



Figure 4: (a) Rating distribution and (b) distribution of pairs over the four POS classes in cross-lingual Multi-SimLex datasets averaged across each of the 66 language pairs (*y* axes plot percentages as the total number of concept pairs varies across different cross-lingual datasets). Minimum and maximum percentages for each rating interval and POS class are also plotted.

intervals contain a substantial number of concept pairs. For instance, the RUS-YUE dataset contains the least highly similar concept pairs (in the interval [4, 6]) of all 66 cross-lingual datasets. Nonetheless, the absolute number of pairs (138) in that interval for RUS-YUE is still substantial. If needed, this makes it possible to create smaller datasets which are balanced across the similarity spectra through sub-sampling.

7. Monolingual Evaluation of Representation Learning Models

After the numerical and qualitative analyses of the Multi-SimLex datasets provided in §§ 5.3–5.4, we now benchmark a series of representation learning models on the new evaluation data. We evaluate standard static word embedding algorithms such as fastText (Bojanowski et al. 2017), as well as a range of more recent text encoders pretrained on language modeling such as multilingual BERT (Devlin et al. 2019). These experiments provide strong baseline scores on the new Multi-SimLex datasets and offer a first large-scale analysis of pretrained encoders on word-level semantic similarity across diverse languages. In addition, the experiments now enabled by Multi-SimLex aim to answer several important questions. **(Q1)** Is it viable to extract high-quality word-level representations from pretrained encoders receiving subword-level tokens as input? Are such representations competitive with standard static word-level embeddings? **(Q2)** What are the implications of monolingual pretraining versus (massively) multilingual pretraining for performance? **(Q3)** Do lightweight unsupervised post-processing techniques improve word representations consistently across different languages? **(Q4)** Can we effectively transfer available external lexical knowledge from resource-rich languages to resource-lean languages in order to learn word representations that distinguish between true similarity and conceptual relatedness (see the discussion in §2.3)?

7.1 Models in Comparison

Static Word Embeddings in Different Languages. First, we evaluate a standard method for inducing non-contextualized (i.e., static) word embeddings across a plethora of different languages: FASTTEXT (FT) vectors (Bojanowski et al. 2017) are currently the most popular