Introduction to Distributional Semantics

Aurélie Herbelot

Centre for Mind/Brain Sciences
University of Trento

Trento 2016

Preliminaries

- Today, broad overview of Distributional Semantics.
- Also preparation for Wednesday's practical.
- My email: aurelie.herbelot@cantab.net!

Introduction

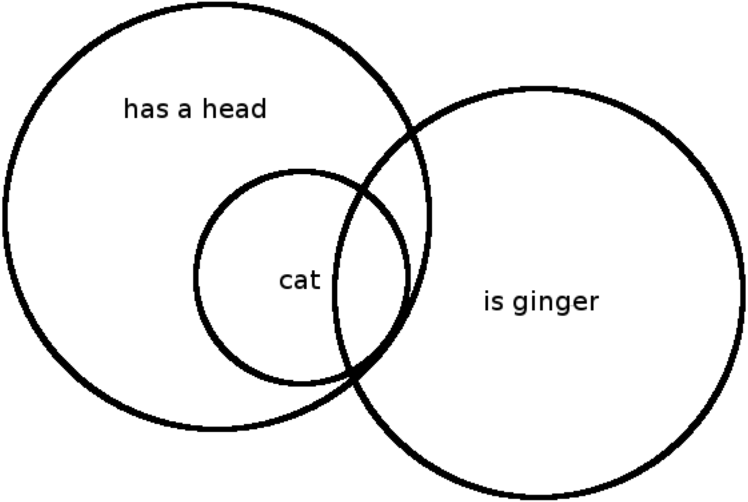
Semantics

- Semantics is the study of meaning.
- What is meaning?
  - No one knows...
  - Go read http://plato.stanford.edu/entries/meaning/
  - Theory of reference (model-theoretic semantics).
  - Meaning as use (distributional semantics).
- Semantics explains how humans (and some animals) communicate *about* the world.

Correspondence theory of meaning

- Tarski: *Snow is white* is true iff snow is white.
- Montague and formal semantics: a set-theoretic theory with the following features:
  - a model of the world;
  - the model consists of sets;
  - words in a language ‘refer’ or ‘denote’ parts of the model;
  - a proposition is true iff it ‘corresponds’ to a state of affairs in the model.

Sets



Problems with set-theoretic semantics

- Poor representation of the semantic content of words. (There is lexical semantics, but it is a semantics for very specific relations: hyponymy, synonymy, antonymy, etc)
- Do humans have sets in their heads? (Cognitive plausibility.)
- Is there truth?
- Where do models come from?

A quick history of distributional semantics

Distributional semantics: a short history

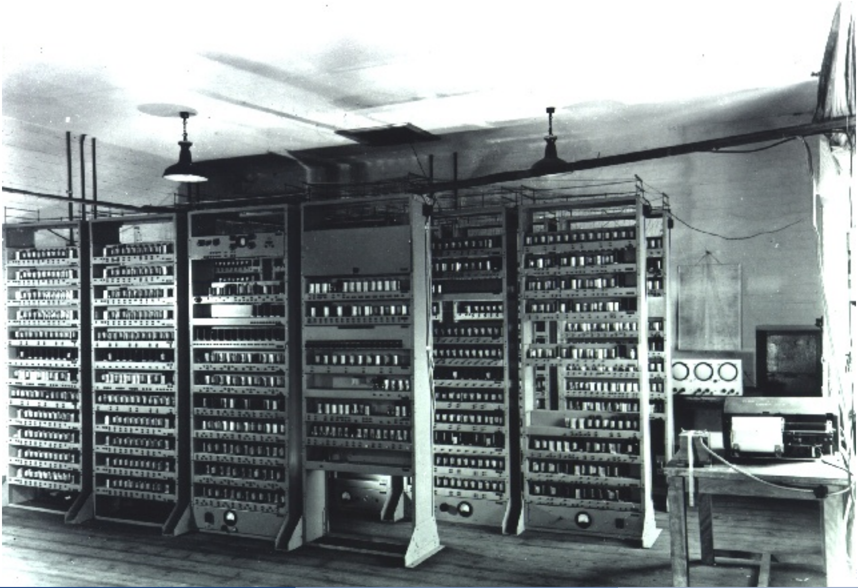


Ludwig Wittgenstein: ‘Meaning is use’: ‘Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache’ (Wittgenstein, 1953. 43)

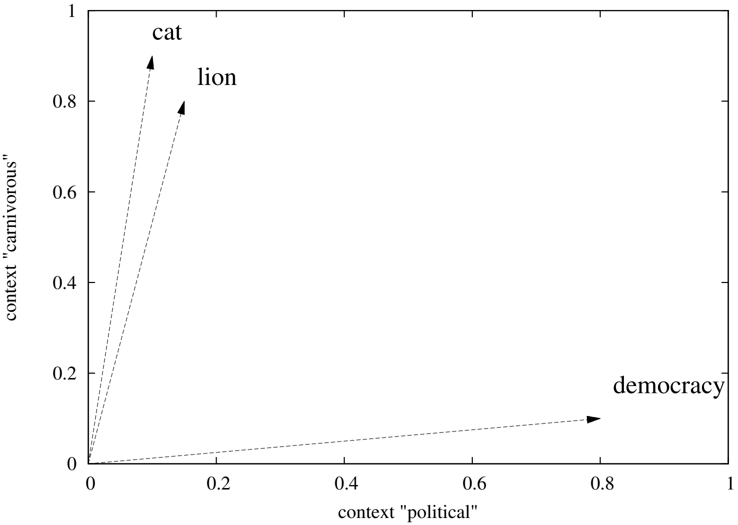
Margaret Masterman: Cambridge Language Research Unit (CLRU: 1955–1986).

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
...

0.115 english-N
0.114 written-A
0.109 grammar-N
0.106 translate-V
0.102 teaching-N
0.097 literature-N
0.096 english-A
0.096 acquisition-N
0.095 communicate-V
0.093 native-A
0.089 everyday-A
0.088 learning-N
0.084 meaning-N
0.083 french-N
0.082 description-N
0.079 culture-N
0.078 speak-V
0.078 foreign-A
0.077 classroom-N
0.077 command-N

0.075 teach-V
0.075 communication-N
0.074 knowledge-N
0.074 polish-A
0.072 speaker-N
0.071 convey-V
0.070 theoretical-A
0.069 curriculum-N
0.068 pupil-N
0.068 level-A
0.067 assessment-N
0.067 use-N
0.067 tongue-N
0.067 medium-N
0.067 spanish-A
0.066 speech-N
0.066 learn-V
0.066 interaction-N
0.065 expression-N
0.064 sign-N

0.064 universal-A
0.064 aspect-N
0.064 german-N
0.063 artificial-A
0.063 logic-N
0.061 understanding-N
0.061 official-A
0.061 formal-A
0.061 complexity-N
0.060 gesture-N
0.060 african-A
0.060 eg-A
0.060 express-V
0.059 implication-N
0.058 distinction-N
0.058 barrier-N
0.057 cultural-A
0.057 literary-A
0.057 variation-N
...

0.129 chocolate-N
0.122 slice-N
0.109 tin-N
0.109 pie-N
0.103 sandwich-N
0.103 decorate-V
0.099 cream-N
0.098 fruit-N
0.097 recipe-N
0.097 bread-N
0.096 oven-N
0.094 birthday-N
0.090 wedding-N
0.087 sugar-N
0.086 cheese-N
0.086 tea-N
0.085 butter-N
0.085 eat-V
0.084 apple-N
0.083 wrap-V

0.083 sweet-A
0.081 mix-N
0.080 mixture-N
0.079 rice-N
0.078 nut-N
0.076 tomato-N
0.076 knife-N
0.075 potato-N
0.075 oz-N
0.075 cook-N
0.075 top-V
0.074 coffee-N
0.073 christmas-N
0.073 ice-N
0.073 orange-N
0.073 layer-N
0.072 packet-N
0.072 roll-N
0.071 brush-V
0.071 meat-N

0.071 salad-N
0.071 piece-N
0.070 line-V
0.070 dry-V
0.069 round-A
0.068 egg-N
0.068 cooking-N
0.066 lb-N
0.066 fat-N
0.064 top-N
0.063 spread-V
0.063 chip-N
0.063 cut-V
0.062 sauce-N
0.062 turkey-N
0.061 milk-N
0.061 plate-N
0.060 remaining-A
0.060 hint-N
...

0.093 coloured-A
0.092 paper-N
0.089 stroke-N
0.089 margin-N
0.089 tip-N
0.085 seize-V
0.077 pig-N
0.077 ltd-A
0.076 drawing-N
0.074 electronic-A
0.072 concrete-A
0.072 portrait-N
0.071 sheep-N
0.068 pocket-N
0.066 code-N
0.066 flow-V
0.066 gardener-N
0.066 sheet-N
0.066 straw-N
0.066 outline-N

0.065 pick-V
0.065 co-N
0.064 palm-N
0.064 writing-N
0.064 jean-N
0.064 literary-A
0.063 writer-N
0.063 write-V
0.063 script-N
0.063 ash-N
0.062 desk-N
0.062 elegant-A
0.061 pause-V
0.061 brush-N
0.060 marine-A
0.060 infant-N
0.059 tape-N
0.059 collapse-N
0.058 cry-N
0.057 delighted-A

0.057 hand-V
0.057 phil-N
0.056 wilson-N
0.056 silver-N
0.056 terror-N
0.055 lower-V
0.055 tap-V
0.055 light-A
0.055 packet-N
0.055 load-V
0.054 cigarette-N
0.054 anxiety-N
0.054 program-N
0.054 complex-N
0.054 ball-N
0.053 rabbit-N
0.053 precious-A
0.052 eg-A
0.052 thanks-N
...

Modelling choices

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):

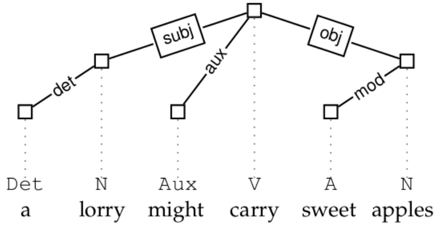
    *... the prime **minister** acknowledged that ...*

  - 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), with or without discounting factor.

$$pmi_{wc} = log(\frac{f_{wc} * f_{total}}{f_w * f_c}) \tag{1}$$

  - Derivatives such PPMI, 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.

What semantic space?

- Singular Value Decomposition (LSA – Landauer and Dumais, 1997): the number of dimensions is reduced by exploiting redundancies in the data. 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 interpretable.
- Other, more esoteric variants...

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

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

- $\emptyset$ **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 (or in the case of a dimensionality-reduced space, the reduced dimensions).
- The values at the intersecton of rows and the columns form a matrix. The values of a row are the vector for a particular lexical item.

‘Real’ distributions

Corpus description

- Obtained from the entire English Wikipedia.
- Corpus parsed with the English Resource Grammar (Copestake & Flickinger, 2000) and converted into DMRS form (Copestake, 2009).
- Dependencies considered include:
    - For nouns: head verbs (+ any other argument of the verb), modifying adjectives, head prepositions (+ any other argument of the preposition).
      e.g. cat: chase_v+mouse_n, black_a, of_p+neighbour_n
    - For verbs: arguments (NPs and PPs), adverbial modifiers.
      e.g. eat: cat_n+mouse_n, in_p+kitchen_n, fast_a
    - For adjectives: modified nouns; rest as for nouns (assuming intersective composition).
      e.g. black: cat_n, chase_v+mouse_n

System description

- Semantic space: top 100,000 contexts.
- Weighting: normalised PMI (Bouma 2009).

$$pmi_{wc} = \frac{log(\frac{f_{wc} * f_{total}}{f_w * f_c})}{-log(\frac{f_{wc}}{f_{total}})} \tag{2}$$

An example noun

- *language*:

0.541816::other+than_p()+English_n
0.525895::English_n+as_p()
0.523398::English_n+be_v
0.48977::english_a
0.481964::and_c+literature_n
0.476664::people_n+speak_v
0.468399::French_n+be_v
0.463604::Spanish_n+be_v
0.463591::and_c+dialects_n
0.452107::grammar_n+of_p()
0.445994::foreign_a
0.445071::germanic_a
0.439558::German_n+be_v
0.436135::of_p()+instruction_n

0.435633::speaker_n+of_p()
0.423595::generic_entity_rel_+speak_v
0.42313::pron_rel_+speak_v
0.42294::colon_v+English_n
0.419646::be_v+English_n
0.418535::language_n+be_v
0.4159::and_c+culture_n
0.410987::arabic_a
0.408387::dialects_n+of_p()
0.399266::part_of_rel_+speak_v
0.397::percent_n+speak_v
0.39328::spanish_a
0.39273::welsh_a
0.391575::tonal_a

An example adjective

- *academic*:

0.517031::Decathlon_n
0.512661::excellence_n
0.449711::dishonesty_n
0.445393::rigor_n
0.426142::achievement_n
0.421246::discipline_n
0.397311::vice_president_n+for_p()
0.391978::institution_n
0.38937::credentials_n
0.378062::journal_n
0.373727::journal_n+be_v
0.372052::vocational_a
0.371873::student_n+achieve_v
0.361359::athletic_a
0.356562::reputation_n+for_p()
0.354674::regalia_n
0.353712::program_n
0.351601::freedom_n
0.347751::student_n+with_p()
0.34621::curriculum_n
0.342008::standard_n
0.34151::at_p()+institution_n
0.340271::career_n
0.337857::Career_n
0.329923::dress_n
0.329358::scholarship_n
0.329281::prepare_v+student_n
0.328009::qualification_n

Issues with the representation

Corpus choice

- As much data as possible?
  - British National Corpus (BNC): 100 m words
  - Wikipedia: 897 m words
  - UKWac: 2 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.

Corpus choice

- Distribution for *unicycle*, as obtained from Wikipedia.

0.448051::motorized_a
0.404372::pron_rel_+ride_v
0.238612::for_p()+entertainment_n
0.235763::half_n+be_v
0.235407::unwieldy_a
0.230275::earn_v+point_n
0.216627::pron_rel_+crash_v
0.190785::man_n+on_p()
0.186325::on_p()+stage_n
0.185063::position_n+on_p()

0.168102::slip_v
0.162611::and_c+1_n
0.159627::autonomous_a
0.155822::balance_v
0.133084::tall_a
0.124242::fast_a
0.106976::red_a
0.0714643::come_v
0.0601987::high_a

Polysemy

- Distribution for *pot*, as obtained from Wikipedia.

0.566454::melt_v
0.442374::pron_rel_+smoke_v
0.434682::of_p()+gold_n
0.40773::porous_a
0.401654::of_p()+tea_n
0.39444::player_n+win_v
0.393812::money_n+in_p()
0.376198::of_p()+coffee_n
0.33117::amount_n+in_p()
0.329211::ceramic_a
0.326387::hot_a
0.323321::boil_v
0.313404::bowl_n+and_c
0.306324::ingredient_n+in_p()
0.301916::plant_n+in_p()

0.298764::simmer_v
0.292397::pot_n+and_c
0.284539::bottom_n+of_p()
0.28338::of_p()+flower_n
0.279412::of_p()+water_n
0.278914::food_n+in_p()
0.262501::pron_rel_+heat_v
0.260375::size_n+of_p()
0.25511::pron_rel_+split_v
0.254363::of_p()+money_n
0.2535::of_p()+culture_n
0.249626::player_n+take_v
0.246479::in_p()+hole_n
0.244051::of_p()+soil_n
0.243797::city_n+become_v

Fixed expressions

- Distribution for *time*, as obtained from Wikipedia.

0.462949::of_p()+death_n
0.448965::same_a
0.446277::1_n+at_p(temp)
0.445338::Nick_n+of_p()
0.423542::spare_a
0.418568::playoffs_n+for_p()
0.416471::of_p()+retirement_n
0.405288::of_p()+release_n
0.397135::pron_rel_+spend_v
0.389886::sand_n+of_p()
0.385954::pron_rel_+waste_v
0.382816::place_n+around_p()
0.37777::of_p()+arrival_n
0.376466::of_p()+completion_n
0.374797::after_p()+time_n
0.374682::of_p()+arrest_n
0.371589::country_n+at_p()
0.370736::age_n+at_p()
0.370626::space_n+and_c
0.370555::in_p()+career_n
0.370464::world_n+at_p()
0.363982::and_c+space_n
0.363241::generic_entity_rel_+mark_v
0.361872::of_p()+introduction_n
0.357929::in_p()+year_n
0.357565::of_p()+appointment_n
0.356229::of_p()+trouble_n
0.355658::of_p()+merger_n
0.354794::on_p()+ice_n
0.353891::practice_n+at_p()
0.351994::of_p()+birth_n
0.351556::full_a
0.348029::of_p()+accident_n
0.34785::state_n+at_p()
0.347753::to_p()+time_n
0.345147::of_p()+election_n
0.345088::area_n+at_p()
0.342571::and_c+money_n
0.342113::time_n+after_p()
0.341877::allotted_a

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

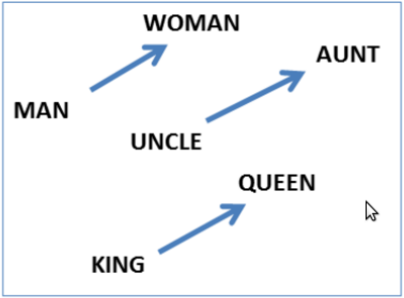
- Reproduce human similarity judgements (expressed as a score on a bounded scale).
- Rubenstein & Goodenough (1965): 65 noun pairs.
- Finkelstein et al (2002): WordSim353.
- Bruni et al (2014): MEN (1000 test pairs).
- Calculate spearman correlation (ρ) between systems results and human judgements. Human correlation on the MEN dataset is 0.68.

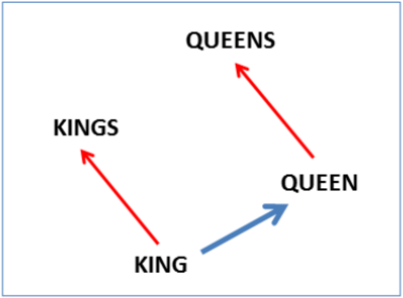
Categorisation

- 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%.

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.





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

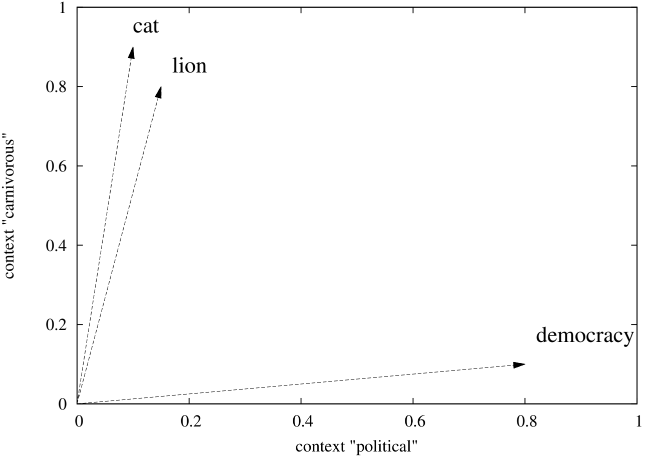
- Landauer & Dumais (1997): knowledge acquisition.
- Lund, Burgess & Atchley (1995): priming.
- Anderson et al (2013): multimodal distributional representations simulate brain activation.

Landauer & Dumais (1997)

- Explain rate of word/concept acquisition in children.
- Children learn new words by reading:
  - the majority of English words are used in print;
  - children are exposed to fewer new words in speech than in print;
  - explicit teaching does not introduce so many new words either.
- But how can a child learn a concept just by seeing it in context?

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:

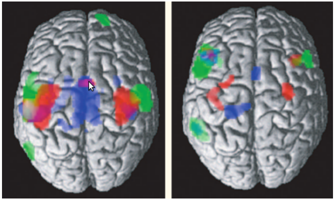
  Stem: levied
  (a) imposed
  (b) believed
  (c) requested
  (d) correlated
  Solution: (a) imposed

- 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.

Hebbian theory



Let us assume that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability.[...] When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

Hebb (1961)

Anderson et al (2013)

- A multi-modal distributional model of word meaning, compared with brain activation data.
- Multi-modal: build distributions not just from text but also from images. Use bounding boxes to identify ‘visual words’ in an image and record context. The information from text and images is concatenated.
- Brain data: fMRI data from Mitchell et al (2008).
- Calculate spearman ρ between similarity figures in brain data and distributional data.

Anderson et al (2013)

Animals	Bear, Cat, Cow, Dog Horse
Building	Apartment, Barn, Church, House
Building parts	Arch, Chimney, Closet, Door, Window
Clothing	Coat, Dress, Pants, Shirt, Skirt
Furniture	Bed, Chair, Desk, Dresser, Table
Insect	Ant, Bee, Beetle, Butterfly, Fly
Kitchen utensils	Bottle, Cup, Glass, Knife, Spoon
Man-made objects	Bell, Key, Refrigerator, Telephone, Watch
Tool	Chisel, Hammer, Screwdriver
Vegetable	Celery, Corn, Lettuce, Tomato
Vehicle	Airplane, Bicycle, Car, Train, Truck

Table : Words represented by brain/distributional models

Results

- $\rho = 0.53$ for whole-brain data at the category level.
- $\rho = 0.17$ for whole-brain data at the word pair level.
- So: strong correlations observed at the category level (i.e. similarity between 'man-made objects/tools' vs 'man-made objects/animals'), but the fine-grained level is not so easy to model.

A linguistic representation

- 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.

Sense disambiguation: Erk & Padó (2008)

- Disambiguating *river bank*:
  1. COMPOUND-LEFT: river COMPOUND-RIGHT: bank
  2. Calculate centroid of word vectors which have COMPOUND-LEFT: river as context (average over *access, basin, boat*, etc)
  3. Compose (multiply) centroid with *bank* vector.

Disambiguating *bank*

bank

COMP^{-}:(compound)robber
COMP:(compound)savings
COMP^{-}:(compound)robbery
COMP:(of)Thames
COMP:(of)rhine
COOR:ditch
VERB:(ARG2)rob
COMP:(compound)sperm
COMP^{-}:(compound)account
COMP^{-}:(compound)Thai
COMP:(compound)Habib
COMP:(of)river
COMP:(of)River
COMP^{-}:(compound)Berhad
COMP:(compound)Deutsche
COMP:(of)Nile
COMP^{-}:(compound)teller
COMP:(compound)HSBC
COMP^{-}:(compound)holiday
COMP^{-}:(compound)Fargo

national bank

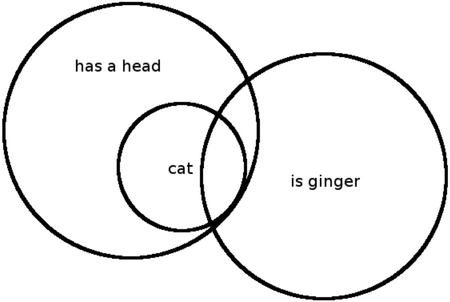
VERB:(ARG1)charge
COOR:strip
COOR:bed
COMP:(in)Philippines
VERB:(ARG2)rob
COOR:firm
VERB:(ARG2)burst
COMP^{-}:(on)section
VERB:(ARG1)borrow
COMP^{-}:(by)place_rel_
VERB:(ARG1)finance
COMP^{-}:(in)money
COOR:account
COMP^{-}:(of)failure
COMP^{-}:(from)money
VERB:(ARG2)bank
VERB:(ARG1)lower
COMP:(in)Hong_Kong
VERB:(ARG1)offset
COMP^{-}:(compound)Ltd

river bank

COMP:(of)stream
COMP(poss):river
COMP:(of)creek
COMP:(of)st
COMP:(of)canal
VERB:(ARG1)lend
COMP:(of)reservoir
COMP:(of)lake
COMP:(compound)river
COMP:(at)mouth
COMP^{-}:(on)village
COMP^{-}:(on)area
COMP:(of)Nile
COMP^{-}:(on)lie
COMP:(about)kilometer
VERB:(ARG2)erode
COMP^{-}:(on)situate
COOR:turn
COMP^{-}:(on)city
COMP:(of)channel

Quantifier entailment: Baroni et al (2012)

- Quantifiers: *no, few, some, many, most, all, more than two...*
- A stronghold of formal semantics: for sure, you need sets to do quantification...



Quantifier entailment: Baroni et al (2012)

- Study of *all, both, each, either, every, few, many, most, much, no, several, some*.
- Learn quantifier entailment by example: observe phrases such as *all cats/some cats* in a corpus, and train an SVM classifier.
- Classify previously unseen quantifier pairs.
- Results: up to 77% precision in detecting entailment. A surprising result.

Do distributions model meaning?

- A model of word meaning:
  - Cats are robots from Mars that chase mice.
  - Dogs are robots from Mars that chase cats.
  - Trees are 3D holograms from Jupiter.
- A similarity-based evaluation of this model would find that cats and dogs are very similar, but both are much less similar to trees.
- A good model of language?

Do distributions model meaning?

- A theory of meaning has to say how language relates to the world. For instance, model-theoretic semantics says that the meaning of *cat* is the set of all cats in a world.
- In distributionalism, meaning is the way we use words to talk *about* the world. No metaphysical assumptions.
- So if we use the words ‘robots from Mars’ to talk about cats, all is fine (see whales and fish).
- Not quite... (stay tuned: next week, ‘Formal Distributional Semantics’)

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 theory...

Other popular models

- Neural Network language models on the rise (Word2Vec: Mikolov et al, 2013).
- Predict models: given a context, predict a word (or the opposite!)
- Excellent results on a range of tasks, but the magic might come from setting some parameters correctly... (Levy & Goldberg 2014).