

Distributional Semantics

Raffaella Bernardi

University of Trento

Acknowledgments

Credits: Some of the slides of today lecture are based on earlier DS courses taught by Marco Baroni, Stefan Evert, Aurelie Herbelot, Alessandro Lenci, Gemma Boleda and Roberto Zamparelli.

Criticisms to Formal Semantics

FS is not a Cognitive Semantics

- ▶ Poor representation of the semantic content of words.
- ▶ Do humans have sets in their heads? (Cognitive plausibility.)
- ▶ Is there truth?
- ▶ Where do models come from?

But these issues were not in the agenda of Formal semanticists.

Background

Recall: Frege and Wittgenstein

Frege:

1. Linguistic signs have a reference and a sense:
 - (i) "Mark Twain is Mark Twain" [same ref. same sense]
 - (ii) "Mark Twain is Samuel Clemens". [same ref. diff. sense]
2. Both the sense and reference of a sentence are built compositionally.

Lead to the Formal (or Denotational) Semantics studies of natural language that focused on "meaning" as "reference".
[Seen last time]

Wittgenstein's claims brought philosophers of language to focus on "meaning" as "sense" leading to the "language as use" view.

Background

Content vs. Grammatical words

The “language as use” school has focused on content words meaning. vs. Formal semantics school has focused mostly on the grammatical words and in particular on the behaviour of the “logical words”.

- ▶ **content words**: are words that carry the content or the meaning of a sentence and are open-class words, e.g. *noun, verbs, adjectives* and most *adverbs*.
- ▶ **grammatical words**: are words that serve to express grammatical relationships with other words within a sentence; they can be found in almost any utterance, no matter what it is about, e.g. such as *articles, prepositions, conjunctions, auxiliary verbs*, and *pronouns*.

Among the latter, one can distinguish the **logical words**, viz. those words that correspond to logical operators

Background

Recall: Formal Semantics: reference

The main questions are:

1. What does a given *sentence* mean?
2. How is its meaning built?
3. How do we infer some piece of information out of another?

Logic view answers: The meaning of a sentence 1. is its truth value, 2. is built from the meaning of its words; 3. is represented by a FOL formula, hence inferences can be handled by logic entailment.

Moreover,

- ▶ The meaning of words is based on the *objects* in the domain – it's the set of entities, or set of pairs/triples of entities, or set of properties of entities.
- ▶ Composition is obtained by function-application and abstraction
- ▶ Syntax guides the building of the meaning representation.

Background

Distributional Semantics: sense

The main questions have been:

1. What is the sense of a given *word*?
2. How can it be induced and represented?
3. How do we relate word senses (synonyms, antonyms, hyperonym etc.)?

Well established answers:

1. The sense of a word can be given by its use, viz. by the *contexts* in which it occurs;
2. It can be induced from (either raw or parsed) corpora and can be represented by *vectors*.
3. *Cosine similarity* captures synonyms (as well as other semantic relations).

Distributional Semantics

pioneers

1. Intuitions in the '50:
 - ▶ Wittgenstein (1953): word usage can reveal semantics flavor (context as physical activities).
 - ▶ Harris (1954): words that occur in similar (linguistic) context tend to have similar meanings.
 - ▶ Weaver (1955): co-occurrence frequency of the context words near a given target word is important for WSD for MT.
 - ▶ Firth (1957): “you shall know a word by the company it keeps”
2. Deerwster et al. (1990): put these intuitions at work.

The distributional hypothesis in everyday life

McDonald & Ramscar (2001)

- ▶ He filled the **wampimuk** with the substance, passed it around and we all drunk some
- ▶ We found a little, hairy **wampimuk** sleeping behind the tree

Just from the contexts a human could guess the meaning of “wampimuk”.

Distributional Semantics

weak and strong version: Lenci (2008)

- ▶ **Weak:** a quantitative method for semantic analysis and lexical resource induction
- ▶ **Strong:** A cognitive hypothesis about the form and origin of semantic representations

Distributional Semantics

Main idea in a picture: The sense of a word can be given by its use (context!).

hotels . 1. God of the morning star 5. How does your garden
, or meditations on the morning star . But we do , as a matte
sing metaphors from the morning star , that the should be pla
ilky Way appear and the morning star rising like a diamond be
and told them that the morning star was up in the sky , they
ed her beauteous as the morning star , Fixed in his purpose
g star is the brightest morning star . Suppose that ' Cicero
radise on the beam of a morning star and drank it out of gold
ey worshipped it as the morning star . Their Gods at on stool
things . The moon , the morning star , and certain animals su

flower , He lights the evening star . " Daisy 's eyes filled
he planet they call the evening star , the morning star . Int
fear it . The punctual evening star , Worse , the warm hawth
of morning star and of evening star . And the fish worship t
are Fair sisters of the evening star , But wait -- if not tod
ie would shine like the evening star . But Richardson 's own
na . As the morning and evening star , the planet Venus was u
l appear as a brilliant evening star in the SSW . I have used
o that he could see the evening star , a star has also been d
il it reappears as an ' evening star ' at the end of May . Tr

Collecting context counts for target word **dog**

The **dog** barked in the park.

The **owner** of the **dog** put him
on the **leash** since he **barked**.

bark	++
park	+
owner	+
leash	+

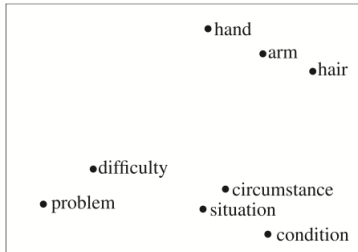
The co-occurrence matrix

	leash	walk	run	owner	pet	bark
dog	3	5	2	5	3	2
cat	0	3	3	2	3	0
lion	0	3	2	0	1	0
light	0	0	0	0	0	0
bark	1	0	0	2	1	0
car	0	0	1	3	0	0

Distributional Semantics Model

It's a quadruple $\langle B, A, S, V \rangle$, where:

- ▶ B is the set of “basis elements” – the dimensions of the space.
- ▶ A is a lexical association function that assigns co-occurrence frequency of words to the dimensions.
- ▶ V is an optional transformation that reduces the dimensionality of the semantic space.
- ▶ S is a similarity measure.

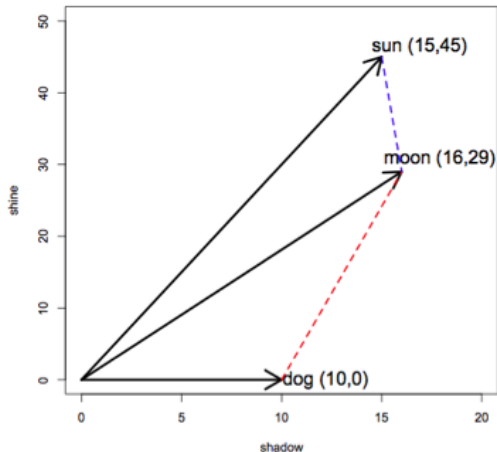


Distributional Semantics Model

Toy example: vectors in a 2 dimensional space

$B = \{shadow, shine, \}$; $A =$ co-occurrence frequency;

S : Euclidean distance. Target words: “moon”, “sun”, and “dog”.



Distributional Semantics Model

Two dimensional space representation

$\vec{moon}=(16,29)$, $\vec{sun}=(15,45)$, $\vec{dog}=(10,0)$ together in a space representation (a matrix dimensions \times target-words):

$$\begin{bmatrix} 16 & 15 & 10 \\ 29 & 45 & 0 \end{bmatrix}$$

The most commonly used representation is the transpose matrix (A^T): target-words \times dimensions:

	shine	shadow
\vec{moon}	16	29
\vec{sun}	15	45
\vec{dog}	10	0

The dimensions are also called “features” or “context”.

A distributional cat (from the British National Corpus)

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

0.115 english-N	0.075 teach-V	0.064 universal-A
0.114 written-A	0.075 communication-N	0.064 aspect-N
0.109 grammar-N	0.074 knowledge-N	0.064 german-N
0.106 translate-V	0.074 polish-A	0.063 artificial-A
0.102 teaching-N	0.072 speaker-N	0.063 logic-N
0.097 literature-N	0.071 convey-V	0.061 understanding-N
0.096 english-A	0.070 theoretical-A	0.061 official-A
0.096 acquisition-N	0.069 curriculum-N	0.061 formal-A
0.095 communicate-V	0.068 pupil-N	0.061 complexity-N
0.093 native-A	0.068 level-A	0.060 gesture-N
0.089 everyday-A	0.067 assessment-N	0.060 african-A
0.088 learning-N	0.067 use-N	0.060 eg-A
0.084 meaning-N	0.067 tongue-N	0.060 express-V
0.083 french-N	0.067 medium-N	0.059 implication-N
0.082 description-N	0.067 spanish-A	0.058 distinction-N
0.079 culture-N	0.066 speech-N	0.058 barrier-N
0.078 speak-V	0.066 learn-V	0.057 cultural-A
0.078 foreign-A	0.066 interaction-N	0.057 literary-A
0.077 classroom-N	0.065 expression-N	0.057 variation-N
0.077 command-N	0.064 sign-N	...

0.129 chocolate-N	0.083 sweet-A	0.071 salad-N
0.122 slice-N	0.081 mix-N	0.071 piece-N
0.109 tin-N	0.080 mixture-N	0.070 line-V
0.109 pie-N	0.079 rice-N	0.070 dry-V
0.103 sandwich-N	0.078 nut-N	0.069 round-A
0.103 decorate-V	0.076 tomato-N	0.068 egg-N
0.099 cream-N	0.076 knife-N	0.068 cooking-N
0.098 fruit-N	0.075 potato-N	0.066 lb-N
0.097 recipe-N	0.075 oz-N	0.066 fat-N
0.097 bread-N	0.075 cook-N	0.064 top-N
0.096 oven-N	0.075 top-V	0.063 spread-V
0.094 birthday-N	0.074 coffee-N	0.063 chip-N
0.090 wedding-N	0.073 christmas-N	0.063 cut-V
0.087 sugar-N	0.073 ice-N	0.062 sauce-N
0.086 cheese-N	0.073 orange-N	0.062 turkey-N
0.086 tea-N	0.073 layer-N	0.061 milk-N
0.085 butter-N	0.072 packet-N	0.061 plate-N
0.085 eat-V	0.072 roll-N	0.060 remaining-A
0.084 apple-N	0.071 brush-V	0.060 hint-N
0.083 wrap-V	0.071 meat-N	...

0.093 coloured-A	0.065 pick-V	0.057 hand-V
0.092 paper-N	0.065 co-N	0.057 phil-N
0.089 stroke-N	0.064 palm-N	0.056 wilson-N
0.089 margin-N	0.064 writing-N	0.056 silver-N
0.089 tip-N	0.064 jean-N	0.056 terror-N
0.085 seize-V	0.064 literary-A	0.055 lower-V
0.077 pig-N	0.063 writer-N	0.055 tap-V
0.077 ltd-A	0.063 write-V	0.055 light-A
0.076 drawing-N	0.063 script-N	0.055 packet-N
0.074 electronic-A	0.063 ash-N	0.055 load-V
0.072 concrete-A	0.062 desk-N	0.054 cigarette-N
0.072 portrait-N	0.062 elegant-A	0.054 anxiety-N
0.071 sheep-N	0.061 pause-V	0.054 program-N
0.068 pocket-N	0.061 brush-N	0.054 complex-N
0.066 code-N	0.060 marine-A	0.054 ball-N
0.066 flow-V	0.060 infant-N	0.053 rabbit-N
0.066 gardener-N	0.059 tape-N	0.053 precious-A
0.066 sheet-N	0.059 collapse-N	0.052 eg-A
0.066 straw-N	0.058 cry-N	0.052 thanks-N
0.066 outline-N	0.057 delighted-A	...

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).

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: $n=2$ (window of size 2):

... 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: $n=2$ (window of size 2):

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

What is “context”?

DOC1: The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

What is “context”?

Documents

DOC1: The silhouette of the **sun** beyond a wide-open bay on the lake; the **sun** still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

What is “context”?

All words in a wide window

DOC1: The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

What is “context”?

Content words only

DOC1: The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

What is “context”?

Content words in a narrower window

DOC1: The silhouette of the sun beyond a wide-open bay on the lake; the sun still glitters although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

What is “context”?

POS-coded content lemmas

DOC1: The silhouette-n of the sun beyond a wide-open-a bay-n on the lake-n; the sun still glitter-v although evening-n has arrive-v in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

What is “context”?

POS-coded content lemmas filtered by syntactic path to the target

DOC1: The silhouette-n of the sun beyond a wide-open bay on the lake; the sun still glitter-v although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

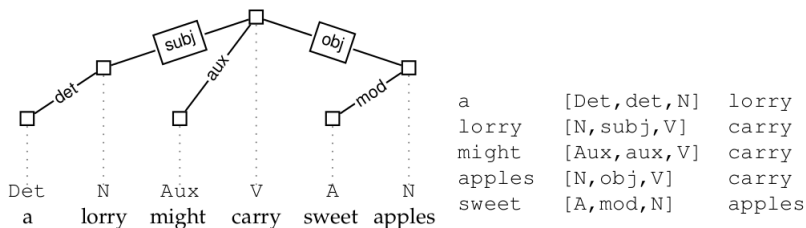
What is “context”?

... with the syntactic path encoded as part of the context

DOC1: The silhouette-n_ppdep of the sun beyond a wide-open bay on the lake; the sun still glitter-v_subj although evening has arrived in Kuhmo. It's midsummer; the living room has its instruments and other objects in each of its corners.

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).



Parsed vs unparsed data: examples

word (unparsed)

meaning_n
derive_v
dictionary_n
pronounce_v
phrase_n
latin_j
ipa_n
verb_n
mean_v
hebrew_n
usage_n
literally_r

word (parsed)

or_c+phrase_n
and_c+phrase_n
syllable_n+of_p
play_n+on_p
etymology_n+of_p
portmanteau_n+of_p
and_c+deed_n
meaning_n+of_p
from_p+language_n
pron_rel_+utter_v
for_p+word_n
in_p+sentence_n

Context weighting

- ▶ Raw context counts typically transformed into scores
- ▶ In particular, association measures to give more weight to contexts that are more significantly associated with a target word
- ▶ General idea: the less frequent the target word and (more importantly) the context element are, the higher the weight given to their observed co-occurrence count should be (because their expected chance co-occurrence frequency is low)
 - ▶ Co-occurrence with frequent context element *time* is less informative than co-occurrence with rarer *tail*
- ▶ Different measures – e.g., Mutual Information, Log Likelihood Ratio – differ with respect to how they balance raw and expectation-adjusted co-occurrence frequencies
 - ▶ Positive Point-wise Mutual Information widely used and pretty robust

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

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

$$pmi_{wc} = \log \frac{p(x, y)}{p(x)p(y)} = \log\left(\frac{f_{wc} * f_{total}}{f_w * f_c}\right) \quad (1)$$

- ▶ 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?

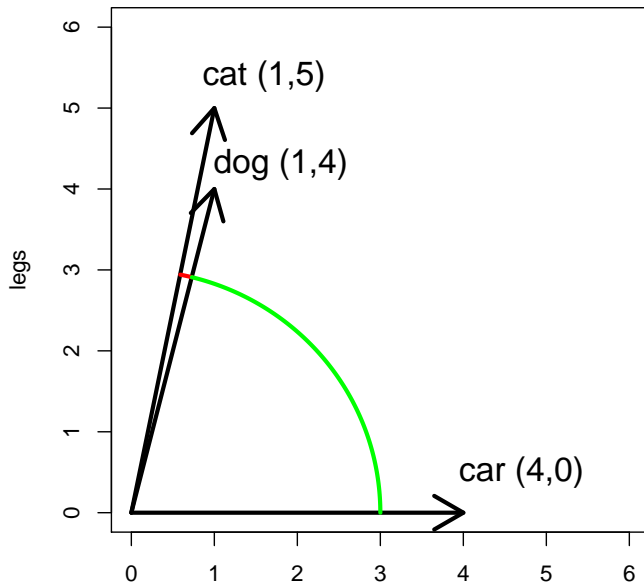
- ▶ Singular Value Decomposition (LSA – Landauer and Dumais, 1997): the number of dimensions is reduced by exploiting redundancies in the data. A new dimension might correspond to a generalisation over several of the original dimensions (e.g. the dimensions for *car* and *vehicle* are collapsed into one).
 - ▶ + Very efficient (200-500 dimensions). Captures generalisations in the data.
 - ▶ - SVD matrices are not interpretable.
- ▶ Other, more esoteric variants...

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.

Distributional Semantics Models

Recall from LD: Similarity measure: Angle



Distributional Semantics Model

Similarity measure: cosine similarity

Cosine is the most common similarity measure in distributional semantics. The similarity of two words is computed as the cosine similarity of their corresponding vectors \vec{x} and \vec{y} or, equivalently, the cosine of the angle between \vec{x} and \vec{y} is:

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x}}{|\vec{x}|} \cdot \frac{\vec{y}}{|\vec{y}|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

- ▶ x_i is the weight of dimension i in x .
- ▶ y_i is the the weight of dimension i in y .
- ▶ $|\vec{x}|$ and $|\vec{y}|$ are the lengths of \vec{x} and \vec{y} . Hence, $\frac{\vec{x}}{|\vec{x}|}$ and $\frac{\vec{y}}{|\vec{y}|}$ are the normlized (unit) vectors.

Cosine ranges from 1 for parallel vectors (perfectly correlated words) to 0 for orthogonal (perpendicular) words/vectors.

In sum: Building a DSM

The “linguistic” steps

Pre-process a corpus (to define targets and contexts)



Select the targets and the contexts

The “mathematical” steps

Count the target-context co-occurrences



Weight the contexts (optional, but recommended)



Build the distributional matrix



Reduce the matrix dimensions (optional)



Compute the vector distances on the (reduced) matrix

Building a DSM

Corpus pre-processing

- ▶ Minimally, corpus must be tokenized
- ▶ POS tagging, lemmatization, dependency parsing. . .
- ▶ Trade-off between deeper linguistic analysis and
 - ▶ need for language-specific resources
 - ▶ possible errors introduced at each stage of the analysis
 - ▶ more parameters to tune

Building a DSM

different pre-processing – Nearest neighbours of *walk*

tokenized BNC

- ▶ stroll
- ▶ walking
- ▶ walked
- ▶ go
- ▶ path
- ▶ drive
- ▶ ride
- ▶ wander
- ▶ sprinted
- ▶ sauntered

lemmatized BNC

- ▶ hurry
- ▶ stroll
- ▶ stride
- ▶ trudge
- ▶ amble
- ▶ wander
- ▶ walk-nn
- ▶ walking
- ▶ retrace
- ▶ scuttle

Building a DSM

different window size – Nearest neighbours of *dog*

2-word window in BNC

- ▶ cat
- ▶ horse
- ▶ fox
- ▶ pet
- ▶ rabbit
- ▶ pig
- ▶ animal
- ▶ mongrel
- ▶ sheep
- ▶ pigeon

30-word window in BNC

- ▶ kennel
- ▶ puppy
- ▶ pet
- ▶ bitch
- ▶ terrier
- ▶ rottweiler
- ▶ canine
- ▶ cat
- ▶ to bark
- ▶ Alsatian

Building a DSM

Syntagmatic relations uses

Syntagmatic relations as (a) **context-filtering functions**: only those words that are linked to the targets by a certain relation are selected, or as (b) **context-typing functions**: relation define the dimensions.
E.g.:

“A dog bites a man. A man bites a dog. A dog bites a man.”

		dog	man
(a) window-based	bite	3	3
(b) dependency based	bite _{sub}	2	1
	bite _{obj}	1	2

- ▶ (a) Dimension-filtering based on (a1) window: e.g. Rapp, 2003, Infomap NLP; (a2) dependency: Padó & Lapata 2007.
- ▶ (b) Dimension-typing based on (b1) window: HAL; (b2) dependency: Grefenstette 1994, Lin 1998, Curran & Moens 2002, Baroni & Lenci 2009.

Evaluation on Lexical meaning

(More on this with Carlo)

Developers of semantic spaces typically want them to be “general-purpose” models of semantic similarity

- ▶ Words that share many contexts will correspond to concepts that share many attributes (*attributional similarity*), i.e., concepts that are taxonomically similar:
 - ▶ Synonyms (*rhino/rhinoceros*), antonyms and values on a scale (*good/bad*), co-hyponyms (*rock/jazz*), hyper- and hyponyms (*rock/basalt*)
- ▶ Taxonomic similarity is seen as the fundamental semantic relation, allowing categorization, generalization, inheritance
- ▶ Evaluation focuses on tasks that measure taxonomic similarity

Evaluation on Lexical meaning

synonyms

DSM captures pretty well synonyms. DSM used over TOEFL test:

- ▶ Foreigners average result: 64.5%
- ▶ Macquarie University Staff (Rapp 2004):
 - ▶ Ave. not native speakers: 86.75%
 - ▶ Ave. native speakers: 97.75%
- ▶ DM:
 - ▶ DM (dimension: words): 64.4%
 - ▶ Padó and Lapata's dependency-filtered model: 73%
 - ▶ Rapp's 2003 SVD-based model trained on lemmatized BNC: 92.5%
- ▶ Direct comparison in Baroni and Lenci 2010
 - ▶ Dependency-filtered: 76.9%
 - ▶ Dependency-typing: 75.0%
 - ▶ Co-occurrence window: 69.4%

Evaluation on Lexical meaning

Other classic semantic similarity tasks

Also used for:

- ▶ The Rubenstein/Goodenough norms: modeling semantic similarity judgments
- ▶ The Almuhareb/Poesio data-set: clustering concepts into categories
- ▶ The Hodgson semantic priming data
- ▶ Baroni & Lenci 2010: general-purpose model for:
 - ▶ concept categorization (car ISA vehicle),
 - ▶ selectional preferences (eat chocolate vs *eat sympathy),
 - ▶ relation classification (exam-anxiety CAUSE-EFFECT relation),
 - ▶ salient properties (car-wheels).
 - ▶ ...

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.

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.
- ▶ So if we use the words 'robots from Mars' to talk about cats, all is fine (see whales and fish).

From counting co-occurrences to predicting the word

Skip-Gram (T. Mikolov)

Nowadays most renowned model Word2Vec. You will use it with LD.

Main change: From counting co-occurrences to predict the word in the context.

(Last 2 years: Bert and ELMo – you will look at them with AH)

Word2Vec

Word2vec is a particularly computationally-efficient predictive model for learning word embeddings from raw text. It comes in two flavors:

- ▶ the Skip-Gram model: predicts source context-words from the target words [better for larger datasets]
- ▶ the Continuous Bag-of-Words model (CBOW): predicts target words (e.g. 'mat') from source context words ('the cat sits on the') [useful for smaller datasets]

Word2Vec

Skip-Gram

Given a specific word in the middle of a sentence (the input word), look at the words nearby and pick one at random. The network is going to tell us the probability for every word in our vocabulary of being the nearby word that we chose.

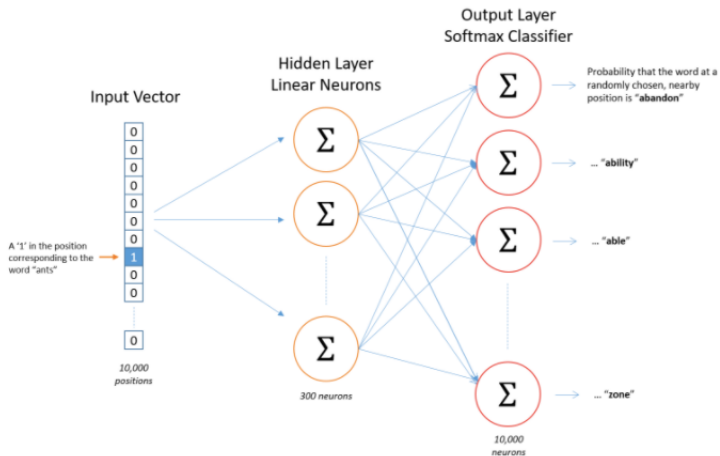
1. build the vocabulary, e.g .10K unique words
2. each word is represented as a “one-hot vector”: 10K-d, all 0 and one 1.
3. output: a 10K-d vector whose value are the probability that a randomly selected nearby word is that vocabulary word.

When training this network on word pairs, the input is a one-hot vector representing the input word and the training output is also a one-hot vector representing the output word.

But when you evaluate the trained network on an input word, the output vector will actually be a probability distribution

Word2Vec

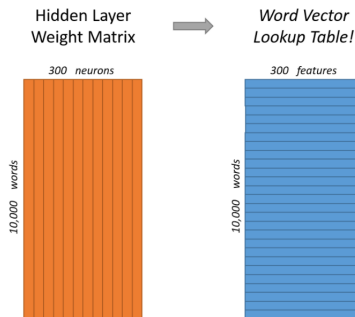
Skip-Gram: architecture



Word2Vec: Skip-Gram

Word Vectors

E.g., we are learning word vectors with 300 features. So the hidden layer is going to be represented by a weight matrix with 10K rows (one for every word in our vocabulary) and 300 columns (one for every hidden neuron).



If you look at the rows of this weight matrix, these are actually what will be our *word vectors* (or *word embeddings*)!

Applications

- ▶ IR: Semantic spaces might be pursued in IR within the broad topic of “semantic search”
- ▶ DSM as supplementary resource in e.g.,:
 - ▶ Question answering (Tomás & Vicedo, 2007)
 - ▶ Bridging coreference resolution (Poesio et al., 1998, Versley, 2007)
 - ▶ Language modeling for speech recognition (Bellegarda, 1997)
 - ▶ Textual entailment (Zhitomirsky-Geffet and Dagan, 2009)

Online query tool

- ▶ To query already built distributional semantics spaces: <http://clic.cimec.unitn.it/infomap-query/>
- ▶ To build semantic spaces <http://clic.cimec.unitn.it/composes/toolkit/>
- ▶ Semantic space based on distributional similarity and Latent Semantic Analysis (LSA), further complemented with semantic relations extracted from WordNet: <http://swoogle.umbc.edu/SimService/>
- ▶ Snaut <http://meshugga.ugent.be/snaut/>. It allows to measure semantic distance between words or documents and explore distributional semantics models through a convenient interface.
- ▶ Available count/predict spaces: <http://clic.cimec.unitn.it/composes/semantic-vectors.html>

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...

Conclusion

So far

The main questions have been:

1. What is the sense of a given *word*?
2. How can it be induced and represented?
3. How do we relate word senses (synonyms, antonyms, hyperonym etc.)?

Well established answers:

1. The sense of a word can be given by its use, viz. by the *contexts* in which it occurs;
2. It can be induced from (either raw or parsed) corpora and can be represented by *vectors*.
3. *Cosine similarity* captures synonyms (as well as other semantic relations).

Conclusion

Next time

- ▶ Can vectors representing phrases be extracted too?
- ▶ What about compositionality of word sense?
- ▶ How do we “infer” some piece of information out of another?

Next steps

- ▶ Next frontal class on the 07.11
- ▶ On the 4.11 **** you**** will tell me what you have learned with Luca about Word2Vec and relate the two courses.

References

Tutorials

- ▶ Slides taken from the nice tutorial by Chris McCormick:
`http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/`
- ▶ Another nice tutorial: `https://www.tensorflow.org/tutorials/word2vec`
- ▶ S. Evert and A. Lenci. “Distributional Semantic Models” ESSLLI 2009. `http://wordspace.collocations.de/doku.php/course:esslli2009:start`

References

Papers

- ▶ M. Baroni, G. Dinu and G. Kruszewski. 2014. *Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors* Proceedings of ACL 2014 (52nd Annual Meeting of the Association for Computational Linguistics), East Stroudsburg PA: ACL, 238-247.
- ▶ Mandera, Keuleers, & Brysbaert, *Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation*. In press
- ▶ Levy and Goldberg 2014
- ▶ Peter D. Turney, and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics". *In Journal of Artificial Intelligence Research*. 37, (2010), 141-188.
- ▶ G. Strang, "Introduction to Linear Algebra". Wellesley Cambridge Press. 2009
- ▶ K. Gimpel "Modeling Topics" 2006

Background: Vector and Matrix

Vector Space

A vector space is a mathematical structure formed by a collection of vectors: objects that may be added together and multiplied (“scaled”) by numbers, called scalars in this context.

Vector an n-dimensional vector is represented by a column:

$$\begin{bmatrix} v_1 \\ \dots \\ v_n \end{bmatrix}$$

or for short as $\vec{v} = (v_1, \dots, v_n)$.

Background: Vectors

Dot product or inner product

$$\vec{v} \cdot \vec{w} = (v_1 w_1 + \dots + v_n w_n) = \sum_{i=1}^n v_i w_i$$

Example We have three goods to buy and sell, their prices are (p_1, p_2, p_3) (price vector \vec{p}). The quantities we are buy or sell are (q_1, q_2, q_3) (quantity vector \vec{q} , their values are positive when we sell and negative when we buy.) Selling the quantity q_1 at price p_1 brings in $q_1 p_1$. The total income is the *dot product*:

$$\vec{q} \cdot \vec{p} = (q_1, q_2, q_3) \cdot (p_1, p_2, p_3) = q_1 p_1 + q_2 p_2 + q_3 p_3$$

Background: Vector

Length and Unit vector

Length $\|\vec{v}\| = \sqrt{\vec{v} \cdot \vec{v}} = \sqrt{\sum_{i=1}^n v_i^2}$

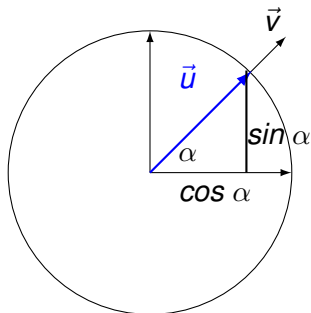
Unit vector is a vector whose length equals one.

$$\vec{u} = \frac{\vec{v}}{\|\vec{v}\|}$$

is a unit vector in the same direction as \vec{v} . (normalized vector)

Background: Vector

Unit vector



$$\vec{u} = \frac{\vec{v}}{\|\vec{v}\|} = (\cos \alpha, \sin \alpha)$$

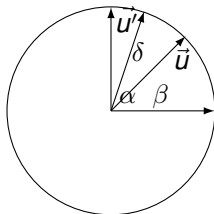
Background: Vector

Cosine formula

Given δ the angle formed by the two unit vectors \vec{u} and \vec{u}' , s.t.

$$\vec{u} = (\cos \beta, \sin \beta) \text{ and } \vec{u}' = (\cos \alpha, \sin \alpha)$$

$$\vec{u} \cdot \vec{u}' = (\cos \beta)(\cos \alpha) + (\sin \beta)(\sin \alpha) = \cos(\beta - \alpha) = \cos \delta$$



Given two arbitrary vectors v and w :

$$\cos \delta = \frac{\vec{v}}{\|\vec{v}\|} \cdot \frac{\vec{w}}{\|\vec{w}\|}$$

The bigger the angle δ , the smaller is $\cos \delta$; $\cos \delta$ is never bigger than 1 (since we used unit vectors) and never less than -1. It's 0 when the angle is 90°