南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：面向相似度量的源代码理解技术

工学硕士(软件工程领域) 专业 2018 级硕士生姓名：邹智鹏

指导教师（姓名、职称）：陈振宇 教授

摘 要

伴随着大数据时代的到来和大量数据知识的积累，现在的软件系统已经逐渐从信息化朝智能化方向发展，智能化软件工程便是如此。精准理解代码含义，能够用于代码分类、缺陷检测、克隆检测、代码检索等众多智能化软件工程任务。但现有的代码理解方法无法全面地捕捉代码文本层次的语义特征，并且处理语法结构特征的方法具有较高的复杂度和较差的鲁棒性，制约甚至忽略了语法特征的提取。此外，现有的抽象语法树处理方法，引入了大量的噪声，显著降低了代码理解的准确性。

源代码理解客观上即是能用文本刻画其实现的功能，并寻找一个映射，用一固定维度的低维稠密向量来刻画代码含义，并在文本语义空间中使用相似性度量方法进行评估的过程。源代码将预先被处理为语言无关的 AST 形态，根据相应的定义和算法构造路径对序列，从中提取语法、语义等特征。综合现有的问题，本文提出了混合编码器，即用于获取语义信息的子序列编码器，和用于获取语法信息的路径对编码器。其中，子序列编码处理源代码中的自然语言片段，丰富了语义特征的建模。本文还使用更简单的 RNN 网络构建路径对编码器，处理代码的语法结构，实现了更鲁棒和更准确的语法理解。特别地，本文还提出基于自注意力机制的路径融合方法 (**S**elf-attention based **P**ath **F**usion, SPF)，将自注意力机制应用到 RNN 的下游，实现更准确的结构特征融合，大幅降低了原始 AST 处理的噪声，增强了代码理解的准确性，降低了问题规模，提升了编码效率。

本文分别在方法名生成和代码同文本语义相似度量两个任务上，对代码理解方法进行验证。实验结果显示，本文的方法在两个任务上的效果均明显超越了基准实验的表现。根据实验结果分析，子序列处理方法解决了 AST 叶子结点和方法名中存在的长尾问题。凭借 RNN 的稳定性和时序性，AST 节点间的依赖关系建模更加准确，提升了语法特征抽取效果。SPF 方法在 RNN 网络上的应用，抑制了噪声数据干扰，动态地得到了高质量的特征组合，最终在两个任务上分别取得了相对基准方法 20% 和 60% 的指标提升。此外，综合损失函数改进和代码 API 信息，使 SPF 方法达到了更优的代码理解目标。

关键词：代码理解，智能化软件工程，语义匹配，自注意力，代码推荐