

well with the English ratings, but we cannot achieve the same level of correlation in Welsh with Welsh FT word embeddings. Difference in performance between two closely related languages, EST (low-resource) and FIN (high-resource), provides additional evidence in this respect.

The highest reported scores with M-BERT and XLM-100 are obtained for Mandarin Chinese and Yue Chinese: this effectively points to the weaknesses of massively multilingual training with a joint subword vocabulary spanning 102 and 100 languages. Due to the difference in scripts, “language-specific” subwords for YUE and CMN do not need to be shared across a vast amount of languages and the quality of their representation remains unscathed. This effectively means that M-BERT’s subword vocabulary contains plenty of CMN-specific and YUE-specific subwords which are exploited by the encoder when producing M-BERT-based representations. Simultaneously, higher scores with M-BERT (and XLM in Table 13) are reported for resource-rich languages such as French, Spanish, and English, which are better represented in M-BERT’s training data. We also observe lower absolute scores (and a larger number of OOVs) for languages with very rich and productive morphological systems such as the two Slavic languages (Polish and Russian) and Finnish. Since Polish and Russian are known to have large Wikipedias and Common Crawl data (Conneau et al. 2019) (e.g., their Wikipedias are in the top 10 largest Wikipedias worldwide), the problem with coverage can be attributed exactly to the proliferation of morphological forms in those languages.

Finally, while Table 12 does reveal that unsupervised post-processing is useful for all languages, it also demonstrates that peak scores are achieved with different post-processing configurations. This finding suggests that a more careful language-specific fine-tuning is indeed needed to refine word embeddings towards semantic similarity. We plan to inspect the relationship between post-processing techniques and linguistic properties in more depth in future work.

Multilingual vs. Language-Specific Contextualized Embeddings. Recent work has shown that—despite the usefulness of massively multilingual models such as M-BERT and XLM-100 for zero-shot cross-lingual transfer (Pires, Schlinger, and Garrette 2019; Wu and Dredze 2019)—stronger results in downstream tasks for a particular language can be achieved by pretraining language-specific models on language-specific data.

In this experiment, motivated by the low results of M-BERT and XLM-100 (see again Table 13), we assess if monolingual pretrained encoders can produce higher-quality word-level representations than multilingual models. Therefore, we evaluate language-specific BERT and XLM models for a subset of the Multi-SimLex languages for which such models are currently available: Finnish (Virtanen et al. 2019) (BERT-BASE architecture, uncased), French (Le et al. 2019) (the FlauBERT model based on XLM), English (BERT-BASE, uncased), Mandarin Chinese (BERT-BASE) (Devlin et al. 2019) and Spanish (BERT-BASE, uncased). In addition, we also evaluate a series of pretrained encoders available for English: (i) BERT-BASE, BERT-LARGE, and BERT-LARGE with whole word masking (WWM) from the original work on BERT (Devlin et al. 2019), (ii) monolingual “English-specific” XLM (Conneau and Lample 2019), and (iii) two models which employ parameter reduction techniques to build more compact encoders: ALBERT-B uses a configuration similar to BERT-BASE, while ALBERT-L is similar to BERT-LARGE, but with an $18\times$ reduction in the number of parameters (Lan et al. 2020).¹⁶

¹⁶ All models and their further specifications are available at the following link:
<https://huggingface.co/models>.