Manifold-Aware Embeddings for Cross-Lingual Dense Retrieval on XQuAD

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Do manifold-aware embeddings derived from Wikipedia-based semantic relatedness metrics improve cross-lingual dense retrieval performance on XQuAD compared to standard cosine similarity, as measured. Dense Passage Retrieval (DPR) typically relies on Euclidean or cosine distance to measure query-passage relevance in embedding space, which is effective when embeddings lie on a linear manifold. However, our experiments across DPR benchmarks suggest that embeddings often lie on. 18 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MA-DPR: Manifold-aware Distance Metrics for Dense Passage Retrieval. Research question: Do manifold-aware embeddings derived from Wikipedia-based semantic relatedness metrics improve cross-lingual dense retrieval performance on XQuAD compared to standard cosine similarity, as measured by MRR@10 and exact match accuracy?.

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 4.2/10.

3 Results

13 papers retrieved. 18 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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
All codes and results are available online at github.com/QianfengWen/Manifold_Distance_Retrieval.git	×	0.05
System specifications include CPU—Intel(R) Core(TM) i7-14700HX and GPU—NVIDIA GeForce RTX 4070 Laptop GPU	×	0.02
Average CPU utilization during measurement was $\sim$5%	×	0.00
DPR with dEuclidean + NCA is one of the baselines used in the evaluation	×	0.05
MS MARCO is used as a benchmark dataset	×	0.03
NFCorpus is used as a benchmark dataset	×	0.04
SciDocs is used as a benchmark dataset	×	0.04
ANTIQUE is used as a benchmark dataset	×	0.04
The embedding model used is 'msmarco-distilbert-base-tas-b (tas-b)'	×	0.04
The embedding model used is 'SciNCL'	×	0.04
MS MARCO is the in-distribution dataset for the tas-b model	×	0.02
SciDocs is the in-distribution dataset for the SciNCL model	×	0.03
All embeddings are 2-normalized	×	0.04
Retrieval results are evaluated using Recall, MAP, and nDCG for the top 20 ranked assignments	×	0.06
The evaluation includes NFCorpus, SciDocs, and ANTIQUE datasets	×	0.02
The manifold hypothesis is validated by examining the relationship between dManifold and dEuclidean across relevant and	×	0.08
In a perfectly linear embedding space, manifold-aware distance induced by dKNN Euclidean + cDC should align with standar	✓	0.19
Non-linear structure causes divergence between manifold-aware distance and Euclidean distance	✓	0.22

References

- http://arxiv.org/abs/2305.03950v1
- http://arxiv.org/abs/2502.14620v1

- http://arxiv.org/abs/2509.13562v1