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.