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?
- 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 compositionaly.

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

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

- **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 *quantifiers*, *prepositions*, *conjunctions*, *negation*, *auxiliary* verbs, and *pronouns*.

The formal semantics school has focused mostly on the grammatical words and in particular on the behaviour of the “logical words” (eg. quantifiers, negation).
The “language as use” school has focused on content words meaning.

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.

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

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”.

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	+

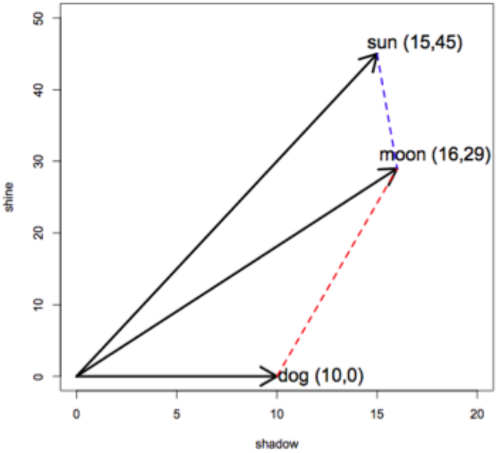
The co-occurrence matrix

	leash	shadow	shine	owner	park	bark
dog	3	10	0	5	3	2
cat	0	3	3	2	3	0
sun	0	15	45	0	1	0
moon	0	16	29	0	0	0
bark	1	0	0	2	1	0
car	0	0	1	3	0	0

leash, shadow, shine etc. are called the context/features/basis elements. They form the semantic space.
dog, cat, sun etc. are called the target words represented by vectors. They are said to “live in the space”, in this case in a 6-dimensional space.

Toy example: vectors in a 2-dimensional space

$B = \{shadow, shine,\}$; A= co-occurency frequency;
S: Euclidean distance. Target words: “moon”, “sun”, and “dog”.



A distributional cat (from the British National Corpus)

0.124 pet-N
0.123 mouse-N
0.099 rat-N
0.097 owner-N
0.096 dog-N
0.092 domestic-A
0.090 wild-A
0.090 duck-N
0.087 tail-N
0.084 leap-V
0.084 prey-N
0.083 breed-N
0.080 rabbit-N
0.078 female-A
0.075 fox-N
0.075 basket-N
0.075 animal-N
0.074 ear-N
0.074 chase-V
0.074 smell-V

0.074 tiger-N
0.073 jump-V
0.073 tom-N
0.073 fat-A
0.071 spell-V
0.071 companion-N
0.070 lion-N
0.068 breed-V
0.068 signal-N
0.067 bite-V
0.067 spring-V
0.067 detect-V
0.067 bird-N
0.066 friendly-A
0.066 odour-N
0.066 hunting-N
0.066 ghost-N
0.065 rub-V
0.064 predator-N
0.063 pig-N

0.063 hate-V
0.063 asleep-A
0.063 stance-N
0.062 unfortunate-A
0.061 naked-A
0.061 switch-V
0.061 encounter-V
0.061 creature-N
0.061 dominant-A
0.060 black-A
0.059 chocolate-N
0.058 giant-N
0.058 sensitive-A
0.058 canadian-A
0.058 toy-N
0.058 milk-N
0.057 human-N
0.057 devil-N
0.056 smell-N
...

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

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

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

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

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

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

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

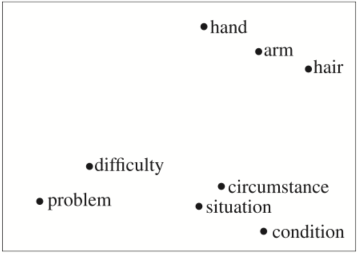
0.065 pick-V
0.065 co-N
0.064 palm-N
0.064 writing-N
0.064 jean-N
0.064 literary-A
0.063 writer-N
0.063 write-V
0.063 script-N
0.063 ash-N
0.062 desk-N
0.062 elegant-A
0.061 pause-V
0.061 brush-N
0.060 marine-A
0.060 infant-N
0.059 tape-N
0.059 collapse-N
0.058 cry-N
0.057 delighted-A

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

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.



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.

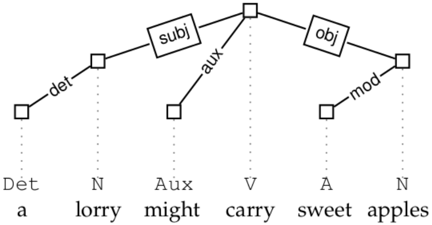
What is “context”?

. . . 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).



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

- 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

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):

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

  - Derivatives such PPMI, PLMI, etc.

What semantic space?

- Entire vocabulary.
  - + All information included – even rare, but important contexts
  - - Inefficient (100,000s dimensions). Noisy (e.g. *002.png—thumb—right—200px—graph_n*)
- Top *n* words with highest frequencies.
  - + More efficient (5000-10000 dimensions). Only ‘real’ words included.
  - - May miss out on infrequent but relevant contexts.

What semantic space?

- Singular Value Decomposition (SVD) (eg used in 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.

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.

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, $\frac{\vec{x}}{|x|}$ and $\frac{\vec{y}}{|y|}$ are the normilized (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)

$\Downarrow$

Select the targets and the contexts

The "mathematical" steps

Count the target-context co-occurrences

$\Downarrow$

Weight the contexts (optional, but recommended)

$\Downarrow$

Build the distributional matrix

$\Downarrow$

Reduce the matrix dimensions (optional)

$\Downarrow$

Compute the vector distances on the (reduced) matrix

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%

From counting co-occurrences to predicting the word

Skip-Gram (T. Mikolov)

Nowadays most renowned model Word2Vec.
Main change: From counting co-occourences to predict the word in the context.
(Last 2 years: BERT and ELMo)

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

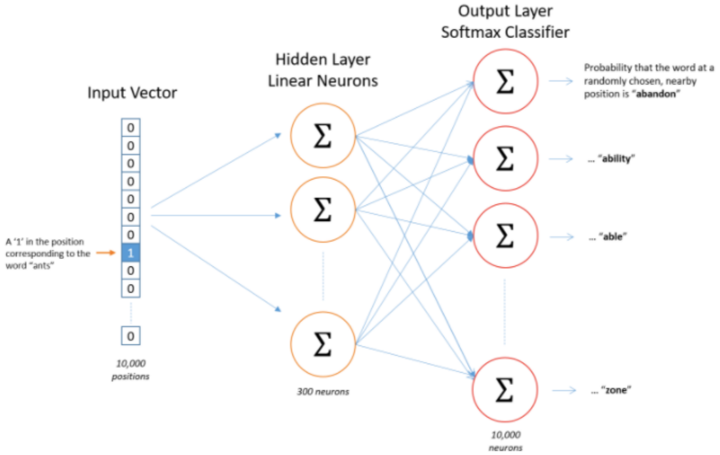
Given a specific word in the middle of a sentence (the input word), look at the words nearby and pick one at random. The network is going to tell us the probability for every word in our vocabulary of being the nearby word that we chose.

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

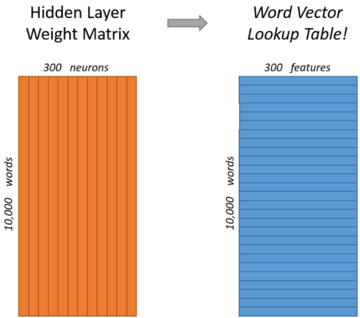
Skip-Gram: architecture



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

- 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.
- Evaluation against psycholinguistic data shows that DS can model at least *some* phenomena.
- A powerful computational semantics tool, with surprising results.