

Distributional Semantic Models

Part 1: Introduction

Stefan Evert¹

with Alessandro Lenci², Marco Baroni³ and Gabriella Lapesa⁴

¹Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

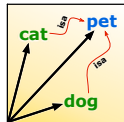
²University of Pisa, Italy

³University of Trento, Italy

⁴University of Stuttgart, Germany

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

Copyright © 2009–2018 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
- ▶ “You shall know a word by the company it keeps!”
— J. R. Firth (1957)
- ▶ Distributional hypothesis: difference of meaning correlates with difference of distribution (Zellig Harris 1954)
- ▶ “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)

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

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)
- ▶ “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)

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**”?

Can we infer meaning from usage?

What is the meaning of “**bardiwac**”?

Can we infer meaning from usage?

- ▶ He handed her her glass of **bardiwac**.

What is the meaning of “**bardiwac**”?

Can we infer meaning from usage?

- ▶ He handed her her glass of **bardiwac**.
- ▶ Beef dishes are made to complement the **bardiwacs**.

What is the meaning of “**bardiwac**”?

Can we infer meaning from usage?

- ▶ 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**.

What is the meaning of “**bardiwac**”?

Can we infer meaning from usage?

- ▶ 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.

What is the meaning of “**bardiwac**”?

Can we infer meaning from usage?

- ▶ 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**.


What is the meaning of “**bardiwac**”?

Can we infer meaning from usage?

- ▶ 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.

What is the meaning of “**bardiwac**”?

Can we infer meaning from usage?

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

All examples from British National Corpus (handpicked and slightly edited).

Word sketch of “cat”

Can we infer meaning from collocations?














cat British National Corpus freq = 5381

<https://the.sketchengine.co.uk/>

object of 964 2.0	and/or 1056 1.7	pp obj like-p 106 28.9	possessor 91 1.9	possession 232 4.7
skin 9 7.91	dog 208 8.49	grin 11 7.63	Schrödinger 8 10.87	cradle 24 9.91
diddle 7 7.85	cat 68 8.01	fight 9 4.62	witch 4 6.82	whisker 9 8.92
stroke 10 7.09	kitten 13 8.01	smile 4 4.24	gardener 4 6.0	paw 5 7.44
torture 5 6.57	fiddle 9 7.71	look 11 2.04	Henry 8 4.91	fur 9 7.14
feed 22 6.34	mouse 29 7.68		neighbour 5 4.28	tray 4 5.34
rain 4 6.3	monkey 15 7.55	pp among-p 17 14.8		tail 5 4.91
chase 9 6.27	budgie 4 6.74	pigeon 15 8.66		tongue 5 4.89
rescue 7 6.15	rabbit 12 6.48			ear 5 4.0

subject of 842 3.3	adj subject of 142 2.6	pp obj of-p 324 1.3	modifier 1622 1.2	modifies 610 0.5
purr 7 7.76	asleep 4 6.09	moral 4 7.06	pussy 76 10.42	flap 16 8.39
miaow 5 7.57	alive 4 5.06	breed 6 5.77	Cheshire 45 8.9	litter 15 8.15
mew 4 7.18	concerned 4 2.94	signal 4 3.89	stray 25 8.7	phobia 5 7.64
jump 20 6.95	black 4 2.36	sight 4 3.77	siamese 17 8.35	burglar 8 7.55
scratch 8 6.84	likely 4 1.96	species 5 3.36	tabby 17 8.35	faeces 6 7.47
leap 10 6.78		game 9 3.14	wild 53 7.94	assay 10 7.38
stalk 4 6.56		picture 6 2.99	pet 31 7.92	Hastings 7 6.91
react 4 5.33		death 7 2.71	tom 12 7.8	scan 4 6.59

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

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














$$\text{sim}(\text{semi-circle, vertical line, triangle}, \text{wavy line, knife, wavy line}) = 0.770$$

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














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

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

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

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

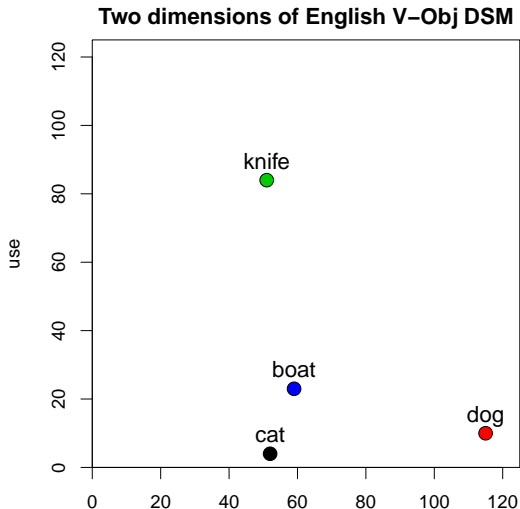
- ▶ row vector \mathbf{x}_{dog} describes usage of word *dog* in the corpus
- ▶ can be seen as coordinates of point in n -dimensional Euclidean space

	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 \mathbf{M}

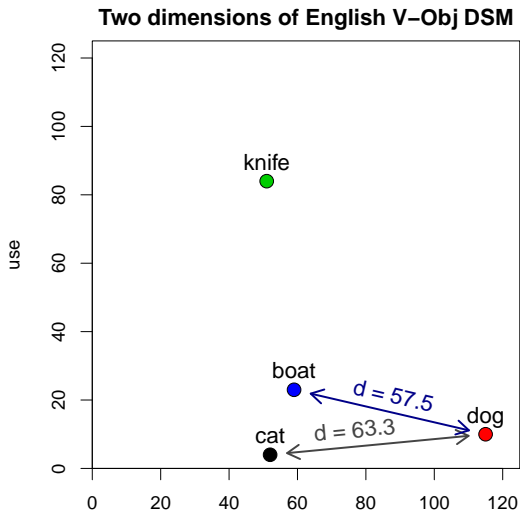
Geometric interpretation

- ▶ 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

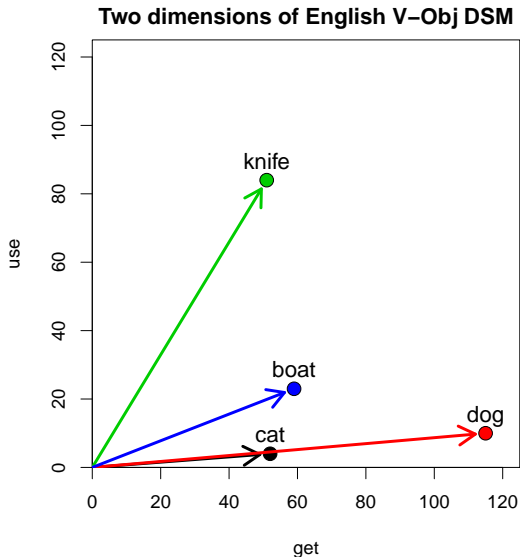
- ▶ similarity = spatial proximity (Euclidean dist.)
- ▶ location depends on frequency of noun ($f_{\text{dog}} \approx 2.7 \cdot f_{\text{cat}}$)



get

Geometric interpretation

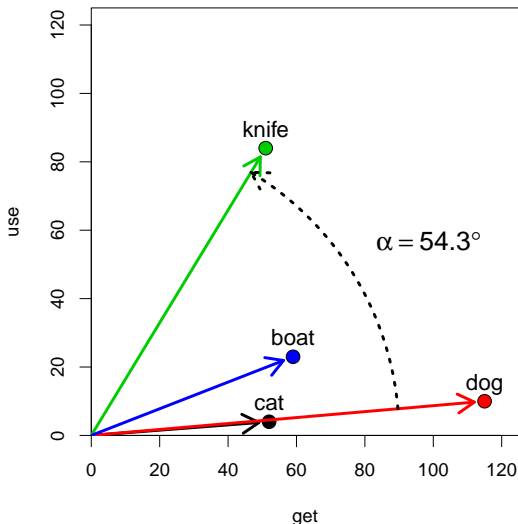
- ▶ vector can also be understood as arrow from origin
- ▶ direction more important than location



Geometric interpretation

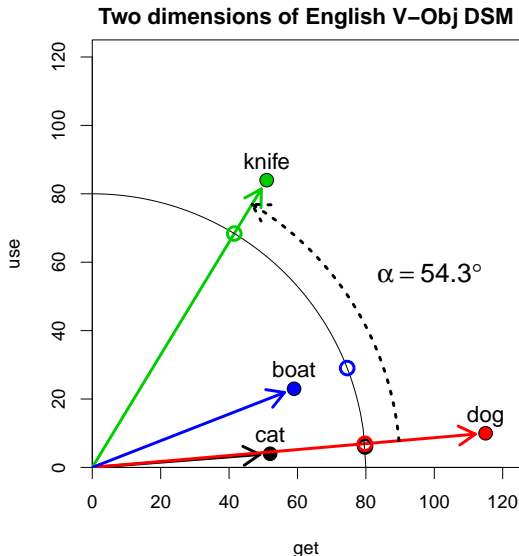
- ▶ vector can also be understood as arrow from origin
- ▶ direction more important than location
- ▶ use angle α as distance measure

Two dimensions of English V-Obj DSM



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 $\|\mathbf{x}_{\text{dog}}\|$ 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.

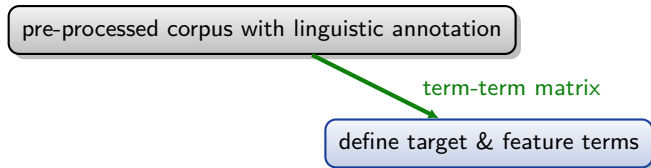
	get	see	use	hear	eat	kill
knife	0.027	-0.024	0.206	-0.022	-0.044	-0.042
cat	0.031	0.143	-0.243	-0.015	-0.009	0.131
dog	-0.026	0.021	-0.212	0.064	0.013	0.014
boat	-0.022	0.009	-0.044	-0.040	-0.074	-0.042
cup	-0.014	-0.173	-0.249	-0.099	-0.119	-0.042
pig	-0.069	0.094	-0.158	0.000	0.094	0.265
banana	0.047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

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

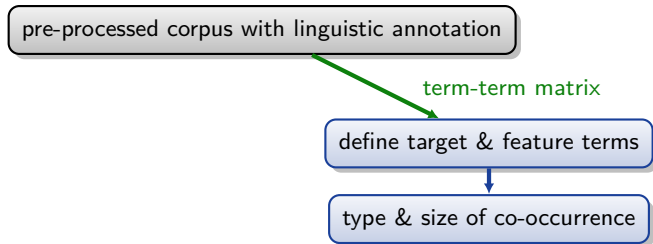
Building a distributional model

pre-processed corpus with linguistic annotation

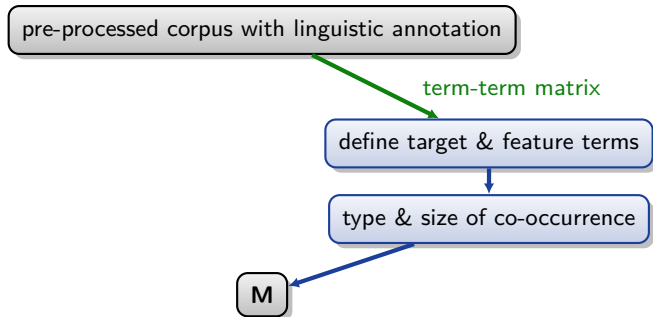
Building a distributional model



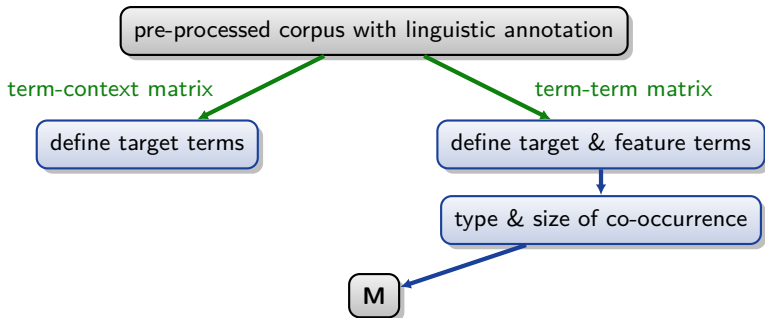
Building a distributional model



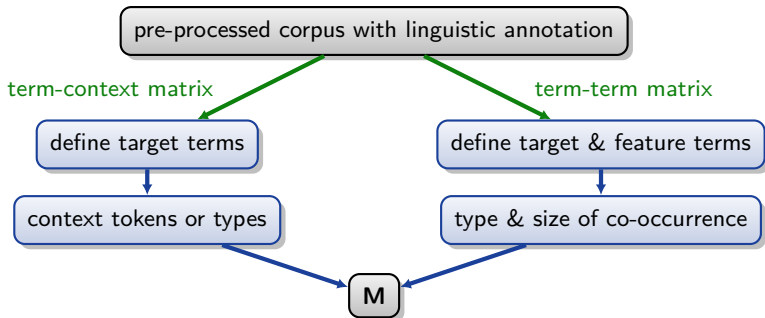
Building a distributional model



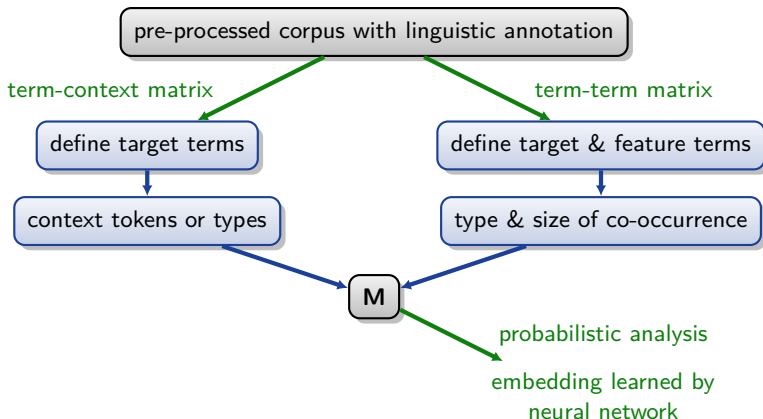
Building a distributional model



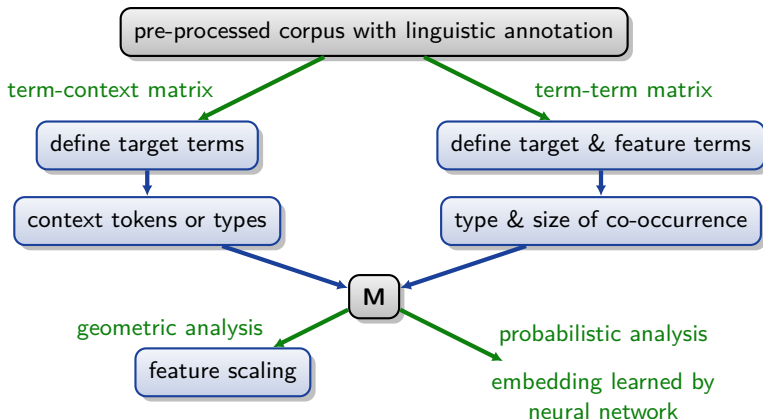
Building a distributional model



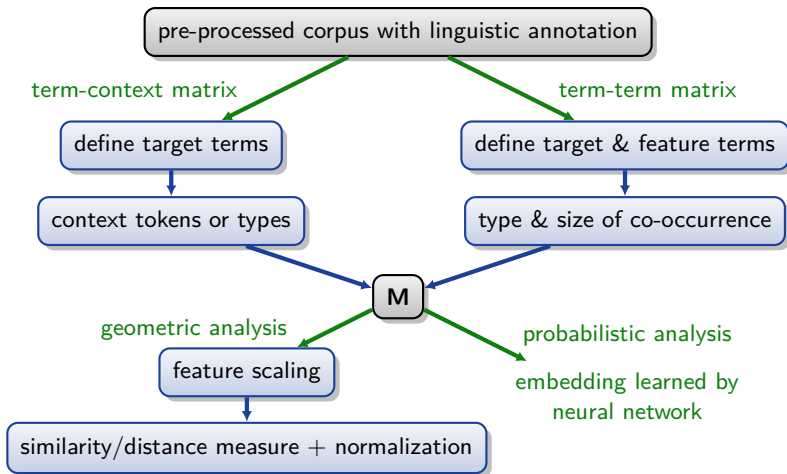
Building a distributional model



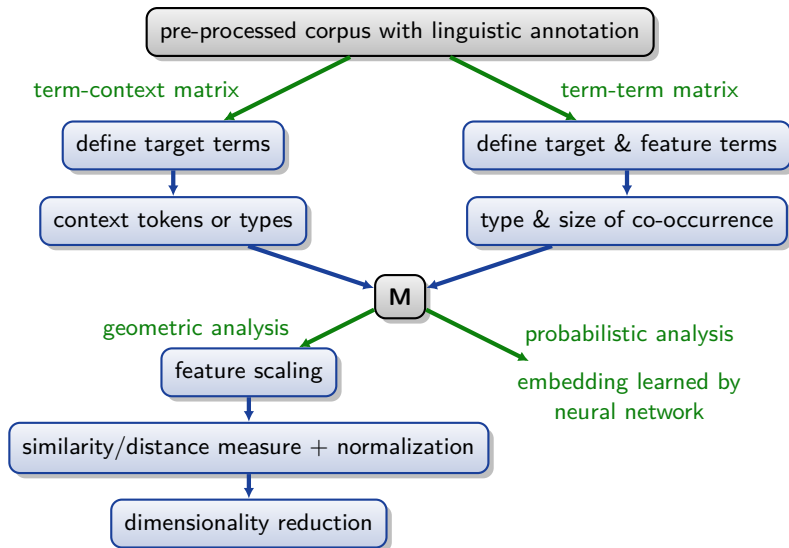
Building a distributional model



Building a distributional model



Building a distributional model



Nearest neighbours

DSM based on verb-object relations from BNC, reduced to 100 dim. with SVD

Neighbours of **trousers** (cosine angle):

☞ shirt (18.5), blouse (21.9), scarf (23.4), jeans (24.7), skirt (25.9), sock (26.2), shorts (26.3), jacket (27.8), glove (28.1), coat (28.8), cloak (28.9), hat (29.1), tunic (29.3), overcoat (29.4), pants (29.8), helmet (30.4), apron (30.5), robe (30.6), mask (30.8), tracksuit (31.0), jersey (31.6), shawl (31.6), ...

Nearest neighbours

DSM based on verb-object relations from BNC, reduced to 100 dim. with SVD

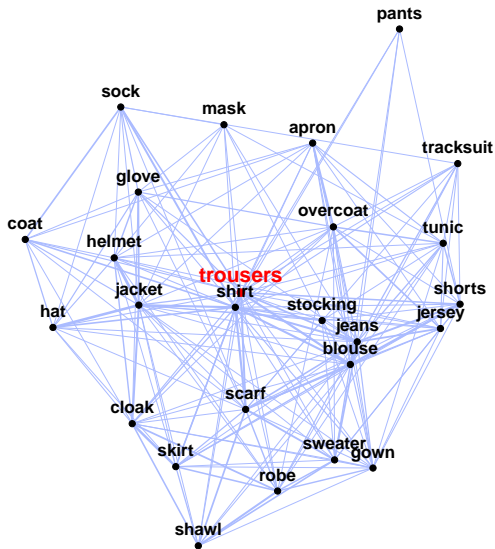
Neighbours of **trousers** (cosine angle):

👉 shirt (18.5), blouse (21.9), scarf (23.4), jeans (24.7), skirt (25.9), sock (26.2), shorts (26.3), jacket (27.8), glove (28.1), coat (28.8), cloak (28.9), hat (29.1), tunic (29.3), overcoat (29.4), pants (29.8), helmet (30.4), apron (30.5), robe (30.6), mask (30.8), tracksuit (31.0), jersey (31.6), shawl (31.6), ...

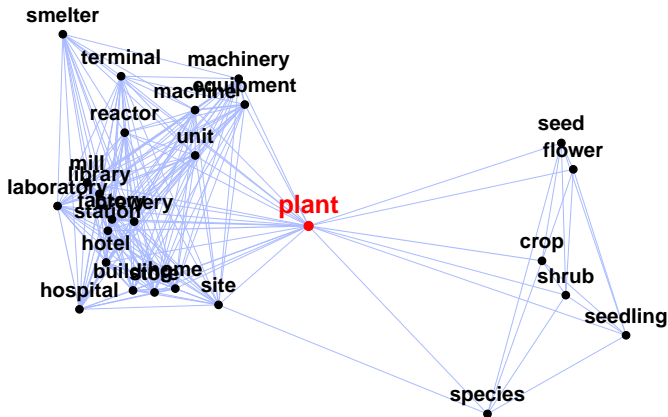
Neighbours of **rage** (cosine angle):

👉 anger (28.5), fury (32.5), sadness (37.0), disgust (37.4), emotion (39.0), jealousy (40.0), grief (40.4), irritation (40.7), revulsion (40.7), scorn (40.7), panic (40.8), bitterness (41.6), resentment (41.8), indignation (41.9), excitement (42.0), hatred (42.5), envy (42.8), disappointment (42.9), ...

Nearest neighbours with similarity graph

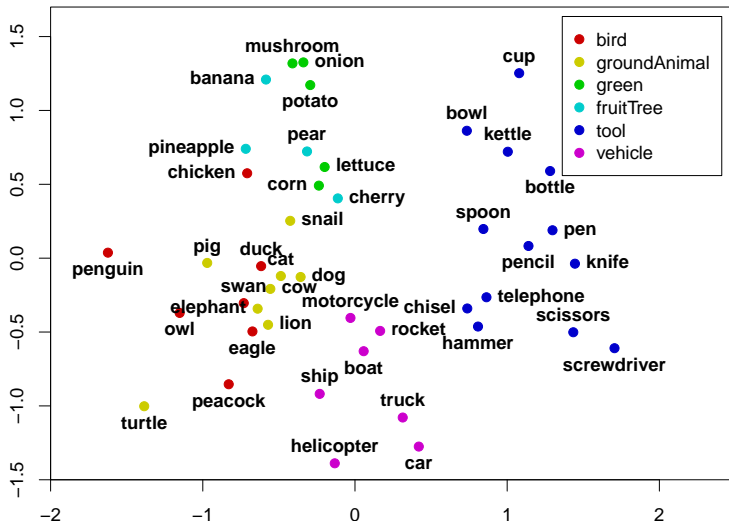


Nearest neighbours with similarity graph

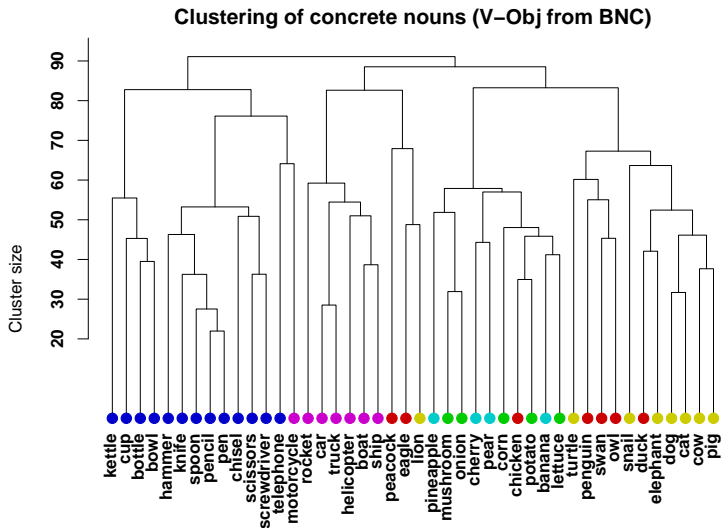


Semantic maps

Semantic map (V-Obj from BNC)

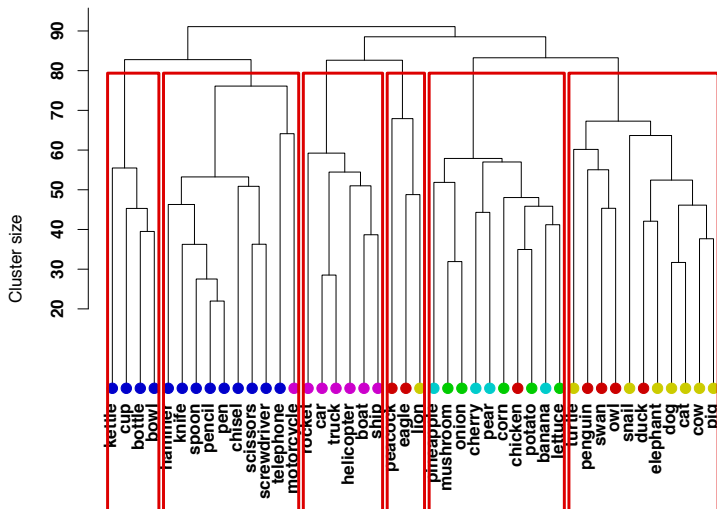


Clustering

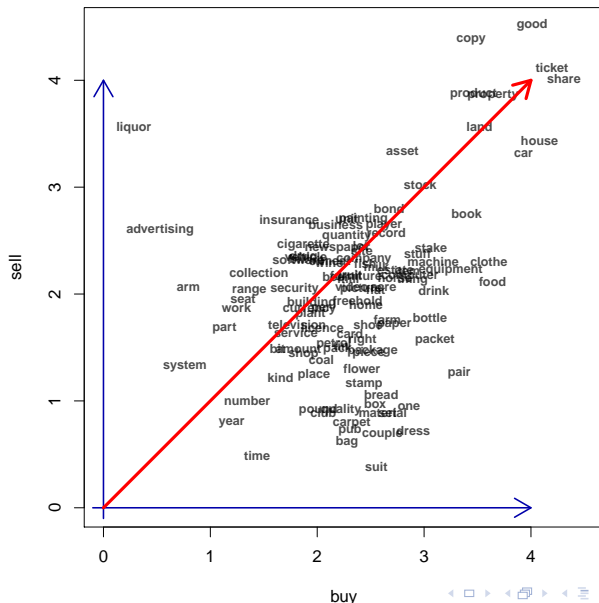


Clustering

Clustering of concrete nouns (V-Obj from BNC)



Latent “meaning” dimensions



Word embeddings

DSM vector as sub-symbolic meaning representation

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

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

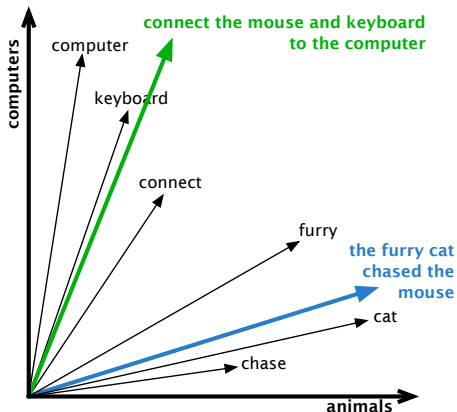
Word embeddings

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

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:
 $\text{distributional similarity/distance} \sim \text{semantic similarity/distance}$

► **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**

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)

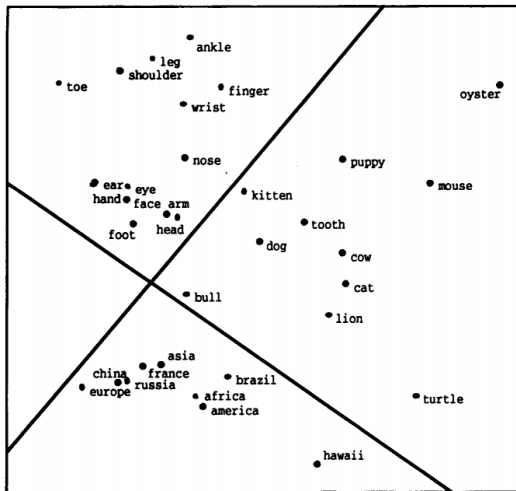


Figure 2. Multidimensional scaling of co-occurrence vectors.

Many parameters . . .

- ▶ Enormous range of DSM parameters and applications
- ▶ Examples showed three entirely different models, each tuned to its particular application

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: Matrix algebra & SVD
 - ▶ part 5: Understanding distributional semantics

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: Matrix algebra & SVD
 - ▶ 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

- ▶ Query expansion in information retrieval (Grefenstette 1994)
- ▶ Unsupervised part-of-speech induction (Schütze 1995)
- ▶ Word sense disambiguation (Schütze 1998; Rapp 2004b)
- ▶ Thesaurus compilation (Lin 1998; Rapp 2004a)
- ▶ Attachment disambiguation (Pantel and Lin 2000)
- ▶ Probabilistic language models (Bengio *et al.* 2003)
- ▶ Translation equivalents (Sahlgren and Karlgren 2005)
- ▶ Ontology & wordnet expansion (Pantel *et al.* 2009)
- ▶ Language change (Sagi *et al.* 2009; Hamilton *et al.* 2016)
- ▶ Multiword expressions (Kiehl and Clark 2013)
- ▶ Analogies (Turney 2013; Gladkova *et al.* 2016)
- ▶ Sentiment analysis (Rothe and Schütze 2016; Yu *et al.* 2017)
- 👉 Input representation for neural networks & machine learning

Recent workshops and tutorials

- ▶ 2007: CoSMo Workshop (at Context '07)
- ▶ 2008: ESSLLI Wshp & Shared Task, Italian J of Linguistics
- ▶ 2009: GeMS Wshp (EACL), DiSCo Wshp (CogSci), ESSLLI
- ▶ 2010: 2nd GeMS (ACL), ESSLLI Wshp, Tutorial (NAACL),
J Natural Language Engineering
- ▶ 2011: 2nd DiSCo (ACL), 3rd GeMS (EMNLP)
- ▶ 2012: DiDaS Wshp (ICSC), ESSLLI Course
- ▶ 2013: CVSC Wshp (ACL), TFDS Wshp (IWCS), Dagstuhl
- ▶ 2014: 2nd CVSC (EACL), DSM Wshp (Insight)
- ▶ 2015: VSM4NLP (NAACL), ESSLLI Course, TAL Journal
- ▶ 2016: DSALT Wshp (ESSLLI), Tutorial (COLING), Tutorial
(Konvens), ESSLLI Course, Computational Linguistics
- ▶ 2017: ESSLLI Course
- ▶ 2018: Tutorial (LREC), ESSLLI Course₁ & Course₂

click on Workshop name to open Web page



Software packages

Infomap NLP	C	<i>classical LSA-style DSM</i>
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>
Vecto	Python	<i>framework for count & predict models</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 wordspace wiki at <http://wordspace.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/wordspace/>
- ▶ 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)

Outline

Introduction

The distributional hypothesis

Distributional semantic models

Three famous examples

Getting practical

Software and further information

R as a (toy) laboratory

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 3.5 or newer from <http://www.r-project.org/>
- ▶ RStudio from <http://www.rstudio.com/>
- ▶ R packages from CRAN (through RStudio menu):
[sparsesvd](#), **[wordspace](#)** (optional: **[tm](#)**, **[quanteda](#)**, **[Rtsne](#)**)
 - ▶ if you are attending a course, you may also be asked to install the **[wordspaceEval](#)** package with some non-public data sets
- ▶ Get data sets, precompiled DSMs and **[wordspaceEval](#)** from <http://wordspace.collocations.de/doku.php/course:material>

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
```

	<i>breed</i>	<i>tail</i>	<i>feed</i>	<i>kill</i>	<i>important</i>	<i>explain</i>	<i>likely</i>
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))
```


Explorations

While you wait for part 2,
you can explore some DSM similarity networks online:

- ▶ <https://corpora.linguistik.uni-erlangen.de/shiny/workspace/>
- ▶ built in R with `workspace` and `shiny`

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.
- Gladkova, Anna; Drozd, Aleksandr; Matsuoka, Satoshi (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California.
- Grefenstette, Gregory (1994). *Explorations in Automatic Thesaurus Discovery*, volume 278 of *Kluwer International Series in Engineering and Computer Science*. Springer, Berlin, New York.

References II

- Hamilton, William L.; Clark, Kevin; Leskovec, Jure; Jurafsky, Dan (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 595–605, Austin, TX.
- Harris, Zellig (1954). Distributional structure. *Word*, **10**(23), 146–162.
- Kiela, Douwe and Clark, Stephen (2013). Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1427–1432, Seattle, WA.
- 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.
- 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.

References III

- 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.
- 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 (2004a). 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.

References IV

- Rapp, Reinhard (2004b). A practical solution to the problem of automatic word sense induction. In *Proceedings of the ACL-2004 Interactive Posters and Demonstrations Sessions*, pages 194–197, Barcelona, Spain. Association for Computational Linguistics.
- Rothe, Sascha and Schütze, Hinrich (2016). Word embedding calculus in meaningful ultradense subspaces. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–517, Berlin, Germany.
- Sagi, Eyal; Kaufmann, Stefan; Clark, Brady (2009). Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (GEMS)*, pages 104–111, Athens, Greece.
- Sahlgren, Magnus and Karlgren, Jussi (2005). Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, **11**, 327–341.
- 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.

References V

- 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. (2013). Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *Transactions of the Association for Computational Linguistics*, **1**, 353–366.
- Turney, Peter D. and Pantel, Patrick (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, **37**, 141–188.
- Yu, Liang-Chih; Wang, Jin; Lai, K. Robert; Zhang, Xuejie (2017). Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark.