Typological Features in Source Language Selection for Cross-Lingual NER on Low-Resource African Languages

Assignee Research

June 27, 2026

Abstract

Cross-lingual transfer learning enables NLP for low-resource languages by leveraging labeled data from higher-resource sources, yet existing comparisons of source language selection strategies do not control for total training data, confounding language selection effects with data quantity effects. We introduce Budget-Xfer, a framework that formulates multi-source cross-lingual transfer as a budget-constrained resource allocation problem. Given a fixed annotation budget B, our framework jointly optimizes which source languages to include and how much data to allocate from each. We evaluate fou

1 Introduction

This paper examines: Budget-Xfer: Budget-Constrained Source Language Selection for Cross-Lingual Transfer to African Languages. Research question: Does incorporating typological features into the selection of multiple source languages reduce the performance degradation of cross-lingual NER models on low-resource African languages compared to random source selection?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 9 claims extracted; 7 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
cosine_gap predicts cross-lingual transfer with a correlation coefficient (ρ) of 0.4–0.6.	×	0.11
FLORES-200 provides 1,012 parallel sentences per language.	×	0.12
MasakhaNER 2.0 provides entity-annotated data in 20 African languages with standard categories (PER, ORG, LOC, DATE).	✓	0.20
AfriSenti covers 14 African languages with three classes (positive, negative, neutral).	✓	0.17
The evaluation metric for NER is entity-level F1 (micro-averaged) computed with seqeval.	✓	0.20
The evaluation metric for sentiment analysis is weighted F1 to account for class imbalance.	✓	0.18
Multi-source transfer significantly outperforms single-source transfer with Cohen's d effect sizes ranging from 0.80 to	✓	0.24
Among multi-source strategies, differences are modest and non-significant.	✓	0.25
Random selection outperforms similarity-based selection for NER but not sentiment analysis.	✓	0.21

References

- http://arxiv.org/abs/2106.09063v4
- http://arxiv.org/abs/2603.27651v1
- http://arxiv.org/abs/2501.18750v1