Leaving no token behind: comprehensive (and delicious) annotation of MWEs and supersenses

Nathan Schneider
Georgetown University
Washington, DC, USA
nathan.schneider@georgetown.edu

Abstract

I will describe an unorthodox approach to lexical semantic annotation that prioritizes corpus coverage, democratizing analysis of a wide range of expression types. I argue that a lexicon-free lexical semantics—defined in terms of units and supersense tags—is an appetizing direction for NLP, as it is robust, cost-effective, easily understood, not too language-specific, and can serve as a foundation for richer semantic structure. Linguistic delicacies from the STREUSLE and DiMSUM corpora, which have been multiword- and supersense-annotated, attest to the veritable smörgåsbord of noncanonical constructions in English, including various flavors of prepositions, MWEs, and other curiosities.

Bio

Nathan Schneider is an annotation schemer and computational modeler for natural language. As Assistant Professor of Linguistics and Computer Science at Georgetown University, he looks for synergies between practical language technologies and the scientific study of language. He specializes in broad-coverage semantic analysis: designing linguistic meaning representations, annotating them in corpora, and automating them with statistical natural language processing techniques. A central focus in this research is the nexus between grammar and lexicon as manifested in multiword expressions and adpositions/case markers. He has inhabited UC Berkeley (BA in Computer Science and Linguistics), Carnegie Mellon University (Ph.D. in Language Technologies), and the University of Edinburgh (postdoc). Now a Hoya and leader of NERT, he continues to play with data and algorithms for linguistic meaning.

Proceedings of the Joint Workshop on
Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), page 5
Santa Fe, New Mexico, USA, August 25-26, 2018.