on monolingual Wikipedia corpora of 102 languages (comprising all Multi-SimLex languages) with a 12-layer Transformer network, and yields 768-dimensional representations. Since the concept pairs in Multi-SimLex are lowercased, we use the uncased version of M-BERT.[10] M-BERT comprises all Multi-SimLex languages, and its evident ability to perform cross-lingual transfer (Pires, Schlinger, and Garrette 2019; Wu and Dredze 2019; Wang et al. 2020) also makes it a convenient baseline model for cross-lingual experiments later in §8. The second multilingual model we consider, XLM-100,[11] is pretrained on Wikipedia dumps of 100 languages, and encodes each concept into a $1,280$-dimensional representation. In contrast to M-BERT, XLM-100 drops the next-sentence prediction objective and adds a cross-lingual masked language modeling objective. For both encoders, the representations of each concept are computed as averages over the last $H = 4$ hidden layers in all experiments, as suggested by Wu et al. (2019).[12]

Besides M-BERT and XLM, covering multiple languages, we also analyze the performance of "language-specific" BERT and XLM models for the languages where they are available: Finnish, Spanish, English, Mandarin Chinese, and French. The main goal of this comparison is to study the differences in performance between multilingual "one-size-fits-all" encoders and language-specific encoders. For all experiments, we rely on the pretrained models released in the Transformers repository (Wolf et al. 2019).[13]

Unsupervised post-processing steps devised for static word embeddings (i.e., mean-centering, ABTT, UNCOVEC) can also be applied on top of contextualized embeddings if we predefine a vocabulary of word types V that will be represented in a word vector space $\mathbf{X}$. We construct such V for each language as the intersection of word types covered by the corresponding CC+Wiki fastText vectors and the (single-word or multi-word) expressions appearing in the corresponding Multi-SimLex dataset.

Finally, note that it is not feasible to evaluate a full range of available pretrained encoders within the scope of this work. Our main intention is to provide the first set of baseline results on Multi-SimLex by benchmarking a sample of most popular encoders, at the same time also investigating other important questions such as performance of static versus contextualized word embeddings, or multilingual versus language-specific pretraining. Another purpose of the experiments is to outline the wide potential and applicability of the Multi-SimLex datasets for multilingual and cross-lingual representation learning evaluation.

7.2 Results and Discussion

The results we report are Spearman's ρ coefficients of the correlation between the ranks derived from the scores of the evaluated models and the human scores provided in each Multi-SimLex dataset. The main results with static and contextualized word vectors for all test languages are summarized in Table 12. The scores reveal several interesting patterns, and also pinpoint the main challenges for future work.

[10] https://github.com/google-research/bert/blob/master/multilingual.md

[11] https://github.com/facebookresearch/XLM

[12] In our preliminary experiments on several language pairs, we have also verified that this choice is superior to: a) using the output of only the last hidden layer (i.e., $H = 1$) and b) averaging over all hidden layers (i.e., $H = 12$ for the BERT-BASE architecture). Likewise, using the special prepended '[CLS]' token rather than the constituent sub-words to encode a concept also led to much worse performance across the board.

[13] github.com/huggingface/transformers. The full list of currently supported pretrained encoders is available here: huggingface.co/models.