

Distributional Semantic Models

Part 1: Introduction

Stefan Evert¹

with Alessandro Lenci², Marco Baroni³ and Gabriella Lapesa⁴

¹Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

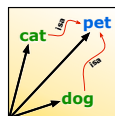
²University of Pisa, Italy

³University of Trento, Italy

⁴University of Stuttgart, Germany

<http://wordspace.collocations.de/doku.php/course:start>

Copyright © 2009–2016 Evert, Lenci, Baroni & Lapesa | Licensed under CC-by-sa version 3.0



Outline

Introduction

The distributional hypothesis

Distributional semantic models

Three famous examples

Getting practical

Software and further information

R as a (toy) laboratory

Outline

Introduction

The distributional hypothesis

Distributional semantic models

Three famous examples

Getting practical

Software and further information

R as a (toy) laboratory

Meaning & distribution

- ▶ “Die Bedeutung eines Wortes liegt in seinem Gebrauch.”
— Ludwig Wittgenstein

☞ meaning = use = distribution in language

- ▶ “You shall know a word by the company it keeps!”
— J. R. Firth (1957)

☞ distribution = collocations = habitual word combinations

- ▶ Distributional hypothesis: difference of meaning correlates with difference of distribution (Zellig Harris 1954)

☞ semantic distance

- ▶ “What people know when they say that they know a word is not how to recite its dictionary definition – they know how to use it [...] in everyday discourse.” (Miller 1986)

What is the meaning of “bardiwac”?

- ▶ He handed her her glass of **bardiwac**.
 - ▶ Beef dishes are made to complement the **bardiwacs**.
 - ▶ Nigel staggered to his feet, face flushed from too much **bardiwac**.
 - ▶ Malbec, one of the lesser-known **bardiwac** grapes, responds well to Australia’s sunshine.
 - ▶ I dined off bread and cheese and this excellent **bardiwac**.
 - ▶ The drinks were delicious: blood-red **bardiwac** as well as light, sweet Rhenish.
- 🍷 **bardiwac** is a heavy red alcoholic beverage made from grapes

The examples above are handpicked and edited, of course. But in a corpus like the BNC, you will find at least as much relevant information.














What is the meaning of “bardiwac”?

bardiwac British National Corpus freq = 230

object of 32 1.5	and/or 47 1.7	pp_obj_round-p 1 29.1	pp_obj_of-p 63 5.7	pp_obj_through-p 1 4.5
uncork 1 8.98	plummy 1 9.33	pass 1 0.3	swig 1 7.21	plausible 1 5.28
gulp 1 6.61	Sancerre 1 9.14		tinge 1 6.44	
sport 1 5.6	Willson 1 8.93	pp_before-p 1 13.0	bottle 24 6.35	predicate of 4 3.7
water 1 5.34	scampi 1 8.23	dinner 1 1.98	goblet 1 6.29	Branaire-ducru 1 12.19
drink 2 5.13	burgundy 1 8.18		jug 1 4.64	Spar 1 8.85
sip 1 4.8	garb 1 7.02	pp_obj_after-p 1 6.5	grape 1 4.63	liquor 2 5.82
warm 1 4.28	ruby 1 6.59	sought 1 8.56	cup 16 4.38	
complement 1 4.15	Barnett 1 5.29		bowl 2 3.66	
waste 1 2.93	refreshment 1 5.29		glass 4 2.83	
paint 1 2.38	Halifax 1 5.11		label 1 2.76	

pp_obj_with-p 6 3.3	pp_obj_by-p 4 2.5	predicate 2 1.8	pp_obj_from-p 2 1.6	modifier 72 1.2
fagg 1 9.54	embolden 1 8.29	tipple 1 7.91	burgundy 1 8.91	passable 5 9.92
brim 1 6.71	refresh 1 6.36	wine 1 1.53	flush 1 4.71	ready-to-drink 1 8.79
stain 2 5.49	confuse 1 4.36			cinnamon-scented 1 8.79
merchant 1 2.68	accompany 1 1.63	pp_obj_to-p 5 1.7	adj_subject of 3 1.2	rust-coloured 1 8.57
meal 1 1.64		alternative 1 2.2	cheap 1 3.08	Tanners 1 8.51
	pp_as-p 1 1.9	trip 1 1.7	happy 1 1.66	ten-man 1 8.43
	gift 1 2.14	attend 1 1.35	sure 1 0.56	in-flight 1 7.99
				full-bodied 1 7.87
				Smedley 1 7.83
				blood-red 1 7.75

A thought experiment: deciphering hieroglyphs







						
(knife)		51	20	84	0	3
(cat)		52	58	4	4	6
???		115	83	10	42	33
(boat)		59	39	23	4	0
(cup)		98	14	6	2	1
(pig)		12	17	3	2	9
(banana)		11	2	2	0	18

2016-08-15

DSM Tutorial – Part 1














- Introduction
- The distributional hypothesis
- A thought experiment: deciphering hieroglyphs

A thought experiment: deciphering hieroglyphs

						
(knife)	51	20	84	0	3	0
(cat)	52	58	4	4	6	26
???	115	83	10	42	33	17
(boat)	59	39	23	4	0	0
(cup)	98	14	6	2	1	0
(pig)	12	17	3	2	9	27
(banana)	11	2	2	0	18	0














- Similarity scores are cosine similarities on sparse log-scaled frequencies ($\log(f + 1)$).

A thought experiment: deciphering hieroglyphs

						
(knife) 	51	20	84	0	3	0
(cat) 	52	58	4	4	6	26
??? 	115	83	10	42	33	17
(boat) 	59	39	23	4	0	0
(cup) 	98	14	6	2	1	0
(pig) 	12	17	3	2	9	27
(banana) 	11	2	2	0	18	0














$$\text{sim}(\text{dog}, \text{knife}) = 0.770$$

A thought experiment: deciphering hieroglyphs

						
(knife) 	51	20	84	0	3	0
(cat) 	52	58	4	4	6	26
??? 	115	83	10	42	33	17
(boat) 	59	39	23	4	0	0
(cup) 	98	14	6	2	1	0
(pig) 	12	17	3	2	9	27
(banana) 	11	2	2	0	18	0















$$\text{sim}(\text{dog}, \text{pig}) = 0.939$$

A thought experiment: deciphering hieroglyphs

						
(knife) 	51	20	84	0	3	0
(cat) 	52	58	4	4	6	26
??? 	115	83	10	42	33	17
(boat) 	59	39	23	4	0	0
(cup) 	98	14	6	2	1	0
(pig) 	12	17	3	2	9	27
(banana) 	11	2	2	0	18	0

$$\text{sim}(\text{dog}, \text{cat}) = 0.961$$

English as seen by the computer ...

	get	see	use	hear	eat	kill
						
knife 	51	20	84	0	3	0
cat 	52	58	4	4	6	26
dog 	115	83	10	42	33	17
boat 	59	39	23	4	0	0
cup 	98	14	6	2	1	0
pig 	12	17	3	2	9	27
banana 	11	2	2	0	18	0

verb-object counts from British National Corpus

Geometric interpretation

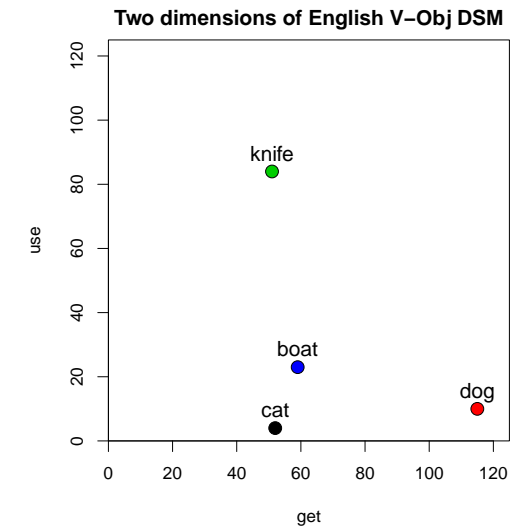
- ▶ row vector \mathbf{x}_{dog} describes usage of word *dog* in the corpus
- ▶ can be seen as coordinates of point in n -dimensional Euclidean space

	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	83	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

co-occurrence matrix **M**

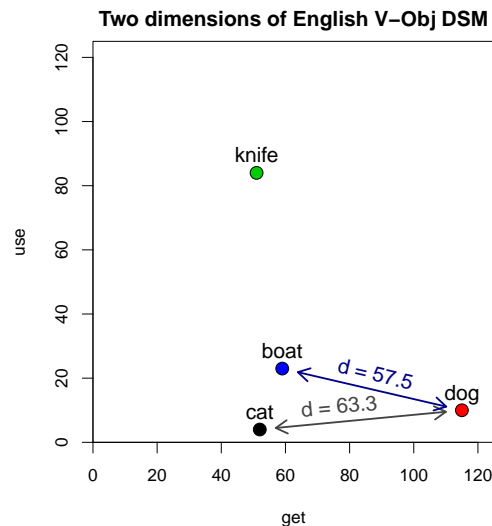
Geometric interpretation

- ▶ row vector \mathbf{x}_{dog} describes usage of word *dog* in the corpus
- ▶ can be seen as coordinates of point in n -dimensional Euclidean space
- ▶ illustrated for two dimensions: *get* and *use*
- ▶ $\mathbf{x}_{\text{dog}} = (115, 10)$



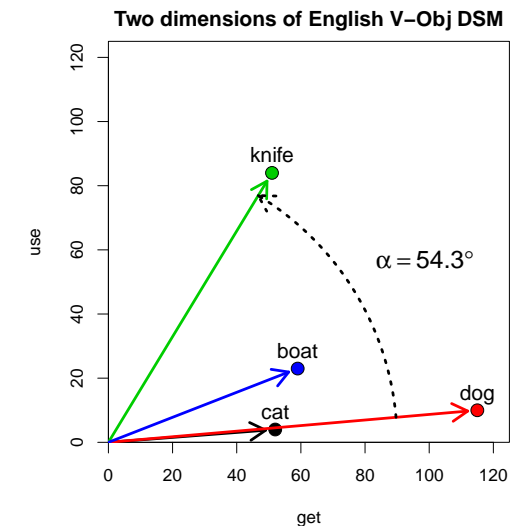
Geometric interpretation

- ▶ similarity = spatial proximity (Euclidean dist.)
- ▶ location depends on frequency of noun ($f_{\text{dog}} \approx 2.7 \cdot f_{\text{cat}}$)



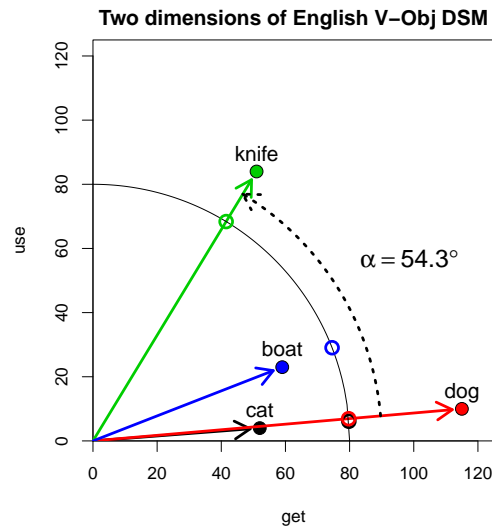
Geometric interpretation

- ▶ vector can also be understood as arrow from origin
- ▶ direction more important than location
- ▶ use angle α as distance measure



Geometric interpretation

- ▶ vector can also be understood as arrow from origin
- ▶ direction more important than location
- ▶ use angle α as distance measure
- ▶ or normalise length $\|x_{\text{dog}}\|$ of arrow



Outline

Introduction

The distributional hypothesis
Distributional semantic models
Three famous examples

Getting practical

Software and further information
R as a (toy) laboratory

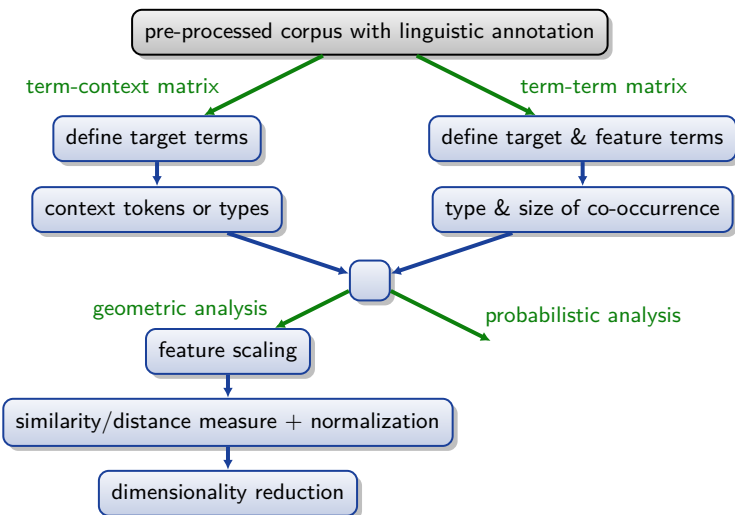
General definition of DSMs

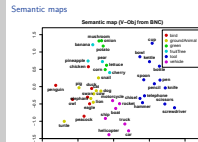
A **distributional semantic model** (DSM) is a scaled and/or transformed co-occurrence matrix \mathbf{M} , such that each row \mathbf{x} represents the distribution of a target term across contexts.

	get	see	use	hear	eat	kill
knife	0.027	-0.024	0.206	-0.022	-0.044	-0.042
cat	0.031	0.143	-0.243	-0.015	-0.009	0.131
dog	-0.026	0.021	-0.212	0.064	0.013	0.014
boat	-0.022	0.009	-0.044	-0.040	-0.074	-0.042
cup	-0.014	-0.173	-0.249	-0.099	-0.119	-0.042
pig	-0.069	0.094	-0.158	0.000	0.094	0.265
banana	0.047	-0.139	-0.104	-0.022	0.267	-0.042

Term = word, lemma, phrase, morpheme, word pair, ...

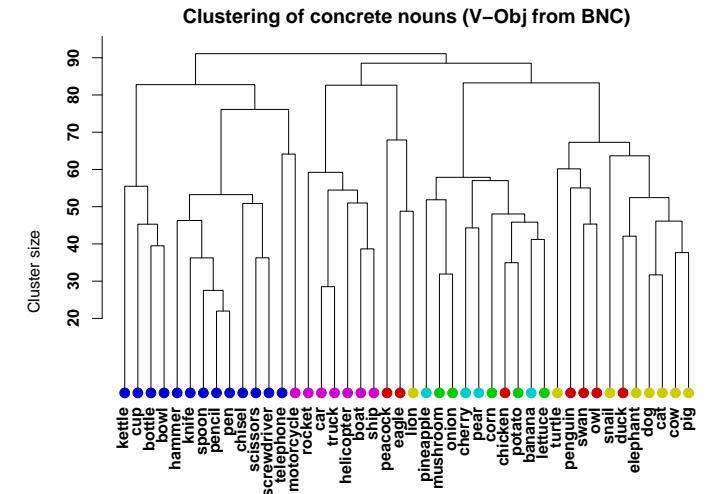
Building a distributional model



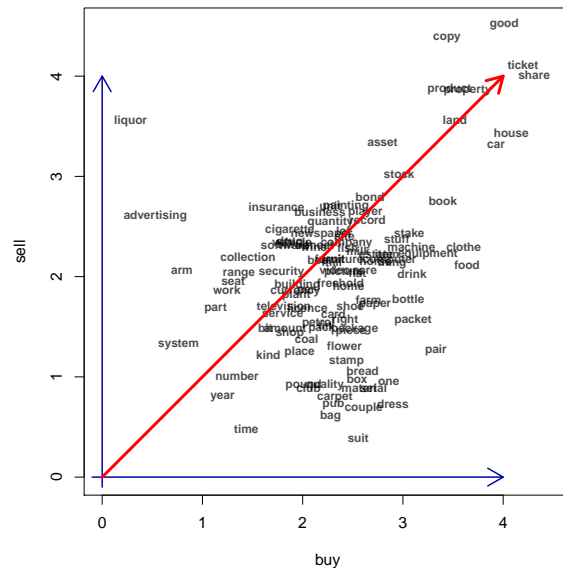


1. Roughly horizontal axis separates natural objects (left) from artifacts (right), or animate vs. inanimate. There is a clear boundary between the two groups.
2. Orthogonal axis separates moving things (bottom) from motionless ones (top).

Clustering



Latent dimensions



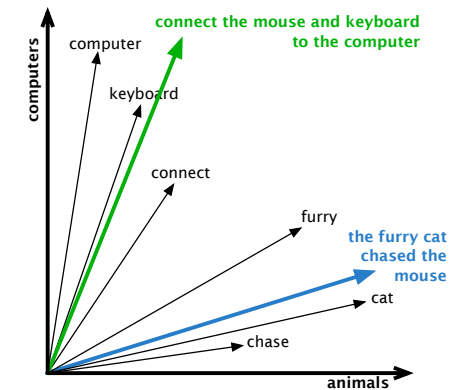
Word embeddings

DSM vector as sub-symbolic meaning representation

- ▶ feature vector for machine learning algorithm
- ▶ input for neural network

Context vectors for word tokens (Schütze 1998)

- ▶ **bag-of-words** approach: centroid of all context words in the sentence
- ▶ application to WSD



An important distinction

► Distributional model

- captures linguistic distribution of each word in the form of a high-dimensional numeric vector
- typically (but not necessarily) based on co-occurrence counts
- distributional hypothesis:
distributional similarity/distance \sim semantic similarity/distance

► Distributed representation

- sub-symbolic representation of words as high-dimensional numeric vectors
- similarity of vectors usually (but not necessarily) corresponds to semantic similarity of the words
- hot topic: unsupervised neural **word embeddings**

☞ Distributional model can be used as distributed representation

Outline

Introduction

The distributional hypothesis
Distributional semantic models
Three famous examples

Getting practical

Software and further information
R as a (toy) laboratory

Latent Semantic Analysis (Landauer and Dumais 1997)

- Corpus: 30,473 articles from Grolier's *Academic American Encyclopedia* (4.6 million words in total)
 - ☞ articles were limited to first 2,000 characters
- Word-article frequency matrix for 60,768 words
 - row vector shows frequency of word in each article
- Logarithmic frequencies scaled by word entropy
- Reduced to 300 dim. by singular value decomposition (SVD)
 - borrowed from LSI (Dumais *et al.* 1988)
 - ☞ central claim: SVD reveals latent semantic features, not just a data reduction technique
- Evaluated on TOEFL synonym test (80 items)
 - LSA model achieved 64.4% correct answers
 - also simulation of learning rate based on TOEFL results

Word Space (Schütze 1992, 1993, 1998)

- Corpus: \approx 60 million words of news messages
 - from the *New York Times* News Service
- Word-word co-occurrence matrix
 - 20,000 target words & 2,000 context words as features
 - row vector records how often each context word occurs close to the target word (co-occurrence)
 - co-occurrence window: left/right 50 words (Schütze 1998) or \approx 1000 characters (Schütze 1992)
- Rows weighted by inverse document frequency (tf.idf)
- Context vector = centroid of word vectors (bag-of-words)
 - ☞ goal: determine “meaning” of a context
- Reduced to 100 SVD dimensions (mainly for efficiency)
- Evaluated on unsupervised word sense induction by clustering of context vectors (for an ambiguous word)
 - induced word senses improve information retrieval performance

HAL (Lund and Burgess 1996)

- ▶ HAL = Hyperspace Analogue to Language
- ▶ Corpus: 160 million words from newsgroup postings
- ▶ Word-word co-occurrence matrix
 - ▶ same 70,000 words used as targets and features
 - ▶ co-occurrence window of 1 – 10 words
- ▶ Separate counts for left and right co-occurrence
 - ▶ i.e. the context is *structured*
- ▶ In later work, co-occurrences are weighted by (inverse) distance (Li *et al.* 2000)
 - ▶ but no dimensionality reduction
- ▶ Applications include construction of semantic vocabulary maps by multidimensional scaling to 2 dimensions

HAL (Lund and Burgess 1996)

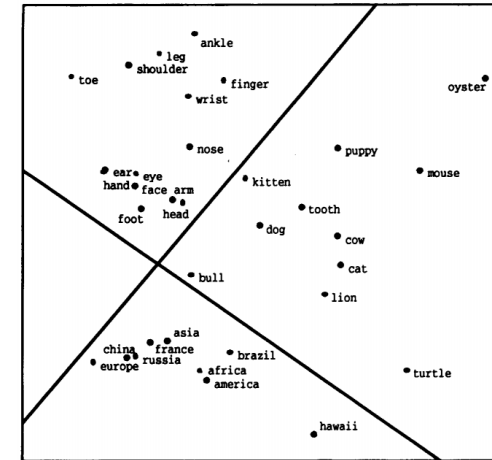


Figure 2. Multidimensional scaling of co-occurrence vectors.

Many parameters . . .

- ▶ Enormous range of DSM parameters and applications
- ▶ Examples showed three entirely different models, each tuned to its particular application
- ➡ Need overview of DSM parameters & understand their effects
 - ▶ part 2: The parameters of a DSM
 - ▶ part 3: Evaluating DSM representations
 - ▶ part 4: The mathematics of DSMs
 - ▶ part 5: Understanding distributional semantics
- ➡ Distributional semantics is an empirical science

Outline

Introduction

The distributional hypothesis
 Distributional semantic models
 Three famous examples

Getting practical

Software and further information
 R as a (toy) laboratory

Some applications in computational linguistics

- ▶ Unsupervised part-of-speech induction (Schütze 1995)
- ▶ Word sense disambiguation (Schütze 1998)
- ▶ Query expansion in information retrieval (Grefenstette 1994)
- ▶ Synonym tasks & other language tests (Landauer and Dumais 1997; Turney *et al.* 2003)
- ▶ Thesaurus compilation (Lin 1998; Rapp 2004)
- ▶ Ontology & wordnet expansion (Pantel *et al.* 2009)
- ▶ Attachment disambiguation (Pantel and Lin 2000)
- ▶ Probabilistic language models (Bengio *et al.* 2003)
- ▶ Sub-symbolic input representation for neural networks
- ▶ Many other tasks in computational semantics: entailment detection, noun compound interpretation, identification of noncompositional expressions, . . .

Recent conferences and workshops

- ▶ **2007:** CoSMo Workshop (at Context '07)
- ▶ **2008:** ESSLLI Lexical Semantics Workshop & Shared Task, Special Issue of the Italian Journal of Linguistics
- ▶ **2009:** GeMS Workshop (EACL 2009), DiSCo Workshop (CogSci 2009), ESSLLI Advanced Course on DSM
- ▶ **2010:** 2nd GeMS (ACL 2010), ESSLLI Workshop on Compositionality and DSM, DSM Tutorial (NAACL 2010), Special Issue of JNLE on Distributional Lexical Semantics
- ▶ **2011:** 2nd DiSCo (ACL 2011), 3rd GeMS (EMNLP 2011)
- ▶ **2012:** DiDaS (at ICSC 2012)
- ▶ **2013:** CVSC (ACL 2013), TFDS (IWCS 2013), Dagstuhl
- ▶ **2014:** 2nd CVSC (at EACL 2014)

click on Workshop name to open Web page

2016-08-15

DSM Tutorial – Part 1

Getting practical

Software and further information

Recent conferences and workshops

1. CoSMo = Contextual Information in Semantic Space Models
2. ESSLLI = European Summer School in Logic, Language and Information
3. GeMS = Geometrical Models of Natural Language Semantics
4. DiSCo = Distributional Semantics beyond Concrete Concepts
5. JNLE = Journal of Natural Language Engineering
6. DiSCo 2 = Distributional Semantics and Compositionality
7. DiDaS = Workshop on Distributional Data Semantics
8. CVSC = Continuous Vector Space Models and their Compositionality
9. TFDS = Towards a Formal Distributional Semantics

Recent conferences and workshops

- 2007: CoSMo Workshop (at Context '07)
- 2008: ESSLLI Lexical Semantics Workshop & Shared Task, Special Issue of the Italian Journal of Linguistics
- 2009: GeMS Workshop (EACL 2009), DiSCo Workshop (CogSci 2009), ESSLLI Advanced Course on DSM
- 2010: 2nd GeMS (ACL 2010), ESSLLI Workshop on Compositionality and DSM, DSM Tutorial (NAACL 2010), Special Issue of JNLE on Distributional Lexical Semantics
- 2011: 2nd DiSCo (ACL 2011), 3rd GeMS (EMNLP 2011)
- 2012: DiDaS (at ICSC 2012)
- 2013: CVSC (ACL 2013), TFDS (IWCS 2013), Dagstuhl
- 2014: 2nd CVSC (at EACL 2014)

click on Workshop name to open Web page

Software packages

HiDEx	C++	<i>re-implementation of the HAL model (Lund and Burgess 1996)</i>
SemanticVectors	Java	<i>scalable architecture based on random indexing representation</i>
S-Space	Java	<i>complex object-oriented framework</i>
JoBimText	Java	<i>UIMA / Hadoop framework</i>
Gensim	Python	<i>complex framework, focus on parallelization and out-of-core algorithms</i>
DISSECT	Python	<i>user-friendly, designed for research on compositional semantics</i>
wordspace	R	<i>interactive research laboratory, but scales to real-life data sets</i>

click on package name to open Web page

Further information

- ▶ Handouts & other materials available from workspace wiki at <http://workspace.collocations.de/>
 - based on joint work with Marco Baroni and Alessandro Lenci
- ▶ Tutorial is open source (CC), and can be downloaded from <http://r-forge.r-project.org/projects/workspace/>
- ▶ Review paper on distributional semantics:

Turney, Peter D. and Pantel, Patrick (2010). *From frequency to meaning: Vector space models of semantics*. *Journal of Artificial Intelligence Research*, **37**, 141–188.
- ▶ I should be working on textbook *Distributional Semantics* for *Synthesis Lectures on HLT* (Morgan & Claypool)

Prepare to get your hands dirty ...

- ▶ We will use the statistical programming environment **R** as a toy laboratory in this tutorial
 - but one that scales to real-life applications

Software installation

- ▶ **R** version 3.3 or newer from <http://www.r-project.org/>
- ▶ RStudio from <http://www.rstudio.com/>
- ▶ R packages from CRAN (through RStudio menu):

sparsesvd, **workspace**

 - ▶ if you are attending a course, you may also be asked to install the **workspaceEval** package with some non-public data sets
- ▶ Data sets from <http://www.collocations.de/data/#dsm>

Outline

Introduction

The distributional hypothesis
 Distributional semantic models
 Three famous examples

Getting practical

Software and further information
 R as a (toy) laboratory

First steps in R

Start each session by loading the workspace package.

```
> library(workspace)
```

The package includes various example data sets, some of which should look familiar to you.

```
> DSM_HieroglyphsMatrix
      get see use hear eat kill
knife  51  20  84   0   3   0
cat    52  58   4   4   6  26
dog    115  83  10  42  33  17
boat   59  39  23   4   0   0
cup    98  14   6   2   1   0
pig    12  17   3   2   9  27
banana 11   2   2   0  18   0
```

Term-term matrix

Term-term matrix records co-occurrence frequencies with feature terms for each target term

```
> DSM_TermTermMatrix
```

	breed	tail	feed	kill	important	explain	likely
cat	83	17	7	37	–	1	–
dog	561	13	30	60	1	2	4
animal	42	10	109	134	13	5	5
time	19	9	29	117	81	34	109
reason	1	–	2	14	68	140	47
cause	–	1	–	4	55	34	55
effect	–	–	1	6	60	35	17

Term-context matrix

Term-context matrix records frequency of term in each individual context (e.g. sentence, document, Web page, encyclopaedia article)

```
> DSM_TermContextMatrix
```

	Felidae	Pet	Feral	Bloat	Philosophy	Kant	Back pain
cat	10	10	7	–	–	–	–
dog	–	10	4	11	–	–	–
animal	2	15	10	2	–	–	–
time	1	–	–	–	2	1	–
reason	–	1	–	–	1	4	1
cause	–	–	–	2	1	2	6
effect	–	–	–	1	–	1	–

Some basic operations on a DSM matrix

```
# apply log-transformation to de-skew co-occurrence frequencies
> M <- log2(DSM_HieroglyphsMatrix + 1) # see part 2
> round(M, 3)

# compute semantic distance (cosine similarity)
> pair.distances("dog", "cat", M, convert=FALSE)
dog/cat
0.9610952

# find nearest neighbours
> nearest.neighbours(M, "dog", n=3)
      cat      pig      cup
16.03458 20.08826 31.77784

> plot(nearest.neighbours(M, "dog", n=3, dist.matrix=TRUE))
```

References I

- Bengio, Yoshua; Ducharme, Réjean; Vincent, Pascal; Jauvin, Christian (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137–1155.
- Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Deerwester, S.; Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *CHI '88: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, pages 1–32. The Philological Society, Oxford.
- Grefenstette, Gregory (1994). *Explorations in Automatic Thesaurus Discovery*, volume 278 of *Kluwer International Series in Engineering and Computer Science*. Springer, Berlin, New York.
- Harris, Zellig (1954). Distributional structure. *Word*, 10(23), 146–162.
- 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.

References II

- Li, Ping; Burgess, Curt; Lund, Kevin (2000). The acquisition of word meaning through global lexical co-occurrences. In E. V. Clark (ed.), *The Proceedings of the Thirtieth Annual Child Language Research Forum*, pages 167–178. Stanford Linguistics Association.
- 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.
- Lund, Kevin and Burgess, Curt (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, **28**(2), 203–208.
- Miller, George A. (1986). Dictionaries in the mind. *Language and Cognitive Processes*, **1**, 171–185.
- Pantel, Patrick and 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.

References IV

- Turney, Peter D.; Littman, Michael L.; Bigham, Jeffrey; Shnayder, Victor (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pages 482–489, Borovets, Bulgaria.

References III

- Pantel, Patrick; Crestan, Eric; Borkovsky, Arkady; Popescu, Ana-Maria; Vyas, Vishnu (2009). Web-scale distributional similarity and entity set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 938–947, Singapore.
- 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 (1992). Dimensions of meaning. In *Proceedings of Supercomputing '92*, pages 787–796, Minneapolis, MN.
- Schütze, Hinrich (1993). Word space. In *Proceedings of Advances in Neural Information Processing Systems 5*, pages 895–902, San Mateo, CA.
- 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. and Pantel, Patrick (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, **37**, 141–188.