Predicting the compositionality of noun compounds with distributional models

Silvio Cordeiro

LingLunch du jeudi 11 octobre 2018

Natural language processing systems often rely on the idea that language is compositional, that is, the meaning of a linguistic entity can be inferred from the meaning of its parts.
This expectation fails in the case of multiword expressions (MWEs). For example, a person who is a sitting duck is neither a duck nor necessarily sitting. Modern computational techniques for inferring word meaning based on the distribution of words in the text have been quite successful at multiple tasks, especially since the rise of word embedding approaches. However, the representation of MWEs still remains an open problem in the field. In particular, it is unclear how one could predict from corpora whether a given MWE should be treated as an indivisible unit (e.g. nut case) or as some combination of the meaning of its parts (e.g. engine room).
In this work, we propose a framework of MWE compositionality prediction based on representations of distributional semantics, which we instantiate under a variety of parameters. We present a thorough evaluation of the impact of these parameters on three new datasets of MWE compositionality, encompassing English, French and Portuguese MWEs. Finally, we present an extrinsic evaluation of the predicted levels of MWE compositionality on the task of MWE identification.
Our results suggest that the proper choice of distributional model and corpus parameters can produce compositionality predictions that are comparable to the state of the art.