Impact of LLM-Generated Synthetic Data on Zero-Shot Cross-Lingual NER Entity Span Accuracy Across Typologically Diverse Languages

Assignee Research

June 22, 2026

Abstract

Natural language tasks like Named Entity Recognition (NER) in the clinical domain on non-English texts can be very time-consuming and expensive due to the lack of annotated data. Cross-lingual transfer (CLT) is a way to circumvent this issue thanks to the ability of multilingual large language models to be fine-tuned on a specific task in one language and to provide high accuracy for the same task in another language. However, other methods leveraging translation models can be used to perform NER without annotated data in the target language, by either translating the training set or test set.

1 Introduction

This paper examines: Multilingual Clinical NER: Translation or Cross-lingual Transfer?. Research question: What is the impact of incorporating synthetic data generated by large language models on the zero-shot cross-lingual transfer performance of NER models, as evaluated by entity span accuracy across typologically diverse 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 7.3/10.

3 Results

15 papers retrieved. 22 claims extracted; 16 independently verified. Quality review score: 7.3/10.

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 MedNERF dataset contains 100 sentences and 406 entities.	✓	0.15
The MedNERF dataset is available at https://huggingface.co/datasets/Posos/MedNERF.	✓	0.19
The GERNERMED test dataset consists of 30 sentences and 119 entities.	×	0.15
The n2c2 dataset contains 16,656 sentences and 65,495 entities.	×	0.12
The MedNERF dataset uses the same annotation guidelines as the n2c2 dataset.	×	0.09
The MedNERF dataset does not include the ADE, ROUTE, and REASON labels.	×	0.11
The MedNERF dataset includes the labels DRUG, STRENGTH, FREQUENCY, DURATION, DOSAGE, and FORM.	✓	0.22
The MedNERF dataset was built using a sample of French medical prescriptions.	✓	0.21
The MedNERF dataset was annotated with dosage instructions obtained from a private set of scanned typewritten drug presc	✓	0.23
The MedNERF dataset was created using a state-of-the-art Optical Character Recognition (OCR) software.	×	0.11
The MedNERF dataset was manually curated to identify sentences containing dosage instructions.	✓	0.18
The MedNERF dataset is intended to be a test and not a training dataset.	✓	0.17
The GERNERMED test dataset was released by Frei et al. (2022).	✓	0.18
The GERNERMED test dataset consists of 30 sentences from physicians.	✓	0.17
The GERNERMED test dataset is annotated with the same guidelines as n2c2.	×	0.15
The translation-based methods require a translation and an alignment model.	✓	0.16
The selection of translation and alignment algorithms should not be based on downstream cross-lingual abilities.	✓	0.20
The translate-train approach consists in constructing a translated version of the n2c2 dataset and training a NER algori	✓	0.33
XLM-R Base is preferred over mBERT as it has the same number of layers but outperforms it on multilingual benchmarks.	✓	0.23
XLM-R Large is used to evaluate the impact of model size.	✓	0.16
distilmBERT is used to give insights about what is possible with less resources.	✓	0.15
With CLT, models will only see English NER	✓	0.26

References

- http://arxiv.org/abs/2003.11080v5
- http://arxiv.org/abs/2505.17125v1
- http://arxiv.org/abs/2306.04384v1