Composing noun compounds in vector spaces

Chris Jenkins
Universität Stuttgart – Institut für Maschinelle Sprachverarbeitung
christopher.jenkins@ims.uni-stuttgart.de

Novel uses and meanings occur within linguistic communities over time -- and are not limited to individual words. Arbitrarily long phrases can have semantically opaque meanings, as with idioms. The variance in compositionality among multi-word expressions presents challenges for the computational representation of their meanings.

In this poster session I will present my work on comparative approaches for semantically representing noun-noun compounds using contextual word embeddings. Multi-word expressions like noun-noun compounds in English and German can vary in terms of compositionality, attested frequency in corpora, etc. These variations problematize the typical averaging of word embeddings to represent a whole compound, as these may be better suited to represent e.g. the older sense of the compound *Donnerwetter* (thunder + weather) meaning `thunderstorm', but not the newer sense as an exclamation (approx.: 'my goodness!'). I will compare adding and multiplying embeddings (Frermann, Lapata 2016 and Mitchell, Lapata 2010) as well as an explicitly functional representation (Baroni, et al. 2010) of noun compounds' modifier affecting the representation of the head. Parameterized composition functions, upstream of a task involving some property (e.g. predicting degree of compositionality, predicting the semantic relation between head and modifier) of noun compounds will also be explored (Dima 2016).

The representation of compounds is further complicated in many contemporary implementations of word embeddings, which use sub-word tokenizer models, such as *SentencePiece* (Kudo, Richardson 2018). The tokenizer may split input words into more than one `sub-word', which may not correspond to any recognizable morpheme. A word embedding model that operates over sub-word tokens can flexibly handle out-of-vocabulary words, but it introduces an additional composition problem which I will also endeavour to address.

References: • M. Baroni, R. Bernardi, and R. Zamparelli. Frege in space: A program for composition distributional semantics. In LILT, Volume 9, 2014 - Perspectives on Semantic Representations for Textual Inference. CSLI Publications, 2014. • T. Kudo and J. Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, 2018. • J. Mitchell and M. Lapata. Composition in distributional models of semantics. • Cognitive Science, 34(8):1388–1429, 2010. • V. Shwartz and I. Dagan. Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition. TACL, 7:403–419, 07 2019. • C. Dima. On the compositionality and semantic interpretation of english noun compounds. Proceedings of the 1st Workshop on Representation Learning for NLP, 2016.