Impact of Mixed-Domain Code-Text Fine-Tuning on JaCoText Generalization in Cross-Language Benchmarks

Assignee Research

June 11, 2026

Abstract

Pretrained transformer-based models have shown high performance in natural language generation task. However, a new wave of interest has surged: automatic programming language generation. This task consists of translating natural language instructions to a programming code. Despite the fact that well-known pretrained models on language generation have achieved good performance in learning programming languages, effort is still needed in automatic code generation. In this paper, we introduce JaCoText, a model based on Transformers neural network. It aims to generate java source code from natura

1 Introduction

This paper examines: JaCoText: A Pretrained Model for Java Code-Text Generation. Research question: What is the impact of incorporating mixed-domain (Java + other programming languages) code-text pairs during fine-tuning on JaCoText's generalization capability, as evaluated by pass@1 on cross-language benchmarks such as HumanEval-X?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

13 papers retrieved. 13 claims extracted; 10 independently verified. Quality review score: 7.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
JaCoText is a model based on Transformers neural network designed to generate Java source code from natural language tex	✓	0.29
JaCoText was initialized from powerful pretrained models, explored additional pretraining on a Java dataset, and scaled	✓	0.21
Experiments conducted on the CONCODE dataset show that JaCoText achieves new state-of-the-art results.	✓	0.26
CodeGPT is trained from scratch on the CodeSearchNet dataset.	×	0.15
CodeGPT-adapted is initialized from GPT-2 pretrained weights.	✓	0.18
PLBART uses the same architecture as BART-base and employs three noising strategies: token masking, token deletion, and t	✓	0.22
CoTexT uses the same architecture as T5base and is trained on unimodal and bimodal data using the CodeSearchNet Corpus a	✓	0.24
On the CONCODE dataset, the T5-base model achieved a BLEU score of 32.74, an EM score of 18.65, and a CodeBLEU score of	×	0.13
On the CONCODE dataset, the CoTexT-1CC model achieved a BLEU score of 37.40, an EM score of 20.10, and a CodeBLEU score	✓	0.17
On the CONCODE dataset, the JaCoText-L-2CC-PL model achieved a BLEU score of 39.87, an EM score of 22.45, and a CodeBLEU	✓	0.15
JaCoText-L-2CC-PL achieved the highest BLEU, EM, and CodeBLEU scores among all models listed in Table II.	×	0.10
Reference [17] used a BiLSTM encoder and an RNN decoder to generate syntactically valid parse trees.	✓	0.23
Reference [26] used a Bi-LSTMs encoder to compute contextual representations of natural language and an LSTM-based RNN d	✓	0.22

References

- http://arxiv.org/abs/2504.16584v1
- http://arxiv.org/abs/2303.12869v1
- http://arxiv.org/abs/2510.18904v1