Logic and Natural Language Semantics: Distributional Semantics

RAFFAELLA BERNARDI

DISI, UNIVERSITY OF TRENTO

E-MAIL: BERNARDI@DISI.UNITN.IT

Contents

1. Formal Semantics Applications

A software based on Categorial Grammar and λ calculus ideas is:

http://svn.ask.it.usyd.edu.au/trac/candc

it's and implementation of CCG:

http://groups.inf.ed.ac.uk/ccg/publications.html

It can parse huge documents.

Has been used for e.g. textual entailment (see lecture 4 (not done) on the web site.)

2. Back to philosophy of language

Frege:

1. Linguistic signs have a reference and a sense:

   (i) "Mark Twain is Mark Twain" vs. (ii) "Mark Twain is Samuel Clemens".

   (i) same sense and same reference vs. (ii) different sense and same reference.

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.

2.1. Back to Linguistics

But,

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** or open class: 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** or closed class: 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 corresponds to logical operators: negation, conjunction, disjunction, quantifiers.

2.2. 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.

2.3. Distributional Models: 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 row or parsed) corpora and can be represented by **vectors**.
3. **Cosine similarity** captures synonyms (as well as other semantic relations).

2.4. New questions within DS: “incomplete expressions”

More recent questions:

4. What about “incomplete” expressions (functions) (e.g. verbs, adjectives)?
5. How sense are put together to build the sense of phrases?
6. How do we “infer” some piece of information out of another?

Recent results:

4. a complete expression (e.g. noun) is represented by a vector vs. an “incomplete” expression is represented by a **matrix**.
5. Words are composed by applying a matrix to a vector (viz. **matrix product**).
6. New “similarity measures” have been defined to capture lexical entailment.

For an overview of DS see Turney & Pantel (2010).

2.5. Our Current work within DS: logical words

7. What about logical words?
8. Can their sense be induced from corpora?
9. How can they be represented?

3. Distributional Semantic: main idea

The sense of a word can be given by its use, viz. by the contexts in which it occurs;

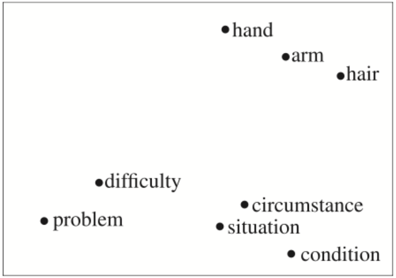
You can tell a word by the company it keeps (Firth, 1957)

```
he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
 surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
 in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
 But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
 they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
 the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind
```

3.1. DS model

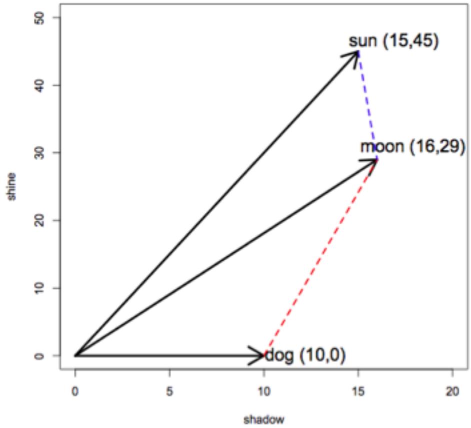
It's a quadruple $\langle B, A, S, V \rangle$, where:

- 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.
- S is a similarity measure.
- V is an optional transformation that reduces the dimensionality of the semantic space.



3.2. Toy example: vectors in a 2 dimensional space

$B = \{shadow, shine, \}$; A= frequency; S: angle measure (or Euclidean distance.)



	moon	sun	dog
shine	16	15	10
shadow	29	45	0

Smaller is the angle, more similar are the terms. (Cosine Similarity)

3.3. Space, dimensions, co-occurrence frequency

(Many) space dimensions Usually, the space dimensions are the **most k frequent words** (minus stop words.). They can be plain words, words with their PoS, words with their syntactic relation (e.g. our derivations)

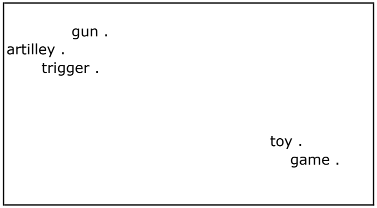
Co-occurrence frequency Instead of plain counts, the values can be more significant weights that take into account frequency and relevance of the words within the corpus. (e.g. tf-idf, mutual information, log-likelihood ratio etc.).

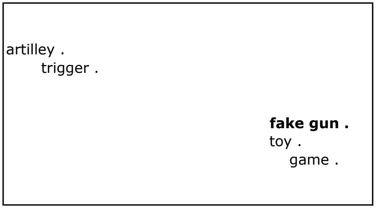
3.4. DM success on Lexical meaning

DM captures pretty well synonyms. DM used over TOEFL test:

- Foreigners average result: 64.5%
- Macquarie University Staff (Rapp 2004):
  - Ave. 5 not native speakers: 86.75%
  - Ave. 5 native speakers: 97.75%
- DM:
  - DM (dimension: words): 64.4%
  - Best system: 92.5%

3.5. Compositionality in DS

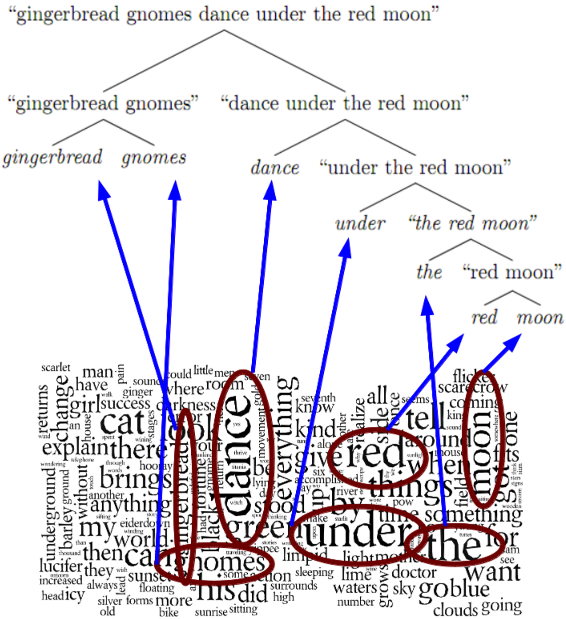




Focus on words, only recently on composition of words into phrases.

4. DS new research line: “incomplete expressions”

The sense of a sentence is built out of the senses of its words:



4.1. Formal Semantics and Distributional Semantics

	Formal	Distributional
expressions denote	in different domains of different types	in different space of different types
an expression of an atomic type is represented by	a constant	a vector
an expression of a function type is represented by	an n-argument function	a n-by-n-matrix
composition is performed by	function-application	by matrix product

Ideas imported into DM

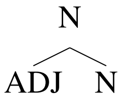
(a) Meaning flows from the words;

(b) "Complete" (argument) vs. Incomplete (function) words;

(c) meaning representations are guided by the syntactic structure.

4.2. Adjective noun composition in FS

Syntax:



FS: ADJ is a function (n/n) that modifies a noun (n): $[\![\text{Red}]\!] \cap [\![\text{Moon}]\!]$

$$(\lambda Y.\lambda x.\mathtt{Red}(x) \land Y(x))(\mathtt{Moon}) \leadsto \lambda x.\mathtt{Red}(x) \land \mathtt{Moon}(x)$$

4.3. Adj noun composition in DS

Distributional Semantics (e.g. 2 dimensional space):

N/N: matrix

red	d1	d2
d1	$n1$	$n2$
d2	$m1$	$m2$

N: vector

	moon
d1	$k1$
d2	$k2$

Function app. by the matrix product and returns a vector:

$\overrightarrow{red\ moon} = \sum_{i=1}^{n} red_i \times moon_i$

N: vector

	red moon
d1	$(n1 \times k1) + (n2 \times k2)$
d2	$(m1 \times k1) + (m2 \times k2)$

4.4. Vector vs. Matrix computation

The vector representation of a word living in a space of an atomic type can be induced by the corpus.[Well established]

The matrix representation of a word living in a space of a functional type can be learned via regression (Baroni & Zamparelli (2010)):

```
n and the moon shining i
 with the moon shining s
rainbowed moon . And the
 crescent moon , thrille
in a blue moon only , wi
now , the moon has risen
d now the moon rises , f
y at full moon , get up
 crescent moon . Mr Angu

f a large red moon , Campana
, a blood red moon hung over
 glorious red moon turning t
The round red moon , she 's
l a blood red moon emerged f
n rains , red moon blows , w
monstrous red moon had climb
 . A very red moon rising is
under the red moon a vampire
```

“red” is learned, using linear regression, from the pairs (N, red-N).

4.5. Linear regression

From the vectors input pairs, linear regression gives us the values of the “red” matrix

input pairs				
	moon	red moon	army	red army
d1	301	11	...	...
d2	93	90	...	...

Learned matrix		
red	d1	d2
d1	n1 ($301 \leadsto 11, \ldots$)	n2 ($301 \leadsto 90, \ldots$
d2	m1 ($93 \leadsto 11, \ldots$)	m2 ($11 \leadsto 90, \ldots$

Recall: function app. by the matrix product and returns a vector:

$$\overrightarrow{red\ moon} = \sum_{i=1}^{n} red_i \times moon_i$$

To double check the validity of the approach: the result of the matrix product has been compared to the vector induced from the corpus: positive results.

4.6. DS Composition: "function application"

Baroni & Zamparelli 2010, they have

- trained separate models for each adjective;
- (a) composed the learned matrix (function) with a noun vector (argument) by matrix product ($\otimes$) – the adjective weight matrix with the noun vector value;
- composed adjectives with nouns using: (b) additive and (c) multiplicative model –starting from adjective and noun vectors;
- harvested vectors for "adjective-noun" from the corpus;
- compared (a) "learned_matrix $\otimes$ vector_noun" ("function application") vs. (b) "vector_adj + vector_noun" vs. (c) "vector_adj $\times$ vector_noun";
- shown that – among (a), (b), (c) – (a) gives results more similar to the "harvested vector_adj-noun" than the other two methods.

For an overview on DS and compositionality see Mitchell & Lapata (2010).

4.7. Adjectives: observation

In Formal semantics, one meaning, e.g. "red" $\lambda Y.\lambda x.\mathtt{Red}(x) \wedge Y(x)$.

But different uses (Pustejovsky 1995):

- red Ferrari [the outside]
- red watermelon [the inside]
- red traffic light [only the signal]
- ..

Baroni & Zampareli approach predict these differences.

4.8. Adjectives in DS

Take an abstract noun ("activist") and a concrete noun ("moon") "red activist" is an abstract noun again, and "red moon" a concrete noun again. The matrix "red" composes with

- abstract noun: increases the values of abstract dimensions and leaves unchanged the values of the concrete dimensions;
- concrete noun: increases the value of the concrete dimensions and leaves unchanged the one of the abstract dimensions.

red	d1:shine	d2:soviet
d1: shine	n1	0
d2: soviet	0	m2

	moon	activist
d1: shine	k1	0
d2: soviet	0	p2

$\leadsto$

	red moon	red activist
d1: shine	(n1 x k1) + (0 x 0)	0 (=(n1 x 0) + (0 x p2))
d2: soviet	0 (=(0 x k1) + (m2 x 0))	(0 x 0) + (m2 x p2)

4.9. Back to Lambek calculus

$$\dfrac{\dfrac{X_1 : n \vdash X_1 : n \quad Y_1 : n \vdash Y_1 : Nn}{(X_2 : n/n \otimes X_1 : n) \vdash X_2\, X_1 : n}\,(/L) \qquad Y_2 : n \vdash Y_2 : n}{X_3 : n/n \otimes (X_2 : n/n \otimes X_1 : n) \vdash X_3\, (X_2\, X_1) : n}\,(/L)$$

Syntax: instantiate the categories with one of the word belonging to them e.g. “black young dog”

Semantics: the final meaning representation of the actual string is obtained by replacing the corresponding proof-term variables with the actual meaning representation.

Meaning: word meaning is represented by lambda terms (representing the set-theoretical interpretation), hence replace

X_3 with $\lambda X.\lambda y.\mathtt{black}(y) \wedge X(y)$, X_2 with $\lambda Y.\lambda x.\mathtt{young}(x) \wedge Y(x)$, X_1 with $\lambda z.\mathtt{dog}(z)$

$\rightsquigarrow \lambda x.\mathtt{black}(x) \wedge \mathtt{young}(x) \wedge \mathtt{dog}(x)$

Sense: word sense is represented by matrices and vectors, hence

replace X_3, X_2 with the matrices representing “black” and “young”, and X_1 with the vector representing “dog”, and yield $\overrightarrow{\textit{black young dog}}$

4.10. DS: Logical words

Logical words have been treated as stop-words, viz. simply ignored.

No results have been published so far on them from the DS view.

Ed Hoevy has spoken about them in some talks.

There are ongoing works on QP by Louise McNally and Gemma Boleda (Barcellona) and Marco Baroni, Roberto Zamparelli and me (Trento) with Chieng Chang Shan (from US) and EMLCT students (Q. T. Do and X. Gutiérrez)

Quantifiers

4.11. Summing up: FS and DS main interest

We can think of the following classes of words:

- Content words: nouns, adjectives, verbs [focus of DS]
- Grammatical words: preposition, articles, quantifiers, coordination, auxiliary verbs, pronouns and negation. [focus of FS]

DS research has obtained satisfactory results on content words by evaluating them on different lexical semantic tasks.

New research is "importing" in the DS framework some of the understanding achieved within the FS school.

FS starting point is logical entailment between a set of premises and a conclusion. Is DS good for this too?

5. Entailment in DS

- Lexical entailment: already some successful results.
- Phrase entailment: a few studies done.
- Sentential entailment: none.

5.1. DM success on Lexical entailment

Lexical entailment Cosine similarity has shown to be a valid measure for the synonymy relation, but it does not capture the "is-a" relation – e.g. it's symmetric!

Kotlerman, Dagan, Szpektor and Zhitomirsky-Geffet 2010 see is-a relation as "feature inclusion" (where "features" are the space dimensions) and propose an asymmetric measure. Intuition behind their measure:

1. Is-a score higher if included features are ranked high for the narrow term.
2. Is-a score higher if included features are ranked high in the broader term vector as well.
3. Is-a score is lower for short feature vectors.

Very positive results compared to WordNet-based measures.

They have focused on nouns.

5.2. FS Entailment

Entailment in FS Partially ordered domains

$$
\begin{array}{ll}
[\![\text{tall student}]\!] \leq_{(e\to t)} [\![\text{student}]\!] & \text{iff } \forall \alpha \in D_e \\
[\![\text{tall student}(\alpha)]\!] \leq_t [\![\text{student}(\alpha)]\!] & \text{iff} \\
[\![\text{tall student}]\!]([\![\alpha]\!]) \leq_t [\![\text{student}]\!]([\![\alpha]\!]) & \text{iff} \\
[\![\text{tall student}]\!]([\![\alpha]\!]) = 0 \text{ or } [\![\text{student}]\!]([\![\alpha]\!]) = 1. &
\end{array}
$$

Lesson: (a) different entailment relations for different domains; (b) same entailment relation for words and phrases belonging to the same category (e.g. "dog $\leq_{(e\to t)}$ animal" and also "small dog $\leq_{(e\to t)}$ animal")

Entailment in DS Dagan et al. measure (fine tuned on nouns) generalize to

- words of other categories?
- phrases?
- sentences?

5.3. Entailment at phrasal level in DS: Preliminary results

Work done with M. Baroni, C.C. Shan and T. N. Q. Do.

- Degan et. al. measure
  - does generalize to expressions of the noun category, tested on N1 $\leq$ N2 and ADJ N1 $\leq$ N1.
  - does not generalize to expressions of other categories, tested on QPs.
- Still DS models do contain information needed to detect the entailment relation among other categories too, tested on QP using Machine Learning methods.

Questions: which are the dimensions involved in the entailment relation for the various categories? Can we hope to find an abstract definition based on atomic and function types as in FS?

6. Back to FS & DS: what else?

In FS, 1. The meaning of a sentence is its truth value, 2. is built from **the meaning of its words**; 3. is represented by a FOL formula, hence we use logic entailment to handle inferences. Moreover,

- Composition is obtained by **function-application**.
- **Syntax guides** the building of the meaning representation. Lambek: function application (elimination) and abstraction (introduction rule).

Open questions in DS view What's the meaning of a sentence? What's the meaning of "entities", e.g., "John". Which is the DS representation corresponding to a higher order function, e.g. QP? What's the linear algebra operation corresponding to lambda abstraction – how can structure be de-composed in a DS representation (e.g. relative clauses)? Can DS representation capture "entailment"?

7. Who, what, where

Lambek Calculus and the alike

- Michael Moortgat (Utrecht)
- Chris Barker (NYU)
- Glyn Morrill (Barcelona)
- Philippe de Groote (Nancy)
- Christian Retoré (Bourdeaux)
- Mark Steedman (Edinburgh) – not logical approach (see before)
- ...

- Marco Baroni, Roberto Zamparelli and me (Trento) with Ken Shan (Cornell Uni.)
- Alessandro Lenci (Pisa)
- Louise McNally and Gemma Boleda (Barcellona)
- Mirella Lapata (Edinburgh)
- Stephen Clark, Ed Grefenstette, Mehrnoosh Sadrzadeh, Stephen Pulman, Oxford and Cambridge
- Katrin Erk (Texas)
- ...

8. Background: Matrix and vector

8.1. Linear equation

A linear equation is an algebraic equation in which each term is either a constant or the **product of a constant and a single** variable. E.g. a two variables x and y is $y = mx + b$, where m and b designate constants.

The origin of the name "**linear**" comes from the fact that the set of solutions of such an equation forms a **straight line** in the plane.

The general linear equation in n variables is: $a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$

In this form, $a_1, a_2, \ldots, a_n$ are the coefficients, $x_1, x_2, \ldots, x_n$ are the variables, and b is the constant.

Such an equation will represent an $(n-1)$-dimensional hyperplane in n-dimensional Euclidean space (or in our case n-dimensional vector space)

8.2. 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 $\vec{v} = (v_1, \dots v_n)$.

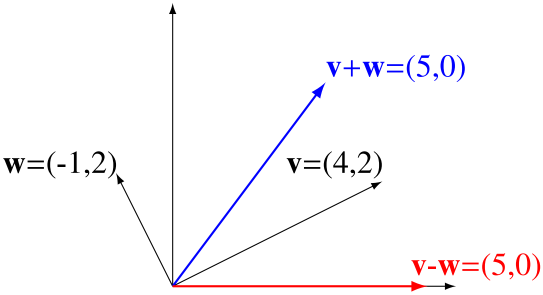
Operation on vectors Vector addition and vector difference: $\vec{v} + \vec{w} = (v_1 + w_1, \dots v_n + w_n)$, similarly for the $-$.

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

The addition of $c\vec{v}$ and $d\vec{w}$ is a linear combination of $\vec{v}$ and $\vec{w}$.

8.3. Vector visualization

Vectors are visualized by arrows.



vector addition produces the diagonal of a parallelogram.

vectors corresponds to points (the point where the arrow ends.)

8.4. Vector equation

Given two vectors $\vec{v} = (2, -1)$, $\vec{w} = (-1, 2)$ and the vector equation

$$c\vec{v} + d\vec{w} = (1, 0)$$

the solution is given by the two scalar equation:

$$2c - d = 1 \ \text{ and } \ -c + 2d = 0$$

$c = \frac{2}{3}$, $d = \frac{1}{3}$

8.5. Dot product or inner product

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

Example We've three goods to buy and sell, their prices are (p_1, p_2, p_3) (price vector $\vec{p}$). The quantity 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

$$\vec{q} \cdot \vec{p} = (q_1, q_2, q_3) \cdot (p_1, p_2, p_3) = q_1 p_1 + q_2 p_2 + q_3 p_3$$

8.6. Length and Unit vector

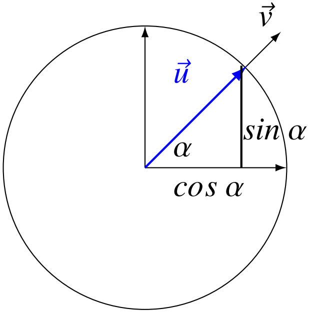
Length $||\vec{v}|| = \sqrt{\vec{v} \cdot \vec{v}} = \sqrt{\sum_{i=1}^{n} v_i^2}$

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 $\vec{v}$. (normilized vector)

8.7. Unit vector

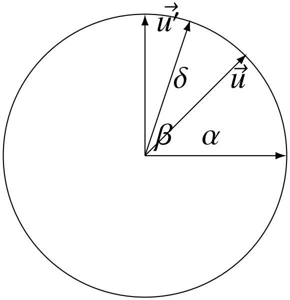


$\vec{v} = (1, 1)$, $\vec{u} = \frac{\vec{v}}{\|\vec{v}\|} = (cos\ \alpha, sin\ \alpha)$

8.8. 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 $\vec{u'} = (cos\,\alpha, sin\,\alpha)$

$$\vec{u} \cdot \vec{u'} = (cos\,\beta) \times (cos\,\alpha) + (sin\,\beta) \times (sin\,\alpha) = cos(\beta - \alpha) = cos\,\delta$$



$$cos\,\alpha = \frac{\vec{v}}{||\vec{v}||} \cdot \frac{\vec{w}}{||\vec{w}||}$$

The bigger the angle α, the smaller is $cos\,\alpha$; $cos\,\alpha$ is never bigger than 1 (since we used unit vectors) and never less than -1. It's 0 when the angle is 90^o

8.9. Matrices

Matrices encode **linear maps**:

9. Cosine Similarity

$$cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}$$

- x_i is the tf-idf weight of term i in x.
- y_i is the tf-idf weight of term i in y
- $|\vec{x}|$ and $|\vec{y}|$ are the lengths of $\vec{x}$ and $\vec{y}$.
- This is the cosine similarity of $\vec{x}$ and $\vec{y}$ or, equivalently, the cosine of the angle between $\vec{x}$ and $\vec{y}$.

In short, the similarity between two vectors is computed by the cosine of the angle between them.

10. QP: Ideas from FS and Cognitive analysis

- Formal Semantics, due to their mathematical character:
  - QPs differ w.r.t. the relation they establish holding between the NP and the VP they quantify over.
  - QPs effect reasoning.
  - Different QPs co-occur with different Polarity Items.
  - Not all QPs can be negated.
  - Not all QPs can be coordinated and not by the same coordination.
- Cognitive Science:
  - QPs have different scalar strength.
  - QPs have different pragmatic effects.
  - QPs differ w.r.t. the relation they establish between (the VP of) consecutive sentences (and the anaphoric pronouns in it).

10.1. Conjecture on QP in DS

- Content words effect the topic of the sentences they occur in.
  - Vectors representing nouns have been extracted from corpora considering grammatical words as stop-words.
  - Adjectives modify nouns. Matrix of adjectives have been learned from noun vectors (i.e. ignoring grammatical words).
- QPs effect the reasoning that can be drawn from the sentences they occur in. In the matrix of a QP important role could be played by:
  - Grammatical words.
  - The frequency relation among the NP and the V related by the QP.
  - content-words polarity.