Distributional semantics for linguists

Ann Copestake and Aurélie Herbelot

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

August 2012

Session 1a: 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:

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

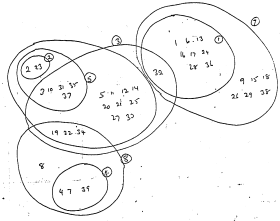
where V_1, V_2 are cooccurring sets, F_1, F_2 are the frequencies of the nouns in the corpus.
Spärck Jones (1967): Harper's similarity coefficient is 'of doubtful propriety'. Instead clustered ('clumped') by Jaccard:

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

Spärck Jones (1967)

Figure 14

Illustration of clumps listed in Figure 13



gure 13

mps obtained for Data 3 using Cohesion Function 1

ump	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 ccpper metal proton silver alloy uranium
2	question problem 2
3	expression calculation 1 problem 1 measurement study consideration investigation presence determination ratio absence 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 calcuiation 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_
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.

Introduction

History

Underlying assumptions

Course 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!)