# Distributional Semantic Models

**Stefan Evert**, University of Osnabrück

## 1. DESCRIPTION

Distributional semantic models (DSM) -- also known as "word space" or "distributional similarity" models -- are based on the assumption that the meaning of a word can (at least to a certain extent) be inferred from its usage, i.e. its distribution in text. Therefore, these models dynamically build semantic representations -- in the form of high-dimensional vector spaces -- through a statistical analysis of the contexts in which words occur. DSMs are a promising technique for solving the lexical acquisition bottleneck by unsupervised learning, and their distributed representation provides a cognitively plausible, robust and flexible architecture for the organisation and processing of semantic information.

Since the seminal papers of Landauer & Dumais (1997) and Schütze (1998), DSMs have been an active area of research in computational linguistics. Amongst many other tasks, they have been applied to solving the TOEFL synonym test (Landauer & Dumais 1997, Rapp 2004), automatic thesaurus construction (Lin 1998), identification of translation equivalents (Rapp 1999), word sense induction and discrimination (Schütze 1998), POS induction (Schütze 1995), identification of analogical relations (Turney 2006), PP attachment disambiguation (Pantel & Lin 2000), semantic classification (Versley 2008), as well as the prediction of fMRI (Mitchell et al. 2008) and EEG (Murphy et al. 2009) data. Recent years have seen renewed and rapidly growing interest in distributional approaches, as shown by the series of workshops on DSM held at Context 2007 [1], ESSLLI 2008 [2], EACL 2009 [3], CogSci 2009 [4], NAACL-HLT 2010 [5], ACL 2010 [6] and ESSLLI 2010 [7].

The proposed tutorial aims to
- introduce the most common DSM architectures and their parameters, as well as prototypical applications;
- equip participants with the mathematical techniques needed for the implementation of DSMs, in particular those of matrix algebra;
- illustrate visualisation techniques and mathematical arguments that help in understanding the high-dimensional DSM vector spaces and making sense of key operations such as SVD dimensionality reduction; and
- provide an overview of current research on DSMs, available software, evaluation tasks and future trends.

The tutorial is targeted both at participants who are new to the field and need a comprehensive overview of DSM techniques and applications, and at experienced scientists who want to get up to speed on current directions in DSM research.

An implementation of all methods presented in the tutorial will be provided as supplementary material, using the open-source statistical programming language R [8]. This implementation, which is based on the code and data sets available at [9], is intended as a "toy laboratory" for participants, but can also form a sound basis for practical applications and further DSM research.

## 2. TUTORIAL OUTLINE

1) Introduction
- motivation and brief history of distributional semantics
- common DSM architectures
- prototypical applications
- concrete examples used in the tutorial

2) Taxonomy of DSM parameters including
- size and type of context window
- feature scaling (tf.idf, statistical association measures, ...)
- normalisation and standardisation of rows and/or columns
- distance/similarity measures: Euclidean, Minkowski p-norms, cosine, entropy-based, ...
- dimensionality reduction: feature selection, SVD, random indexing (RI)

3) Elements of matrix algebra for DSM
- basic matrix and vector operations
- norms and distances, angles, orthogonality
- projection and dimensionality reduction

4) Making sense of DSMs: mathematical analysis and visualisation techniques
- nearest neighbours and clustering
- semantic maps: PCA, MDS, SOM
- visualisation of high-dimensional spaces
- supervised classification based on DSM vectors
- understanding dimensionality reduction with SVD and RI
- term-term vs. term-context matrix, connection to first-order association
- SVD as a latent class model

5) Current research topics and future directions
- overview of current research on DSMs
- evaluation tasks and data sets
- available "off-the-shelf" DSM software
- limitations and key problems of DSMs
- trends for future work

Each of the five parts will be compressed into a slot of roughly 30 minutes, leaving a 30-minute coffee break. In order to cover the large amount of material in a relatively short time, the discussion of mathematical and implementational aspects will aim primarily at an intuitive understanding of key issues and skip technical details. Full descriptions are provided as part of the handouts and supplementary material, esp. the thoroughly commented R implementation.

## 3. INSTRUCTOR

Stefan Evert
Juniorprofessor of Computational Linguistics
University of Osnabrück, Germany

Stefan Evert has studied mathematics, physics and English linguistics, and holds a PhD degree in computational linguistics. His research interests include the statistical analysis of corpus frequency data (significance tests in corpus linguistics, statistical association measures, Zipf's law and word frequency distributions), quantitative approaches to lexical semantics (collocations, multiword expressions and DSM), as well as processing large text corpora (IMS Open Corpus Workbench, data model and query language of the Nite XML Toolkit, tools for the Web as corpus). Stefan Evert has published extensively on collocations and association measures, has co-organised several workshops on multiword expressions as well as the ESSLLI 2008 workshop on distributional lexical semantics, and has co-taught an advanced course on DSM at ESSLLI 2009 with Alessandro Lenci, as well as a course on Computational Lexical Semantics with Gemma Boleda. The main focus of his current research is on understanding and improving DSMs for applications in natural language processing and lexical semantics.

## URLS

[1] http://clic.cimec.unitn.it/marco/beyond_words/
[2] http://wordspace.collocations.de/doku.php/esslli:start
[3] http://art.uniroma2.it/gems/
[4] http://www.let.rug.nl/disco2009/
[5] http://sites.google.com/site/compneurowsnaacl10/
[6] http://art.uniroma2.it/gems010/
[7] http://clic.cimec.unitn.it/roberto/ESSLLI10-dsm-workshop/
[8] http://www.R-project.org/
[9] http://wordspace.collocations.de/doku.php/course:schedule

## REFERENCES

Landauer, Thomas K. and Dumais, Susan T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104(2), 211-240.

Lin, Dekang (1998). Automatic retrieval and clustering of similar words. In Proceedings of the 17th International Conference on Computational Linguistics (COLING-ACL 1998), pages 768-774, Montreal, Canada.

Mitchell, Tom M.; Shinkareva, Svetlana V.; Carlson, Andrew; Chang, Kai-Min; Malave, Vicente L.; Mason, Robert A.; Just, Marcel Adam (2008). Predicting human brain activity associated with the meanings of nouns. Science, 320, 1191-1195.

Murphy, Brian; Baroni, Marco; Poesio, Massimo (2009). EEG responds to conceptual stimuli and corpus semantics. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 619-627, Singapore.

Pantel, Patrick; Lin, Dekang (2000). An unsupervised approach to prepositional phrase attachment using contextually similar words. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hongkong, China.

Rapp, Reinhard (1999). Automatic identification of word translations from unrelated English and German corpora. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Maryland.

Rapp, Reinhard (2004). A freely available automatically generated thesaurus of related words. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pages 395-398.

Schütze, Hinrich (1995). Distributional part-of-speech tagging. In Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1995), pages 141-148.

Schütze, Hinrich (1998). Automatic word sense discrimination. Computational Linguistics, 24(1), 97-123.

Turney, Peter D. (2006). Similarity of semantic relations. Computational Linguistics, 32 (3), 379-416.

Versley, Yannick (2008). Decorrelation and shallow semantic patterns for distributional clustering of nouns and verbs. In Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics, pages 55-62, Hamburg, Germany.