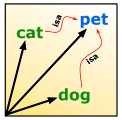
Distributional Semantic Models

Part 1: Introduction

Stefan Evert[1]
with contributions from Marco Baroni[2] and Alessandro Lenci[3]

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

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



Outline

Outline

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**"?

- 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, of course. But in a corpus like the BNC, you will find at least as many informative sentences.

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

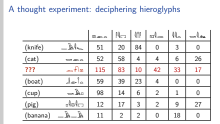
modifier	72	1.2
passable	5	9.92
ready-to-drink	1	8.79
cinnamon-scented	1	8.79
rust-coloured	1	8.57
Tanners	1	8.51
ten-man	1	8.43
in-flight	1	7.99
full-bodied	1	7.87
Smedley	1	7.83
blood-red	1	7.75

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

2014-09-16

DSM Tutorial – Part 1
└Introduction
  └The distributional hypothesis
    └A thought experiment: deciphering hieroglyphs



1. 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

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

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

sim(,) = 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

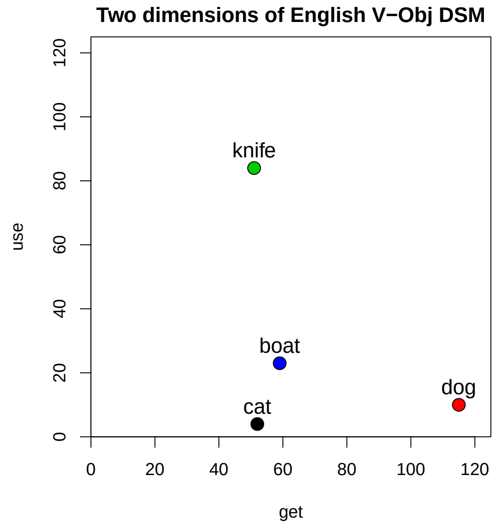
- row vector $\mathbf{x}_{\text{dog}}$ describes usage of word *dog* in the corpus
- can be seen as coordinates of point in n-dimensional Euclidean space

	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 M

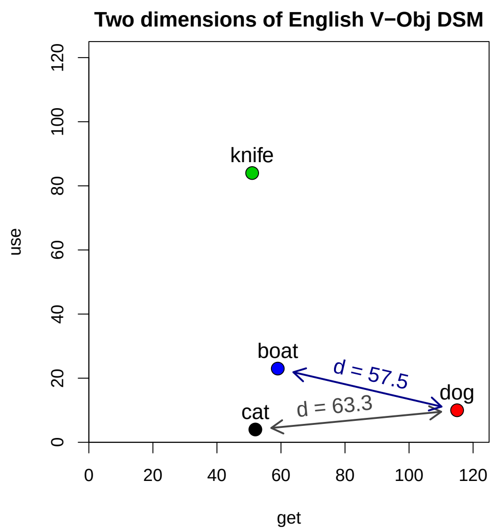
Geometric interpretation

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



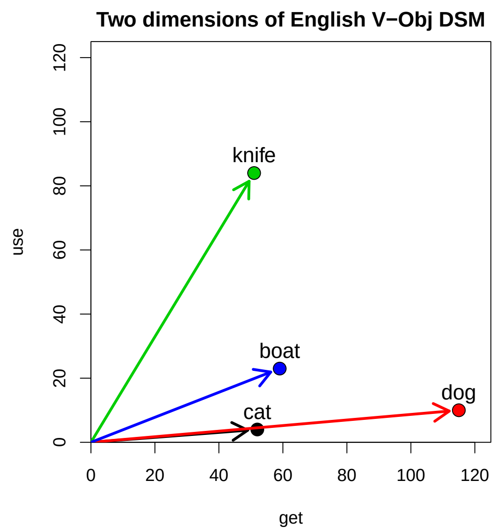
Geometric interpretation

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



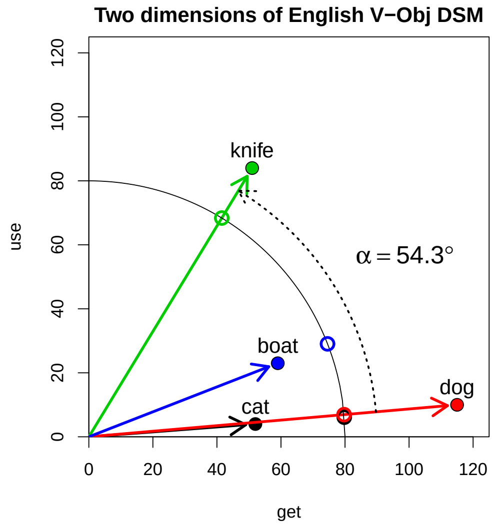
Geometric interpretation

- 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



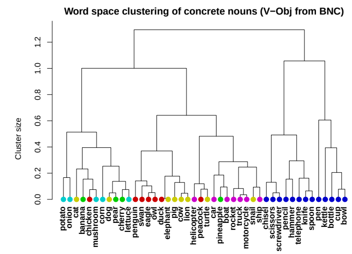
Geometric interpretation

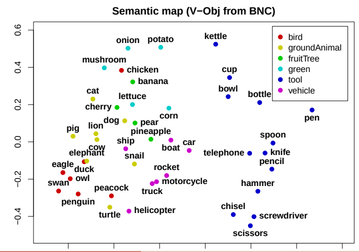
- similarity = spatial proximity (Euclidean dist.)
- location depends on frequency of noun ($f_{\text{dog}} \approx 2.7 \cdot f_{\text{cat}}$)
- direction more important than location
- normalise "length" $\|\mathbf{x}_{\text{dog}}\|$ of vector
- or use angle α as distance measure



Semantic distances

- main result of distributional analysis are "semantic" distances between words
- typical 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
  - detection of multiword expressions





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 1998a; 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

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

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

Term-context vs. term-term matrix
⇓
Definition of terms & linguistic pre-processing
⇓
Size & type of context
⇓
Geometric vs. probabilistic interpretation
⇓
Feature scaling
⇓
Normalisation of rows and/or columns
⇓
Similarity / distance measure
⇓
Dimensionality reduction

Outline

Overview of DSM parameters

Term-context vs. term-term matrix

⇓

Definition of terms & linguistic pre-processing

⇓

Size & type of context

⇓

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 contxt** (specific syntagmatic relation)
- 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

Overview of DSM parameters

Term-context vs. term-term matrix

⇓

Definition of terms & linguistic pre-processing

⇓

Size & type of context

⇓

Geometric vs. probabilistic interpretation

⇓

Feature scaling

⇓

Normalisation of rows and/or columns

⇓

Similarity / distance measure

⇓

Dimensionality reduction

Corpus pre-processing

- Minimally, corpus must be tokenised ➜ identify terms
- Linguistic annotation
  - part-of-speech tagging
  - lemmatisation / stemming
  - word sense disambiguation (rare)
  - shallow syntactic patterns
  - dependency parsing
- Generalisation of terms
  - often lemmatised to reduce data sparseness: *go, goes, went, gone, going* ➜ *go*
  - POS disambiguation (*light*/N vs. *light*/A vs. *light*/V)
  - word sense disambiguation ($bank_{\text{river}}$ vs. $bank_{\text{finance}}$)
- Trade-off between deeper linguistic analysis and
  - need for language-specific resources
  - possible errors introduced at each stage of the analysis

Effects of pre-processing

Nearest neighbours of *walk* (BNC)

word forms

- stroll
- walking
- walked
- go
- path
- drive
- ride
- wander
- sprinted
- sauntered

lemmatised corpus

- hurry
- stroll
- stride
- trudge
- amble
- wander
- walk-nn
- walking
- retrace
- scuttle

Effects of pre-processing

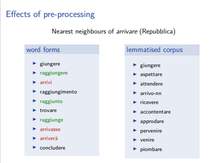
Nearest neighbours of *arrivare* (Repubblica)

word forms

- giungere
- raggiungere
- arrivi
- raggiungimento
- raggiunto
- trovare
- raggiunge
- arrivasse
- arriverà
- concludere

lemmatised corpus

- giungere
- aspettare
- attendere
- arrivo-nn
- ricevere
- accontentare
- approdare
- pervenire
- venire
- piombare



1. Colours seem to indicate inflected forms belonging to the same lemma.

Overview of DSM parameters

Term-context vs. term-term matrix

⇓

Definition of terms & linguistic pre-processing

⇓

Size & type of context

⇓

Geometric vs. probabilistic interpretation

⇓

Feature scaling

⇓

Normalisation of rows and/or columns

⇓

Similarity / distance measure

⇓

Dimensionality reduction

Surface context

Context term occurs within a window of *k* words around target.

The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

Parameters:

- window size (in words or characters)
- symmetric vs. one-sided window
- uniform or "triangular" (distance-based) weighting
- window clamped to sentences or other textual units?

Effect of different window sizes

Nearest neighbours of *dog* (BNC)

2-word window	30-word window
cat	kennel
horse	puppy
fox	pet
pet	bitch
rabbit	terrier
pig	rottweiler
animal	canine
mongrel	cat
sheep	to bark
pigeon	Alsatian

Textual context

Context term is in the same linguistic unit as target.

The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

Parameters:

- type of linguistic unit
  - sentence
  - paragraph
  - turn in a conversation
  - Web page

Syntactic context

Context term is linked to target by a syntactic dependency (e.g. subject, modifier, ...).

The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

Parameters:

- types of syntactic dependency (Padó and Lapata 2007)
- direct vs. indirect dependency paths
  - direct dependencies
  - direct + indirect dependencies
- homogeneous data (e.g. only verb-object) vs. heterogeneous data (e.g. all children and parents of the verb)
- maximal length of dependency path

“Knowledge pattern” context

Context term is linked to target by a lexico-syntactic pattern (text mining, cf. Hearst 1992, Pantel & Pennacchiotti 2008, etc.).

In Provence, Van Gogh painted with bright colors such as red and yellow. These colors produce incredible effects on anybody looking at his paintings.

Parameters:

- inventory of lexical patterns
  - lots of research to identify semantically interesting patterns (cf. Almuhareb & Poesio 2004, Veale & Hao 2008, etc.)
- fixed vs. flexible patterns
  - patterns are mined from large corpora and automatically generalised (optional elements, POS tags or semantic classes)

Structured vs. unstructured context

- In **unstructered** models, context specification acts as a **filter**
  - determines whether context tokens counts as co-occurrence
  - e.g. linked by specific syntactic relation such as verb-object
- In **structured** models, context words are **subtyped**
  - depending on their position in the context
  - e.g. left vs. right context, type of syntactic relation, etc.

Structured vs. unstructured surface context

A dog bites a man. The man’s dog bites a dog. A dog bites a man.

unstructured	bite
dog	4
man	3

A dog bites a man. The man’s dog bites a dog. A dog bites a man.

structured	bite-l	bite-r
dog	3	1
man	1	2

Structured vs. unstructured dependency context

A dog bites a man. The man’s dog bites a dog. A dog bites a man.

unstructured	bite
dog	4
man	2

A dog bites a man. The man’s dog bites a dog. A dog bites a man.

structured	bite-subj	bite-obj
dog	3	1
man	0	2

Comparison

- Unstructured context
  - data less sparse (e.g. *man kills* and *kills man* both map to the *kill* dimension of the vector $\mathbf{x}_{\text{man}}$)
- Structured context
  - more sensitive to semantic distinctions (*kill-subj* and *kill-obj* are rather different things!)
  - dependency relations provide a form of syntactic "typing" of the DSM dimensions (the "subject" dimensions, the "recipient" dimensions, etc.)
  - important to account for word-order and compositionality

Overview of DSM parameters

Term-context vs. term-term matrix
⇓
Definition of terms & linguistic pre-processing
⇓
Size & type of context
⇓
Geometric vs. probabilistic interpretation
⇓
Feature scaling
⇓
Normalisation of rows and/or columns
⇓
Similarity / distance measure
⇓
Dimensionality reduction

Geometric vs. probabilistic interpretation

- Geometric interpretation
  - row vectors as points or arrows in n-dim. space
  - very intuitive, good for visualisation
  - use techniques from geometry and linear algebra
- Probabilistic interpretation
  - co-occurrence matrix as observed sample statistic
  - "explained" by generative probabilistic model
  - recent work focuses on hierarchical Bayesian models
  - probabilistic LSA (Hoffmann 1999), Latent Semantic Clustering (Rooth *et al.* 1999), Latent Dirichlet Allocation (Blei *et al.* 2003), etc.
  - explicitly accounts for random variation of frequency counts
  - intuitive and plausible as topic model

☞ focus on geometric interpretation in this tutorial

Overview of DSM parameters

Term-context vs. term-term matrix
⇓
Definition of terms & linguistic pre-processing
⇓
Size & type of context
⇓
Geometric vs. probabilistic interpretation
⇓
Feature scaling
⇓
Normalisation of rows and/or columns
⇓
Similarity / distance measure
⇓
Dimensionality reduction

Feature scaling

Feature scaling is used to "discount" less important features:

- Logarithmic scaling: $x' = \log(x+1)$
  (cf. Weber-Fechner law for human perception)
- Relevance weighting, e.g. tf.idf (information retrieval)
- Statistical **association measures** (Evert 2004, 2008) take frequency of target word and context feature into account
  - the less frequent the target word and (more importantly) the context feature are, the higher the weight given to their observed co-occurrence count should be (because their expected chance co-occurrence frequency is low)
  - different measures – e.g., mutual information, log-likelihood ratio – differ in how they balance observed and expected co-occurrence frequencies

Association measures: Mutual Information (MI)

word_1	word_2	f_{obs}	f_1	f_2
dog	*small*	855	33,338	490,580
dog	*domesticated*	29	33,338	918

Expected co-occurrence frequency:

$$f_{\text{exp}} = \frac{f_1 \cdot f_2}{N}$$

Mutual Information compares observed vs. expected frequency:

$$\text{MI}(w_1, w_2) = \log_2 \frac{f_{\text{obs}}}{f_{\text{exp}}} = \log_2 \frac{N \cdot f_{\text{obs}}}{f_1 \cdot f_2}$$

Disadvantage: MI overrates combinations of rare terms.

Other association measures

word_1	word_2	f_{obs}	f_{exp}	MI	local-MI	t-score
dog	*small*	855	134.34	2.67	2282.88	24.64
dog	*domesticated*	29	0.25	6.85	198.76	5.34
dog	*sgjkj*	1	0.00027	11.85	11.85	1.00

The log-likelihood ratio (Dunning 1993) has more complex form, but its "core" is known as local MI (Evert 2004).

$$\text{local-MI}(w_1, w_2) = f_{\text{obs}} \cdot \text{MI}(w_1, w_2)$$

The t-score measure (Church and Hanks 1990) is popular in lexicography:

$$\text{t-score}(w_1, w_2) = \frac{f_{\text{obs}} - f_{\text{exp}}}{\sqrt{f_{\text{obs}}}}$$

Details & many more measures: `http://www.collocations.de/`

Overview of DSM parameters

Term-context vs. term-term matrix

⇓

Definition of terms & linguistic pre-processing

⇓

Size & type of context

⇓

Geometric vs. probabilistic interpretation

⇓

Feature scaling

⇓

Normalisation of rows and/or columns

⇓

Similarity / distance measure

⇓

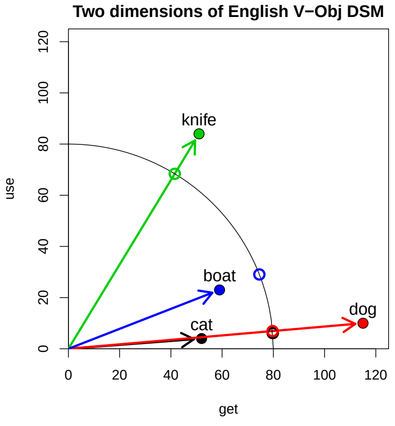
Dimensionality reduction

Normalisation of row vectors

- geometric distances only make sense if vectors are normalised to unit length
- divide vector by its length:

$$\mathbf{x}/\|\mathbf{x}\|$$

- normalisation depends on distance measure!
- special case: scale to relative frequencies with $\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n|$ ➜ probabilistic interpretation



Scaling of column vectors

- In statistical analysis and machine learning, features are usually centred and scaled so that

$$\text{mean} \quad \mu = 0$$
$$\text{variance} \quad \sigma^2 = 1$$

- In DSM research, this step is less common for columns of $\mathbf{M}$
  - centring is a prerequisite for certain dimensionality reduction and data analysis techniques (esp. PCA)
  - scaling may give too much weight to rare features
- $\mathbf{M}$ cannot be row-normalised and column-scaled at the same time (result depends on ordering of the two steps)

Overview of DSM parameters

Term-context vs. term-term matrix

⇓

Definition of terms & linguistic pre-processing

⇓

Size & type of context

⇓

Geometric vs. probabilistic interpretation

⇓

Feature scaling

⇓

Normalisation of rows and/or columns

⇓

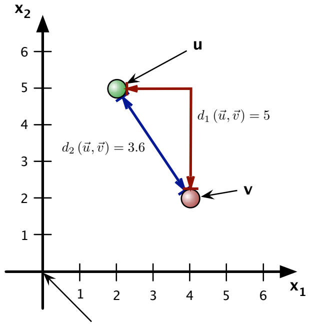
Similarity / distance measure

⇓

Dimensionality reduction

Geometric distance

- **Distance** between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ ➜ (dis)**similarity**
  - $\mathbf{u} = (u_1, \ldots, u_n)$
  - $\mathbf{v} = (v_1, \ldots, v_n)$
- **Euclidean** distance $d_2(\mathbf{u}, \mathbf{v})$
- "City block" **Manhattan** distance $d_1(\mathbf{u}, \mathbf{v})$
- Both are special cases of the **Minkowski** p-distance $d_p(\mathbf{u}, \mathbf{v})$ (for $p \in [1, \infty]$)



$$d_p(\mathbf{u}, \mathbf{v}) := \left(|u_1 - v_1|^p + \cdots + |u_n - v_n|^p\right)^{1/p}$$

$$d_\infty(\mathbf{u}, \mathbf{v}) = \max\{|u_1 - v_1|, \ldots, |u_n - v_n|\}$$

Other distance measures

- Information theory: **Kullback-Leibler** (KL) **divergence** for probability vectors (non-negative, $\|\mathbf{x}\|_1 = 1$)

$$D(\mathbf{u}\|\mathbf{v}) = \sum_{i=1}^{n} u_i \cdot \log_2 \frac{u_i}{v_i}$$

- Properties of KL divergence
  - most appropriate in a probabilistic interpretation of $\mathbf{M}$
  - zeroes in $\mathbf{v}$ without corresponding zeroes in $\mathbf{u}$ are problematic
  - not symmetric, unlike geometric distance measures
  - alternatives: skew divergence, Jensen-Shannon divergence
- A symmetric distance measure (Endres and Schindelin 2003)

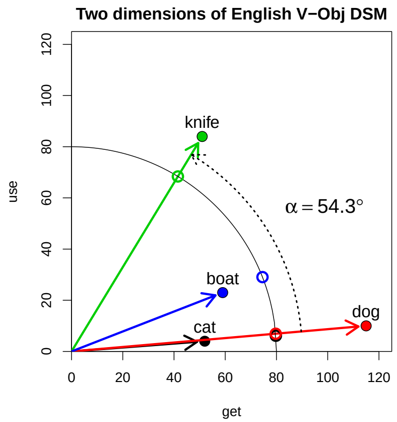
$$D_{\mathbf{uv}} = D(\mathbf{u}\|\mathbf{z}) + D(\mathbf{v}\|\mathbf{z}) \quad \text{with} \quad \mathbf{z} = \frac{\mathbf{u}+\mathbf{v}}{2}$$

Similarity measures

- angle α between two vectors $\mathbf{u}, \mathbf{v}$ is given by

$$\cos\alpha = \frac{\sum_{i=1}^{n} u_i \cdot v_i}{\sqrt{\sum_i u_i^2} \cdot \sqrt{\sum_i v_i^2}} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}$$

- **cosine** measure of similarity: $\cos\alpha$
  - $\cos\alpha = 1$ ➜ collinear
  - $\cos\alpha = 0$ ➜ orthogonal



Overview of DSM parameters

Term-context vs. term-term matrix

⇓

Definition of terms & linguistic pre-processing

⇓

Size & type of context

⇓

Geometric vs. probabilistic interpretation

⇓

Feature scaling

⇓

Normalisation of rows and/or columns

⇓

Similarity / distance measure

⇓

Dimensionality reduction

Dimensionality reduction = model compression

- Co-occurrence matrix $\mathbf{M}$ is often unmanageably large and can be extremely sparse
  - Google Web1T5: 1M × 1M matrix with one trillion cells, of which less than 0.05% contain nonzero counts (Evert 2010)

➥ Compress matrix by reducing dimensionality (= rows)

- **Feature selection**: columns with high frequency & variance
  - measured by entropy, chi-squared test, ...
  - may select correlated (➜ uninformative) dimensions
  - joint selection of multiple features is useful but expensive
- **Projection** into (linear) subspace
  - principal component analysis (PCA)
  - independent component analysis (ICA)
  - random indexing (RI)
  - ☞ intuition: preserve distances between data points

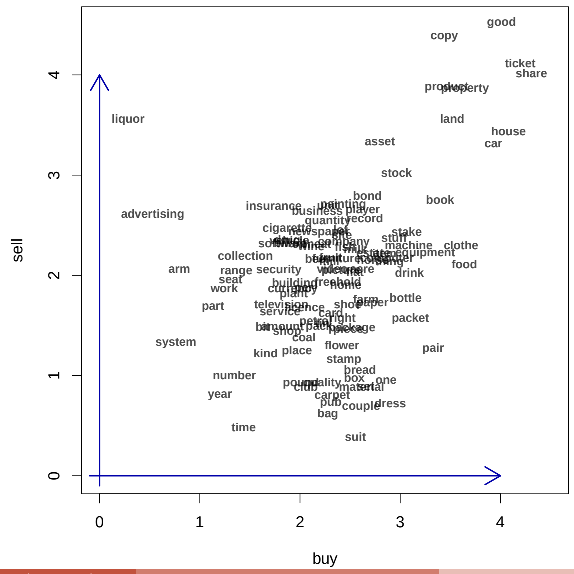
Dimensionality reduction & latent dimensions

Landauer and Dumais (1997) claim that LSA dimensionality reduction (and related PCA technique) uncovers **latent dimensions** by exploiting correlations between features.

- Example: term-term matrix
- V-Obj cooc's extracted from BNC
  - targets = noun lemmas
  - features = verb lemmas
- feature scaling: association scores (modified log Dice coefficient)
- $k = 111$ nouns with $f \geq 20$ (must have non-zero row vectors)
- $n = 2$ dimensions: *buy* and *sell*

noun	*buy*	*sell*
bond	0.28	0.77
cigarette	-0.52	0.44
dress	0.51	-1.30
freehold	-0.01	-0.08
land	1.13	1.54
number	-1.05	-1.02
per	-0.35	-0.16
pub	-0.08	-1.30
share	1.92	1.99
system	-1.63	-0.70

Dimensionality reduction & latent dimensions

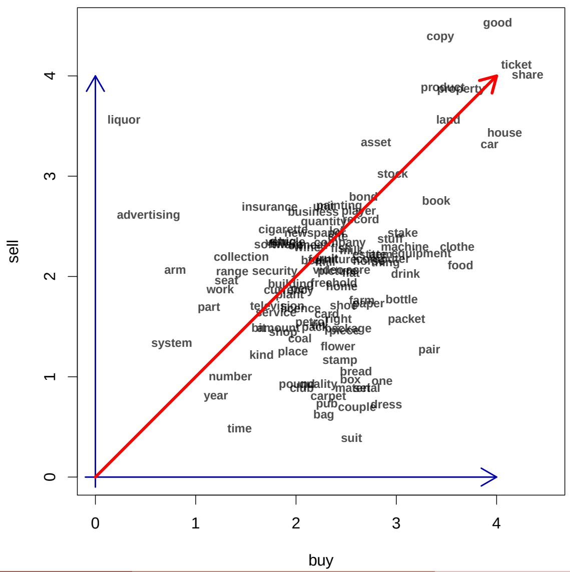


Motivating latent dimensions & subspace projection

- The **latent property** of being a commodity is "expressed" through associations with several verbs: *sell*, *buy*, *acquire*, …
- Consequence: these DSM dimensions will be **correlated**

- Identify **latent dimension** by looking for strong correlations (or weaker correlations between large sets of features)
- Projection into subspace V of $k < n$ latent dimensions as a "**noise reduction**" technique ➜ **LSA**
- Assumptions of this approach:
  - "latent" distances in V are semantically meaningful
  - other "residual" dimensions represent chance co-occurrence patterns, often particular to the corpus underlying the DSM

The latent "commodity" dimension



Outline

Some well-known DSM examples

Latent Semantic Analysis (Landauer and Dumais 1997)

- term-context matrix with document context
- weighting: log term frequency and term entropy
- distance measure: cosine
- dimensionality reduction: SVD

Hyperspace Analogue to Language (Lund and Burgess 1996)

- term-term matrix with surface context
- structured (left/right) and distance-weighted frequency counts
- distance measure: Minkowski metric ($1 \leq p \leq 2$)
- dimensionality reduction: feature selection (high variance)

Some well-known DSM examples

Infomap NLP (Widdows 2004)

- term-term matrix with unstructured surface context
- weighting: none
- distance measure: cosine
- dimensionality reduction: SVD

Random Indexing (Karlgren and Sahlgren 2001)

- term-term matrix with unstructured surface context
- weighting: various methods
- distance measure: various methods
- dimensionality reduction: random indexing (RI)

Some well-known DSM examples

Dependency Vectors (Padó and Lapata 2007)

- term-term matrix with unstructured dependency context
- weighting: log-likelihood ratio
- distance measure: information-theoretic (Lin 1998b)
- dimensionality reduction: none

Distributional Memory (Baroni and Lenci 2010)

- term-term matrix with structured and unstructered dependencies + knowledge patterns
- weighting: local-MI on type frequencies of link patterns
- distance measure: cosine
- dimensionality reduction: none

Outline

Nearest neighbours

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

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

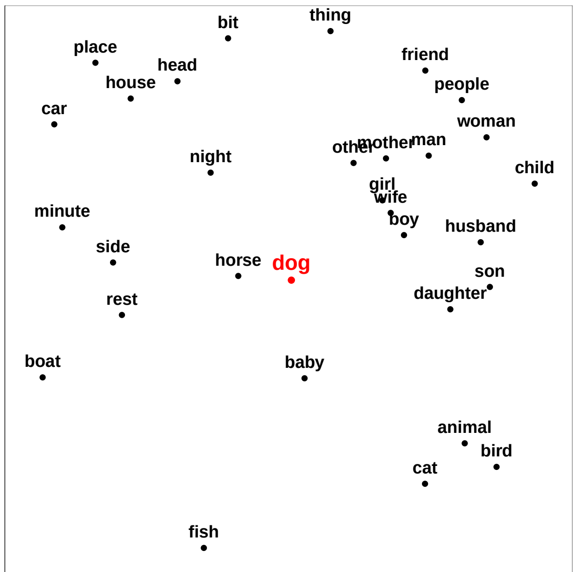
☞ girl (45.5), boy (46.7), horse(47.0), wife (48.8), baby (51.9), daughter (53.1), side (54.9), mother (55.6), boat (55.7), rest (56.3), night (56.7), cat (56.8), son (57.0), man (58.2), place (58.4), husband (58.5), thing (58.8), friend (59.6), ...

Neighbours of **school**:

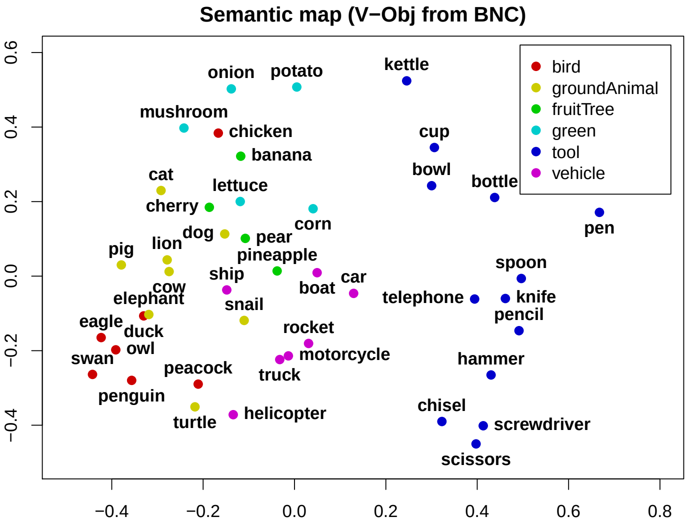
☞ country (49.3), church (52.1), hospital (53.1), house (54.4), hotel (55.1), industry (57.0), company (57.0), home (57.7), family (58.4), university (59.0), party (59.4), group (59.5), building (59.8), market (60.3), bank (60.4), business (60.9), area (61.4), department (61.6), club (62.7), town (63.3),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

1. Neighbours and neighbourhood plots from BNC verb-object DSM, reduced to 100 dimensions by SVD.

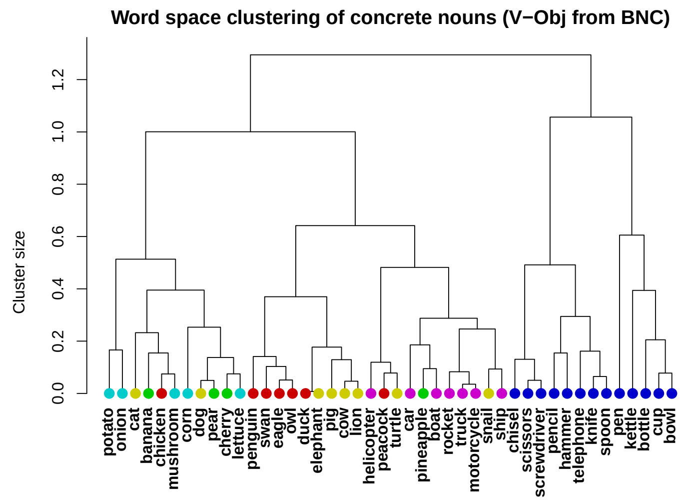
Nearest neighbours



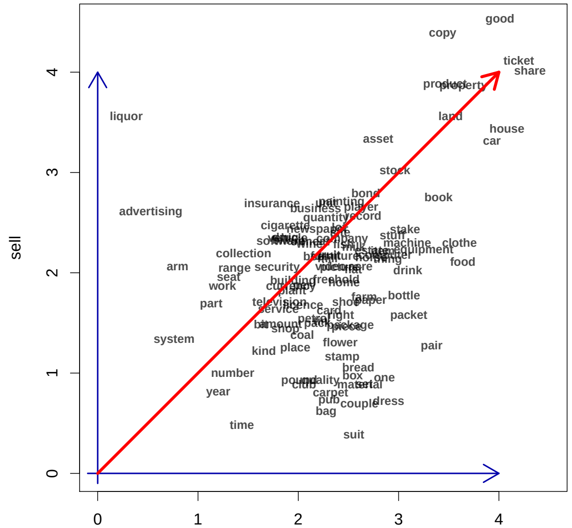
Semantic maps



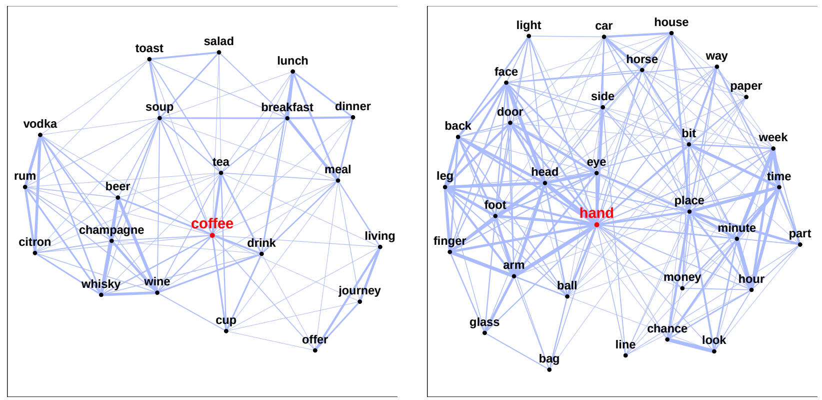
Clustering



Latent dimensions



Semantic similarity graph (topological structure)



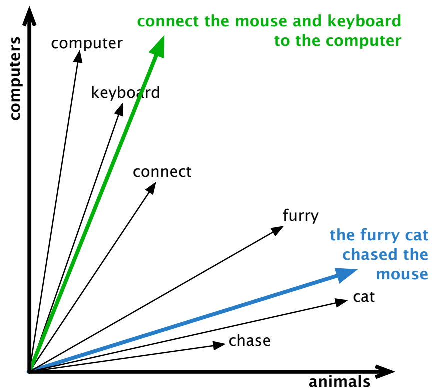
Context vectors (Schütze 1998)

Distributional representation only at type level

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

Context vector approximates meaning of individual token

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



Outline

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 ($\mathbf{c}_1 \dots \mathbf{c}_n$)
  2. measure the distance between $\mathbf{t}$ and $\mathbf{c}_i$, with $1 \le i \le n$
  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%
- 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%

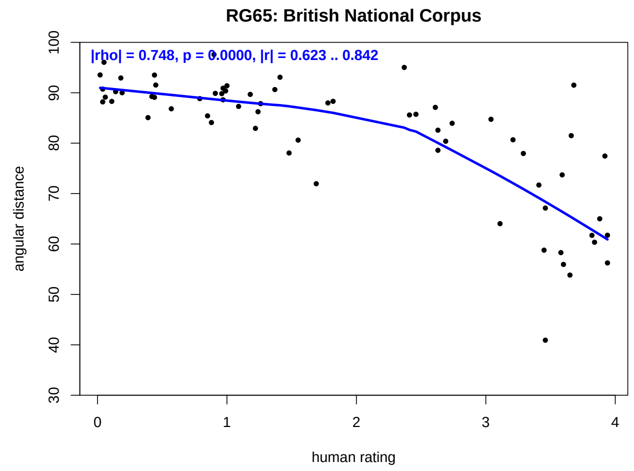
Semantic similarity judgments

- 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

- 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

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

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

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

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.

Blei, David M.; Ng, Andrew Y.; Jordan, Michael, I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

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.

Church, Kenneth W. and Hanks, Patrick (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1), 22–29.

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.

Dunning, Ted E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

References II

Endres, Dominik M. and Schindelin, Johannes E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, **49**(7), 1858–1860.

Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from `http://www.collocations.de/phd.html`.

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

Evert, Stefan (2010). Google Web 1T5 n-grams made easy (but not for the computer). In *Proceedings of the 6th Web as Corpus Workshop (WAC-6)*, pages 32–40, Los Angeles, CA.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, pages 1–32. The Philological Society, Oxford. Reprinted in Palmer (1968), pages 168–205.

Grefenstette, Gregory (1994). *Explorations in Automatic Thesaurus Discovery*, volume 278 of *Kluwer International Series in Engineering and Computer Science*. Springer, Berlin, New York.

References III

Harris, Zellig (1954). Distributional structure. *Word*, **10**(23), 146–162. Reprinted in Harris (1970, 775–794).

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.

Hoffmann, Thomas (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*.

Karlgren, Jussi and Sahlgren, Magnus (2001). From words to understanding. In Y. Uesaka, P. Kanerva, and H. Asoh (eds.), *Foundations of Real-World Intelligence*, chapter 294–308. CSLI Publications, Stanford.

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 IV

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.

Lin, Dekang (1998a). 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.

Lin, Dekang (1998b). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML-98)*, pages 296–304, Madison, WI.

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.

References V

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 (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.

Rooth, Mats; Riezler, Stefan; Prescher, Detlef; Carroll, Glenn; Beil, Franz (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 104–111.

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.

References VI

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.

Widdows, Dominic (2004). *Geometry and Meaning*. Number 172 in CSLI Lecture Notes. CSLI Publications, Stanford.