Distributional Compositionality
Intro to Distributional Semantics

Raffaella Bernardi

University of Trento

February 14, 2012

Acknowledgments

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

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 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

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

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: Vector and Matrix

Operations on vectors

Vector addition:

$$\vec{v} + \vec{w} = (v_1 + w_1, \ldots v_n + w_n)$$

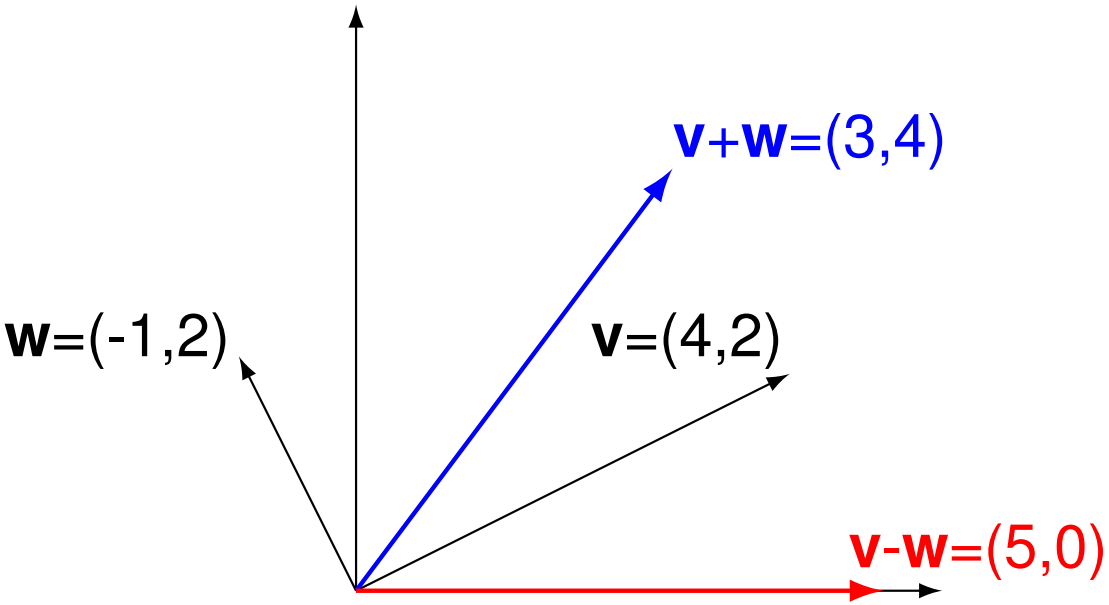
similarly for the $-$.

Scalar multiplication: $c\vec{v} = (cv_1, \ldots cv_n)$ where c is a "scalar".

Background: Vector and Matrix

Vector visualization

Vectors are visualized by arrows. They correspond to points (the point where the arrow ends.)



vector addition produces the diagonal of a parallelogram.

Background: Vector and Matrix

Dot product or inner product

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

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 and Matrix

Length and Unit vector

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

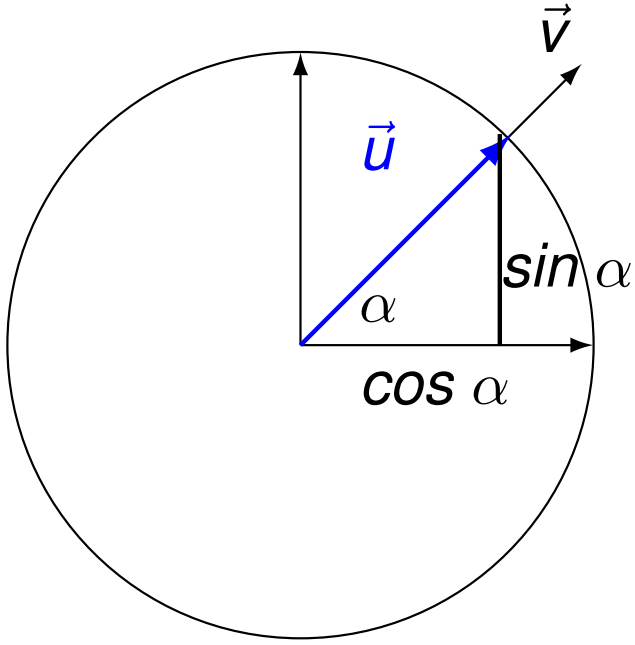
Unit vector is a vector whose length equals one.

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

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

Background: Vector and Matrix

Unit vector



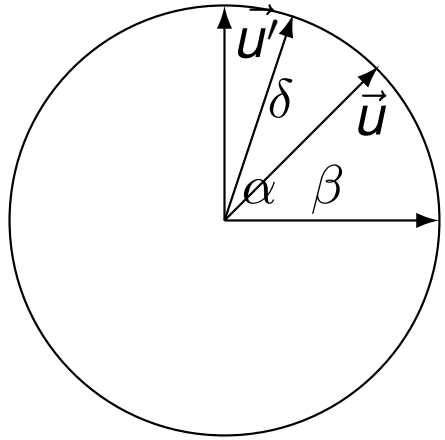
$$\vec{u} = \frac{\vec{v}}{||\vec{v}||} = (\cos \alpha, \sin \alpha)$$

Background: Vector and Matrix

Cosine formula

Given δ the angle formed by the two unit vectors $\vec{u}$ and $\vec{u'}$, s.t. $\vec{u} = (cos\ \beta, sin\ \beta)$ 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* δ; *cos* δ is never bigger than 1 (since we used unit vectors) and never less than -1. It's 0 when the angle is 90^o

Background: Vector and Matrix

Matrices multiplication

A matrix is represented by [nr-rows x nr-columns].
Eg. for a 2 x 3 matrix, the notation is:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

a_{ij} i stands for the row nr, and j stands for the column nr.

The multiplication of two matrices is obtained by

Rows of the 1st matrix x columns of the 2nd.

A matrix with m-columns can be multiplied only by a matrix of m-rows:

[n x m] x [m x k] = [n x k].

Background: Vector and Matrix

A matrix acts on a vector

Example of 2 x 2 matrix multiplied by a 2 x 1 matrix (viz. a vector). Take A and $\vec{x}$ to be as below.

$$A\vec{x} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} (1,0)\cdot(x_1,x_2) \\ (-1,1)\cdot(x_1,x_2) \end{bmatrix} = \begin{bmatrix} 1(x_1)+0(x_2) \\ -1(x_1)+1(x_2) \end{bmatrix} =$$

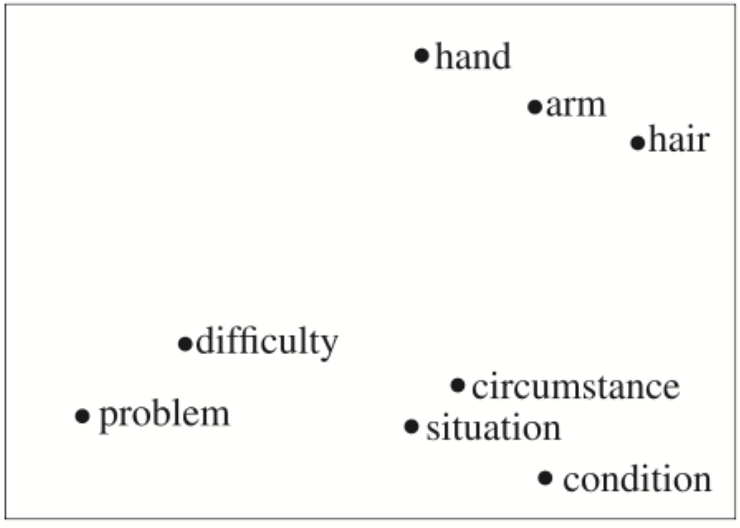
$$= \begin{bmatrix} x_1 \\ x_2 - x_1 \end{bmatrix} = \vec{b}$$

A is a "difference matrix": the output vector $\vec{b}$ contains differences of the input vector $\vec{x}$ on which "the matrix has acted."

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.

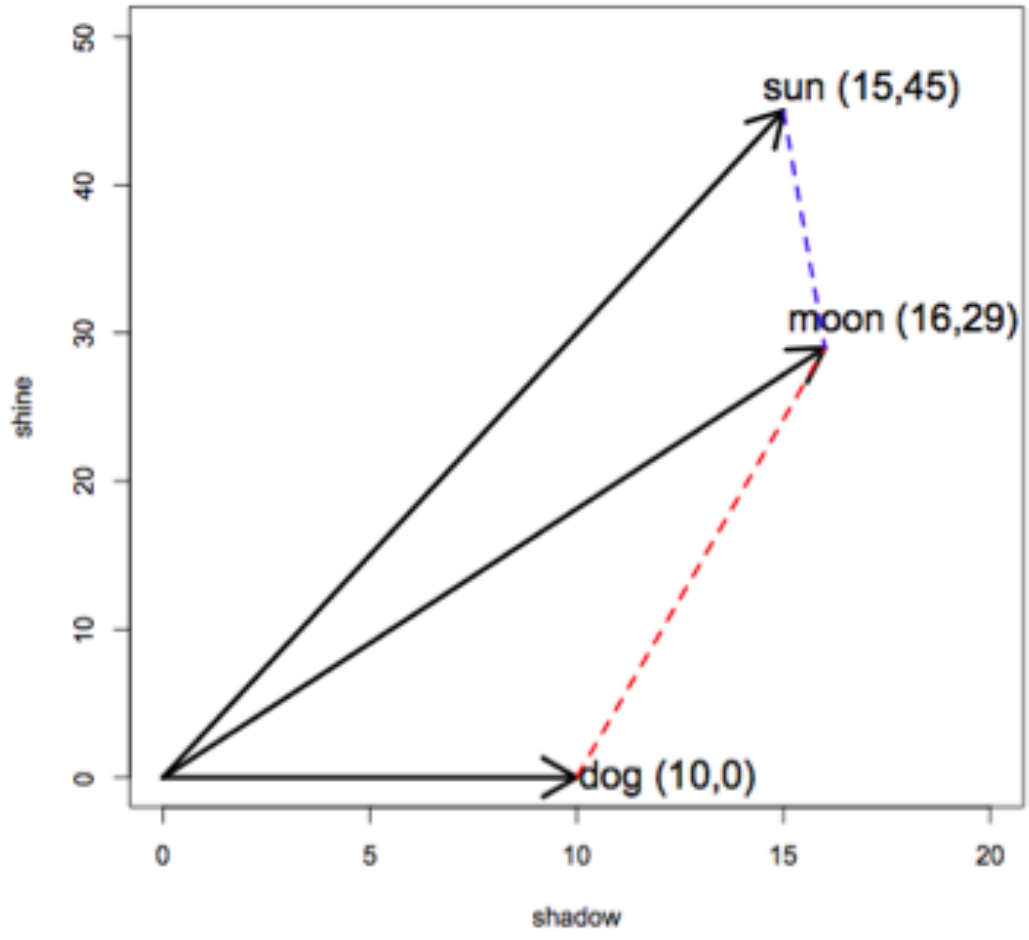


Distributional Semantics Model

Toy example: vectors in a 2 dimensional space

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

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



Distributional Semantics Model

Two dimensional space representation

$\overrightarrow{moon}$=(16,29), $\overrightarrow{sun}$= (15,45), $\overrightarrow{dog}$=(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
$\overrightarrow{moon}$	16	29
$\overrightarrow{sun}$	15	45
$\overrightarrow{dog}$	10	0

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

Distributional Semantics Model

One space and many dimensions

- **One Space** Usually, words are taken to be all in the same space.
- **Many space dimensions** Usually, the space dimensions are the *most k frequent words*, minus the “stop-words”, viz. high-frequency words with relatively low information content, such us grammatical words (e.g. of, the, and, them, . . .). Hence, they may be around 2k-30K or even more.
- **What in the dimensions** They can be plain words, words with their PoS, words with their syntactic relation, or even documents. Hence, a text needs to be: tokenized, normalized (e.g., capitalization and stemming), annotated with PoS tags (N, J, etc.), and if required also parsed.

Distributional Semantics Model

Lexical association function

Instead of plain counts, the values can be more significant weights of the co-occurrence frequency:

- **tf-idf** (term frequency (tf) $\times$ inverse document frequency (idf)): an element gets a high weight when the corresponding term is frequent in the corresponding document (tf is high), but the term is rare in other documents of the corpus (df is low, idf is high.) [Spärk Jones, 1972]
- **PMI** (pointwise mutual information): measure how often two events x and y occur, compared with what we would expect if they were independent [Church & Hankes, 1989]

Distributional Semantics Model

Dimensionality reduction

Reduce the dimension-by-word matrix to a lower dimensionality matrix (a matrix with less – linearly independent – dimensions). Two main reasons:

- Smoothing: capture “latent dimensions” that generalize over sparser surface dimensions (SVD)
- Efficiency/space: sometimes the matrix is so large that you don’t even want to construct it explicitly (Random Indexing)

Distributional Semantics Model

Dimensionality reduction: SVD

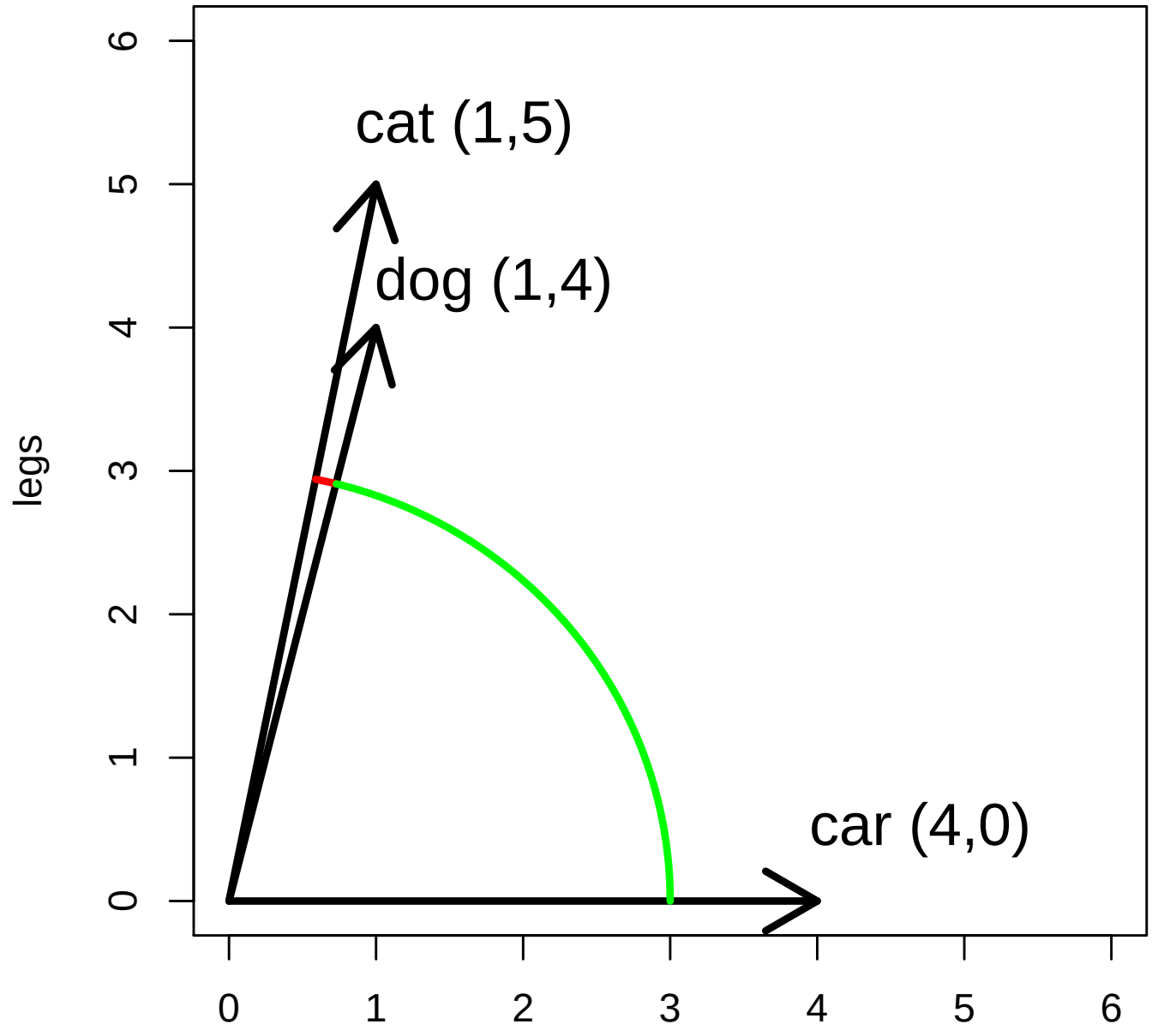
- General technique from Linear Algebra
- given a matrix of $n \times m$ dimensionality, construct a $k \times m$ matrix, where $k << n$ (and $k < m$)
  - e.g., from a 10K dimensions of 20K words matrix to a 300 "latent dimensions" x 20K words matrix
  - k is an arbitrary choice
- the new matrix preserves most of the variance in the original matrix

SVD: Pros and cons

- Pros:
  - Good performance (in most cases)
  - At least some indication of robustness against data sparseness
  - Smoothing as generalization
  - Smoothing also useful to generalize features to words that do not co-occur with them in the corpus
- Cons:
  - Non-incremental
  - Latent dimensions are difficult to interpret
  - Does not scale up well (but see recent developments. . .)

Distributional Semantics Models

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, $\frac{\vec{x}}{|x|}$ 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.

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

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.

Building a DSM

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.

Building a DSM

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.

Building a DSM

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.

Building a DSM

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.

Building a DSM

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.

Building a DSM

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.

Building a DSM

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.

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

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).
  - ...

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)

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

New research questions

- Do all words live in the same space?
- What about grammatical words?
- Can vectors representing phrases be extracted too?
- What about compositionality of word sense?
- How do we “infer” some piece of information out of another?

References

- 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
- S. Evert and A. Lenci. “Distributional Semantic Models” ESSLLI 2009. `http://wordspace.collocations.de/doku.php/course:esslli2009:start`
- K. Gimpel “Modeling Topics” 2006