27 Giugno 2006

Semantica Lessicale

Applicazioni

Marco Pennacchiotti

pennacchiotti@info.uniroma2.it

Tel. 06 7259 7717

Ing.dell'Informazione, *stanza P1B-03* (nuova ala Ing.Inf, primo piano)

Programma

- **Breve introduzione all'NLP**
  - Linguaggi Naturali e Linguaggi Formali
  - Complessità
- **Morfologia**
  - *Teoria:* Morfologia del Linguaggio Naturale
  - *Strumenti:* Automi e Trasduttori
  - *Analisi Morfologica:* con automi e trasduttori
- **Part of Speech Tagging**
  - *Teoria:* Le classi morfologiche
  - *Strumenti a Analisi:* modelli a regole e statistici
- **Sintassi**
  - *Teoria:* Sintassi del Linguaggio Naturale
  - *Strumenti:* CFG
  - *Analisi Sintattica:* parsing top-down, bottom-up, Early
- **Semantica**
  - Distributional Lexical Semantics
  - Sentence Semantics

Info

Sommario

- Semantica

- **Introduzione**
- Semantica lessicale distribuzionale
  - Concetti di base
  - La *Distributional Hypothesys*
  - Misure distribuzionali
- Applicazioni di semantica lessicale distribuzionale
  - Costruzione di un Thesaurus
  - Paraphrasing

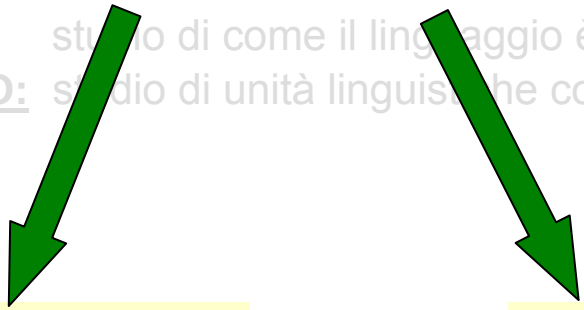
Le lezioni di queste slides sono in parte adattate dal corso di *Lexical Semantics* di **Patrick Pantel** (ISI-USC):

Lexical Semantics e applicazioni

- **FONETICA:** studio dei suoni linguistici
- **MORFOLOGIA:** studio delle componenti significative di una parola
- **SINTASSI:** studio delle strutture relazionali tra le parole

- **SEMANTICA:** **studio del significato delle parole e di come esse si combinano per formare il significato delle frasi**

- **PRAGMATICA:** studio di come il linguaggio è usato per raggiungere obiettivi
- **ANALISI DEL DISCORSO:** studio di unità linguistiche complesse



LEXICAL SEMANTICS

Studio del significato delle parole

SENTENCE SEMANTICS

Studio del significato di intere frasi

-Studio delle relazioni lessicali (sinonimia, iperonimia,meronimia, antinomia, entailment, causa,…)

-Il significato di una parola è contenuto nella parola stessa?

Lexical Semantics

Lexical Semantics e Classical Sentence Semantics

- *Lexical Semantics*
  - studio del significato delle *parole* e delle relazioni tra di esse
  - studio di semplici *espressioni linguistiche* e delle loro relazioni
  - non necessita di formalismi sofisticati: poco formalizzata
  - basata su conoscenza ontologica o statistica

- *Classical Sentence Semantics*
  - studio del significato di intere *frasi*
  - generalmente fortemente formalizzata in formalismi logici e grammatiche
  - storicamente legata ad un'interpretazione deterministica del significato della lingua

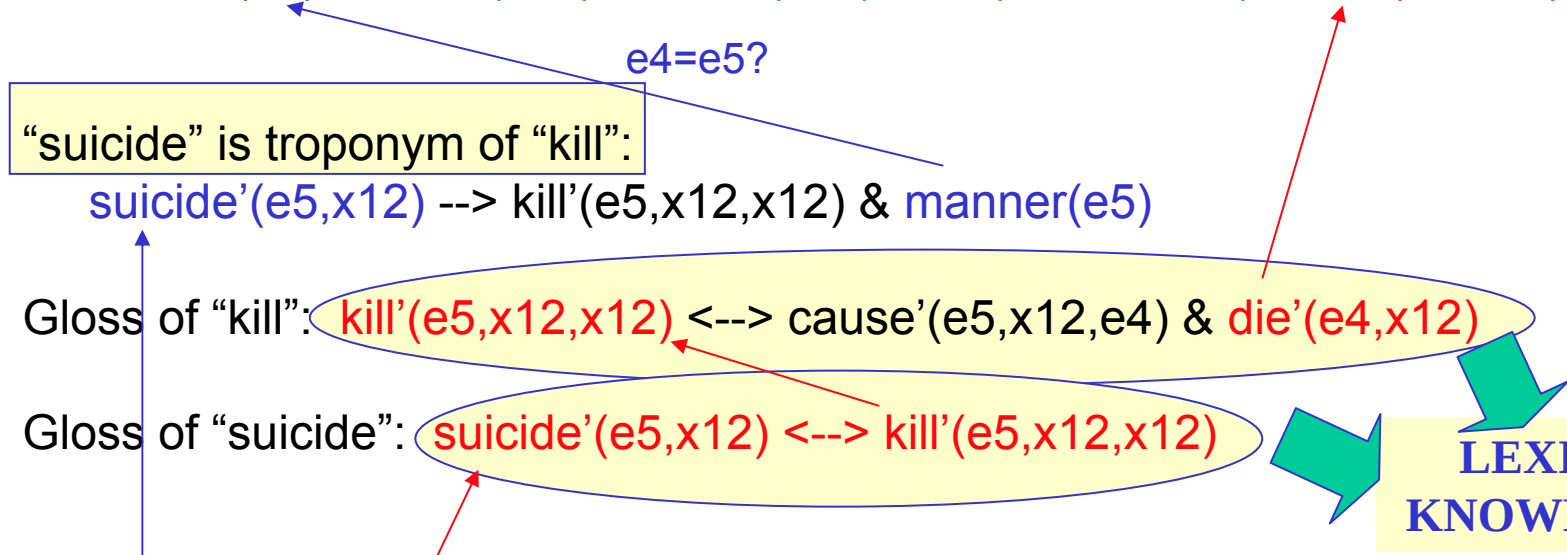
UTILIZZABILI INSIEME?

Lexical Semantics

Lexical Semantics + Classical Semantic per QA : *Hobbs Logical form [Hobbs,2005]*

Q: "How did Adolf Hitler die?"

QLF: manner(e4) & Adolf(x10) & Hitler(x11) & nn(x12,x10,11) & die'(e4,x12)



ALF: it(x14) & be'(e1,x14,x2) & Zhukov(x1) & 's(x2,x1) & soldier(x2)
& plant'(e2,x2,x3) & Soviet(x3) & flag(x3) & atop(e2,x4) & Reichstag(x4)
& on(e2,x8) & May(x5) & 1(x6) & 1945(x7) & nn(x8,x5,x6,x7) & day(x9)
& Adolf(x10) & Hitler(x11) & nn(x12,x10,x11) & commit'(e3,x12,e5)
& suicide'(e5,x12)

A: "It was Zhukov's soldiers who planted a Soviet flag atop the Reichstag on May 1, 1945, a day after Adolf Hitler committed suicide."

Lexical Semantics e applicazioni

LEXICAL SEMANTICS E APPLICAZIONI

Come può la *semantica lessicale* aiutare nelle applicazioni di NLP ?

relazioni tra parole o termini

- *relazioni generiche :* similarità / correlazione
- *relazioni specifiche :* iperonimia, meronimia, etc.
  - Applicazioni tipiche;
    - Costruzione di Thesausus
    - Question Answering, Information Extraction

relazioni tra espressioni linguistiche complesse

- paraphrasing ("*X* wrote *Y*" ≡ "*X* is the author of *Y*")
- textual entailment ("*X* kill *Y*" → "*Y* die")
  - Applicazioni tipiche:
    - Question Answering
    - Text Summarization
    - information Extraction

Lexical Semantics e applicazioni

- Che metodologie utilizzare ?
  - **metodologie distribuzionali** (basate unicamente su *corpora)*
    - approcci statistici non supervisionati (*knowledge harvesting*)
    - fortemente basate su studi statistico-distribuzionali delle parole
    - uso di nessun o semplici strumenti di NLP (es, shallow parsing)
    - adattabili *no-cost* a differenti lingue
    - non garantiscono una analisi semantica approfondita (*relazioni semplici*)
  - **metodologie basate su *conoscenza***
    - approcci con analisi di strutture ontologiche o reti semantiche (es,*WordNet*)
    - uso di misure di distanza all'interno della rete
    - non portabili a differenti lingue se non esiste una rete per essa
    - garantiscono un'analisi semantica approfondita e precisa tanto quanto la rete è semanticamente espressiva (*relazioni complesse*)

Similarità VS Correlazione

Che tipo di relazioni possono esistere tra due parole ?

- *Semplici:* correlazione, similarità
- *Complesse :* is-a, part-of, causa, … …

RELAZIONI SEMPLICI

Correlazione (*C*)

Due parole w1 e w2 si dicono *semanticamente correlate* se sono legate da una qualsiasi relazione semantica

Esempio

- *delfino-mare* `vive_in(delfino,mare)`
- *uomo-testa* `part_of(testa,uomo)`

Similarità (*S*)

Due parole si dicono *semanticamente simili* se sono vicine in una in una gerarchia IS-A

Esempio

- *gatto-cane* `is_a(cane,anim_dom) , is_a(gatto,anim_dom)`
- *gatto-mammifero* `is_a(gatto,mammifero)`

Similarità VS Correlazione

- La correlazione *include* la similarità

- Il grado di correlazione/similarità tra due parole (dette *target words*) viene misurato da appropriate misure:
  - misura di corellazione: $C(w_1,w_2)$
  - misura di similarità: $S(w_1,w_2)$
  - Valori più alti indicano un grado maggiore di correlazione/similarità

- In letteratura sono state proposte differenti misure di correlazione e di similarità, riconducibili a due classi:
  - *misure distribuzionali*: basate sulla *Distributional Hypothesis*
  - *misure non-distribuzionali*: basate sull'uso di ontologie e reti semantiche

- Distanza sematica: misure inverse di correlazione e similarità:
  - $D_C(w_1,w_2)= 1 / C(w_1,w_2)$
  - $D_S(w_1,w_2)= 1 / S(w_1,w_2)$

Sommario

- Semantica

- Introduzione
- **Semantica lessicale distribuzionale**
  - Concetti di base
  - La *Distributional Hypothesys*
  - Misure distribuzionali
- Applicazioni di semantica lessicale distribuzionale
  - Costruzione di un Thesaurus
  - Paraphrasing

Distributional Hypothesis

DOMANDA… *Il significato di una parola è contenuto nella parola stessa, oppure nelle parole con cui occorre ?*

Differenti filosofi, semiotici e linguistici darebbero ognuno una risposta opposta all'altro… ma per noi *"ingegneri"* ?

ESEMPIO :

DUGONGO

- soluzione 1 : guardo in un dizionario!
  ma se il dizionario non c'è, o non contiene la parola?

- soluzione 2 : proviamo qualche acrobazia morfologica:
  du – gongo
  una band formata da due gonghisti? …poco probabile

Distributional Hypothesis

DUGONGO

- soluzione 3 : vado su Internet e guardo il contesto in cui si trova la parola:

  -*“Le informazioni raccolte in queste pagine derivano dall'osservazione diretta di due esemplari di Dugongo che ho avuto la fortuna di incontrare in **Mar** Rosso”*

  -*“Bella la spiaggetta con il dugongo e bella l'**escursione** con i delfini.”*

  -*“se sarete fortunati vedrete anche il Dugongo,vero tormentone della nostra compagnia, che si può osservare in una **escursione** che costa circa 15 euro”*

  -*il dugongo vive quasi esclusivamente in **mare**.*

- Quali altre parole occorrono con *“mare”, “escursione”, “esemplare”, “spiaggia”*…?
  - ***Foca***
  - ***Traghetto***
  - ***Leone marino***
  - ***Focena***
- Quindi forse il *dugongo* è una sorta di mammifero marino …

Distributional Hypothesis

DUGONGO

“Mammifero marino erbivoro dei Sireni, con largo muso a setole intorno alla bocca (Dugong dugong) ”

Lexical Semantics e applicazioni

I sistemi di NLP sono pieni di dugonghi !!

- i lessici semantici (es. *WordNet*) non possono modellare tutte le parole:
  - neologismi vengono create tutti i giorni
  - alcuni domini hanno parole molto specifiche
  - alcune volte utilizziamo espressioni inesistenti
  - non esistono lessici semantici per tutte le lingue

- servono quindi approcci che:
  - modellino il significato delle parole non utilizzando dizionari codificati
  - siano *universali* : basati su principi indipendenti dalla lingua,che siano portabili e *cross-language*
  - possano modellare “imprevisti” (neologismi ed espressioni sconosciute)

Sommario

- Semantica

- Introduzione
- **Semantica lessicale distribuzionale**
  - **Concetti di base**
  - La *Distributional Hypothesys*
  - Misure distribuzionali
- Applicazioni di semantica lessicale distribuzionale
  - Costruzione di un Thesaurus
  - Paraphrasing

Concetti di base

CO-OCCORRENZA

- Le parole che si trovano in una certa finestra di una *target word* *t* sono dette co-occorrenze
  - la finestra può comprendere un dato numero di parole vicine, una frase, un paragrafo, un documento

- L'insieme delle co-occorrenze di *t* è detto contesto *C(t)*
  - nozioni più complesse di contesto possono comprendere co-occorrenze che sono in una certa **relazione sintattica** con la target word (es. verbo della target word, ecc…) oppure solo parole appartenti ad un certa Part of Speach (es. Nome, verbo…)

- **ESEMPIO:**

Finestra di 4 parole → C(dugongo) = {*fortunati, vedrete, anche, il, vero, tormentone, della, nostra*}
Relazione V-Sog → C(dugongo) = {*vedrete*}

*se sarete fortunati vedrete anche il **Dugongo**, vero tormentone della nostra compagnia.*

w_{-4} w_{-3} w_{-2} w_{-1} **t** w_{+1} w_{+2} w_{+3} w_{+4}

Concetti di base

MISURE DI ASSOCIAZIONE TRA PAROLE

- Associazione Binaria (B)
  - Se due parole *x* e *y* co-occorrono almeno una volta in un corpus *D*, allora il loro grado di associazione è 1, altrimenti è 0
  - Problema: non distingue tra parole che co-occorrono spesso e parole che co-occorrono raramente

$$B(x,y) = \begin{cases} 0 & \text{se non co}-\text{occorrono} \\ 1 & \text{se co}-\text{occorrono} \end{cases}$$

Concetti di base

MISURE DI ASSOCIAZIONE TRA PAROLE

- Frequenza (F)
  - Più volte due parole *x* e *y* co-occorrono più il loro grado di associazione è alto
  - Prende in considerazione il comportamento delle parole nel testo
  - Problema: parole molto frequenti hanno grado di associazione più alto rispetto a parole poco frequenti !!

$$F(x,y) = \frac{c_{xy}}{N}$$

c_{xy}= numero di occorrenze della co-occorrenza *xy* in un corpus *D*
N = numero di occorrenze totale di tutte le parole di un corpus *D*

Concetti di base

MISURE DI ASSOCIAZIONE TRA PAROLE

- Pointwise Mutual Information (I) (PMI)
  - Due parole *x* e *y* che co-occorrono spesso rispetto alle loro occorrenze in un corpus *D*, hanno un alto grado di associazione
  - Vantaggio rispetto a F: Due parole che co-occorrono spesso ma che sono molto frequenti hanno associazione minore rispetto a parole che co-occorrono lo stesso numero di volte ma che sono meno frequenti
  - Definita originariamente in *Information Theory* [Fano,1961] come verifica della *null hypothesis of independence*

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

P(x)= probabilità della'evento *x*
P(y)= probabilità dell'evento *y*
P(x,y)= probabilità congiunta degli eventi *x* e *y*

Concetti di base

MISURE DI ASSOCIAZIONE TRA PAROLE

- Pointwise Mutual Information (I) (PMI)
  - La definizione di *I* viene adattata all'NLP da [Church and Hanks, 1989], considerando:

    P(x) = probabilità della parola *x* nel linguaggio

    P(y) = probabilità della parola *y* nel linguaggio

    P(x,y) = probabilità che *x* co-occorra con *y*

  - e stimando le probabilità utilizzando MLE (*Maximum Likelihood Estimation*):

$$I(x,y) \approx \log_2 \frac{\frac{c_{xy}}{N}}{\frac{c_x}{N} \times \frac{c_y}{N}}$$

c_i= numero di occorrenze di *i* in un corpus *D*

c_{ij}= numero di occorrenze della co-occorrenza *ij* in un corpus *D*

N = numero di occorrenze totale di tutte le parole di un corpus *D*

Concetti di base

MISURE DI ASSOCIAZIONE TRA PAROLE

- Pointwise Mutual Information (I) (PMI)
  - Può essere intesa come verifica della *null hypothesis di* indipendenza:

$$P(x,y) = P(x) \times P(y)$$

  - $I(x,y) >> 0$ *associazione alta*
    - *x* e *y* compaiono in *D* quasi sempre insieme
  - $I(x,y) = 0$ *x e y sono indipendenti, non associati*
    - $P(x,y) = P(x) \times P(y)$ (*null hypothesis verificata*)
    - *x* e *y* non co-occorrono mai in *D* se non per puro caso
  - $I(x,y) < 0$ *associazione negativa*
    - *x* e *y* co-occorrono meno di quanto dovrebbero in base al caso

Concetti di base

MISURE DI ASSOCIAZIONE TRA PAROLE

- Pointwise Mutual Information (I) (PMI)
  - Problema: *I* predice un grado di associazione troppo alto per eventi con *basse frequenze*
    - In caso di perfetta dipendenza vengono favorite le coppie con bassa frequenza *I(x,y)=log 1/P(x)*
    - E'una buona misura di indipendenza, ma non buona misura di dipendenza
    - Per ovviare al problema si utilizza un fattore correttivo [Pantel Lin, 2002]:

$$I_{corrected}(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \times \frac{\min(freq(x),freq(y))}{\min(freq(x),freq(y))+1}$$

Concetti di base

MISURE DI ASSOCIAZIONE TRA PAROLE

- Probabilità condizionata (PC)
  - Indica quanto una parola *y* tende ad associarsi con una parola *x*

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

  - stimando le probabilità con MLE:

$$P(x \mid y) \approx \frac{\frac{c_{xy}}{N}}{\frac{c_y}{N}}$$

  - Non è stato ancora accertato se in NLP è più efficace l'approccio informativo di *I* o quello puramente probabilistico di *PC*

Sommario

- Semantica

- Introduzione
- **Semantica lessicale distribuzionale**
  - Concetti di base
  - **La *Distributional Hypothesys***
  - Misure distribuzionali
- Applicazioni di semantica lessicale distribuzionale
  - Costruzione di un Thesaurus
  - Paraphrasing

Distributional Hypothesis

DISTRIBUTIONAL HYPOTHESIS

Parole che occorrono nello stesso contesto tendono ad avere un simile significato (Harris,1968)

La definizione è molto potente, ma per questo anche molto generica:

- Cosa si intende per "*simile significato*"?
  - parole che hanno qualche relazione tra loro? (*correlazione*)
  - parole sinonimi o *quasi-sinonimi*? (*similarità*)
- Cosa si intende per *"contesto"* ?
  - un documento? Un paragrafo? Una frase?
  - una particolare struttura sintattica ?
- Perché limitarsi a "*parole*", invece di espressioni linguistiche più complesse?

Distributional Hypothesis

CORRELAZIONE DISTRIBUZIONALE

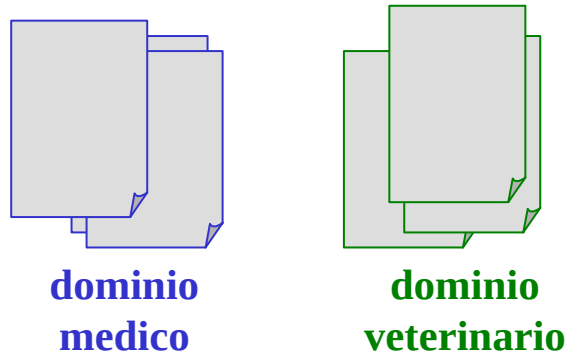
Due parole w_1 e w_2 si dicono *distribuzionalmente correlate* se hanno molte co-occorrenze comuni, e queste co-occorrenze non hanno nessuna restrizione sintattica sulla loro relazione con w_1 e w_2.

Due parole w_1 e w_2 distribuzionalmente correlate sono *semanticamente correlate.*

- Parole dello stesso *dominio* sono distribuzionalmente correlate, in quanto occorrono negli stessi contesti (stessi documenti, pagine web, ecc.)
- Parole relazionate che *non* fanno parte dello stesso dominio non sono distribuzionalmente correlate

ESEMPIO:

- correlate: *dottore, ospedale, malattia, medicina, cura, sintomo*
- non correlate: *dottore*, *veterinario*



Distributional Hypothesis

SIMILARITA' DISTRIBUZIONALE

Due parole w_1 e w_2 si dicono *distribuzionalmente simili* se hanno molte co-occorrenze comuni, e queste co-occorrenze sono relazionate a w_1 e w_2 dalla stessa relazione sintattica.

Due parole w_1 e w_2 distribuzionalmente simili sono *semanticamente simili.*

- Parole dello stesso *dominio e* con le stesse *proprietà* sintattiche, sono distribuzionalmente simili:
  - generalmente stessa Part Of Speech
  - stesse relazioni sintattiche

- ESEMPIO:

simili: *dottore, infermiere*

co-occorrenze comuni:

"…X lavora in ospedale…" (lavora , V-Sog, X)

"…X cura paziente…" (cura, V-Sog, X)

"…la prognosi di X…" (prognosi, NP-PP, X)

correlate e non-simili: *dottore, guarire*

co-occorrenze comuni (*paziente,ospedale*):

"il paziente guarisce in ospedale"

"il paziente del dottore è nell'ospedale "

Sommario

- Semantica

- Introduzione
- **Semantica lessicale distribuzionale**
  - Concetti di base
  - La *Distributional Hypothesys*
  - **Misure distribuzionali**
- Applicazioni di semantica lessicale distribuzionale
  - Costruzione di un Thesaurus
  - Paraphrasing

Misure Distribuzionali

- Le misure distribuzionali misurano il grado di correlazione o di similarità tra due parole sfruttando il principio della Distributional Hypothesis

- L'inverso di una misura distribuzionale è detta *distanza distribuzionale*

- Tipologie principali:

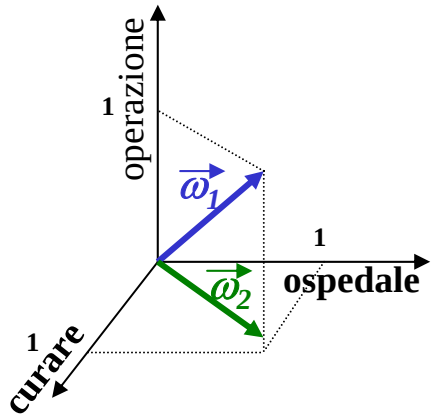
misure su spazi	*coseno, Manhattan distance, distanza euclidea*
misure insiemistiche	*Jaccard, Dice factor*
misure basate su MI	*Hindle, Lin*
misure di Information Retrieval	*CRM*

Misure su spazi

- La similarità/correlazione viene modellata su uno *spazio multidimensionale* in cui:
  - le dimensioni rappresentano tutte parole del lessico ***L***
  - una target word w_1 è rappresentata da un punto dello spazio
  - il vettore $\vec{\omega}_1$ dall'origine al punto è detto *feature vector* di w_1 ed ha come componenti i valori di associazione tra w_1 e le parole $w \in L$. Tali valori sono calcolati da una *misura di associazione* ***A(w1,w)*** (frequenza,mutua informazione, ecc…)

- La similarità/correlazione tra due target word w_1 e w_2 viene misurata come **distanza** tra i due feature vectors $\vec{\omega}_1$ e $\vec{\omega}_2$,

- **ESEMPIO**:



L ={operazione, curare,ospedale}

w_1 = dottore

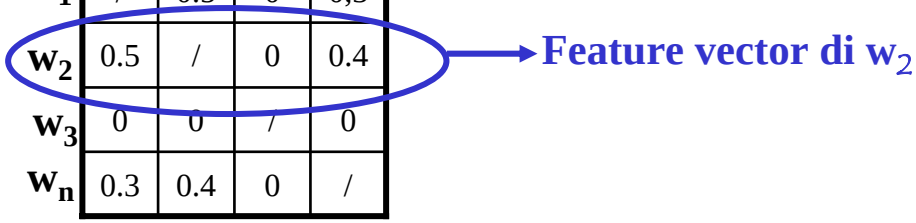
w_2 = infermiere

$$B(w_1, w) = \begin{cases} 0 & \text{se } w_1 \text{ e } w \text{ non co-occorrono} \\ 1 & \text{se } w_1 \text{ e } w \text{ co-occorrono} \end{cases}$$

Rappresentazione a features

- Dato un lessico *L* estratto da un corpus *D*, è possibile rappresentare i valori di co-occorrenza tra tutte le parole di *L*, tramite una matrice delle co-occorrenze:
  - la matrice ha come elementi i *valori di associazione* tra ogni parola del corpus
  - i valori di associazione sono tutti calcolati utilizzando la stessa funzione (binaria,frequenza,mutua informazione o probabilità condizionata…)
  - ogni riga indica il *feature vector* di una parola
  - ogni *feature* del vettore esprime l'associazione tra una target word e una parola del lessico

	w_1	w_2	w_3	w_n
w_1	/	0.5	0	0,3
w_2	0.5	/	0	0.4
w_3	0	0	/	0
w_n	0.3	0.4	0	/



- La matrice è generalmente sparsa:
  - Tecniche di LSI (Latent Semantic Indexing) possono essere utilizzate per ridurre a dimensione della matrice

Misure su spazi

COSENO

- La similarità/correlazione tra due parole w_1 e w_2 è calcolata come coseno tra i feature vectors
- Valori tra 0 e 1, più alti se la similarità/correlazione è maggiore

$$Cos(w_1, w_2) = \frac{\vec{w}_1 . \vec{w}_2}{|\vec{w}_1| \times |\vec{w}_2|}$$

- Versione probabilistica

$$Cos(w_1, w_2) = \frac{\sum_{w \in C(w_1) \cup C(w_2)} (P(w|w_1) \times P(w|w_2))}{\sqrt{\sum_{w \in C(w_1)} P(w|w_1)^2} \times \sqrt{\sum_{w \in C(w_2)} P(w|w_2)^2}}$$

- Versione PMI

$$Cos^{MI}(w_1, w_2) = \frac{\sum_{w \in C(w_1) \cup C(w_2)} (I(w, w_1) \times I(w, w_2))}{\sqrt{\sum_{w \in C(w_1)} I(w, w_1)^2} \times \sqrt{\sum_{w \in C(w_2)} I(w, w_2)^2}}$$

Misure su spazi

MANHATTAN DISTANCE

- La similarità/correlazione tra due parole w_1 e w_2 è calcolata in base alle differenze tra i valori di associazione con le co-occorrenze
- Valori tra 0 e ∞, più alti se la similarità/correlazione è *minore*

- Versione probabilistica

$$L_1(w_1, w_2) = \sum_{w \in C(w_1) \cup C(w_2)} | P(w|w_1) - P(w|w_2) |$$

- Versione PMI

$$L_1^{MI}(w_1, w_2) = \sum_{w \in C(w_1) \cup C(w_2)} | I(w, w_1) - I(w, w_2) |$$

Misure su spazi

DISTANZA EUCLIDEA

- La similarità/correlazione tra due parole w_1 e w_2 è calcolata in base alla *norma2* delle differenze tra i valori di associazione con le co-occorrenze
- Valori tra 0 e ∞, più alti se la similarità/correlazione è *minore*

- Versione probabilistica

$$L_2(w_1, w_2) = \sqrt{\sum_{w \in C(w_1) \cup C(w_2)} (P(w|w_1) - P(w|w_2))^2}$$

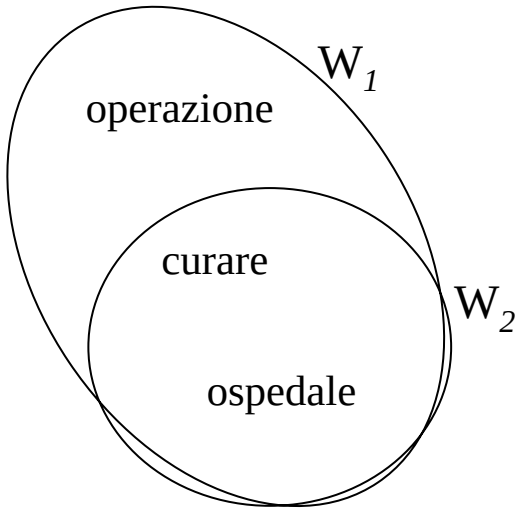
- Versione PMI

$$L_2^{MI}(w_1, w_2) = \sqrt{\sum_{w \in C(w_1) \cup C(w_2)} (I(w, w_1) - I(w, w_2))^2}$$

Misure insiemistiche

- La Distributional Hypothesis viene interpretata a livello insiemistico, considerando l'insieme W_1 delle parole che co-occorrono con w_1 e l'insieme W_2 delle parole che co-occorrono con w_2
- Viene calcolata la similarità/correlazione tra i due insiemi W_1 e W_2

- **ESEMPIO:**



w_1 = dottore W_1 ={operazione, curare,ospedale}

w_2 = infermiere W_2 ={curare,ospedale}

Misure insiemistiche

JACCARD

$$Jaccard(w_1, w_2) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|}$$

- Versione probabilistica

$$Jaccard^{CP}(w_1, w_2) = \frac{\sum_{w \in C(w_1) \cup C(w_2)} \min(P(w|w_1), P(w|w_2))}{\sum_{w \in C(w_1) \cup C(w_2)} \max(P(w|w_1), P(w|w_2))}$$

- Versione PMI

$$Jaccard^{MI}(w_1, w_2) = \frac{\sum_{w \in C(w_1) \cup C(w_2)} \min(I(w, w_1), I(w, w_2))}{\sum_{w \in C(w_1) \cup C(w_2)} \max(I(w, w_1), I(w, w_2))}$$

Misure insiemistiche

DICE FACTOR

$$Dice(w_1, w_2) = \frac{2 \times |W_1 \cap W_2|}{|W_1| + |W_2|}$$

- Versione probabilistica

$$Dice^{CP}(w_1, w_2) = \frac{2 \times \sum_{w \in C(w_1) \cup C(w_2)} \min(P(w|w_1), P(w|w_2))}{\sum_{w \in C(w_1)} P(w|w_1) + \sum_{w \in C(w_2)} P(w|w_2)}$$

- Versione PMI

$$Dice^{MI}(w_1, w_2) = \frac{2 \times \sum_{w \in C(w_1) \cup C(w_2)} \min(I(w, w_1), I(w, w_2))}{\sum_{w \in C(w_1)} I(w, w_1) + \sum_{w \in C(w_2)} I(w, w_2)}$$

Misure di correlazione e di similarità

- Le misure sin'ora descritte possono essere utilizzate sia come misure di correlazione, sia come misure di similarità

- **MISURE DISTRIBUZIONALI DI CORRELAZIONE**
  - misurano quanto sono correlate da una qualsiasi relazione semantica due target words
  - due parole distribuzionalmente correlate hanno simili co-occorrenze senza alcuna restrizione sintattica
  - Il contesto della di una target word è quindi rappresentato da tutte le parole che co-occorrono
  - Il *feature vector* di una target word ha quindi come elementi diversi da 0 tutte le parole del lessico che co-occorrono con essa

- **MISURE DISTRIBUZIONALE DI SIMILARITA'**
  - misurano quanto sono distanti in una gerarchia ISA due target words
  - due parole distribuzionalmente simili hanno simili co-occorrenze nelle stesse relazioni sintattiche
  - Il contesto di una target word è quindi rappresentato da tutte le parole che co-occorrono con essa in uno specifico ruolo sintattico
  - Il *feature vector* di una target word ha quindi come elementi diversi da 0 tutte le parole del lessico che co-occorrono con essa in un certo ruolo sintattico

Misure di correlazione e di similarità

- **ESEMPIO**

  Parole relazionate a “*music*”

  - **LSA (misura di correlazione)**

composer:0.972126176021433
beethoven:0.9649838573998915
orchestra:0.9649287214135739
musician:0.9649182070617937
tchaikovsky:0.963975389785644
string_quartet:0.9639576255386102
soloist:0.9630438894238251
english_music:0.9606993759301515
musical_style:0.9598944130975287
chamber_music:0.9597416219885411
dance_music:0.95833900150136
stravinsky:0.9581701461315594
good_music:0.957943340234033
symphony:0.9568055701201995
quartet:0.9559454812888393
double_violin_concerto:0.9529596999102421

  - **LIN DISTRIBUTIONAL APPROACH (misura di similarità)**

song 0.276
jazz 0.192
dance 0.191
art 0.191
tune 0.188
sound 0.183
opera 0.182
melody 0.178
poetry/poetry 0.177
classical music 0.177
rhythm 0.162
entertainment 0.162
lyric 0.160
recording 0.157,
Film 0.157
video 0.155

Sommario

- Semantica

- Introduzione
- Semantica lessicale distribuzionale
  - Concetti di base
  - La *Distributional Hypothesys*
  - Misure distribuzionali
- **Applicazioni di semantica lessicale distribuzionale**
  - **Costruzione di un Thesaurus**
  - Paraphrasing

Thesaurus distribuzionale [Lin, 1998]

- **COS'E' UN THESAURUS**

  Un thesaurus è un vocabolario organizzato formalmente in modo da rendere esplicite le relazioni semantiche (es. iperonimia, meronimia) tra i termini di una lingua (detti *descrittori*)

- **A COSA SERVE UN THESAURUS IN NLP**
  - I sistemi di NLP hanno necessità di sapere quali relazioni esistono tra quali termini (o parole)
  - Thesaura codificati *manualmente* possono non contenere tutte le parole. Inoltre si riferiscono solo a domini generici o predefiniti
  - Theasura per un desiderato dominio possono essere acquisiti *automaticamente* utilizzando un corpus di dominio *D*

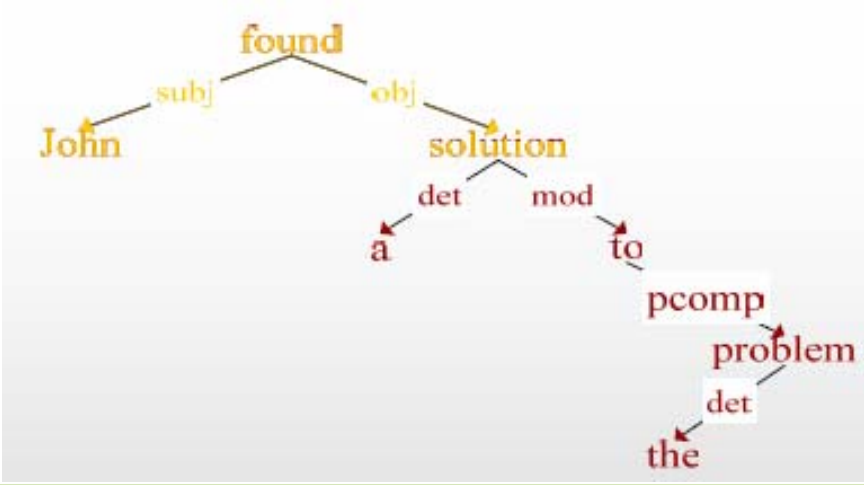
- E' possibile costruire automaticamente un thesaurus utilizzando:
  - un corpus D
  - una strategia per l'identificazione del contesto (frase, documento, finestra, sintassi…)
  - una misura per calcolare il grado di associazione tra parole (binaria, frequenza,…)
  - una misura distribuzionale di similarità tra parole (coseno, Jaccard…)
  - strumenti per l'analisi sintattica (chunker + POS tagger + parser)

Thesaurus distribuzionale

SCELTA DEL CONTESTO (*features*)

- Poichè il thesaurus contiene parole *simili*, è necessario applicare misure di similarità
- Il contesto di una target word ***t*** è quindi formato dalle co-occorrenze ***w*** in qualche relazione sintattica ***r*** con essa (*contesto grammaticale*)
- Una feature ***f*** di ***t*** è quindi una coppia *f*=(*r*|*w*)
- Il feature vector di ***t*** è quindi $\vec{t}$*{(r1|w1), (r2|w2), (r3|w3)...}*

- **ESEMPIO**



Features di “found”:

(V:subj:N|John),(V:obj:N|solution)

Features di “John”:

(-V:subj:N|find)

Thesaurus distribuzionale

SCELTA DEL CONTESTO (*features*)

METODO

1. Parsing sintattico superficiale a dipendenze *(Minipar)* su un corpus *D*
2. Per ogni parola estrai il contesto grammaticale
3. Conta le occorrenze delle features (*synt*|*c*) in *D*

degree (-V:obj:N)	
earn	371
have	336
receive	234
get	199
hold	103
obtain	59
pursue	51
complete	38
offer	34
finish	33
require	31
confer	21
achieve	21
need	20
award	16
show	16
take	16

Features **(-V:obj:N | *)** per la parola "*degree*":

f1 = (-V:obj:N | earn)

f2 = (-V:obj:N | have)

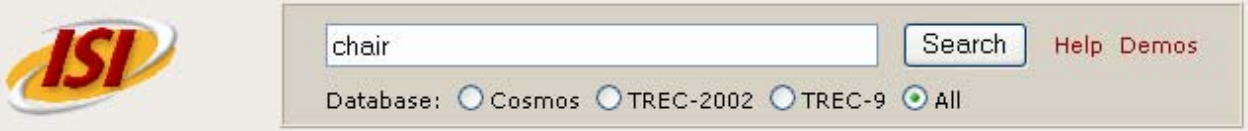
f3 = (-V:obj:N | receive)

f4 = (-V:obj:N | get)

… …

On-line feature DB demo

Thesaurus distribuzionale



chair

-N:nn:N 68 times:
rail 8, cushion 4, leg 4, wonder 4, all 2, back 2, height 2, molding 2, seat 2, umpire 2, arm 1, Bob Packwood 1, college 1, color 1, contributor 1, demand 1, department 1, designer 1, finish 1, frame 1, friend 1, list 1, manufacturer 1, meeting 1, moment 1, opposite 1, part 1, Price 1, reading 1, scrape 1, small 1, snag 1, stack 1, theme 1, ump 1, upside 1, vinyl 1, Westly 1, work 1, _SPEC 1

N:nn:N 382 times:
_NUM 61, leather 25, office 15, Vice 13, desk 12, Wicker 11, witness 10, department 8, massage 8, Dining 7, committee 5, wheel 5, Adirondack 4, club 4, kitchen 4, recliner 4, dining room 3, gold 3, party 3, patio 3, Queen Anne 3, redwood 3, William E. Simon 3, worn 3, _SPEC 3, anchor 2, arm 2, bean 2, corner 2, council 2, fold-out 2, French 2, hotel 2, kindergarten 2, pull-up 2, Quad 2, Racing 2, schoolroom 2, university 2, wingback 2, wrought iron 2, ag 1, agency 1, Alexander Hamilton 1, bag 1, Balinese 1, Barcelona 1, bent 1, Big Bird 1, blood bank 1, board 1, Breuer 1, cafeteria 1, caucus 1, change 1, cocktail 1, concert 1, conference 1, cotton 1, Courtside 1, dinner 1, director 1, donor 1, examination 1, faculty 1, FCC 1, glass fiber 1, gliding 1, harp 1, hiring 1, His-and-hers 1, hole 1, iceberg 1, jury 1, laser 1, library 1, locker room 1, loss 1, Louis 1, makeup 1, molding 1, mud 1, N 1, oak 1, olive green 1, orchestra 1, Paris 1, park 1, plaid 1, playing 1, plywood 1, poetry 1, poolside 1, press 1, Pullman 1, push 1, ragtag 1, rattan 1, reading 1, Reagan-Bush 1, red 1, relaxation 1, restaurant 1, Rockin 1, Santa 1, search committee 1, selection 1, shaker 1, shape 1, shelter 1, sideline 1, Ski 1, slipper 1, Slovak 1, sofa 1, stack 1, starch 1, study 1, style 1, subcommittee 1, suite 1, sun 1, tambourine 1, term 1, timeout 1, tripod 1, turquoise 1, use 1, Vatican 1, village 1, Waiting 1, waiting room 1, wall 1, walnut 1, wedding 1, wrestle 1, writing 1

-N:conj:N 248 times:
table 51, sofa 24, desk 21, bed 8, bench 5, book 4, couch 4, mattress 3, professor 3, Tables 3, typewriter 3, wall 3, base 2, chaise 2, chaise longue 2, dining table 2, dish 2, drape 2, kitchen table 2, love seat 2, pillow 2, rug 2, seating 2, sheet 2, supply 2, wheelchair 2, writing table 2, Airlines 1, aisle 1, antique 1, appointment 1, article 1, artwork 1, back 1, belt 1, bleachers 1, bottle 1, cabin 1, carpet 1, cash register 1, chain 1, chair 1, chest 1, chrome 1, cloth 1, conceit 1, Congressional Black Caucus 1, console table 1, cost 1, counter 1, cup 1, cushion 1, Dean 1, dinner 1, dresser 1, drum 1, electric typewriter 1, everything 1,

Thesaurus distribuzionale

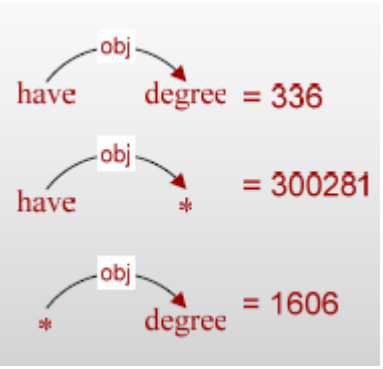
MISURA DI ASSOCIAZIONE (*valore delle features*)

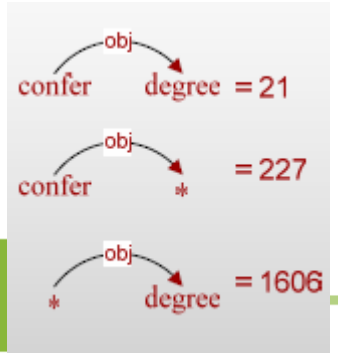
- Il feature vector rappresenta le caratteristiche di ***t*** *:* quindi ogni feature ***f*** deve rappresentare una informazione rilevante che caratterizzi ***t***
- E' quindi necessario scegliere una buona misura di associazione per il calcolo dei valori delle features:
  - frequenza *VS* Mutua Informazione
  - Mutua Informazione è più informativa !!

- **ESEMPIO**

Valori delle features **(-V:obj:N | *)** per la parola "*degree*" utilizzando *F* o *I*

w	F	I
earn	371	7.0528
have	336	3.61231
receive	234	5.48029
Get	199	4.293
hold	103	4.46028
obtain	59	5.5881
pursue	51	5.7867
complete	38	4.5451
offer	34	3.71229
finish	33	4.56256
require	31	4.0958
confer	21	7.88787
achieve	21	4.77196
need	20	2.96236
award	16	4.60602
show	16	2.91462
take	16	1.62973





Il verbo "***confer"*** è molto più caratteristico di "***have"*** per la parola "*degree*"

Thesaurus distribuzionale

MISURA DI ASSOCIAZIONE (*valore delle features*)

METODO

1. Accedi al database contenente le frequenze di co-occorrenza grammaticale
2. Per ogni parola ***t*** calcola il valore delle features ***f*** utilizzando la formula stimata della ***I*** :

$$I(t, f) \approx \log_2 \frac{\frac{c_{tf}}{N}}{\frac{\sum_{i=1}^{n} c_{if}}{N} \times \frac{\sum_{j=1}^{m} c_{tj}}{N}}$$

Dove:

- *n* è il numero di parole nel corpus
- *m* è il numero di features
- *N* è il numero di tutte le features di tutte le parole

Thesaurus distribuzionale

MISURA DI SIMILARITA'

- A questo punto ogni parola ***t*** è rappresentata da un *feature vector* di valori di co-occorrenza grammaticale calcolato utilizzando PMI

- Per creare il thesaurus è ora quindi necessario:
  - per ogni parola $\mathbf{t_1}$ calcolare la similarità con ogni altra parola $\mathbf{t_2}$ del corpus utilizzando la misura distribuzionale **coseno**
  - per ogni parola $\mathbf{t_1}$ ordinare in ordine decrescente le parole più simili in base al coseno

$$sim(t_1, t_2) = \cos(t_1, t_2) = \frac{\sum_f I(t_1, f) \times I(t_2, f)}{\sqrt{I(t_1, f)^2} \times \sqrt{I(t_2, f)^2}}$$

Thesaurus distribuzionale

MISURA DI SIMILARITA'

- **PROBLEMA:**
  - Il costo del calcolo della similarità su tutte le parole di un corpus è elevatissimo
  - Esempio:

    per un corpus di 200K parole (dimensioni medie) è necessario calcolare:

    200K parole $*$ 200K parole $*$ 12M features !!!!

- **SOLUZIONE:**
  - poiché i *feature vectors* sono generalmente sparsi, è possibile utilizzare *reverse-indexing* sulle feartures
  - grazie al reverse-indexing è possibile rintracciare le parole che hanno certe features
  - per ogni parola $\mathbf{t}_1$ basta ordinare le features in base a *I*, e considerare solo le features con i valori più alti
  - calcolare la similarità di $\mathbf{t}_1$ solo con le parole che hanno tali features

Thesaurus distribuzionale

- **COSA *IDENTIFICA* IL THESAURUS DISTRIBUZIONALE ?**
  - SI : parole simili (stesso POS, stesso contesto grammaticale)
  - NO: sinonimi (similarità≠sinonimia)
  - NO: parole correlate non simili
- **QUALI SONO I LIMITI?**
  - Il thesaurus contiene genericamente *parole simili*, non distinguendo ed isolando sinonimi, antinomi ed altre relazioni
  - Questo limite è insito nella *distributional hypothesis:*

    La Distributional Hypothesis non è abbastanza potente da consentire distinzioni semantiche tra parole simili
  - Sarebbero necessarie altri tecniche e risorse per raffinare il thesaurus (es. utilizzo di patterns linguistici)

THESAURUS DISTRIBUZIONALE ON-LINE DEMO (Dekang Lin and Patrick Pantel)

http://www.isi.edu/~pantel/Content/Demos/LexSem/thesaurus.htm

Thesaurus distribuzionale

chair Search Help Demos

Database: Cosmos TREC-2002 TREC-9 All

chair

N

sofa 0.224, armchair 0.212, furniture/furniture 0.200, table 0.190, couch 0.177, desk 0.168, stool 0.167, lamp 0.159, easy chair 0.156, folding chair 0.152, bed 0.147, lounge chair 0.145, pillow 0.144, coffee table 0.140, rug 0.140, mattress/mattress 0.137, recliner 0.136, carpet 0.136, seat 0.134, box 0.131, wall 0.131, armoire/armoire 0.130, bookcase 0.130, rocking chair 0.128, swivel chair 0.127, tile 0.125, cot 0.122, floor/floor 0.122, curtain 0.121, settee 0.121, chandelier 0.120, furnishings 0.117, bookshelf 0.116, carpeting 0.114, stove 0.114, blanket 0.113, love seat 0.113, frame 0.113, paneling 0.112, fireplace 0.112, window/window 0.112, bench 0.111, shelf 0.110, cloth 0.109, object 0.108, banquette 0.107, room 0.107, lawn chair 0.106, Mirror 0.105, door 0.104, tent 0.102, cushion 0.102, refrigerator 0.102, booth 0.102, television set 0.101, glass 0.101, tray 0.100, wallpaper 0.099, sculpture 0.099, staircase 0.099, tablecloth 0.099, bag 0.098, screen 0.098, vase 0.098, towel 0.097, sheet 0.097, bottle 0.097, shoe 0.097, camera 0.096, drape 0.096, toy 0.096, piece 0.095, roof 0.095, plate 0.095, dining table 0.095, wheelchair 0.094, file cabinet 0.093, upholstery 0.093, bedspread 0.093, drawer 0.093, crate 0.093, cupboard 0.092, bicycle 0.092, drapery 0.092, pot 0.092, tree 0.092, flooring 0.092, toilet 0.091, jacket 0.091, rack 0.091, seating 0.090, clothing 0.090, brick 0.090, bar 0.089, coffin 0.089, dresser 0.089, fence 0.089, lectern 0.089, balcony 0.089, shirt 0.089, clock 0.089, machine 0.088, headboard 0.088, quilt 0.088, crib 0.088, bathtub 0.087, lantern 0.087, candle 0.087, closet 0.087, hat 0.087, car 0.087, suitcase 0.087, fabric 0.086, Doll 0.086, card table 0.086, mat 0.086, tub 0.086, statue 0.086, pipe 0.086, picture frame 0.085, fixture 0.084, container 0.084, appliance 0.084, railing 0.084, awning 0.084, typewriter 0.084, sleeping bag 0.083, ceiling 0.083, pedestal 0.083, lighting 0.083, antique 0.083, ladder 0.082, umbrella 0.082, trash can 0.082, cubicle 0.082, cart 0.082, coat 0.082, flower 0.082, kitchen 0.081, chairmanship 0.081, Pew 0.081, sweater 0.081, bathroom 0.081, accessory 0.081, wall hanging 0.081, plaque 0.081, dress 0.080, bunk bed 0.080, painting 0.080, ornament 0.080, trailer 0.080, tv set 0.080, Slab 0.080, pad 0.080, cabinet 0.080, bleachers 0.080, pool table 0.080, light 0.079, locker 0.079, bedding 0.079, display case 0.079, basket 0.079, computer 0.079, phone 0.079, T-shirt 0.079, bucket 0.079, jewelry 0.078, dish 0.078, bike 0.078, commode/commode 0.078, lampshade 0.078, banner 0.078, blackboard 0.077, building 0.077, flag 0.077, casket 0.077, pottery 0.076, poster 0.076, filing cabinet 0.076, bunk 0.076, decoration 0.076, easel 0.076, utensil 0.076, ashtray 0.076, piece of furniture 0.075, podium 0.075, stair 0.075, doorway 0.075, tire 0.075, number 0.075, mantel 0.075, grand piano 0.075, rock 0.075, deck 0.075, napkin 0.075, boat 0.075, hammock 0.075, stone 0.075, briefcase 0.074, molding 0.074

Thesaurus

Mutua Informazione e coseno

Che differenza c'è tra queste due funzioni?

$$A(w_1, w_2) = \cos(\vec{w}_1, \vec{w}_2)$$

$$B(w_1, w_2) = I(w_1, w_2)$$

Quale valore semantico hanno queste funzioni?