Impact of Cross-Lingual Query Generation on Zero-Shot Retrieval Accuracy in the XTREME Benchmark

Assignee Research

June 12, 2026

Abstract

Effective cross-lingual dense retrieval methods that rely on multilingual pre-trained language models (PLMs) need to be trained to encompass both the relevance matching task and the cross-language alignment task. However, cross-lingual data for training is often scarcely available. In this paper, rather than using more cross-lingual data for training, we propose to use cross-lingual query generation to augment passage representations with queries in languages other than the original passage language. These augmented representations are used at inference time so that the representation can enco

1 Introduction

This paper examines: Augmenting Passage Representations with Query Generation for Enhanced Cross-Lingual Dense Retrieval. Research question: How does augmenting passage representations with cross-lingual query generation impact zero-shot retrieval accuracy on the XTREME benchmark compared to standard multilingual dense retrieval baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.0/10.

3 Results

12 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 7.0/10.

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
The xDR model initialized with mBERT outperforms the xDR model initialized with XLM-R, achieving an average R@2kt score	✓	0.31
The xQG passage embedding augmentation approach improves the XLM-R xDR, achieving an average score of 29.8, which is a s	✓	0.41
The mBERT's effectiveness improves with xQG, achieving an average score of 46.2, which is also a statistically significa	✓	0.41
The zero-shot mBERT model achieves an average R@2kt of 33.0; this also improves when combined with xQG, achieving an ave	✓	0.44
Using more generated queries is beneficial for both R@2tk and R@5tk, with improvements becoming statistically significan	✓	0.30
mBERT performs better than XLM-R for both R@2kt and R@5kt.	✓	0.23
The use of xQG embedding augmentation statistically significantly improves the effectiveness of both XLM-R and mBERT bac	✓	0.24
The xQG improves almost all models across all languages with the exceptions of mBERT's R@2kt for Japanese (Ja) and mBERT	✓	0.29

References

- http://arxiv.org/abs/2305.03950v1
- http://arxiv.org/abs/2511.19325v1
- http://arxiv.org/abs/2204.07496v4