Scalability of Cross-Lingual Query vs. Passage Generation for MLQA Retrieval Accuracy

Assignee Research

June 16, 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 the scalability of cross-lingual query generation compare to cross-lingual passage generation in terms of retrieval accuracy on MLQA when using varying numbers of target languages?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.0/10.

3 Results

9 papers retrieved. 24 claims extracted; 20 independently verified. Quality review score: 8.0/10.

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 training set contains annotated relevant passage-query pairs and the dev set contains 2k passage-answer pairs.	✓	0.21
Queries in both the train and dev sets are in seven typologically diverse languages (Ar, Bn, Fi, Ja, Ko, Ru, and Te), wh	✓	0.35
There are about 18M passages in the corpus.	×	0.10
Table 1 presents the effectiveness of xDR models initialized with the XLM-R and mBERT backbone PLMs and trained on the X	✓	0.31
Zero-shot denotes the models trained only on the English subset of the NQ dataset.	✓	0.22
The xDR initialized with mBERT outperforms the xDR initialized with XLM-R, achieving an average R@2kt score of 44.1, whi	✓	0.35
The xQG passage embedding augmentation approach improves the XLM-R xDR, achieving an average score of 29.8, which is a s	✓	0.41
mBERT's effectiveness improves with xQG, achieving an average score of 46.2, which is also a statistically significant i	✓	0.42
The zero-shot mBERT model achieves an average R@2kt of 33.0; this also improves when combined with xQG, achieving an ave	✓	0.45
Similar trends are found for R@5kt.	×	0.14
Overall, we find that our xQG can significantly improve all investigated xDR models.	✓	0.21
In terms of per language effectiveness, xQG improves almost all models across all languages with the exceptions of mBERT	✓	0.32
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 backbones.	✓	0.26
Figure 2 reports the impact of using different amounts of generated queries to augment passage embeddings when using mBE	✓	0.32
The results suggest that using more generated queries is beneficial for both R@2tk and R@5tk.	✓	0.26
The improvements become statistically significant when 4 or more generated queries are used.	✓	0.22
Table [Table (p3)] shows the performance of XLM-R and mBERT models with and without xQG for R@2kt.	×	0.07
Table [Table (p3)] shows the performance of XLM-R and mBERT models with and without xQG for R@5kt.	×	0.06
The development of cross-lingual pre-trained	✓	0.24

References

- http://arxiv.org/abs/2511.19325v1
- http://arxiv.org/abs/2305.03950v1
- http://arxiv.org/abs/2508.09516v1