Synergistic Effects of Vocabulary Augmentation and Script Transliteration on POS Tagging Accuracy in Low-Resource Languages

Assignee Research

June 28, 2026

Abstract

Pretrained multilingual language models have become a common tool in transferring NLP capabilities to low-resource languages, often with adaptations. In this work, we study the performance, extensibility, and interaction of two such adaptations: vocabulary augmentation and script transliteration. Our evaluations on part-of-speech tagging, universal dependency parsing, and named entity recognition in nine diverse low-resource languages uphold the viability of these approaches while raising new questions around how to optimally adapt multilingual models to low-resource settings.

1 Introduction

This paper examines: Specializing Multilingual Language Models: An Empirical Study. Research question: Does combining vocabulary augmentation with script transliteration yield synergistic improvements in part-of-speech tagging accuracy for low-resource languages over using either adaptation alone?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

1 papers retrieved. 15 claims extracted; 11 independently verified. Quality review score: 7.5/10.

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
Vocabulary augmentation improves performance on a small set of low-resource languages.	✓	0.16
Transliteration quality may be a confounding factor in the performance of script transliteration methods.	×	0.11
Vocabulary augmentation offers an appealing balance of performance and cost.	✓	0.21
Vocabulary augmentation improves vocabulary coverage of the target language.	✓	0.16
Vocabulary augmentation results hold to an even more pronounced degree in unseen low-resource languages with non-Latin s	✓	0.24
LAPT achieves 95.74 $\pm$ 0.44 performance.	×	0.09
VA achieves 95.28 $\pm$ 0.51, 97.15 $\pm$ 0.04, and 93.28 $\pm$ 0.19 performance.	✓	0.20
MBERT achieves 71.83 $\pm$ 0.90 performance.	×	0.09
LAPT achieves 72.77 $\pm$ 1.12 performance.	×	0.11
VA achieves 73.22 $\pm$ 1.23 and 91.62 $\pm$ 0.23 performance.	✓	0.15
FASTT achieves 84.26 $\pm$ 0.86, 87.98 $\pm$ 0.76, 67.21 $\pm$ 4.30, and 33.53 $\pm$ 17.89 performance.	✓	0.25
BERT achieves 88.08 $\pm$ 0.62, 90.31 $\pm$ 0.20, 76.58 $\pm$ 0.98, 54.64 $\pm$ 3.51, 61.54 $\pm$ 3.70, and 92.85 $\pm$ 2.04 performance.	✓	0.32
MBERT achieves 91.13 $\pm$ 0.07, 92.56 $\pm$ 0.09, 82.82 $\pm$ 0.57, 61.86 $\pm$ 2.60, 50.76 $\pm$ 1.86, 94.60 $\pm$ 0.34, 92.13 $\pm$ 0.27, 61.85 $\pm$	✓	0.56
LAPT achieves 91.61 $\pm$ 0.74, 92.96 $\pm$ 0.13, 84.13 $\pm$ 0.78, 81.53 $\pm$ 2.33, 56.76 $\pm$ 4.91, 95.17 $\pm$ 0.29, 92.41 $\pm$ 0.15, 59.17 $\pm$	✓	0.60
VA achieves 91.38 $\pm$ 0.56, 92.70 $\pm$ 0.11, 84.82 $\pm$ 1.00, 80.00 $\pm$ 2.77, 68.93 $\pm$ 3.30, 95.43 $\pm$ 0.22, 92.43 $\pm$ 0.16, 64.23 $\pm$ 3.	✓	0.59

References

- http://arxiv.org/abs/2106.09063v4