Distributional Semantic Models

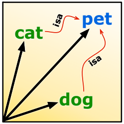
Part 1: Introduction

Stefan Evert[1]

with Alessandro Lenci[2], Marco Baroni[3] and Gabriella Lapesa[4]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2]University of Pisa, Italy
[3]University of Trento, Italy
[4]University of Stuttgart, Germany

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



Outline

Introduction
- The distributional hypothesis
- Three famous examples

Distributional semantic models
- Definition & overview
- Using DSM distances
- Quantitative evaluation
- Software and further information

Outline

Introduction
- The distributional hypothesis
- Three famous examples

Distributional semantic models
- Definition & overview
- Using DSM distances
- Quantitative evaluation
- Software and further information

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)

What is the meaning of "**bardiwac**"?

What is the meaning of "**bardiwac**"?

- He handed her her glass of bardiwac.

What is the meaning of "**bardiwac**"?

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

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.

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.

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.

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.

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
uncork	1	8.98
gulp	1	6.61
sport	1	5.6
water	1	5.34
drink	7	5.13
sip	1	4.8
warm	1	4.28
complement	1	4.15
waste	1	2.93
paint	1	2.38

and/or	47	1.7
plummy	1	9.33
Sancerre	1	9.14
Willson	1	8.93
scampi	1	8.23
burgundy	1	8.18
garb	1	7.02
ruby	1	6.59
Barnett	1	5.29
refreshment	1	5.29
Halifax	1	5.11

pp_obj_round-p	1	29.1
pass	1	0.3

pp_before-p	1	13.0
dinner	1	1.98

pp_obj_after-p	1	6.5
sought	1	8.56

pp_obj_of-p	63	5.7
swig	1	7.21
tinge	1	6.44
bottle	24	6.35
goblet	1	6.29
jug	1	4.64
grape	1	4.63
cup	16	4.38
bowl	2	3.66
glass	4	2.83
label	1	2.76

pp_obj_through-p	1	4.5
plausible	1	5.28

predicate_of	4	3.7
Branaire-ducru	1	12.19
Spar	1	8.85
liquor	2	5.82

pp_obj_with-p	6	3.3
fagg	1	9.54
brim	1	6.71
stain	2	5.49
merchant	1	2.68
meal	1	1.64

pp_obj_by-p	4	2.5
embolden	1	8.29
refresh	1	6.36
confuse	1	4.36
accompany	1	1.63

pp_as-p	1	1.9
gift	1	2.14

predicate	2	1.8
tipple	1	7.91
wine	1	1.53

pp_obj_to-p	5	1.7
alternative	1	2.2
trip	1	1.7
attend	1	1.35

pp_obj_from-p	2	1.6
burgundy	1	8.91
flush	1	4.71

adj_subject_of	3	1.2
cheap	1	3.08
happy	1	1.66
sure	1	0.56

modifier	72	1.2
passable	5	9.92
ready-to-drink	1	8.79
cinnamon-scented	1	8.79
rust-coloured	1	8.57
Tanners	1	8.51
ten-man	1	8.43
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	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

$$\mathrm{sim}(\ , \ ) = 0.770$$

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

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

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

$$\text{sim}(\;\cdot\; , \;\cdot\;) = 0.961$$

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

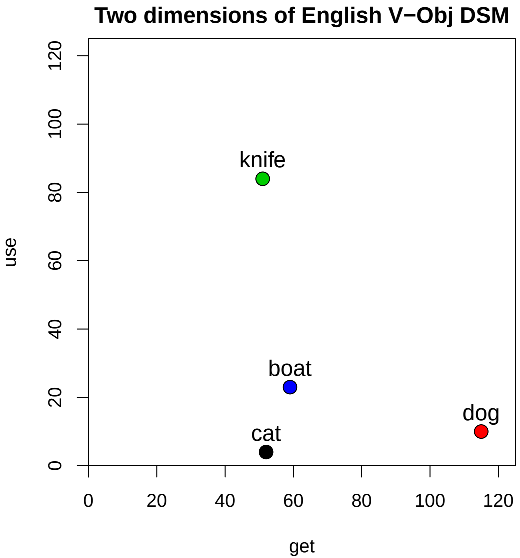
- row vector $\mathbf{x}_{\text{dog}}$ describes usage of word *dog* in the corpus
- can be seen as coordinates of point in n-dimensional Euclidean space

	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 $\mathbf{M}$

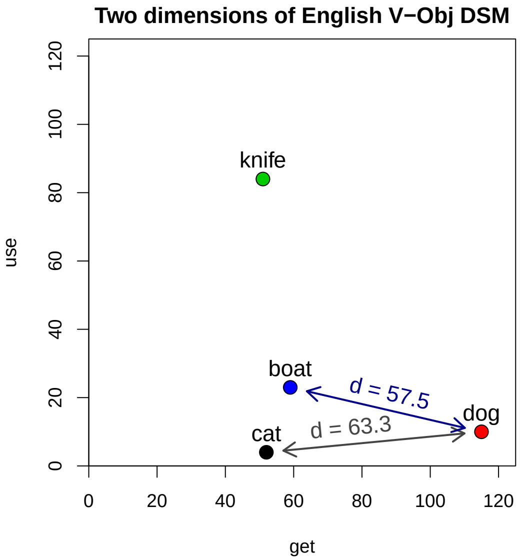
Geometric interpretation

- row vector $\mathbf{x}_{\text{dog}}$ 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*
- $\mathbf{x}_{\text{dog}} = (115, 10)$



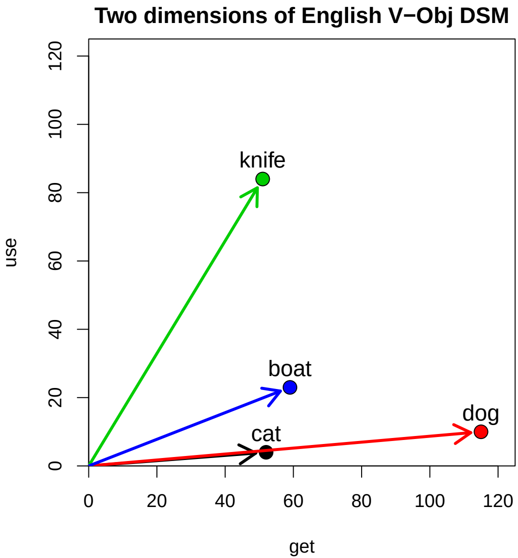
Geometric interpretation

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



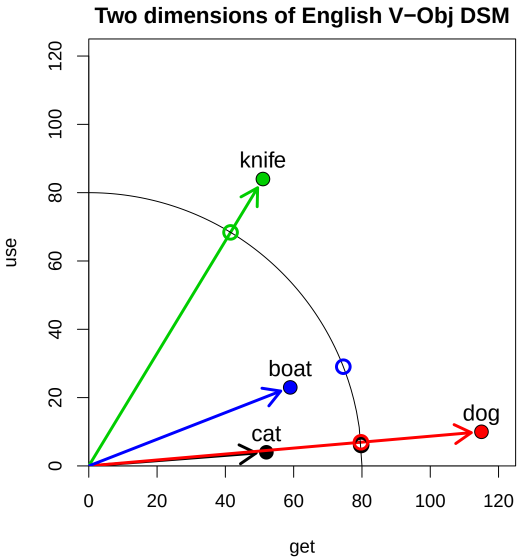
Geometric interpretation

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



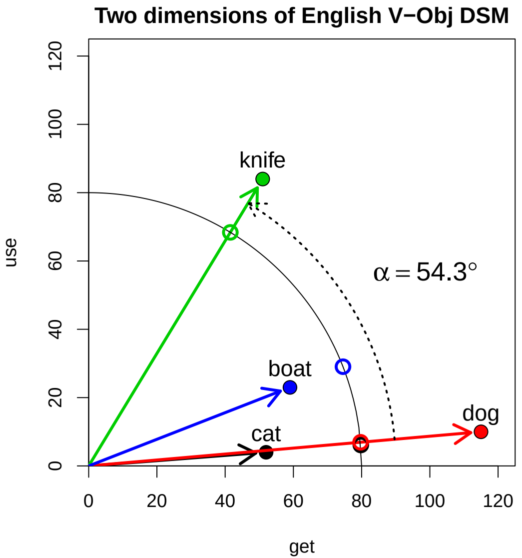
Geometric interpretation

- similarity = spatial proximity (Euclidean dist.)
- location depends on frequency of noun ($f_{dog} \approx 2.7 \cdot f_{cat}$)
- direction more important than location
- normalise "length" $\|\mathbf{x}_{dog}\|$ of vector



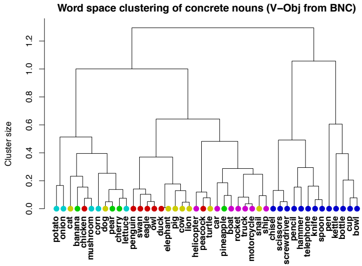
Geometric interpretation

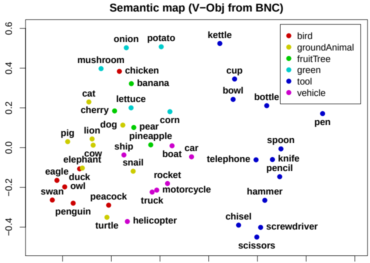
- similarity = spatial proximity (Euclidean dist.)
- location depends on frequency of noun ($f_{\text{dog}} \approx 2.7 \cdot f_{\text{cat}}$)
- direction more important than location
- normalise "length" $\|\mathbf{x}_{\text{dog}}\|$ of vector
- or use angle α as distance measure



Semantic distances

- main result of distributional analysis are "semantic" distances between words
- immediate applications
  - nearest neighbours
  - clustering of related words
  - construct semantic map
- other applications require clever use of the distance information
  - semantic relations
  - relational analogies
  - word sense disambiguation
  - identification of multiword expressions





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:
    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**

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

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)
- Subsymbolic input representation for neural networks
- Many other tasks in computational semantics:
  entailment detection, noun compound interpretation, identification of noncompositional expressions, . . .

Outline

Introduction
The distributional hypothesis
Three famous examples

Distributional semantic models
Definition & overview
Using DSM distances
Quantitative evaluation
Software and further information

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: ≈ 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 ≈ 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)
- Applications include construction of semantic vocabulary maps by multidimensional scaling to 2 dimensions

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

Outline

Introduction
The distributional hypothesis
Three famous examples

Distributional semantic models
Definition & overview
Using DSM distances
Quantitative evaluation
Software and further information

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, . . .

General definition of DSMs

Mathematical notation:

- $k \times n$ co-occurrence matrix $\mathbf{M}$ (example: 7×6 matrix)
  - k rows = target terms
  - n columns = features or **dimensions**

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \\ \vdots & \vdots & & \vdots \\ m_{k1} & m_{k2} & \cdots & m_{kn} \end{bmatrix}$$

- distribution vector $\mathbf{m}_i$ = i-th row of $\mathbf{M}$, e.g. $\mathbf{m}_3 = \mathbf{m}_{\text{dog}}$
- components $\mathbf{m}_i = (m_{i1}, m_{i2}, \ldots, m_{in})$ = features of i-th term:

$$\begin{aligned} \mathbf{m}_3 &= (-0.026, 0.021, -0.212, 0.064, 0.013, 0.014) \\ &= (m_{31}, m_{32}, m_{33}, m_{34}, m_{35}, m_{36}) \end{aligned}$$

Overview of DSM parameters

pre-processed corpus with linguistic annotation

Overview of DSM parameters

pre-processed corpus with linguistic annotation

$\Downarrow$

term-context vs. term-term matrix

Overview of DSM parameters

pre-processed corpus with linguistic annotation

$\Downarrow$

term-context vs. term-term matrix

$\swarrow$ $\searrow$

define targets & contexts | define targets & features

Overview of DSM parameters

pre-processed corpus with linguistic annotation

$\Downarrow$

term-context vs. term-term matrix

$\swarrow$ $\searrow$

define targets & contexts | define targets & features

$\Downarrow$ $\Downarrow$

type of context | type & size of co-occurrence window

Overview of DSM parameters

pre-processed corpus with linguistic annotation

$\Downarrow$

term-context vs. term-term matrix

$\swarrow$ $\searrow$

define targets & contexts | define targets & features

$\Downarrow$ $\Downarrow$

type of context | type & size of co-occurrence window

$\searrow$ $\swarrow$

geometric vs. probabilistic interpretation

Overview of DSM parameters

pre-processed corpus with linguistic annotation

$\Downarrow$

term-context vs. term-term matrix

$\swarrow$ $\searrow$

define targets & contexts | define targets & features

$\Downarrow$ | $\Downarrow$

type of context | type & size of co-occurrence window

$\searrow$ $\swarrow$

geometric vs. probabilistic interpretation

$\Downarrow$

feature scaling

Overview of DSM parameters

pre-processed corpus with linguistic annotation

$\Downarrow$

term-context vs. term-term matrix

$\swarrow$ $\searrow$

define targets & contexts | define targets & features

$\Downarrow$ $\Downarrow$

type of context | type & size of co-occurrence window

$\searrow$ $\swarrow$

geometric vs. probabilistic interpretation

$\Downarrow$

feature scaling

$\Downarrow$

normalisation of rows and/or columns

Overview of DSM parameters

pre-processed corpus with linguistic annotation

$\Downarrow$

term-context vs. term-term matrix

$\swarrow$ $\searrow$

define targets & contexts	define targets & features
$\Downarrow$	$\Downarrow$
type of context	type & size of co-occurrence window
$\searrow$	$\swarrow$

geometric vs. probabilistic interpretation

$\Downarrow$

feature scaling

$\Downarrow$

normalisation of rows and/or columns

$\Downarrow$

similarity / distance measure

Overview of DSM parameters

pre-processed corpus with linguistic annotation

⇓

term-context vs. term-term matrix

↙ ↘

define targets & contexts | define targets & features

⇓ ⇓

type of context | type & size of co-occurrence window

↘ ↙

geometric vs. probabilistic interpretation

⇓

feature scaling

⇓

normalisation of rows and/or columns

⇓

similarity / distance measure

⇓

dimensionality reduction

Term-context matrix

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

$$\mathbf{F} = \begin{bmatrix} \cdots & \mathbf{f}_1 & \cdots \\ \cdots & \mathbf{f}_2 & \cdots \\ & \vdots & \\ & \vdots & \\ \cdots & \mathbf{f}_k & \cdots \end{bmatrix}$$

	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	–

Term-context matrix

Some footnotes:

- Features are usually context tokens, i.e. individual instances
- Can also be generalised to context types, e.g.
  - bag of content words
  - specific pattern of POS tags
  - n-gram of words (or POS tags) around target
  - subcategorisation pattern of target verb
- Term-context matrix is often very **sparse**

Term-term matrix

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

$$\mathbf{M} = \begin{bmatrix} \cdots & \mathbf{m}_1 & \cdots \\ \cdots & \mathbf{m}_2 & \cdots \\ & \vdots & \\ & \vdots & \\ \cdots & \mathbf{m}_k & \cdots \end{bmatrix}$$

	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

☞ we will usually assume a term-term matrix in this tutorial

Term-term matrix

Some footnotes:

- Often target terms $\neq$ feature terms
  - e.g. nouns described by co-occurrences with verbs as features
  - identical sets of target & feature terms ➜ symmetric matrix
- Different types of contexts (Evert 2008)
  - **surface context** (word or character window)
  - **textual context** (non-overlapping segments)
  - **syntactic context** (specific syntagmatic relation)
  - additional data: "marginal" frequencies of targets and features
- Can be seen as smoothing of term-context matrix
  - average over similar contexts (with same context terms)
  - data sparseness reduced, except for small windows
  - we will take a closer look at the relation between term-context and term-term models later in this tutorial

Outline

Introduction
The distributional hypothesis
Three famous examples

Distributional semantic models
Definition & overview
Using DSM distances
Quantitative evaluation
Software and further information

Nearest neighbours

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

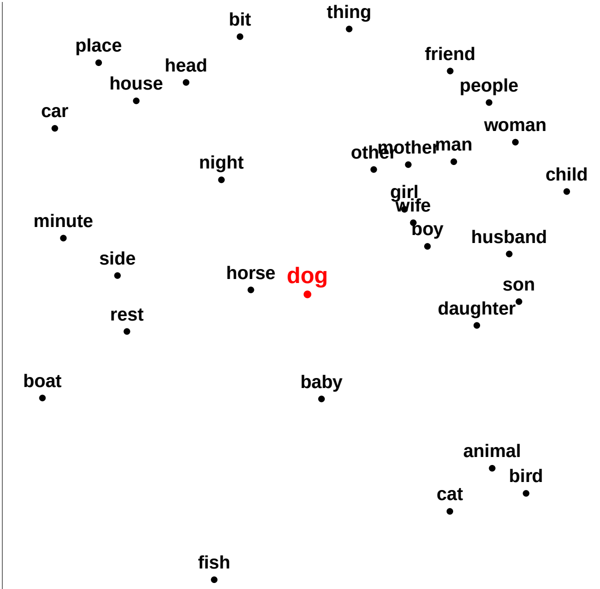
Neighbours of **dog** (cosine angle):

☞ girl (45.5), boy (46.7), horse(47.0), wife (48.8), baby (51.9), daughter (53.1), side (54.9), mother (55.6), boat (55.7), rest (56.3), night (56.7), cat (56.8), son (57.0), man (58.2), place (58.4), husband (58.5), thing (58.8), friend (59.6), . . .

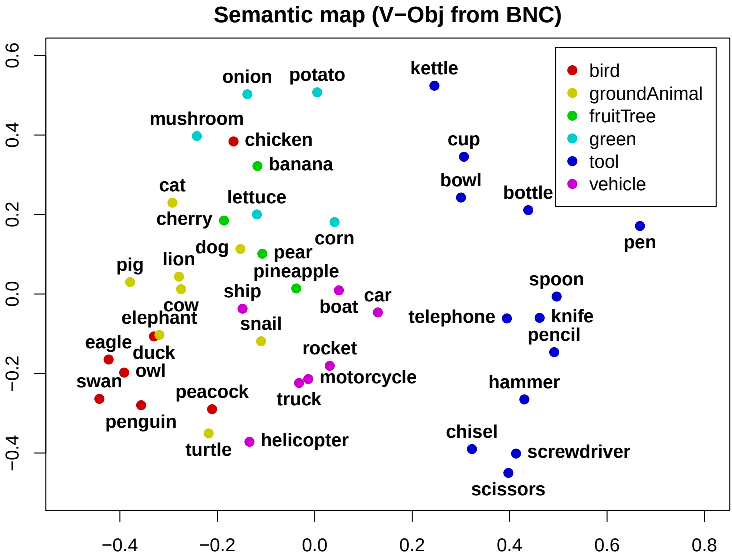
Neighbours of **school**:

☞ country (49.3), church (52.1), hospital (53.1), house (54.4), hotel (55.1), industry (57.0), company (57.0), home (57.7), family (58.4), university (59.0), party (59.4), group (59.5), building (59.8), market (60.3), bank (60.4), business (60.9), area (61.4), department (61.6), club (62.7), town (63.3), library (63.3), room (63.6), service (64.4), police (64.7), . . .

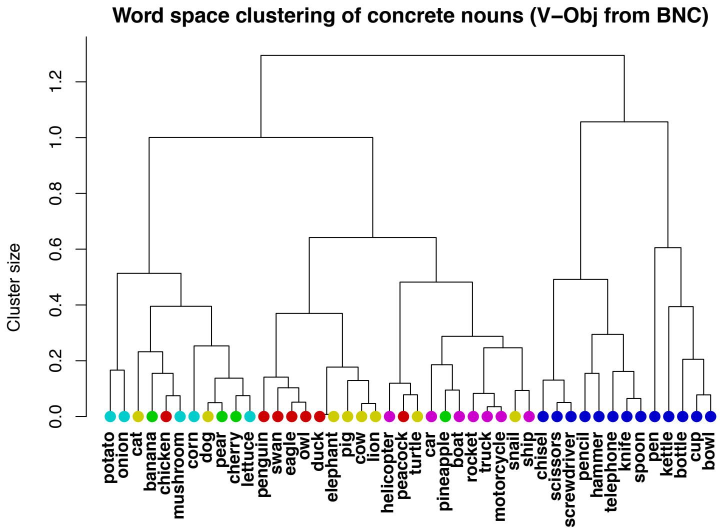
Nearest neighbours



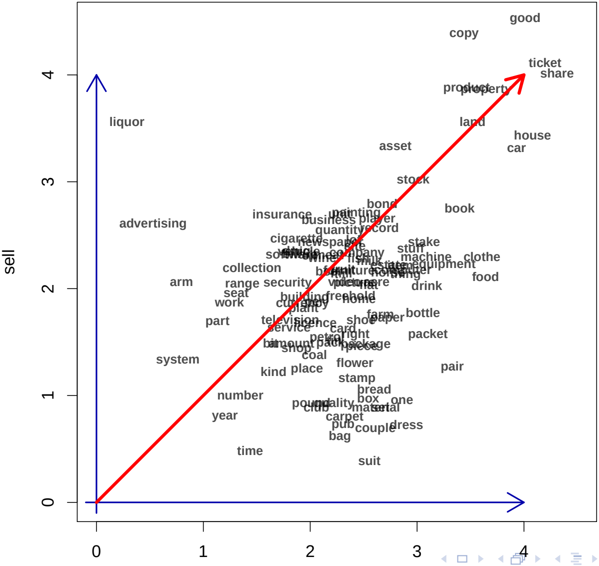
Semantic maps



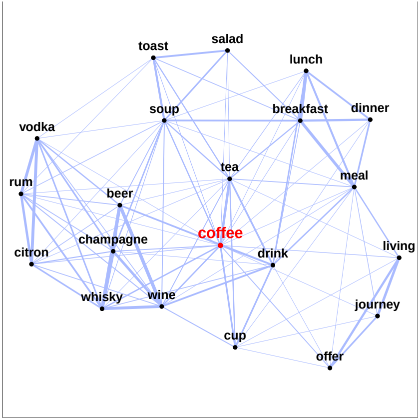
Clustering


Word space clustering of concrete nouns (V–Obj from BNC)
Cluster size
0.0
0.2
0.4
0.6
0.8
1.0
1.2
potato
onion
cat
banana
chicken
mushroom
corn
dog
pear
cherry
lettuce
penguin
swan
eagle
owl
duck
elephant
pig
cow
lion
helicopter
peacock
turtle
car
pineapple
boat
rocket
truck
motorcycle
snail
ship
chisel
scissors
screwdriver
pencil
hammer
telephone
knife
spoon
pen
kettle
bottle
cup
bowl

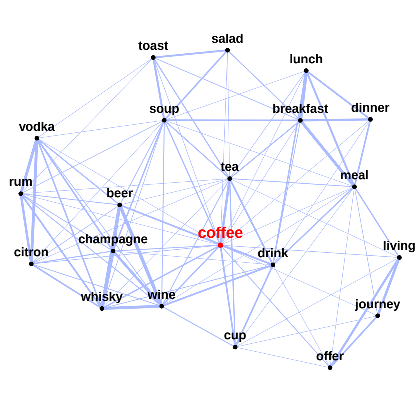
Latent dimensions

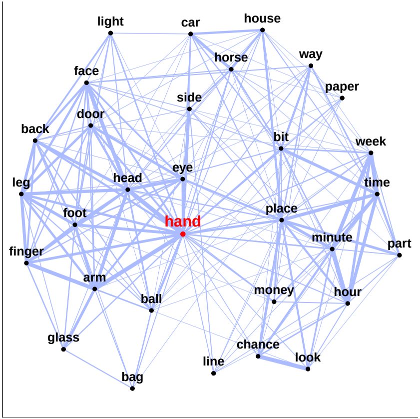


Semantic similarity graph (topological structure)



Semantic similarity graph (topological structure)





Context vectors (Schütze 1998)

Distributional representation only at type level

- ☞ What is the “average” meaning of *mouse*? (computer vs. animal)

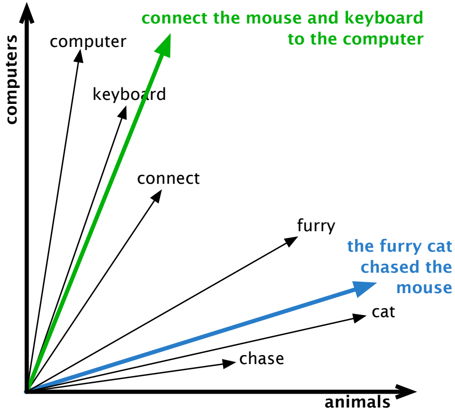
Context vectors (Schütze 1998)

Distributional representation only at type level

☞ What is the "average" meaning of *mouse*? (computer vs. animal)

Context vector approximates meaning of individual token

- **bag-of-words** approach: centroid of all context words in the sentence



Outline

Introduction
- The distributional hypothesis
- Three famous examples

Distributional semantic models
- Definition & overview
- Using DSM distances
- Quantitative evaluation
- Software and further information

The TOEFL synonym task

- The TOEFL dataset
  - 80 items
  - Target: *levied*
    Candidates: *believed*, *correlated*, *imposed*, *requested*

The TOEFL synonym task

- The TOEFL dataset
  - 80 items
  - Target: *levied*
    Candidates: *believed*, *correlated*, *imposed*, *requested*

The TOEFL synonym task

- The TOEFL dataset
  - 80 items
  - Target: *levied*
    Candidates: *believed*, *correlated*, *imposed*, *requested*
  - Target *fashion*
    Candidates: *craze*, *fathom*, *manner*, *ration*

The TOEFL synonym task

- The TOEFL dataset
  - 80 items
  - Target: *levied*
    Candidates: *believed*, *correlated*, *imposed*, *requested*
  - Target *fashion*
    Candidates: *craze*, *fathom*, *manner*, *ration*

The TOEFL synonym task

- The TOEFL dataset
  - 80 items
  - Target: *levied*
    Candidates: *believed*, *correlated*, *imposed*, *requested*
  - Target *fashion*
    Candidates: *craze*, *fathom*, *manner*, *ration*

- DSMs and TOEFL
  1. take vectors of the target ($\mathbf{t}$) and of the candidates ($\mathbf{c}_1 \ldots \mathbf{c}_n$)
  2. measure the distance between $\mathbf{t}$ and $\mathbf{c}_i$, with $1 \leq i \leq n$
  3. select $\mathbf{c}_i$ with the shortest distance in space from $\mathbf{t}$

Humans vs. machines on the TOEFL task

- Average foreign test taker: 64.5%

Humans vs. machines on the TOEFL task

- Average foreign test taker: 64.5%
- Macquarie University staff (Rapp 2004):
  - Average of 5 non-natives: 86.75%
  - Average of 5 natives: 97.75%

Humans vs. machines on the TOEFL task

- Average foreign test taker: 64.5%
- Macquarie University staff (Rapp 2004):
  - Average of 5 non-natives: 86.75%
  - Average of 5 natives: 97.75%
- Distributional semantics
  - Classic LSA (Landauer and Dumais 1997): 64.4%
  - Padó and Lapata's (2007) dependency-based model: 73.0%
  - Distributional memory (Baroni and Lenci 2010): 76.9%
  - Rapp's (2004) SVD-based model, lemmatized BNC: 92.5%
  - Bullinaria and Levy (2012) carry out aggressive parameter optimization: 100.0%

Semantic similarity judgments

- Rubenstein and Goodenough (1965) collected similarity ratings for 65 noun pairs from 51 subjects on a 0–4 scale

w_1	w_2	avg. rating
car	*automobile*	3.9
food	*fruit*	2.7
cord	*smile*	0.0

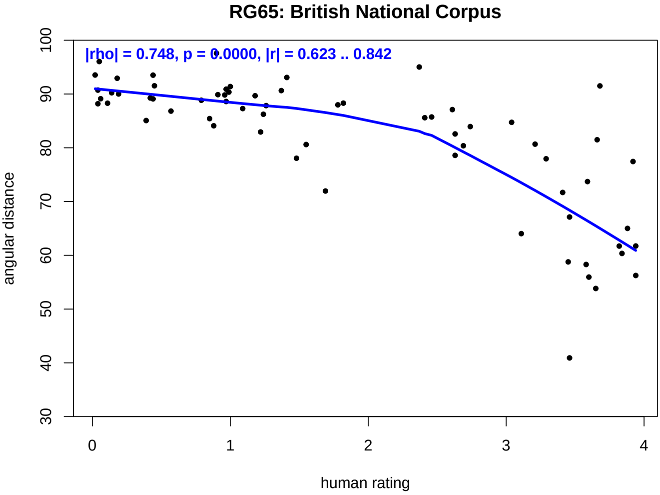
Semantic similarity judgments

- Rubenstein and Goodenough (1965) collected similarity ratings for 65 noun pairs from 51 subjects on a 0–4 scale

w_1	w_2	avg. rating
car	*automobile*	3.9
food	*fruit*	2.7
cord	*smile*	0.0

- DSMs vs. Rubenstein & Goodenough
  1. for each test pair (w_1, w_2), take vectors $\mathbf{w}_1$ and $\mathbf{w}_2$
  2. measure the distance (e.g. cosine) between $\mathbf{w}_1$ and $\mathbf{w}_2$
  3. measure (Pearson) correlation between vector distances and R&G average judgments (Padó and Lapata 2007)

Semantic similarity judgments: example



Semantic similarity judgments: results

Results on RG65 task:

- Padó and Lapata's (2007) dependency-based model: 0.62
- Dependency-based on Web corpus (Herdağdelen *et al.* 2009)
  - without SVD reduction: 0.69
  - with SVD reduction: 0.80
- Distributional memory (Baroni and Lenci 2010): 0.82
- Salient Semantic Analysis (Hassan and Mihalcea 2011): 0.86

Outline

Introduction
The distributional hypothesis
Three famous examples

Distributional semantic models
Definition & overview
Using DSM distances
Quantitative evaluation
Software and further information

Software packages

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

click on package name to open Web page

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

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)

References I

Baroni, Marco and Lenci, Alessandro (2010). Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4), 673–712.

Bengio, Yoshua; Ducharme, Réjean; Vincent, Pascal; Jauvin, Christian (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, **3**, 1137–1155.

Bullinaria, John A. and Levy, Joseph P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, **44**(3), 890–907.

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.

Evert, Stefan (2008). Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, chapter 58, pages 1212–1248. Mouton de Gruyter, Berlin, New York.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, pages 1–32. The Philological Society, Oxford.

References II

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.

Hassan, Samer and Mihalcea, Rada (2011). Semantic relatedness using salient semantic analysis. In *Proceedings of the Twenty-fifth AAAI Conference on Artificial Intelligence*.

Herdağdelen, Amaç; Erk, Katrin; Baroni, Marco (2009). Measuring semantic relatedness with vector space models and random walks. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 50–53, Suntec, Singapore.

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-occurences. In E. V. Clark (ed.), *The Proceedings of the Thirtieth Annual Child Language Research Forum*, pages 167–178. Stanford Linguistics Association.

References III

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.

Padó, Sebastian and Lapata, Mirella (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, **33**(2), 161–199.

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.

References IV

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.

Rubenstein, Herbert and Goodenough, John B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, **8**(10), 627–633.

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.

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.