基准。载于 *ACL*论文集。4443–4458。[7] Lucas Graves。2018年。理解自动事实核查的前景与局限。*路透社新闻学研究所（2018年）*。[8] Zhijiang Guo、Michael Schlichtkrull 和 Andreas Vlachos。2022年。自动事实核查综述。*计算语言学协会会刊* 10（2022年），178–206。[9] Ashim Gupta 和 Vivek Srikumar。2021年。X-Fact：用于多语言事实核查的新基准数据集。载于 *ACL*论文集。675–682。[10] Xuming Hu、Zhijiang Guo、GuanYu Wu、Aiwei Liu、Lijie Wen 和 S Yu Philip。2022年。CHEF：基于证据的事实核查中文试验性数据集。载于 *NAACL-HLT*论文集。3362–3376。[11] Xuming Hu、Lijie Wen、Yusong Xu、Chenwei Zhang 和 Philip Yu。2020年。SelfORE：面向开放关系抽取的自监督关系特征学习。载于 *EMNLP*论文集。线上，3673–3682。[12] Xuming Hu、Chenwei Zhang、Fukun Ma、Chenyao Liu、Lijie Wen 和 S Yu Philip。2021年。通过增量元自训练进行半监督关系抽取。载于 *EMNLP* 论文集。487–496。[13] Xuming Hu 、 Chenwei Zhang 、 Yawen Yang、Xiaohe Li、Li Lin、Lijie Wen 和 S Yu Philip。2021年。面向低资源关系抽取的梯度模仿强化学习。载于 *EMNLP*论文集。2737–2746。[14] Alon Jacovi 和 Yoav Goldberg。2020年。迈向忠实且可解释的NLP系统：应如何定义和评估忠实性？载于 *计算语言学协会第五十八届年会论文集*。计算语言学协会，线上，4198–4205。https://doi.org/10.18653/v1/ 2020.acl-main.386[15] Diederik P Kingma 和 Jimmy Ba。2015年。Adam：一种随机优化方法。载于 *ICLR*论文集。[16] Tao Lei、Regina Barzilay 和 Tommi Jaakkola。2016年。神经预测的合理化。载于EMNLP论文集。107–117。[17] Aiwei Liu、Xuming Hu、Li Lin 和 Lijie Wen。2022年。通过迭代学习模式链接图实现语义增强的Text-to-SQL解析。载于KDD论文集。1021–1030。[18] Aiwei Liu、Xuming Hu、Lijie Wen 和 Philip S Yu。2023年。对ChatGPT零样本Text-to-SQL能力的全面评估。arXiv预印本arXiv:2303.13547（2023年）。[19] Shuliang Liu、Xuming Hu、Chenwei Zhang、Shu'ang Li、Lijie Wen 和 Philip S. Yu。2022年。HiURE：面向无监督关系抽取的层次化示例对比学习。载于NAACL-HLT论文集。5970–5980。[20] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar 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。arXiv预印本arXiv:1907.11692（2019年）。[21] Zhenghao Liu、Chenyan Xiong、Maosong Sun 和 Zhiyuan Liu。2020年。利用核图注意力网络进行细粒度事实核查。载于ACL论文集。计算语言学协会，线上，7342–7351。[22] Dan S Nielsen 和 Ryan McConville。2022年。Mumin：大规模多语言多模态事实核查错误信息社交网络数据集。载于SIGIR论文集。3141–3153。[23] Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002年。Bleu：一种用于机器翻译自动评估的方法。载于ACL论文集。311–318。[24] Ronak Pradeep、Xueguang Ma、Rodrigo Nogueira 和 Jimmy Lin。2021年。Vera：用于减少消费者健康搜索中有害错误信息的预测技术。载于SIGIR论文集。2066–2070。[25] Hannah Rashkin、Eunsol Choi、Jin Yea Jang、Svitlana Volkova 和 Yejin Choi。2017年。真相的不同色调：分析虚假新闻与政治事实核查中的语言。载于EMNLP论文集。计算语言学协会，哥本哈根，丹麦，2931–2937。[26] Tal Schuster、Roei Schuster、Darsh J. Shah 和 Regina Barzilay。2020年。文体计量学检测机器生成虚假新闻的局限性。计算语言学46，2（2020年），499–510。[27] Shaden Shaar、Nikolay Babulkov、Giovanni Da San Martino 和 Preslav Nakov。2020年。这是一个众所周知的谎言：检测此前经过事实核查的主张。载于ACL论文集。计算语言学协会，线上，3607–3618。[28] James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018年。FEVER：用于事实提取与验证的大规模数据集。载于NAACL-HLT论文集。计算语言学协会，新奥尔良，路易斯安那州，809–819。[29] James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018年。FEVER：用于事实提取与验证的大规模数据集。载于NAACL-HLT论文集。809–819。[30] James Thorne、Andreas Vlachos、Oana Cocarascu、Christos Christodoulopoulos 和 Arpit Mittal。2018年。FEVER2.0共享任务。载于第二届事实提取与验证研讨会（FEVER）论文集。[31] Nicolas Turenne。2018年。传闻光谱。PloS one 13，1（2018年），e0189080。[32] Nguyen Vo 和 Kyumin Lee。2018年。守护者的崛起：用于打击虚假新闻的事实核查URL推荐。载于SIGIR论文集。275–284。[33] Soroush Vosoughi、Deb Roy 和 Sinan Aral。2018年。真实与虚假新闻在网络上的传播。Science 359，6380（2018年），1146–1151。[34] Zhi-Xiu Ye 和 Zhen-Hua Ling。2019年。用于少样本关系分类的多级匹配与聚合网络。载于ACL论文集。2872–2881。[35] Wenpeng Yin 和 Dan Roth。2018年。TwoWingOS：一种用于基于证据的主张验证的双翼优化策略。载于EMNLP论文集，Ellen Riloff、David Chiang、Julia Hockenmaier 和 Jun'ichi Tsujii（编）。计算语言学协会，线上，

[36] 105 114 10 18653 1 18 1010 –。https://doi.org/。/2021 v /d - Xia Zeng、Amani S Abumansour 和 Arkaitz Zubiaga。。自动事实核查：综述。*Language and Linguistics Compass* 15，10（2021年），e12438。[37] Xin Zhang、Yong Jiang、Xiaobin Wang、Xuming Hu、Yueheng Sun、Pengjun Xie 和 Meishan Zhang。2022年。通过检索相关样本实现领域特定NER。载于COLING论文集。2398–2404。