A Comparison of Language Modeling and Translation as Multilingual Pretraining Objectives

Zihao Li*[1], Shaoxiong Ji*[1], Timothee Mickus[1], Vincent Segonne[2], and Jörg Tiedemann[1]

[1]University of Helsinki
[2]Université Bretagne Sud
firstname.lastname@[1]helsinki.fi, [2]univ-ubs.fr

Abstract

Pretrained language models (PLMs) display impressive performances and have captured the attention of the NLP community. Establishing best practices in pretraining has, therefore, become a major focus of NLP research, especially since insights gained from monolingual English models may not necessarily apply to more complex multilingual models. One significant caveat of the current state of the art is that different works are rarely comparable: they often discuss different parameter counts, training data, and evaluation methodology. This paper proposes a comparison of multilingual pretraining objectives in a controlled methodological environment. We ensure that training data and model architectures are comparable, and discuss the downstream performances across 6 languages that we observe in probing and fine-tuning scenarios. We make two key observations: (1) the architecture dictates which pretraining objective is optimal; (2) multilingual translation is a very effective pretraining objective under the right conditions. We make our code, data, and model weights available at https://github.com/Helsinki-NLP/lm-vs-mt.

1 Introduction

The release of BERT (Devlin et al., 2019) has marked a paradigm shift in the NLP landscape and has ushered in a thorough investment of the NLP research community in developing large language models that can readily be adapted to novel situations. The design, training, and evaluation of these models has become a significant enterprise of its own. In recent years, that sustained interest has shifted also to encompass multilingual models (e.g., Muennighoff et al., 2022; Alves et al., 2024). There is considerable variation as to how such models are trained: For instance, some rely on datasets comprising multiple languages without explicit cross-lingual supervision (e.g., Liu et al., 2020), and some use explicit supervision (Xue et al., 2021).

One complication that arises from this blossoming field of study is that much of the work being carried out is not directly comparable beyond the raw performances on some well-established benchmark, a procedure which may well be flawed (Gorman and Bedrick, 2019). Avoiding apples-to-oranges comparison requires a methodical approach in strictly comparable circumstances, which is the stance we adopt in this paper.

In short, we focus on two variables—model architecture and pretraining objectives—and set out to train five models in strictly comparable conditions and compare their monolingual performances in three downstream applications: sentiment analysis, named entity recognition, and POS-tagging. The scope of our study spans from encoder-decoder machine translation models, to decoder-only causal language models and encoder-only BERT-like masked language models. We categorize them into double-stacks (encoder-decoder) and single-stacks (encoder-only or decoder-only) models. We intend to answer two research questions: (i) Does the explicit cross-lingual train-

*Equal contribution and corresponding authors.

ing signal of translation objectives foster better downstream performances in monolingual tasks? (ii) Is the optimal choice of architecture independent of the training objective?

2 Methods and Settings

We start our inquiry by adopting a principled stance: We train strictly comparable models with MT and LM objectives before contrasting their performances on monolingual tasks.

2.1 Models and objectives

To allow a systematic evaluation, we train models with various neural network architectures and learning objectives. All models are based on the transformer architecture (Vaswani et al., 2017) and implemented in fairseq (Ott et al., 2019). We consider both double-stacks (encoder-decoder) and single-stacks (encoder-only or decoder-only) models.

The two double-stack models are variants of the BART architecture of (Lewis et al., 2020); they are trained either on a straightforward machine translation (MT) objective, using language tokens to distinguish the source, or on the original denoising auto-encoder objective of Lewis et al. We refer to these two models as 2-LM and 2-MT respectively.

We also consider three single-stack models: (i) an encoder-only model trained on the masked language modeling objective (MLM) of Devlin et al. (2019); (ii) an autoregressive causal language model (CLM), similar to Radford et al. (2019); and (iii) an autoregressive model trained to generate a sentence, followed by its translation in the language specified by a given control token, known as a translation language model (TLM) as proposed by Conneau and Lample (2019).

2.2 Pretraining conditions

Our core focus is on guaranteeing comparable conditions across the different pretraining objectives we consider. This entails that our datasets need to be doubly structured: both in documents for CLM pretraining; and as aligned bitexts for MT pretraining. Two datasets broadly match these criteria: the UNPC (Ziemski et al., 2016) and OpenSubtitles (OpSub; Tiedemann, 2012) corpora. The choice also narrows down the languages considered in this study: we take the set of languages present in both resources, namely the six languages in UNPC: Arabic (AR), Chinese (ZH), English (EN), French (FR), Russian (RU), and Spanish (ES).

2.3 Downstream evaluation

The evaluations encompassed both sequence-level and token-level classification tasks using datasets tailored for sentiment analysis (SA), named entity recognition (NER), part-of-speech (POS) tagging, and natural language inference (NLI).

3 Results

Double-stack models consistently show that 2-MT outperforms 2-LM across all languages after fine-tuning and probing. For single-stack models, CLM is most effective in probing, while MLM generally ranks first in fine-tuning for POS, NER, and NLI.

4 Conclusion

This paper conducts an empirical study of how pretraining conditions of multilingual models impact downstream performances. We observe that translation objectives can be highly effective for model pretraining, specifically for double-stack architectures.

References

[1] Alves et al. 2024. Tower: An open multilingual large language model for translation-related tasks.

[2] Conneau and Lample. 2019. Cross-lingual language model pretraining.

[3] Devlin et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

[4] Ziemski et al. 2016. The United Nations parallel corpus v1.0.

A Overview of pretraining objectives

Example data point [MISSING]

Table 1: Example data point for all pretraining objectives.