Impact of Multilingual Pretraining Objectives on Zero-Shot Cross-Lingual NER Performance

Assignee Research

July 19, 2026

Abstract

Named entity recognition (NER) suffers from the scarcity of annotated training data, especially for low-resource languages without labeled data. Cross-lingual NER has been proposed to alleviate this issue by transferring knowledge from high-resource languages to low-resource languages via aligned cross-lingual representations or machine translation results. However, the performance of cross-lingual NER methods is severely affected by the unsatisfactory quality of translation or label projection. To address these problems, we propose a Cross-lingual Entity Projection framework (CROP) to enable

1 Introduction

This paper examines: CROP: Zero-shot Cross-lingual Named Entity Recognition with Multilingual Labeled Sequence Translation. Research question: What is the impact of incorporating multilingual pretraining objectives (e.g., MLM vs. TLM vs. XLMR-style) in the teacher model on the zero-shot F1 performance of student models in cross-lingual NER across high-resource and low-resource language pairs in XTREME-NER?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

11 papers retrieved. 13 claims extracted; 10 independently verified. Quality review score: 7.4/10.

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
The labeled multilingual model is tuned on the CCaligned dataset, which includes parallel data in English and 39 other l	✓	0.18
The valid and test sets are from the FLORES-101 dataset.	✓	0.18
The cross-lingual dataset is constructed from CoNLL-2002 for Spanish (es) and Dutch (nl) NER, CoNLL-2003 for English (en	✓	0.29
All entities are classified into 4 entity types in BOI-2 format, including LOC, MISC, ORG, and PER.	✓	0.21
The proposed method is evaluated on the cross-lingual NER dataset from the XTREME benchmark, with named entities annotat	✓	0.28
The synthetic data is post-processed by filtering translated sentences with incorrect language, deleting sequences excee	✓	0.21
The method is compared with baselines initialized by cross-lingual pretrained models including mBERT and XLM-R, and expe	✓	0.16
UniTrans unifies both model transfer and data transfer for cross-lingual NER.	×	0.11
MulDA proposes a sequence translation method to translate labeled training data of the source languages to other languag	✓	0.28
The metrics used for evaluation are entity-level precision, recall, and F1 scores.	✓	0.16
The CoNLL-5 dataset includes training, dev, and test sets with detailed statistics provided in Table 1.	×	0.13
The XTREME-40 dataset uses the same split for the training, dev, and test set as previous work.	✓	0.19
The performance of the proposed method (CROP) is compared with various baselines and shows improvements in F1 scores acr	×	0.09

References

- `http://arxiv.org/abs/2210.07022v1`

- http://arxiv.org/abs/2004.12440v2
- http://arxiv.org/abs/2509.02160v2