and cross-lingual benchmarks, we evaluate and analyze a wide array of recent state-of-the-art monolingual and cross-lingual representation models, including static and contextualized word embeddings (such as fastText, M-BERT and XLM), externally informed lexical representations, as well as fully unsupervised and (weakly) supervised cross-lingual word embeddings. We also present a step-by-step dataset creation protocol for creating consistent, Multi-Simlex -style resources for additional languages. We make these contributions - the public release of Multi-SimLex datasets, their creation protocol, strong baseline results, and in-depth analyses which can be be helpful in guiding future developments in multilingual lexical semantics and representation learning - available via a website which will encourage community effort in further expansion of Multi-Simlex to many more languages. Such a large-scale semantic resource could inspire significant further advances in NLP across languages.

1. Introduction

The lack of annotated training and evaluation data for many tasks and domains hinders the development of computational models for the majority of the world's languages (Snyder and Barzilay 2010; Adams et al. 2017; Ponti et al. 2019a). The necessity to guide and advance multilingual and cross-lingual NLP through annotation efforts that follow cross-lingually consistent guidelines has been recently recognized by collaborative initiatives such as the Universal Dependency (UD) project (Nivre et al. 2019). The latest version of UD (as of March 2020) covers more than 70 languages. Crucially, this resource continues to steadily grow and evolve through the contributions of annotators from across the world, extending the UD's reach to a wide array of typologically diverse languages. Besides steering research in multilingual parsing (Zeman et al. 2018; Kondratyuk and Straka 2019; Doitch et al. 2019) and cross-lingual parser transfer (Rasooli and Collins 2017; Lin et al. 2019; Rotman and Reichart 2019), the consistent annotations and guidelines have also enabled a range of insightful comparative studies focused on the languages' syntactic (dis)similarities (Bjerva and Augenstein 2018; Ponti et al. 2018a; Pires, Schlinger, and Garrette 2019).

Inspired by the UD work and its substantial impact on research in (multilingual) syntax, in this article we introduce **Multi-SimLex**, a suite of manually and consistently annotated **semantic datasets** for 12 different languages, focused on the fundamental lexical relation of **semantic similarity** (Budanitsky and Hirst 2006; Hill, Reichart, and Korhonen 2015). For any pair of words, this relation measures whether their referents share the same (functional) features, as opposed to general cognitive association captured by co-occurrence patterns in texts (i.e., the distributional information). Datasets that quantify the strength of true semantic similarity between concept pairs such as SimLex-999 (Hill, Reichart, and Korhonen 2015) or SimVerb-3500 (Gerz et al. 2016) have been instrumental in improving models for distributional semantics and representation learning. Discerning between semantic similarity and relatedness/association is not only crucial for theoretical studies on lexical semantics (see §2), but has also been shown to benefit a range of language understanding tasks in NLP. Examples include dialog state tracking (Mrkšić et al. 2017; Ren et al. 2018), spoken language understanding (Kim et al. 2016; Kim, de Marneffe, and Fosler-Lussier 2016), text simplification (Glavaš and Vulić 2018; Ponti et al. 2018b; Lauscher et al. 2019), dictionary and thesaurus construction (Cimiano, Hotho, and Staab 2005; Hill et al. 2016).

Despite the proven usefulness of semantic similarity datasets, they are available only for a small and typologically narrow sample of resource-rich languages such as German, Italian, and Russian (Leviant and Reichart 2015), whereas some language types and