基于深度文本摘要的开源软件缺陷挖掘研究

汶东震(sewen@mail.dlut.edu.cn) 张帆 Mesfin Tadesse Michael 徐博 林鸿飞 杨亮 林原
大连理工大学计算机科学与技术学院

论文摘要

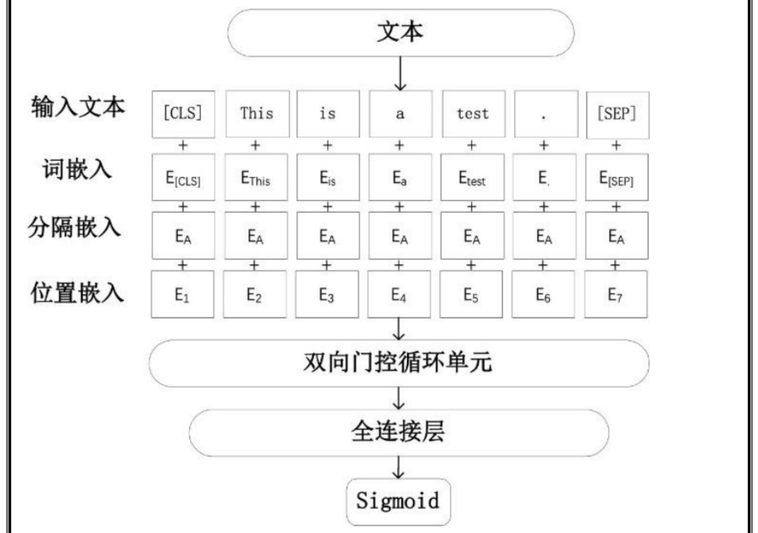
软件缺陷(**Software Defeat/Bug**)对于用户体验影响巨大，在一些重要场合中甚至会产生严重后果和经济损失。针对软件缺陷的快速定位与修复是软件工程领域一项重要研究内容。现代软件开发过程中严格的软件测试流 程能够显著降低软件错误同时提升软件质量，但软件开发成本也随之提升。开源软件运动的兴起革新了软件开发和测 试流程，贡献者通过邮件列表、社区讨论等方式对开源软件质量进行探讨，软件缺陷往往描述潜藏在用户交流文本中。

如何从海量开源软件缺陷报告文本中快速定位软件缺陷内容是该文的研究重点。该文首先研究了自然语言预训练技术 在开源软件缺陷数据挖掘中的效果。在分析开源软件文本特点基础上，提出一种基于预训练自然语言模型的深度文本 摘要模型，用以从海量文本中定位和抽取软件缺陷描述的重要内容。相关对比实验表明该文提出的模型在开源软件缺 陷报告文本挖掘中有着良好的效果，相关结论为开源软件缺陷报告挖掘任务提供了一定指引。

论文简介

本文提出了一种基于预训练自然语言模型的深度文本摘要模型，应用于开源软件缺陷报告挖掘中。相关实验表明，其针对性的提高了开源软件缺陷报告中缺陷识别挖掘的效果。

算法原理



实验仿真

表 3 不同实验方案对比结果

数据集	方案	Acc	Pre	Rec	F1	微调
SDS	单句	0.661	0.53	0.456	0.47	否
	单句	0.643	0.499	0.424	0.44	是
	前后句	0.593	0.413	0.341	0.361	否
	IR 方案	0.578	0.388	0.323	340	否
	主题	0.669	0.544	0.464	0.481	否
	单句+G	**0.696**	**0.591**	**0.516**	**0.527**	否
ADS	单句	0.679	0.477	0.465	0.44	否
	单句	0.715	0.539	0.546	0.509	是
	前后句	0.633	0.395	0.373	0.359	否
	IR 方案	0.635	0.397	0.37	0.358	否
	主题	0.686	0.49	0.47	0.45	否
	单句+G	**0.722**	**0.551**	**0.56**	**0.521**	否

表 4 不同分类后端结果对比

数据集	实验方案	Acc	Pre	Rec	F1
SDS	单句	0.661	0.53	0.456	0.47
	单句+G	**0.696**	**0.591**	**0.516**	**0.527**
	单句+L	0.687	0.574	0.499	0.511
ADS	单句	0.679	0.477	0.465	0.44
	单句+G	**0.722**	**0.551**	**0.56**	**0.521**
	单句+L	0.72	0.547	0.552	0.515

论文结论

软件缺陷报告数据具备更多领域特定内容，在处理时需要针对领域特定词语进行处理。同时，在结合较强的预训练自然语言模型表示之后，使用循环神经网络再次进行语义编码后结果会有进一步提升。

最后，本文仅仅初步针对软件缺陷挖掘任务结合自然语言预训练模型在两个规模较小的数据集上进行了探索。从实验也可以看到，在数据集上训练词向量效果略微有所降低，其中原因可能 与数据量较低有一定关系。文中两个数据集均来源于开源项目，后续研究中会在大量数据集上进行相关探索，对开源软件缺陷报告文本挖掘相关理论进行进一步完善。

西安电子科技大学 XIDIAN UNIVERSITY