Natural Language Processing 1

Lecture 5: Lexical and distributional semantics

Katia Shutova

ILLC
University of Amsterdam

12 November 2018

Semantics

Compositional semantics:

- studies how meanings of phrases are constructed out of the meaning of individual words
- principle of compositionality: meaning of each whole phrase derivable from meaning of its parts
- sentence structure conveys some meaning: obtained by syntactic representation

Lexical semantics:

- studies how the meanings of individual words can be represented and induced

What is lexical meaning?

- recent results in psychology and cognitive neuroscience give us some clues
- but we don't have the whole picture yet
- different representations proposed, e.g.
  - formal semantic representations based on logic,
  - *or* taxonomies relating words to each other,
  - *or* distributional representations in statistical NLP
- but none of the representations gives us a complete account of lexical meaning

How to approach lexical meaning?

- Formal semantics: set-theoretic approach
  e.g., cat′: the set of all cats; bird′: the set of all birds.
- meaning postulates, e.g.

$$\forall x[\text{bachelor}'(x) \rightarrow \text{man}'(x) \wedge \text{unmarried}'(x)]$$

- Limitations, e.g. *is the current Pope a bachelor?*
- Defining concepts through enumeration of all of their features in practice is highly problematic
- How would you define e.g. *chair, tomato, thought, democracy*? – impossible for most concepts
- Prototype theory offers an alternative to set-theoretic approaches

How to approach lexical meaning?

- Formal semantics: set-theoretic approach
  e.g., cat′: the set of all cats; bird′: the set of all birds.
- meaning postulates, e.g.

$$\forall x[\text{bachelor}'(x) \rightarrow \text{man}'(x) \wedge \text{unmarried}'(x)]$$

- Limitations, e.g. *is the current Pope a bachelor?*
- Defining concepts through enumeration of all of their features in practice is highly problematic
- How would you define e.g. *chair, tomato, thought, democracy*? – impossible for most concepts
- Prototype theory offers an alternative to set-theoretic approaches

Prototype theory

- introduced the notion of graded semantic categories
- no clear boundaries
- no requirement that a property or set of properties be shared by all members
- certain members of a category are more central or prototypical (i.e. instantiate the prototype)

furniture: chair is more prototypical than stool

Eleanor Rosch 1975. *Cognitive Representation of Semantic Categories* (J Experimental Psychology)

Prototype theory (continued)

- Categories form around prototypes; new members added on basis of resemblance to prototype
- Features/attributes generally graded
- Category membership a matter of degree
- Categories do not have clear boundaries

Semantic relations

Hyponymy: IS-A

dog is a hyponym of *animal*
animal is a hypernym of *dog*

- hyponymy relationships form a taxonomy
- works best for concrete nouns
- multiple inheritance: e.g., is *coin* a hyponym of both *metal* and *money*?

Other semantic relations

Meronomy: PART-OF e.g., *arm* is a meronym of *body*, *steering wheel* is a meronym of *car* (piece vs part)

Synonymy e.g., *aubergine*/*eggplant*.

Antonymy e.g., *big*/*little*

Also:

Near-synonymy/similarity e.g., *exciting*/*thrilling*
e.g., *slim*/*slender*/*thin*/*skinny*

WordNet

- large scale, open source resource for English
- hand-constructed
- wordnets being built for other languages
- organized into synsets: synonym sets (near-synonyms)
- synsets connected by semantic relations

```
S: (v) interpret, construe, see (make sense of;
 assign a meaning to) - "How do you interpret his
  behavior?"
S: (v) understand, read, interpret, translate (make
 sense of a language) "She understands French";
 "Can you read Greek?"
```

Polysemy and word senses

The children **ran** to the store
If you see this man, **run**!
Service **runs** all the way to Cranbury
She is **running** a relief operation in Sudan
the story or argument **runs** as follows
Does this old car still **run** well?
Interest rates **run** from 5 to 10 percent
Who's **running** for treasurer this year?
They **ran** the tapes over and over again
These dresses **run** small

Polysemy

- homonymy: unrelated word senses. *bank* (raised land) vs *bank* (financial institution)
- *bank* (financial institution) vs *bank* (in a casino): related but distinct senses.
- regular polysemy and sense extension
  - zero-derivation, e.g. *tango* (N) vs *tango* (V), or *rabbit, turkey, halibut* (meat / animal)
  - metaphorical senses, e.g. *swallow* [food], *swallow* [information], *swallow* [anger]
  - metonymy, e.g. he played *Bach*; he drank his *glass*.
- vagueness: *nurse, lecturer, driver*
- cultural stereotypes: *nurse, lecturer, driver*

No clearcut distinctions.

Word sense disambiguation

- Needed for many applications
- relies on context, e.g. collocations: *striped bass* (the fish) vs *bass guitar*.

Methods:

- supervised learning:
  - Assume a predefined set of word senses, e.g. WordNet
  - Need a large sense-tagged training corpus (difficult to construct)
- semi-supervised learning (Yarowsky, 1995)
  - bootstrap from a few examples
- unsupervised sense induction
  - e.g. cluster contexts in which a word occurs

WSD by semi-supervised learning

Yarowsky, David (1995) *Unsupervised word sense disambiguation rivalling supervised methods*

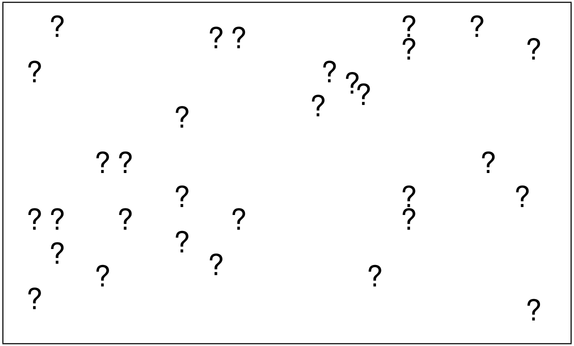
Disambiguating *plant* (factory vs vegetation senses):

1. Find contexts in training corpus:

sense	training example
?	company said that the *plant* is still operating
?	although thousands of *plant* and animal species
?	zonal distribution of *plant* life
?	company manufacturing *plant* is in Orlando
	etc

Yarowsky (1995): schematically

Initial state

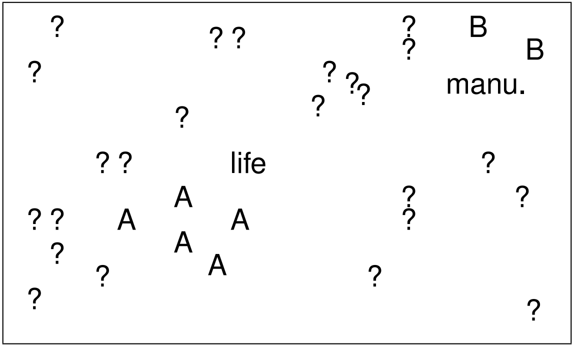


2. Identify some seeds to disambiguate a few uses:

‘*plant* life’ for vegetation use (A)
‘manufacturing *plant*’ for factory use (B)

sense	training example
?	company said that the *plant* is still operating
?	although thousands of *plant* and animal species
A	zonal distribution of *plant* life
B	company manufacturing *plant* is in Orlando
	etc

Seeds



3. Train a decision list classifier on Sense A/Sense B examples.

Rank features by log-likelihood ratio:

$$\log \left(\frac{P(Sense_A | f_i)}{P(Sense_B | f_i)} \right)$$

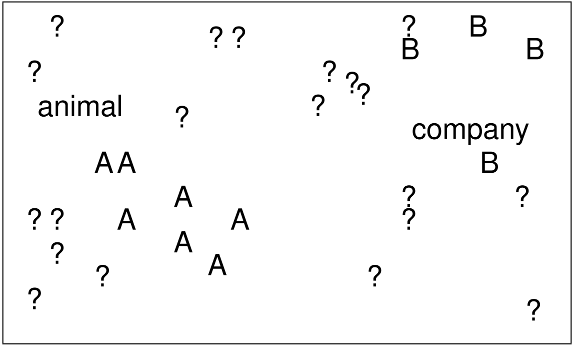
reliability	criterion	sense
8.10	*plant* life	A
7.58	manufacturing *plant*	B
6.27	*animal* within 10 words of *plant*	A
	etc	

4. Apply the classifier to the training set and add reliable examples to A and B sets.

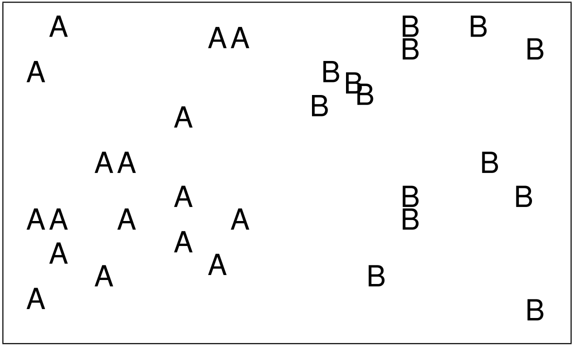
sense	training example
?	company said that the *plant* is still operating
A	although thousands of *plant* and animal species
A	zonal distribution of *plant* life
B	company manufacturing *plant* is in Orlando
	etc

5. Iterate the previous steps 3 and 4 until convergence

Iterating:



Final:



6. Apply the classifier to the unseen test data

- 'one sense per discourse': can be used as an additional refinement
- Yarowsky's experiments were nearly all on homonyms: these principles may not hold as well for sense extension.

Problems with WSD as supervised classification

Yarowsky reported an accuracy of 95%, but ...

- on 'easy' homonymous examples
- real performance around 75% (supervised)
- need to predefine word senses (not theoretically sound)
- need a very large training corpus (difficult to annotate, humans do not agree)
- learn a model for individual words — no real generalisation

Better way:

- unsupervised sense induction (but a very hard task)

Distributional hypothesis

You shall know a word by the company it keeps (Firth)

The meaning of a word is defined by the way it is used (Wittgenstein).

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Cornish Scrumpy Medium Dry. £19.28 - Case

Distributional hypothesis

You shall know a word by the company it keeps (Firth)

The meaning of a word is defined by the way it is used (Wittgenstein).

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Cornish Scrumpy Medium Dry. £19.28 - Case

Distributional hypothesis

You shall know a word by the company it keeps (Firth)

The meaning of a word is defined by the way it is used (Wittgenstein).

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Cornish Scrumpy Medium Dry. £19.28 - Case

Distributional hypothesis

You shall know a word by the company it keeps (Firth)

The meaning of a word is defined by the way it is used (Wittgenstein).

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Cornish Scrumpy Medium Dry. £19.28 - Case

Distributional hypothesis

You shall know a word by the company it keeps (Firth)

The meaning of a word is defined by the way it is used (Wittgenstein).

it was authentic scrumpy, rather sharp and very strong

we could taste a famous local product — scrumpy

spending hours in the pub drinking scrumpy

Cornish Scrumpy Medium Dry. £19.28 - Case

Scrumpy


WESTONS
CIDERS INSPIRED BY TRADITION
SCRUMPY
CLOUDY CIDER
Alc 7.5% Vol

Distributional hypothesis

This leads to the distributional hypothesis about word meaning:

- the context surrounding a given word provides information about its meaning;
- words are similar if they share similar linguistic contexts;
- semantic similarity $\approx$ distributional similarity.

Distributional semantics

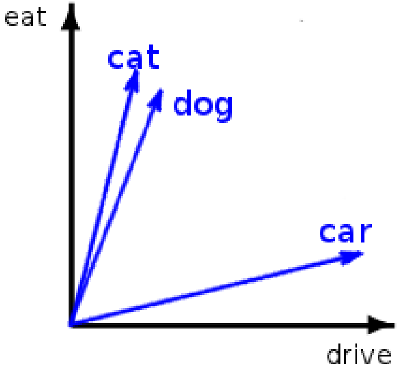
Distributional semantics: family of techniques for representing word meaning based on (linguistic) contexts of use.

1. Count-based models:
   - Vector space models
   - dimensions correspond to elements in the context
   - words are represented as vectors, or higher-order tensors
2. Prediction models:
   - Train a model to predict plausible contexts for a word
   - learn word representations in the process

Count-based approaches: the general intuition

- The **semantic space** has dimensions which correspond to possible contexts – features.
- 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).
- *scrumpy* [...pub 0.8, drink 0.7, strong 0.4, joke 0.2, mansion 0.02, zebra 0.1...]

Vectors



Feature matrix

	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	$f_{1,m}$	$f_{2,m}$		$f_{n,m}$

The notion of context

1. Word windows (unfiltered): *n* words on either side of the lexical item.
   Example: n=2 (5 words window):

   | *The prime* ***minister*** *acknowledged the* | *question.*

   minister [the 2, prime 1, acknowledged 1, question 0]

Context

2 Word windows (filtered): *n* words on either side removing some words (e.g. function words, some very frequent content words). Stop-list or by POS-tag.
Example: n=2 (5 words window), stop-list:

| *The prime* ***minister*** *acknowledged the* | *question.*

minister [prime 1, acknowledged 1, question 0]

Context

3 Lexeme window (filtered or unfiltered); as above but using stems.
Example: n=2 (5 words window), stop-list:

| *The prime **minister** acknowledged the* | *question.*

minister [prime 1, acknowledge 1, question 0]

Context

4 Dependencies (directed links between heads and dependents). Context for a lexical item is the dependency structure it belongs to (various definitions).
Example:

*The prime **minister** acknowledged the question.*

minister [prime_a 1, acknowledge_v 1]

minister [prime_a_mod 1, acknowledge_v_subj 1]

minister [prime_a 1, acknowledge_v+question_n 1]

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

Dependency vectors

word (Subj)	**word (Dobj)**
come_v	use_v
mean_v	say_v
go_v	hear_v
speak_v	take_v
make_v	speak_v
say_v	find_v
seem_v	get_v
follow_v	remember_v
give_v	read_v
describe_v	write_v
get_v	utter_v
appear_v	know_v
begin_v	understand_v
sound_v	believe_v
occur_v	choose_v

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}...

Characteristic model

- Weights given to the vector components express how *characteristic* a given context is for word w.
- Pointwise Mutual Information (PMI)

$$PMI(w,c) = \log \frac{P(w,c)}{P(w)P(c)} = \log \frac{P(w)P(c|w)}{P(w)P(c)} = \log \frac{P(c|w)}{P(c)}$$

$$P(c) = \frac{f(c)}{\sum_k f(c_k)}, \qquad P(c|w) = \frac{f(w,c)}{f(w)},$$

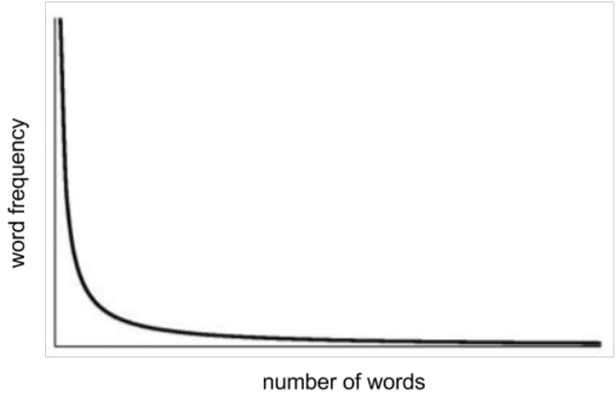
$$PMI(w,c) = \log \frac{f(w,c)\sum_k f(c_k)}{f(w)f(c)}$$

$f(w,c)$: frequency of word w in context c
$f(w)$: frequency of word w in all contexts
$f(c)$: frequency of context c

What semantic space?

- Entire vocabulary.
  - + All information included – even rare contexts
  - - Inefficient (100,000s dimensions). Noisy (e.g. *002.png|thumb|right|200px|graph_n*). Sparse
- Top *n* words with highest frequencies.
  - + More efficient (2000-10000 dimensions). Only ‘real’ words included.
  - - May miss out on infrequent but relevant contexts.

Word frequency: Zipfian distribution



What semantic space?

- Entire vocabulary.
  - + All information included – even rare contexts
  - - Inefficient (100,000s dimensions). Noisy (e.g. *002.png|thumb|right|200px|graph_n*). Sparse.
- Top *n* words with highest frequencies.
  - + More efficient (2000-10000 dimensions). Only ‘real’ words included.
  - - May miss out on infrequent but relevant contexts.

What semantic space?

- Singular Value Decomposition (SVD): the number of dimensions is reduced by exploiting redundancies in the data.
  - + Very efficient (200-500 dimensions). Captures generalisations in the data.
  - - SVD matrices are not interpretable.
- Non-negative matrix factorization (NMF)
  - Similar to SVD in spirit, but performs factorization differently

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, PMI-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.

- Assume only keep open-class words.
- **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.

- $\emptyset$ $\emptyset$ **the** major difference
- $\emptyset$ 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 3	possibly 2	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 (PPMIs):**

difference 0	impossible 0	thing 0
get 0	major 0	turns 0
go 0.70	possibly 0.70	usually 0.70
goes 1	repair 0	wrong 0.40