Computational Semantics and Pragmatics

Autumn 2012

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

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*]

The underlying assumption is that word meaning depends, at least in part, on the contexts in which words are used:

- 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

Origins of Distributional Semantics

- Currently, distributional semantics is especially 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.
- Harris goes one step farther and claims that *distributions* should be taken as an explanans *for meaning* itself.
  - only this can turn semantics into a proper part of the *linguistic science*.

Beyond Structuralism

Some traditions that developed after Structuralism are critical of DS:

- Generative linguistics: focus on I-language — internalised competence of ideal speakers — and dismissal of language use.
- Formal semantics: model-theoretic and referential tradition, focus on denotational semantics; meaning is anchored in the world, not language-internal.
- Cognitive psychology: some proponents of a conceptual view of meaning find DSMs too "external"

In contrast, other traditions embrace DS:

- Corpus linguistics and lexicography: distributional semantics is the main methodological principle for semantic analysis.
- Cognitive Psychology: *Contextual Hypothesis* by Miller and Charles (1991) distributions as a way to explain cognitive semantic representations and how they are built by learners.

Essence of Distributional Semantics

Again, the main general assumption behind DSMs is that *word meaning depends on the contexts in which words are used*.

There are three main aspects that characterise distributional semantic representations and make them very different from representations in lexical and formal semantics. They are:

- inherently context-based and hence context-dependent
  - the linguistic contexts in which words are observed enter into their semantic constitution;
- inherently dynamic
  - meaning derives from the way a word interacts with different contexts (dimensions) - from its global distributional history, which is constantly evolving;
- inherently quantitative and gradual
  - meaning is represented in terms of statistical distribution in various linguistic contexts.

Other important aspects linked to DSMs

- Use of linguistic corpora: 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.
- Use of statistical techniques: Statistical and mathematical techniques are key tools for DS:
  - used to create an abstract contextual representation over usages;
  - formal and empirically testable semantics models.

DSMs make use of mathematical and computational techniques to turn the informal DH into empirically testable semantic models.

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.

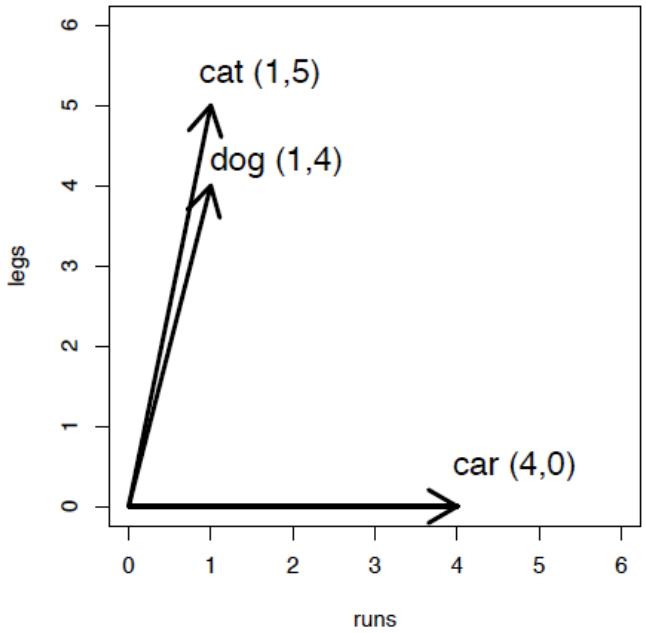
A DSM allows us to measure semantic similarity between words.

Vectors and Similarity

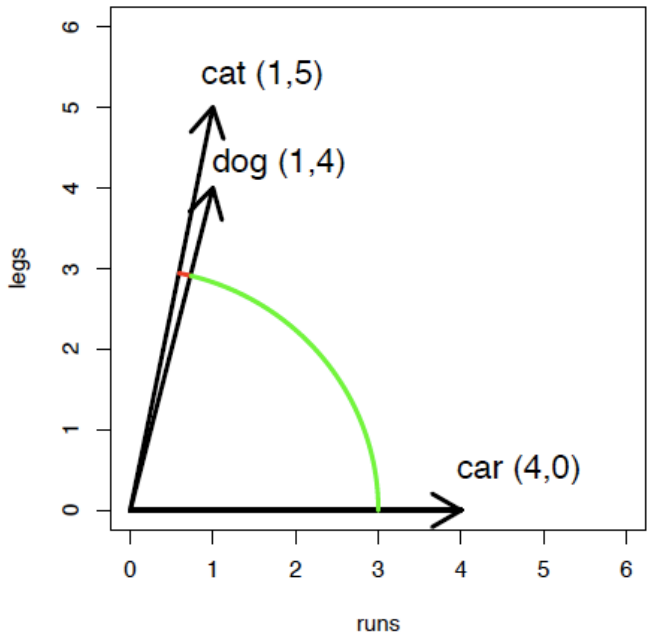
Vectors can be displayed in a vector space. This is easier to visualise if we look at two dimensions only, e.g. at two dimensional spaces.

	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

- Building a matrix of frequencies
- Weighting or scaling the features
- Smoothing the matrix: dimensionality reduction
- Measuring similarity / distance between vectors

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 couple of examples of weighting measures:

- idf: the *inverse document frequency* of a lemma l is calculated as follows, where N is the total number of documents in the corpus and df_l (document frequency) is the number of documents in the corpus that contain term l.

$$\text{idf}_l = \log \frac{N}{\text{df}_l}.$$

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)

- another classic measure is mutual information

observed co-occurrence frequency (f_{obs})

	small	domesticated
dog	855	29

$f_{\text{dog}} = 33.338$
$f_{\text{small}} = 490.580$
$f_{\text{domest.}} = 918$
$N =$ 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)
$\longrightarrow$ 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
  - ...

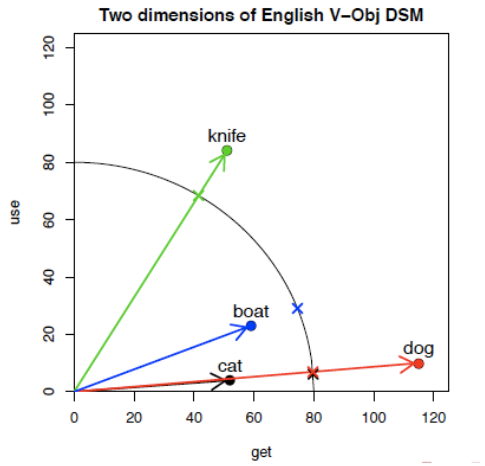
  [*we will not cover the details of these techniques; see references*]

Step 3.4: Similarity/Distance Measures

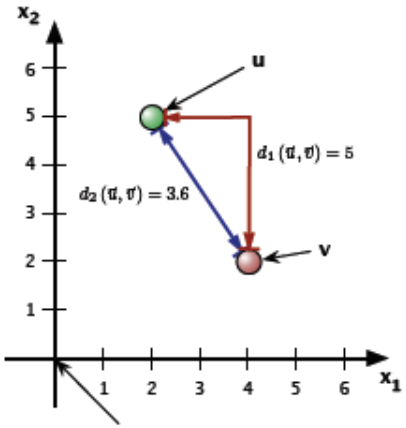
- cosine measure of similarity: angle between two vectors

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

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.)

Interpreting DSMs

What aspects of meaning are encoded in DSMs? 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 a few model parameters, the interface returns the top semantic neighbours of t in m.the target.

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

Read the documentation page to find out which parameters are being used by each model and experiment with a few target words.