Emerging Cross-lingual Structure in Pretrained Language Models

Shijie Wu[♠*] **Alexis Conneau**[♡*]
Haoran Li[♡] **Luke Zettlemoyer**[♡] **Veselin Stoyanov**[♡]
[♠]Department of Computer Science, Johns Hopkins University
[♡]Facebook AI
shijie.wu@jhu.edu, aconneau@fb.com
{aimeeli,lsz,ves}@fb.com

Abstract

We study the problem of multilingual masked language modeling, i.e. the training of a single model on concatenated text from multiple languages, and present a detailed study of several factors that influence why these models are so effective for cross-lingual transfer. We show, contrary to what was previously hypothesized, that transfer is possible even when there is no shared vocabulary across the monolingual corpora and also when the text comes from very different domains. The only requirement is that there are some shared parameters in the top layers of the multi-lingual encoder. To better understand this result, we also show that representations from monolingual BERT models in different languages can be aligned post-hoc quite effectively, strongly suggesting that, much like for non-contextual word embeddings, there are universal latent symmetries in the learned embedding spaces. For multilingual masked language modeling, these symmetries are automatically discovered and aligned during the joint training process.

1 Introduction

Multilingual language models such as mBERT (Devlin et al., 2019) and XLM (Lample and Conneau, 2019) enable effective cross-lingual transfer — it is possible to learn a model from supervised data in one language and apply it to another with no additional training. Recent work has shown that transfer is effective for a wide range of tasks (Wu and Dredze, 2019; Pires et al., 2019). These work speculates why multilingual pretraining works (e.g. shared vocabulary), but only experiment with a single reference mBERT and is unable to systematically measure these effects.

In this paper, we present the first detailed empirical study of the effects of different masked language modeling (MLM) pretraining regimes on cross-lingual transfer. Our first set of experiments is a detailed ablation study on a range of zero-shot cross-lingual transfer tasks. Much to our surprise, we discover that language universal representations emerge in pretrained models without the requirement of any shared vocabulary or domain similarity, and even when only a subset of the parameters in the joint encoder are shared. In particular, by systematically varying the amount of shared vocabulary between two languages during pretraining, we show that the amount of overlap only accounts for a few points of performance in transfer tasks, much less than might be expected. By sharing parameters alone, pretraining learns to map similar words and sentences to similar hidden representations.

To better understand these effects, we also analyze multiple monolingual BERT models trained independently. We find that monolingual models trained in different languages learn representations that align with each other surprisingly well, even though they have no shared parameters. This result closely mirrors the widely observed fact that word embeddings can be effectively aligned across languages (Mikolov et al., 2013). Similar dynamics are at play in MLM pretraining, and at least in part explain why they aligned so well with relatively little parameter tying in our earlier experiments.

This type of emergent language universality has interesting theoretical and practical implications. We gain insight into why the models transfer so well and open up new lines of inquiry into what properties emerge in common in these representations. They also suggest it should be possible to adapt pretrained models to new languages with little additional training and it may be possible to better align independently trained representations without having to jointly train on all of the (very large) unlabeled data that could be gathered. For example, concurrent work has shown that a pre-

[*]Equal contribution. Work done while Shijie was interning at Facebook AI.