Semantic distances & LSA

Stefan Trausan-Matu

"Politehnica" University of Bucharest

and

Romanian Academy Research Institute for Artificial Intelligence

Bucharest, Romania

stefan.trausan@cs.pub.ro

http://www.racai.ro/~trausan

Lexical chains

- Venture capitalists have become culture heroes in the New Economy. They're the visionary investors who sink millions of dollars into risky start-up companies, take seats on their boards to mentor the tech nerds--and then cash out at a huge profit. The name "Sand Hill Road," the street in Menlo Park where Kleiner Perkins and some of the other famous venture capitalists have their offices, has assumed an almost magical quality, as if the partners were dispensing fairy dust rather than cash.

Applications of lexical chains

- Verification of text cohesion
- Text segmentation
- Summarization
- Word sense disambiguation
- Determination of discourse structure
- Automatic hypertext generation
- Intelligent spelling checking
- Information retrieval

Building lexical chains

- Text scanning and detection of semantically related words – small distance
- Constructing a set of potential lexical chains
- Computational problems

Low semantic distance = high similarity

- *bank–money* *bank–river*
- *apple–fruit* *orange-fruit*
- *pen–paper* *pen–pencil*
- *men-crowd* *red-blue*
- *man-human* *woman-human*
- *man-men* *men-women*
- *man-bike* *wheel–bike*
- *sweet-bitter* *sweet-dessert*
- *desert-storm* *desert-defect*
- *singer-song* *hacker-soft*

Closeness relations

- Synonym
- Hyponym/hypernim
- Meronym/Holonym
- Antonym
- Entailment
- Typicality

High semantic distance

- bike-cat
- software-dog
- dog-amoeba
- idea-sleep
- runner-paint

Semantic distance according to:

- Dictionaries (Kozima and Ferugori, Kozima and Ito)
- Thesauri (e.g. Roget – Morris and Hirst; Bunrui–goihyo Japanese thesaurus – Okamura and Honda)
- Semantic networks (e.g. MeSH –Medical Subject Headings – Rada)
- Ontologies – WordNet, FrameNet

Chains of Wordnet senses

- [venture capitalist(1), hero(1), visionary(1), investor(1), mentor(1), nerd(1), profit(2), venture capitalist(1), quality(1), partner(1), fairy(1)]

Rada et al.

- $dist_R(c1,c2)$=min nr of edges between c1 and c2

Hirst and St'Onge

- Morris and Hirst – applied for WordNet
- Directions:
  - *downward* (cause, hyponym, holonym, entailment)
  - *upward* (hypernym, meronym)
  - *horizontal* (similar, participle_of, see_also, antonyme, attribute)

Hirst and St'Onge

- Relations:
  - *extra-strong* – word repetition
  - *strong*
    - same synset (desert–defect)
    - antonyms (cold-hot)
    - sub-phrase (school – elementary school)
  - *medium-strength*

Medium strength

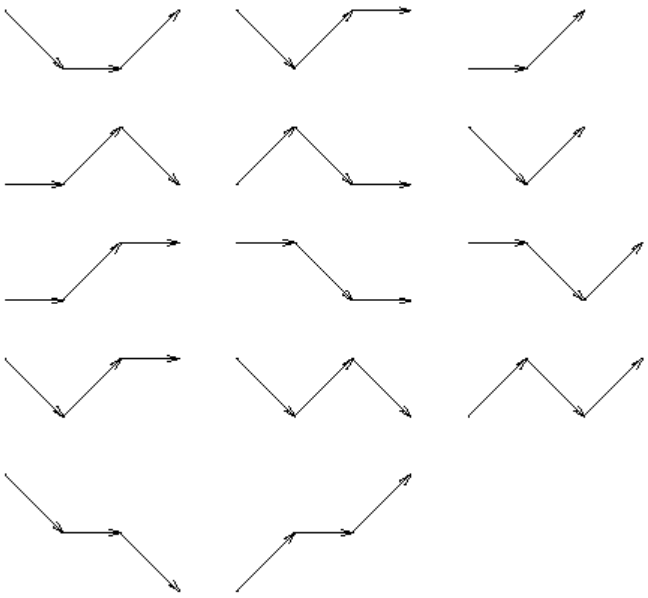
rel_{HS} =

$$\begin{cases} 3C, & \text{for extra-strong relations} \\ 2C, & \text{for strong relations} \\ C - path\_length - (k * \#\ changes\_in\_direction), & \text{for medium strength relations} \\ 0 & \text{otherwise} \end{cases}$$

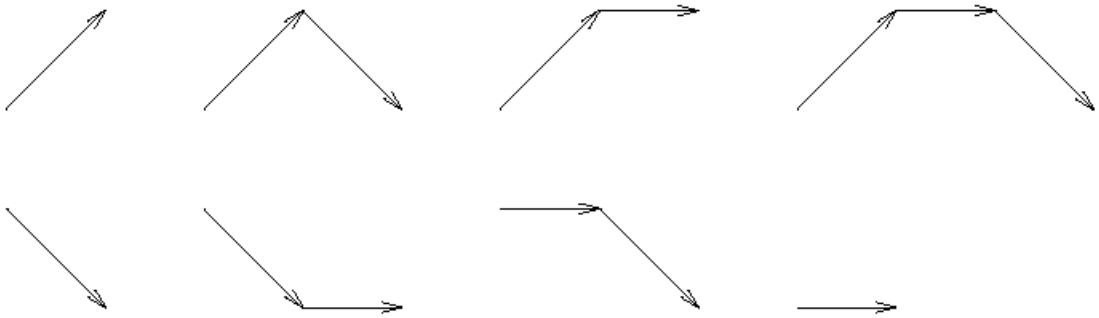
Forbidden sequences

- No other direction can precede an upward direction
- Only one direction change is allowed
- Exception: upward – horizontal - downward

Forbidden sequences



Allowed sequences



Sussna

- Weight of edge depends on fanout *nr*: the number of arcs leaving *c*.
  - $w(c1 \rightarrow c2) = 2 - 1 / nr(c1)$
- If *c*1 and *c*2 are adjacent:
  - $Dist_S(c1,c2)=(w(c1 \rightarrow c2)+w(c2 \rightarrow c1)) / 2d$
  - d = the depth of the edge in the ontology
- If not adiacent – add distances to shortest paths

Wu si Palmer

- $c = lso(c1;c2)$, lso=*lowest super ordinate*
- $\mathbf{sim_{WP}(c1,c2)=(2xN)/(N1+N2+2xN)}$,
- Ni – the path from ci to c
- N – the path from c to the root
- $\mathbf{dist_{WP}(c1,c2)=(N1+N2)/(N1+N2+2xN)}$,

Leacock si Chodorow

- **sim_{LC}(c1,c2)=-log(1+length(c1,c2))/(2xD))**
- length(c1,c2) is Rada's shortest path
- D is the height of the ontology

Resnik

- The similarity of two concepts is their shared information → information content of the lowest superordinate
- Information content of *c* is given by -log $p(c)$ → the less frequent it is, the more information it contains.
- **$sim_R(c1,c2) = -\log p(lso(c1;c2))$**

Lin

- The similarity between arbitrary objects *A* and *B* is measured by the ratio between the amount of information needed to state their commonality and that needed to fully describe what they are.

Lin

- **$sim_L(c1,c2)=2\log p(lso(c1,c2))/(\log p(c1)+\log p(c2))$**

Jiang and Conrath

- **$dist_{JC}(c1,c2)=2\log p(lso(c1,c2)) - (\log p(c1)+\log p(c2))$**

Problems in detecting lexical chains

- Under-chaining – the difficulty of finding semantically related words
- Over-chaining – wrong determination of semantically unrelated words

Under-chaining – Named entitities recognition

- Venture capitalists have become culture heroes in the New Economy. They're the visionary investors who sink millions of dollars into risky start-up companies, take seats on their boards to mentor the tech nerds--and then cash out at a huge profit. The name "Sand Hill Road," the street in Menlo Park where Kleiner Perkins and some of the other famous venture capitalists have their offices, has assumed an almost magical quality, as if the partners were dispensing fairy dust rather than cash.

Under-chaining – Coreference resolution

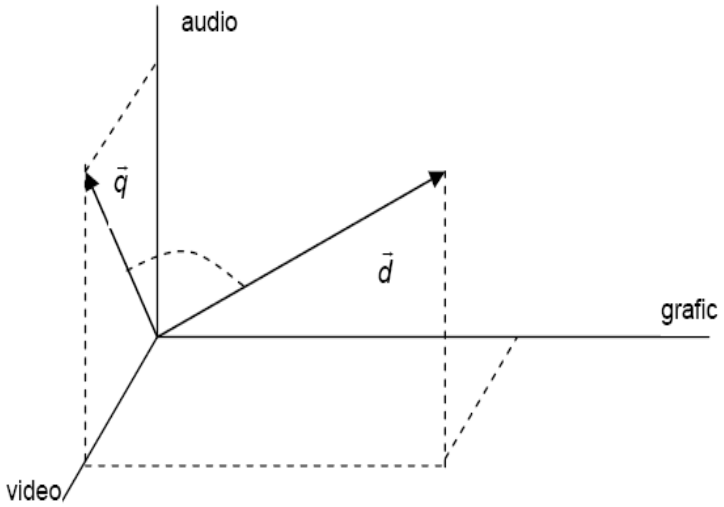
- Venture capitalists have become culture heroes in the New Economy. They're the visionary investors who sink millions of dollars into risky start-up companies, take seats on their boards to mentor the tech nerds--and then cash out at a huge profit. The name "Sand Hill Road," the street in Menlo Park where Kleiner Perkins and some of the other famous venture capitalists have their offices, has assumed an almost magical quality, as if the partners were dispensing fairy dust rather than cash.

Over-chaining

- Venture capitalists have become culture heroes in the New Economy. They're the visionary investors who sink millions of dollars into risky start-up companies, take seats on their boards to mentor the tech nerds--and then cash out at a huge profit. The name "Sand Hill Road," the street in Menlo Park where Kleiner Perkins and some of the other famous venture capitalists have their offices, has assumed an almost magical quality, as if the partners were dispensing fairy dust rather than cash.

Semantic spaces in Latent Semantic Indexing (LSI)

Vector space model



The LSI idea

- Reducing the dimensionality of the vector space, similarly to the *least squares method*
- The effect is the creation of semantic spaces containing semantically related words
- http://lsa.colorado.edu

Terms-documents array

(ex. from Manning and Schutze, 1999)

$$A = \left(\begin{array}{c|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline \cos\, monaut & 1 & 0 & 1 & 0 & 0 & 0 \\ astronaut & 0 & 1 & 0 & 0 & 0 & 0 \\ moon & 1 & 1 & 0 & 0 & 0 & 0 \\ car & 1 & 0 & 0 & 1 & 1 & 0 \\ truck & 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right)$$

Singular value decomposition (SVD)

$$A_{txd} = T_{txn} S_{nxn} D^{T}_{dxn}$$

n=min(t,d)

T

$$T^T = \begin{pmatrix} & \cos monaut & astronaut & moon & car & truck \\ \dim 1 & -0.44 & -0.13 & -0.48 & -0.70 & -0.26 \\ \dim 2 & -0.30 & -0.33 & -0.51 & 0.35 & 0.65 \\ \dim 3 & 0.57 & -0.59 & -0.37 & 0.15 & -0.41 \\ \dim 4 & 0.58 & 0.00 & 0.00 & -0.58 & 0.58 \\ \dim 5 & 0.25 & 0.73 & -0.61 & 0.16 & -0.09 \end{pmatrix}$$

S

$$S = \begin{pmatrix} 2.16 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.59 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.28 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.39 \end{pmatrix}$$

D

$$D^T = \left(\begin{array}{c|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline \dim 1 & -0.75 & -0.28 & -0.20 & -0.45 & -0.33 & -0.12 \\ \dim 2 & -0.29 & -0.53 & -0.19 & 0.63 & 0.22 & 0.41 \\ \dim 3 & 0.28 & -0.75 & 0.45 & -0.20 & 0.12 & -0.33 \\ \dim 4 & 0.00 & 0.00 & 0.58 & 0.00 & -0.58 & 0.58 \\ \dim 5 & -0.53 & 0.29 & 0.63 & 0.19 & 0.41 & -0.22 \end{array}\right)$$

Properties of SVD

- SVD is unique
- T, D are orthonormal:

$$T^T T = D^T D = I$$

- S values are sorted

Reduced A

$$\Delta = \| A - \hat{A} \|_2$$

- By SVD on maps the n-dimension space on a k-dimension one, with n >>k
- Common values for k are 100 and 150.

B

$$B = S_{2x2} D^T_{dx2}$$

$$B = \begin{array}{c|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline \dim 1 & -1.62 & -0.60 & -0.04 & -0.97 & -0.71 & -0.26 \\ \dim 2 & -0.46 & -0.84 & -0.30 & 1.00 & 0.35 & 0.65 \end{array}$$

Document correlation

(Manning and Schutze, 1999)

$$A^T A = (TSD^T)^T TSD^T = DS^T T^T TSD^T = (DS)(DS)^T = (SD^T)^T (SD^T) = B^T B$$

	d_1	d_2	d_3	d_4	d_5	d_6
d_1	1.00					
d_2	0.78	1.00				
d_3	0.40	0.88	1.00			
d_4	0.47	−0.18	−0.62	1.00		
d_5	0.74	0.16	−0.32	0.94	1.00	
d_6	0.10	−0.54	−0.87	0.93	0.74	1.00

Term correlation

$$AA^T = TSD^T (TSD^T)^T = TSD^T DS^T T^T = (TS)(TS)^T$$