Artificial Code-Switching in Pre-Training for Cross-Lingual Zero-Shot Transfer

Assignee Research

July 8, 2026

Abstract

Multilingual pre-trained models have achieved remarkable performance on cross-lingual transfer learning. Some multilingual models such as mBERT, have been pre-trained on unlabeled corpora, therefore the embeddings of different languages in the models may not be aligned very well. In this paper, we aim to improve the zero-shot cross-lingual transfer performance by proposing a pre-training task named Word-Exchange Aligning Model (WEAM), which uses the statistical alignment information as the prior knowledge to guide cross-lingual word prediction. We evaluate our model on multilingual machine rea

1 Introduction

This paper examines: Bilingual Alignment Pre-Training for Zero-Shot Cross-Lingual Transfer. Research question: How does artificial code-switching during pre-training impact cross-lingual zero-shot transfer performance as measured by F1 score on XNLI compared to standard multilingual pre-training?.

2 Methodology

Systematic literature search across multiple databases yielded 7 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

7 papers retrieved. 17 claims extracted; 12 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
The mBERT+TLM model outperforms mBERT by a large margin in the zero-shot setting.	✓	0.27
The mBERT+TLM model is not as good as the mBERT in the translate-train setting.	✓	0.24
The mBERT+WEAM model improves the scores in the zero-shot setting.	✓	0.20
The mBERT+WEAM model outperforms mBERT in the translate-train setting.	✓	0.27
The mBERT+WEAM model can exceed the performance of translate-train even with zero-shot training.	✓	0.25
The mBERT+TLM and word-aligned mBERT achieved similar improvements on XNLI compared to mBERT.	✓	0.24
The mBERT+WEAM model has significantly outperformed both mBERT+TLM and word-aligned mBERT on XNLI.	✓	0.17
The mBERT+WEAM result is slightly lower but close to the translate-train result on XNLI.	✓	0.24
The examples in XNLI have shorter input sequences and thus have fewer translation noises.	✓	0.17
The masking probability is set to 0.3 during the pre-training steps.	×	0.15
The learning rate is set to 5e-5.	×	0.08
The batch size is set to 32.	×	0.09
The max sequence length is set to 128.	×	0.10
The number of pre-training epochs is set to 2.	✓	0.16
λ is set to 1.	×	0.07
The WEAM model uses FastAlign to identify bilingual word pairs in parallel bilingual sentence pairs.	✓	0.24
The WEAM model performs multilingual prediction and cross-lingual prediction for each masked token.	✓	0.19

References

- http://arxiv.org/abs/2406.13361v1
- http://arxiv.org/abs/2106.01732v2

- http://arxiv.org/abs/2102.12407v1