哈尔滨工业大学计算机科学与技术学院

《算法设计与分析》

课程报告

|  |  |
| --- | --- |
| 学号 | 2022111189 |
| 姓名 | 李博文 |
| 班级 | 2203202 |
| 专业 | 信息安全 |
| 授课教师 | 张开旗 |
| 报告日期 | 2023年 11 月 10 日 |

论文题目

作者：Masoumi, S., Amirkhani, H., Sadeghian, N. *et al.* 标题：Natural language processing (NLP) to facilitate abstract review in medical research: the application of BioBERT to exploring the 20-year use of NLP in medical research. *Syst Rev* **13**, 107 (2024). 链接：<https://doi.org/10.1186/s13643-024-02470-y> 出处：VLDB

# 问题定义

本文主要解决了医学研究中系统性和范围性文献综述中的摘要审阅步骤耗时费力的问题。作者通过将自然语言处理 (NLP) 应用于特定文献综述问题，即过去 20 年中 NLP 在医学研究中的应用趋势，来展示自动摘要审阅模型的性能，使用了 BERT 模型和逻辑回归模型实现了自动摘要审阅。BERT 模型能够有效地提取文本的特征，并将其转换为可用于分类的向量表示。逻辑回归模型使用 BERT 模型的输出作为输入特征，并预测文本属于哪个类别。通过这种方式，可以自动地对摘要进行分类，从而提高文献综述的效率。

# 算法描述和分析

本文使用了基于 BERT 的语言模型 BioBERT 来进行摘要分类。BioBERT 在生物医学文本上进行预训练，并在一些生物医学数据集上优于其他预训练语言模型。作者对 BioBERT 进行了微调，使其能够根据三个变量（使用场景、文本来源和主要研究领域）对摘要进行分类。以及逻辑回归模型： 用于根据 BERT 模型的输出对文本进行分类。

BERT 模型是一种基于 Transformer 的预训练语言模型，它通过双向编码的方式对文本进行理解。BERT 模型在大型文本语料库上进行预训练，并学习到丰富的语言知识。这使得 BERT 模型能够对各种 NLP 任务进行微调，例如文本分类、命名实体识别、问答等。BERT 模型能够有效地提取文本的特征，并将其转换为可用于分类的向量表示。这使得逻辑回归模型能够更准确地预测文本属于哪个类别。此外，BERT 模型还可以进行微调，使其能够适应不同的 NLP 任务。实验结果表明，该方法能够有效地对摘要进行分类，并取得了良好的性能。

作者从 PubMed、Embase、PsycINFO 和 CINAHL 数据库中检索相关摘要，并手动审阅了 485 篇摘要，并根据三个变量对其进行分类，然后使用 BioBERT 模型对训练数据进行微调，使其能够根据三个变量对摘要进行分类，最后使用手动注释的摘要对模型进行评估，并计算微 f1 分数和准确率，通过这些步骤，体现出BioBERT 在摘要审阅方面表现出良好的性能，微 f1 分数分别为 77.35%（使用场景）、76.24%（文本来源）和 85.64%（主要研究领域）。

逻辑回归模型是一种经典的分类模型，它通过学习输入特征与目标类别之间的概率关系来进行分类。在本文中，逻辑回归模型使用 BERT 模型的输出作为输入特征，并预测文本属于哪个类别。

逻辑回归模型有逻辑回归模型结构简单，易于理解和实现、可以解释模型的决策过程，例如哪些特征对分类结果影响最大等特点，可以用于二分类和多分类任务，例如，将文本分类到不同的类别中，如情感分析、主题分类等，识别垃圾邮件，预测用户的信用风险等

假设需要根据“使用场景”对一篇摘要进行分类。使用训练好的模型对摘要进行编码，然后使用模型预测摘要属于哪个类别，将摘要输入到 BERT 模型中，然后 BERT 模型对摘要进行编码，将其转换为向量表示，最后使用逻辑回归模型对 BERT 模型的输出进行分类，预测摘要属于哪个类别。

假设摘要内容为：“本研究使用 NLP 技术对电子病历进行语义分析，以识别患者的临床特征。”BERT 模型会将该摘要编码为一个向量表示，例如：[0.1, 0.2, 0.3, ...]。然后，逻辑回归模型会根据该向量表示预测摘要属于“临床决策支持”类别，通过这种方式，可以自动地对摘要进行分类，从而提高文献综述的效率。

# 讨论和分析

论文工作总结：

本文展示了 BioBERT 在自动摘要审阅方面的潜力，并提供了关于 NLP 在医学研究中应用趋势的见解，总而言之，BioBERT 是一种很有前途的工具，可以帮助研究人员自动进行摘要审阅。

已有算法的不足：

训练 BioBERT 模型需要大量的手动标注数据，可能会耗费时间和资源，而且本文仅使用了英文摘要，无法应用于其他语言的文献综述，并且模型的性能可能会受到训练数据中潜在偏差的影响。

可能的改进：

可以使用主动学习技术来减少手动标注数据的需求，并且开发能够处理多种语言的 NLP 模型，还能提高模型的透明度和可解释性，以便研究人员更好地理解模型的决策过程。

不适用的情况：

如果研究主题非常狭窄，可能无法找到足够的训练数据来训练有效的模型，并且 如果文本高度专业化，可能需要开发专门的 NLP 模型才能对其进行处理。

**参考文献**

1. Masoumi, S., Amirkhani, H., Sadeghian, N. *et al.* Natural language processing (NLP) to facilitate abstract review in medical research: the application of BioBERT to exploring the 20-year use of NLP in medical research. *Syst Rev* 13, 107 (2024). <https://doi.org/10.1186/s13643-024-02470-y>
2. Ahammad, S.H., Kalangi, R.R., Nagendram, S. *et al.* Improved neural machine translation using Natural Language Processing (NLP). *Multimed Tools Appl* **83**, 39335–39348 (2024). <https://doi.org/10.1007/s11042-023-17207-7>
3. Liu B, Lane I (2016) Attention-based recurrent neural network models for joint intent detection and slot filling. arXiv preprint arXiv:1609.01454
4. Klimova B, Pikhart M, Benites AD, Lehr C, Sanchez-Stockhammer C (2023) Neural machine translation in foreign language teaching and learning: a systematic review. Educ Inf Technol 28(1):663–682