Computational Semantics and Pragmatics

Autumn 2013

Raquel Fernández
Institute for Logic, Language & Computation
University of Amsterdam

Issues in Lexical Semantics

- How to characterise word meaning:
  - by their contribution to sentence meaning?
  - with semantic primitives? logical relations? plain defintions?
  - with structured templates including "qualia" components?
- Psychological theories of concepts / word meaning:
  - concepts are fuzzy (can't be captured with necessary properties)
  - they give rise to typicality effects
- Ambiguity: most words have several senses
  - does it make sense to enumerate them all in the lexicon?
  - the generative lexicon can capture regular polysemy to some extent
  - continuum between regular polysemy, polysemy, homonomy...

In NLP, the task of word sense disambiguation (WSD) takes for granted an inventory of word senses (e.g. WordNet). But the inventory and the notion of word sense itself do not seem well-founded.

Towards More Objective Representations

Given the lack of clear principles for characterising word meanings and the lexicon, some researchers started to be sceptical about the notion of word meaning itself...

Adam Kilgarriff (1997) I don't believe in word senses, *Computers and the Humanities*, 31:91–113.

Patrick Hanks (2000) Do Word Meanings Exist?, *Computers and the Humanities*, 34:205-215.

Their alternative proposal is that word meaning depends, at least in part, on the contexts in which words are used:

$\Rightarrow$ usage-based view of meaning.

An example by Stefan Evert: what's the meaning of *'bardiwac'*?

- He handed her her glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- Nigel staggered to his feet, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.
- I dined on bread and cheese and this excellent bardiwac.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

$\Rightarrow$ *'bardiwac'* is a heavy red alcoholic beverage made from grapes

Distributional Sematic Models (DSMs) or Vector Space Models aim to make precise the intuition that context tells us a good deal about word meaning.

Distributional Semantic Models

DSMs are motivated by the so-called Distributional Hypothesis:

> "The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear." [Z. Harris (1954) *Distributional Structure*]

- DSMs make use of mathematical and computational techniques to turn the informal DH into empirically testable semantic models.
- Contextual semantic representations from data about language usage: an abstraction over the linguistic contexts in which a word is encountered.

	see	use	hear	...
boat	39	23	4	...
cat	58	4	4	...
dog	83	10	42	...

→ Distributional vector of 'dog': $x_{dog} = (83, 10, 42, \ldots)$

Origins of Distributional Semantics

- Currently, distributional semantics is extremely popular in computational linguistics.
- However, its origins are grounded in the linguistic tradition:
  - American *structural linguistics* during the 1940s and 50s, especially the figure of Zellig Harris (influenced by Sapir and Bloomfield).
- Harris proposed the method of distributional analysis as a scientific methodology for linguistics:
  - introduced for phonology, then methodology for all linguistic levels.
- Structuralists don't consider meaning an *explanans* in linguistics: too subjective and vague a notion to be methodologically sound.
  - linguistic units need to be determined by formal means: by their distributional structure.

Origins of Distributional Semantics

Harris goes one step farther and claims that *distributions* should be taken as an explanans *for meaning* itself:
$\rightarrow$ only this can turn semantics into a proper part of the *linguistic science*.

Vector Space Models use linguistic corpora and statistical techniques to turn these ideas into empirically testable semantic models.

Currently DS is corpus-based, however DS $\neq$ corpus linguistics: the DH is not by definition restricted to linguistic context

- but current corpus-based methods are more advanced than available methods to process extra-linguistic context.
- corpus-based methods allow us to investigate how *linguistic* context shapes meaning.

General Definition of DSMs

A distributional semantic model (DSM) is a co-occurrence matrix $\mathbf{M}$ where rows correspond to *target terms* and columns correspond to *context* or *situations* where the target terms appear.

	see	use	hear	...
boat	39	23	4	...
cat	58	4	4	...
dog	83	10	42	...

- Distributional vector of 'dog': $x_{dog} = (83, 10, 42, \ldots)$
- Each value in the vector is a feature or dimension.
- The values in a matrix are derived from event frequencies.

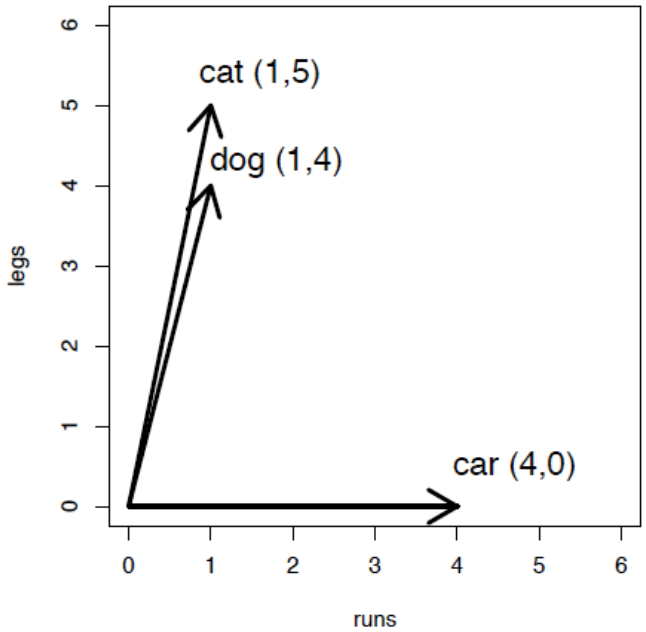
A DSM allows us to measure semantic similarity between words.

Vectors and Similarity

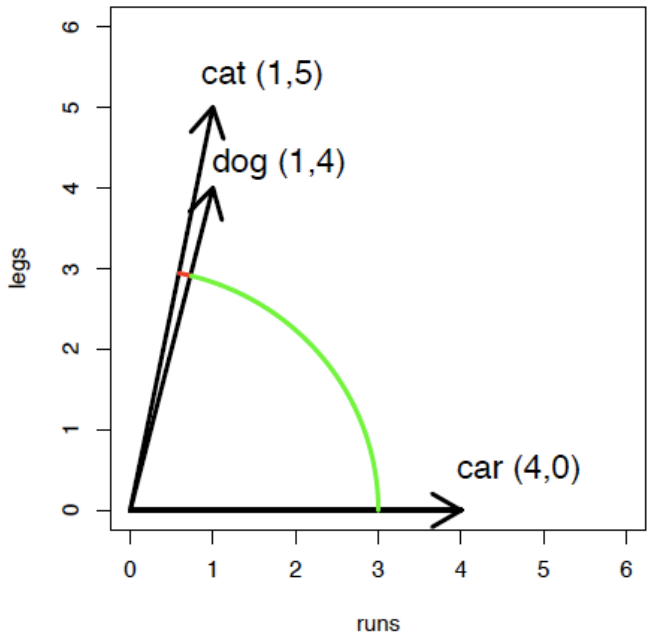
Vectors can be displayed in a vector space. This is easier to visualise if we look at two dimensions only, e.g. at two dimensional spaces.

	run	legs
dog	1	4
cat	1	5
car	4	0

semantic space



semantic similarity as angle between vectors



Generating a DSM

Assuming we have a corpus, creating a DSM involves these steps:

- Step 1: Define target terms (rows) and contexts (columns)
- Step 2: Linguistic processing: pre-process the corpus used as data
- Step 3: Mathematical processing: build up the matrix

We need to evaluate the resulting semantic representations.

Step 1: Rows and Columns

Decide what the target terms (rows) and the contexts or situations where the target terms occur (columns) are. Some examples:

- Word-based matrix: typically restricted to content words; the matrix may be symmetric (same words in rows and columns) or non-symmetric.
- Syntax-based matrix: the part of speech of the words or the syntactic relation that holds between them may be taken into account.
- Pattern-based matrix: rows may be pairs of words (*mason:stone*, *carpenter:wood*) and columns may correspond to patterns where the pairs occur (*X cuts Y*, *X works with Y*).

Step 2: Linguistic Processing

- The minimum processing required is tokenisation
- Beyond this, depending on what our target terms/contexts are, we may have to apply:
  - stemming
  - lemmatisation
  - POS tagging
  - parsing
  - semantic role labeling
  - ...

Step 3: Mathematical Processing

1. Building a matrix of frequencies
2. Weighting or scaling the features
3. Smoothing the matrix: dimensionality reduction

Step 3.1: Building the Frequency Matrix

Building the frequency matrix essentially involves counting the frequency of *events* (e.g. *how often does "dog" occur in the context of "see"?*)

In order to do the counting, we need to decide on the size or type of context where to look for occurrences. For instance:

- within a window of k words around the target
- within a particular linguistic unit:
  - a sentence
  - a paragraph
  - a turn in a conversation
  - ...

The mean distance of the Sun from the Earth is approximately 149.6 million kilometers, though the distance varies as the Earth moves from perihelion in January to aphelion in July. At this average distance, light travels from the Sun to Earth in about 8 minutes and 19 seconds. The Sun does not have a definite boundary as rocky planets do, and in its outer parts the density of its gases drops exponentially with increasing distance from its center.

Step 3.2: Feature Weighting/Scaling

Once a matrix has been created, typically the features (i.e. the frequency counts in the cells) are scaled and/or weighted.

Scaling: used to compress wide range of frequency counts to a more manageable size

- *logarithmic scaling*: we substitute each value x in the matrix for $\log(x+1)$ [*we add* $+1$ *to avoid zeros and negative counts*].

$\log_y(x)$: how many times we have to multiply y with itself to get x
$\log_{10}(10000) = 4 \quad \log_{10}(10000+1) = 4.0004$

- arguably this is consistent with the Weber-Fechner law about human perception of differences between stimulus

Step 3.2: Feature Weighting/Scaling

Weighting: used to give more weight to surprising events than to expected events $\rightarrow$ the less frequent the target and the context, the higher the weight given to the observed co-occurrence count (because their expected chance co-occurrence is low)

- a classic measure is mutual information

observed co-occurrence frequency (f_{obs})

	small	domesticated
dog	855	29

$f_{\text{dog}} = 33338$
$f_{\text{small}} = 490580$
$f_{\text{domest.}} = 918$
$N =$ total # or words in corpus

$*$ expected co-occurrence frequency between word$_1$ and word$_2$: $f_{exp} = \frac{f_{w1} \cdot f_{w2}}{N}$

$*$ mutual information compares observed vs. expected frequency:

$$\text{MI}(w1, w2) = \log_2 \frac{f_{obs}}{f_{exp}}$$

There are many other types of weighting measures (see references).

Step 3.3: Dimensionality Reduction

The co-occurrence frequency matrix is often unmanageably large and can be extremely sparse (many cells with 0 counts)
$\rightarrow$ we can compress the matrix by reducing its dimensionality, i.e. reducing the number of columns.

- Feature selection: we typically want to keep those columns that have high frequency and high variance.
  - we may eliminate correlated dimensions because they are uninformative.
- Projection into a subspace: several sophisticated mathematical techniques from linear algebra can be used, e.g.:
  - principal component analysis
  - singular value decomposition
  - ...

  [*we will not cover the details of these techniques; see references*]

Comparing Vectors

Once our DSM has been generated, we have at our disposal a matrix where word meanings are modelled as vectors: points in a highly mutidimensional space.

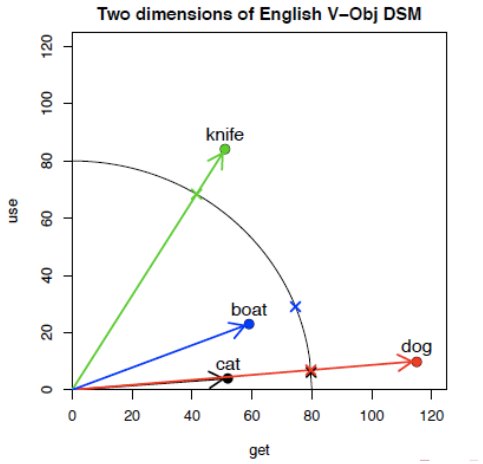
The most obvious thing we can do with them is to quantify how similar two meanings are by measuring the distance between them in vector space.

Similarity/Distance Measures

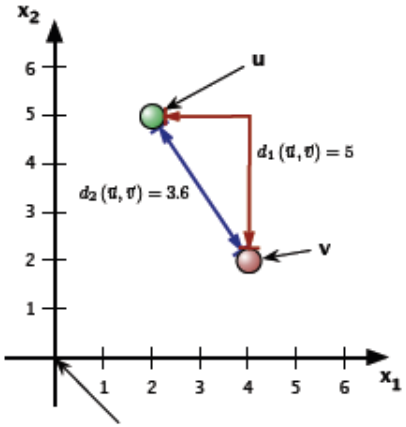
- Cosine measure of similarity: angle between two vectors

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}}{||\mathbf{x}||} \cdot \frac{\mathbf{y}}{||\mathbf{y}||}$$

vectors need to be normalised to unit length (dividing the vector by its length) - what matters is the angle



- Other popular distance measures include:



  - Euclidean distance
  - "City block" Manhattan distance

Several other types of similarity measures have been proposed (see refs.)

What's in a DSM?

What aspects of meaning are encoded in DSMs? What is *semantic similarity*? Semantic neighbours in DSMs have different types of semantic relations with the target.

The web interface of Infomap allows you to query several DSMs. Given a target word and a few model parameters, the interface returns the top semantic neighbours of t in m.the target.

http://clic.cimec.unitn.it/infomap-query/

The documentation page gives you details of the parameters used by each model. You can experiment with a few target words and different models.

Evaluating DSMs

The most common way of evaluating a DSMs consists in testing how well it captures semantic similarity, broadly understood.

Some classic evaluation methods:

- Synonym identification
- Modeling semantic similarity judgments
- Semantic priming

Synonym Identification: the TOEFL task

The TOEFL dataset: 80 target items with candidate synonyms.

Target: *levied* Candidates: *imposed*, *believed*, *requested*, *correlated*

DSMs and TOEFL:

1. take vectors of the target (**t**) and of the candidates ($\mathbf{c}_1 \ldots \mathbf{c}_n$)
2. measure the distance between **t** and $\mathbf{c}_i$, with $1 \geq i \geq n$
3. select $\mathbf{c}_i$ with the shortest distance in space from **t**

- Humans
  - Average foreign test taker: 64.5%
  - Macquarie University staff (Rapp 2003): non-natives 86.75%; natives: **97.75%**
- DSMs
  - Latent Semantic Analysis (Landauer & Dumais 1997): 64.4%
  - Padó and Lapata's (2007) dependency-based model: 73%
  - Rapp's (2003) model trained on lemmatized BNC: **92.5%**

R. Rapp (2003) Discovering the meanings of an ambiguous word by searching for sense descriptors with complementary context patterns, in *Proceedings of TIA 2003*.

Semantic Similarity Judgements

Can DSMs model human semantic similarity judgements?

- Dataset: Rubenstein and Goodenough (1965) (R&G) 65 noun pairs rated by 51 subjects on a 0-4 similarity scale

car	automobile	3.9
food	fruit	2.7
cord	smile	0.0

- DSMs and R&G:
  1. for each test pair (w_1,w_2), take vectors $\mathbf{w}_1$ and $\mathbf{w}_2$
  2. measure the distance (e.g. cosine) between $\mathbf{w}_1$ and $\mathbf{w}_2$
  3. measure (with Pearson's r) the correlation between vector distances and R&G average judgments

  Padó and Lapata (2007) show there are strong correlations between the distances in their dependency-based DSM and the human judgements ($r = 0.8$).

S. Padó & M. Lapata, Dependency-Based Construction of Semantic Space Models, *Computational Linguistics*, 33(2):161-199.

Semantic Priming

Hearing/reading some words facilitates access to other words in various lexical tasks (naming, lexical decision, reading): the word *pear* is recognized/accessed faster if it is heard/read after *apple*.

- Psychologists have found similar amounts of priming for different semantic relations between words in a single word lexical decision task (deciding whether a stimulus is a word or not).
  - synonyms: to dread/to fear
  - antonyms: short/tall
  - coordinates (co-hyponyms): train/truck
  - super- and subordinate pairs (hypernyms): container/bottle
  - free association pairs: dove/peace
  - phrasal associates: vacant/building

Semantic Priming

How can we evaluate DSMs against data from semantic priming?

1. for each related prime-target pair, measure cosine-based similarity between pair items (e.g. to dread/to fear)
2. to estimate unrelated primes, take average of cosine-based similarity of target with other primes from same relation data-set (e.g. to value/to fear)
3. similarity between related items should be significantly higher than average similarity between unrelated items

- McDonald & Brew (2004), Padó & Lapata (2007) found significant effects ($p < .01$) for all semantic relations.
- The stronger effects were found for synonyms, antonyms, and coordinates.

S. McDonald; C. Brew (2004) A Distributional Model of Semantic Context Effects in Lexical Processing, in *Proceedings of ACL 2004*.

Summing Up

- Usage-based view of word meaning: the context where words occur tell us a good deal about what they mean (*determine their meaning*).
- DSMs/VSMs make the distributional hypothesis precise, giving quantitative predictions that can be tested.
- They are typically evaluated against human perceptions (judgements, priming) of semantic similarity.

Tomorrow

- More on pros and cons of DSMs
- A bit on WSD methods with vector spaces
- An example of a research paper:

Katja Abramova, Raquel Fernández, and Federico Sangati (2013) Automatic Labeling of Phonesthemic Senses. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pp. 1696-1701. Berlin, Germany.

  * have a look at this paper by tomorrow
  * have a look at the homework by tomorrow