Introduction to Distributional Semantics

Aurélie Herbelot

University of Trento/Geneva

Geneva 2016

Distributional semantics

cat

Distributional semantics: a short history

- J.R. Firth: *You shall know a word by the company it keeps* (1957).
- Zelig Harris: *Words that appear in the same contexts are semantically similar* (1954).

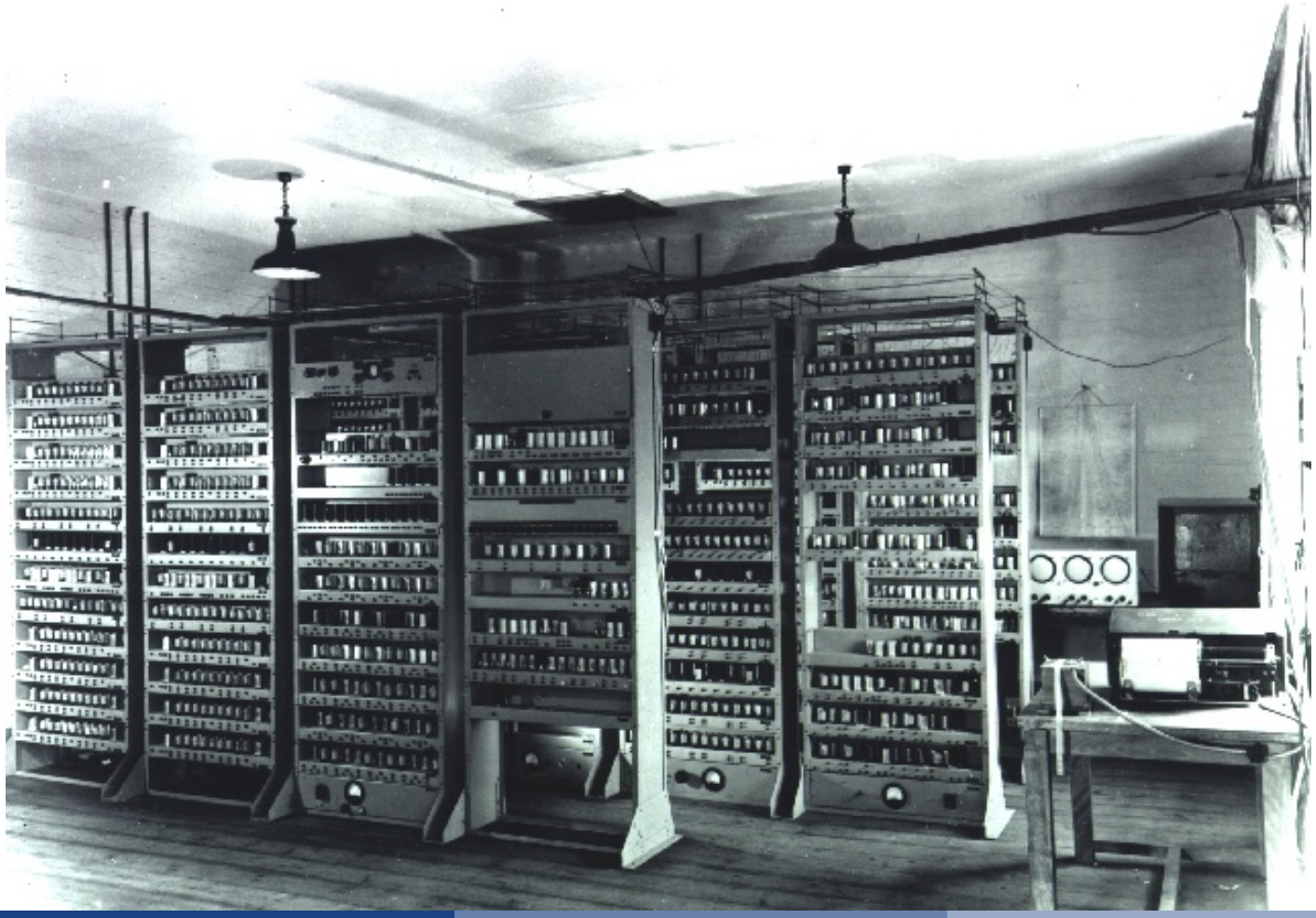
Distributional semantics: a short history

Ludwig Wittgenstein: 'Meaning is use': 'Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache' (Wittgenstein, 1953. 43)

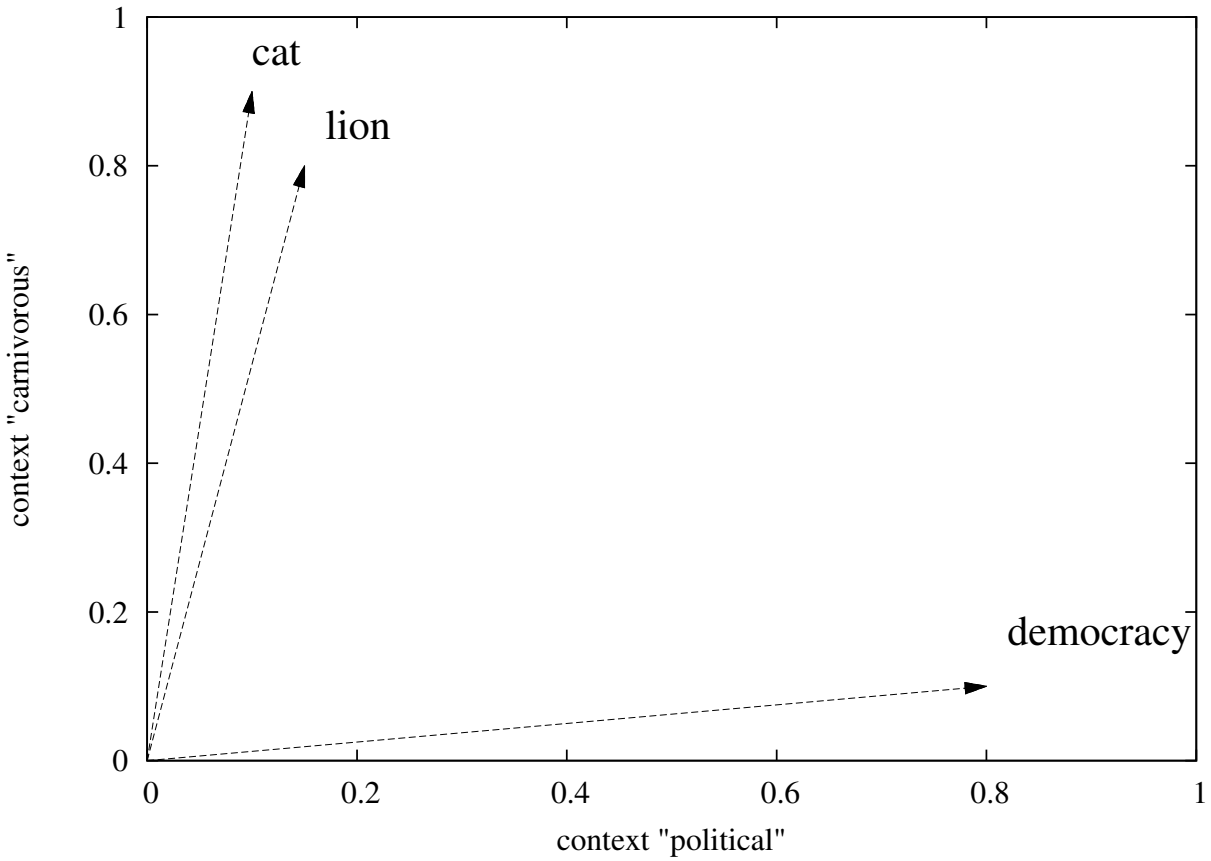
Margaret Masterman: Cambridge Language Research Unit (CLRU: 1955–1986).

Karen Spärck-Jones: Early experiments on distributional semantics: 1963, 1967.

‘The’ computer: the EDSAC



The semantic space



The components of distributional representations

- Contexts: other words in the close vicinity of the target (*eat, mouse, sleep*), or syntactic/semantic relations (*eat(x), chase(x,mouse), like(x,sleep)*).
- Weights: usually a measure of how characteristic the context is for the target (e.g. Pointwise Mutual Information).
- A semantic space: a vector space in which dimensions are the contexts with respect to which the target is expressed. The target word is a vector in that space (vector components are given by the weights of the distribution).

A distributional cat (from the British National Corpus)

0.124 pet-N
0.123 mouse-N
0.099 rat-N
0.097 owner-N
0.096 dog-N
0.092 domestic-A
0.090 wild-A
0.090 duck-N
0.087 tail-N
0.084 leap-V
0.084 prey-N
0.083 breed-N
0.080 rabbit-N
0.078 female-A
0.075 fox-N
0.075 basket-N
0.075 animal-N
0.074 ear-N
0.074 chase-V
0.074 smell-V
0.074 tiger-N
0.073 jump-V
0.073 tom-N
0.073 fat-A
0.071 spell-V
0.071 companion-N
0.070 lion-N
0.068 breed-V
0.068 signal-N
0.067 bite-V
0.067 spring-V
0.067 detect-V
0.067 bird-N
0.066 friendly-A
0.066 odour-N
0.066 hunting-N
0.066 ghost-N
0.065 rub-V
0.064 predator-N
0.063 pig-N
0.063 hate-V
0.063 asleep-A
0.063 stance-N
0.062 unfortunate-A
0.061 naked-A
0.061 switch-V
0.061 encounter-V
0.061 creature-N
0.061 dominant-A
0.060 black-A
0.059 chocolate-N
0.058 giant-N
0.058 sensitive-A
0.058 canadian-A
0.058 toy-N
0.058 milk-N
0.057 human-N
0.057 devil-N
0.056 smell-N
...

Modelling choices

Corpus choice

- As much data as possible?
  - British National Corpus (BNC): 100 m words
  - Wikipedia: 897 m words
  - UKWac: 2 bn words
  - ENCOW: 10 bn words
  - ...
- In general preferable, *but*:
  - More data is not necessarily the data you want.
  - More data is not necessarily realistic from a psycholinguistic point of view. We perhaps encounter 50,000 words a day. BNC = 5 years' text exposure.

The notion of context

- **Context:** if the meaning of a word is given by its context, what does 'context' mean?
  - Word windows (unfiltered): *n* words on either side of the lexical item under consideration (unparsed text).
    Example: n=2 (5 words window):

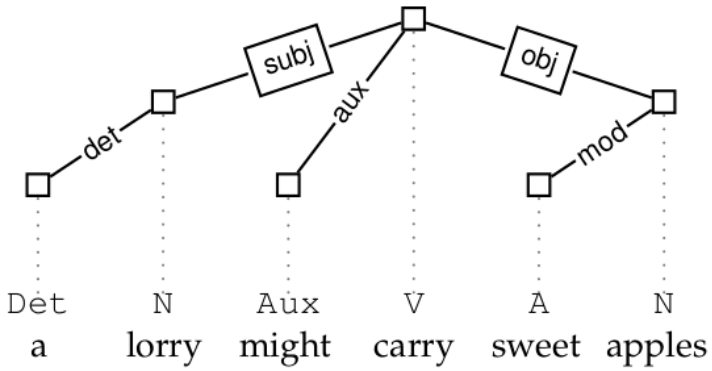
    *... the prime **minister** acknowledged that ...*

  - Word windows (filtered): *n* words on either side of the lexical item under consideration (unparsed text). Some words are not considered part of the context (e.g. function words, some very frequent content words). The stop list for function words is either constructed manually, or the corpus is POS-tagged.
    Example: n=2 (5 words window):

    *... the prime **minister** acknowledged that ...*

The notion of context

- Dependencies: syntactic or semantic. The corpus is converted into a list of directed links between heads and dependents. Context for a lexical item is the dependency structure it belongs to. The length of the dependency path can vary according to the implementation (Padó and Lapata, 2007).




```
a        [Det,det,N]   lorry
lorry    [N,subj,V]    carry
might    [Aux,aux,V]   carry
apples   [N,obj,V]     carry
sweet    [A,mod,N]     apples
```

Parsed vs unparsed data: examples

word (unparsed)	**word (parsed)**
meaning_n	or_c+phrase_n
derive_v	and_c+phrase_n
dictionary_n	syllable_n+of_p
pronounce_v	play_n+on_p
phrase_n	etymology_n+of_p
latin_j	portmanteau_n+of_p
ipa_n	and_c+deed_n
verb_n	meaning_n+of_p
mean_v	from_p+language_n
hebrew_n	pron_rel_+utter_v
usage_n	for_p+word_n
literally_r	in_p+sentence_n

Context weighting

- Binary model: if context *c* co-occurs with word *w*, value of vector $\vec{w}$ for dimension *c* is 1, 0 otherwise.

  ... [a long long long ***example*** *for a distributional semantics] model... (n=4)*

  ... {a 1} {dog 0} {long 1} {sell 0} {semantics 1}...

- Basic frequency model: the value of vector $\vec{w}$ for dimension *c* is the number of times that *c* co-occurs with *w*.

  ... [a long long long ***example*** *for a distributional semantics] model... (n=4)*

  ... {a 2} {dog 0} {long 3} {sell 0} {semantics 1}...

Context weighting

- Characteric model: the weights given to the vector components express how *characteristic* a given context is for w. Functions used include:
  - Pointwise Mutual Information (PMI):

$$pmi_{wc} = \log \frac{p(w,c)}{p(w)p(c)} = log(\frac{f_{wc} * f_{total}}{f_w * f_c}) \quad (1)$$

    where frequencies are defined *over co-occurrences* (i.e. f_w is the sum of all $f_{wc_{i\ldots k}}$ and f_c is the sum of all $f_{w_{1\ldots k}c}$).
  - Derivatives such Positive Pointwise Mutual Information (PPMI), Pointwise Local Mutual Information (PLMI), etc.

What semantic space?

- Entire vocabulary.
  - + All information included – even rare, but important contexts
  - - Inefficient (100,000s dimensions). Noisy (e.g. *002.png|thumb|right|200px|graph_n*)
- Top *n* words with highest frequencies.
  - + More efficient (5000-10000 dimensions). Only 'real' words included.
  - - May miss out on infrequent but relevant contexts.
- Dimensionality reduction (e.g. SVD).

Getting distributions from text

Our reference text

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- **Example:** Produce distributions using a word window, frequency-based model

The semantic space

Douglas Adams, *Mostly harmless*

the_DT major_JJ difference_NN between_IN a_DT thing_NN that_WDT might_MD go_VB wrong_JJ and_CC a_DT thing_NN that_WDT can_MD not_RB possibly_RB go_VB wrong_JJ be_VBZ that_IN when_WRB a_DT thing_NN that_WDT can_MD not_RB possibly_RB go_VB wrong_JJ go_VBZ wrong_JJ it_PRP usually_RB turn_VBZ out_RP to_TO be_VB impossible_JJ to_TO get_VB at_IN or_CC repair_NN

- We assume that we only keep nouns, verbs, adjectives and adverbs in the semantic space.
- **Dimensions:**

go_V	not_R	impossible_J
wrong_J	difference_N	out_R
thing_N	turn_V	repair_V
possibly_R	usually_R	
be_V	major_J	

Frequency counts...

Douglas Adams, *Mostly harmless*

major_J difference_N thing_N go_V wrong_J thing_N not_R possibly_R go_V wrong_J be_V thing_N not_R possibly_R go_V wrong_J go_V wrong_J usually_R turn_V out_R be_V impossible_J get_V repair_N

- **Counts:**

4 go_V	2 not_R	1 impossible_J
4 wrong_J	1 difference_N	1 out_R
3 thing_N	1 turn_V	1 repair_V
2 possibly_R	1 usually_R	
2 be_V	1 major_J	

Conversion into 3-word windows...

Douglas Adams, *Mostly harmless*

major_J difference_N thing_N go_V wrong_J thing_N not_R possibly_R go_V wrong_J be_V thing_N not_R possibly_R go_V wrong_J go_V wrong_J usually_R turn_V out_R be_V impossible_J get_V repair_N

- ∅ **major** difference
- major **difference** thing
- difference **thing** go
- thing **go** wrong
- ...

Distribution for *wrong*

Douglas Adams, *Mostly harmless*

major_J difference_N thing_N [go_V wrong_J thing_N] not_R possibly_R [go_V wrong_J be_V] thing_N not_R possibly_R [go_V wrong_J [go_V] wrong_J usually_R] turn_V out_R be_V impossible_J get_V repair_N

- **Distribution (frequencies):**

5.0 go_V	0.0 possibly_R	0.0 impossible_J
1.0 thing_N	0.0 difference_N	0.0 out_R
1.0 usually_R	0.0 turn_V	0.0 repair_N
1.0 be_V	0.0 get_V	0.0 not_R
0.0 wrong_J	0.0 major_J	

Distribution for *wrong*

Douglas Adams, *Mostly harmless*

major_J difference_N thing_N [go_V wrong_J thing_N] not_R possibly_R [go_V wrong_J be_V] thing_N not_R possibly_R [go_V wrong_J [go_V] wrong_J usually_R] turn_V out_R be_V impossible_J get_V repair_N

- **Distribution (PPMIs):**

0.748490106304 go_V	0.0 possibly_R	0.0 impossible_J
0.6221273278 usually_R	0.0 difference_N	0.0 out_R
0.229608686181 be_V	0.0 turn_V	0.0 repair_N
0.0 thing_N	0.0 get_V	0.0 not_R
0.0 wrong_J	0.0 major_J	

The output of a DS system

- Some ‘row’ labels: the vocabulary of the system.
- Some ‘column’ labels: the contexts (in the case of a dimensionality-reduced space, the reduced dimensions don’t have a label, though).
- The values at the intersection of rows and the columns form a matrix. The values of a row are the vector for a particular lexical item:
  wrong 0.748490106304 0.6221273278 0.229608686181 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

Dimensionality reduction

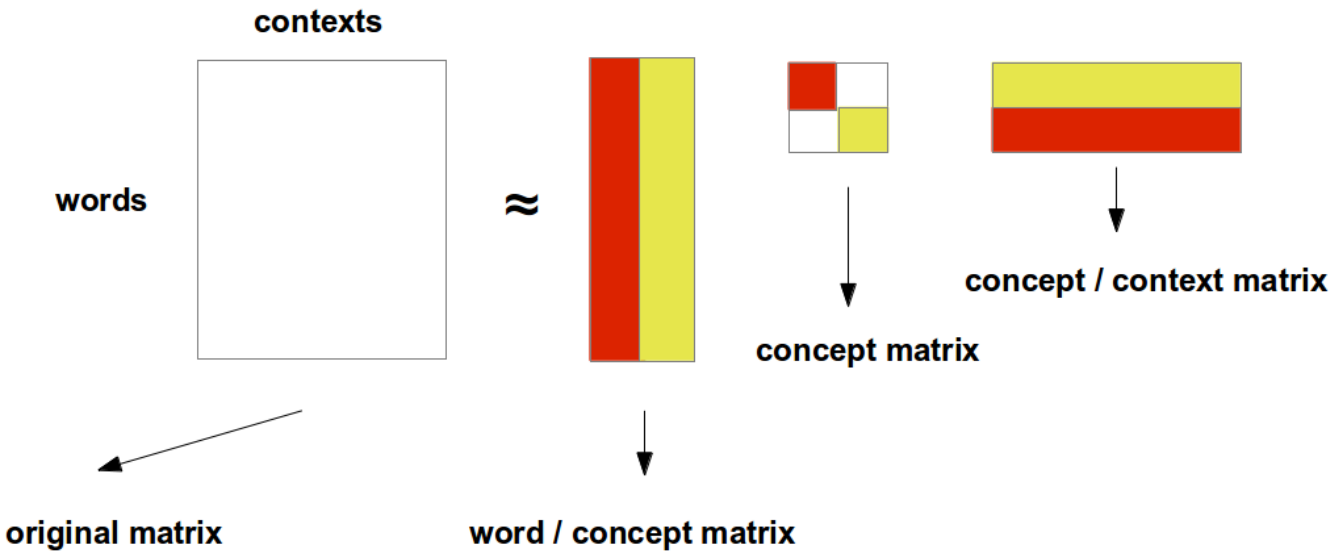
Singular Value Decomposition

- SVD is a matrix factorisation method which expresses a matrix in terms of three other matrices:

$$A = U\Sigma V^T \tag{2}$$

- U and V are orthogonal: they are matrices such that
  - $UU^T = U^T U = I$
  - $VV^T = V^T V = I$
- Σ is a diagonal matrix (only the diagonal entries are non-zero).

Singular Value Decomposition



The SVD derivation

- From our definition, $A = U\Sigma V^T$, it follows that...
- $A^T = V\Sigma^T U^T$
- $A^T A = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T$
  (Recall that $U^T U = I$ because U is orthogonal.)
- $A^T AV = V\Sigma^2$
  (The inverse V^{-1} of an orthogonal matrix is V^T, since $V^T V = I$.)
- Similarly, $AA^T U = U\Sigma^2$.

SVD and eigenvectors

- An *eigenvector* of a linear transformation is a non-zero vector that doesn't change its direction when that linear transformation is applied to it:

$$Av = \lambda v \tag{3}$$

  v is the eigenvector, *lambda* is the eigenvalue.
- Let's consider again the end of our derivation: $A^T AV = V\Sigma^2$.
- The columns of V are the eigenvectors of $A^T A$. (Similarly, the columns of U are the eigenvectors of AA^T.)

The singular values of SVD

- If V contains the eigenvectors of $A^T A$, Σ^2 encapsulates its eigenvalues:
  $A^T AV = V\Sigma^2$.
- Σ itself contains the square roots of the eigenvalues, also known as *singular values*.

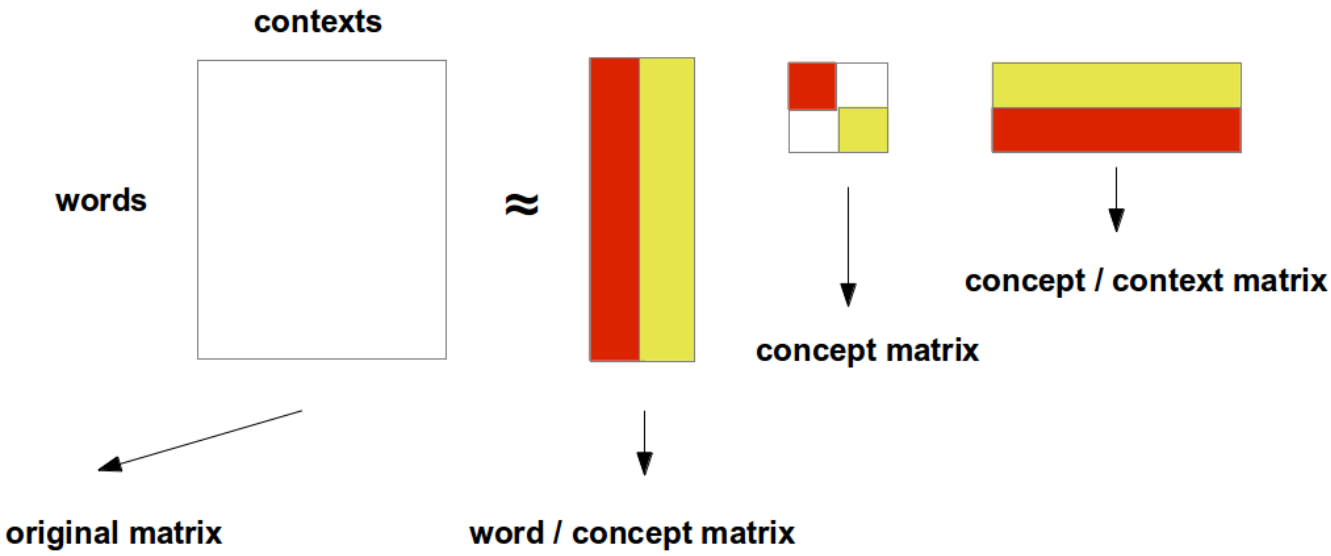
SVD at a glance

- Calculate $A^T A$.
- Calculate the eigenvalues of $A^T A$ and sort them in descending order. Take their square roots to obtain the singular values of $A^T A$ (i.e. the matrix Σ).
- Use the eigenvalues to compute the eigenvectors of $A^T A$. These eigenvectors become the columns of V.
- Compute $U = AV\Sigma^{-1}$.

Finally... dimensionality reduce!

- Now we know the value of U, Σ and V.
- To obtain a reduced representation of A, choose the top k singular values in Σ and multiply the corresponding columns in U by those values.
- Example: the original matrix is 10000 x 5000. The SVD produces:
  - U: 10000 x 10000
  - Σ: 10000 x 5000
  - V: 5000 x 5000
- We can now take the first 300 singular value of Σ, and the corresponding 300 columns of U. We are now multiplying a matrix of 10000 x 300 by a matrix of 300 x 300.

Singular Value Decomposition



What semantic space?

- Singular Value Decomposition (LSA – Landauer and Dumais, 1997). A new dimension might correspond to a generalisation over several of the original dimensions (e.g. the dimensions for *car* and *vehicle* are collapsed into one).
  - + Very efficient (200-500 dimensions). Captures generalisations in the data.
  - - SVD matrices are not straightforwardly interpretable.

SVD for visualisation

Evaluation

Evaluating a semantic space

- How good is your semantic space?
- It depends on what what you want it to be (i.e. which theory of meaning you are supporting.)
- So far, cognitive plausibility has been the main test: can we reproduce human linguistic judgements?

Similarity-based evaluation

- Reproduce human similarity judgements (expressed as a score on a bounded scale).
- Rubenstein & Goodenough (1965): 65 noun pairs.
- Finkelstein et al (2002): WordSim353.
- Bruni et al (2014): MEN (1000 test pairs).
- Calculate spearman correlation (ρ) between system results and human judgements. Current human correlation on the MEN dataset is well over 0.7.

Similarity-based evaluation

Human output

```
sun sunlight 50.000000
automobile car 50.000000
river water 49.000000
stair staircase 49.000000
...
green lantern 18.000000
painting work 18.000000
pigeon round 18.000000
...
muscle tulip 1.000000
bikini pizza 1.000000
bakery zebra 0.000000
```

System output

```
stair staircase 0.913251552368
sun sunlight 0.727390960465
automobile car 0.740681924959
river water 0.501849324363
...
painting work 0.448091435945
green lantern 0.383044261062
...
bakery zebra 0.061804313745
bikini pizza 0.0561356056323
pigeon round 0.028243620524
muscle tulip 0.0142570835367
```

Categorisation

- Cluster concepts into categories: e.g. *cat* and *giraffe* under ANIMAL, *car* and *motorcycle* under VEHICLE (Almuhareb 2006)
- Evaluated in terms of 'purity': if all the concepts in one automatically-produced cluster are from the same category, purity is 100%.

Categorisation and purity

- Given clustered data, purity is defined as

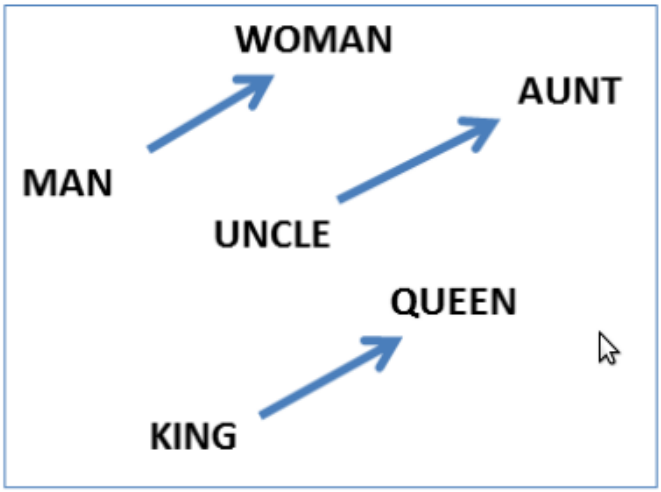
$$P(S_r) = \frac{1}{n_r} max_i(n_r^i) \tag{4}$$

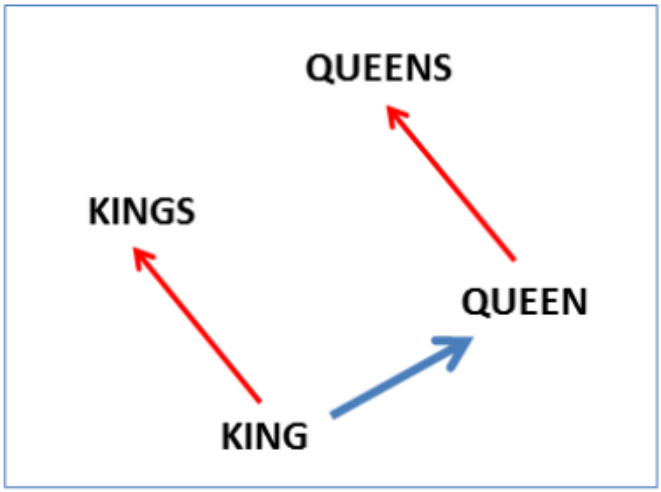
- Example: given the cluster S_1 {A A A T A A A T T A}:

$P(S_1) = \frac{1}{10} max(7, 3) = 0.7$

Analogy

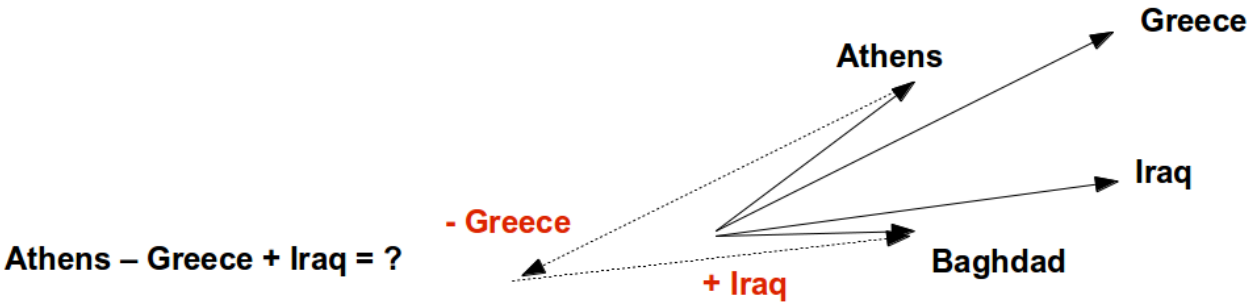
- Answer semantic and morphological analogy questions of the type *Rome is to Italy what Tokyo is to ...* (Mikolov et al 2013)
- Evaluated in terms of accuracy.





Analogy

- Athens Greece Baghdad Iraq
- ...
- happy happily cheerful cheerfully
- ...
- decrease decreasing jump jumping
- ...



The many faces of DS

Distributional semantics in 2016

Linguistic representation: disambiguation, adjective semantics, quantifiers, phrasal composition, *meaning* of words.

Cognitive representation: simulates language acquisition, priming, fMRI measurements.

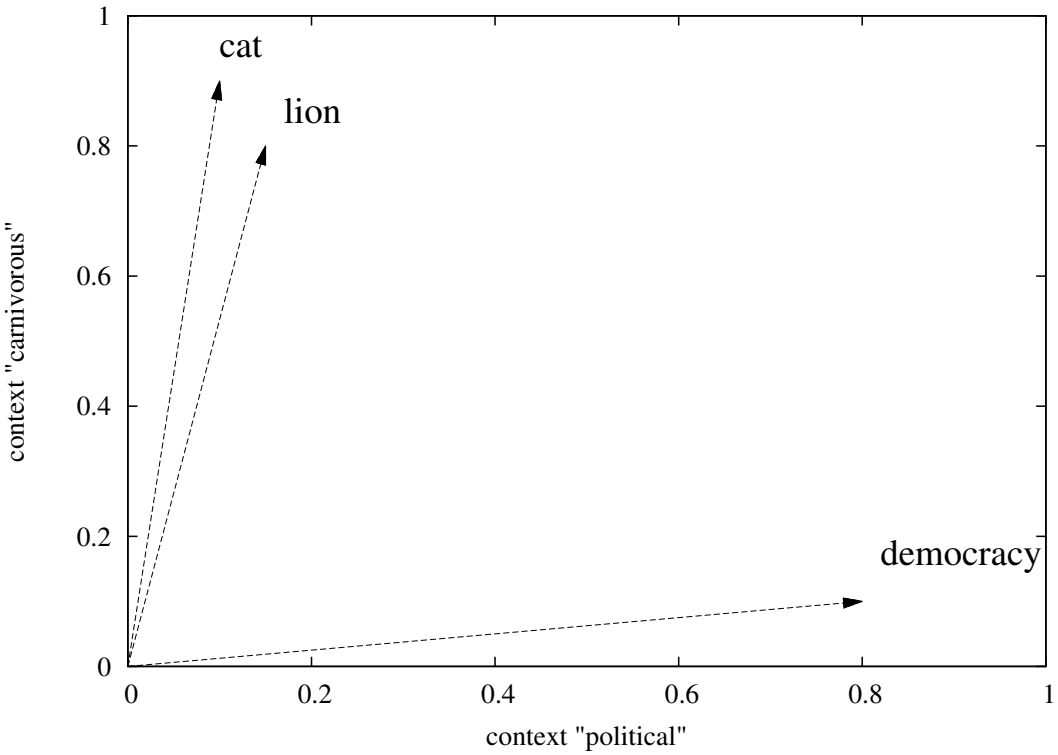
Useful hack: representation of the lexicon for NLP applications.

A cognitive representation

- Landauer & Dumais (1997): knowledge acquisition.
- Lund, Burgess & Atchley (1995): priming.
- Anderson et al (2013): multimodal distributional representations simulate brain activation.

Landauer & Dumais (1997)

- Implicit learning:



Landauer & Dumais (1997)

- The semantic space is built via a dimensionality-reduced word-document matrix.
- Corpus: the Grolier's Academic American Encyclopedia.
- Evaluation: Test of English as a Foreign Language (TOEFL) – synonymy test:

  Stem: levied
  (a) imposed
  (b) believed
  (c) requested
  (d) correlated
  Solution: (a) imposed
- Best performance around 300 dimensions.

Lund, Burgess & Atchley (1995)

- HAL: Hyperspace Analogue to Language.
- Priming: subjects are asked to recognise whether a string of letter is a word or not.
- Subjects' response time is faster if the target word is preceded by a similar item: *doctor/hospital* vs *doctor/kangaroo*.
- Priming effects can be simulated using similarity information from a semantic space.
- But: relatedness (*cradle/baby*) is not enough to produce priming effects.

A linguistic representation (next week!)

- Account for the composition of short phrases: find a function $f(\vec{u}, \vec{v})$ which returns the meaning of the composition of $\vec{u}$ and $\vec{v}$.
- Sense disambiguation: re-weight a vector in context to get the various senses of the word it represents.
- Capture some inferential properties of language: if Molly is a cat, Molly is an animal, *many cats* entails *some cats*.
- Work on affixes, mass/count distinction, relative pronouns, negation, etc, etc.

Conclusion

Conclusion

- Distributional semantics is *one* possible semantic theory, which has experimental support – both in linguistics and cognitive science.
- Various models for distributional systems, with various consequences on the output.
- Known issues: corpus-dependence (which notion of concept is at play here?), word senses are collapsed (perhaps not such a bad thing...), fixed expressions create noise in the data.

Conclusion

- Evaluation against psycholinguistic data shows that DS can model at least *some* phenomena.
- A powerful computational semantics tool, with surprising results.
- But a tool without a fully-fledged linguistic theory...