

and robust choice given 1) the availability of pretrained vectors in a large number of languages (Grave et al. 2018) trained on large Common Crawl (CC) plus Wikipedia (Wiki) data, and 2) their superior performance across a range of NLP tasks (Mikolov et al. 2018). In fact, FASTTEXT is an extension of the standard word-level CBOW and skip-gram word2vec models (Mikolov et al. 2013) that takes into account subword-level information, i.e. the constituent character n-grams of each word (Zhu, Vulić, and Korhonen 2019). For this reason, FASTTEXT is also more suited for modeling rare words and morphologically rich languages.⁶

We rely on 300-dimensional FT word vectors trained on CC+Wiki and available online for 157 languages.⁷ The word vectors for all languages are obtained by CBOW with position-weights, with character n-grams of length 5, a window of size 5, 10 negative examples, and 10 training epochs. We also probe another (older) collection of FT vectors, pretrained on full Wikipedia dumps of each language.⁸ The vectors are 300-dimensional, trained with the skip-gram objective for 5 epochs, with 5 negative examples, a window size set to 5, and relying on all character n-grams from length 3 to 6. Following prior work, we trim the vocabularies for all languages to the 200K most frequent words and compute representations for multi-word expressions by averaging the vectors of their constituent words.

Unsupervised Post-Processing. Further, we consider a variety of *unsupervised post-processing* steps that can be applied post-training on top of any pretrained input word embedding space *without* any external lexical semantic resource. So far, the usefulness of such methods has been verified only on the English language through benchmarks for lexical semantics and sentence-level tasks (Mu, Bhat, and Viswanath 2018). In this paper, we assess if unsupervised post-processing is beneficial also in other languages. To this end, we apply the following post-hoc transformations on the initial word embeddings:

- 1) *Mean centering* (MC) is applied after unit length normalization to ensure that all vectors have a zero mean, and is commonly applied in data mining and analysis (Bro and Smilde 2003; van den Berg et al. 2006).
- 2) *All-but-the-top* (ABTT) (Mu, Bhat, and Viswanath 2018; Tang, Mousavi, and de Sa 2019) eliminates the common mean vector and a few top dominating directions (according to PCA) from the input distributional word vectors, since they do not contribute towards distinguishing the actual semantic meaning of different words. The method contains a single (tunable) hyper-parameter dd_A which denotes the number of the dominating directions to remove from the initial representations. Previous work has verified the usefulness of ABTT in several English lexical semantic tasks such as semantic similarity, word analogies, and concept categorization, as well as in sentence-level text classification tasks (Mu, Bhat, and Viswanath 2018).
- 3) UNCOVEC (Artetxe et al. 2018) adjusts the similarity order of an arbitrary input word embedding space, and can emphasize either syntactic or semantic information in the transformed vectors. In short, it transforms the input space \mathbf{X} into an adjusted space $\mathbf{X}\mathbf{W}_\alpha$ through a linear map \mathbf{W}_α controlled by a single hyper-parameter α . The n^{th} -

⁶ We have also trained standard word-level CBOW and skip-gram with negative sampling (SGNS) on full Wikipedia dumps for several languages, but our preliminary experiments have verified that they under-perform compared to FASTTEXT. This finding is consistent with other recent studies demonstrating the usefulness of subword-level information (Vania and Lopez 2017; Mikolov et al. 2018; Zhu, Vulić, and Korhonen 2019; Zhu et al. 2019). Therefore, we do not report the results with CBOW and SGNS for brevity.

⁷ <https://fasttext.cc/docs/en/crawl-vectors.html>

⁸ <https://fasttext.cc/docs/en/pretrained-vectors.html>