Robustness of Cross-Lingual Retrieval Models via Optimal Transport Distillation Under Domain Shifts in 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: To what extent does the robustness of cross-lingual retrieval models trained via optimal transport distillation degrade under domain shifts in low-resource languages, and how does this compare to models trained with conventional fine-tuning methods, as evaluated by F1-score and precision-recall curves on domain-shifted datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.7/10.

3 Results

10 papers retrieved. 11 claims extracted; 10 independently verified. Quality review score: 7.7/10.

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
WikiCLIR is a large cross-lingual retrieval collection based on linked foreign language articles from Wikipedia pages.	✓	0.22
Relevant judgments in WikiCLIR are synthetically generated based on mutual links across pages.	✓	0.17
mMARCO is a multilingual passage ranking dataset built by translating queries and passages in MS MARCO into the target l	✓	0.26
Relevant judgments in mMARCO are more credible than those in WikiCLIR because MS MARCO is generated from query logs.	✓	0.18
OPTICAL is a knowledge distillation framework that formulates the cross-lingual token alignment task as an optimal trans	✓	0.19
In OPTICAL, the cost matrix for the optimal transport problem is defined as the cross-lingual token cosine distance.	✓	0.19
In OPTICAL, the loss is defined as the Frobenius inner product of the transportation plan and the cost matrix.	✓	0.25
OPTICAL only requires bitext data for distillation training.	×	0.14
Experiments were performed on seven language pairs, including four low-resource languages and three medium or high-resou	✓	0.18
The proposed method achieved a 13.7% improvement in Mean Average Precision (MAP) over a method based on neural machine t	✓	0.21

References

- `http://arxiv.org/abs/2506.15415v1`
- `http://arxiv.org/abs/2301.12566v1`
- `http://arxiv.org/abs/2408.11942v1`