京都工芸繊維大学

氏　　名	ほん ちょんとん HONG JUNTONG
学位(専攻分野)	博　士（工学）
学位記番号	博　1 1 4 2　号
学位授与の日付	令和 6 年 9 月 25 日
学位授与の要件	学位規則第 4 条第 1 項該当
研究科・専攻	工芸科学研究科　設計工学専攻
学位論文題目	Bridging Source Code and Natural Language for Code Search （コード検索のためのソースコードと自然言語の接続手法の提案）
審査委員	(主査)教授　水野 修 教授　澁谷 雄 教授　馬　強 准教授　杜 偉薇

論文内容の要旨

コード検索は，自然言語で書かれたクエリに基づいて，大規模なコードリポジトリから関連性の高いコードを探し出すタスクであり，ソフトウェア工学において重要な役割を果たしている．このタスクは，ソフトウェア開発において多角的な改善を可能にする．例えば，ターゲット機能を実装する際に既存のコードを少し修正して利用したり，同様の機能の実装にインスピレーションを得たりすることができる．また，高品質なリポジトリから適切に書かれたコードを見つけることで，大規模プロジェクトの機能モジュールを再利用する際の迅速な特定が可能となる．そのため，コード検索の性能向上が近年注目されている．

コード検索モデルは，コードとクエリの間に存在するギャップを埋めるという大きな課題に直面している．このギャップは，プログラミング言語で書かれたコードと自然言語で書かれたクエリが本質的に異なる性質を持つためである．そのため，コードとクエリの間の関係を効果的に把握し，確立することが重要である．従来の情報探索(IR)ベースの方法は，主に情報探索技術に依存し，外部知識データベースを通じて特徴を拡張し，テキスト類似性技術を用いてアライメント(語の並び)を捉えようとするが，クエリの理解や単語特徴の適切な表現に課題がある．これらの課題に対処するために，深層学習(DL)技術を用いたコード検索方法が注目されている．DL ベースの方法は，コードやクエリのセマンティクスを学習し，未見の特徴をより適切に扱うことで，IR ベースの方法の限界を克服する可能性がある．しかし，DL ベースの方法にも課題が存在し，特に異なるエンコーダによって異なるコードトークンの表現を生成する問題や，新しいトークンを作成する際に語彙の問題が生じることがある．さらに，コードをプレーンテキストとして扱うと語彙が過剰となる問題が発生し，階層的な構造情報を学習することが難しくなる．

このような課題に対処するため，本論文では以下の点に焦点を当てる．1）コードとクエリのアライメント学習の探求，2）特徴エンジニアリングへの過度の依存とコード埋め込みの表現強化の探求，3）語彙なしでコード埋め込みを学習する新しいアプローチの探求．これらの課題に対応するため，統合アライメントモデルとグローバルアライメントモデルを提案し，またコードを画像として表現するモデルも提案する．実験結果は，これらのアプローチがコード理解を強化し，新

しい洞察を提供することを示している．

論文審査の結果の要旨

本論文は，ソフトウェア開発において自然言語で書かれたクエリに基づいて大規模なコードリポジトリから関連性の高いコードを探し出すタスク「コード検索」の精度を向上させる手法についての研究を行ったものである．従来の手法は主に情報探索(IR)技術に依存していたが，近年は深層学習(DL)技術を応用した手法が見られるようになった. DL 技術ではコードやクエリのセマンティクスや意味を学習することができ，より精緻な検索を可能にすると期待されている．しかし，一方で，DL ベース手法にはエンコーダによる内部表現の違いや，多量の語彙を扱うことができなくなる問題が知られており，解決が必要であった．

多くの DL ベース手法は，コード特徴の拡張に依存してコードとクエリの間のアライメント(語の並び)学習を促進しようとするのに対し，提案する統合アライメントモデルは，限られた特徴エンジニアリングの下でコード検索性能を向上させることを目指した．

そのために，コードとクエリの間のアライメントを捕捉するための新しいアライメント学習アプローチを提案した．また，クロスモーダルアライメントを提案し，コード検索の性能を向上させる効果を示している．クロスモーダルモデルの有効性と改善を測定するための評価実験をいくつか実施している．評価実験の結果，統合アライメントモデルが限られた特徴の下で従来の研究よりも良好な性能を発揮することを示した．

また，コードとクエリの埋め込みを多様で特徴的な表現に変換し，密なグラフ畳み込みネットワークを提案することでグローバルアライメントモデルを提案した．評価実験から，提案モデルがコード検索の性能を大幅に向上させ，以前の研究を大きく上回ることを確認した．

また，コードやクエリをプレーンテキストで扱うことに起因する過剰な語彙の問題に対応するためにコードをテキストではなく画像で扱う手法の検討も行っている．コードを画像として扱うことで語彙を考慮せずに学習データとして扱うことができることを示した．

以上の通り，本論文において提案された手法は，コード検索技術の精度向上に寄与しており，ソフトウェア開発におけるコスト削減や信頼性向上につながることが期待できる．

本論文は，以下に示す 3 編の査読を経た論文を基礎としており，これらは全て申請者が筆頭著者である．

[1] Juntong Hong, Eunjong Choi, Kinari Nishiura, and Osamu Mizuno, “Global Alignment Learning for Code Search,” In Proc. of 22nd International Conference on Software Engineering Research, Management and Applications (SERA2024), (in press), 6 pages, May 2024. (Honolulu, HI, USA)

[2] Juntong Hong, Eunjong Choi, and Osamu Mizuno, "A Combined Alignment Model for Code Search," IEICE Transactions on Information and Systems, E107-D(3), pp. 257-267, March 2024.

[3] Juntong Hong, Osamu Mizuno, and Masanari Kondo, "An Empirical Study of Source Code Detection Using Image Classification," In Proc. of 10th International Workshop on Empirical Software Engineering in Practice (IWESEP 2019), 6 pages, December 2019. (Tokyo, Japan)