Distributional semantics for linguists

Ann Copestake and Aurélie Herbelot

Computer Laboratory, University of Cambridge
and
Department Linguistik, Universität Potsdam

August 2012

Session 1a: Outline

Introduction

History

Underlying assumptions

Course outline

Outline.

Distributional semantics

Distributional semantics: family of techniques for representing word meaning based on (linguistic) contexts of use.

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Distributional semantics

Distributional semantics: family of techniques for representing word meaning based on (linguistic) contexts of use.

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Distributional semantics

Distributional semantics: family of techniques for representing word meaning based on (linguistic) contexts of use.

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Distributional semantics: the intuitions

- Humans typically learn word meanings (concepts) from context: sometimes perceptually grounded, sometimes not.
- Possibly processed to some different representation, but perhaps mental representation directly reflects context?
- Distributional semantics uses linguistic context to represent meaning (partially).
- Meaning seen as a space, with dimensions corresponding to elements in the context (features).
- Computational techniques generally use vectors (semantic space models, vector space models).

Distributional semantics: the intuitions

- Humans typically learn word meanings (concepts) from context: sometimes perceptually grounded, sometimes not.
- Possibly processed to some different representation, but perhaps mental representation directly reflects context?
- Distributional semantics uses linguistic context to represent meaning (partially).
- Meaning seen as a space, with dimensions corresponding to elements in the context (features).
- Computational techniques generally use vectors (semantic space models, vector space models).

Distributional semantics: the intuitions

- Humans typically learn word meanings (concepts) from context: sometimes perceptually grounded, sometimes not.
- Possibly processed to some different representation, but perhaps mental representation directly reflects context?
- Distributional semantics uses linguistic context to represent meaning (partially).
- Meaning seen as a space, with dimensions corresponding to elements in the context (features).
- Computational techniques generally use vectors (semantic space models, vector space models).

Distributional semantics: the intuitions

- Humans typically learn word meanings (concepts) from context: sometimes perceptually grounded, sometimes not.
- Possibly processed to some different representation, but perhaps mental representation directly reflects context?
- Distributional semantics uses linguistic context to represent meaning (partially).
- Meaning seen as a space, with dimensions corresponding to elements in the context (features).
- Computational techniques generally use vectors (semantic space models, vector space models).

Distributional semantics: the intuitions

- Humans typically learn word meanings (concepts) from context: sometimes perceptually grounded, sometimes not.
- Possibly processed to some different representation, but perhaps mental representation directly reflects context?
- Distributional semantics uses linguistic context to represent meaning (partially).
- Meaning seen as a space, with dimensions corresponding to elements in the context (features).
- Computational techniques generally use vectors (semantic space models, vector space models).

Outline.

Some history

- Early discussion: Osgood (1952), Zelig Harris (1954).
- Firth (1957): 'You shall know a word by the company it keeps'.
- 'distributional semantics' by 1960s: e.g., Garvin (1962).
- Spärck Jones (1964): PhD thesis 'Synonymy and Semantic Classification' (dictionaries for context).
- First experiments on sentential contexts: Harper (1965) inspired by Harris; Spärck Jones (1967).
- Grefenstette (1994), Schütze (1998); Landauer and Dumais (1997) 'Latent Semantic Analysis' (LSA).
- Huge proliferation of papers in computational linguistics (CL) once corpora (and large scale parsing) become available.

Vector representations and clustering

Words represented as vectors of features:

	feature_1	feature_2	...	feature_n
word_1	$f_{1,1}$	$f_{2,1}$		$f_{n,1}$
word_2	$f_{1,2}$	$f_{2,2}$		$f_{n,2}$
...				
word_m	$f_{1,m}$	$f_{2,m}$		$f_{n,m}$

Features: co-occur with word_n in some window, co-occur with word_n as a syntactic dependent, occur in paragraph_n, occur in document_n ...

First computational application: Spärck Jones (1964)

Words co-occurring with words

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	1	0	1	1	0	0	1	0
information	1	0	1	1	0	0	1	0

(example from Jurafsky and Martin, 2008)

apricot: { boil, large, sugar, water }
pineapple: { boil, large, sugar, water }
digital: { arts, data, function, summarized }
information: { arts, data, function, summarized }

Clustering: group together words with ‘similar’ vectors.

Words co-occurring with words

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	1	0	1	1	0	0	1	0
information	1	0	1	1	0	0	1	0

(example from Jurafsky and Martin, 2008)

apricot: { boil, large, sugar, water }
pineapple: { boil, large, sugar, water }
digital: { arts, data, function, summarized }
information: { arts, data, function, summarized }

Clustering: group together words with 'similar' vectors.

Words co-occurring with words

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	1	0	1	1	0	0	1	0
information	1	0	1	1	0	0	1	0

(example from Jurafsky and Martin, 2008)

apricot: { boil, large, sugar, water }
pineapple: { boil, large, sugar, water }
digital: { arts, data, function, summarized }
information: { arts, data, function, summarized }

Clustering: group together words with ‘similar’ vectors.

Early clustering

Harper (1965): cooccurrence data for 40 nouns from 120,000 words of Russian scientific text: adjective dependents, noun dependents, noun governors.

Harper clustered by:

$$\frac{|V_1 \cap V_2|}{F_1 F_2}$$

where V_1, V_2 are cooccurring sets, F_1, F_2 are the frequencies of the nouns in the corpus.

Spärck Jones (1967): Harper's similarity coefficient is 'of doubtful propriety'. Instead clustered ('clumped') by Jaccard:

$$\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$$

Early clustering

Harper (1965): cooccurrence data for 40 nouns from 120,000 words of Russian scientific text: adjective dependents, noun dependents, noun governors.
Harper clustered by:

$$\frac{|V_1 \cap V_2|}{F_1 F_2}$$

where V_1, V_2 are cooccurring sets, F_1, F_2 are the frequencies of the nouns in the corpus.

Spärck Jones (1967): Harper's similarity coefficient is 'of doubtful propriety'. Instead clustered ('clumped') by Jaccard:

$$\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$$

Early clustering

Harper (1965): cooccurrence data for 40 nouns from 120,000 words of Russian scientific text: adjective dependents, noun dependents, noun governors.
Harper clustered by:

$$\frac{|V_1 \cap V_2|}{F_1 F_2}$$

where V_1, V_2 are cooccurring sets, F_1, F_2 are the frequencies of the nouns in the corpus.
Spärck Jones (1967): Harper's similarity coefficient is 'of doubtful propriety'. Instead clustered ('clumped') by Jaccard:

$$\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$$

Spärck Jones (1967)

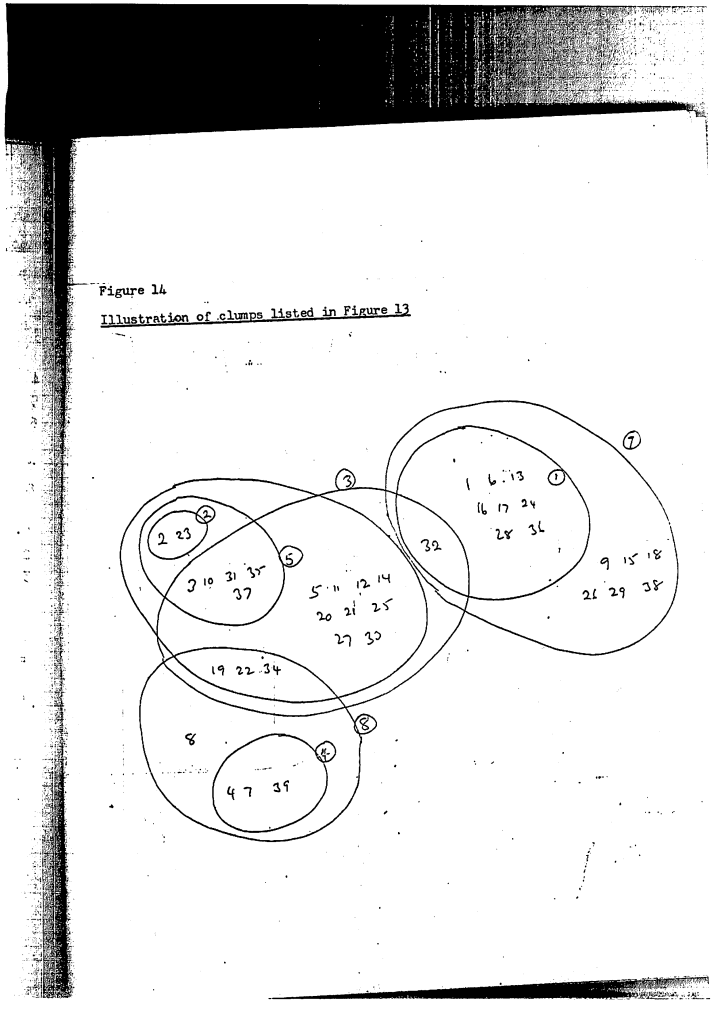
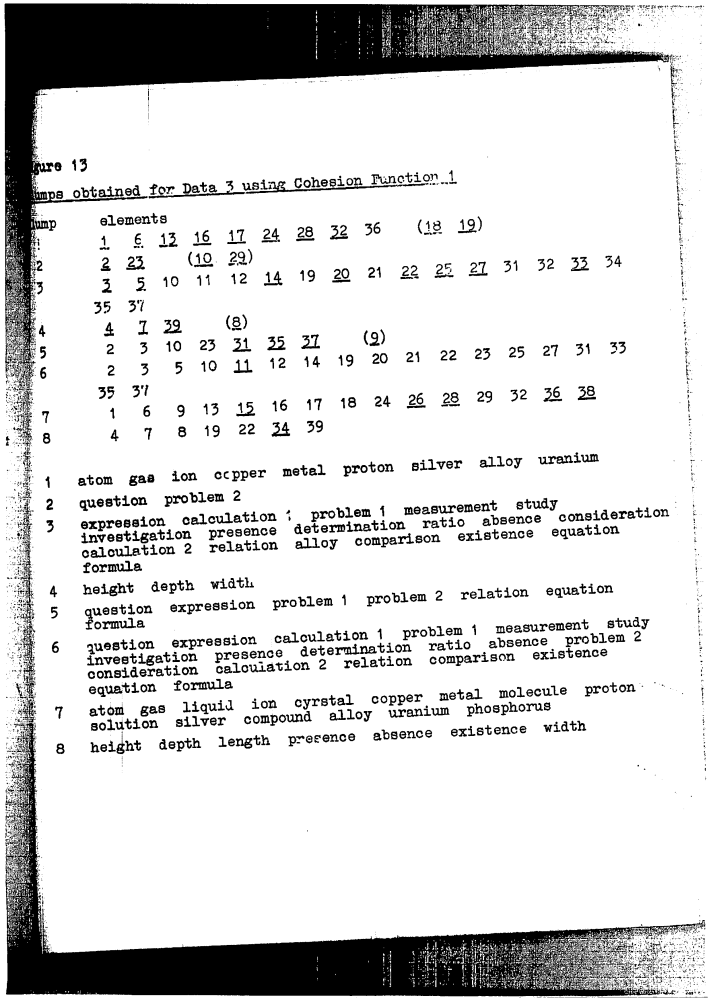


Figure 14

Illustration of clumps listed in Figure 13



ure 13

mps obtained for Data 3 using Cohesion Function 1

lump	elements
1	1 6 13 16 17 24 28 32 36 (18 19)
2	2 23 (10 29)
3	3 5 10 11 12 14 19 20 21 22 25 27 31 32 33 34 35 37
4	4 7 39 (8)
5	2 3 10 23 31 35 37 (9)
6	2 3 5 10 11 12 14 19 20 21 22 23 25 27 31 33 35 37
7	1 6 9 13 15 16 17 18 24 26 28 29 32 36 38
8	4 7 8 19 22 34 39

1 atom gas ion copper metal proton silver alloy uranium

2 question problem 2

3 expression calculation 1 problem 1 measurement study investigation presence determination ratio absence consideration calculation 2 relation alloy comparison existence equation formula

4 height depth width

5 question expression problem 1 problem 2 relation equation formula

6 question expression calculation 1 problem 1 measurement study investigation presence determination ratio absence problem 2 consideration calculation 2 relation comparison existence equation formula

7 atom gas liquid ion cyrstal copper metal molecule proton solution silver compound alloy uranium phosphorus

8 height depth length presence absence existence width

CS history and distributional semantics

- Early distributional work not followed up:
  - limitations of computers and available corpora.
  - 1966 ALPAC report led to diminished funding for CL.
  - “It must be recognized that the notion ‘probability of a sentence’ is an entirely useless one, under any known interpretation of this term.” (Chomsky 1969)
  - KSJ and others switched to Information Retrieval: KSJ (inspired by classification experiments) and Robertson develop tf*idf measure.
- Early 1990s: influence from IR: large corpora, computer memory, disk space make simple distributional techniques practical.
- Early 2000s: large scale, robust parsing makes more complex notions of context practical.

Characteristic contexts: beer

0.484118::can_n+of_p()	0.323999::and_c+drink_n
0.470041::and_c+wine_n	0.323292::alcoholic_a
0.451887::brand_n+of_p()	0.321707::tear_n+in_p()
0.444771::pron_rel_+drink_v	0.321004::and_c+brewery_n
0.407286::wine_n+and_c	0.31969::and_c+beverage_n
0.403163::duff_a	0.317467::bread_n+and_c
0.392823::and_c+cigarette_n	0.315654::recipe_n+for_p()
0.388944::liter_n+of_p()	0.312405::premium_a
0.38283::sweat_n+and_c	0.306168::rye_a
0.364612::wheat_a	0.30428::have_v+taste_n
0.341821::seasonal_a	0.301791::lite_a
0.3409::in_p()+Hell_n	0.300422::in_p()+glass_n
0.333707::or_c+spirit_n	0.299759::style_n+of_p()
0.325886::for_p()+horse_n	0.297687::stale_a
0.324157::drink_n+and_c	0.297159::be_v+drink_n

Characteristic contexts: ?

0.532551::and_c+Perry_n	0.224517::homemade_a
0.475489::sparkle_v	0.217018::ferment_v
0.462226::beer_n+and_c	0.215903::pron_rel_+drink_v
0.324184::be_v+drink_n	0.215738::and_c+wine_n
0.313665::alcoholic_a	0.212648::in_p()+Denmark_n
0.295653::hard_a	0.199628::fruit_n+and_c
0.272322::brand_n+of_p()	0.183856::eat_v+and_c
0.268747::wine_n+and_c	0.18323::and_c+apple_n
0.264604::for_p()+star_n	0.183142::and_c+grape_n
0.256199::in_p()+branch_n	0.182793::from_p()+Wales_n
0.255403::and_c+beer_n	0.182706::have_v+density_n
0.246708::liter_n+of_p()	0.180874::to_p()+productio
0.243786::and_c+spice_n	0.180084::in_p()+layer_n
0.241399::cloudy_a	0.178431::hazy_a
0.239619::gallon_n+of_p()	0.178213::Tech_n+and_c

Outline.

Psycholinguistics

- Latent Semantic Analysis (LSA) popular as a technique for investigating lexical semantics.
- Neural basis of word meaning: functional web of neurons associated with a lexeme connects recognizers, semantics and articulators (e.g. Pulvermüller 2002).
- Hebbian learning principle: paraphrased as “Neurons that fire together wire together”.
- Under these assumptions: if two lexemes co-occur frequently this would necessarily lead to strong associations between their functional webs.

Assumptions about lexical semantics

1. Limited (if any) role for semantic primitives (*kill* not CAUSE(x (DIE(y))) or similar).
2. No hard boundary between linguistic knowledge and world knowledge.
3. Acquisition must be considered.
4. Word meaning is fuzzy, speakers negotiate meaning.
5. Senses (other than homonyms) are not discrete.

Why ‘Distributional semantics for linguists’?

- Part of an approach to meaning representation?
- More modestly:
  - Semantic classification for investigation of syntax-semantic interface.
  - Investigative tool for sociolinguists etc.
- Practicalities: free/cheap corpora and ordinary computer hardware are now fully adequate for most experiments.

Outline.

Course outline

1. Introduction
   a. Introduction, historical overview, course structure.
   b. Basic distributional models.
2. a. Classical lexical semantics versus distributional semantics.
   b. Collocation. Polysemy. Some linguistic applications.
3. a. Composition of distributions.
   b. Deeper distributional semantics? ‘Lexicalised compositionality’.
4. The Generative Lexicon and distributional semantics.
5. a. Quantification and distributional semantics.
   b. General discussion (time permitting!)

Distributional semantics for linguists

Lecture 1b

Aurelie Herbelot

Universität Potsdam
Department Linguistik

ESSLLI 2012

Outline

Overview

- Models: which choices must be made when designing a distributional semantics system?
- Building the system: step-by-step example.
- Looking at real distributions.
- Issues: corpus choice, polysemy, fixed expressions.

Outline

The general intuition

- **Distributions** are vectors in a multidimensional semantic space, that is, objects with a magnitude (length) and a direction.
- The **semantic space** has dimensions which correspond to possible contexts.
- For our purposes, a distribution can be seen as a point in that space (the vector being defined with respect to the origin of that space).
- *cat* [...dog 0.8, eat 0.7, joke 0.01, mansion 0.2, zebra 0.1...]

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):

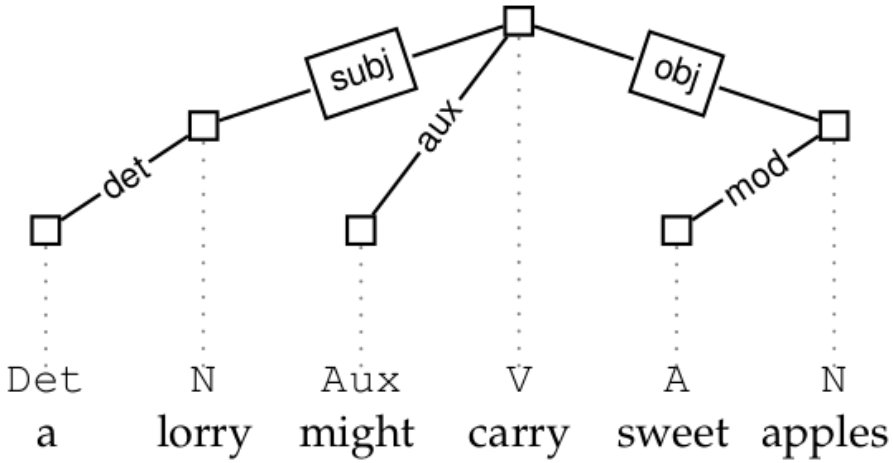
    *... the prime **minister** acknowledged that ...*

  - 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.

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

  - Derivatives such as Mitchell and Lapata's (2010) weighting function (PMI without the log).

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...

Outline

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 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.

- We assume that we only keep content words in the semantic space.
- **Dimensions:**

difference	impossible	thing
get	major	turns
go	possibly	usually
goes	repair	wrong

Frequency counts...

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.

- **Counts:**

difference 1	impossible 1	thing 3
get 1	major 1	turns 1
go 3	possibly 2	usually 1
goes 1	repair 1	wrong 4

Conversion into 5-word windows...

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.

- ∅ ∅ **the** major difference
- ∅ the **major** difference between
- the major **difference** between a
- major difference **between** a thing
- ...

Distribution for *wrong*

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.

- **Distribution (frequencies):**

difference 0	impossible 0	thing 0
get 0	major 0	turns 0
go 1	possibly 1	usually 1
goes 2	repair 0	wrong 2

Distribution for *wrong*

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.

- **Distribution (PMIs):**

difference 0	impossible 0	thing 0
get 0	major 0	turns 0
go 0.22184875	possibly 0.397940009	usually 0.698970004
goes 1	repair 0	wrong 0.397940009

Outline

Corpus description

- Obtained from the entire English Wikipedia.
- Corpus parsed with the English Resource Grammar (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 2007).

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

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

Outline

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

Outline

Conclusion

- 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.

Distributional semantics for linguists

Lecture 2a: Distributional semantics vs. classical lexical semantics

Aurelie Herbelot

Universität Potsdam
Department Linguistik

ESSLLI 2012

Outline

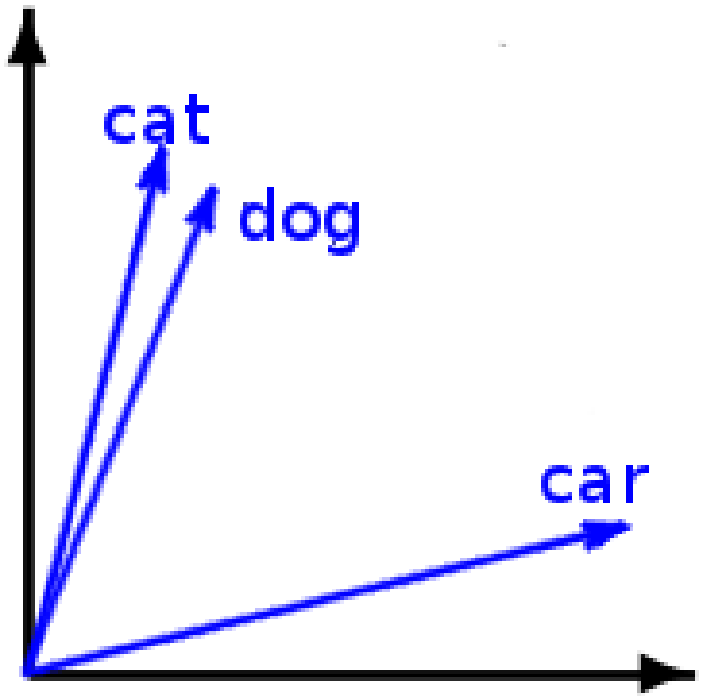
Overview

- How to calculate similarity in a distributional semantic space? What do we mean by ‘similarity’?
- A review of some standard lexical relations: synonymy, antonymy and hyponymy. How can they be translated in distributional terms?
- A summary of the (lack of) explanatory power in current distributional semantics theories.

Outline

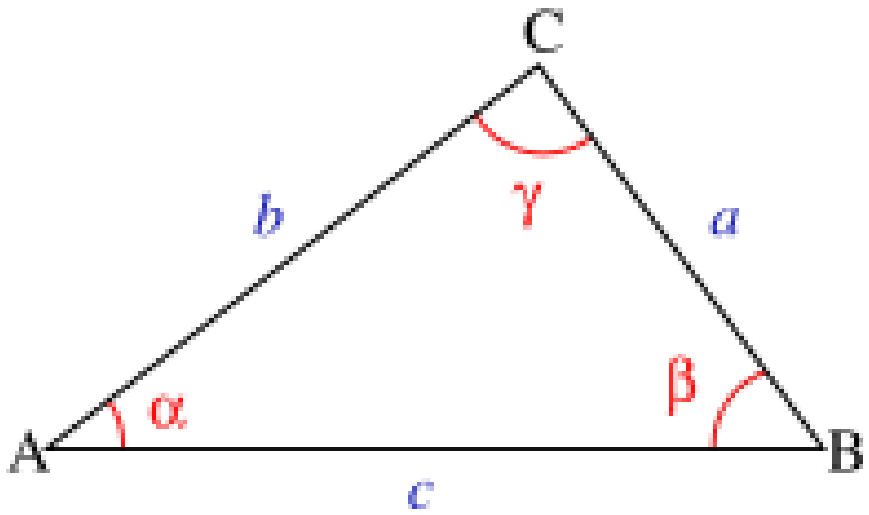
Calculating similarity in a distributional space

- Distributions are vectors, so distance can be calculated.



Some trigonometry

- Law of cosines: $c^2 = a^2 + b^2 - 2ab\cos\gamma$



Measuring similarity

- Cosine:

$$\frac{\sum v1_k * v2_k}{\sqrt{\sum v1_k^2} * \sqrt{\sum v2_k^2}} \quad (1)$$

- The cosine measure calculates the angle between two vectors and is therefore length-independent. This is important, as frequent words have longer vectors than less frequent ones.
- Other measures include Jaccard, Lin... (For an overview: see Weeds, 2004).

Some numbers

- The scale of similarity...
  - house – building 0.428354
  - gem – jewel 0.306866
  - capitalism – communism 0.294677
  - motorcycle – bike 0.29329
  - test – exam 0.269151
  - school – student 0.250291
  - singer – academic 0.168105
  - horse – farm 0.133888
  - man –accident 0.0885102
  - tree – auction 0.0234772
  - cat –county 0.00731196

Example

- Words most similar to *cat*, as chosen from the 5000 most frequent nouns in Wikipedia.

1 cat	0.276042 man	0.230953 person
0.4512 dog	0.275582 cow	0.229124 pet
0.357814 animal	0.264269 fox	0.228973 lizard
0.336883 rat	0.260912 girl	0.228406 chicken
0.331284 rabbit	0.26071 sheep	0.223872 monster
0.329772 pig	0.258142 boy	0.218094 people
0.309073 monkey	0.255272 elephant	0.216812 tiger
0.307839 bird	0.248803 deer	0.215497 mammal
0.302241 horse	0.247423 woman	0.212786 bat
0.296586 mouse	0.245761 fish	0.2122 duck
0.292734 wolf	0.243787 squirrel	0.209441 cattle
0.292047 creature	0.243725 dragon	0.208839 dinosaur
0.287286 human	0.243714 frog	0.207969 character
0.286601 goat	0.234795 baby	0.207257 kid
0.282235 snake	0.233694 child	0.206511 turtle
0.279406 bear	0.231072 lion	0.2049 robot

But what is similarity?

- In distributional semantics, very broad notion. Includes synonyms, near-synonyms, hyponyms, taxonomical siblings, antonyms, etc.
- The broad notion does correlate with a psychological reality. One of the favourite tests of the distributional semantics community is the calculation of correlation between a distributional similarity system and human judgments on the Miller & Charles (1991) test set.

Miller & Charles 1991

3.92 automobile-car	3.05 bird-cock	0.84 forest-graveyard
3.84 journey-voyage	2.97 bird-crane	0.55 monk-slave
3.84 gem-jewel	2.95 implement-tool	0.42 lad-wizard
3.76 boy-lad	2.82 brother-monk	0.42 coast-forest
3.7 coast-shore	1.68 crane-implement	0.13 cord-smile
3.61 asylum-madhouse	1.66 brother-lad	0.11 glass-magician
3.5 magician-wizard	1.16 car-journey	0.08 rooster-voyage
3.42 midday-noon	1.1 monk-oracle	0.08 noon-string
3.11 furnace-stove	0.89 food-rooster	
3.08 food-fruit	0.87 coast-hill	

- Miller & Charles experiment: re-run of Rubenstein & Goodenough (1965). Correlation coefficient = 0.97.

Distributional methods are discursive

- Distributions are a good conceptual representation if you believe that ‘the meaning of a word is given by its usage’.
- Corpus-dependent, culture-dependent, register-dependent. Example: similarity between *policeman* and *cop*: 0.232632.

Distributions are register-dependent

policeman
0.586482::ball_n+poss_rel
0.47911::and_c+civilian_n
0.424271::soldier_n+and_c
0.409217::and_c+soldier_n
0.384081::secret_a
0.370919::people_n+include_v
0.36834::corrupt_a
0.358544::uniformed_a
0.352538::uniform_n+poss_rel
0.349553::civilian_n+and_c
0.315058::iraqi_a
0.311442::lot_n+poss_rel
0.307535::chechen_a
0.303514::laugh_v
0.286281::and_c+criminal_n
0.285162::incompetent_a
0.284202::pron_rel_+shoot_v
0.279526::hat_n+poss_rel
0.276776::terrorist_n+and_c
0.272654::and_c+crowd_n
0.271465::military_a

cop
0.450031::crooked_a
0.448631::corrupt_a
0.439307::maniac_a
0.380065::dirty_a
0.373174::honest_a
0.357623::uniformed_a
0.350859::tough_a
0.327847::pron_rel_+call_v
0.320139::funky_a
0.317952::bad_a
0.29243::veteran_a
0.290737::and_c+robot_n
0.285521::and_c+criminal_n
0.279318::bogus_a
0.276689::talk_v+to_p()+pron_rel_
0.272944::investigate_v+murder_n
0.257574::on_p()+force_n
0.251643::parody_n+of_p()
0.249137::Mason_n+and_c
0.246172::pron_rel_+kill_v
0.246089::racist_a

Outline

Extension

- In set-theoretic semantics, the meaning or **extension** of *cat*, cat′, is the set of all cats in some world.
- Sets intersect, so the meaning of *black cat* is $\text{cat}'(x) \wedge \text{black}'(x)$, the intersection of the set of cats and the set of black things.
- Some entities will be in several sets.

Outline

The classical account

- Difference between full synonymy (*eggplant/aubergine*) and near-synonymy (*city, town*).
- The extensions of two full synonyms are identical sets. $\text{eggplant}' = \text{aubergine}'$
- The extensions of two near synonyms have a high (whatever that means...) overlap. i.e. with respect to a specific context, near-synonyms will often be substitutable.

Some facts about synonymy

- Near-synonymy is frequent, absolute synonymy relates to dialect etc. (*eggplant/aubergine*)
- Word sense assumptions affect synonymy assumptions.
- Language learners tend to assume non-synonymy.
  e.g., “labeling entities with distinct words leads infants to create representations of two distinct individuals” (Carey, 2009:p 277)

Near-synonymy and meaning acquisition

- Readers only need a few uses to obtain a working idea of a new word's meaning. (Rice, 1990)
- Hypothesis: understanding a new word (without definition) can be modelled by two-phase comparison:
  - initial approximation: e.g., *rancid* is similar to *off*
  - acquisition of differentiating information **characteristic contexts**: e.g., *rancid* tends to appear with fatty foods (or dairy foods, or . . .)
- People's beliefs about low-to-medium frequency words may differ but approximation is usually good enough for communication.

Are *frumpy* and *dowdy* synonyms?

Ann's intuition (pre data check): both negative, both refer to women/women's clothing, *dowdy* implies *dull*, *frumpy* implies *tasteless*.
BNC:

- frumpy: 17 total. 8 clothing, 9 people.
- dowdy: 73 total. 35% people, 10% clothing, 20% abstract, 15% location/organisation.
- Conjoined adjectives
  frumpy: *old* (twice), *new*
  dowdy: plain; solid; nondescript; gauche; second-rate; unkempt; unpleasant, stupid

slightly dowdy elegance — if there could be such a thing

Are *frumpy* and *dowdy* synonyms?

Ann's intuition (pre data check): both negative, both refer to women/women's clothing, *dowdy* implies *dull*, *frumpy* implies *tasteless*.

BNC:

- frumpy: 17 total. 8 clothing, 9 people.
- dowdy: 73 total. 35% people, 10% clothing, 20% abstract, 15% location/organisation.
- Conjoined adjectives
  frumpy: *old* (twice), *new*
  dowdy: plain; solid; nondescript; gauche; second-rate; unkempt; unpleasant, stupid

slightly dowdy elegance — if there could be such a thing

Are *frumpy* and *dowdy* synonyms?

Ann's intuition (pre data check): both negative, both refer to women/women's clothing, *dowdy* implies *dull*, *frumpy* implies *tasteless*.
BNC:

- frumpy: 17 total. 8 clothing, 9 people.
- dowdy: 73 total. 35% people, 10% clothing, 20% abstract, 15% location/organisation.
- Conjoined adjectives
  frumpy: *old* (twice), *new*
  dowdy: plain; solid; nondescript; gauche; second-rate; unkempt; unpleasant, stupid

slightly dowdy elegance — if there could be such a thing

Are *frumpy* and *dowdy* synonyms?

Ann's intuition (pre data check): both negative, both refer to women/women's clothing, *dowdy* implies *dull*, *frumpy* implies *tasteless*.
BNC:

- frumpy: 17 total. 8 clothing, 9 people.
- dowdy: 73 total. 35% people, 10% clothing, 20% abstract, 15% location/organisation.
- Conjoined adjectives
  frumpy: *old* (twice), *new*
  dowdy: plain; solid; nondescript; gauche; second-rate; unkempt; unpleasant, stupid

slightly dowdy elegance — if there could be such a thing

Are *frumpy* and *dowdy* synonyms?

Ann's intuition (pre data check): both negative, both refer to women/women's clothing, *dowdy* implies *dull*, *frumpy* implies *tasteless*.
BNC:

- frumpy: 17 total. 8 clothing, 9 people.
- dowdy: 73 total. 35% people, 10% clothing, 20% abstract, 15% location/organisation.
- Conjoined adjectives
  frumpy: *old* (twice), *new*
  dowdy: plain; solid; nondescript; gauche; second-rate; unkempt; unpleasant, stupid

slightly dowdy elegance — if there could be such a thing

Full synonymy and meaning acquisition

- Full synonyms are probably acquired differently from near-synonyms, generally by (relatively) explicit definition:

  The aubergine (eggplant) has to be one of my favourite vegetables.

- Full synonyms may be different vocalisations for the same concept (their lexemes share a single semantic functional web in the brain).
- Contrast with near-synonyms which are separate concepts.

Frequency and synonymy

- Speakers use the most frequent term in their experience to convey a particular idea (frequency assumed to correlate with strength of neural connections).
- More frequent words tend to have broader meanings (more 'senses' . . .)
- Two words of very different frequency are unlikely to cover exactly the same semantic space.
- Many words are of too low frequency for hearers to make reliable decisions about synonymy.

Synonymy: requisites for an ideal distributional account

- Distinguishing between near-synonyms and full synonyms.
- No hard line between near-synonyms and non-synonyms.
- Degree of synonymy between two lexemes will vary between individuals.

The distribution of synonyms

- Similarity between *egglant/aubergine*: 0.114024
  Relatively low cosine. Partly due to frequency (222 for *eggplant*, 56 for *aubergine*).
- Similarity between *policeman/cop*: 0.232632
  To be expected: *policeman* and *cop* are discursively very different.
- Similarity between *city/town*: 0.735319
- So... similarity does not tell us how to distinguish between full and near-synonymy.

Outline

The classical account

- Three basic types of antonymy:
  - gradable (opposite ends of a scale: *cold/hot*, modifiable with *very*, etc)
  - non-gradable (discrete opposition: *dead/alive*)
  - multiple (non-gradable, discontinuous scale: *lecturer, reader, professor*).
- In terms of extension: the same entity cannot be described as both X and its antonym Y in a given situation. i.e. for a micro-world corresponding to a situation where I drink tea, the tea cannot be in the set of cold things and in the set of hot things.

Distributions of antonyms

- Similarities between:
  - cold/hot 0.287398
  - dead/alive 0.242078
  - large/small 0.6783
  - colonel/general 0.333739

Identifying antonyms

- Antonyms have a high distributional similarity. It is hard to distinguish them from near-synonyms.
- The identification of antonyms usually requires some heuristics to be applied to pairs of highly similar distributions.
- For instance, it has been observed that antonyms are frequently coordinated while synonyms are not:
  - a selection of cold and hot drinks
  - wanted dead or alive
  - lectures, readers and professors are invited to attend

Outline

The classical account

- Relationship between a more general term and a more specific term (*dog/poodle*).
- The extension of the more general includes the extension of the more specific (all poodles are dogs).
- The intension of the more specific includes the intension of the more general (all that can be said about dogs can be said about poodles)... in an essentialist account (see penguins).

Distributions of hyponyms

- No clear inclusion relationship. The set of contexts recorded for *cat* and *animal* overlap, but they are by no means subsets.
- Kotlerman et al (2010), however, demonstrated that in general, if X is a hyponym of Y, features with high values in X tend to have a high value in Y.
- Baroni et al (2012) learn hyponymy from distributions for adjective-noun phrases (a black cat is a cat). But they do not report on the features used by the classifier.
- Similarity between *cat* and *animal*: 0.357814.

Outline

Issues

- There is no formal definition for the standard lexical relations in distributional semantics.
- The standard definitions rely on the idea of extension, but there is no obvious correspondence between the corpora used to produce distributions and the real world.

Outline

Conclusion

- A vague notion of similarity can be captured from distributions.
- Standard lexical relations tend to have medium to high similarity but there is very much variations, some of which due to frequency effects, some of which idiosyncratic.
- Identifying pairs of words in a specific lexical relation relies on heuristics.
- There is, in the standard distributional account, no formal specification for lexical relations.