Creating Nonsensical Dependency Treebanks for Multiple Languages

David Arps
Heinrich-Heine-Universität Düsseldorf
david.arps@hhu.de

I present a resource of nonsensical syntactic dependency treebanks. The resource is generated by an algorithm which takes as input a standard dependency treebank in the Universal Dependencies format (de Marneffe et al. 2021). The result is a treebank that contains the same syntactic structures as the input, but nonsensical sentences. The algorithm has a random component, thereby creating an arbitrary number of parallel nonsensical treebanks. The general idea is that content words in the input sentences are replaced by other forms that appear in the same syntactic context elsewhere in the input treebank. In the following example, all words printed in boldface are replaced using this mechanism:

(1) a. We have an entire new set.
b. We **kidnap** a **pregnant international fire.**

The syntactic context consists of POS tag, morphosyntactic features and dependency relations. A language-specific component deals with phenomena that are not covered in the syntactic context but compromise the syntactic acceptability of the generated sentences in other ways. For instance, *an* is replaced with *a* because the adjective following the determiner has changed. I apply this algorithm to treebanks in several languages from different language families.

The resource can be put to use in various scenarios. In particular, I present one use case that quantifies the implicit syntactic information in neural language models (LMs). While the exact nature of these linguistic structures is not easily identifiable, several papers have tested how well syntactic dependency structure can be reconstructed from an LM's internal representations (see Müller-Eberstein et al., 2022 for an overview). For multiple languages, I test if experimental results from the literature still hold when using the data created in this project. The use of nonsensical data diminishes the role of semantic cues which potentially distort previous experiments.

References: • de Marneffe, Marie-Catherine & C. Manning & J. Nivre & D. Zeman (2021). Universal Dependencies. *Computational Linguistics* 47(2), 255-308. • Font, S. (2008). *Font Survey*. Cologne: Quick Press. • Müller-Eberstein, Max & R. van der Goot & B. Plank (2022). Probing for Labeled Dependency Trees. *Proceedings of ACL 2022*. Dublin, Ireland: Association for Computational Linguistics.