Synthetic Data Diversity and Robustness in Teacher-Student NER Models for Low-Resource Languages

Assignee Research

July 4, 2026

Abstract

Named Entity Recognition(NER) for low-resource languages aims to produce robust systems for languages where there is limited labeled training data available, and has been an area of increasing interest within NLP. Data augmentation for increasing the amount of low-resource labeled data is a common practice. In this paper, we explore the role of synthetic data in the context of multilingual, low-resource NER, considering 11 languages from diverse language families. Our results suggest that synthetic data does in fact hold promise for low-resource language NER, though we see significant variatio

1 Introduction

This paper examines: Does Synthetic Data Help Named Entity Recognition for Low-Resource Languages?. Research question: What is the impact of synthetic data diversity on the robustness of teacher-student NER models when evaluated on low-resource target languages?.

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 8.5/10.

3 Results

16 papers retrieved. 22 claims extracted; 20 independently verified. Quality review score: 8.5/10.

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
Seed data is a reasonable choice when there is no available labeled data, and is better than using entirely automaticall	✓	0.22
A small amount of high-quality data, coupled with cross-lingual transfer from a related language, always offers better p	✓	0.23
The study focuses on 11 languages from three distinct language families: Tamil, Kannada, Malayalam, Telugu (Dravidian),	✓	0.32
Igbo, Yoruba, and Kinyarwanda are not among the 100 languages in the XLM-Roberta pretraining corpus.	✓	0.22
The Universal NER dataset is used as the high-quality, manually annotated dataset for Swedish, Danish, and Slovak.	✓	0.26
MasakhaNER2 is used for Kinyarwanda, Swahili, Igbo, and Yoruba.	✓	0.19
The Naamapadam dataset is used for Tamil, Kannada, Malayalam, and Telugu.	✓	0.21
The Naamapadam dataset's train and validation splits are constructed using parallel corpora and contain some noise.	✓	0.23
The Naamapadam dataset's test sets contain 500-1000 datapoints per language and are completely manually annotated.	✓	0.25
The Naamapadam dataset remains the largest NER resource for Tamil, Kannada, Malayalam, and Telugu.	✓	0.25
All datasets cover largely identical NER categories, allowing for comparisons between them.	✓	0.20
The Universal NER and Naamapadam datasets cover persons, locations, and organizations as categories.	✓	0.25
MasakhaNER2 data covers persons, locations, organizations, and dates as categories.	✓	0.16
Models trained entirely on LLM-generated data are compared with those trained using WikiANN, a large, automatically crea	✓	0.21
WikiANN covers the 11 languages studied and represents a different form of synthetic data generated from scraping knowle	✓	0.22
WikiANN has no manual annotations but is frequently used as a standard low-resource NER benchmark.	✓	0.19
The average percentage of usable training datapoints from GPT-4.1, Llama-3.1, and aya-expanse are 82.6%, 59.7%, and 11.1	✓	0.27
GPT-4.1 is assumed to be the state of the art among the LLMs used.	×	0.09
Llama-3.1-8B-Instruct is a much smaller, open-	✓	0.20

References

- http://arxiv.org/abs/2303.09306v2
- http://arxiv.org/abs/2505.16814v3
- http://arxiv.org/abs/2004.12440v2