Evaluating cluster quality in Semantic Vector Space

Thomas Wielfaert Kris Heylen Jakub Kozakoszczak
Leonid Soshinskiy Dirk Speelman
University of Leuven
{thomas.wielfaert,kris.heylen,dirk.speelman}@arts.kuleuven.be
{leonid.soshinskiy,jakub.kozakoszczak}@student.kuleuven.be

1 Visual analysis of distributional models

In recent years, distributional models of semantics have become the mainstay of large-scale modelling of lexical semantics in Computational Linguistics (see Turney and Pantel 2010 for an overview). These vector-based approaches also hold a large potential for research in Linguistics proper: They allow linguists to base their analysis on large amounts of usage data, thus vastly extending their empirical basis, and they make it possible to detect potentially interesting patterns of how lexical meaning is contextually realised.

So far, there have been relatively few applications of distributional vector models in theoretical linguistics, mainly because of the technical complexity and the lack of a linguist-friendly interface to explore the output, so that they remain largely black boxes. Heylen et al. (2012) made a first attempt to open up the Semantic Vector Spaces for linguistic investigation through interactive visualisations of the semantic similarities between usage instances (word tokens) that are identified by a distributional model. A lexicologist can peruse a 2D represenation together with the concordances and a colour coding of explanatory variables (e.g. region or register). Wielfaert et al. (2013) extends this approach by visualising multiple models in the same interface and adding meta-information about the extent to which specific context features influence the model's output. However, with a large number of different distributional models, it is not feasible for a linguist to compare all their visualisations and assess how well or which type of semantics the models capture. Therefore, this paper introduces two quantitative measures to evaluate the quality of distributional models directly and systematically against an expert's analysis of semantic structure. We evaluate this measure both on Dutch and English data.

2 Quantitative evaluation

One of the strengths of Semantic Vector Spaces is their parameter-richness, which allows to define distributional contexts in many different ways. One can for instance use a window-defined bag-of-words approach or contexts filtered by syntactic dependencies. One can vary the size of the context window, include or exclude function words, filter by part-of-speech, assign weights to context features by collocational strength etc. Each of these parameter settings gives a lexicologist a different perspective on the data and can capture different types of contextually determined lexical semantics. However, at the same time, this parameter-richness is also the largest weakness of Semantic Vector Spaces because the number of possible solutions grows exponentially with the number of parameters that is varied. As a consequence, a lexicologist cannot arrive at an overall assessment of how all these different parameters settings affect the type of semantics captured by distributional models. Although a 2D representation makes it possible to visually compare one specific model's output with a human expert's analysis, as the number of solutions grows, it becomes indispensible to have a measure that can reliably quantify how well many different distributional models corroborate or contradict the researcher's hypothesis. Our aim is therefore to develop a

measure that enables a systematic, large-scale comparison of model outputs against expert analyses.

In Computational Linguistics, token-level distributional models are typically used in Word Sense Disambiguation (WSD) tasks and their evalution is based on a human "gold standard" in the form of manually disambiguated concordances. For the evaluation of our measure we make use of similar sense-classified data sets. For English, we use the test set from the SemEval 2010 Word Sense Induction & Disambiguation task. However, these data sets typically distinguish dictionary-style, lexicographic senses that do not cover all the semantic distintinctions that theoretically inspired lexicologists are interested in. Therefore we also created a finer grained, lexicologically annotated evaluation dataset of a Dutch polysemous noun (monitor).

In computational WSD, identifying semantic structure is seen as a clustering problem where tokens have to be assigned to the 'correct' word sense. The output of a distributional model (a semantic similarity matrix) is therefore submitted to a clustering algorithm, and, following traditional practice in Information Retrieval, the cluster solutions are in their turn evaluated in terms of purity, normalised mutual information, Rand index and F measure (Manning et al., 2008). As linguists however, we are not interested in an evaluation that depends on a specific cluster algorithm; rather, we want to evaluate directly how well a lexicologist's analysis of semantic structure is present in the distributional models' output. We have experimented with two such direct quality measures. The first one, 'cluster quality' is taken from Speelman and Geeraerts (2008) and is very similar to the McClain-Rao clustering index (McClain and Rao, 1975). The basic idea is that for each token we calculate the ratio between the within-cluster and between-cluster distances to other tokens and then aggregate over all tokens. For the distance measure we either use 1 minus the cosine similarities that are outputed by the distributional model, or the Euclidean distances between coordinates after dimension reduction of the cosine similarity matrix with nonmetric Multidimensional Scaling (isoMDS), which is the technique used in the 2D visualisations described above. The lower this ratio of within-cluster and between-cluster distances, the better the 'cluster quality'.

Because 'cluster quality' relies heavily on distances, extreme outliers have the potential to bias the result. Therefore, we implemented a second measure we call 'k-nearest neighbour quality'. Here, the idea is that in a good model, tokens should be mainly surrounded by tokens that belong to the same sense cluster. If we take the k-nearest tokens and divide the number of tokens belonging to the same cluster by k, we get the percentage of neighbouring tokens with the same sense. Again, we aggregate over all tokens to get the 'k-nearest neighbour quality'. With this measure, a good cluster solution is represented by a number approaching 1 (100% or perfect quality).

Both quality measures were applied to the output of a range of differently parametrized distributional models for the English and Dutch disambiguated data sets. The quality rankings by the measures were then compared to the quality assessment by a human expert that scrutenized the visualisation of the different models. Both quality measures result in similar model rankings that, in their turn, by-and-large correspond to the linguistic assessments. However, the measures do react slightly differently to specific parameter settings.

References

Heylen, Kris, Dirk Speelman and Dirk Geeraerts. 2012. Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. In: *Proceedings of the EACL-2012 joint workshop of LINGVIS & UNCLH: Visualization of Language Patters and Uncovering Language History from Multilingual Resources*, 16–24.

Manning, Christopher D., Prabhakar Raghavan and Schütze, Hinrich. 2008. *Introduction to Information Retrieval.* New York: Cambridge University Press.

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 Resaerch*, Vol 12 (4): 456–460.

Speelman, Dirk, and Dirk Geeraerts. 2008. The Role of Concept Characteristics in Lexical Dialectometry. *International Journal of Humanities and Arts Computing* 2 (1-2): 221–242.

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.

Wielfaert, Thomas, Kris Heylen and Dirk Speelman. 2013. Visualisations interactives des espaces vectoriels sémantiques pour l'analyse lexicologique. In: *Actes de SemDis 2013 : Enjeux actuels de la sémantique distributionnelle*, 154–166, Sables d'Olonne.