Clustering Granularity and Downstream Performance in Multilingual Subword Language Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the quantitative trade-off between clustering granularity (e.g., vocabulary sizes of 50 vs. 200) and downstream task performance (measured by word error rate) in multilingual SLMs. 8 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. Research question: What is the quantitative trade-off between clustering granularity (e.g., vocabulary sizes of 50 vs. 200) and downstream task performance (measured by word error rate) in multilingual SLMs?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

12 papers retrieved. 8 claims extracted; 4 independently verified. Quality review score: 6.5/10.

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
Replacing the large cross-lingual embedding matrix of XLM-R and re-initializing it for a new, language-specific vocabula	✓	0.24
LAPT with trainable transformer layers is denoted as LAPT-FULL, and training with the transformer frozen (but trainable	×	0.02
Training data for all languages is obtained from OSCAR v.22.01 (Abadji et al., 2022).	×	0.03
For targets with less than 1GB of data, the entire dataset is used for training new Sentencepiece models. For those with	×	0.04
Embedding-replacement techniques proposed in the monolingual model adaptation literature are inadequate for adapting mul	✓	0.24
Replacing large cross-lingual vocabularies with smaller language-specific ones provides a computationally-efficient meth	✓	0.28
The simple re-initialization techniques proposed, based on script-wise embedding subdistributions, rival techniques suc	✓	0.26
Figure 5b in the Appendix verifies that the clusters capture the initial vs. medial token distinction.	×	0.01

References

- http://arxiv.org/abs/2304.00649v1
- http://arxiv.org/abs/2309.04679v2
- http://arxiv.org/abs/2508.06621v1