Multilingual Corpus Diversity and Zero-Shot Transfer in Low-Resource Languages

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the impact of multilingual pre-training corpus diversity on zero-shot transfer performance for low-resource languages in the XTREME benchmark. 7 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Finding Universal Grammatical Relations in Multilingual BERT. Research question: What is the impact of multilingual pre-training corpus diversity on zero-shot transfer performance for low-resource languages in the XTREME benchmark?.

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 4.5/10.

3 Results

4 papers retrieved. 7 claims extracted; 0 independently verified. Quality review score: 4.5/10.

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
Multilingual BERT is pretrained on corpora in 104 languages.	×	0.08
The syntactic probe recovers syntactic trees across all the languages investigated, achieving on average an improvement	×	0.05
The probe achieves significantly higher UUAS (on average, 9.3 points better on absolute performance and 6.7 points bette	×	0.03
The structural probe most effectively recovers tree structure from the 7th or 8th mBERT layer.	×	0.03
Increasing the probe maximum rank beyond approximately 64 or 128 gives diminishing returns.	×	0.01
The structural probe method of Hewitt and Manning (2019) probes for syntactic trees by finding a linear transformation u	×	0.06
The linear transformation of the structural probe is interpreted as defining a syntactic subspace.	×	0.04

References

- `http://arxiv.org/abs/2402.14743v2`
- `http://arxiv.org/abs/2010.08275v1`
- `http://arxiv.org/abs/2005.04511v2`