Impact of Pretraining Data Volume on Zero-Shot Cross-Lingual Code Generation Performance

Assignee Research

July 6, 2026

Abstract

Zero-shot cross-lingual transfer is a central task in multilingual NLP, allowing models trained in languages with more sufficient training resources to generalize to other low-resource languages. Earlier efforts on this task use parallel corpora, bilingual dictionaries, or other annotated alignment data to improve cross-lingual transferability, which are typically expensive to obtain. In this paper, we propose a simple yet effective method, SALT, to improve the zero-shot cross-lingual transfer of the multilingual pretrained language models without the help of such external data. By incorporati

1 Introduction

This paper examines: Self-Augmentation Improves Zero-Shot Cross-Lingual Transfer. Research question: To what extent does increasing pretraining data volume for low-resource languages improve zero-shot cross-lingual transfer performance on code generation tasks compared to architecture scaling?.

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 7.5/10.

3 Results

13 papers retrieved. 11 claims extracted; 8 independently verified. Quality review score: 7.5/10.

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
SALT achieves 1.4% improvement on 15 languages of XNLI and 4.1% on 6 languages of PAWS-X in terms of average accuracy co	✓	0.19
SALT adopts a modified masked language modeling (MLM) method to distill cross-lingual token translation pairs.	✓	0.25
SALT makes two modifications to the original MLM learning objective: predicting tokens only in a specific target languag	✓	0.24
Generated token-level cross-lingual substitutions with high enough predicted probabilities are used for code-switching.	✓	0.25
SALT predicts synonyms in the source language with a different threshold for synonym prediction.	×	0.12
SALT proposes an online self-augmentation technique based on cross-lingual embedding mixup.	✓	0.15
The mixed-up token embedding is generated as hi = r $\cdot$ hsi + (1 - r) $\cdot$ hti, where r = {rj}, rj $\in$ [0, 1] is a random vecto	✓	0.27
The interpolation coefficient vector r is dynamically sampled in each step of training for each instance.	✓	0.16
SALT achieves an average accuracy of 69.8 on XNLI with 15 different random seeds.	×	0.10
SALT achieves an average accuracy of 66.8 on XNLI without English.	×	0.07
SALT achieves significant improvements in comparison with mBERT baseline by t-test (p $\leq$ 0.05).	✓	0.18

References

- http://arxiv.org/abs/2508.09516v1
- http://arxiv.org/abs/2309.10891v1
- http://arxiv.org/abs/2310.09917v3