ChatGPT による Java プログラムのバグ修正の分析

1240361　藤川開斗　【 ソフトウェア検証・解析学研究室 】

1　はじめに

ChatGPT が登場してからまだ 13ヶ月ほどしかたっていないにも関わらず仕事場や日常生活にすでに浸透しており, 業務や勉強等に関することについて質問をすることに用いられている．ソフトウェア作成の場合は, プログラムのバグ修正の目的でも利用可能と考えられる．Sobania ら [1] は，バグ修正に対する ChatGPT の有用性を実験的に示している．しかし， [1] では Python プログラムに関する実験が行われており，他の言語についての傾向はわかっていない. 実際，同様のバグのあるプログラムでも，Python プログラムに対しては簡単な少数の問い合わせでバグの特定と修正が行えるのに対し，Java プログラムに対しては Python と同じ問い合わせでは目的の結果が得られないことがある．

そこで， [1] のバグ修正に関する実験と同様の実験を Java プログラムに対して行い，傾向のちがいを分析した．

2　実験方法

バグのあるプログラム集として，[1] と同様に QuixBugs ベンチマーク[1]を使用する．QuixBugs は，40 個のバグのあるプログラムをそれぞれ Python と Java で記述したものからなる．実験手順も [1] とほぼ同様で，40 個の各プログラムについて以下を行う．

1. ChatGPT に「このプログラムにバグはありますか？どのように修正すればいいですか？ (改行後にプログラムコードを記述する)」と質問する．

2. 提供されたコードにメインメソッドが含まれているか調べる. 含まれていない場合は,「メインメソッドは?」と質問して提供されたメインメソッドを追加してコンパイルと実行を行う．

3. コンパイルと実行を行ったときにコンパイルエラーが発生するかまたは想定される出力結果が得られたか調べる．

- 「コンパイルエラーが発生」または「想定される出力結果を得ることができなかった」場合は, 表示されたコンパイルエラー全文あるいは出力された結果をコピー&ペーストして再度質問する．
- 6 回質問を繰り返すまでに修正ができなかった場合は, ChatGPT に Python に翻訳してもらい, 提供された Python のコードを使用して実験を行う．提供された Python のコードを実行する場合は, できたかどうかを確認したら調査を終了する．

[1]https://github.com/jkoppel/QuixBugs

Java と Python それぞれに対して同じ手順を行う（「Python に翻訳」以降は Java に対してのみ行う）．質問を繰り返す回数を， [1] では 4 回としていたが本研究では 6 回に増やした．実際，5 回目または 6 回目に成功する例もあった．

3　実験結果

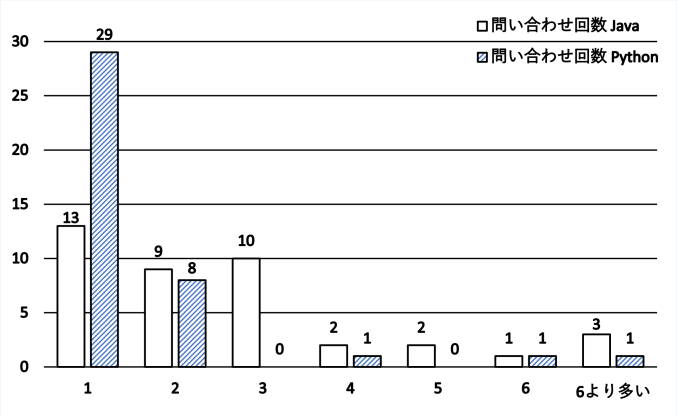

図 1　Java プログラムと Python プログラムの問い合わせ回数

目的の結果が得られるまでの Java, Python それぞれの必要問い合わせ回数を図 1 に示す．平均問い合わせ回数は, Java の場合は 2.65 回，Python の場合は 1.63 回だった．全体的に Python の方が問い合わせ回数が少なく済んでいる．

また，入力されたプログラムが何を行うものか ChatGPT が正しく推定できているかどうか調べるために，返答中に「このプログラムは…を行うプログラムです」のような文章が含まれるかどうか調べた．そのような文章が返答に含まれ，かつ正しい内容だった割合は，Java の場合は 35.0%，Python の場合は 82.5%であり，Python の方が正しく推定できている割合が高かった．

4　まとめ

本研究では， [1] で述べられているバグ修正作業における ChatGPT の有用性について，プログラミング言語による傾向のちがいを調べるため，Java プログラムに対する傾向を調査した．

参考文献

[1] D. Sobania, C. Hanna, M. Briesch, J. Petke, “An Analysis of the Automatic Bug Fixing Performance of ChatGPT”. arXiv preprint arXiv:2301.08653, 20 Jan 2023.