Computational Semantics and Pragmatics

Autumn 2011

Raquel Fernández

Institute for Logic, Language & Computation
University of Amsterdam

Plan for Today

- We'll continue to look into the properties of DSMs, including how they can be evaluated.
  - based on material from Stefan Evert, A. Lenci, and Marco Baroni
  - for further details, see
    - `http://www.wordspace.collocations.de/`
    - P. Turney and P. Pantel (2010) From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research*, 37:141-188.
- Discussion of the philosophical implications of DSMs based on:
  - A. Lenci (2008) Distributional Semantics in Linguistic and Cognitive Research, in Lenci (ed.), *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science*, special issue of the *Italian Journal of Linguistics*, 20(1):1-30.

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.

Generating a DSM

- 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 the matrix and compare the resulting vectors

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
  - a Webpage

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)

- recall idf (inverse document frequency)
- 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
  - ...

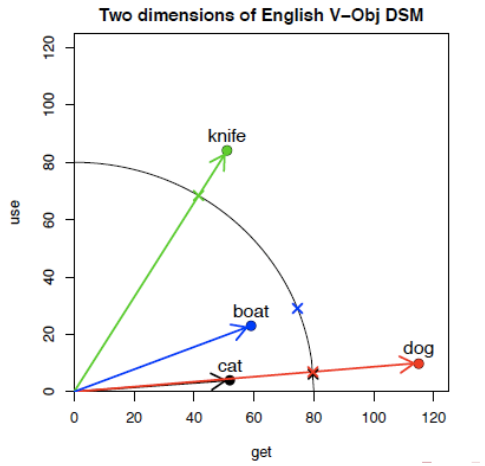
  [*we will not cover the details of these techniques; see references*]

Step 3.4: Similarity/Distance Measures

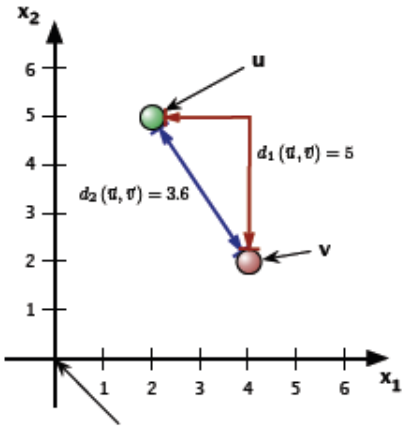
- cosine measure of similarity: angle between two vectors

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

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? Neighbors in DSMs have different types of semantic relations with the target

car (InfomapNLP on BNC; n = 2)

- van co-hyponym
- vehicle hyperonym
- truck co-hyponym
- motorcycle co-hyponym
- driver related entity
- motor part
- lorry co-hyponym
- motorist related entity
- cavalier hyponym
- bike co-hyponym

car (InfomapNLP on BNC; n = 30)

- drive function
- park typical action
- bonnet part
- windscreen part
- hatchback part
- headlight part
- jaguar hyponym
- garage location
- cavalier hyponym
- tyre part

Web Infomap [http://clic.cimec.unitn.it/infomap-query/]

Interpreting DSMs

We can distinguish between two broad types of semantic relations:

- Attributional similarity: two words sharing a high number of salient features (attributes)
  - synonymy (car/automobile)
  - hypernymy (car/vehicle)
  - co-hyponymy (car/van/truck)
- Semantic relatedness: two words semantically associated without being necessarily similar
  - function (car/drive)
  - meronymy (car/tyre)
  - location (car/road)
  - attribute (car/fast)

Evaluation of Attributional Similarity

Most DSMs encode attributional similarity. How can we evaluate them? Some possibilities include:

- 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 2004): 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. 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*.

Semantic Relatedness & Relational Similarity

Attributional similarity can be modeled with word-based or syntax-based DSMs that have single words in rows/columns:

	die	kill	gun
teacher	109.4	0.0	0.0
victim	1335.2	22.4	0.0
soldier	4547.5	1306.9	105.9
policeman	68.6	38.2	30.5

To distinguish between different types of semantic relations, we can use pattern-based matrices:

- rows: word pairs
- columns: syntagmatic links between the word pairs

	in	at	with	use
teacher : school	11894.4	7020.1	28.9	0.0
teacher : handbook	2.5	0.0	3.2	10.1
soldier : gun	2.8	10.3	105.9	41.0

Semantic Relatedness & Relational Similarity

	in	at	with	use
teacher : school	11894.4	7020.1	28.9	0.0
teacher : handbook	2.5	0.0	3.2	10.1
soldier : gun	2.8	10.3	105.9	41.0

- look at direct co-occurrences of word pairs (when we talk about a concept, we are likely to also mention its parts, function, etc.)
- use the contexts of pairs to measure pair similarity (relational similarity)
- group pairs into coherent relation types by their contexts (semantic relatedness)
- pairs that occur in similar contexts (i.e. connected by similar words and structures) will tend to be related, with the shared contexts acting as a cue to the nature of their relation

Relational similarity can be evaluated with a relational equivalent of the TEOFL task (Turney 2006).

P. Turney (2006) Similarity of Semantic Relations, *Computational Linguistics*, 33(3):379-416.

Philosophical Implications

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.

In contrast, other traditions embrace DS:

- Corpus linguistics and lexicography: distributional semantics is the main methodological principle for semantic analysis.
  $\rightsquigarrow$ *recall the paper by Kilgarriff we discussed.*
- 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 (I)

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 distributed and 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.

Essence of Distributional Semantics (II)

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.

Essence of Distributional Semantics (III)

Where does DS stand within the nativist vs. empiricist debate?

- Nativism: part of our *language faculty* is innate. The human brain comes equipped with a limited set of choices; when learning, children select the correct options using their parents' speech, in combination with the context.
- Empiricism: emphasis on *learning* from usage. There isn't an innate language structure, but general and perhaps language-specific learning capabilities part of our cognitive apparatus.
- Lenci points out that DS is indeed empiricist, but not inherently anti-nativist:
  - some DSMs extract meaning features from raw data;
  - others may include higher level information such as syntax.

Status of Contextual Representations

The core assumption behind DSMs — once more: *word meaning depends on the contexts in which words are used* — can be interpreted in different ways.

- **Key issue**: is the hypothesized dependency between contexts (word distributions) and semantics (word meaning) simply a correlation or is there a causal relation between them?

Answers to this question give rise to two versions of the Distributional Hypothesis, which differ on the status they assign to contextual representations:

⇒ *"Weak" DH vs. "Strong" DH*

The Weak Distributional Hypothesis

- Only assumes a correlation between semantic content and contextual distributions:
  - by examining distributions and exploiting their correlation with semantics, we get at a better understanding of lexical meaning;
  - word meaning (whatever this might is) determines the distribution of words in context;
  - we can try to uncover semantic content by inspecting a significant number of distributions.
- This weak version of the hypothesis is compatible with different research programmes:
  - find paradigmatic classes of e.g. verbs
  - empirical foundations for lexical semantic theories, such as the Generative Lexicon.

The Strong Distributional Hypothesis

- Assumes that distributions have a causal role in the creation of semantic representations at a cognitive level:
  - the distributional behaviour of a word is a way to *explain* its semantic content;
  - the environments where a word appears have an effect on its cognitive semantic representation.
- Evidence for this strong version of the hypothesis comes from the possibility of modelling psychological phenomena with distributional representations, such as:
  - human similarity judgements
  - semantic priming

Unsatisfactory Aspects of DSMs

The strong version of the DH is committed to the cognitive plausibility of DSMs. However, some core aspects of semantics are not satisfactorily treated by these models:

- semantic relations and lexical entailment
- compositionality
- reference, symbol grounding and embodiment

This has raised criticisms: skeptics point out that whatever distributions can tell us about a word, this cannot be its meaning.

Key issue: Do these weak points depend on features of current models, or are they inherent to the essence of the DH?

Semantics Relations & Lexical Entailment

Knowing the meaning of a word involves recognising the validity of inferences that hold between sentences that include that word:

(1) a. Google bought a new company $\rightarrow$ Google purchased a new company
b. Ann drives a car $\rightarrow$ Ann drives a vehicle

(2) a. Google purchased a new company $\rightarrow$ Google bought a new company
b. Ann drives a vehicle $\not\rightarrow$ Ann drives a car
c. Ann drives a car $\not\rightarrow$ Ann drives a van
Ann drives a van $\not\rightarrow$ Ann drives a car

- (1a)-(2a) synonymy (1b)-(2b) hyponymy (2c) co-hyponymy

DSMs can recognise the attributional similarity between these words, but can't distinguish between different (asymmetric) types of relations.

Compositionality

- DS is concerned with lexical meaning – compositionality is typically not the focus (as with most lexical semantic theories).
- However, arguably any semantic theory should be able to explain how the meaning of a complex expression can be built up from the meanings of its components.
- Can DSMs provide a satisfactory account of compositionality?
- What is needed is a way to compose distributional information. But this is not straightforward:
  - Landauer & Dumais (2007) propose *vector summation*: the distributional meaning of *'the dog bit the man'* is the sum of the vectors of each of the words in the sentence.
  - But this can't distinguish it from *'The man bit the dog'* !
- Recent approaches adopt more sophisticated models of vector composition that include syntactic dependencies.
- Compositionality is an open issue in current DSM research.

Reference and Embodiment

DSM are often refused as plausible cognitive models of meaning for two reasons:

- They are regarded as ungrounded symbolic representations. As such they fall under the "symbol grounding problem" (Harnad 1990) and the "Chinese Room argument" (Searle 1980)
  - Chinese Room: thought experiment by John Searle against the possibility of strong AI (Turing test).
  - To the extent that DSMs claim that meaning can be derived by pure symbol manipulation, without direct reference to the world, they are subject to this problem.

Reference and Embodiment

The second reason why DSMs are refused as cognitive models is tied to the Embodied Cognition Hypothesis.

- According to the ECH, conceptual representations are grounded in the sensory-motor systems .
  - concepts/meanings are not amodal, formal symbols but perceptual symbols represented within the perceptual systems we acquire them
  - embodied simulations proposed by Barsalou: knowing the meaning of the word *'turtle'* implies being able to re-enact our perceptual experiences with turtles.
  - some findings in neuroscience back up these claims.
- To the extent that distributional contextual representations are not embodied, they are not cognitively plausible according to this line of research. Linguistic distributions are seen as a *product* of embodied conceptualisation.

Reference, Embodiment – and Symbolic Context

Is it right to assume that DS cannot play any substantial role in a cognitive explanation of meaning? According to A. Lenci:

- It doesn't seem to make sense to completely reduce meaning and concepts to representations grounded in sensory modalities;
- as it probably doesn't make sense to reduce everything to symbol manipulation.

$\Rightarrow$ Both aspects play a role in the processes leading to meaning formation.

- There is a growing trend of proposing dual models that combine embodied and distributional information.

- These issues are not limited to DSM research: what processes create meaning remains a big open question in philosophy and cognitive science.

What's Next

Word sense disambiguation (WSD): the task of determining which sense of a word is being used in a particular context.

- supervised vs. unsupervised methods

$\Rightarrow$ Recall that HW#2 is due on Monday 17 October 2011