软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(12):5720−5738 [doi: 10.13328/j.cnki.jos.007415] [CSTR: 32375.14.jos.007415]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于常识推理问答的多模态题文不符检测*

余建兴[1,3,4], 王世祺[1], 陈 祺[1], 赖韩江[2], 饶洋辉[2], 苏勤亮[2], 印 鉴[1,3]

[1](中山大学 人工智能学院, 广东 珠海 519082)

[2](中山大学 计算机学院, 广东 广州 510006)

[3](可持续旅游智能评测技术文化和旅游部重点实验室 (中山大学), 广东 珠海 510006)

[4](人工智能与数字经济广东省实验室 (广州), 广东 广州 510330)

通信作者: 印鉴, E-mail: issjyin@mail.sysu.edu.cn

摘 要: 主要研究题文不符的社交推文检测任务. 这些推文往往通过欺骗性的标题或封面图来误导读者点击与之无关的低质内容, 以让其广泛传播和带来点击量等商业利益. 为了规避检测, 恶意的创作者还会使用各种窍门将题文不符的推文伪装成合法的, 譬如添加无关易混淆的合法内容来干扰检测器. 检测这种推文需要对细节反复推敲, 甚至还要借助外部的常识进行多步推理验证. 然而, 传统方法一般把推文看成是一堆词语符号并简单灌入神经网络做分类, 忽略对其内在隐含的虚假细节进行分析, 导致漏判和误判. 而且这种黑盒子般的模型缺乏可解释性. 为了解决这些问题, 提出一种问答引导的新检测器, 通过质疑-验证的方式对细节逐一分析, 以发现潜在的不一致和虚假点. 首先利用多模态检索增强技术提取推文中的细节点, 然后通过提问的方式来质疑每个点. 为了充分验证事实和其复杂关系, 不仅覆盖简单的浅层匹配提问, 还有深层次常识推理的高阶提问. 每个提问可以从推文中找到字面答案. 但是该答案可能是虚构和不准确的. 为此, 通过开放域的问答模型借助外部知识源来交叉验证, 推导出相对可信的答案. 当两个答案不同时, 推文很可能存在虚假内容. 这种不一致可以作为有效的特征, 并与其他多模态的语义特征结合, 以提高检测模型的判别能力和鲁棒性. 此外, 这可以把复杂的检测任务分解为一系列问答步骤, 便于找出不一致细节来解释引起题文不符的原因. 在 3 个主流数据集上做了充分的实验, 验证了该方法的有效性.

关键词: 题文不符检测; 常识推理; 提问生成

中图法分类号: TP181

中文引用格式: 余建兴, 王世祺, 陈祺, 赖韩江, 饶洋辉, 苏勤亮, 印鉴. 基于常识推理问答的多模态题文不符检测. 软件学报, 2025, 36(12): 5720–5738. http://www.jos.org.cn/1000-9825/7415.htm

英文引用格式: Yu JX, Wang SQ, Chen Q, Lai HJ, Rao YH, Su QL, Yin J. Multi-modal Clickbait Detection by Asking Commonsense Reasoning Questions to Infer Inconsistencies. Ruan Jian Xue Bao/Journal of Software, 2025, 36(12): 5720–5738 (in Chinese). http://www.jos.org.cn/1000-9825/7415.htm

Multi-modal Clickbait Detection by Asking Commonsense Reasoning Questions to Infer Inconsistencies

YU Jian-Xing[1,3,4], WANG Shi-Qi[1], CHEN Qi[1], LAI Han-Jiang[2], RAO Yang-Hui[2], SU Qin-Liang[2], YIN Jian[1,3]

[1](School of Artificial Intelligence, Sun Yat-sen University, Zhuhai 519082, China)

[2](School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China)

[3](Key Laboratory of Intelligent Assessment Technology for Sustainable Tourism, Ministry of Culture and Tourism (Sun Yat-sen University), Zhuhai 510006, China)

* 基金项目: 国家自然科学基金 (62276279, 62372483, 62276280, U2001211, U22B2060); 广东省基础与应用基础研究基金 (2024B1515020032); 广州市科技计划 (2023B01J0001, 2024B01W0004)
收稿时间: 2024-09-25; 修改时间: 2024-12-22; 采用时间: 2025-02-18; jos 在线出版时间: 2025-09-10
CNKI 网络首发时间: 2025-09-11

[4](Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou), Guangzhou 510330, China)

Abstract: This study investigates the task of clickbait detection in social media posts. These posts often employ deceptive headlines or thumbnails to mislead readers into clicking on irrelevant or undesirable content, thus enabling widespread dissemination and generating commercial benefits such as increased clicks. To evade detection, malicious creators frequently disguise clickbait posts as legitimate ones, using techniques such as adding irrelevant or misleading content to deceive the detector. Detecting such posts requires a detailed analysis and complex multi-step reasoning using commonsense knowledge to identify inconsistencies. However, existing methods typically treat a post as a simple text span and feed it into a neural network for classification, neglecting the analysis of inherent false details, which leads to misjudgments. Moreover, these black-box models lack explainability. To address this issue, a new question-guided detector is proposed, which systematically analyzes the details through a doubt-then-verify approach to uncover potential inconsistencies and falsehoods. Specifically, a multi-modal retrieval-augmented technique is used to extract detailed clues from the content of the post, followed by questioning each clue. To ensure thorough verification of facts and their complex relationships, both simple matching questions and deep commonsense reasoning questions with varying levels of complexity are employed. Each question yields a plausible answer from the post, but the answer may be fabricated or inaccurate. Therefore, an open-domain QA model is utilized for cross-verification, leveraging external knowledge to derive a more reliable answer. When discrepancies are found between answers, the post is likely to contain false content. This inconsistency serves as a valuable feature and can be combined with other multi-modal features indicative of clickbait, improving the discriminative power of the detection model. By breaking down the complex clickbait detection task into a series of question-guided verification steps, inconspicuous inconsistencies can be identified to explain the underlying reasons for clickbait. Extensive experiments on three popular datasets demonstrate the effectiveness of the proposed approach.

Key words: clickbait detection; commonsense reasoning; question generation

随着社交媒体的快速发展, 推文已成为人们获取信息和分享感悟的重要渠道. 优质的推文可以捕获大量的用户点击量和阅读量, 这会给推文的创作者和社交平台带来盈利. 为了获得竞争优势, 恶意创作者通常使用不正当的方式在推文的标题、封面图和正文中添加一些诱人或欺骗性的内容, 以吸引读者的关注. 当读者受好奇心驱使[1]而点开相关的链接, 却发现内含大量如夸张、低俗等低质内容, 进而产生被骗、失望和厌恶的感觉. 受利益的驱使, 这种题文不符的推文在社交平台上日渐猖獗. 这不但损害了用户体验, 而且还会给社交平台甚至社会都带来严重的负面影响, 譬如传播错误信息、误导公众舆论、侵蚀用户信任和社交平台的声誉等[2]. 因此, 研究如何有效地检测这种题文不符的推文具有重要的商业价值和广泛的应用前景[3].

为了检测题文不符推文, 社交平台采取了很多的方法, 其中一种方式是人工审核. 然而, 这种方式成本高、耗时大, 难以应对大量题文不符推文的快速传播. 因此, 机器自动检测成为当前的重点研究方向. 根据判别依据的来源, 机器检测方法可以归纳为两种类型. 第 1 种是基于社交行为[4], 即通过学习推文在网络上传播的特性来判别, 如题文不符推文通常有高点击量和高分享次数, 但由于质量差, 对应的阅读时间却很短. 但这种行为特征依赖于用户的反馈数据, 存在延迟和遗漏等问题, 且检测时效性弱. 另一种方法是分析推文的内容, 根据题文不符推文常见的欺骗性词语、句法、情感、写作风格[5], 甚至是标点符号[6]等语言特征来判断. 此外, 题文不符推文的各个组成部分之间通常存在不一致, 例如标题与正文无关, 所描述的内容虚假并与实际情况相矛盾等. 为了分析内容的质量, 早期方法依赖于人工构建的规则, 成本高且可扩展性差. 后期使用数据驱动的神经网络模型, 通过学习内容与标签之间的关联来预测. 虽然可用性变强了, 但这种黑箱般的神经模型难以解释决策过程[7]. 可解释性对于建立可信人工智能至关重要[8]. 此外, 仅基于简单关联, 恶意创作者很容易利用一些技巧来伪装, 例如添加一些无害的无关内容以制造假关联来欺骗检测模型. 此外, 恶意创作者还会引入多种模态的虚假内容, 如使用视觉模态的与正文无关封面图来诱惑读者点击. 这比纯文本内容更难判断, 因为需要分析各种模态内容细节的真实性和相关性.

所谓质疑是发现真理的重要途径. 针对这些复杂的题文不符推文, 我们发现人们往往通过质疑来分析出细节中的不一致, 即对推文中的各个可疑点进行发问和深度验证, 并结合推文字面的含义和常识推理来发现矛盾点, 实现去伪存真. 如图 1 所示, 推文带有一条吸引人的标题, 恶意创作者在正文中注入了不少虚假内容, 让其跟真实内容混杂在一起, 力求让其伪装成合法的, 从而引起检测模型漏判和误判. 图 1 中的红框标记了诱骗的内容, 下划线和箭头表示多步的推理过程. 针对题文不符推文, 人们往往通过质疑来揣摩细节和发现矛盾. 例如发问“由在基尔大学实习的 15 岁男孩取得的名为 WASP-142b 的新发现是什么?” 人们发现从推文字面中得到答案 1, 但依赖外部

常识却推理出另一个答案 2. 这种矛盾表明推文中存在不实内容, 属于题文不符. 这种质疑和双重验证的方式不但可以实现高效的判断, 还能向用户展示其判断过程, 解释引起题文不符的原因. 受此启发, 我们提出通过问和答的技术来模拟人类的这种质疑-验证的过程, 从而提高检测器的准确性和可解释性. 然而, 如何提出一个切中要害的问题很关键, 但却不容易. 一个好提问可以加速不一致线索的发现, 而低质量提问则会陷入无效的验证中. 现有方法往往根据发问相关的内容直接转换或者映射出对应的提问. 对应的答案通常是给定文本中的片段, 可以通过直接匹配获得[9]. 这些简单提问仅能验证单步的虚假关系, 难以识别复杂且伪装的题文不符情况. 这些复杂推文经常涉及语境中多个事物之间、内部和外部常识之间、多跳间接关系之间的不一致. 这迫切需要可以生成复杂提问的方法. 考虑到题文不符推文的鉴别难度有差异, 还需对复杂提问的推理难度进行精细化控制. 此外, 传统方法主要是围绕文本内容进行发问, 难以检测图像和文本等跨模态之间的矛盾点. 为此, 我们需要根据推文的多模态内容构建高阶的常识推理提问. 这些提问不但难度可控, 而且有广泛的覆盖度, 既涉及推文语境的理解, 还能应对需要外部常识验证的场景.

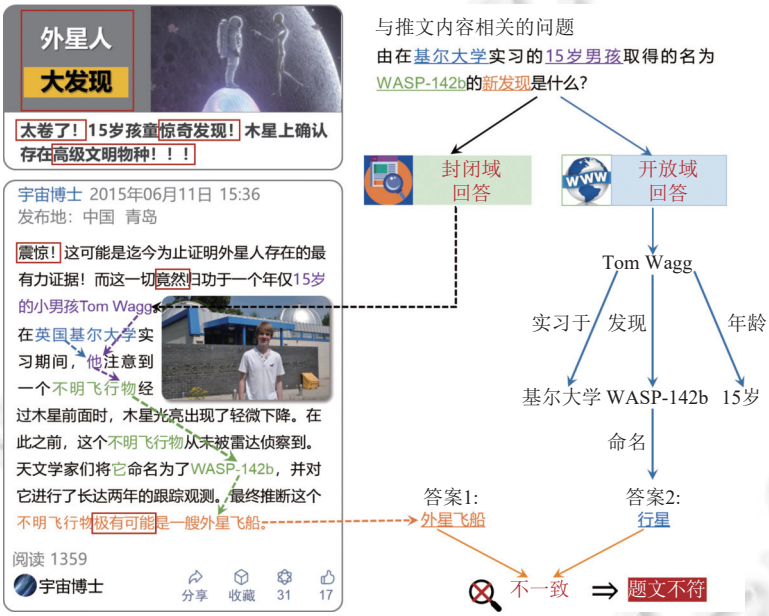


图 1　存在虚假和不一致内容的题文不符推文

为了应对这些挑战, 本文提出一种基于问答引导的题文不符检测框架. 该框架通过生成一系列与推文相关且难度可控的优质提问, 以质疑-验证的方式来发现虚假和不一致内容. 具体地, 我们首先使用检索增强技术提取出推文中图、文等不同种模态内容的相关线索, 以此在统一模态下进行质疑验证. 然后, 利用发问模板将这些线索转换成符合语法规范的基础子提问. 随后, 通过迭代的方式把这些子提问组装成更复杂的提问. 提问的难度由推理步数决定. 通过增加推理步骤数, 就能形成更长的推理链. 这种组合方式不仅能增强生成过程的可解释性, 还能实现对生成结果推理难度的定量控制. 其中复杂提问的组装是借助大语言模型 (large language model, LLM) 强大的低资源学习能力来实现. 为了提升提问的多样性, 我们构造了示例池, 并在组装时从示例池中选取不同示例来构成多样的提示. 把这些提示输入 LLM 就可以生成多个结果. 此外, 我们设计校验器对生成结果的质量进行评估, 从而筛除因错误传递等原因产生的低质量提问. 针对每个通过质量校验的提问, 我们分别根据推文内容和开放域知识进行解答和交叉验证. 当两类答案不相符时, 推文可能存在虚假内容. 这种不一致作为一种有效的特征, 可以与其他推文的多模态特征融合来提升检测模型的判别性能. 此外, 当发现题文不符时, 我们可以列举出对应的质疑提问和验证答案, 为判别过程提供可解释性. 在 3 个流行数据集上的大量实验验证了方法的有效性.

本文的主要贡献包括以下 3 点.

(1) 揭示了题文不符检测领域中存在的复杂伪装问题, 指出了复杂推导和可解释性等难点.

(2) 以一种问答的方式来检测题文不符, 从易到难地生成一系列高质量提问, 通过质疑推文细节来发现潜在矛盾. 生成的提问难度可控, 能够灵活地验证各种多跳间接关系的真实性和内外部知识的一致性.

(3) 进行了广泛的实验, 充分地检验所提出方法在题文不符检测领域中的性能. 结果表明, 本文方法能够有效检测多模态题文不符, 并且对结果提供了可解释的推理过程, 还对伪装的复杂推文具有很好的判别能力.

1 相关工作

在社交网络中, 推文的点击率、阅读量越高通常意味着广告等收益越大. 各大媒体和推文的创作者为了追求高收益往往会想方设法使推文更具吸引力. 一些不正当的创作者会使用色情、夸张、恐怖、冒犯性和欺骗性等内容诱惑读者点击. 这导致大量题文不符的低质推文在网络上迅速传播, 严重损害了用户体验. 传统的人工审核方法无力应对社交平台每日产生的海量推文. 因此, 依靠机器自动检测是目前较为可行的解决方案, 这已成为热门的研究话题. 一些学者尝试通过分析社交行为特点来检测, 这种行为包括创作者的发帖历史[10], 以及由读者产生的分享、评论、点击量和阅读时间等推文状态数据[11]. 然而, 这些方法存在冷启动问题, 即无法对没有社交历史行为数据的新推文做判断. 另一个可行的方向是分析推文的内容[12], 归纳出题文不符常见的用词、情感、标点符号等语言特征来预测[13]. 早期工作是通过人工构建的规则作为特征, 存在对专家知识依赖度高、扩展难等缺陷. 为解决这些不足, 学者们逐步转向数据驱动的神经网络方法. 这种方法无须人工定义特征, 而是通过编码-解码网络直接学习推文内容和题文不符标签的映射关系[14]. 具体而言, 题文不符推文往往具有诱人的标题和与之无关且带有欺骗内容的正文. 为了发现这种不一致性, 一些研究提出使用 n-gram 字面匹配或孪生网络[15]来衡量标题与正文的相关度. 然而, 标题是短文本而正文却是长文本, 两者信息量并不对等. 简单的字面匹配很容易由于长文本中的无关噪音而错误计算相关度. 为了处理这个问题, Yi 等人[16]将正文归纳为短文本摘要, 然后再与标题匹配. 但是, 该方法的性能受限于摘要模型, 存在误差积累问题. 此外, 不一致不仅存在于单一的文本模态之内, 也存在于不同模态之间[17], 例如配图和文并无关联. 一些研究者提出基于数据融合方法, 通过提取和拼接多种模态的特征来预测[18]. 然而, 这些方法大多是黑箱模型, 并不能向用户展示决策的过程和依据.

不同于以往研究, 本文把推文检测转换成一个问答任务, 以一种可解释的方式来发现不一致. 即通过发问来质疑推文中潜在的线索, 并通过文中和文外等多维度的知识来交叉验证. 这种基于问答的验证技术具有良好的可解释性[19], 可用于核查事实[20]、检测虚假[21]、错误纠正[22]等. 然而, 提出一个高质量且一针见血的问题并不容易[23]. 早期方法是通过人工制定的规则[24]或模板[25]来生成提问, 但这依赖于专家经验, 成本高且不容易扩展. 后续的研究工作大多转向数据驱动的神经网络模型[26], 它们往往利用编解码器框架及其变体学习映射关系, 直接把输入文本转成提问[27], 如 Seq2Seq[28]、预训练模型[29]、双学习模型[30]、图的模型[31]以及对抗网络模型[32]等. 由于缺乏对解答过程的建模, 生成的结果往往是浅层提问, 即通过字面匹配而直接获得答案[33]. 这种简单提问难以发现涉及多个事物多步关系的复杂题文不符. 一些工作提出引入解答的推理链来辅助生成复杂的多跳提问[34], 但这种提问通常仅能验证推文内部的实体和关系, 难以应对涉及推文以外更宽泛的常识矛盾场景[35]. 为此, Jia 等人[36]提出从知识图谱[37]中检索相关的三元组作为外部知识编码进模型, 尝试提升常识感知能力, 但这种大杂烩式的编码方式较为粗糙, 生成的结果多是单跳提问, 并不具备常识可推理能力[38]. 为解决这一不足, Yu 等人[39]通过对抗训练[40]来约束模型生成常识推理提问, 但生成过程缺少对提问复杂度的控制. 针对可控性问题, 传统工作主要根据提问是否可回答可解来定义难度. 这个定义过于粗粒, 难以精细地刻画难度的等级. Gao 等人[41]使用潜在变量来控制复杂度, 但这些变量是隐含的, 解释性差. Kumar 等人[42]根据提问中实体的流行度以及其在知识图谱中对应实体的链接置信度来衡量提问的难度. 这些方法大多数从问的角度着手, 而它的复杂度却主要体现在如何答, 即一个含有多个条件很冗长的提问也有可能是简单的, 因为它的答案可以通过匹配直接获得[43]. 因此, 本文提出从答的角度来实现刻画和控制提问的难度, 即获取提问解答过程涉及的文中和文外常识的多个实体和关系, 并将其作为先验知识引导模型生成, 以迭代的方式由易到难地生成一系列可推理的常识提问, 并灵活控制难度.

2 方法和框架

本文旨在学习一个模型 $F(y \mid x)$ 来检测输入推文 x 的质量 y ($y=1$ 表示题文不符, $y=0$ 则为正常), 其中 x 包括文本标题、封面图及其链接的多模态正文. x 的内容可能存在欺骗、矛盾和不一致, 例如含有诱人却与推文无关的封面图, 或者正文含有虚假内容. 此外, 恶意创作者会使用各种窍门和言辞将题文不符的推文加以伪装来规避检测. 检测这些推文并不容易, 需要对推文多模态内容的细节进行深入分析, 甚至还需要引入外部常识进行多步骤推理. 考虑到质疑是人们验证细节真实性的有效方式, 我们引入了问答技术, 把一个检测任务转换成问答任务来解决. 问答包括提问 (question generation, QG) 和作答 (question answering, QA) 两个技术. 从简单到复杂提出一系列问题 $\{Q_1,\dots,Q_n\}$ 来质疑 x 的细节 C. 针对这些提问, 我们可以从 x 中找到相应的答案 $\{A_1^C,\dots,A_n^C\}$. 考虑到这些答案不一定准确, 我们可以借助开放域问答模型从 x 以外的知识源中检索相关证据进行交叉验证, 由此推导出另一组答案 $\{A_1^O,\dots,A_n^O\}$. 当这两类答案 (A_i^C,A_i^O) 不同时, 推文很可能存在虚假内容. 这些线索可以用作判别特征来提升检测器 $F(\cdot)$ 的判别能力. 此外, 这种方式可以将复杂的检测任务分解为多个小的质疑-验证步骤, 还容易解释导致题文不符的步骤和矛盾点. 如图 2 所示, 本方法由 3 个模块组成: (1) 检索模块, 用于提取与推文各个模态内容相关的线索; (2) 提问生成模块, 由易到难地生成常识推理提问来质疑每条线索; (3) 回答模块, 对提问逐一验证以发现潜在的不一致特征, 并联合其他多模态的题文不符判别特征来预测结果. 下面分段阐述模型的具体构造.

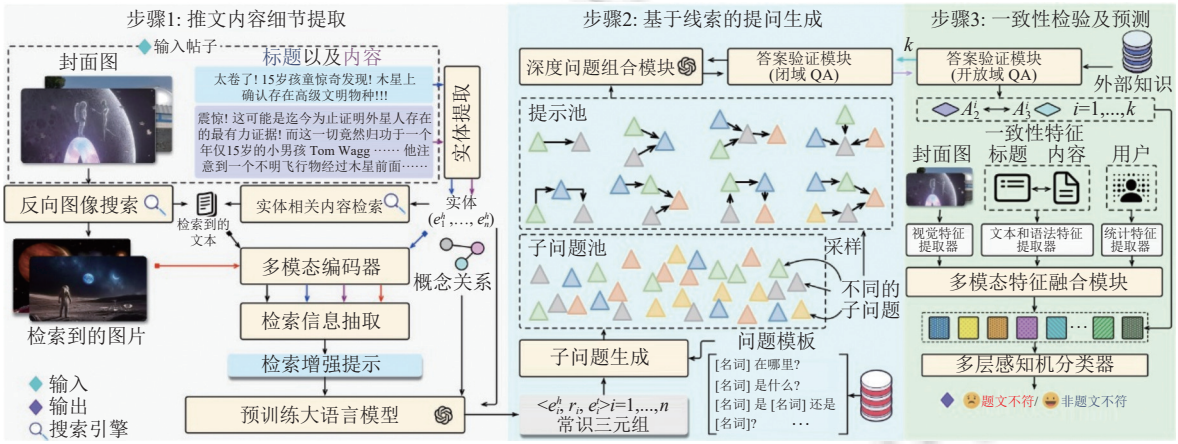


图 2 模型总体框架图

2.1 推文内容细节提取

推文所涉及的虚假和不一致往往体现在细节中, 如各种实体和关系. 为了能够更全面地验证这些细节, 我们提取推文之中和之外两个维度的相关细节 $(o,r,s)\in S$, o 和 s 分别表示头实体和尾实体, r 是关系, $S=S_P\cup S_C$, S_P 表示推文之内的细节, S_C 是没显式出现在推文中却很重要的外部常识.

2.1.1 推文内多模态线索检索

推文往往包含多个模态的内容, 而模态之间存在异构鸿沟, 即不同模态实体和关系之间难以直接进行关联和推理. 传统的方法主要通过分析文本内容来检测, 难以发现复杂的涉及多个模态的不一致. 为了解决这个问题, 我们把各个模态的内容转换为文本描述, 便于从中抽取出实体和关系等细节. 由于文本具有表达丰富、语义明确的特点, 它能够描述视觉等模态的内容, 进而灵活地将各种模态的信息联系在一起. 这就可以将复杂的多模态任务转换为单模态数据, 使得跨模态的关联和推理变得更加便捷和高效. 具体来说, 我们首先利用语言-图像预训练模型 ALIP[44]为推文中的每张图片 $v\in V$ 生成对应的描述文本 v_{cap}, 以便于在统一的文本模态中进行质疑和推理. 不同于普通的文本大模型 LLM, ALIP 是一种多模态大模型. 考虑到图片中嵌入的文字也经常存在题文不符现象, 我们还使用光学字符识别技术 (optical character recognition, OCR)[45]提取图中的文本 v_{ocr}. 最终, 将图像描述文本 v_{cap}、v_{ocr}

和推文文本 $t \in T$ 输入依存解析器[46], 抽取细粒度实体和关系, 形成 S_P.

2.1.2 推文之外常识线索检索

判断题文不符经常涉及常识, 即文内和文外部常识之间的不一致. 这里常识指大多数人所共享, 包括关于物理世界、历史、名人、事件等的事实. 尽管推文中没有直接提及这些知识, 但它们却是推理链中不可或缺的组成部分. 人们通过后天的学习能自然地关联和补充这些隐含的知识, 但机器却不具备这种能力. 因此, 我们需要借助额外知识源获取与推文相关的常识线索 S_C. 一种常用的检索常识策略是借助于知识图谱 (knowledge graph, KG), 但图谱的构建涉及大量的人工经验, 覆盖范围有限[47]. 此外, KG 还可能存在过时知识无法匹配推文中讨论的新事物. 由于 LLM 比 KG 覆盖更广泛的常识事实, 我们提出运用提示学习引导 LLM 生成常识线索, 并采用检索增强技术缓解 LLM 的幻觉问题. 下面介绍检索过程的 3 个步骤.

(1) 从互联网等大知识库中检索与推文相关的内容, 将其作为上下文输入 LLM, 让大模型有据可依, 促使生成更加准确、相关且丰富的常识. 具体地, 我们以推文中的每个图像 $v \in V$ 作为查询, 利用搜索引擎 Google Reverse Image 检索相关度排名前 μ 的文档. 我们抽取文档中的文本, 其描述了与 v 相关的事件细节, 包括背景、地点、参与者等. 类似地, 我们以推文中的每段文本 $t \in T$ 为查询, 利用 Google Search API 检索并保留排名前 ξ 的文档文本. 然后, 将这些文本输入 RECOMP 模型[48]提炼语义并形成精简摘要, 实现去除冗余和噪音的目的.

(2) 从步骤 1 检索获得的结果中提取相关信息用于提示学习. 首先利用多模态编码器 MUSTIE[49]将每个检索结果 g、推文文本 t 及图像 v 分别表征为 d_e 维度的特征向量 e_g、e_t 及 e_v. MUSTIE 考虑了不同模态的结构特性, 擅长建模模态特征之间的关系, 具有很好的多模态表征能力. 然后, 我们计算每个检索结果 g 与推文 x 的相关度为 $sim(g,x)=\lambda e_g \bar{e}_t+(1-\lambda)e_t \bar{e}_g$, 并保留前 N 个最相关的检索结果, 其中 λ 是预设权重值.

(3) 每个检索结果 g 是一段与推文相关的常识事件描述文本. 我们将 g 输入到依存解析器中提取其中的实体集合 E^g. 随后, 我们利用根据 g 和 E^g 构造的提示指令引导 LLM 输出常识线索. 提示模板如表 1 所示, 包括 [Context]、[Entity-1] 和 [Relation] 这 3 个输入的占位符, 它们分别由 g、E^g 中的实体 e_i^g 以及典型的常识关系 r 来填充. LLM 根据填充后的提示语句给占位符 [Entity-2] 填充合适的实体 e, 由此我们可以获得一条常识线索 $(e_i^g, r, e) \in S_C$, 继续将 e 填充到 [Entity-1] 则能获得提示, 让 LLM 输出更高阶的常识实体 s. 重复这一过程 k 轮, 我们能收集 k 阶的常识线索三元组. 如表 2 所示, 为了确定典型的常识关系, 我们对常见的常识推理提问所涉及的关系进行了分析与归纳, 并从知识图谱 *ConceptNet*-5[50]中找出对应的 12 组关系.

表 1 用于引出常识知识的提示模板

Prompt Samples
Based on the provided context: [Context]. Given the entity [Entity-1] mentioned in the context and the commonsense relation type [Relation], please identify a commonsense entity [Entity-2] that relates to the previous entity through the relation.

表 2 相关的典型常识关系类型

编号	具体关系
1	*RelatedTo*, *Synonym*, *SymbolOf*, *SimilarTo*, *ExternalURL*
2	*FormOf*, *DerivedFrom*, *Etymo-logicallyRelatedTo*, *EtymologicallyDerivedFrom*
3	*UsedFor*, *CapableOf*
4	*PartOf*, *HasA*, *HasContext*
5	*AtLocation*, *LocatedNear*
6	*Causes*, *HasSubevent*, *HasFirstSubevent*, *HasLastSubevent*, *HasPrerequisite*, *MotivatedByGoal*, *ObstructedBy*, *CausesDesire*
7	*HasProperty*
8	*Desires*
9	*CreatedBy*, *MadeOf*
10	*Antonym*, *DistinctFrom*
11	*IsA*, *DefinedAs*, *MannerOf*
12	*ReceivesAction*

2.2 基于线索的提问生成

为了从推文中找到能反映题文不符的判别证据, 我们生成一组切中要害的提问来质疑 S 中的每条线索. 所生成的提问难度可控, 能灵活地验证可疑线索集 S 中各种间接和高阶关系. 此处的推理难度被定义为解答提问所需的推理步骤数. 提问生成过程包含 3 个步骤, 首先根据每个线索 $s \in S$ 生成 1 跳的子提问, 然后将它们组合以获得多跳的复杂提问. 最终, 这些提问由一个校验器来评估质量, 以筛选出一组满足语言流畅性、语法有效性和推理难度等要求的优质提问.

2.2.1 基础子提问生成

复杂提问通常由一组相较简单的子提问按照某种推理逻辑聚合而成, 而这些子提问又由更小的提问组成[51]. 如果把每个单跳的基础子提问看成一个推理步骤, 我们就可以将复杂提问分解成一系列由推理链关联的基础子问题. 反之, 这些基础子提问也应能重新组装成相应的复杂提问. 受此启发, 我们先生成一组单跳的子提问作为用于组装的基元. 具体地, 针对每个三元组线索 $s \in S$, 我们以其中一个实体作为答案, 另一实体及关系作为发问点, 使用 BART 模型生成一个单跳的子提问. 该模型擅长于这种简单事实类的提问生成任务. 为了增强结果的流畅度, 我们还从大量无标注的提问数据中提取提问的表述模式[52], 并将其作为模板指导提问的生成过程. 具体来说, 对于每条提问数据, 先替换同义词, 便于概括表述形式的同时保留核心语义. 然后, 用 POS 标签替换提问中的短语, 例如用 [*NP*] 替换名词, 用 [*V*] 替换动词等, 以此进一步归纳为通用的发问句法模板.

对于每个三元组, 我们根据关系类型 r 检索相匹配的提问模板. 譬如与三元组 (Beijing, AtLocation, China) 相匹配的模板为“Which city is located in [*NP*]?”和“Where is [*NP*]?”, 以此生成的提问为“Which city is located in China?”和“Where is Beijing?”. 基于检索得到的提问模板, 我们利用 HuggingFace API[53]中的 BART-large 模型为三元组生成与答案对应的提问. 该模型由 24 个隐藏层组成, 每层包含 16 个注意力头, 隐藏层的维度为 4 096. 在经典的问答数据集 SQuAD[54]上进行训练后, 该模型能够准确地将三元组转换为单跳的子提问.

为了减少计算成本, 我们使用前缀调优技术[55], 即冻结预训练得到的参数并仅学习少量的前缀参数 pr, 使得 pr 能有效地学习“根据提问模板将三元组转换成提问”这一指令来引导模型生成高质量结果. 具体地, 我们首先将前缀 pr、三元组 s 的原始文本和提问模板 u_s 拼接为 $z=[pr;s;u_s]$, 并将其输入 BART 编码器中. 通过使用教师强制 (teacher forcing) 算法, 模型能够学习最优的前缀来生成期望的结果. 解码过程参考公式 (1), 其中 P_θ、$|P_{\text{idx}}|$、z_i、$\boldsymbol{h}_i$ 分别表示可学习的参数矩阵、前缀长度、输入中的第 i 个词以及该词的隐藏状态. 在训练过程中, 模型参数 ϕ 是固定的, 仅前缀参数 θ 需要训练. 最终, 我们收集所有的生成结果来构建出子提问池 $QP=\{qp_1,\ldots,qp_{|S|}\}$.

$$\begin{cases} \boldsymbol{h}_i = \begin{cases} P_\theta[i,:], & \text{if } i \in P_{\text{idx}} \\ \text{BART}_\phi^{\text{Encoder}}(z_i, \boldsymbol{h}_{<i}), & \text{otherwise} \end{cases} \\ qp = \text{BART}_\phi^{\text{Decoder}}(\boldsymbol{h}_1,\ldots,\boldsymbol{h}_n) \end{cases} \tag{1}$$

2.2.2 复杂提问组合式生成

接下来, 我们利用这些基础子提问作为基块, 通过类似堆积木的方式逐步组装成复杂提问. 组装过程主要有两个阶段, 首先寻找可组合的子提问对, 然后将其合并为更复杂的提问. 组合结果可以用于迭代地生成高阶提问. 这种组合式的方法可以提高可解释性, 并使提问生成器有效地控制中间推理过程, 从而获取高质量的提问.

(1) 考虑到并不是所有子提问都能组合成合理的提问, 一些缺乏逻辑关联的子提问对容易生成不合理且无法回答的结果. 为此, 我们采用“枚举+验证”的策略, 从子提问池 QP 中筛选合适的可组合提问对. 具体而言, 两个子提问以及它们的答案 (qp_i,A_i) 和 (qp_j,A_j) 可组合的条件之一是 qp_j 提及 A_i. 考虑到一些同义实体可能有多种表达形式, 因此使用工具包 SpaCy[56]来对齐同义实体. 此外, 我们还要求 A_j 没有被 qp_i 提及, 避免循环推理.

(2) 具备常识等高阶理解能力的提问往往包含复杂的推理结构, 这种结果指引了发问的方向. 因此, 我们设计了几种典型的结构作为先验知识来引导生成. 这些结构涵盖了主流社交媒体平台和数据集中几乎所有 2–4 跳推理的类型, 每个结构都是一个有向无环图, 其节点和边分别表示子提问和推理关系. 要将一组可组合的子提问 $QP_k \subseteq QP$ 转换为复杂提问, 可以利用经典的 Seq2Seq 模型. 然而, 这种模型严重依赖大量的训练数据, 且还需要对

这些数据的每个推理步骤进行标注. 但目前这种标注资源的数据集很稀缺. 此外, 手动标注成本高昂. 为了解决这个问题, 我们借助大模型 LLM, 其具备强大的小样本学习能力. 如图 3 所示, 我们设计思维链的提示样例来输入 LLM, 利用其强大的上下文学习能力, 融合多个子提问形成一个复杂提问. 第 k 个推理结构的提示由说明、一些示例、输入和输出占位符组成. 考虑到每个提问可以有多种同义的表达方式, 我们构造多个提示语句来模拟这种多样性能力. 图 4 展示了部分利用思维链合成复杂提问的提示样例. 图 4 中, 带颜色的下划线文本表示每个子提问的答案. 子提问 qp_i 的答案被另一个子提问 qp_j 提及, 表明这些子提问可以通过它们的答案关联, 从而形成一个有效的推理过程. 具体地, 我们为第 k 个推理类型收集了一组示例 E_k. 通过从 E_k 中抽取子集 E_i 作为输入, 我们可以构造多样化的提示语句, 而非单一且固定的提示. 基于公式 (2), 进一步为每组子提问 QP_k 生成多个复杂提问, 其中 m 是采样数量, $p_G(\cdot|\cdot)$ 表示使用核采样的 GPT-3 模型, 其中 $p=0.5$ [57], E 是示例集.

$$\begin{cases} Y_k=\bigcup_{i=1}^{m}\{p_G(y_i|prompt_k(QP_k,E_i))\} \\ E_i=\{e_i:e_i\ E_k\};E_k\subseteq E;QP_k\subseteq QP \end{cases} \tag{2}$$

Chain-of-thought Prompt

Given two sub-questions and compose them into a complex question. Example:
Sub-question-1: Who succeeded the first president of South Africa? Answer: Nelson Mandela
Sub-question-2: Who succeeded Nelson Mandela? Answer: Cyril Ramaphosa
Let's generate step by step:
Step 1: Identify the entity "first president of South Africa" in Sub -question-1.
Step 2: Note Sub-question-2 contains the answer to Sub-question-1.
Step 3: Substitute the "Nelson Mandela" from Sub-question-2 into the entity "the first president of South Africa" in Sub -question-1.
Step 4: Compose the questions into the final result: Who succeeded the first president of South Africa? Answer: Cyril Ramaphosa.
Now, we provide another two sub-questions,
Sub-question-1: [placeholder for Sub-question 1]
Sub-question-2: [placeholder for Sub-question 2]
You should compose them into a complex question like the above example and output the result:

图 3 用于生成 2 跳推理提问的思维链提示样例

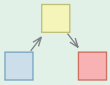
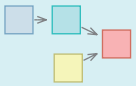
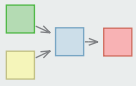
Structure	Composable Sub-Questions	Reasoning Question
	Who succeeded the first president of South Africa? [Nelson Mandela] Who succeeded [Nelson Mandela]? [Cyril Ramaphosa]	Who succeeded the first president of South Africa? [Cyril Ramaphosa]
	Who wrote the music for the ballet The Nutcracker? [Tchaikovsky] Where did Lenin pass away? [Moscow] When did [Tchaikovsky] live in [Moscow]? [1890s]	When did the composer of the music in the ballet The Nutcracker live in the city where Lenin passed away? [1890s]
	What city was Vincent Van Gogh born in? [Zundert] What country is [Zundert] located in? [Netherlands] What is the official language of the [Netherlands]? [Dutch]	What is the main language spoken where Vincent Van Gogh was born? [Dutch]
	Who starred in Space Jam? [Michael Jordan] Where was [Michael Jordan] born? [Brooklyn] What author was born in [Brooklyn]? [Walt Whitman] What is the name of [Walt Whitman]? [Mark Twain]	What is the name of the author born in the city where the basketball player who starred in Space Jam was born? [Mark Twain]
	What is the largest rainforest in the world? [The Amazon] What country is [The Amazon] rainforest located in? [Brazil] What most popular language do people speak in [Brazil]? [Portuguese] How many people speak [Portuguese] in [Brazil]? [200 million]	How many people speak the most popular language in the South American country that has the largest rainforest? [200 million]
	Who directed *The Dark Knight*? [Christopher Nolan] What country is Hugh Jackman from? [Australia] What movie directed by [Christopher Nolan] stars an actor from [Australia]? [Inception] Which year was [Inception] released? [2010]	Which year was the movie directed by *The Dark Knight*'s director, that stars an actor from Hugh Jackman's country, released? [2010]

图 4 基于思维链提示中使用的部分提问示例

图 4 基于思维链提示中使用的部分提问示例 (续)

2.2.3 提问质量校验

上述组合后的结果可能只是简单提问, 而不是期望的常识推理提问. 例如, 对于由多个从句构成的组合提问, 其答案可以通过匹配从文本中直接找到, 而不需要借助隐藏的常识知识进行复杂推理, 甚至无需对文中的语义进行理解[58]. 此外, 组合提问可能没达到所需的推理难度. 一些提问可能存在推理捷径, 答案可以直接通过匹配获得, 而无须多跳推理. 为此, 我们设计一个校验器来全面地测试组合提问的质量. 它是一个加权的分类器. 每个提问 Q_i 的得分记为 $\omega(Q_i)=\sum_{j=1}^{n}\beta_j\cdot v_j(QS_i,A_i,Q_i)$, 其中 β_j 是可学习权重, QS_i 是生成 Q_i 时参考的线索文本片段, A_i 是预期答案, $v_j(\cdot)$ 是评判函数, 得分超过阈值 λ_2 的 Q_i 将被认为是优质的提问. 我们从 3 个方面设计 $v_j(\cdot)$.

(1) 答案可解性. 为测试提问 Q_i 可解, 我们使用封闭域的问答模型 SG-Net[59]预测答案 $A_i^1=SG\_Net(QS_i,Q_i)$. 该模型具有较好的性能, 在典型的问答数据集 SQuAD 2.0[60]上, 其 $F1$ 得分比人类仅仅低 1.79%. 这可以作为较好的验证工具. 通过比对 A_i^1 和预期答案 A_i, 可以衡量 Q_i 的可解性, 即 $v_1=g(\boldsymbol{e}_{A_i},\boldsymbol{e}_{A_i^1})$, 其中 $g(\cdot)$ 为余弦相似度函数, $\boldsymbol{e}_{A_i},\boldsymbol{e}_{A_i^1}$ 为由 DeBERTa-v3-large[61]得到的 A_i 和 A_i^1 的嵌入向量.

(2) 推理的难度. 一个好的提问 Q_i 的推理难度应符合预期的推理跳数, 且没有推理捷径, 这需要分析 Q_i 中间的推理过程. 为此, 我们首先将 Q_i 解析成一棵 AMR 树[51], 每棵子树对应一个子提问 q_j, 而树的总高度 d' 即为推理步骤. 我们计算 d' 与预期推理难度 d 的差异 $v_2=1/\log_2(2+|d'-d|)$, 差异越小则 v_2 的值越大. 为判断是否存在推理捷径, 通常需要一条推理链作为参照, 但在实际情景中, 获取这种推理链的标注成本较高. 考虑到在提问生成过程中附带输出了预期答案, 我们转而利用问答模型 GA[62]进行评估, 该模型仅能回答简单的匹配类提问, 并不具备高阶的常识推理能力. 如果 Q_i 可以被 GA 准确解答, 则很可能存在推理捷径. 基于这一思路, 我们计算 GA 得到的答案 $A_i^2=GA(QS_i,Q_i)$ 与预期答案 A_i 的相似度 $v_3=1-g(\boldsymbol{e}_{A_i},\boldsymbol{e}_{A_i^2})$. 类似地, 用相同的方法检测组成 Q_i 的每个子提问是否包含推理捷径. 为便于分离分析中间推理步骤, 我们将上一个推理步骤的答案替换为相应的子树, 从而阻断当前推理步与之前推理步骤的联系, 形成独立的子提问 q_j. 如公式 (3) 所示, 我们将由 SG-Net 模型得到的相对可信答案 $a_j^1=SG\_Net(QS_i,q_j)$ 与 GA 模型得到的答案比较, 得到中间推理步骤感知的推理难度得分 v_4, 其中 l 为子提问数量.

$$v_4=\sum_{j=1}^{l}\left[1-g\left(\boldsymbol{e}_{a_j^1},\boldsymbol{e}_{a_j^2}\right)\right]. \tag{3}$$

(3) 上下文相关性. 对于一个优质的提问来说, 解答所需要的证据应该来源于输入的上下文中[63]. 我们设计分类器 $v_5=\sigma(\boldsymbol{W}_4[\boldsymbol{e}_x;\boldsymbol{e}_{Q_i}])$ 来判断 Q_i 与输入上下文是否契合与相关, 其中 $\boldsymbol{W}_4$ 是权重矩阵, $[;]$ 为连接操作, 而 $\sigma(\cdot)$ 是用于预测上下文相关性的逻辑函数. $\boldsymbol{e}_x$、$\boldsymbol{e}_{Q_i}$ 则是提问感知的上下文、上下文感知的提问表征. 它们由 BiDAF 模型[64]编码得到, BiDAF 擅长通过交叉注意力捕捉相关性.

校验器的训练分为两个阶段. 首先, 我们使用一些标注数据对进行初步训练. 该过程需要构建正、负样例作为训练样本, 而每个样本 (SQ_i,Q_i,A_i) 都可以视为一个正样例. 我们观察到, 提问中的实体词在确定推理方向上起着重要的作用. 改变这些实体词, 对应的答案通常也会变化. 因此, 通过替换实体词来生成负样例. 考虑到各个提问中

的实体通常不同, 我们将 Q_i 中的实体替换为其他提问 Q_j 中的实体, 从而产生多个负样例. 基于这些正、负样例, 我们使用交叉熵损失来训练校验器. 为弥补训练样本数量不足的缺陷, 我们在第 2 阶段使用来自生成器的增强训练样本逐渐更新校验器参数. 具体地, 我们使用带有各种示例的思维链提示生成多个候选样本, 并将由校验器评判后得分最高的样本视为正样例. 针对这些伪正样例, 继续通过前述的实体替换方法来构造负样例. 这些增强数据可以更好地促进校验器的训练. 同时, 这些经过验证的反馈可以帮助生成器过滤低质量噪音, 输出优质的增强数据. 经过多次迭代, 我们可以得出最优的校验器和生成器.

2.3 一致性验证及预测

针对生成的提问 Q_i, 往往可以从推文 x 中找到一个答案 A_i^c. 然而, 题文不符推文包含欺骗性内容, 致使 A_i^c 可能并非 Q_i 的正确答案. 因此, 我们可以通过验证答案 A_i^c 的真实性来反推出题文不符. 为了验证, 我们引入具备较强常识推理能力的问答模型 VE[65]. 针对提问 Q_i, VE 能够从互联网等开放域知识源中检索相关的常识知识, 并通过推理来获得答案 A_i^o. 当基于全局常识的 A_i^o 和基于文本局部语境的 A_i^c 不匹配的时候, 推文很可能存在题文不符内容. 这可以作为一种有效的特征, 通过联合其他经典的题文不符特征来提升检测器的鲁棒性, 作为更客观预测.

2.3.1 答案比较

为了比较 A_i^c 和 A_i^o, 一种简单的方法是利用 $F1$ 指标来评测单词的匹配度. 但这种基于字面匹配的方法不适用于无重叠词语的两个近义答案计算. 为了解决这个问题, 我们微调一个 Transformer 模型来进行匹对. 首先使用 BERT 模型将推文上下文线索 S_P、生成的提问 $\{Q_1,\ldots,Q_n\}$ 和答案对 $\{(A_1^c,A_1^o),\ldots,(A_n^c,A_n^o)\}$ 分别编码为 $\boldsymbol{e}_P$、$\{\boldsymbol{z}_1^q,\ldots,\boldsymbol{z}_n^q\}$ 和 $\{\boldsymbol{z}_1^a,\ldots,\boldsymbol{z}_n^a\}$. 随后, 线索、提问和答案对由特殊标记<CLS> 间隔, 而在答案对内部还插入了分隔标记<SEP>, 以此作为预训练模型 BERT 的输入. 如公式 (4) 所示, 我们根据推文内容与提问的相似度评估相应答案的重要程度, 其中 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 是可学习权重集、$\langle\cdot\rangle$ 为余弦相似度, α_i 是注意力机制. 由于不同提问具有不同的难度和推理方向, 它们对题文不符判别的重要性也存在差异. 为此, 我们使用该机制针对每个提问学习其重要性权重.

$$\begin{cases} att(\boldsymbol{e}_P,\boldsymbol{z}_i^q)=\langle \boldsymbol{W}_1\boldsymbol{e}_P,\boldsymbol{W}_2\boldsymbol{z}_i^q\rangle \\ \alpha_i=Softmax(att(\boldsymbol{e}_P,\boldsymbol{z}_i^q)) \\ \boldsymbol{z}=\sum_{i=1}^{m}\alpha_i\boldsymbol{z}_i^a \end{cases} \tag{4}$$

2.3.2 多特征联合预测

为了提高题文不符判别的鲁棒性和准确性, 我们从推文中提取 5 类经典的多模态特征, 包括视觉特征、文本特征、跨模态特征、语言特征和作者画像特征, 连同上述的答案一致性特征共 6 类特征来构建检测器. 这些特征具有良好的判别性. 通过拼接这 6 类特征, 我们得到向量 $\boldsymbol{o}$, 并将其输入到多层感知机分类器来预测结果 $\hat{y}=Softmax(\boldsymbol{W}_3\boldsymbol{o}+b)$, 其中 $\boldsymbol{W}_3$ 是可学习的参数矩阵、b 为偏置项. 如果 $\hat{y}$ 判别为题文不符, 或所生成提问已达到预设的最大推理难度 d, 我们终止提问生成过程, 并认为目标推文合规. 否则, 我们继续生成更高阶提问来验证推文中的疑点. 模型通过二元交叉熵损失进行训练, 即 $L=BCE(y,\hat{y})$. 下文详细介绍提取 5 类多模态特征的方式.

(1) 视觉特征. 推文一般包含封面图和正文插图两类视觉内容, 我们利用 Swin Transformer[66]模型提取对应的特征. 该模型以预训练 Transformer 作为骨架, 基于滑动窗口自注意力机制提取图像的层次特征. 此外, 考虑到推文图像中常包含与题文不符的人物, 我们分别使用 DNN 和 RetinaNet 网络检测图像中的人物和面部特征[67].

(2) 文本特征. 我们首先对推文中的标题和正文分词, 并通过预训练模型 BERTbase[68]给每个词嵌入编码. 考虑到封面图中也可能包含诱骗性文字, 我们还使用光学识别模型 OCR 提取、编码封面图中的文字.

(3) 跨模态特征. 题文不符推文往往在内容上存在不一致, 例如封面图与正文无关联. 因此推文各部分内容的一致性是题文不符判别的重要依据之一. 考虑到各部分数据模态不同, 所获取的多模态特征处于不同的语义空间, 无法直接进行匹对. 为此, 我们引入 ViLBERT 模型[69], 通过多头注意力网络捕捉视觉特征 $\boldsymbol{V}^s$ 和文本特征 $\boldsymbol{T}^b$ 的关联. 它根据所输入的数据类型区分不同输入特征, 最终输出文本感知的视觉特征 $\boldsymbol{F}^{vt}=CT((\boldsymbol{T}^b\boldsymbol{W}^t),(\boldsymbol{V}^s\boldsymbol{W}^v))$, 以及视觉感知的文本特征 $\boldsymbol{F}^{tv}=CT((\boldsymbol{V}^s\boldsymbol{W}^v),(\boldsymbol{T}^b\boldsymbol{W}^t))$, 其中 $\boldsymbol{W}^t$、$\boldsymbol{W}^v$ 为权重矩阵. 依赖于多头注意力机制, $\boldsymbol{F}^{vt}$ 和 $\boldsymbol{F}^{tv}$ 能够从

多个角度刻画题文不一致的诱惑行为.

(4) 语言特征. 题文不符推文在字面表述上具有鲜明的语言特色和修辞套路, 我们从 6 个角度提取语言特征, 包括 (a) 正文-封面图一致性, 表示为由预训练 CLIP 模型[70]提取的正文特征 $\boldsymbol{b}^c$ 和封面图特征 $\boldsymbol{t}^c$, 即 $[\boldsymbol{b}^c;\boldsymbol{t}^c]$; (b) 正文-标题一致性, 首先利用 BERT 编码正文和标题, 并使用带注意力机制的双向 GRU 分别获取其上下文特征 $\boldsymbol{b}$ 和 $\boldsymbol{h}$, 通过孪生网络[71]评估一致性 $sim_{bh} = Siamese(\boldsymbol{b},\boldsymbol{h})$; (c) 封面图-标题一致性, 首先使用在 MS COCO 数据集[72]上预训练的图像描述模型[73]将封面图转化为文本, 然后基于 BERT 编码来计算该文本与标题的余弦相似度; (d) 标题情感极性, 利用情感分类器[74]输出的两维特征, 分别表示情感极性 (积极/中立/消极) 以及强度值, 每一维的取值被标准化至范围 [0, 1]; (e) 标题词法分析, 通过统计特征记录推文标题中出现标点符号 (如“!”“?”“~”)、表情符号、代词、肯定词、模糊词等出现的次数; (f) 常用诱骗词分析, 构建一个常用词统计特征, 用于记录题文不符推文常用词[75](如“震惊”“艳照”“肢解”等) 的出现频次.

(5) 作者画像特征: 推文的作者能一定程度反应推文的质量. 若一个作者在过往历史中频繁发布题文不符的推文, 则其很可能是恶意创作者, 它后续发布的推文也很大可能是带有偏见和诱骗性的. 为了刻画每个作者 u_j 的口碑度, 我们根据其作者个人资料来提取画像特征, 包括作者的账号年龄、昵称、被关注数、关注数、推文发布数、过往历史中低质量推文发布数、第 1 篇推文发布距今时间和最后一篇推文发布距今时间 (按天数计算). 昵称项被输入至 BERTbase 获取词嵌入特征, 其余项则都被表征为一维的数值特征. 通过将这些特征逐项拼接, 我们得以获取表征作者画像的特征.

3 实验分析

我们进行了定性和定量分析的实验, 以全面评估所本文所提出方法的性能.

3.1 数据集与实验设置

我们在 3 个经典数据集上进行了评估, 包括 CLDInst[76]、Clickbait17[77]和 FakeNewsNet[78], 它们分别包含 4k/3k、9k/29k 和 5k/17k 的题文不符合正常推文样本. 这些数据集都是通过众包方式收集的, 对应的样本来自 Twitter 和 Instagram 等流行社交媒体平台. 相当一部分样本需要借助复杂的多跳推理和隐藏的常识才能分析出结果, 这与我们的任务很契合. 表 3 展示了这 3 个数据集的统计结果, 下文对它们进行详细介绍.

表 3　3 个评估数据集的统计特征

数据集	总样本数	题文不符样本数	非题文不符样本数
CLDInst	7769	4260	3509
Clickbait17	38517	9276	29241
FakeNewsNet	23196	5755	17441

(1) CLDInst: 涵盖了从 Instagram 爬取的 7769 篇时尚相关推文. 这些推文由众包网站聘用的标注工作者进行判别. 数据集共有 4260 篇推文被标注为题文不符样本.

(2) Clickbait17: 包含来自 27 家美国主要新闻出版商的 38517 篇 Twitter 推文. 为了避免偏见, 在数据集构建时每天在每个出版商处最多抽取 10 篇推文, 其中 9276 篇推文被标注为题文不符样本.

(3) FakeNewsNet: 一个大规模的多模态新闻数据集, 包含来自 PolitiFact 和 GossipCop 网站的 23000 多篇带有假/真标签的推文. 该数据集需要丰富的常识能力才能判别, 其中来自 PolitiFact 的虚假样本有 432 个, 真实样本有 624 个; 来自 GossipCop 的虚假样本有 5323 个, 真实样本有 16817 个. 与题文不符推文类似, 虚假样本通常存在不相关、不一致内容. 该数据集可用于评估模型识别诱骗类推文的能力.

为了评估检测性能, 我们采用了 4 个分类领域中典型的指标, 包括准确率 (ACC)、精确率 (PRE)、召回率 (REC) 和 $F1$-score ($F1$). 模型的检测性能越强, 这些指标的值则会越大. 考虑到这些数据集中题文不符、非题文不符推文数据量显著不平衡, 我们使用过采样技术训练所有评估模型. 为确保实验结果的显著性, 我们重复执行了 5 次测试, 并报告了平均性能.

3.2 与基准模型的性能比较实验

我们将所提出方法与 6 个主流的基准模型进行了比较, 包括: ① HPFN[79], 根据帖子在社交网络上的传播行为特征做判断; ② FCQA[80], 生成关于标题的提问来验证真实性; ③ MCAN[81], 通过堆叠的协同注意力层捕获文本和视觉特征的相关度来判断; ④ CPDM[82], 使用集成分类器来判断内容、标题和封面图之间的不一致性; ⑤ CCD[83], 基于因果干预和反事实推理来判断; ⑥ VLP[84], 基于多模态预训练特征进行检测.

表 4 展示了在 4 个评估指标上的比较结果及其方差. 使用统计 t 检验, p 值<0.005, 表明提升显著. 我们发现在 ACC 方面, 本文方法的得分比在 CLDInst、Clickbait17 和 FakeNewsNet 等上比最佳的基准模型 (如 VLP、CCD) 还高出 6.16%、6.27% 和 6.85%. 这表明了本文方法的有效性, 通过提出切中要点合理的提问可以有效地检测各种可疑线索. 通过学习常识推理能力, 本文方法可以通过复杂的问答来验证各种线索的真实性, 这有助于发现诱骗信息. 为了评估常识知识的必要性, 我们从每个数据集中随机选择了 600 个样本进行人工标注. 结果表明, 大约 32%、35% 和 36% 的推文借助外部常识推理才能判断. 我们观察到本文模型优于传统的问答方法 FCQA. FCQA 仅聚焦标题的真实性, 而本文方法则对推文中各个模态中的线索逐一进行提问和验证. 我们还发现 MCAN、CPDM、CCD 和 VLP 等多模态方法优于 FCQA、HPFN 等单模态方法. 这表明各种模态之间普遍存在不一致性, 它可以作为题文不符检测的重要判别依据. 与传统的特征融合方法不同, 我们从多种模态中检索线索, 并针对每个线索提出提问, 更好地发现所有潜藏的不一致.

表 4 本文方法与基准模型的性能比较结果 (%)

基准模型与本文方法	CLDInst				Clickbait17				FakeNewsNet			
	ACC↑	*PRE*↑	*REC*↑	*F*1↑	*ACC*↑	*PRE*↑	*REC*↑	*F*1↑	*ACC*↑	*PRE*↑	*REC*↑	*F*1↑
HPFN	73.15	72.12	70.23	71.16	81.68	84.25	82.35	83.29	84.82	84.43	85.68	85.05
FCQA	76.49	76.86	75.74	76.30	82.47	84.83	82.44	83.62	82.11	84.64	81.49	83.04
MCAN	79.21	77.45	77.80	77.62	84.51	85.56	82.74	84.13	83.82	85.13	82.21	83.64
CPDM	80.04	77.89	79.31	78.59	86.14	86.30	83.07	84.65	84.27	85.18	82.44	83.79
CCD	82.77	83.13	82.91	83.02	88.36	87.61	86.46	87.03	87.72	85.96	86.46	86.21
VLP	85.56	84.25	83.87	84.06	88.70	87.34	86.02	86.67	88.02	87.25	86.23	86.74
Ours	**90.83**	**89.41**	**89.25**	**89.33**	**94.26**	**93.25**	**92.47**	**92.86**	**94.05**	**93.67**	**93.10**	**93.38**

注: 带下划线、加粗数据分别表示基准模型、所有对比模型在各评分指标中的最优得分

3.3 消融实验

为了验证本文方法中各模块的有效性, 我们为 6 个关键模块设置相应的消融实验, 以评估它们对总体性能的影响, 包括 (1) w/o *MulR*, 只关注文本内容而丢弃图像模态; (2) w/o *RaG*, 去除从外部知识源中检索增强获得的内容, 直接从大模型 LLM 中获取常识知识; (3) w/o *PCO*, 丢弃常识线索, 只根据推文中的内容来提问; (4) w/o *CQG*, 剔除提问生成模块, 使用检索到的线索作为提示输入大模型 LLM 直接进行预测; (5) w/o *VVM*, 删除提问质量校验器; (6) w/o *CVM*, 删除答案比较模块, 只根据推文内容的多模型特征进行预测.

如图 5 所示, 各消融实验组的性能均下降超过 5%. 这表明本文方法的每个组成模块在线索检索和常识推理中都发挥了重要作用, 让模型能更容易发现题文不符. 删除 *CQG* 模块对性能的影响最大, 在 3 个评测数据集上的 *ACC* 指标分别下降了 6.48%、7.23% 和 7.41%. 缺少该模块将使模型难以实现跨模态推理. 去除 *MulR* 模块也会导致性能大幅下降, 其 *ACC* 和 *F*1 指标分别下降超过 6.27% 和 6.31%. 这表明基于封闭域和开放域的问答模型来验证答案匹配程度可以很好地检测出不一致性. *RaG* 和 *PCO* 模块为我们的模型提供了常识线索. 剔除它们也会引起判断依据不足, 导致性能下降. 所有消融结果都验证了我们框架的合理性和有效性.

3.4 生成提问质量的定性评估

所生成提问的质量对于检测题文不符的各种不一致和虚假内容至关重要. 除了定量评测提问质量, 我们采用人工方法进行定性评估, 即将本文方法与基于问答的基准模型 FCQA 比较. 具体地, 我们从每个数据集的测试集中随机选取 600 个样本, 并收集评测方法的预测结果. 将结果交由众包平台 Figure-Eight 进行评级, 包括 3 个指标: (1) Syntax, 衡量提问在语法层面是否合理和流畅; (2) Relevance, 评估提问是否与推文相关; (3) Usefulness, 判断提

问是否有助于发现题文不符. 我们计算了评级的平均累积分数作为各评测方法的性能得分. 这些分数范围为 1–10, 其中 1 表示最差, 10 表示最好. 评级者的一致性通过 Randolph 的自由边际 Kappa 指标来衡量[85].

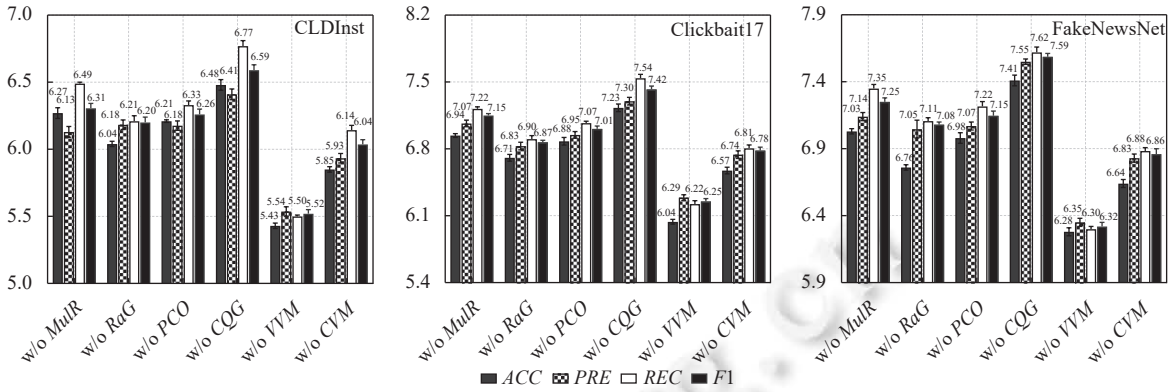


图 5 消融实验结果, 减少某模块后的性能下降幅度

表 5 使用统计 t 检验, p 值<0.005 时, 表明提升显著. 本文方法在所有指标上都显著超过了基准方法 FCQA, 尤其是在 Usefulness 指标上. 这与第 3.2 节中定量评测结果一致. 所有指标在 Kappa 一致性指标上均被评为中等[86]. 可以推断, 本文方法使用由易到难组合范式的可控框架, 可以生成具备常识推理能力的高质量提问. 这有助于找到推文中隐式和伪装的诱饵.

表 5 人工评估生成提问质量的结果

方法	CLDInst			Clickbait17			FakeNewsNet		
	Syntax	Relevance	Usefulness	Syntax	Relevance	Usefulness	Syntax	Relevance	Usefulness
FCQA	4.96	4.64	4.27	5.33	4.87	4.46	5.37	4.91	4.60
Ours	6.49	7.07	7.21	6.85	7.21	7.47	6.97	7.40	7.58

此外, 我们还评估了检测题文不符推文所需的提问推理难度. 对于 CLDInst、Clickbait17 和 FakeNewsNet 数据集中的每个题文不符样本, 我们记录了提问生成过程结束时的轮数. 这些轮数表明了推理步骤和复杂程度. 如图 6 所示, 我们观察到第 1 轮生成的提问只能识别出 7.3%–8.8% 的题文不符样本. 而大多数样本则是在 2–3 轮生成后才能被识别, 这一阶段生成的提问涉及复杂的多跳推理. 这说明题文不符检测是一项具有挑战性的任务, 需要进行复杂的推理分析才能发现不一致和虚假的内容.

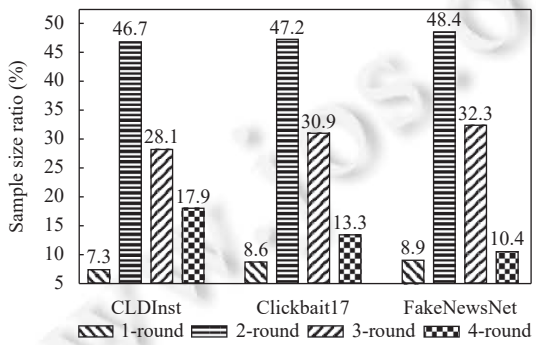


图 6 识别题文不符推文所需的提问推理难度评估结果

3.5 案例分析与讨论

为了进一步探究所提出方法的优缺点, 我们进行了案例分析. 图 7 所示的推文通过封面图和标题误导读者, 造

成火龙果对降低血压有极大功效的假象. 针对这篇推文, 使用我们的方法检索了一系列上下文和常识事实的三元组, 并以此作为线索来构建多个子提问. 基于先验推理结构, 我们组装成复杂具有常识推理能力的提问. 这个组合过程是可控的, 所生成提问的推理步骤数随着轮数增加而递增, 每个提问都从一个新角度验证潜在的题文不符方面. 我们展示了不一致得分最高的提问. 要回答这个复杂提问, 需要对多个事实进行推理, 例如水果中含有钾元素、钾是一类矿物质元素以及钾元素对降低血压有功效. 根据推文中的上下文内容和外部开放域的知识, 我们的模型分别得出两个答案. 通过比较它们可以找出不一致之处, 从而更好地帮助检测器做出正确判断, 还可以对决策给出合理的解释.

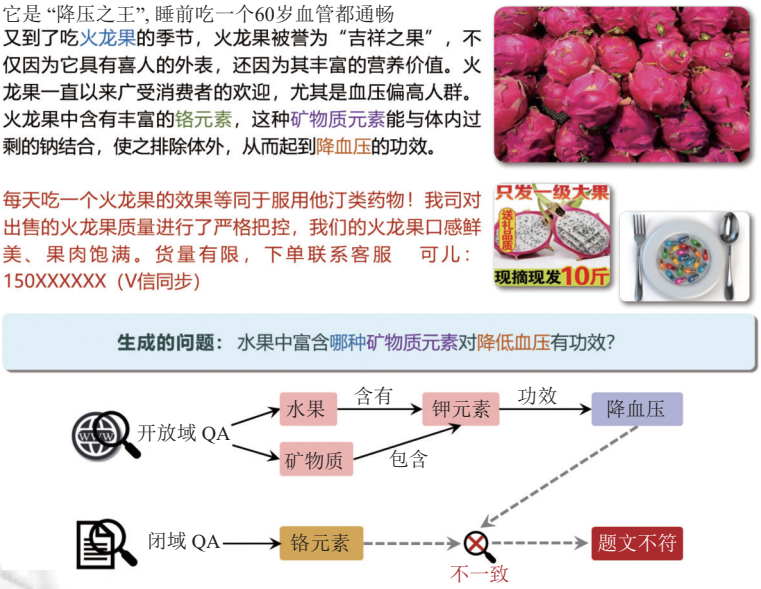


图 7 本文方法判断题文不符推文的案例展示

4 总 结

本文旨在检测社交媒体平台上传播的题文不符推文. 这些推文各组成部分内容通常存在不一致, 例如用与推文内容不符的封面图来诱惑用户点击, 这会损害用户体验. 此外, 这些推文会使用窍门进行伪装, 以规避检测系统. 这项任务具有挑战性, 需要对推文细节全方位分析, 甚至借助外部常识才能推断潜在的不一致. 现有方法大多基于神经网络, 这种黑匣子框架的推理能力和可解释性较弱. 为了解决这个问题, 我们提出了提问-验证的检测框架. 具体来说, 我们首先使用多模态检索增强技术来收集与推文相关的各类潜在线索. 然后, 我们可控地生成一系列相关提问来质疑每个线索, 并通过解答这些提问来验证线索的真实性. 与简单的基于匹配的提问不同, 我们方法生成的提问具备多跳常识推理能力, 它们有助于灵活地检查复杂且隐式的虚假事实和不一致关系. 通过多知识源多维交叉验证答案的可靠性, 我们可以发现推文中的虚假、伪装等不一致内容, 并为预测结果提供可解释的推理过程. 在 3 个数据集上进行的充分的实验, 结果表面了本方法的有效性.

References:

[1] Indurthi V, Syed B, Gupta M, Varma V. Predicting clickbait strength in online social media. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020. 4835–4846. [doi: 10.18653/v1/2020.coling-main.425]

[2] Ecker UKH, Lewandowsky S, Chang EP, Pillai R. The effects of subtle misinformation in news headlines. Journal of Experimental Psychology: Applied, 2014, 20(4): 323–335. [doi: 10.1037/xap0000028]

[3] Immorlica N, Jagadeesan M, Lucier B. Clickbait vs. quality: How engagement-based optimization shapes the content landscape in online platforms. In: Proc. of the 2024 ACM Web Conf. Singapore: ACM, 2024. 36–45. [doi: 10.1145/3589334.3645353]

[4] Agarwal B, Agarwal A, Harjule P, Rahman A. Understanding the intent behind sharing misinformation on social media. Journal of Experimental & Theoretical Artificial Intelligence, 2023, 35(4): 573–587. [doi: 10.1080/0952813X.2021.1960637]

[5] Zhou XY, Jain A, Phoha VV, Zafarani R. Fake news early detection: A theory-driven model. Digital Threats: Research and Practice, 2020, 1(2): 12. [doi: 10.1145/3377478]

[6] Coste CI, Bufnea D. Advances in clickbait and fake news detection using new language-independent strategies. Journal of Communications Software and Systems, 2021, 17(3): 270–280. [doi: 10.24138/jcomss-2021-0038]

[7] Ju TJ, Liu GS, Zhang ZS, Zhang R. A review of probe interpretable methods in natural language processing. Chinese Journal of Computers, 2024, 47(4): 733–758 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2024.00733]

[8] Dai SC, Hsu YL, Xiong AP, Ku LW. Ask to know more: Generating counterfactual explanations for fake claims. In: Proc. of the 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2022. 2800–2810. [doi: 10.1145/3534678.3539205]

[9] Fan YF, Zou BW, Xu QT, Li ZF, Hong Y. Survey on commonsense question answering. Ruan Jian Xue Bao/Journal of Software, 2024, 35(1): 236–265 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6913.htm [doi: 10.13328/j.cnki.jos.006913]

[10] Shu K, Zhou XY, Wang SH, Zafarani R, Liu H. The role of user profiles for fake news detection. In: Proc. of the 2019 IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining. Vancouver: ACM, 2019. 436–439. [doi: 10.1145/3341161.3342927]

[11] Ma J, Gao W, Wong KF. Detect rumors in microblog posts using propagation structure via kernel learning. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017. 708–717. [doi: 10.18653/v1/P17-1066]

[12] Wu Y, Cao MP, Zhang YZ, Jiang Y. Detecting clickbait in Chinese social media by prompt learning. In: Proc. of the 26th Int'l Conf. on Computer Supported Cooperative Work in Design. Rio de Janeiro: IEEE, 2023. 369–374. [doi: 10.1109/CSCWD57460.2023.10152690]

[13] Popat K. Assessing the credibility of claims on the web. In: Proc. of the 26th Int'l Conf. on World Wide Web Companion. Perth: ACM, 2017. 735–739. [doi: 10.1145/3041021.3053379]

[14] Cheng MX, Nazarian S, Bogdan P. VRoC: Variational autoencoder-aided multi-task rumor classifier based on text. In: Proc. of the 2020 Web Conf. Taipei: ACM, 2020. 2892–2898. [doi: 10.1145/3366423.3380054]

[15] Meng Q, Liu B, Sun XG, Yan H, Liang CY, Cao JX, Lee RKW, Bao X. Attention-fused deep relevancy matching network for clickbait detection. IEEE Trans. on Computational Social Systems, 2023, 10(6): 3120–3131. [doi: 10.1109/TCSS.2022.3207479]

[16] Yi XY, Zhang JR, Li WH, Wang XT, Xie X. Clickbait detection via contrastive variational modelling of text and label. In: Proc. of the 31st Int'l Joint Conf. on Artificial Intelligence. Vienna, 2022. 4475–4481. [doi: 10.24963/ijcai.2022/621]

[17] Shu K, Cui LM, Wang SH, Lee D, Liu H. dEFEND: Explainable fake news detection. In: Proc. of the 25th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. Anchorage: ACM, 2019. 395–405. [doi: 10.1145/3292500.3330935]

[18] Kumar V, Khattar D, Gairola S, Kumar Lal Y, Varma V. Identifying clickbait: A multi-strategy approach using neural networks. In: Proc. of the 41st Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval. Ann Arbor: ACM, 2018. 1225–1228. [doi: 10.1145/3209978.3210144]

[19] Wang SY, Wei ZY, Fan ZH, Huang ZF, Sun WJ, Zhang Q, Huang XJ. PathQG: Neural question generation from facts. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 9066–9075. [doi: 10.18653/v1/2020.emnlp-main.729]

[20] Pan LM, Lu XY, Kan MY, Nakov P. QACheck: A demonstration system for question-guided multi-hop fact-checking. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations. Singapore: Association for Computational Linguistics, 2023. 264–273. [doi: 10.18653/v1/2023.emnlp-demo.23]

[21] Liao H, Peng JH, Huang ZY, Zhang W, Li GH, Shu K, Xie X. MUSER: A multi-step evidence retrieval enhancement framework for fake news detection. In: Proc. of the 29th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. Long Beach: ACM, 2023. 4461–4472. [doi: 10.1145/3580305.3599873]

[22] Thorne J, Vlachos A. Evidence-based factual error correction. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. ACL, 2021. 3298–3309. [doi: 10.18653/v1/2021.acl-long.256]

[23] Liang YY, Wang JN, Zhu HL, Wang L, Qian WN, Lan YS. Prompting large language models with chain-of-thought for few-shot knowledge base question generation. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 4329–4343. [doi: 10.18653/v1/2023.emnlp-main.263]

[24] Mazidi K, Nielsen RD. Linguistic considerations in automatic question generation. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014. 321–326. [doi: 10.3115/v1/P14-

2053]

[25] Willert N, Thiemann J. Template-based generator for single-choice questions. Journal of Technology, Knowledge and Learning, 2024, 29(1): 355–370. [doi: 10.1007/s10758-023-09659-5]

[26] Yu JX, Quan XJ, Su QL, Yin J. Generating multi-hop reasoning questions to improve machine reading comprehension. In: Proc. of the Web Conf. 2020. Taipei: ACM, 2020. 281–291. [doi: 10.1145/3366423.3380114]

[27] Ang BH, Gollapalli SD, Ng SK. Socratic question generation: A novel dataset, models, and evaluation. In: Proc. of the 17th Conf. of the European Chapter of the Association for Computational Linguistics. Dubrovnik: Association for Computational Linguistics, 2023. 147–165. [doi: 10.18653/v1/2023.eacl-main.12]

[28] Du XY, Shao JR, Cardie C. Learning to ask: Neural question generation for reading comprehension. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017. 1342–1352. [doi: 10.18653/v1/P17-1123]

[29] Su D, Xu Y, Winata GI, Xu P, Kim H, Liu ZH, Fung P. Generalizing question answering system with pre-trained language model fine-tuning. In: Proc. of the 2nd Workshop on Machine Reading for Question Answering. Hong Kong: Association for Computational Linguistics, 2019. 203–211. [doi: 10.18653/v1/D19-5827]

[30] Wang WC, Feng S, Wang DL, Zhang YF. Answer-guided and semantic coherent question generation in open-domain conversation. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 5066–5076. [doi: 10.18653/v1/D19-1511]

[31] Yang JY, Dong YH, Qian JB. Research progress of few-shot learning methods based on graph neural networks. Journal of Computer Research and Development, 2024, 61(4): 856–876 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.202220933] [doi: 10.7544/issn1000–1239.202220933]

[32] Bao JW, Gong YY, Duan N, Zhou M, Zhao TJ. Question generation with doubly adversarial nets. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2018, 26(11): 2230–2239. [doi: 10.1109/TASLP.2018.2859777]

[33] Wang JY, Li JL, Zhao H. Self-Prompted chain-of-thought on large language models for open-domain multi-hop reasoning. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023. 2717–2731. [doi: 10.18653/v1/2023.findings-emnlp.179]

[34] Wang LY, Xu ZH, Lin ZB, Zheng HT, Shen Y. Answer-driven deep question generation based on reinforcement learning. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020. 5159–5170. [doi: 10.18653/v1/2020.coling-main.452]

[35] Kulshreshtha S, Rumshisky A. Reasoning circuits: Few-shot multi-hop question generation with structured rationales. In: Proc. of the 1st Workshop on Natural Language Reasoning and Structured Explanations. Toronto: Association for Computational Linguistics, 2023. 59–77. [doi: 10.18653/v1/2023.nlrse-1.6]

[36] Jia X, Wang H, Yin DW, Wu YF. Enhancing question generation with commonsense knowledge. In: Proc. of the 20th China National Conf. on Chinese Computational Linguistics. Huhhot: Springer, 2021. 145–160. [doi: 10.1007/978-3-030-84186-7_10]

[37] Li ZP, Cao Z, Li PF, Zhong Y, Li SB. Multi-Hop question generation with knowledge graph-enhanced language model. Applied Sciences, 2023, 13(9): 5765. [doi: 10.3390/app13095765]

[38] Yu WH, Zhu CG, Qin LH, Zhang ZH, Zhao T, Jiang M. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In: Findings of the Association for Computational Linguistics: ACL 2022. Dublin: Association for Computational Linguistics, 2022. 1896–1906. [doi: 10.18653/v1/2022.findings-acl.149]

[39] Yu JX, Wang SQ, Zheng LB, Su QL, Liu W, Zhao BQ, Yin J. Generating deep questions with commonsense reasoning ability from the text by disentangled adversarial inference. In: Findings of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics, 2023. 470–486. [doi: 10.18653/v1/2023.findings-acl.30]

[40] Zhao WM, Alwidian S, Mahmoud QH. Adversarial training methods for deep learning: A systematic review. Journal of Algorithms, 2022, 15(8): 283. [doi: 10.3390/a15080283]

[41] Gao YF, Bing LD, Chen W, Lyu MR, King I. Difficulty controllable generation of reading comprehension questions. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao, 2019. 4968–4974.

[42] Kumar V, Hua YC, Ramakrishnan G, Qi GL, Gao LL, Li YF. Difficulty-controllable multi-hop question generation from knowledge graphs. In: Proc. of the 18th Int'l Semantic Web Conf. Auckland: Springer, 2019. 382–398. [doi: 10.1007/978-3-030-30793-6_22]

[43] Bi S, Liu JY, Miao ZY, Min QZ. Difficulty-controllable question generation over knowledge graphs: A counterfactual reasoning approach. Information Processing & Management, 2024, 61(4): 103721. [doi: 10.1016/j.ipm.2024.103721]

[44] Yang KC, Deng JK, An X, Li JW, Feng ZY, Guo J, Yang J, Liu TL. ALIP: Adaptive language-image pre-training with synthetic caption.

In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 2910–2919. [doi: 10.1109/ICCV51070.2023.00273]

[45] Li MH, Lv TC, Chen JY, Cui L, Lu YJ, Florencio D, Zhang C, Li ZJ, Wei FR. TROCR: Transformer-based optical character recognition with pre-trained models. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI, 2023. 13094–13102. [doi: 10.1609/aaai.v37i11.26538]

[46] Dozat T, Manning CD. Deep biaffine attention for neural dependency parsing. arXiv:1611.01734, 2017.

[47] Du XY, Liu MW, Shen LW, Peng X. Survey on representation learning methods of knowledge graph for link prediction. Ruan Jian Xue Bao/Journal of Software, 2024, 35(1): 87–117 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6902.htm [doi: 10.13328/j.cnki.jos.006902]

[48] Xu FY, Shi WJ, Choi E. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. arXiv:2310.04408, 2023.

[49] Wang QF, Wang JG, Quan XJ, Feng FL, Xu ZL, Nie SL, Wang SN, Khabsa M, Firooz H, Liu DF. MUSTIE: Multimodal structural Transformer for Web information extraction. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics, 2023. 2405–2420. [doi: 10.18653/v1/2023.acl-long.135]

[50] Speer R, Chin J, Havasi C. ConceptNet 5.5: An open multilingual graph of general knowledge. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI, 2017. 4444–4451. [doi: 10.1609/aaai.v31i1.11164]

[51] Deng Z, Zhu Y, Chen Y, Witbrock M, Riddle P. Interpretable AMR-based question decomposition for multi-hop question answering. In: Proc. of the 31st Int'l Joint Conf. on Artificial Intelligence. Vienna, 2022. 4093–4099.

[52] Cao SY, Wang L. Controllable open-ended question generation with a new question type ontology. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and Proc. of the 11th Int'l Joint Conf. on Natural Language Processing. Association for Computational Linguistics, 2021. 6424–6439. [doi: 10.18653/v1/2021.acl-long.502]

[53] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu CW, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush A. Transformers: State-of-the-art natural language processing. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, 2020. 38–45. [doi: 10.18653/v1/2020.emnlp-demos.6]

[54] Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100, 000+ questions for machine comprehension of text. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016. 2383–2392. [doi: 10.18653/v1/D16-1264]

[55] Li XL, Liang P. Prefix-tuning: Optimizing continuous prompts for generation. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. Association for Computational Linguistics, 2021. 4582–4597. [doi: 10.18653/v1/2021.acl-long.353]

[56] Honnibal M, Montani I. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. Journal of To appear, 2017, 7(1): 411–420.

[57] Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. arXiv:1904.09751, 2020.

[58] Shaheer S, Hossain I, Sarna SN, Mehedi MHK, Rasel AA. Evaluating question generation models using qa systems and semantic textual similarity. In: Proc. of the 13th IEEE Annual Computing and Communication Workshop and Conf. Las Vegas: IEEE, 2023. 431–435. [doi: 10.1109/CCWC57344.2023.10099244]

[59] Zhang ZS, Wu YW, Zhou JR, Duan SF, Zhao H, Wang R. SG-Net: Syntax-guided machine reading comprehension. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 9636–9643. [doi: 10.1609/aaai.v34i05.6511]

[60] Rajpurkar P, Jia R, Liang P. Know what you don't know: Unanswerable questions for SQuAD. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 784–789. [doi: 10.18653/v1/P18-2124]

[61] He PC, Liu XD, Gao JF, Chen WZ. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv:2006.03654, 2021.

[62] Dhingra B, Liu HX, Yang ZL, Cohen W, Salakhutdinov R. Gated-attention readers for text comprehension. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017. 1832–1846. [doi: 10.18653/v1/P17-1168]

[63] Dara S, Srinivasulu CH, Babu CHM, Ravuri A, Paruchuri T, Kilak AS, Vidyarthi A. Context-aware auto-encoded graph neural model for dynamic question generation using NLP. ACM Trans. on Asian and Low-resource Language Information Processing, 2023. [doi: 10.1145/3626317]

[64] Seo MJ, Kembhavi A, Farhadi A, Hajishirzi H. Bidirectional attention flow for machine comprehension. arXiv:1611.01603, 2018.

[65] Zhao RC, Li XX, Joty S, Qin CW, Bing LD. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In: Proc. of the 61st

Annual Meeting of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics, 2023. 5823–5840. [doi: 10.18653/v1/2023.acl-long.320]

[66] Liu Z, Lin YT, Cao Y, Hu H, Wei YX, Zhang Z, Lin S, Guo BN. Swin Transformer: Hierarchical vision Transformer using shifted windows. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 9992–10002. [doi: 10.1109/ICCV48922.2021.00986]

[67] Lin TY, Goyal P, Girshick RB, He KM, Dollár P. Focal loss for dense object detection. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2999–3007. [doi: 10.1109/ICCV.2017.324]

[68] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[69] Lu JS, Batra D, Parikh D, Lee S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 13–23.

[70] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Pamela M, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In Proc. of the 38th Int'l Conf. on Machine Learning. 2021. 8748–8763.

[71] Neculoiu P, Versteegh M, Rotaru M. Learning text similarity with siamese recurrent networks. In: Proc. of the 1st Workshop on Representation Learning for NLP. Berlin: Association for Computational Linguistics, 2016. 148–157. [doi: 10.18653/v1/W16-1617]

[72] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision–ECCV. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[73] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RS, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: JMLR.org, 2015. 2048–2057.

[74] Hossain A, Karimuzzaman M, Hossain MM, Rahman A. Text mining and sentiment analysis of newspaper headlines. Information, 2021, 12(10): 414. [doi: 10.3390/info12100414]

[75] Alcântara C, Moreira V, Feijo D. Offensive video detection: Dataset and baseline results. In: Proc. of the 12th Language Resources and Evaluation Conf. Marseille: European Language Resources Association, 2020. 4309–4319.

[76] Ha Y, Kim J, Won D, Cha M, Joo J. Characterizing clickbaits on instagram. In: Proc. of the 12th Int'l AAAI Conf. on Web and Social Media. Stanford: AAAI, 2018. 92–101. [doi: 10.1609/icwsm.v12i1.15019]

[77] Potthast M, Gollub T, Komlossy K, Schuster S, Wiegmann M, Fernandez EPG, Hagen M, Stein B. Crowdsourcing a large corpus of clickbait on twitter. In: Proc. of the 27th Int'l Conf. on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018. 1498–1507.

[78] Shu K, Mahudeswaran D, Wang SH, Lee D, Liu H. FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data, 2020, 8(3): 171–188. [doi: 10.1089/big.2020.0062]

[79] Shu K, Mahudeswaran D, Wang SH, Liu H. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In: Proc. of the 14th Int'l AAAI Conf. on Web and Social Media. AAAI, 2020. 626–637. [doi: 10.1609/icwsm.v14i1.7329]

[80] Yang J, Vega-Oliveros D, Seibt T, Rocha A. Explainable fact-checking through question answering. In: Proc. of the 2022 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 8952–8956. [doi: 10.1109/ICASSP43922.2022.9747214]

[81] Wu Y, Zhan PW, Zhang YJ, Wang LM, Xu Z. Multimodal fusion with co-attention networks for fake news detection. In: Findings of the Association for Computational Linguistics. Association for Computational Linguistics, 2021. 2560–2569. [doi: 10.18653/v1/2021.findings-acl.226]

[82] Mowar P, Jain M, Goel R, Vishwakarma DK. Clickbait in youtube prevention, detection and analysis of the bait using ensemble learning. arXiv:2112.08611, 2021.

[83] Chen ZW, Hu LM, Li WX, Shao YX, Nie LQ. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics, 2023. 627–638. [doi: 10.18653/v1/2023.acl-long.37]

[84] Wang JP, Ge YX, Yan R, Ge YY, Lin KQ, Tsutsui S, Lin XD, Cai GY, Wu JP, Shan Y, Qie XH, Shou MZ. All in one: Exploring unified video-language pre-training. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 6598–6608. [doi: 10.1109/CVPR52729.2023.00638]

[85] Randolph JJ. Free-marginal multirater Kappa (multirater kfree): An alternative to Fleiss' fixed-marginal multirater Kappa. 2005. https://www.researchgate.net/publication/224890485_Free-Marginal_Multirater_Kappa_multirater_kfree_An_Alternative_to_Fleiss_Fixed-Marginal_Multirater_Kappa

[86] Viera AJ, Garrett JM. Understanding interobserver agreement: The Kappa statistic. Family Medicine, 2005, 37(5): 360–363.

附中文参考文献:

[7] 鞠天杰, 刘功申, 张倬胜, 张茹. 自然语言处理中的探针可解释方法综述. 计算机学报, 2024, 47(4): 733–758. [doi: 10.11897/SP.J.1016.2024.00733]

[9] 范怡帆, 邹博伟, 徐庆婷, 李志峰, 洪宇. 常识问答研究综述. 软件学报, 2024, 35(1): 236–265. http://www.jos.org.cn/1000-9825/6913.htm [doi: 10.13328/j.cnki.jos.006913]

[31] 杨洁祎, 董一鸿, 钱江波. 基于图神经网络的小样本学习方法研究进展. 计算机研究与发展, 2024, 61(4): 856–876. [doi: 10.7544/issn1000–1239.202220933]

[47] 杜雪盈, 刘名威, 沈立炜, 彭鑫. 面向链接预测的知识图谱表示学习方法综述. 软件学报, 2024, 35(1): 87–117. http://www.jos.org.cn/1000-9825/6902.htm [doi: 10.13328/j.cnki.jos.006902]

余建兴(1985－), 男, 博士, 副教授, 博士生导师, CCF 杰出会员, 主要研究领域为自然语言处理, 知识图谱, 机器推理.

饶洋辉(1986－), 男, 博士, 副教授, 博士生导师, CCF 高级会员, 主要研究领域为机器学习, 情感计算, 主题建模, 事件检测.

王世祺(1997－), 男, 博士生, 主要研究领域为多模态问答, 虚假信息检测.

苏勤亮(1986－), 男, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 深度学习, 统计学习, 贝叶斯分析, 自然语言处理.

陈祺(2001－), 男, 硕士生, 主要研究领域为自然语言处理, 数理问答.

印鉴(1968－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为人工智能, 机器学习, 大数据分析与处理, 数据库与数据挖掘.

赖韩江(1987－), 男, 博士, 副教授, 博士生导师, CCF 高级会员, 主要研究领域为多模态数据处理和分析, 视觉检索.