CodeBERT を用いたソースコード作成者の習熟度の推定

1250358　林　晃太郎　【 ソフトウェア検証・解析学研究室 】

1　はじめに

プログラミングスキルの客観的評価は教育現場での学習成果の測定や企業の採用活動において重要である．

現在の習熟度評価手法としてコーディングテストが広く行われている．しかし，コーディングテストは解答の正否に依存しており，テストを解くために必要な知識を保持しているかによって正答率が大きく変化する．そのため短時間のテストではプログラミングスキルの全体像を把握しきれないという課題が存在する．この問題の解決策として，ソースコードの特徴を解析し，コード作成者のスキルを推定する手法が有効と考えられる．

既存研究として，競技プログラミングのレーティングとソースコードの品質の関係を調査する研究が行われている．槇原らはソースコードの特徴量を用いて，機械学習によりソースコードの品質を判定する手法を提案した [2].

一方近年，ソフトウェア開発の効率化と品質向上のために，事前学習モデルを活用したプログラム解析が注目されている．本研究では，事前学習モデルである CodeBERT を用いてソースコード作成者の習熟度を推定する手法を提案する．

2　競技プログラミング

競技プログラミングは，与えられた問題を正確かつ効率的に解く能力を競うプログラミング競技である．いくつかの競技プログラミング大会では，ユーザーの習熟度を示す指標としてレーティングが採用されており，コンテストの成績に応じて変動する．本研究では，レーティングが採用されており，初級者から上級者まで多くのプログラマーが参加している AtCoder を利用する．

3　CodeBERT

CodeBERT[1] は自然言語処理における深層学習モデルとして広く使われる Transformer をベースにした事前学習モデルであり，自然言語と 6 種のプログラミング言語のペアで事前学習されたマルチプログラミング言語モデルである．CodeBERT はコード検索，コード補完，コードクローン検出，コード要約，バグ検出などに応用されている．

4　方法

本研究では，競技プログラミングサイトの提出コードと提出者の当時のレーティングを対象にし，CodeBERT をベースとしたモデルを構築する．AtCoder の競技プログラミングデータを収集し，参加者のレーティングを習熟度の指標として利用する．

本研究の予備実験では，AtCoder Beginner Contest 385 の大会中に提出されたソースコードのみを対象として学習および評価を実施する．全体で 53,239 個のソースコードを収集し，その中からランダムに 47,916 個を学習用データセット，残りの 5,323 個を評価用データセットとして抽出する．さらに，各コードの提出者のレーティングを基に，400 未満と 400 以上の 2 群に分類し，二値分類モデルを構築・評価する．

5　予備実験の結果

エポック（学習回数）ごとの学習時の損失関数の推移を図 1 に示す．この結果では，学習が進むにつれて損失が減少しており，特に最初の数エポックで大きく低下している．エポック 7 付近からは傾きが緩やかになっており，学習の収束が近づいていることが分かる．

エポックごとの検証データに対する正答率の推移を図 2 に示す．この結果では，学習初期は 80.6%の精度であったが，正答率はエポックが進むにつれて上昇し，エポック 9 で 89.7%，エポック 10 で 89.8%まで向上している．

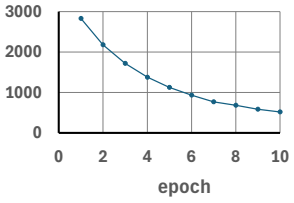

図 1　訓練損失

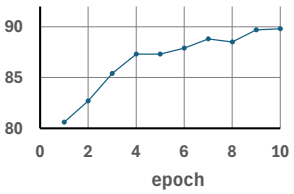

図 2　正答率

6　まとめ

予備実験の結果から，CodeBERT で競技プログラミングのデータを学習することで，ソースコードの特徴から作成者の習熟度を推定できる可能性が示された．

本実験では，対象コンテストを増やし，より多くの問題について学習を行う．また，習熟度の推定精度を向上させるため，習熟度レベルの分類を細分化する．もしくは連続値として近似予測を行う手法を検討する．これにより，より実用的なソースコード作成者の習熟度の推定を目指す．

参考文献

[1] Z. Feng, et al. “CodeBERT: A Pre-Trained Model for Programming and Natural Languages”. arXiv preprint arXiv:2002.08155，2020.

[2] 槇原，松下，井上．“ソースコード特徴量を用いた機械学習によるソースコード品質の評価手法”．信学研報 SS. Vol. 119, No. 113, pp. 105–110，2019.