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

arXiv:1911.01464v3 [cs.CL] 7 May 2020

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.

trained MLM model can be rapidly fine-tuned to another language (Artetxe et al., 2019).

This paper offers the following contributions:

- We provide a detailed ablation study on cross-lingual representation of bilingual BERT. We show parameter sharing plays the most important role in learning cross-lingual representation, while shared BPE, shared softmax and domain similarity play a minor role.
- We demonstrate even without any shared subwords (anchor points) across languages, cross-lingual representation can still be learned. With bilingual dictionary, we propose a simple technique to create more anchor points by creating synthetic code-switched corpus, benefiting especially distantly-related languages.
- We show monolingual BERTs of different language are similar with each other. Similar to word embeddings (Mikolov et al., 2013), we show monolingual BERT can be easily aligned with linear mapping to produce cross-lingual representation space at each level.

2 Background

Language Model Pretraining Our work follows in the recent line of language model pretraining. ELMo (Peters et al., 2018) first popularized representation learning from a language model. The representations are used in a transfer learning setup to improve performance on a variety of downstream NLP tasks. Follow-up work by Howard and Ruder (2018); Radford et al. (2018) further improves on this idea by fine-tuning the entire language model. BERT (Devlin et al., 2019) significantly outperforms these methods by introducing a masked-language model and next-sentence prediction objectives combined with a bi-directional transformer model.

The multilingual version of BERT (dubbed mBERT) trained on Wikipedia data of over 100 languages obtains strong performance on zero-shot cross-lingual transfer without using any parallel data during training (Wu and Dredze, 2019; Pires et al., 2019). This shows that multilingual representations can emerge from a shared Transformer with a shared subword vocabulary. Cross-lingual language model (XLM) pretraining (Lample and Conneau, 2019) was introduced concurrently to mBERT. On top of multilingual masked language models, they investigate an objective based on parallel sentences as an explicit cross-lingual signal. XLM shows that cross-lingual language model pretraining leads to a new state of the art on XNLI (Conneau et al., 2018), supervised and unsupervised machine translation (Lample et al., 2018). Other work has shown that mBERT outperforms word embeddings on token-level NLP tasks (Wu and Dredze, 2019), and that adding character-level information (Mulcaire et al., 2019) and using multi-task learning (Huang et al., 2019) can improve cross-lingual performance.

Alignment of Word Embeddings Researchers working on word embeddings noticed early that embedding spaces tend to be shaped similarly across different languages (Mikolov et al., 2013). This inspired work in aligning monolingual embeddings. The alignment was done by using a bilingual dictionary to project words that have the same meaning close to each other (Mikolov et al., 2013). This projection aligns the words outside of the dictionary as well due to the similar shapes of the word embedding spaces. Follow-up efforts only required a very small seed dictionary (e.g., only numbers (Artetxe et al., 2017)) or even no dictionary at all (Conneau et al., 2017; Zhang et al., 2017). Other work has pointed out that word embeddings may not be as isomorphic as thought (Søgaard et al., 2018) especially for distantly related language pairs (Patra et al., 2019). Ormazabal et al. (2019) show joint training can lead to more isomorphic word embeddings space.

Schuster et al. (2019) showed that ELMo embeddings can be aligned by a linear projection as well. They demonstrate a strong zero-shot cross-lingual transfer performance on dependency parsing. Wang et al. (2019) align mBERT representations and evaluate on dependency parsing as well.

Neural Network Activation Similarity We hypothesize that similar to word embedding spaces, language-universal structures emerge in pretrained language models. While computing word embedding similarity is relatively straightforward, the same cannot be said for the deep contextualized BERT models that we study. Recent work introduces ways to measure the similarity of neural network activation between different layers and different models (Laakso and Cottrell, 2000; Li et al., 2016; Raghu et al., 2017; Morcos et al., 2018; Wang et al., 2018). For example, Raghu et al.

(2017) use canonical correlation analysis (CCA) and a new method, singular vector canonical correlation analysis (SVCCA), to show that early layers converge faster than upper layers in convolutional neural networks. Kudugunta et al. (2019) use SVCCA to investigate the multilingual representations obtained by the encoder of a massively multilingual neural machine translation system (Aharoni et al., 2019). Kornblith et al. (2019) argues that CCA fails to measure meaningful similarities between representations that have a higher dimension than the number of data points and introduce the centered kernel alignment (CKA) to solve this problem. They successfully use CKA to identify correspondences between activations in networks trained from different initializations.

3 Cross-lingual Pretraining

We study a standard multilingual masked language modeling formulation and evaluate performance on several different cross-lingual transfer tasks, as described in this section.

3.1 Multilingual Masked Language Modeling

Our multilingual masked language models follow the setup used by both mBERT and XLM. We use the implementation of Lample and Conneau (2019). Specifically, we consider continuous streams of 256 tokens and mask 15% of the input tokens which we replace 80% of the time by a mask token, 10% of the time with the original word, and 10% of the time with a random word. Note the random words could be foreign words. The model is trained to recover the masked tokens from its context (Taylor, 1953). The subword vocabulary and model parameters are shared across languages. Note the model has a softmax prediction layer shared across languages. We use Wikipedia for training data, preprocessed by Moses (Koehn et al., 2007) and Stanford word segmenter (for Chinese only) and BPE (Sennrich et al., 2016) to learn subword vocabulary. During training, we sample a batch of continuous streams of text from one language proportionally to the fraction of sentences in each training corpus, exponentiated to the power 0.7.

Pretraining details Each model is a Transformer (Vaswani et al., 2017) with 8 layers, 12 heads and GELU activiation functions (Hendrycks and Gimpel, 2016). The output softmax layer is tied with input embeddings (Press and Wolf, 2017). The embeddings dimension is 768, the hidden dimension of the feed-forward layer is 3072, and dropout is 0.1. We train our models with the Adam optimizer (Kingma and Ba, 2014) and the inverse square root learning rate scheduler of Vaswani et al. (2017) with 10^{-4} learning rate and 30k linear warmup steps. For each model, we train it with 8 NVIDIA V100 GPUs with 32GB of memory and mixed precision. It takes around 3 days to train one model. We use batch size 96 for each GPU and each epoch contains 200k batches. We stop training at epoch 200 and select the best model based on English dev perplexity for evaluation.

3.2 Cross-lingual Evaluation

We consider three NLP tasks to evaluate performance: natural language inference (NLI), named entity recognition (NER) and dependency parsing (Parsing). We adopt the **zero-shot cross-lingual transfer** setting, where we (1) fine-tune the pretrained model on English and (2) directly transfer the model to target languages. We select the model and tune hyperparameters with the English dev set. We report the result on average of best two set of hyperparameters.

Fine-tuning details We fine-tune the model for 10 epochs for NER and Parsing and 200 epochs for NLI. We search the following hyperparameter for NER and Parsing: batch size $\{16, 32\}$; learning rate $\{$2e-5, 3e-5, 5e-5$\}$. For XNLI, we search: batch size $\{4, 8\}$; encoder learning rate $\{$1.25e-6, 2.5e-6, 5e-6$\}$; classifier learning rate $\{$5e-6, 2.5e-5, 1.25e-4$\}$. We use Adam with fixed learning rate for XNLI and warmup the learning rate for the first 10% batch then decrease linearly to 0 for NER and Parsing. We save checkpoint after each epoch.

NLI We use the cross-lingual natural language inference (XNLI) dataset (Conneau et al., 2018). The task-specific layer is a linear mapping to a softmax classifier, which takes the representation of the first token as input.

NER We use WikiAnn (Pan et al., 2017), a silver NER dataset built automatically from Wikipedia, for English-Russian and English-French. For English-Chinese, we use CoNLL 2003 English (Tjong Kim Sang and De Meulder, 2003) and a Chinese NER dataset (Levow, 2006), with realigned Chinese NER labels based on the Stanford word segmenter. We model NER as BIO tagging. The task-specific layer is a linear mapping to a softmax