Few-Shot Cross-Lingual NER Robustness to Domain Shifts in Low-Resource Languages

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How robust are few-shot cross-lingual NER performance gains from large autoregressive models to domain shifts, as evaluated on the WikiANN benchmark in low-resource languages. 14 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Specializing Multilingual Language Models: An Empirical Study. Research question: How robust are few-shot cross-lingual NER performance gains from large autoregressive models to domain shifts, as evaluated on the WikiANN benchmark in low-resource languages?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

15 papers retrieved. 14 claims extracted; 3 independently verified. Quality review score: 4.8/10.

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
Chau et al. (2020) augment the model's vocabulary to more effectively tokenize text and then pretrain on a small amount	×	0.03
Chau et al. (2020) report significant performance improvements on a small set of low-resource languages.	×	0.09
Muller et al. (2021) propose to transliterate text in the target language to Latin script to be better tokenized by the	×	0.04
Muller et al. (2021) observe mixed results and note that transliteration quality may be a confounding factor.	×	0.04
The study verifies the performance of vocabulary augmentation on three tasks in a diverse set of nine low-resource langu	✓	0.16
Gains from vocabulary augmentation are associated with improved vocabulary coverage of the target language.	×	0.06
There is a negative interaction between vocabulary augmentation and transliteration in the framework for specializing mu	✓	0.20
Vocabulary augmentation offers an appealing balance of performance and cost.	×	0.06
The study expands on the dependency parsing evaluations of Chau et al. (2020) by additionally considering named entity r	✓	0.18
The study computes the CWR for each token as a weighted sum of the activations at each MBERT layer.	×	0.02
In the reported results, the VA method achieved a score of 95.28 $\pm$ 0.51 on one metric, compared to 95.74 $\pm$ 0.44 for LAPT	×	0.01
In the reported results, the VA method achieved a score of 73.22 $\pm$ 1.23, outperforming MBERT (71.83 $\pm$ 0.90) and LAPT (72	×	0.02
In the reported results, the VA method achieved a score of 68.93 $\pm$ 3.30 on a specific task, outperforming BERT (54.64 $\pm$	×	0.04
In the reported results, the VA method achieved a score of 83.74 on an aggregate metric, outperforming LAPT (81.72) and	×	0.02

References

- http://arxiv.org/abs/2311.14544v1
- http://arxiv.org/abs/2106.09063v4
- http://arxiv.org/abs/2305.14857v1