计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(3):239−247 [doi: 10.15888/j.cnki.csa.009776] [CSTR: 32024.14.csa.009776]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于大语言模型的回译式抄袭检测①

解　勉, 陈　刚, 余晓晗

(中国人民解放军陆军工程大学 指挥控制工程学院, 南京 210007)
通信作者: 陈　刚, E-mail: 13376067283@189.cn

摘　要: 随着信息技术的发展, 诸如借助翻译工具的回译式抄袭行为越发复杂隐蔽, 对抄袭检测方法提出了更高的要求. 为此, 提出一种基于提示工程 (prompt engineering) 的抄袭检测方法. 该方法通过设计提示词, 引导大语言模型 (large language model, LLM) 在语义层面关注句子文本中的潜在相似性, 能够有效识别出语义高度相似的内容. 首先, 回顾了现有的抄袭检测技术和提示工程的应用, 在此基础上设计基于提示工程的回译式抄袭行为检测流程. 其次, 设计提示模版, 通过合并缩减待检测句子对的方式, 提出句子压缩比的抄袭检测指标. 最后, 通过实验证明基于提示工程的抄袭检测方法与传统方法相比, 在检测回译式抄袭行为上具有显著优势.

关键词: 抄袭检测; 提示工程; 大语言模型; 回译

引用格式: 解勉,陈刚,余晓晗.基于大语言模型的回译式抄袭检测.计算机系统应用,2025,34(3):239–247. http://www.c-s-a.org.cn/1003-3254/9776.html

Back Translation Plagiarism Detection Based on Large Language Model

XIE Mian, CHEN Gang, YU Xiao-Han

(College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China)

Abstract: With the development of information technology, back translation plagiarism, such as through the use of translation tools, becomes increasingly complex and covert, posing higher requirements for plagiarism detection methods. For this reason, a plagiarism detection method based on prompt engineering is proposed. This method guides large language model (LLM) to pay attention to potential similarities in sentence texts at the semantic level by designing prompt words, which can effectively identify highly semantically similar content. Firstly, the existing plagiarism detection technologies and the application of prompt engineering are reviewed. Based on this, a backtracking plagiarism behavior detection process based on prompt engineering is designed. Secondly, a prompt template is designed to propose a plagiarism detection index based on sentence compression ratio by merging and reducing the pairs of sentences to be detected. Finally, experiments demonstrate that the plagiarism detection method based on prompt engineering has significant advantages over traditional methods in detecting back translation plagiarism behavior.

Key words: plagiarism detection; prompt engineering; large language model (LLM); back translation

随着互联网和数字出版业的迅猛发展, 大量的文本数据被生成和共享, 使得抄袭成本大幅降低, 抄袭乱象也层出不穷[1]. 同时, 随着人工智能的不断进步, 抄袭者采用的手段也变得更加复杂和隐蔽[2], 例如通过翻译工具掩盖抄袭痕迹, 即借用翻译工具将原文本翻译成另一种语言, 再重新翻译回原始语言, 这一操作称为“回译”, 在这一过程中通常会发生同义词替换、句子重组等现象, 从而在表面上使文本看上去与原作不同, 然而并没有真正改变原文的核心内容和信息. “回译”常用于检测翻译准确性, 但用于掩盖抄袭痕迹则违反了学术诚信, 也对传统的抄袭检测算法构成了挑战.

目前, 抄袭检测方法依赖于文本相似度检测算法,

① 收稿时间: 2024-08-06; 修改时间: 2024-08-27; 采用时间: 2024-09-10; csa 在线出版时间: 2025-01-16
CNKI 网络首发时间: 2025-01-17

通常会将句子两两比较, 设定相似度阈值, 将句子相似度与阈值作对比来判断是否存在抄袭嫌疑[3−5], 这些算法可以划分为两大类: 一类是基于表层字符匹配的方法, 另一类则是基于深入语义理解的方法[6]. 基于表层字符匹配的方法有许多较为成熟的算法, 例如 Jaccard 系数[7]、TF-IDF 算法[8]等, 但它们通常会忽视上下文信息, 也无法有效处理多义词和同义词的问题. 基于深入语义理解的方法主要是通过 BERT[9]、SimCSE[10]等深度学习模型得到表示整个文本语义的向量, 进而使用余弦相似度等方法计算两个句子向量之间的相似度, 该方法通常需要大量的标签数据和计算资源对模型进行预训练, 且训练好的模型通常泛化能力有限, 面对与训练数据分布不同的新领域时句嵌入表征能力可能会下降[11], 造成上下文理解受限, 难以应对实际情况中复杂多样的抄袭策略. 如图 1 所示, 原文本经过中文翻译之后再翻译回英文, 这一过程中出现了多处同义替换 (如图 1 中上标 1 所示)、主题前置 (如图 1 中上标 2 所示)、合并句子 (如图 1 中上标 3 所示) 等技巧使得翻译后的文本与原文本语义相同, 但表达方式、句子结构都发生了一定的变化, 涉及多种抄袭策略, 现有的抄袭检测方法难以应对这些挑战, 需要新的抄袭检测方法识别出更为隐蔽复杂的抄袭行为.

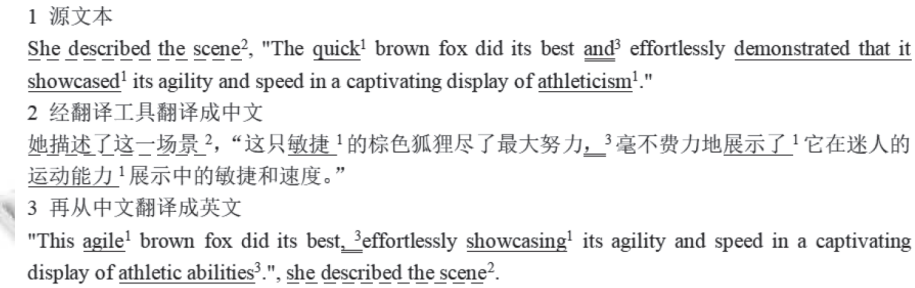

图 1 通过翻译工具掩盖抄袭的示例

随着 ChatGPT[12]等大语言模型的兴起, 自然语言处理的研究范式正逐步向通用人工智能转变, 不需要模型在特定的数据集上进行预训练和微调, 而是采用提示加问答的方式来完成各种任务, 在少样本学习 (few-shot learning) 甚至零样本学习 (zero-shot learning) 条件下依然具备深层次的语义理解和准确的上下文感知能力, 能够有效执行各类自然语言处理任务[13], 例如情感分析、机器阅读理解等, 充分展示了大语言模型在上下文理解和细微信息捕捉方面的显著能力. 相应的, 这些能力也是抄袭检测任务的基础.

因此, 针对通过翻译工具掩盖抄袭痕迹这一复杂的抄袭策略, 本文基于提示工程实现了一种准确且高效的抄袭检测方法, 将文档分句处理, 以最大程度减少大语言模型产生事实错误和幻觉问题为前提设计提示模版, 使得大语言模型可以从全局进行语义理解, 在提示引导下通过问答的方式判断原文句子与疑似抄袭的句子是否存在语义重复的内容, 作为衡量文本相似度的依据, 从而使最终的结果准确率超过 0.9. 这种方法不需要进行微调, 省去了大量标注数据的工作, 同时不依赖于特定的数据集训练, 使得在不改变方法流程的情况下, 可以用于识别其他抄袭策略.

1 相关工作

目前, 抄袭检测主要通过基于字符串、词向量、深度学习以及预训练模型的方法进行, 通常流程为使用某种方法计算两个文本之间的相似度, 并设定阈值, 以此作为判断两个文本是否存在抄袭行为的依据. 随着大语言模型的兴起和不断发展, 基于提示工程开发与优化提示词的方法应运而生, 为解决自然语言处理任务提供了新的方向.

基于字符串的方法通过计算字符重复度来判断抄袭嫌疑, 但忽视了单词的含义和上下文关系. Grozea 等[14]采用 N-gram 匹配方法将 16 个连续的字符看作整体反应两个文本序列之间的相似度; Ekbal 等[15]基于向量空间模型 (vector space model, VSM) 利用 TF-IDF 算法将文本表示成向量, 使用余弦相似度计算向量之间的距离, 以度量原文本和可疑文本之间的相似性, 作为抄袭的评判标准; 田星等[16]提出了基于词向量的 Jaccard 算法, 通过计算词向量相似度确定文本交集并计算 Jaccard 系数, 与阈值比较判断文本相似性, 改进了传统 Jaccard 算法; 申峻宇等[17]使用最小哈希 (MinHash) 构建局部敏感哈希 (locality sensitive hash, LSH) 算法计算哈希值选取可能相似的候选数据对, 再计算其 Jaccard 相似

度作为判断数据对相似的依据之一. 基于字符串的方法只关注字符表层, 融入词向量后虽然考虑到了词语含义, 但仍忽视了文本单元间的顺序和关系, 限制了算法精度和应用范围.

近年来, 深度学习和预训练模型迅速发展, 广泛应用于各种领域. 基于深度学习的方法一般通过神经网络提取文本的特征信息, 利用提取的特征比较不同文本之间的相似度, 最终得到抄袭检测结果. Ichida 等[18]采用基于 LSTM 的孪生神经网络 (Siamese neural network) 架构, 通过度量学习来衡量两个句子之间的语义相似性; Afzal 等[19]采用深度结构语义模型 (deep structured semantic model, DSSM) 提取文本的语义特征, 利用余弦相似度计算文本之间的语义相似度; 郭江华等[20]基于 SimCSE 提出一种改进的无监督句嵌入方法提高了模型的句嵌入表征能力, 利用余弦相似度计算两个句子的语义相似度; Wahle 等[21]对比分析了 5 个预训练词嵌入模型与机器学习分类器在抄袭检测任务上的有效性, 其中 Longformer 模型表现最优; Arabi 等[22]采用了 WordNet 和 FastText 预训练词嵌入网络形成的语义矩阵和加权 TF-IDF 的方法形成的结构矩阵, 计算可疑文本和源文本的语义和结构相似度, 并对两种相似度进行线性结合来判断抄袭现象; 王雪岭[23]通过单句语义特征提取模块和词汇特征提取模块有效提取文本的多种特征并利用 Bi-LSTM 序列模型融合, 结合上下文特征来检测发生抄袭的句子. 尽管基于深度学习的文本相似度检测算法表现出色, 但仍受训练数据集的影响, 需要精心收集构造数据以及花费大量时间进行标注, 对数据的质量以及多样性都有一定的要求, 且训练后通常泛化能力有限, 只适用于数据集相关的特定领域下.

基于预训练模型的方法是指在大量未标注的数据上进行初步训练, 以学习数据的通用特征和规律, 在应用于抄袭检测任务时通常会在文本相似度数据集上进行进一步的微调, 以提高模型表现[24], 有效缓解模型对数据集的依赖. Hambi 等人[25]提出一个 3 层抄袭检测系统, 前 2 层分别使用 doc2vec 和孪生 LSTM 对文本做表征学习, 第 3 层利用卷积神经网络 (convolutional neural network, CNN) 对是否存在抄袭行为作出判断; Qiu 等[26]使用 BERT 预训练模型提取短文的语义特征结合上下文加权数 (context tree weighting, CTW) 算法计算短文本语义相似度; Sujet 等[27]针对特定领域微调开源嵌入模型后显著增强了模型性能; Li 等[28]通过动态硬负样本挖掘和跨 GPU 平衡损失来提高文本嵌入模型的性能. 该类文本嵌入模型不仅考虑单词本身, 还考虑其在句子中的上下文, 能够更好地区分原文和其轻微变体, 从而提高抄袭检测的准确性, 但在理解复杂上下文和泛化能力方面可能有所不足. 随着规模的不断扩大, 预训练模型涌现出了惊人的泛化能力和零样本学习能力[29], 逐渐向大语言模型发展, 通过设计适当准确的提示词使得大语言模型生成高质量的文本内容, 并出现了提示工程 (prompt engineering) 这一关注提示词开发和优化的学科[30]. 例如胡志强等[31]通过大语言模型, 利用精心设计的提示词以问答和上下文推理的方式, 可以在零样本的情况下有效解决政策文本相似度高, 模型学到的特征向量相似度也高, 分类器无法准确预测政策工具的类别这一问题. 尽管提示工程在文本分类等领域得到了一定的关注和应用, 但其在抄袭检测方面的应用尚未被充分探索.

提示工程的出现改变了传统的预训练加微调的方式, 不再局限于提取文本特征信息后利用余弦相似度等方法衡量文本相似度这一抄袭检测方式, 也不再依赖标签数据, 而是可以通过设计和优化输入大语言模型的提示模版, 采用零样本学习的方式, 利用大语言模型强大的语义理解能力判断原文句子与疑似抄袭的句子在内容上是否存在语义相同的部分, 如果语义相同的部分超过一定阈值, 则判为抄袭. 此外, ChatGPT 等大语言模型目前仅能够通过在线 API 访问, 这种模式可能带来数据隐私泄露和结果复现性差等问题. 基于此, 本文采取了在本地部署大语言模型的方法, 以实现句子级别的抄袭检测.

2 研究方法

给定一段待检测的可疑文本可以看成一个句子集合$S=\{s_1,s_2,\cdots,s_m\}$, 与已有原文本集合$S'=\{s'_1,s'_2,\cdots,s'_n\}$中的句子两两配对, 形成新的句子对$s_is'_j(i=1,\cdots,m;j=1,\cdots,n)$, 抄袭检测的目的就是判断句子对是否构成抄袭. 把句子对作为提示模版的组成部分, 输入进大语言模型中进行判断. 对于两个句子, 从人类的逻辑角度来看, 如果语义完全相同, 那么实际上可以合并缩减为一个句子表达, 相较于合并前的句子对长度会大幅下降; 如果语义存在部分重叠, 则可以通过省略重复的部分并适当使用连词等方式表达句子原意, 合并后的

句子长度主要取决于重叠部分的多少, 重叠部分越多, 越有可能被判为抄袭, 同时合并前后句子长度变化也更明显; 而语义完全不同的两个句子, 则必须完整的保留各自的表述才能清晰的传达两个句子的信息, 相较于合并前句子对的长度基本不变. 基于此, 通过设计适当的提示模版使得大语言模型遵循这一逻辑将句子对合并, 保留语义不同的部分, 删除语义相同的部分, 通过比较合并前后句子长度的变化则可以判定句子对在语义层面的重复度, 这一变化可以用比值来衡量.

如图 2 所示, 整个流程可以分为句子对合并、后处理、抄袭判定等步骤. 首先将原文本和待检测文本切割成句子组成句子对 $s_i s'_j$ ($i = 1,\cdots,m; j = 1,\cdots,n$), 和提示模版一起输入大语言模型中, 通过设计提示模版中的提示词引导大语言模型在不丢失句子含义的前提下对句子进行合并、简化, 得到输出结果后进行后处理, 即通过对结果的分析识别大语言模型是否输出幻觉内容, 从而判断是否需要修改或者重新生成, 得到最终的合并结果 r_i; 最后, 将该句子长度 $l(r_i)$ 与合并前的句子对长度 $l(s_i s'_j)$ 进行对比, 使用句子压缩比的抄袭检测指标对该结果进行定量描述, 即 $l(r_i)/l(s_i s'_j)$, 衡量句子对合并简化前后长度的变化. 如果超过阈值 T 则判定为抄袭, 反之不存在抄袭行为.

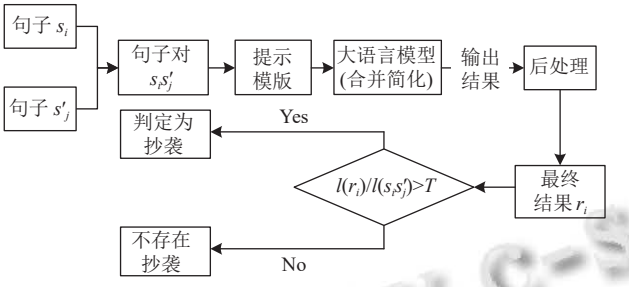

图 2 基于提示工程的抄袭检测流程图

2.1 提示模版的设计

随着模型规模和预训练数据集的不断扩大, 大语言模型能以问答方式执行人类指令, 抄袭检测范式正转向通用人工智能, 不再局限于特定任务, 而是多任务共享固定参数的大语言模型, 实现了更普适灵活的应用. 时宗彬等[32]提出基于提示的上下文学习 (in-context learning, ICL) 实现过程可以表示为式 (1), 同样适用于零样本学习的情况下.

$$P(Y|X_{\text{Prompt}} + X_{\text{Context}}) = \sum_i \log P(y_i|X_{\text{Prompt}} + X_{\text{Context}}) \tag{1}$$

对于本文而言, X_{Prompt} 表示针对具体任务设计的提示模版, X_{Context} 表示源句子和可疑句子组成的句子对, 和提示模版一起输入大语言模型中. 可以看出, 大语言模型的生成结果不仅和需要处理的对象相关, 也受提示模版的影响, 针对同一任务设计的不同的提示模版可能会得到差距很大的结果. 因此, 本文在不对模型参数进行微调的情况下, 提示模型的设计显著的影响句子对合并简化的结果, 应遵循一定的设计准则使得结果尽可能准确, 同时尽量减少大语言模型输出幻觉内容[33].

具体内容如图 3 所示. 其中, Instruct 为任务描述, 首先直接给出任务要求“请合并、简化下面两个句子” (见图 3 中①), 力求具体精炼, 减轻模型的理解负担; 其次, 细化任务, 要求大语言模型“在保留信息完整性的前提下删除重复信息” (见图 3 中②), 避免大语言模型受图 3 中①中的“简化”一词影响, 错误删除信息; 最后, 设计大语言模型自我检测结果是否满足“语义清晰、信息完整、无重复信息” (见图 3 中③) 的要求, 使得生成结果在保留两个句子语义完整的前提下, 将语义重复的信息删除, 只保留一处, 如果重复的信息过多, 则会造成合并前后的句子长度有显著变化, 利用该点可以进行抄袭检测. Input 为待合并简化的句子对. Output Format 为输出格式的设定, 设定“The combined sentence is: ”作为生成结果的固定开头, 可以再次强调此次任务为“合并句子”, 以减少大语言模型的幻觉输出, 也便于后续处理与使用生成结果.

[Instruct]
① Please merge and simplify the following two sentences. ② You should remove duplicate information while maintaining the integrity of the information. ③ Carefully review the merged sentences to ensure clear meaning, retain the information contained in both sentences, and express the information only once.
[Input]
{sentence1}
{sentence2}
[Output Format]
The output format is:
The combined sentence is:...

图 3 提示模版

2.2 后处理

在自然语言生成 (natural language generation, NLG) 及其下游任务中, 幻觉问题一直是研究的焦点. 大语言模型在海量数据上进行大规模的预训练, 但数据往往采集于互联网, 可能包含大量伪造的、过时的内容, 使得模型本身就记忆了错误的知识[34]; 同时, 区

别于传统模型只面向单一任务, 大语言模型可以应用于多任务、多语言、多领域的通用场景, 在生成过程中可能会错误关联不同部分的数据, 导致大语言模型可能生成与提示模版完全无关的内容, 最终影响抄袭检测的判断结果. 基于此, 本文根据实际应用中对存在幻觉问题的生成结果进行分析, 添加了后处理这一步骤, 设定以下 3 条规则进一步筛选处理结果.

(1) 生成结果没有按照提示模版中的输出格式进行输出, 重新利用大语言模型生成合并后的句子, 再次利用规则筛选. 违背提示模版中的任务指示是幻觉问题的典型表现之一, 生成的内容有潜在的风险, 可能夹杂着不真实的信息. 因此为了保证结果的准确性和可靠性, 必须采取适当的纠正措施.

(2) 生成结果中存在空行, 直接将空行以及空行后的内容删除. 经分析发现, 空行后的内容往往是大语言模型试图进一步阐述原因的附加信息. 虽然该部分内容与主句无关, 但会影响本文基于句子长度变化进行抄袭检测的判断. 因此, 为了确保判断的准确性, 删除这些与主句无关且位于换行符之后的内容.

(3) 生成结果中存在异常字符. 采用正则表达式匹配识别生成结果中是否存在与英文无关的字符, 例如中文, 如果有说明模型错误的关联了其他部分的数据, 存在幻觉问题, 需要重新生成.

规则具体内容及示例如表 1 所示. 在得到生成结果后, 将生成结果中的引导词“The combined sentence is:”删除, 与原句子对以单词为基本单位进行切分, 计算单词个数作为句子长度, 以此比较合并前后句子长度的变化, 将合并后的句子长度与原句子对长度之比作为一个特征, 比值越小, 说明经过合并操作后句子长度越小, 原句子对中语义重复的内容越多, 判定为抄袭的可能性越大, 反之亦然. 最后, 借助线性分类器进行超参搜索寻找最佳准确率, 并得到相应的阈值.

表 1 后处理规则内容及示例

具体表现	处理方式	示例
未按照提示模版中的输出格式进行输出	重新生成	“The woman was taken to Charing Cross Hospital and remains critically ill.”
生成结果存在空行	删除空行及空行后的内容	The combined sentence is: Nissan North America Inc. said Wednesday it was moving production of the Pathfinder sport utility vehicle to Tennessee, adding 800 jobs. Explanation: The combined sentence provides …
生成结果存在异常字符	重新生成	The combined sentence is: “Lewis, the WBC champion, can still fight June 21 at the斯台普斯中心in Los Angeles against another opponent.”

3 实验

3.1 实验数据

为了验证本文提出的基于提示工程的抄袭检测方法的有效性, 本文借助微软提出的 MSRP 数据集, 进一步构建了一个利用翻译工具以掩饰抄袭行为的数据集. MSRP 数据集是一个常用于释义识别的语义相似度公开数据集, 包含从互联网新闻源中提取的 4 076 个训练句子对, 以及人工注释, 指示每对是否捕获了语义等价关系, 数据集示例如表 2 所示. 其中, “Quality”表示句子是否语义等价, 取值为{0, 1}, 其中 0 表示句子的语义不等价, 1 表示句子的语义等价; “#1 String”和“#2 String”是待检测的句子对.

基于该数据集中的训练数据, 本文进行了两项调整以构建一个新的数据集, 旨在验证基于提示工程的方法针对那些利用翻译工具来隐蔽抄袭的策略的检测能力. 调整内容及原因如下.

(1) 去除“#1 String”和“#2 String”中包含数字的数据. 由于大语言模型自身模型架构设计和预训练目标的局限性, 生成结果的过程中更多地关注预测下一个词或句子的概率, 力求语言的流畅性和一致性, 而不是理解和计算数字等结构化数据的具体含义或数值之间的关系, 因此大语言模型对数字的准确性把握不足, 本文并未深入判别与数值相关的语义重复.

(2) 在 MSRP 数据集的基础上重新构建“Quality”为“1”和“0”的数据. 对于“Quality”为“1”的数据, 调用百度翻译 API 将“#1 String”翻译成中文, 再翻译回英文构成本文数据集中的“#2 String”, 模拟利用翻译工具来隐蔽抄袭的情况. 考虑到 “Quality”为“0”的数据是为了提高机器学习模型性能而特意设计的, 许多句子对都是十分相似的不同句子, 有很高的迷惑性. 本文对“Quality”为“0”的数据进行了改造, 从“Quality”为“1”的数据中选出句子替换掉“Quality”为“0”句子对中的“#2 String”, 得到完全不同的两个句子.

表 2 MSRP 和本文数据集示例

数据集名称	Quality	#1 String	#2 String
MSRP数据集	1	Amrozi accused his brother, whom he called “the witness”, of deliberately distorting his evidence.	Referring to him as only “the witness”, Amrozi accused his brother of deliberately distorting his evidence.
	0	Yucaipa owned Dominick’s before selling the chain to Safeway in 1998 for $2.5 billion.	Yucaipa bought Dominick’s in 1995 for $693 million and sold it to Safeway for $1.8 billion in 1998.
本文数据集	1	Amrozi accused his brother, whom he called “the witness”, of deliberately distorting his evidence.	Amrozi claimed that he was just a “witness”, accusing his brother of deliberately distorting his evidence. 来源: 调用百度翻译API将“#1 String”翻译成中文后再翻译回英文
	0	Legislation making it harder for consumers to erase their debts in bankruptcy court won overwhelming House approval in March.	Amrozi accused his brother, whom he called “the witness”, of deliberately distorting his evidence. 来源: “Quality”为“1”的“ #1 String”

经过上述调整, 原始数据集转变为专门用于检测通过翻译工具进行抄袭的数据集, 更加符合实际应用场景, 在不改变大语言模型参数的情况下, 更好地模拟通常情况下使用翻译工具进行抄袭的行为, 探索更通用的方法有效的识别出这一学术不端的行为. 调整后的新数据集共 2 518 条数据, 其中“Quality”为“1”的数据有 1 812 条, “Quality”为“0”的数据有 706 条. MSRP 和本文数据集示例如表 2 所示.

3.2 实验配置

实验使用了本地部署的 ChatGLM2-6B 作为主干大语言模型, 通过提示工程与大语言模型交互进行实验测试, 利用逻辑回归分类器 (logistic regression, LR) 使用网格搜索算法进行超参数寻优得到最优准确率为 92%, 此时阈值为 0.8, 同样的对比模型的相似度阈值也设为 0.8.

为了验证本文提出的方法在识别利用翻译工具隐藏抄袭行为的策略上的性能表现, 本文采用宏平均值 (*Macro-F*1) 和准确率 (*accuracy*) 作为实验的评价指标, 其定义如下.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$precision = \frac{TP}{TP+FP},\ recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2\times precision\times recall}{precision+recall}$$

$$Macro\text{-}F1 = \frac{1}{C}\sum_{i=1}^{c} F1_i$$

其中, TP为模型预测为正类, 即“Quality”为“1”, 实际上也是正类的样本数量; TN为预测为负类, 即“Quality”为“0”, 实际上也是负类的样本数量; FP为预测为正类但实际上为负类的样本数量; FN为预测为负类但实际上为正类的样本数量. C代表类别总数, $F1_i$代表第i个类别的 $F1$ 值.

对比模型参考了抄袭检测领域的相关文献, 选取了其中基于字符串的算法、基于词向量的算法、基于大规模语料库的预训练模型以及基于深度学习的模型作为基线模型进行对比实验. 具体的, 选取的基线模型如下.

(1) Jaccard 算法[7]: 是衡量两集合相似度的经典指标, 通过计算交集与并集元素数量之比来量化相似性. 该系数认为两个句子相似度取决于共有词汇多少, 特别适用于评估文本中句子的相似程度.

(2) MinHash and LSH[35]: 是一种 LSH 算法的具体实现方法, 用于在高维空间中快速搜索相似项, 主要思想是通过将集合映射到 MinHash 签名来进行相似度搜索.

(3) TF-IDF[36]: 是一种在文本挖掘和信息检索领域中广泛使用的加权技术, 通过词频和逆文档频率衡量词语在特定文档中的重要性. 单词在特定文档中出现次数多且在其他文档中少, 则更具代表性. 基于此计算权重, 用特征和权值表示文本向量, 再计算余弦相似度, 得出两个文本的相似度.

(4) Longformer[21]: 能够有效检测到机器改写的文本, 对比 5 种预训练词嵌入模型以及 Bart 等 8 个神经语言模型表现最佳.

(5) BERT[9]: 是一个自编码语言模型, 通过两个任务进行预训练, 可以支持阅读理解等多种下游任务. 该模型能够在多个层次上提取特征, 得到表示整个句子语义的向量, 对文本向量进行进一步的计算处理即可得到文本相似度. 本文选择基于 BERT-base-uncased 模型在 paws 数据集上进一步微调得到的 nlp_bert_sentence-similarity_english-base 模型作为对比模型之一.

(6) SimCSE[10]: 是一种通过对比学习方法进行句子嵌入的模型, 可以在有监督和无监督的情况下生成高质量的句子向量, 主要利用对比学习和 dropout 来提高句子表示能力. 本文选择 unsup-simcse-bert-base-uncased 这一无监督模型作为对比模型之一, 该模型的训练数据为从英语维基百科中抽取的 100 万个句子, Batch size 为 64, Learning rate 为 3E–5.

(7) Sentence-T5[37]: 是基于 T5 (text-to-text transfer Transformer) 模型的一种变体, 专门用于生成句子的嵌入表示, 这些嵌入可以用于语义文本相似性、搜索和同义词挖掘等任务. 本文选择 sentence-t5-large-quora-text-similarity 模型作为对比模型之一, batch size 为 8, learning rate 为 2E–5.

(8) MEEB[27]: 全称 Marsilia-Embeddings-EN-Base, 基于 BGE-base-en-v1.5 模型在特定领域进一步微调得到的文本嵌入模型, 具有强大通用性的同时针对特定领域的语言特点进行了优化和精细调整.

3.3 实验结果与分析

为了验证本文提出的方法在检测特定抄袭策略任务上的效果, 本文在构建的数据集上与 6 个模型进行详细的对比实验, 对比模型包括基于字符串的方法以及各类预训练模型. 表 3 中展示了具体的实验结果.

表 3 实验结果

指标	Jaccard	MinHash and LSH	TF-IDF	Longformer	BERT	SimCSE	Sentence-T5	MEEB	ours
accuracy (%)	28.83	29.11	40.67	71.68	83.71	84.63	86.93	91.62	92.30
*Macro-F*1	13.93	14.50	35.11	60.31	68.46	85.37	87.52	91.90	92.33

由表 3 可知, 在检测使用翻译工具掩盖抄袭行为的策略时, Jaccard、MinHash and LSH、TF-IDF 等传统的基于字符串的方法准确率和宏平均值均低于 50%, 显著低于其他深度学习模型, 表明了该方法主要关注字面上的字符匹配以及依赖于表面的文本特征, 例如字符或词汇的出现频率, 可能无法识别出词汇或短语之间同义或多义的关系, 导致即使是表达相同意思的不同词汇或句子, 仅因为字面上的不同而被判定为不相似. 同样地, 当两个句子在用词上高度相似但语义不同时, 基于字符串的方法也可能会忽略这种语义层面的差异性, 导致误判.

Longformer 模型通过结合窗口局部与全局自注意力机制, 提高了对不同重写工具生成文本的泛化能力, 但专注于长距离依赖可能会忽视局部信息的捕捉, 对于依赖细节匹配的抄袭检测可能不够敏感. paws 数据集是谷歌发布的一个支持英语的释义识别对抗性数据集, 针对词序、句法结构导致的语义改变专门设计了格式良好、具有高度重叠词汇的句子对, 可以提高模型对抄袭检测任务的精度, 对比模型中选取的 BERT 预训练模型是基于 BERT-base-uncased 模型在 paws 数据集上进一步微调得到的, 但仍略次于未经特定数据集微调的本文模型. 本文在准确率和宏平均值上相较于该模型分别提高了 8.59% 和 23.87%, 证明了本文在不依赖特定的标签数据训练的情况下, 仅利用提示工程采用零样本学习的方法也可以有效识别抄袭现象. 本文方法相较于 SimCSE 和 Sentence-T5 在准确率和宏平均值上均提高了 4.8% 以上, 其中 SimCSE 利用对比学习和 dropout 相较于提升句子向量表示质量, 能够在没有标注数据的情况下有效地从大量文本中学习到句子的语义表示; Sentence-T5 利用 T5 模型的预训练能力来生成句子的向量表示, 其中效果最好的方法是对 T5 的 encoder 输出进行平均池化 (mean pooling), 均优于 BERT 预训练模型, 但仍略逊于本文提出的方法. 这表明, 设计良好的提示词可以充分利用大语言模型强大的语义理解和生成能力, 深入理解上下文和复杂的概念, 尤其是在处理复杂或模糊的语义关系时. 相比之下, SimCSE 以及 Sentence-T5 可能无法达到同等级别的语义理解深度.

Marsilia-Embeddings-EN-Base 是一个基于通用文本嵌入模型训练得到的英文预训练嵌入模型, 继承了 BGE 的优良特性, 并且针对特定的语境进行了优化和调整, 以适应文本处理的需求, 但在准确率和宏平均值上稍次于本文模型, 这表明微调模型旨在提升特定领域的性能, 而大语言模型通过提示词的设计, 可以更灵活地适应不同的任务和场景, 同时提供了一种更加经济有效的问题解决方案, 节省了大量的计算资源和时间.

4 实例分析

为了进一步分析本文提出的基于大语言模型利用提示工程在抄袭检测任务上的有效性, 本文在对比模型未准确识别出是否为抄袭的句子对中挑选了 4 条样例, 对比结果如表 4 所示.

表 4 实例分析

序号	例句	标签
1	She told the jury, “So Sebastian did his best and convincingly admitted that he did not commit any crimes for the sake of survival.”. “So Sebastian did his best to convincingly confess to a crime that he didn't commit in order to survive,” she told jurors.	1
2	Randall Simon faced off on the right side in the second round of the competition. Randall Simon singled to right for the second run.	1
3	ICANN has criticised the changes and asked VeriSign to voluntarily suspend them. The company has expanded into providing other services for buyers, including payment services.	0
4	A former teammate, Carlton Dotson, has been charged with the murder. A federal magistrate in Fort Lauderdale ordered him held without bail.	0

由表 4 可知, 选取的标签为“1”的例句虽然在主题和表达内容上一致, 但经过翻译工具反复加工后, 使得词语使用、句子结构都发生了一定的变化, 例如第 1 对例句中“for the sake of survival”被替换成“in order to survive”, 主句“She told the jury”也被移到了句子最后, 涉及同义替换等多种抄袭策略, 而本文提出的方法充分利用了大语言模型强大的上下文理解能力和一定的推理能力, 可以在句子级别上有效检测出这种复杂多变的抄袭现象. 选取的标签为“0”的例句, 尽管这两句话的主题和内容不同, 但它们可能在某些方面存在相似之处, 例如第 3 对例句中都涉及公司、服务和业务等词汇. 这种相似性可能导致 BERT 等预训练模型在某种程度上认为它们是重复的. 然而, 从人类的角度来看, 这两句话的语义确实是不同的.

通过以上实例分析, 可以看出本文利用提示工程在不进行微调的情况下, 借助大语言模型强大的上下文语义理解能力和生成能力, 能够有效识别潜在的语义重复现象, 从而得到正确的抄袭判断结果.

5 结束语

本文采用巧妙设计的提示工程, 在未经过特定数据集微调的情况下, 使用本地大语言模型实现了对借助翻译工具掩盖抄袭行为的有效检测. 首先, 通过精心设计的提示词将原句和疑似抄袭的句子进行合并简化, 进而比较合并前后句子长度的变化; 然后将合并前后句子长度的比值作为重要特征输入线性分类器中, 寻找最佳阈值. 相较于传统的微调方法, 采用对话式的方式更为简便高效, 并且无需依赖标注数据, 为抄袭检测提供了一种新的方向.

参考文献

1 陈滔, 张庆国, 何金波, 等. 基于多算法融合的文本抄袭检测的特征提取算法研究. 湖北民族大学学报 (自然科学版), 2022, 40(1): 67–72. [doi: 10.13501/j.cnki.42-1908/n.2022.03.011]

2 刘宏更. 基于小样本学习的文档查重系统的设计与实现 [硕士学位论文]. 北京: 北京邮电大学, 2023. [doi: 10.26969/d.cnki.gbydu.2023.000915]

3 Vrbanec T, Meštrović A. Corpus-based paraphrase detection experiments and review. Information, 2020, 11(5): 241. [doi: 10.3390/info11050241]

4 雷歆, 周蕾越, 周兰江. 融合语法及结构特征的汉老双语句子相似度计算方法. 中文信息学报, 2023, 37(9): 73–82.

5 丁海兰, 祁坤钰. 基于 TextRank 算法和相似度的中文文本主题句自动提取. 吉林大学学报 (工学版), 1–9. https://doi.org/10.13229/j.cnki.jdxbgxb.20240121. [2024-05-06].

6 魏嵬, 丁香香, 郭梦星, 等. 文本相似度计算方法综述. 计算机工程, 2024, 50(9): 18–32. [doi: 10.19678/j.issn.1000-3428.0068086]

7 Real R, Vargas JM. The probabilistic basis of Jaccard’s index of similarity. Systematic Biology, 1996, 45(3): 380–385. [doi: 10.1093/sysbio/45.3.380]

8 杨宏伟, 张红梅, 张骥, 等. 基于 TF-IDF 加权文本语义相似度算法的变电站一键顺控测试方法研究. 电力科学与技术学报, 2023, 38(5): 269–278. [doi: 10.19781/j.issn.1673-9140.2023.05.028]

9 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

10 Gao TY, Yao XC, Chen DQ. SimCSE: Simple contrastive learning of sentence embeddings. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. ACL, 2021. 6894–6910. [doi: 10.18653/v1/2021.emnlp-main.552]

11 Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 3982–3992. [doi: 10.18653/v1/D19-1410]

12 Ouyang L, Wu J, Jiang X, *et al*. Training language models to follow instructions with human feedback. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc.,

2022. 2011.
13 Kojima T, Gu SS, Reid M, *et al*. Large language models are zero-shot reasoners. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1613.
14 Grozea C, Popescu M. Encoplot-tuned for high recall (also proposing a new plagiarism detection score). Proceedings of the 2012 Conference and Labs of the Evaluation Forum. Rome: Springer, 2012.
15 Ekbal A, Saha S, Choudhary G. Plagiarism detection in text using vector space model. Proceedings of the 12th International Conference on Hybrid Intelligent Systems (HIS). Pune: IEEE, 2012. 366–371. [doi: 10.1109/HIS.2012.6421362]
16 田星, 郑瑾, 张祖平. 基于词向量的 Jaccard 相似度算法. 计算机科学, 2018, 45(7): 186–189. [doi: 10.11896/j.issn.1002-137X.2018.07.032]
17 申峻宇, 李东闻, 钟震宇, 等. 一种基于局部敏感哈希的文本数据去重算法及其实现. 南开大学学报 (自然科学版), 2023, 56(6): 29–35.
18 Ichida AY, Meneguzzi F, Ruiz DD. Measuring semantic similarity between sentences using a Siamese neural network. Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro: IEEE, 2018. 1–7. [doi: 10.1109/IJCNN.2018.8489433]
19 Afzal N, Wang YS, Liu HF. MayoNLP at semeval-2016 task 1: Semantic textual similarity based on lexical semantic net and deep learning semantic model. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego: ACL, 2016. 674–679. [doi: 10.18653/v1/S16-1103]
20 郭江华, 苑迎春, 王克俭, 等. 基于改进 SimCSE 的无监督句嵌入方法. 计算机工程与设计, 2023, 44(8): 2382–2388. [doi: 10.16208/j.issn1000-7024.2023.08.019]
21 Wahle JP, Ruas T, Foltýnek T, *et al*. Identifying machine-paraphrased plagiarism. Proceedings of the 17th International Conference on Information. Cham: Springer, 2022. 393–413. [doi: 10.1007/978-3-030-96957-8_34]
22 Arabi H, Akbari M. Improving plagiarism detection in text document using hybrid weighted similarity. Expert Systems with Applications, 2022, 207: 118034. [doi: 10.1016/j.eswa.2022.118034]
23 王雪岭. 基于多特征提取和多句融合的抄袭检测方法 [硕士学位论文]. 杭州: 浙江工商大学, 2023. [doi: 10.27462/d.cnki.ghzhc.2023.001078]
24 苏静, Ahmed M. 一种基于半监督的句子情感分类模型. 重庆大学学报, 1–16. http://kns.cnki.net/kcms/detail/50.1044.n.20240423.1922.004.html. [2024-09-10].
25 Hambi EM, Benabbou F. A new online plagiarism detection system based on deep learning. International Journal of Advanced Computer Science and Applications, 2020, 11(9): 470–478. [doi: 10.14569/ijacsa.2020.0110956]
26 Qiu SJ, Niu Y, Li J, *et al*. Research on semantic similarity of short text based on BERT and time WarPing distance. Journal of Web Engineering, 2021, 20(8): 2521–2544. [doi: 10.13052/jwe1540-9589.20814]
27 Sujet AI, Allaa B, Hamed R. Marsilia-Embeddings-EN-Base: A fine-tuned English embedding model for financial texts. https://hf-mirror.com/sujet-ai/Marsilia-Embeddings-EN-Base. [2024-09-02].
28 Li SY, Tang Y, Chen SZ, *et al*. Conan-embedding: General text embedding with more and better negative samples. arXiv:2408.15710, 2024.
29 Xu HW, Chen YJ, Du YL, *et al*. ZeroPrompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. Proceedings of the 2022 Findings of the Association for Computational Linguistics (EMNLP 2022). ACL, 2022. 4235–4252. [doi: 10.18653/v1/2022.findings-emnlp.312]
30 方海光, 王显闯, 洪心, 等. 面向 AIGC 的教育提示工程学习提示单设计及应用. 现代远距离教育, 2024(2): 62–70. [doi: 10.13927/j.cnki.yuan.20240509.002]
31 胡志强, 李朋骏, 王金龙, 等. 基于 ChatGPT 增强和监督对比学习的政策工具归类研究. 计算机工程与应用, 2024, 60(7): 292–305. [doi: 10.3778/j.issn.1002-8331.2308-0354]
32 时宗彬, 朱丽雅, 乐小虬. 基于本地大语言模型和提示工程的材料信息抽取方法研究. 数据分析与知识发现, 2024, 8(7): 23–31. [doi: 10.11925/infotech.2096-3467.2023.1119]
33 Zhang Y, Li YF, Cui LY, *et al*. Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv:2309.01219, 2023.
34 徐磊, 胡亚豪, 潘志松. 针对大语言模型的偏见性研究综述. 计算机应用研究, 2024, 41(10): 2881–2892. [doi: 10.19734/j.issn.1001-3695.2024.02.0020]
35 Hwang WS, Park J, Kim SW. A method for recommending the latest news articles via MinHash and LSH. Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication. Bali Indonesia: ACM, 2015. 60. [doi: 10.1145/2701126.2701205]
36 Robertson SE, Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Springer, 1994. 232–241. [doi: 10.1007/978-1-4471-2099-5_24]
37 Ni JM, Abrego GH, Constant N, *et al*. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In: Muresan S, Nakov P, Villavicencio A, eds. Findings of the Association for Computational Linguistics: ACL 2022. 2022. 1864–1874.

(校对责编: 张重毅)