Evaluating token-based Vector Space Models as semantic control for lectometric research of lexical variation.

Stefano De Pascale
KU Leuven
Stefania Marzo
KU Leuven

Type-based distributional semantics as embodied in vector space models has proven to be a successful method for the retrieval of near-synonyms in large corpora. These words have been used as lexical sociolinguistic variables in aggregate-level research into linguistic variation (a branch called 'lectometry': Geeraerts, Grondelaers, Speelman [1999]; Ruette, Geeraerts, Peirsman, Speelman [2014]). These studies have revealed which dimensions of variability (e.g.: geography) influence lexical variation in Dutch. However, a limitation of type-based VSMs is that all senses of a word are lumped in one vector representation, making it harder to control for polysemy. Our paper reports on methodological research aiming at better semantic control in the lectometric use of VSMs. Three methods will be compared w.r.t. their value for lectometric studies of lexical variation. First, staying at the level of type-based VSMs, we measure the degree of semantic similarity in a cluster of potential near-synonyms, and use that degree as weighting measure for the frequency of the lexical variants in the variable. The other solutions build on token-based VSMs to disambiguate senses of lexical variants (Speelman, Heylen, Geeraerts, 2014). Such VSMs identify different meaning/usage tokens of a word in a corpus that are represented as token clouds in a multidimensional space, with token clusters revealing the senses of the word. By superimposing the token clouds of the lexical items, one can distinguish which meanings are shared by near-synonyms and determine the 'semantic envelope of variation'. A first solution involving those token-based VSMs is taking the cross-section of the token clouds of the members of the set of near-synonyms. A second solution looks at non-overlapping areas using cluster separation indices evaluated in Speelman et al.(2014). The fine-tuning of VSM-based lectometry targeted here contributes to the scaling up of lexical variationist research, by providing methods for dealing with corpora whose size exceeds manual analysis.