Monolingual Pre-training Data Augmentation for Robust Cross-Lingual Retrieval in Non-Indo-European Low-Resource Languages

Assignee Research

June 18, 2026

Abstract

Benefiting from transformer-based pre-trained language models, neural ranking models have made significant progress. More recently, the advent of multilingual pre-trained language models provides great support for designing neural cross-lingual retrieval models. However, due to unbalanced pre-training data in different languages, multilingual language models have already shown a performance gap between high and low-resource languages in many downstream tasks. And cross-lingual retrieval models built on such pre-trained models can inherit language bias, leading to suboptimal result for low-reso

1 Introduction

This paper examines: Improving Cross-lingual Information Retrieval on Low-Resource Languages via Optimal Transport Distillation. Research question: How does the incorporation of monolingual pre-training data augmentation affect the robustness of cross-lingual retrieval models on non-Indo-European low-resource languages, as measured by MRR and NDCG on BEIR benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 4 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

4 papers retrieved. 15 claims extracted; 13 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
CAL significantly outperforms strong baselines on low-resource languages, including neural machine translation.	✓	0.21
Sasaki et al. proposed a large cross-lingual retrieval collection, WikiCLIR, based on linked foreign language articles f	✓	0.27
The relevant judgments in WikiCLIR are synthetically generated based on mutual links across pages.	✓	0.17
Bonifacio et al. built a multilingual passage ranking dataset, mMARCO, by translating queries and passages in MS MARCO i	✓	0.33
The relevant judgments in mMARCO are more credible than WikiCLIR because MS MARCO is generated from query logs.	✓	0.18
OPTICAL is a novel Optimal Transport-based knowledge distillation framework for low-resource CLIR tasks.	✓	0.20
OPTICAL formulates the cross-lingual token alignment task as an optimal transport problem where the cost matrix is the c	✓	0.23
In OPTICAL, the optimal transportation plan serves as a soft token alignment.	✓	0.21
The loss in OPTICAL is defined as the Frobenius inner product of the transportation plan and the cost matrix.	✓	0.23
OPTICAL only needs bitext data for distillation training.	✓	0.22
Experiments were performed on seven language pairs for CLIR training and evaluation.	✓	0.18
The experimental dataset included four low-resource languages from diverse linguistic families.	×	0.10
The experimental dataset included three medium or high-resource languages as a comparison.	×	0.10
In terms of mean average precision (MAP), the proposed method significantly outperforms several strong baseline methods	✓	0.27
The proposed method achieved a 13.7% improvement over a method based on neural machine translation in terms of MAP.	✓	0.18

References

- http://arxiv.org/abs/2408.10536v1
- http://arxiv.org/abs/2301.12566v1
- http://arxiv.org/abs/2212.09651v4