

情报理论与实践
Information Studies: Theory & Application
ISSN 1000-7490, CN 11-1762/G3

《情报理论与实践》网络首发论文

题目：基于大语言模型的开源情报摘要生成研究
作者：汪伦，艾斯卡尔·艾木都拉，张华平，耿国桐，吕雁飞，雷帅
网络首发日期：2025-02-14
引用格式：汪伦，艾斯卡尔·艾木都拉，张华平，耿国桐，吕雁飞，雷帅. 基于大语言模型的开源情报摘要生成研究[J/OL]. 情报理论与实践.
<https://link.cnki.net/urlid/11.1762.G3.20250214.1507.002>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

●汪 伦^{1,2}, 艾斯卡尔·艾木都拉¹, 张华平³, 耿国桐², 吕雁飞², 雷 帅²

(1.新疆大学计算机科学与技术学院, 新疆 乌鲁木齐 830046; 2.军事科学院军事科学信息研究中心, 北京 100142; 3.北京理工大学计算机学院, 北京 100081)

基于大语言模型的开源情报摘要生成研究

摘 要: [目的/意义] 大语言模型在生成类文本问题上取得显著进展, 为开源情报领域研究人员提供了新思路。将大语言模型应用到开源情报领域, 以探索其在该领域的摘要生成能力。[方法/过程] 通过构建开源情报摘要生成数据集, 使用 Qwen1.5-14B 模型为基座模型, 并利用 LoRA 技术对其进行微调训练, 得到了 Qwen1.5-OSINT 模型。[结果/结论] 将 Qwen1.5-OSINT 模型在开源情报摘要生成任务中与 Qwen1.5-14B 模型进行对比, 结果显示, Qwen1.5-OSINT 模型的 ROUGE-1 指标提升了 0.3842, ROUGE-2 指标提升了 0.2021, ROUGE-L 指标提升了 0.2494, BLEU 指标提升了 0.1348。经微调后的模型在多项评价指标上均超越了基座模型, 展示了出色的摘要生成能力。[局限] 尽管微调后的 Qwen1.5-OSINT 模型在效果上有所提升, 但其依赖的高质量数据集建设仍需大量的人工投入。此外, 模型的泛化能力在不同领域上可能受到限制, 需要进一步的研究和探索。

关键词: 大语言模型; 开源情报; LoRA 微调; 摘要生成

Research on Open Source Intelligence Summary Generation Based on Large Language Models

Wang Lun^{1,2}, Askar Hamdulla¹, Zhang Huaping³, Geng Guotong², Lü Yanfei², Lei Shuai²
(1.School of Computer Science and Technology, Xinjiang University, Xinjiang Urumqi 830046; 2.Institute of Military Science Information, Academy of Military Sciences, Beijing 100142; 3.School of Computer Science, Beijing Institute of Technology, Beijing 100081)

Abstract: [Purpose/significance] Large language models have made significant progress in the problem of generating text, providing new ideas for researchers in the field of open source intelligence. This paper applies large language models to the field of open source intelligence to explore its ability to generate summaries in this field. [Method/process] By constructing an open source intelligence summary generation dataset, using the Qwen1.5-14B model as the base model, and fine-tuning it using the LoRA technology, the Qwen1.5-OSINT model was obtained. [Result/conclusion] The Qwen1.5-OSINT model was compared with the Qwen1.5-14B model in the open source intelligence summary generation task. The results showed that the ROUGE-1 index of the Qwen1.5-OSINT model increased by 0.3842, the ROUGE-2 index increased by 0.2021, the ROUGE-L index increased by 0.2494, and the BLEU index increased by 0.1348. The fine-tuned model surpassed the base model in many evaluation indicators, demonstrating excellent summary generation capabilities. [Limitations] Although the fine-tuned Qwen1.5-OSINT model has improved in performance, the construction of the high-quality dataset it relies on still requires a lot of manual effort. In addition, the generalization ability of the model may be limited in different fields, which requires further research and exploration.

Keywords: large language models; open source intelligence; LoRA fine-tuning; summary generation

0 引言

当今信息化时代, 开源情报已成为获取信息、洞察趋势和辅助决策的重要手段。它涉及从公共渠道, 如互联网、社交媒体、新闻媒体等搜集可公开访问的信息, 并对信息进行分析及利用。但随着数据量的指数级增长, 手动筛选和分析这些信息变得日益困难。因此, 以凝练简洁的语言表达开源情报的核心内容显得尤为重要。这一过程旨在自动为开源情报生成中文简讯, 以支持快速、精准的情报分析和决策制定。

现如今对开源情报的自动化处理主要是运用统计分析、数据挖掘、机器学习等方法, 但在处理大规模、高维度、多源异构数据时存在效率低、准确度有限、难以发现深层次关联等问题。大语言模型(以下简称大模型)^[1-3]的问世, 为各类任务带来了前所未有的机遇, 因此本文尝试将大模型应用到开源情报领域, 提升开源情报处理的性能。

大模型展示了惊人的文本生成和理解能力, 在生成类任务上取得了重大突破, 尤其是在通用知识和广泛技能上表现尤为出色。但这种通过大规模人类通用语料训练的模型, 在处理

开源情报领域数据时面临一定的局限性,如模型可能无法提供精确、专业的军事类知识。这是由于模型在训练时,缺乏开源情报领域相关的数据。

此外,鉴于训练一个特定领域的大模型需要耗费大量的资源、时间和经济成本,我们利用通用大模型丰富的特征表示及较强的泛化能力,采用微调的方式,只需用相对较小的领域数据集对模型进行训练,就可以快速适应新任务,达到较好的性能表现。

针对以上问题,本文利用网上收集的开源情报领域数据构建对话形式数据集 OSINTD (Open Source Intelligence Dataset),通过微调 Qwen1.5-14B 模型,提出了一种用于开源情报领域的大模型 Qwen1.5-OSINT,并在 Qwen1.5-OSINT 模型的基础上设计了简讯生成插件,该插件主要实现一键点击为英文的原始页面生成中文的简讯,以此提高情报工作人员的工作效率,革新情报收集的方式。

综上,本文的主要贡献有以下三点:①构建了高质量开源情报领域的摘要生成数据集 OSINTD。②基于 OSINTD 数据集微调训练了一个开源情报领域的大模型 Qwen1.5-OSINT 模型。③基于 Qwen1.5-OSINT 模型设计了简讯整编插件,旨在提高情报搜集工作的便捷度,同时加快情报搜集效率。

1 研究现状

2013 年 Google Brain 项目发布了深度学习模型 DistBelief^[4],为大规模分布式训练奠定基础。2014 年,被誉为 21 世纪最强大算法模型之一的对抗式生成网络(GAN^[5])诞生,标志着深度学习进入了生成模型研究的新阶段。2017 年,Google 又提出了 Transformer^[6],Transformer 架构是当前大模型领域主流的算法架构基础,由此形成了 GPT^[7]和 BERT^[8]两条主要的技术路线,其中 BERT 最有名的落地项目是谷歌的 AlphaGo^[9]。在 GPT 3.0 发布后,GPT 逐渐成为大模型的主流路线。综合来看,当前几乎所有参数规模超过千亿的大型语言模型都采取 GPT 模式,如百度文心一言,阿里发布的通义千问等。

从应用领域上看,大模型可分为通用大模型和行业大模型两种。通用大模型是具有强大泛化能力,可在不进行微调或少量微调的情况下完成多场景任务,相当于 AI 完成了“通识教育”,ChatGPT、华为的盘古都是通用大模型。行业大模型则是利用行业知识对大模型进行微调,让 AI 完成“专业深造”,以满足在金融、法律、医学等不同领域的需求。

在金融领域,彭博社(Bloomberg)和约翰霍普金斯大学的研究人员构建了一个 3630 亿个 token 的数据集,并增加了来自通用数据集的 3450 亿个 token,经过训练得到一个 500 亿参数的语言模型 BloombergGPT^[10];哥伦比亚大学和纽约大学(上海)的研究人员运用了各种各样的金融训练数据,包括金融新闻、公司公告、社交媒体、金融专业网站、学术资料等构建了一个针对金融行业的开源大型语言模型 FinGPT^[11];复旦大学数据智能与社会计算实验室基于其构建的高质量金融数据集 DISC-Fin-SFT 在通用领域中文大模型 Baichuan-13B-Chat^[12]上进行 LoRA (Low Rand Adaption) 指令微调得到的金融大模型 DISC-FinLLM^[13];度小满在 BLOOM-176B^[14]的基础上针对中文通用领域和金融领域进行了针对性的预训练与微调得到了轩辕大模型^[15]等。

在法律领域,北京大学团队使用论文、新闻、法规、司法解释、法律咨询、法考题、判决书,随后经过清洗、数据增强等来构造训练数据,在 LLaMA 基础上训练得到 ChatLaw^[16]大模型;清华大学利用 ChatGLM^[17]作为基础模型,使用法律问答数据、法律法规和法律参考书籍、法律文书进行微调训练得到 LexiLaw^[18]大模型等。

在医学领域,哈尔滨工业大学团队利用公开和自建的中文医学知识库构建中文医学指令数据集,在 LLaMA 的基础上微调训练得到 Huatuo^[19]大模型;上海科技大学团队利用 ChatGLM 作为基础模型,通过中文医疗对话数据集进行微调,得到 DoctorGLM^[20]大模型;美国德克萨斯大学西南医学中心等 LLaMA 的基础上使用了包含 10 万多真实患者与医生对话数据集 HealthCareMagic,包含 1 万多患者与医生对话数据集 icliniq,包含 5 千多由 GPT 生成的医患对话数据集 GenMedGPT 进行微调训练得到了 ChatDoctor^[21]大模型。

大模型在各个领域的成功应用证明了其强大的学习能力,基于此我们探索大模型在开源情报领域的应用能力。

2 研究方法

为了探索大模型在开源情报领域中的应用能力,本文使用 Qwen1.5-14B 作为基座模型,从网络上收集与军事相关的信息,如军事新闻和军事报告,经过清洗、人工标注等过程,将

非结构化数据处理成对话问答形式的结构化数据，共计 1 万余条，将此数据作为微调的训练数据集，在微调的过程中使用了 LoRA 微调策略来加速微调训练。整个训练过程如图 1 所示。



图 1 训练过程

Fig.1 Training course

2.1 基线模型

本文使用 Qwen1.5-14B 模型作为基座模型，以此探索大模型在开源情报领域中的摘要生成能力。Qwen1.5-14B 使用了基于 tiktoken 的 tokenizer，能够更高效地减少模型的参数量，提高训练和推理的效率。同时 Qwen1.5-14B 加入了 SwiGLU 激活，有助于缓解梯度消失问题。采用多头注意力机制，能够有效捕获不同位置的输入序列之间的多种相关性，增强对上下文的理解能力，这对于生成连贯、语义丰富的文本至关重要。此外，Qwen1.5-14B 位置编码使用的是旋转位置编码，通过复数运算动态调整键值对的表示，以模拟序列中元素的相对位置关系。这种方法在长序列中表现得尤为有效，由于旋转角度是累加的，随着序列长度的增加，旋转的角度也会逐渐累积，旋转位置编码的显著优点是它能够有效地捕捉长程依赖性，从而使得模型能够更容易地识别出远距离位置之间的关系，有助于模型更好地理解 and 生成具有长程依赖性的文本，提高了模型在处理复杂语言结构时的性能。

正是这些设计上的优化和创新，让 Qwen 1.5-14B 具备了强大的语言理解和生成能力，能够在诸如文本生成、对话、问答等多种自然语言处理任务中展现出优越的性能。故我们采用 Qwen1.5-14B 模型对其进行微调，以期能够在开源情报领域展现出较好的效果。

2.2 数据集

为了使模型学习更多开源情报领域的知识和摘要生成能力，本研究搜集整理了大量高质量的英文开源情报，经过严格的审查和人工比对，构建出高质量开源情报领域摘要生成对话数据约 1 万余条，这些对话数据的输入为英文原文、输出为中文简讯，数据的来源包括军事新闻网站、军事官网、军事报告等。数据集的构造流程如图 2 所示。

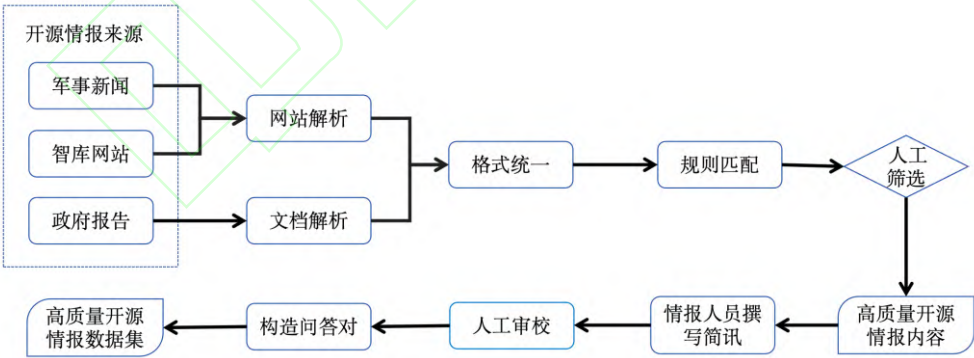


图 2 数据集的构造流程图

Fig.2 Data collection structure flow chart

具体地，本文通过军事新闻、智库网站和政府报告等途径获取开源情报，然后对这些资料进行解析，包括网站解析和文档解析，并将其转化为统一格式的数据。随后，利用预设的规则对这些数据进行匹配，目的是对数据进行过滤，去除杂质内容。

经过上述步骤后，还需要人工介入进行筛选，以确保所选开源情报的质量。然后由专门的情报人员根据筛选后的开源情报编写简讯，从而形成一个高质量的开源情报数据集。数据集统计如表 1 所示。基于上述数据采集和处理流程，本文构建的数据集的格式如表 2 所示。

表 1 数据集统计
Tab.1 Data collection statistics

	训练集	测试集
样本数	10000	500
平均问题单词个数	905	891
平均中文简讯文字个数	377	384

表 2 数据集格式
Tab.2 Data collection format

"Instruction": "为以下内容生成一篇中文简讯"
"input": "IDEX 2023: Steadicopter, Bird Aerosystems add maritime patrol radar to Black Eagle 50H UAS Steadicopter and Bird Aerosystems have integrated the latter's Micro Maritime Patrol Radar (MPR) on the Black Eagle 50H rotary-wing unmanned aircraft system (UAS), and displayed the new configuration at IDEX 2023, held in Abu Dhabi from 20 to 24 February. Steadicopter unveiled a hybrid version of its Black Eagle 50 UAS the Black Eagle 50H in February 2022, and the company has been working to improve the aircraft's maritime intelligence, surveillance, and reconnaissance (ISR) capabilities. The MPR enables the system to conduct long-range ISR operations. Further improvements introduced to the Black Eagle 50H include the ability to integrate an automatic identification system (AIS) and electro-optical/infrared (EO/IR) sensors from Controp. In addition, the operating range has been extended to 200 km."
output": "“黑鹰”50H 无人机在 2 月 20 日至 24 日举行的 2023 年阿布扎比防务展 (IDEX 2023) 上, 以色列 Steadicopter 公司展出了加装微型海上巡逻雷达的“黑鹰”50H 混合动力无人直升机。该雷达由以色列 BIRD 航空系统公司开发, 可使“黑鹰”无人机能进行远距离海上情报、监视和侦察 (ISR), 在恶劣环境中对半径 200 公里内的所有船只进行全天候精确测绘"

2.3 微调训练

利用上述构建的数据集, 对 Qwen1.5-14B 进行微调训练, 使用 LoRA 的方式进行微调。

LoRA 的核心思想是通过低秩分解来近似原本复杂的权重矩阵, 即将大矩阵分解为两个小矩阵的乘积, 这两个小矩阵通常具有远少于原矩阵的参数量。通过这种方式能够有效减少计算资源的消耗并加速训练过程。这也意味着 LoRA 能够对现有的模型进行轻量级的调整或微调, 快速适应特定任务, 同时保留模型原有的大部分知识结构。

要在下游任务微调一个预训练语言模型 (如 Qwen1.5-14B), 则需要更新预训练模型参数, 如公式 (1):

$$W_0 + \Delta W = W_0 + BA \quad (1)$$

式中, W_0 是预训练模型初始化的参数; ΔW 是需要更新的参数, 如果是全参数微调, 则其参数量为 14B, 而对于 LoRA 来说只需要微调 ΔW 。在 LoRA 的训练过程中, W_0 是固定不变的, 只有 A 和 B 是训练参数。

本文在微调过程中的超参数设置如表 3 所示。

表 3 超参数设置
Tab.3 Super triple setting

参数	参数值
微调类型	LoRA
预热步数	20
模板	Qwen
学习率	0.00005
批大小	4
训练轮次	3
Dropout 率	0.1
梯度累积步数	4
学习率调度器类型	cosine
序列最大长度	4096

2.4 评价指标

本文使用 ROUGE 和 BLEU 评测方法作为摘要方法有效性的评判标准。BLEU 通过比较候选摘要或翻译中 n-gram 与参考文本中 n-gram 的匹配情况来计算分数，旨在评价机器翻译输出与人工翻译参考文本之间的相似度。ROUGE 旨在评估系统生成的摘要与人工摘要之间内容的重叠程度。

使用这两种指标作为摘要方法有效性的评判标准，可以综合考量摘要的精确度、召回率以及结构上的相似性，从而获得对摘要质量更为全面的评价。

3 实验分析

为了评估模型的性能，本研究构建了开源情报摘要生成数据集，含有 3000 余条问答数据。实验中，分别对比了 Qwen1.5-14B 模型和微调后的 Qwen1.5-OSINT 模型在同一数据集上的表现，实验结果如表 4 所示。微调前后模型对于相同问题回答的答案对比如表 5 所示。以下是实验结果的详细总结和分析。

表 4 实验结果对比
Tab.4 Comparison of experimental results

Model	ROUGE-1 (%)	ROUGE-2 (%)	ROUGE-L (%)	BLEU (%)
Qwen1.5-OSINT	68.77	35.73	45.16	24.62
Qwen1.5-14B	30.35	15.52	20.22	11.14

表 5 微调前后模型生成的内容对比
Tab.5 Micro-tuning model generative content comparison

Model	The U.S. Air Force has selected BAE Systems, with support from FlexRadio, to provide software defined radios for its Airborne High Frequency Radio Modernization (AHFRM) program. The contract, which has a value of \$176 million, provides a secure alternative to satellite communication methods. ,The AHFRM solution maintains over-the-horizon communications while defeating jamming from potential threats in a drop-in compatible radio design that maximizes FlexRadio’s commercial off-the-shelf technology. Its scalability, modularity, and capacity provisions for future modernization needs and adjacent operational requirements. Advertisement “Our airborne radios are scalable and designed for open architecture applications, making them a solution that’s ready for the Air Force’s future needs,” said Dave Logan, vice president and general manager of C4ISR Solutions at BAE Systems. “The evolving capabilities of threats in the modern battlefield call for an extensible, modernized anti-jam solution, which our product provides.” BAE Systems designs, manufactures, and supplies mission-critical communication systems for the U.S. Department of Defense, allied governments, and large defense aircraft manufacturers. The company has secure, battle-proven communications solutions installed on a broad range of military airborne platforms. AHFRM development will take place at BAE Systems’ state-of-the-art facility in Fort Wayne, Ind.
Qwen1.5-OSINT	美空军选择 BAE 系统公司为“机载高频无线电现代化”（AHFRM）项目提供软件定义无线电设备。相关合同价值 1.76 亿美元，采用 FlexRadio 公司提供的商业现货技术，开发一种基于开放架构应用的可扩展的现代化的抗干扰解决方案，以提供一种安全的基于卫星的通信替代方案，同时保持超视距通信能力。该解决方案的可扩展性、模块化和大容量，将满足空军未来现代化的需求和临近作战需求。AHFRM 将在 BAE 系统公司位于印第安纳州福克斯波特的先进设施中进行开发
Qwen1.5-14B	BAE Systems 与 FlexRadio 合作，获得 1.76 亿美元的合同，为美国空军的 Airborne High Frequency Radio Modernization（AHFRM）项目提供软件定义无线电。该合同提供了一种安全的替代卫星通信方法的解决方案，同时保持了越过地平线的通信，并击败了来自潜在威胁的干扰。该解决方案具

	有可扩展性、模块性和容量，为未来的现代化需求和相邻的操作需求提供了支持。BAE Systems 设计、制造和供应美国国防部、盟国政府和大型国防飞机制造商的使命关键通信系统。该公司拥有安全、经过实战检验的通信解决方案，安装在广泛的军事空中平台上。AHFRM 的开发将在 BAE Systems 位于印第安纳州福尔斯特韦恩的最先进的设施中进行
--	---

从表 4 的实验结果对比中可以看出，在各个评价指标上，微调后的 Qwen1.5-OSINT 模型均显著优于未微调的 Qwen1.5-14B 模型。在 ROUGE-1、ROUGE-2、ROUGE-L 和 BLEU4 个评价指标上，Qwen1.5-OSINT 分别达到了 68.77%、35.73%、45.16%和 24.62%，明显高于 Qwen1.5-14B 的 30.35%、15.52%、20.22%和 11.14%。这些结果表明，经过微调后的模型在生成摘要的准确性、连贯性和完成度方面都有了明显提升。

通过表 5 生成的摘要内容对比，我们可以观察到微调前后模型在实际生成内容上的具体差异。未微调的 Qwen1.5-14B 模型生成的摘要虽然能够传达主要信息，但在语言流畅性和翻译质量上有所欠缺。相比之下，Qwen1.5-OSINT 生成的摘要不仅能够准确传达信息，且在语言表达上更符合人类书写习惯，使得摘要内容更加易读，并且更加突出了关键信息。

4 模型应用

基于 Qwen1.5-OSINT 模型，我们针对军事政府类网站设计了简讯生成插件，插件共三个模块：内容获取模块、内容生成模块及内容展示模块。

4.1 内容获取模块

该模块中主要实现网页链接、标题、内容、发布时间的获取，并利用规则进行过滤，得到过滤后的网页原文内容。

4.2 内容生成模块

该模块中实现调用 Qwen1.5-OSINT 模型，根据过滤后的网页原文内容自动生成中文的简讯；根据网页的链接，解析其域名并获取该域名对应的中文机构名称；根据原文标题生成中文的标题。

4.3 内容展示模块

将前两个模块所获得的内容显示在前台弹窗中，弹窗中包含以下几个字段：原文标题、原文链接、发布时间、中文标题、中文简讯。在弹窗中我们同时加入了三个按钮，点击保存数据按钮可以将原文标题、原文链接、原文内容、发布时间、中文标题、中文简讯、发布机构的中文名称保存到数据库中；点击重新生成按钮可以重新生成页面的简讯；点击导出文档按钮可以将插件中的字段根据提前设计的模板自动生成一个 word 文档。简讯生成插件的流程如图 3 所示，简讯生成插件的具体使用效果如图 4 所示。

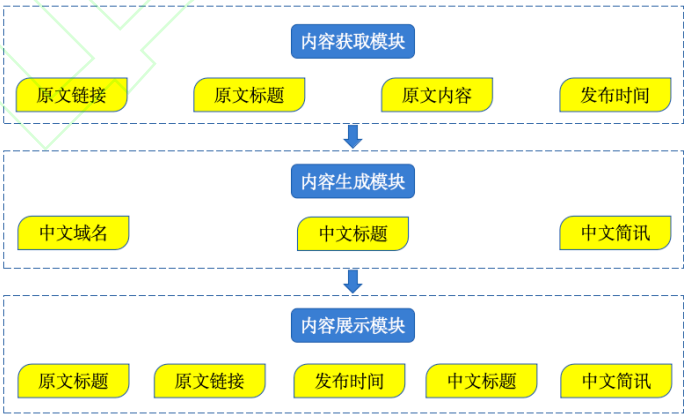


图 3 简讯生成插件流程图

Fig.3 Simple generation plug-in flow chart

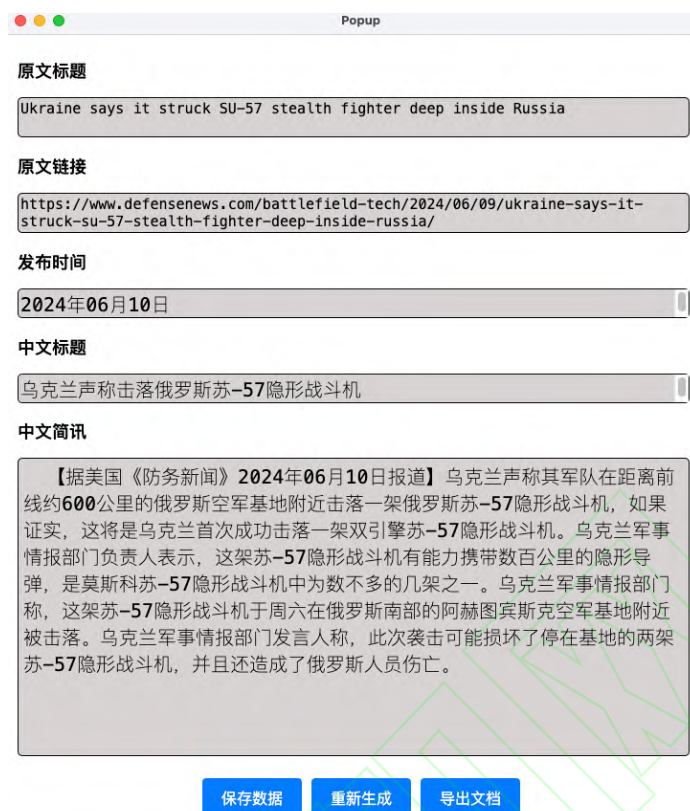


图 4 简讯生成插件效果图

Fig.4 Simple generated plug-in effect image

5 结论

本研究通过构建一个涵盖 1 万余条问答数据的开源情报摘要生成数据集,结合 LoRA 微调得到的 Qwen1.5-OSINT 模型,通过详细的实验对比分析,发现微调后的 Qwen1.5-OSINT 模型在 ROUGE 和 BLEU 等主要评价指标上均显著优于未微调的 Qwen1.5-14B 模型,同时对具体生成的摘要进行分析, Qwen1.5-OSINT 生成的摘要内容不仅更加准确,还更符合人类语言的表达习惯。

进一步,针对军事政府类网站设计了一个实用性插件,能够高效地生成和展示中文简讯。插件的开发和应用进一步验证了 Qwen1.5-OSINT 模型在情报摘要生成上的强大性能。简讯生成插件不仅提高了信息获取和处理的效率,也为用户提供了一种高效、便捷的信息展示和管理工具。

尽管本研究取得了显著的成果,但仍存在一定局限和挑战,如生成内容不准确、存在虚假信息的潜在风险等。此外,数据集的来源和质量也会影响模型的表现。未来研究将引入知识库来减少内容的不准确性,同时进一步扩展和优化数据集,涵盖更多领域和情景,以增强模型的泛化能力和应用广度。进一步提升模型的性能和实际应用价值。□

参考文献

- [1] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.
- [2] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.
- [3] BAI Jinze, BAI Shuai, CHU Yunfei, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.
- [4] DEAN J, CORRADO G S, MONGA R, et al. Large scale distributed deep networks[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2012:1223-1231.
- [5] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. ACM, 2020,63(11):139-144.

- [6] VASWANI A, SHAZEER N, PARMAR N, et al.. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Red Hook, NY, USA: Curran Associates Inc., , 2017: 6000–6010.
- [7] BROWN T B, MANN B, RYDER N, et al.. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Red Hook, NY, USA: Curran Associates Inc., , 2020,159:1877-1901.
- [8] DEVLIN J, CHANG MINGWEI, LEE K, et al.. bert: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics, 2019:4171-4186,.
- [9] GRANTER S R, BECK A H, PAPKE JR D J. AlphaGo, deep learning, and the future of the human microscopist[J]. Archives of Pathology & Laboratory Medicine, 2017, 141(5): 619-621.
- [10] WU Shijie, IRSOY O, LU S, et al. Bloomberggpt: a large language model for finance[J]. arXiv preprint arXiv:2303.17564, 2023.
- [11] LIU Xiaoyang, WANG Guoxuan, YANG Hongyang, et al. FinGPT: democratizing internet-scale data for financial large language models[J]. arXiv preprint arXiv:2307.10485, 2023.
- [12] YANG Aiyuan, XIAO Bin, WANG Bingning, et al. Baichuan 2: open large-scale language models[J]. arXiv preprint arXiv:2309.10305, 2023.
- [13] WEI Chen, WANG Qiushi, LONG Zefei, et al. Disc-finllm: a Chinese financial large language model based on multiple experts fine-tuning[J]. arXiv preprint arXiv:2310.15205, 2023.
- [14] LE SCAO T, FAN A, AKIKI C, et al. Bloom: a 176b-parameter open-access multilingual language model[J]. arXiv preprint arXiv:2211.05100, 2023.
- [15] ZHANG Xuanyu, YANG Qing. XuanYuan 2.0: a large Chinese financial chat model with hundreds of billions parameters[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23). New York, NY, USA: Association for Computing Machinery, 2023:4435-4439.
- [16] CUI Jiayi, LI Zongjian, YAN Yang, et al. Chatlaw: open-source legal large language model with integrated external knowledge bases[J]. arXiv preprint arXiv:2306.16092, 2023.
- [17] ZENG Aohan, LIU Xiao, DU Zhengxiao, et al. Glm-130b: an open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.
- [18] LexiLaw[EB/OL].[2024-12-12].<https://github.com/CSHaitao/LexiLaw>.
- [19] WANG Haochun, LIU Chi, XI Nuwa, et al. Huatuo: tuning llama model with Chinese medical knowledge[J]. arXiv preprint arXiv:2304.06975, 2023.
- [20] XIONG Honglin, WANG Sheng, ZHU Yitao, et al. Doctorglm: fine-tuning your Chinese doctor is not a herculean task[J]. arXiv preprint arXiv:2304.01097, 2023.
- [21] LI Yunxiang, LI Zihan, ZHANG Kai, et al. Chatdoctor: a medical chat model fine-tuned on llama model using medical domain knowledge[J]. arXiv preprint arXiv:2303.14070, 2023.

作者简介：汪伦，男，2000 年生，硕士生。研究方向：自然语言处理。**艾斯卡尔·艾木都拉**，男，1972 年生，博士后，教授。研究方向：自然语言处理。**张华平**（通信作者），男，1978 年生，博士，教授。研究方向：自然语言处理，情报智能分析。**耿国桐**，男，1975 年生，博士，研究员。研究方向：情报智能化。**吕雁飞**，男，1984 年生，博士，副研究员。研究方向：大数据，大模型。**雷帅**，男，1989 年生，硕士，副研究员。研究方向：科技情报。

录用日期：2025-01-20