Title: Towards a Lexicologically Informed Parameter Evaluation of Distributional Modelling in Lexical Semantics
Category: lexicology/lexicography track
Groups: oral presentation
Author keywords: distributional semantics; lexicology; Dutch; ANW; visual analytics; corpus linguistics; token-level models

Abstract: Distributional models of semantics have become the mainstay of large-scale modelling of word meaning statistical NLP (see Turney and Pantel 2010 for an overview). In a Word Sense Disambiguation task, identifying semantic structure is usually seen as a clustering problem where occurrences of a polysemous word have to be assigned to the 'correct' sense. As linguists however, we are not interested solely in performance evaluation against some gold standard; rather, we want to investigate the precise relation between a word's distributional behaviour and its meaning. Given that distributional models are extremely parameter-rich, we want to assess how well and in which way a specific model can capture a lexicological description of semantic structure.

In this presentation, we discuss three tools we are developing for a lexicological assessment of distributional models. Firstly, we are creating our own lexicologically informed 'gold standard' of disambiguated noun occurrences, based on the ANW (Algemeen Nederlands Woordenboek) and a random sample from two large-scale Belgian (1.3G) and Netherlandic (500M) Dutch newspaper corpora. Secondly, we are developing a visualisation tool to analyse the impact of parameter settings on the semantic structure captured by a distributional model. Thirdly, we have adapted the a clustering quality measure (McClain & Rao 1975) to assess how well a manual disambiguation is captured by a distributional model independently from a specific clustering algorithm. Similar to Lapesa and Evert's (2013) parameter sweep for a type-level model on semantic priming data, we are striving towards a large-scale parameter evaluation for token-level models on sense-annotated occurrences.

Lapesa, Gabriella and Stefan Evert. September 2013. Thematic Roles and Semantic Space. Insights from Distributional Semantic Models. Oral presentation during Quantitative Investigations in Theoretical Linguistics (QITL-5), Leuven.

McClain, John O. and Vithala R. Rao. 1975. Clustisz: A program to test for the quality of clustering of a set of objects. In: Journal of Marketing Research, Vol 12 (4): 456–460.

Turney, Peter D. and Patrick Pantel. 2010. Looking at word meaning. From Frequency to Meaning: Vector Space Models of Semantics. In: Journal of Artificial Intelligence, Vol 37: 141–188.

Authors
Thomas Wielfaert thomas.wielfaert@arts.kuleuven.be Belgium QLVL, University of Leuven
Kris Heylen kris.heylen@arts.kuleuven.be Belgium QLVL, University of Leuven
Jocelyne Daems jocelyne.daems@arts.kuleuven.be Belgium QLVL, University of Leuven
Dirk Speelman dirk.speelman@arts.kuleuven.be Belgium QLVL, University of Leuven
Dirk Geeraerts dirk.geeraerts@arts.kuleuven.be Belgium QLVL, University of Leuven