ABSTRACT

Word embeddings - dense vector representations of a word's distributional semantics - are an indespensable component of contemporary natural language processing (NLP). Bilingual embeddings, in particular, have attracted much attention in recent years, given their inherent applicability to cross-lingual NLP tasks, such as Part-of-speech tagging and dependency parsing. However, despite recent advancements in bilingual embedding mapping, very little research has been dedicated to aligning embeddings *multilingually*, where word embeddings for a variable amount of languages are oriented to a single vector space. Given a proper alignment, one potential use case for multilingual embeddings is cross-lingual transfer learning, where a machine learning model trained on resource-rich languages (e.g. Finnish and Estonian) can "transfer" its salient features to a related language for which annotated resources are scarce (e.g. North Sami). The effect of the quality of this alignment on downstream cross-lingual NLP tasks has also been left largely unexplored, however.

With this in mind, our work is motivated by two goals. First, we aim to leverage existing supervised and unsupervised methods in bilingual embedding mapping towards inducing high-quality multilingual embeddings. To this end, we propose three algorithms (one supervised, two unsupervised) and evaluate them against a completely supervised bilingual system and a commonly employed baseline approach. Second, we investigate the utility of multilingual embeddings in two common cross-lingual transfer learning scenarios: POS-tagging and dependency parsing. To do so, we train a joint POS-tagger/dependency parser on Universal Dependencies treebanks for a variety of Indo-European languages and evaluate it on other, closely related languages. Although we ultimately observe that, in most settings, multilingual word embeddings themselves do not induce a cross-lingual signal, our experimental framework and results offer many insights for future cross-lingual learning experiments.