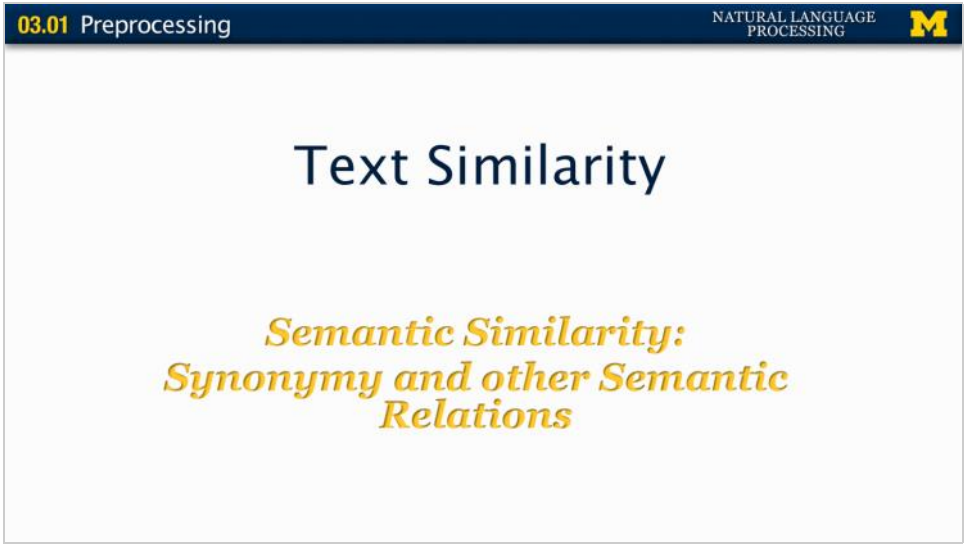
Week 3

2016年7月19日 星期二　　下午1:24

word net, lexical network

3.1



Synonyms and Paraphrases

Different words mean the same

- Example: post-close market announcements

The S&P 500	climbed	6.93, or 0.56 percent, to 1,243.72,	its best close	since June 12, 2001.
The Nasdaq	gained	12.22, or 0.56 percent, to 2,198.44	for its best showing	since June 8, 2001.
The DJIA	rose	68.46, or 0.64 percent, to 10,705.55,	its highest level	since March 15.

Synonyms

近义词.

Synonyms 近义词.

- Different words (and also word compounds) can have similar meanings.
  - For example, the adjectives *tepid* and *lukewarm* have very similar meanings and can be substituted for one another (*tepid water* vs. *lukewarm water).*
- True synonyms are actually relatively rare.
  - For example, even though *big* and *large* are often thought of as synonyms, consider the difference between *Big Leagues* and *Large Leagues*. ☺
- The verbs *sweat* and *perspire* are also near synonyms.
  - However, they differ in their frequency of use and the type of text in which they are likely to appear.

Polysemy 多义词.

Polysemy 多义词

例: book

- Polysemy is the property of words to have multiple senses.
- For example, the noun *book* can refer to the following:
  - A literary work (e.g., "Anna Karenina")
  - A stack of pages (e.g., a notebook)
  - A record of business transactions (think "bookkeeper")
  - A record of bets (think "bookmaker")
  - A list of buy and sell orders in a financial market

Polysemy

- The same word can also have multiple parts of speech, each with its own set of senses. For example, the word *book*, as a verb can mean "make a reservation for" or "occupy".
- The different senses of the same word don't have to be equally frequent.
- Some of the senses may overlap (e.g., the first two senses of *book* on the previous slide). That's partially why different dictionaries list different sets of word senses for the same word.
  - "My favorite books are Anna Karenina and my father's checkbook" ☺
- Some words can be highly polysemous (e.g., the verb "get" has at least 35 different meanings, according to Wordnet).

book n. v.

get

Other Semantic Relations

- Antonymy (near opposites)
  - *raise–lower*
- Hypernymy
  - a *deer* is a hypernym for *elk*
- Hyponymy (the inverse of hypernymy)
- Membership Meronymy:
  - a *flock* includes *sheep* (or *birds*)
- Part Meronymy:
  - a *table* has *legs*

general concept.

Antonymy 反义

Hypernymy.

Membership Meronymy.

Part Meronymy.

Synsets

- Semantic relations hold between word senses, not between words.
- Examples:
  - the antonym of *hot* can be either *mild* or *cold* (or *unattractive*) depending on the specific sense of *hot*.
  - the immediate hypernym of *bar* can be one of the following, among others: *room*, *musical notation*, *obstruction*, *profession*, depending on the sense of *bar*.
- The term *synset* is used to group together all synonyms of the same word. If a word is polysemous, it may be associated with multiple synsets.

Synset : group together all synonyms of the same word

为不同词序.

Text Similarity

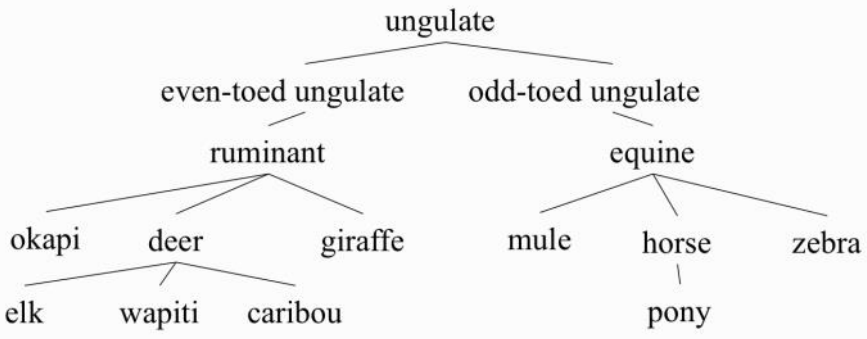
Wordnet

Wordnet

- Wordnet is a project run by George Miller (1920–2012) and Christiane Fellbaum at Princeton University.
- It includes a database of words (mainly nouns and verbs but also adjectives and adverbs) and semantic relations between them.
- The main relation is hypernymy, so the overall structure of the database is more tree-like (see next slide).
- References:
  - George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39–41.
  - Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

George Miller

Tree-like Structure of Wordnet



Wordnet Example (1/6)

The noun bar has 11 senses

1. barroom, bar, saloon, ginmill, taproom -- (a room where alcoholic drinks are served over a counter)
2. bar -- (a counter where you can purchase food or drink)
3. bar -- (a rigid piece of metal)
4. measure, bar -- (notation for a repeating pattern of musical beats; written followed by a vertical bar)
5. bar -- (usually metal placed in windows to prevent escape)
6. prevention, bar -- (the act of preventing)
7. bar -- (a unit of pressure equal to a million dynes per square centimeter)
8. bar -- (a submerged (or partly submerged) ridge in a river or along a shore)
9. legal profession, bar, legal community -- (the body of individuals qualified to practice law)
10. cake, bar -- (a block of soap or wax)
11. bar -- ((law) a railing that encloses the part of the courtroom where the the judges and lawyers sit and the case is tried)

The verb bar has 4 senses

1. bar, debar, exclude -- (prevent from entering; keep out; "He was barred from membership in the club")
2. barricade, block, blockade, block off, block up, bar -- (render unsuitable for passage; "block the way"; "barricade the streets")
3. banish, relegate, bar -- (expel, as if by official decree; "he was banished from his own country")
4. bar -- (secure with, or as if with, bars; "He barred the door")

Wordnet Example (2/6)

Sense 1
barroom, bar, saloon, ginmill, taproom
=> room
=> area
=> structure, construction
=> artifact, artefact
=> object, physical object
=> entity, something

Sense 2
bar
=> counter
=> table
=> furniture, piece of furniture, article of furniture
=> furnishings
=> instrumentality, instrumentation
=> artifact, artefact
=> object, physical object
=> entity, something

Wordnet Example (3/6)

Sense 3
bar
=> implement
=> instrumentality, instrumentation
=> artifact, artefact
=> object, physical object
=> entity, something

Sense 4
measure, bar
=> musical notation
=> notation, notational system
=> writing, symbolic representation
=> written communication, written language
=> communication
=> social relation
=> relation
=> abstraction

Wordnet Example (4/6)

Sense 5
bar
=> obstruction, impediment, impedimenta
=> structure, construction
=> artifact, artefact
=> object, physical object
=> entity, something

Sense 6
prevention, bar
=> hindrance, interference, interfering
=> act, human action, human activity

Sense 7
bar
=> pressure unit
=> unit of measurement, unit
=> definite quantity
=> measure, quantity, amount, quantum
=> abstraction

Wordnet Example (5/6)

Sense 8
bar
=> ridge
=> natural elevation, elevation
=> geological formation, geology, formation
=> natural object
=> object, physical object
=> entity, something
=> barrier
=> mechanism
=> natural object
=> object, physical object
=> entity, something

Wordnet Example (6/6)

Sense 9
legal profession, bar, legal community
=> profession, community
=> occupation, vocation, occupational group
=> body
=> gathering, assemblage
=> social group
=> group, grouping

Sense 10
cake, bar
=> block
=> artifact, artefact
=> object, physical object
=> entity, something

Familiarity and Polysemy

board used as a noun is familiar (polysemy count = 9)

bird used as a noun is common (polysemy count = 5)

cat used as a noun is common (polysemy count = 7)

house used as a noun is familiar (polysemy count = 11)

information used as a noun is common (polysemy count = 5)

retrieval used as a noun is uncommon (polysemy count = 3)

serendipity used as a noun is very rare (polysemy count = 1)

每个词用作某义的常见程度.

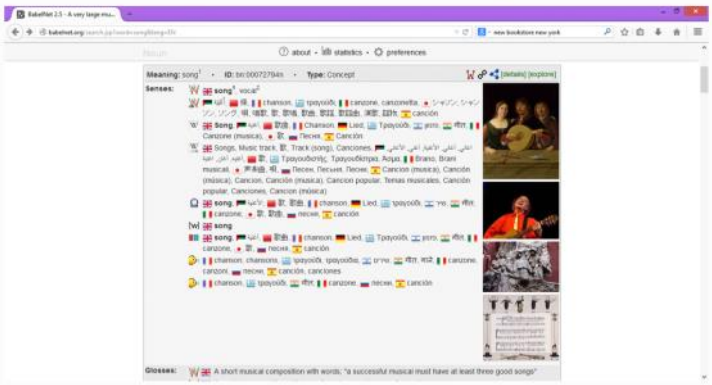
Text Similarity

Other Lexical Networks

External Links

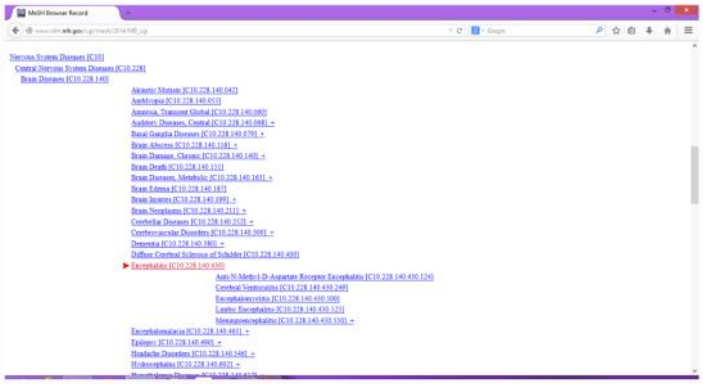
- EuroWordNet
  - http://www.illc.uva.nl/EuroWordNet/
- Open Thesaurus
  - http://www.openthesaurus.de/
- Freebase
  - http://www.freebase.com
- DBPedia
  - http://www.dbpedia.org
- BabelNet
  - http://babelnet.org
- Various thesauri

BabelNet Example



MeSH

Medical Subject Headings



- http://www.nlm.nih.gov/mesh/MBrowser.html

NLP

A specific kind of word similarity measure

Thesaurus-based word similarity: Tree-based (?)

3.2

A specific kind of word similarity measure

Thesaurus-based word similarity:

Tree-based?

NLP

03.02 Thesaurus-Based Word Similarity Methods NATURAL LANGUAGE PROCESSING

Text Similarity

Thesaurus-based Word Similarity Methods

03.02 Thesaurus-Based Word Similarity Methods NATURAL LANGUAGE PROCESSING

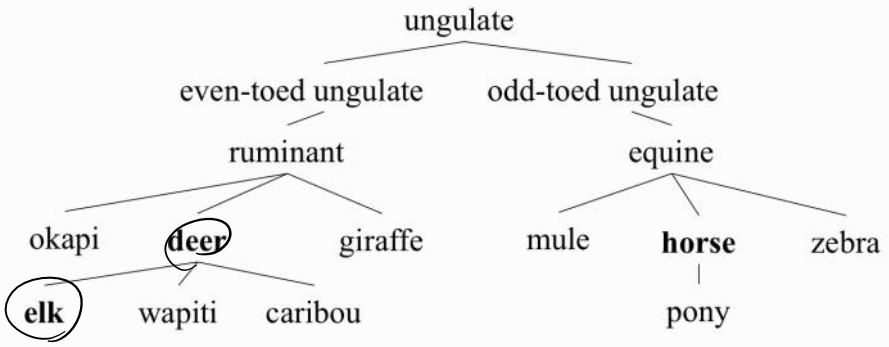
Quiz

- Which pair of words exhibits the greatest similarity?
  - 1. Deer-elk
  - 2. Deer-horse
  - 3. Deer-mouse
  - 4. Deer-roof

Quiz Answer

- Which pair of words exhibits the greatest similarity?
  - **1. Deer-elk** ★ mentioned in the previous lectured
  - 2. Deer-horse
  - 3. Deer-mouse
  - 4. Deer-roof
- Why?
- Remember the Wordnet tree:

Remember Wordnet



Path Similarity

- Version 1
  - Sim (v,w) = - pathlength (v,w)
- Version 2
  - Sim (v,w) = - log pathlength (v,w)

greater distance, smaller similarity.

take log

Problems With This Approach

- There may be no tree for the specific domain or language
- A specific word (e.g., a term or a proper noun) may not be in any tree
- IS-A (hypernym) edges are not all equally apart in similarity space

Problems with this distance calculation

in two trees, but may have close semantic meaning

Path Similarity Between Two Words

- Version 3 (Philip Resnik)

  Sim (v,w) = – log *P*(LCS(v,w))

  where LCS = lowest common subsumer, e.g.

  ungulate for deer and horse

  deer for deer and elk

Philip Resnik

* Lowest common subsumer

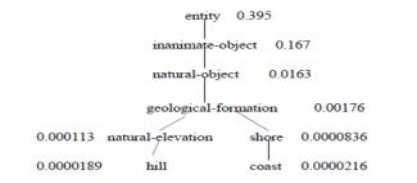
Probability etc.

Information Content

- Version 4 (Dekang Lin)
  - Wordnet augmented with probabilities (Lin 1998)
  - IC(c) = -log P(c)
  - Sim (v,w) = 2 x log *P*(LCS(v,w)) / (log *P*(v) + log *P*(w))

Dekan Lin

→ normalize by the prob of the two individual type.



$$\text{sim}(\text{Hill, Coast}) = \frac{2 \times \log P(\text{Geological-Formation})}{\log P(\text{Hill}) + \log P(\text{Coast})}$$

$$= 0.59$$

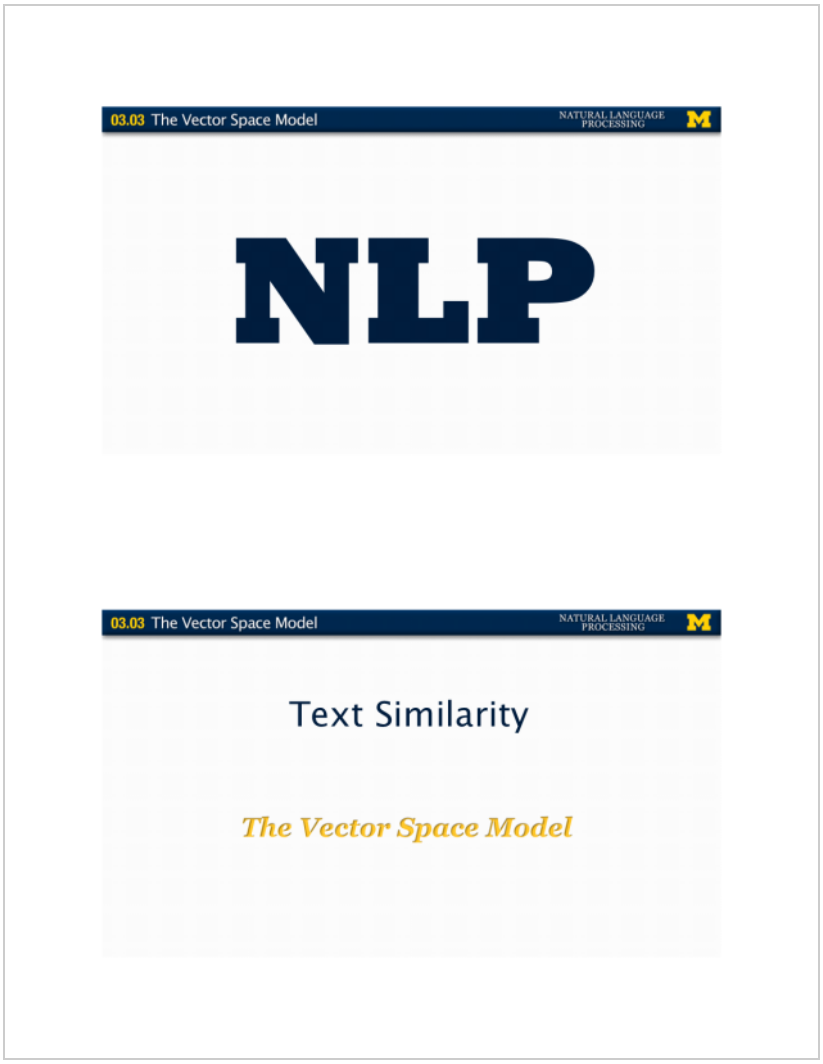
Wordnet Similarity Software

- WordNet::Similarity (Perl)
  - http://www.d.umn.edu/~tpederse/similarity.html
- NLTK (Python)
  - http://www.nltk.org

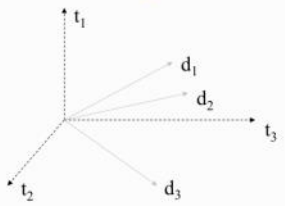
```
>>> dog.lin_similarity(cat, brown_ic)
0.879
>>> dog.lin_similarity(elephant, brown_ic)
0.531
>>> dog.lin_similarity(elk, brown_ic)
0.475
```

03.02 Thesaurus-Based Word Similarity Methods NATURAL LANGUAGE PROCESSING

3.3


03.03 The Vector Space Model
NATURAL LANGUAGE PROCESSING
NLP
03.03 The Vector Space Model
NATURAL LANGUAGE PROCESSING
Text Similarity
The Vector Space Model

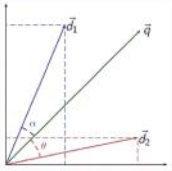
The Vector Space Model



represent document in a space where each term is a dimension.

Document Similarity

- Used in information retrieval to determine which document (d_1 or d_2) is more similar to a given query q.
- Note that documents and queries are represented in the same space.
- Often, the angle between two vectors (or, rather, the cosine of that angle) is used as a proxy for the similarity of the underlying documents.



Cosine Similarity

- The Cosine measure is computed as the normalized dot product of two vectors:

$$\sigma(D,Q) = \frac{|D \cap Q|}{\sqrt{|D||Q|}} = \frac{\sum(d_i q_i)}{\sqrt{\sum(d_i)^2}\sqrt{\sum(q_i)^2}}$$

-> lengths of two vectors.

- A variant of Cosine is the Jaccard coefficient:

$$\sigma(D,Q) = \frac{|D \cap Q|}{|D \cup Q|}$$

Example

- What is the cosine similarity between:
  - D= "cat,dog,dog" = <1,2,0>
  - Q= "cat,dog,mouse,mouse" = <1,1,2>
- Answer:

$$\sigma(D,Q) = \frac{1\times1+2\times1+0\times2}{\sqrt{1^2+2^2+0^2}\sqrt{1^2+1^2+2^2}} = \frac{3}{\sqrt{5}\sqrt{6}} \approx 0.55$$

- In comparison:

$$\sigma(D,D) = \frac{1\times1+2\times2+0\times0}{\sqrt{1^2+2^2+0^2}\sqrt{1^2+2^2+0^2}} = \frac{5}{\sqrt{5}\sqrt{5}} = 1$$

=> largest possible similarity

03.03 The Vector Space Model — NATURAL LANGUAGE PROCESSING

Quiz

- Given the three documents
  - D_1 = <1,3>
  - D_2 = <10,30>
  - D_3 = <3,1>
- Compute the cosine scores
  - $\sigma(D_1,D_2)$
  - $\sigma(D_1,D_3)$
- What do the numbers tell you?

largest possible sim

$= 0.6$

$= \frac{1\times10+3\times30}{\sqrt{1^2+3^2}\sqrt{10^2+30^2}} = \frac{100}{\sqrt{10}\times\sqrt{100+900}} = \frac{100}{\sqrt{10}\times1000} = \frac{100}{100} = 1$

03.03 The Vector Space Model — NATURAL LANGUAGE PROCESSING

Answers to the Quiz

$\sigma(D_1,D_2) = 1$

one of the two documents is a scaled version of the other

$\sigma(D_1,D_3) = 0.6$

swapping the two dimensions results in a lower similarity

Quiz

- What is the range of values that the cosine score can take?

[0, 1)

0 = orthogonal

Answer to the Quiz

- In general, the cosine function has a range of [-1,1]
- However, when the two vectors are both in the first quadrant (since all word counts are non-negative), the range is [0,1].

Vectors in the first quadrants = non negative counts

Text Similarity

The Vector Space Model Applied to Word Similarity

Distributional Similarity

- Two words that appear in similar contexts are likely to be semantically related, e.g.,
  - schedule a test *drive* and investigate Honda's financing options
  - Volkswagen debuted a new version of its front-wheel-*drive* Golf
  - the Jeep reminded me of a recent *drive*
  - Our test *drive* took place at the wheel of loaded Ford EL model
- "You will know a word by the company that it keeps." (J.R. Firth 1957)

Distributional Similarity.

Honda. VW, Jeep. Ford all appear with drive

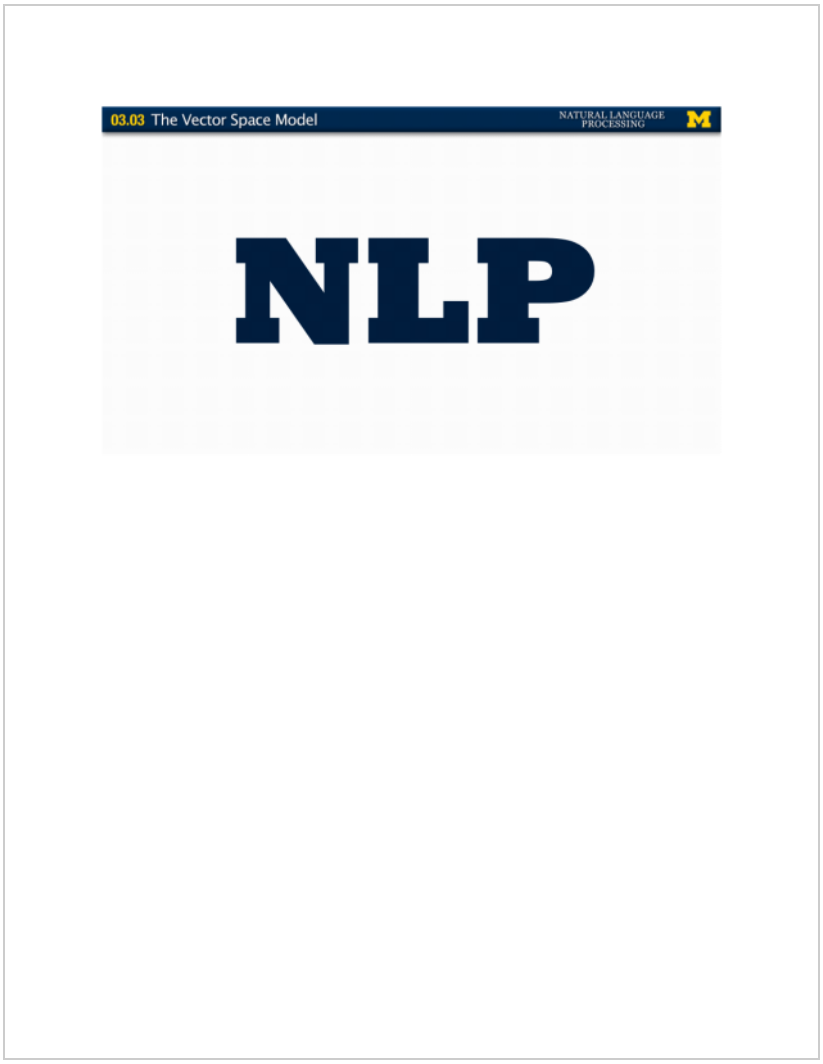
Distributional Similarity

- The context can be any of the following:
  - The word before the target word
  - The word after the target word
  - Any word within *n* words of the target word
  - Any word within a specific syntactic relationship with the target word (e.g., the head of the dependency or the subject of the sentence)
  - Any word within the same sentence
  - Any word within the same document

same sentence
same d.c.

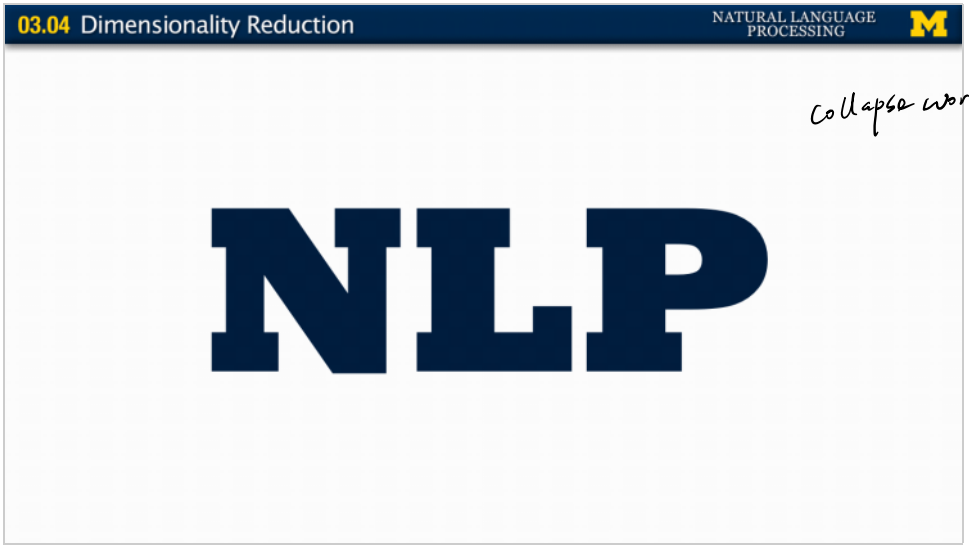
Association Strength

- Frequency matters: we want to ignore spurious word pairings.
- However, frequency alone is not sufficient.
- A common technique is to use pointwise mutual information (PMI).
- Here *w* is a word and *c* is a feature from the context $\text{PMI}(w,c)=\log P(w,c)/P(w)P(c)$



3.4

Dimensionality reduction



collapse words into categories

Text Similarity

Dimensionality Reduction

Problems with the Simple Vector Approaches to Similarity

- Polysemy (sim < cos)
  - bar, bank, jaguar, hot
- Synonymy (sim > cos)
  - building/edifice, large/big, spicy/hot
- Relatedness (people are really good at figuring this)
  - doctor/patient/nurse/treatment
- Sparse matrix
- Needed
  - dimensionality reduction

TOEFL Synonyms and SAT Analogies

- Word similarity vs. analogies

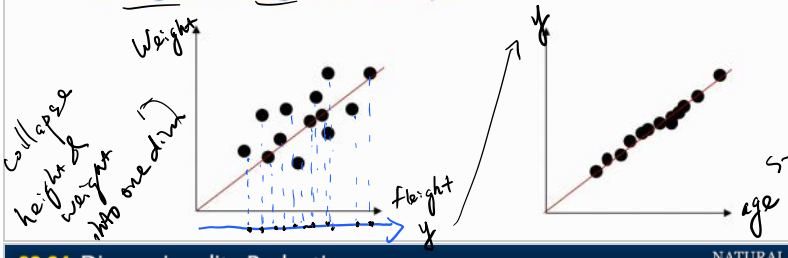
Stem:		levied
Choices:	(a)	imposed
	(b)	believed
	(c)	requested
	(d)	correlated
Solution:	(a)	imposed

Stem:	mason:stone
Choices:	(a) teacher:chalk
	(b) carpenter:wood
	(c) soldier:gun
	(d) photograph:camera
	(e) book:word
Solution:	(b) carpenter:wood

Example from Peter Turney

Dimensionality Reduction

- Looking for hidden similarities in data
- Based on matrix decomposition
- Height/weight example



matrix decomposition

hidden similarity

students older are both taller and heavier

Vectors and Matrices

- A matrix is an *m* x *n* table of objects (in our case, numbers)
- Each row (or column) is a vector.
- Matrices of compatible dimensions can be multiplied together.
- What is the result of the multiplication below?

$$\begin{bmatrix}1 & 2 & 4\\ 2 & 5 & 7\\ 4 & 9 & 14\end{bmatrix}\begin{bmatrix}2\\ 1\\ -1\end{bmatrix} = \begin{bmatrix}0\\ 2\\ 3\end{bmatrix}$$

3×3 3×1 3×1

compatible dimension

Answer to the Quiz

$$\begin{bmatrix}1 & 2 & 4\\ 2 & 5 & 7\\ 4 & 9 & 14\end{bmatrix}\begin{bmatrix}2\\ 1\\ -1\end{bmatrix} = \begin{bmatrix}1\times 2+2\times 1+4\times(-1)\\ 2\times 2+5\times 1+7\times(-1)\\ 4\times 2+9\times 1+14\times(-1)\end{bmatrix} = \begin{bmatrix}0\\ 2\\ 3\end{bmatrix}$$

Matrix decomposition

Eigenvectors and Eigenvalues

- An eigenvector is an implicit "direction" for a matrix

$$A\vec{v} = \lambda\vec{v}$$

- $\vec{v}$ (the eigenvector) is non-zero, though λ (the eigenvalue) can be any complex number in principle.
- Computing eigenvalues: $\det(A-\lambda I)=0$

Eigenvectors and Eigenvalues

- Example:

$$A=\begin{pmatrix}-1 & 3\\ 2 & 0\end{pmatrix} \qquad A-\lambda I=\begin{pmatrix}-1-\lambda & 3\\ 2 & -\lambda\end{pmatrix}$$

- $\det(A-\lambda I) = (-1-\lambda)*(-\lambda)-3*2=0$
- Then: $\lambda+\lambda^2-6=0$; $\lambda_1=2$; $\lambda_2=-3$
- For $\lambda_1=2$:

$$\begin{pmatrix}-3 & 3\\ 2 & -2\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=0$$

- Solutions: $x_1=x_2$

Matrix Decomposition

- If Σ is a square matrix, it can be decomposed into $U\Lambda U^{-1}$, where
  U = matrix of eigenvectors
  Λ = diagonal matrix of eigenvalues

$\Sigma U = U\Lambda$
$U^{-1}\Sigma U = \Lambda$
$\Sigma = U\Lambda U^{-1}$

Example

$$S=\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}, \lambda_1=1, \lambda_2=3$$

$$U=\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}$$

implicit direction of a matrix

eigenvector: implicit direction of a matrix.

$\det(A-\lambda I)=0$

want it to be a square matrix

Derivation

$A=\begin{pmatrix}-1 & 3\\ 2 & 0\end{pmatrix}$

$A\vec{v}=\lambda\vec{v}$

$\Rightarrow \det(A-\lambda I)=0$

$\begin{pmatrix}-1-\lambda & 3\\ 2 & -\lambda\end{pmatrix}=0$

$(-1-\lambda)(-\lambda)-3\cdot 2=0$

$\lambda^2+\lambda-6=0$

$(\lambda+3)(\lambda-2)=0$

$\lambda=-3$ or 2 if $\lambda=2$

$\begin{pmatrix}-1 & 3\\ 2 & 0\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}2x_1\\ 2x_2\end{pmatrix}$

$\begin{pmatrix}-x_1+3x_2\\ 2x_1\end{pmatrix}=\begin{pmatrix}2x_1\\ 2x_2\end{pmatrix}$

$\Rightarrow x_1=x_2$

square matrix can be decomposed.

Derivation

$S\vec{v}=\lambda\vec{v} \Rightarrow \det(S-\lambda I)=0$

$\begin{pmatrix}2-\lambda & 1\\ 1 & 2-\lambda\end{pmatrix}=0$

$\lambda^2-4\lambda+4-1=0$

$(\lambda-3)(\lambda-1)=0$

$\lambda_1=1, \lambda_2=3$ if $\lambda=1$

$\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}x_1\\ x_2\end{pmatrix}$

$A=\begin{bmatrix}a & b\\ c & d\end{bmatrix}$

$$U = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \quad \lambda = 3$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

$UU^{-1} = I$

$$U^{-1} = \begin{pmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

$\lambda = 1 \quad x = 3$

$$S = U\Lambda U^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

new space with igenvectors

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\begin{pmatrix} 2x_1 + x_2 \\ x_1 + 2x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$\Rightarrow x_1 = -x_2$

if $\lambda = 3$

$$\begin{pmatrix} 2x_1 + x_2 \\ x_1 + 2x_2 \end{pmatrix} = \begin{pmatrix} 3x_1 \\ 3x_2 \end{pmatrix}$$

$\Rightarrow x_1 = x_2$

→ convert into a newspace of eigen vectors

When the matrix is not square.

SVD: Singular Value Decomposition

- $A = U\Sigma V^T$
  - U is the matrix of orthogonal eigenvectors of AA^T $m \times m$
  - V is the matrix of orthogonal eigenvectors of A^TA $n \times n$
  - The components of Σ are the eigenvalues of A^TA
- This decomposition exists for all matrices, dense or sparse
- If A has 5 columns and 3 rows, then U will be 5x5 and V will be 3x3
- In Matlab, use *[U,S,V] = svd (A)*

3x5 ? 5x5 ? 3x3

Documnet, term matrices are usually not square.

All matrices have this decomposition

SVD

Example

D1: T6, T9
D2: T1, T2
D3: T2, T5, T8
D4: T1, T4, T6, T8, T9
D5: T1, T7
D6: T3, T7
D7: T1, T3

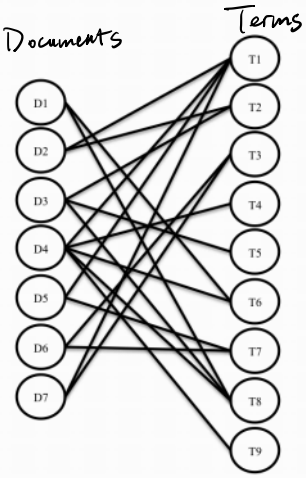
From the Forum

The idea is to map the Term * Document matrix to a lower dimensional space in which both terms and documents can be represented. E.g., if you have a mixture of documents on two topics: politics and medicine, each document or term will be represented as two numbers, one for each topic. The document "elections coming in the Fall" may be represented as (1,0) whereas "treatments for colds" may end up being represented as (0,1). The term "elections" may also end up being (1,0) and "treatment" as (0,1). This is a complete oversimplification.

There is a lecture on this topic in our class. Before we get to it, here are some useful external pointers.

Example

D1: T6, T9
D2: T1, T2
D3: T2, T5, T8
D4: T1, T4, T6, T8, T9
D5: T1, T7
D6: T3, T7
D7: T1, T3



http://glowingpython.blogspot.com/2011/06/svd-decomposition-with-numpy.html

http://radialmind.blogspot.com/2009/11/svd-in-python.html

http://bigdata-madesimple.com/decoding-dimensionality-reduction-pca-and-svd/

http://blog.josephwilk.net/projects/latent-semantic-analysis-in-python.html

http://bl.ocks.org/ktaneishi/9499896#pca.js

Document-Term Matrix

raw 9×7

normalized

Divide by the length of the column.

have to normalize the matrix before decomposition

$$A = \begin{bmatrix} 0&1&0&1&1&0&1 \\ 0&1&1&0&0&0&0 \\ 0&0&0&0&0&1&1 \\ 0&0&0&1&0&0&0 \\ 0&1&1&0&0&0&0 \\ 1&0&0&1&0&0&0 \\ 0&0&0&0&1&1&0 \\ 0&0&1&1&0&0&0 \\ 1&0&0&1&0&0&0 \end{bmatrix} \qquad A^{(n)} = \begin{bmatrix} 0&0.58&0&0.45&0.71&0&0.71 \\ 0&0.58&0.58&0&0&0&0 \\ 0&0&0&0&0&0.71&0.71 \\ 0&0&0&0.45&0&0&0 \\ 0&0.58&0.58&0&0&0&0 \\ 0.71&0&0&0.45&0&0&0 \\ 0&0&0&0&0.71&0.71&0 \\ 0&0&0.58&0.45&0&0&0 \\ 0.71&0&0&0.45&0&0&0 \end{bmatrix}$$

÷ $\sqrt{2}$

Decomposition

can use maybe Matlab to decomposite.

u =

```
-0.6976   -0.0945    0.0174   -0.6950    0.0000    0.0153    0.1442   -0.0000         0
-0.2622    0.2946    0.4693    0.1968   -0.0000   -0.2467   -0.1571   -0.6356    0.3098
-0.3519   -0.4495   -0.1026    0.4014    0.7071   -0.0065   -0.0493   -0.0000    0.0000
-0.1127    0.1416   -0.1478   -0.0734    0.0000    0.4842   -0.8400    0.0000   -0.0000
-0.2622    0.2946    0.4693    0.1968    0.0000   -0.2467   -0.1571    0.6356   -0.3098
-0.1883    0.3756   -0.5035    0.1273   -0.0000   -0.2293    0.0339   -0.3098   -0.6356
-0.3519   -0.4495   -0.1026    0.4014   -0.7071   -0.0065   -0.0493    0.0000   -0.0000
-0.2112    0.3334    0.0962    0.2819   -0.0000    0.7338    0.4659   -0.0000    0.0000
-0.1883    0.3756   -0.5035    0.1273   -0.0000   -0.2293    0.0339    0.3098    0.6356
```

9×9

v =

```
-0.1687    0.4192   -0.5986    0.2261         0   -0.5720    0.2433
-0.4472    0.2255    0.4641   -0.2187    0.0000   -0.4871   -0.4987
-0.2692    0.4206    0.5024    0.4900   -0.0000    0.2450    0.4451
-0.3970    0.4003   -0.3923   -0.1305         0    0.6124   -0.3690
-0.4702   -0.3037   -0.0507   -0.2607   -0.7071    0.0110    0.3407
-0.3153   -0.5018   -0.1220    0.7128   -0.0000   -0.0162   -0.3544
-0.4702   -0.3037   -0.0507   -0.2607    0.7071    0.0110    0.3407
```

7×7

15:20

v is 7×7

Decomposition

Spread on the v1 axis

singular value

9×7

eigen values of AA^T

s =

1.5849	0	0	0	0	0	0
0	1.2721	0	0	0	0	0
0	0	1.1946	0	0	0	0
0	0	0	0.7996	0	0	0
0	0	0	0	0.7100	0	0
0	0	0	0	0	0.5692	0
0	0	0	0	0	0	0.1977
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Related to Dimensionality Reduction

- Singular value decomposition:
  - $A=U\Sigma V^T$
- Dimensionality reduction
  - $A^*=U\Sigma^*V^T$
  - Where Σ^* keeps only the largest eigenvalues

A^* lower dimension version of A.

without losing much of the information

Rank-4 Approximation

4 largest.

s4 =

1.5849	0	0	0	0	0	0
0	1.2721	0	0	0	0	0
0	0	1.1946	0	0	0	0
0	0	0	0.7996	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Rank-4 Approximation

```
u*s4*v'
   -0.0019    0.5985   -0.0148    0.4552    0.7002    0.0102    0.7002
   -0.0728    0.4961    0.6282    0.0745    0.0121   -0.0133    0.0121
    0.0003   -0.0067    0.0052   -0.0013    0.3584    0.7065    0.3584
    0.1980    0.0514    0.0064    0.2199    0.0535   -0.0544    0.0535
   -0.0728    0.4961    0.6282    0.0745    0.0121   -0.0133    0.0121
    0.6337   -0.0602    0.0290    0.5324   -0.0008    0.0003   -0.0008
    0.0003   -0.0067    0.0052   -0.0013    0.3584    0.7065    0.3584
    0.2165    0.2494    0.4367    0.2282   -0.0360    0.0394   -0.0360
    0.6337   -0.0602    0.0290    0.5324   -0.0008    0.0003   -0.0008
```


Rank-4 Approximation

```
u*s4

   -1.1056   -0.1203    0.0207   -0.5558         0         0         0
   -0.4155    0.3748    0.5606    0.1573         0         0         0
   -0.5576   -0.5719   -0.1226    0.3210         0         0         0
   -0.1786    0.1801   -0.1765   -0.0587         0         0         0
   -0.4155    0.3748    0.5606    0.1573         0         0         0
   -0.2984    0.4778   -0.6015    0.1018         0         0         0
   -0.5576   -0.5719   -0.1226    0.3210         0         0         0
   -0.3348    0.4241    0.1149    0.2255         0         0         0
   -0.2984    0.4778   -0.6015    0.1018         0         0         0
```


Rank-4 Approximation

```
s4*v'

   -0.2674   -0.7087   -0.4266   -0.6292   -0.7451   -0.4996   -0.7451
    0.5333    0.2869    0.5351    0.5092   -0.3863   -0.6384   -0.3863
   -0.7150    0.5544    0.6001   -0.4686   -0.0605   -0.1457   -0.0605
    0.1808   -0.1749    0.3918   -0.1043   -0.2085    0.5700   -0.2085
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
```

Rank-2 Approximaton

reserve two largest values.

```
s2 =
   1.5849         0         0         0         0         0         0
        0    1.2721         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
        0         0         0         0         0         0         0
```


Rank-2 Approximation

```
u*s2*v'
    0.1361    0.4673    0.2470    0.3908    0.5563    0.4089    0.5563
    0.2272    0.2703    0.2695    0.3150    0.0815   -0.0571    0.0815
   -0.1457    0.1204   -0.0904   -0.0075    0.4358    0.4628    0.4358
    0.1057    0.1205    0.1239    0.1430    0.0293   -0.0341    0.0293
    0.2272    0.2703    0.2695    0.3150    0.0815   -0.0571    0.0815
    0.2507    0.2412    0.2813    0.3097   -0.0048   -0.1457   -0.0048
   -0.1457    0.1204   -0.0904   -0.0075    0.4358    0.4628    0.4358
    0.2343    0.2454    0.2685    0.3027    0.0286   -0.1073    0.0286
    0.2507    0.2412    0.2813    0.3097   -0.0048   -0.1457   -0.0048
```


Rank-2 Approximation

```
u*s2 - word vector representation in concept space
   -1.1056   -0.1203         0         0         0         0         0
   -0.4155    0.3748         0         0         0         0         0
   -0.5576   -0.5719         0         0         0         0         0
   -0.1786    0.1801         0         0         0         0         0
   -0.4155    0.3748         0         0         0         0         0
   -0.2984    0.4778         0         0         0         0         0
   -0.5576   -0.5719         0         0         0         0         0
   -0.3348    0.4241         0         0         0         0         0
   -0.2984    0.4778         0         0         0         0         0
```

T1 T2 ... Tm

concept 1 concept 2

U x S

$U = AA^T$

r x c c x r => r x r

Word vector representation in concept space

Doc 1 Doc 2 Doc 3 ... Doc N

T1 T2 ... Tm

Rank-2 Approximation

s2*v' - new concept representation of the documents

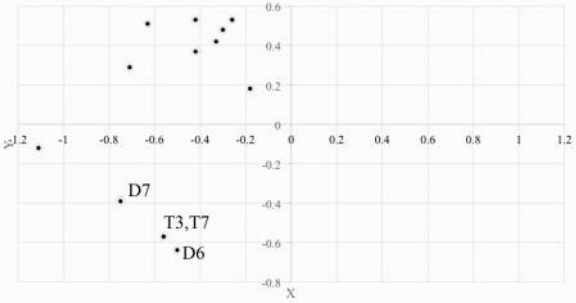
D1	D2	D3	D4	D5	D6	D7
-0.2674	-0.7087	-0.4266	-0.6292	-0.7451	-0.4996	-0.7451
0.5333	0.2869	0.5351	0.5092	-0.3863	-0.6384	-0.3863
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

$S \times V^T$ $V = A^T A$

03.04 Dimensionality Reduction NATURAL LANGUAGE PROCESSING

D1	D2	D3	D4	D5	D6	D7
-0.26	-0.71	-0.42	-0.63	-0.75	-0.5	-0.75
0.53	0.29	0.53	0.51	-0.38	-0.64	-0.39

T1	T2	T3	T4	T5	T6	T7	T8	T9
-1.11	-0.41	-0.56	-0.18	-0.42	-0.3	-0.56	-0.33	-0.3
-0.12	0.37	-0.57	0.18	0.37	0.48	-0.57	0.42	0.48



Two clusters of documents and terms.

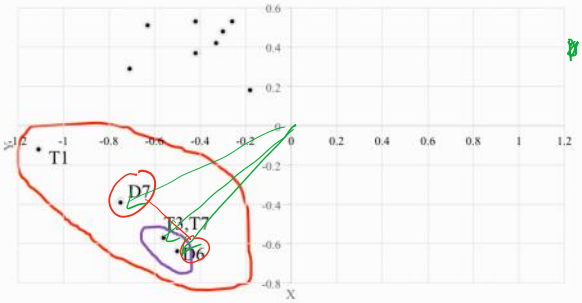
Kill two birds with one stone

* 用 SVD 可以对比出 Document 和 Term 的 similarity

03.04 Dimensionality Reduction NATURAL LANGUAGE PROCESSING

D1	D2	D3	D4	D5	D6	D7
-0.26	-0.71	-0.42	-0.63	-0.75	-0.5	-0.75
0.53	0.29	0.53	0.51	-0.38	-0.64	-0.39

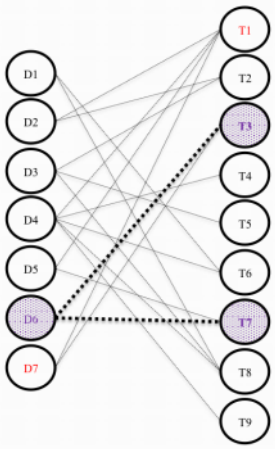
T1	T2	T3	T4	T5	T6	T7	T8	T9
-1.11	-0.41	-0.56	-0.18	-0.42	-0.3	-0.56	-0.33	-0.3
-0.12	0.37	-0.57	0.18	0.37	0.48	-0.57	0.42	0.48



* Similarity calculated from the angles between vectors

Example

D1: T6, T9
D2: T1, T2
D3: T2, T5, T8
D4: T1, T4, T6, T8, T9
D5: T1, T7
D6: T3, T7
D7: T1, T3



D2 is the next similar followed by D4

Documents to Concepts and Terms to Concepts

```
>> A(:,1)'*u*s
-0.4238    0.6784   -0.8541    0.1446   -0.0000   -0.1853    0.0095

>> A(:,1)'*u*s4
-0.4238    0.6784   -0.8541    0.1446         0         0         0

>> A(:,1)'*u*s2
-0.4238    0.6784         0         0         0         0         0

>> A(:,2)'*u*s2
-1.1233    0.3650         0         0         0         0         0

>> A(:,3)'*u*s2
-0.6762    0.6807         0         0         0         0         0
```


Documents to Concepts and Terms to Concepts

```
>> A(:,4)'*u*s2
-0.9972    0.6478         0         0         0         0         0

>> A(:,5)'*u*s2
-1.1809   -0.4914         0         0         0         0         0

>> A(:,6)'*u*s2
-0.7918   -0.8121         0         0         0         0         0

>> A(:,7)'*u*s2
-1.1809   -0.4914         0         0         0         0         0
```

Cont'd

```
>> (s2*v'*A(1,:)')'
-1.7523   -0.1530         0         0         0         0         0         0         0

>> (s2*v'*A(2,:)')'
-0.6585    0.4768         0         0         0         0         0         0         0

>> (s2*v'*A(3,:)')'
-0.8838   -0.7275         0         0         0         0         0         0         0

>> (s2*v'*A(4,:)')'
-0.2831    0.2291         0         0         0         0         0         0         0

>> (s2*v'*A(5,:)')'
-0.6585    0.4768         0         0         0         0         0         0         0
```

03.04 Dimensionality Reduction NATURAL LANGUAGE PROCESSING

Cont'd

```
>> (s2*v'*A(6,:)')'

   -0.4730    0.6078         0         0         0         0         0         0         0

>> (s2*v'*A(7,:)')'

   -0.8838   -0.7275         0         0         0         0         0         0         0

>> (s2*v'*A(8,:)')'

   -0.5306    0.5395         0         0         0         0         0         0         0

>> (s2*v'*A(9,:)')'

   -0.4730    0.6078         0         0         0         0         0         0         0
```

03.04 Dimensionality Reduction NATURAL LANGUAGE PROCESSING

Properties

Document similarity matrix

AA^T

A is a document to term matrix. What is A*A'?

A*A'

9×9

```
   1.5471    0.3364    0.5041    0.2025    0.3364    0.2025    0.5041    0.2025    0.2025
   0.3364    0.6728         0         0    0.6728         0         0    0.3364         0
   0.5041         0    1.0082         0         0         0    0.5041         0         0
   0.2025         0         0    0.2025         0    0.2025         0    0.2025    0.2025
   0.3364    0.6728         0         0    0.6728         0         0    0.3364         0
   0.2025         0         0    0.2025         0    0.7066         0    0.2025    0.7066
   0.5041         0    0.5041         0         0         0    1.0082         0         0
   0.2025    0.3364         0    0.2025    0.3364    0.2025         0    0.5389    0.2025
   0.2025         0         0    0.2025         0    0.7066         0    0.2025    0.7066
```

Properties

What about A'*A?

A'*A

7*7

1.0082	0	0	0.6390	0	0	0
0	1.0092	0.6728	0.2610	0.4118	0	0.4118
0	0.6728	1.0092	0.2610	0	0	0
0.6390	0.2610	0.2610	1.0125	0.3195	0	0.3195
0	0.4118	0	0.3195	1.0082	0.5041	0.5041
0	0	0	0	0.5041	1.0082	0.5041
0	0.4118	0	0.3195	0.5041	0.5041	1.0082

Latent Semantic Indexing (LSI)

Latent Semantic Indexing

or A: Analysis

- Dimensionality reduction = identification of hidden (latent) concepts
- Query matching in latent space

To identify hidden/latent concept in textural spaces.

External Pointers

in some article called LSA.

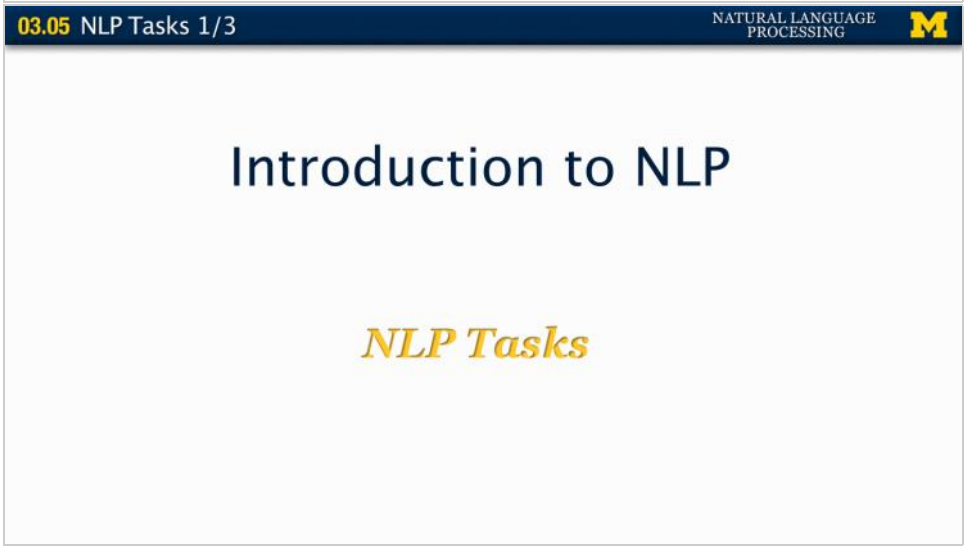
- http://lsa.colorado.edu
- http://www.cs.utk.edu/~lsi

changed.

The active research on this

Tasks that form the core of NLP research

3.5



Part of Speech Tagging

The swimmer is getting ready to run in the final race.

POS tagging

03.05 NLP Tasks 1/3 NATURAL LANGUAGE PROCESSING

Part of Speech Tagging

The swimmer is getting ready to **run** in the final race.

- Run – verb or noun?
- Final – noun or adjective?
- Race – verb or noun?

03.05 NLP Tasks 1/3 NATURAL LANGUAGE PROCESSING

Part of Speech Tagging

The candidate is preparing for his **run** for the presidency.
The swimmer is getting ready to **run** in the final race.

Parsing

subject verb

- Myriam slept.
- Myriam wrote a novel.
- Myriam gave Sally flowers.
- Myriam ate pizza with olives.
- Myriam ate pizza with Sally.
- Myriam ate pizza with a fork.
- Myriam ate pizza with remorse.

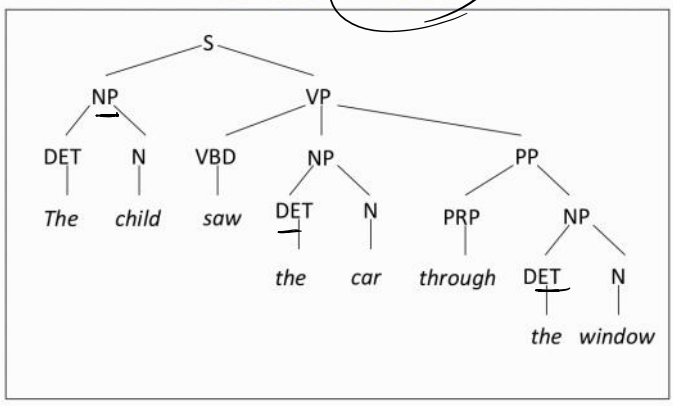
Phrase-Structure Grammar

Constituent structure

```
S   → NP  VP
NP  → DET N
NP  → NP  PP
VP  → VBD
VP  → VBD NP
VP  → VBD NP NP
VP  → VP  PP
PP  → PRP NP
```

```
DET → the
DET → that
DET → a
N   → child
N   → window
N   → car
VBD → found
VBD → ate
VBD → saw
PRP → in
PRP → of
PRP → through
```


Parse Trees



Stanford Parser

Parser Output

tree

```
(ROOT
  (S
    (S
      (NP
        (NP (NN Housing) (NNS starts))
        (, ,)
        (NP
          (NP (DT the) (NN number))
          (PP (IN of)
            (NP
              (NP (JJ new) (NNS homes))
              (VP (VBG being)
                (VP (VBN built))))))
        (, ,))
```

```
(VP (VBD rose)
      (NP (CD 7.2) (NN %))
      (PP (IN in)
        (NP (NNP March)))
      (PP (TO to)
        (NP
          (NP (DT an) (JJ annual) (NN rate))
          (PP (IN of)
            (NP (CD 549,000) (NNS units)))))
      (, ,)
      (ADVP (RB up)
        (PP (IN from)
          (NP
            (NP (DT a) (VBN revised) (CD 512,000))
            (PP (IN in)
              (NP (NNP February)))))))
  (, ,)
  (NP (DT the) (NNP Commerce) (NNP Department))
  (VP (VBD said))
  (. .)))
```


This Problem is Pretty // Easy

- Commercial for a phone company
- Garden path sentences
  - Don't bother coming
  - Don't bother coming early
  - Take the turkey out of the oven at five
  - Take the turkey out of the over at five to four
  - I got canned — something you don't want to hear over the phone
  - I got canned peaches for dinner
  - All Americans need to buy a house
  - All Americans need to buy a house is a lot of money
- Can you think of more such examples?

The problem is pretty/

pretty/easy

pretty

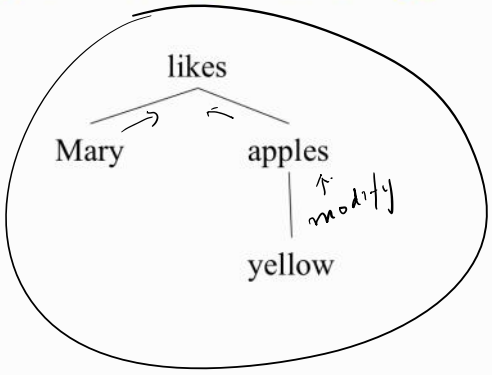
出经典的 Garden Path Sentence

多最后一部分意思会不同

Solution

- This problem is pretty // easy
  - http://www.nacio.cs.cmu.edu/problems2007/N2007-HS.pdf
- Criteria
  - The part before // should be a complete sentence
  - The full sentence has a different meaning than the part before //
  - The part before // should not already be ambiguous

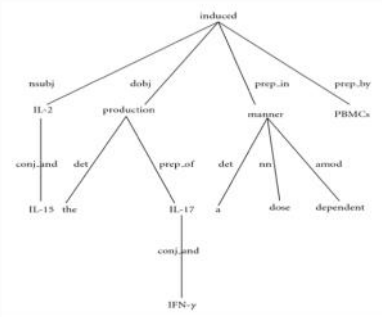
Dependency Parsing



Dependency Parsing

Dependency Parsing

IL-2 and IL-15 induced the production of IL-17 and IFN-γ by PBMCs in a dose dependent manner.



biological natural language processing

Determine whether there's interaction between two proteins in the sentence

Parser Output

nn(starts-2, Housing-1)
nsubj(rose-12, starts-2)
det(number-5, the-4)
appos(starts-2, number-5)
prep(number-5, of-6)
amod(homes-8, new-7)
pobj(of-6, homes-8)
auxpass(built-10, being-9)
partmod(homes-8, built-10)
ccomp(said-36, rose-12)
num(%-14, 7.2-13)
dobj(rose-12, %-14)
prep(rose-12, in-15)
pobj(in-15, March-16)
prep(rose-12, to-17)
det(rate-20, an-18)
amod(rate-20, annual-19)
pobj(to-17, rate-20)
prep(rate-20, of-21)
num(units-23, 549,000-22)
pobj(of-21, units-23)
advmod(rose-12, up-25)
dep(up-25, from-26)
det(512,000-29, a-27)
amod(512,000-29, revised-28)
pobj(from-26, 512,000-29)
prep(512,000-29, in-30)
pobj(in-30, February-31)
det(Department-35, the-33)
nn(Department-35, Commerce-34)
nsubj(said-36, Department-35)

03.05 NLP Tasks 1/3 NATURAL LANGUAGE PROCESSING

NLP

3.6

Information extraction
get important information from a sentence.

NLP

Introduction to NLP

NLP Tasks (cont'd)

03.06 NLP Tasks 2/3 NATURAL LANGUAGE PROCESSING

Information Extraction

Finance Example

name of comp. v. name

- RESEARCH ALERT–Wells Fargo cuts PPD Inc to market perform
- China Southern Air Upgraded To Overweight From Neutral–HSBC
- CITIGROUP RAISES INGERSOLL RAND <IR.N> TO HOLD FROM SELL
- TCF Financial Corp Raised To Overweight From Neutral By JPMorgan
- BAIRD CUTS KIOR INC <KIOR.O> TO UNDERPERFORM RATING
- BRIEF–RESEARCH ALERT–Global Equities Research cuts LinkedIn to equal weight

change from one rating to another.

no from.

Understand different entities on unstructured sentences.

03.06 NLP Tasks 2/3 NATURAL LANGUAGE PROCESSING

Information Extraction

DATE/ TIME	TICKER	COMPANY	SOURCE	OLD	NEW	CHANGE
		PPD Inc	Wells Fargo		market perform	↓
		China Southern Air	HSBC	Neutral	Overweight	↑
	IR.N	INGERSOLL RAND	CITIGROUP	SELL	HOLD	↑
		TCF Financial Corp	JPMorgan	Neutral	Overweight	↑
	KIOR.O	KIOR INC	BAIRD		UNDERPERFORM	↓
		LinkedIn	Global Equities Research		equal weight	↓

False Positives

- Examples of false positives
  - BARCLAYS CUTS FLAGSTONE REINSURANCE <FSR.N> PRICE TARGET TO $9 FROM $11
  - Rimage To Buy Qumu For $52M;; Raises Dividend;; Lowers EPS View
  - S&P rates Ameren Illinois commercial paper 'A-3'
  - BRIEF-Moody's changes otlk for Stirling Water Seafield Finance to positive
  - BRIEF-RESEARCH ALERT-HSBC cuts price targets on European telcos
  - Stifel cuts Philip Morris price target
  - Media General shares plummet on Moody's downgrade
- Explain why these are false positives.

Answers to the Quiz

- BARCLAYS CUTS FLAGSTONE REINSURANCE <FSR.N> PRICE TARGET TO $9 FROM $11
  - Didn't cut the ratings but the price target
- Rimage To Buy Qumu For $52M;; Raises Dividend;; Lowers EPS View
  - Lowers eps view
- S&P rates Ameren Illinois commercial paper 'A-3'
  - Debt rating
- BRIEF-Moody's changes otlk for Stirling Water Seafield Finance to positive
  - Changes outlook
- BRIEF-RESEARCH ALERT-HSBC cuts price targets on European telcos
  - Not a company but a group of companies
- Stifel cuts Philip Morris price target
  - Price target, not rating
- Media General shares plummet on Moody's downgrade
  - Event in the past

Semantics

- First order logic
- Inference
- Semantic analysis

$\forall x,y$: *Mother (x,y)* $\Rightarrow$ *Parent (x,y)*

x is the mother of y $\Rightarrow$ x is the parent of y.

NACLO Problem

- "Bertrand and Russell", 2014 problem by Ben King
  - http://www.naclo.cs.cmu.edu/problems2014/N2014-H.pdf

NACLO Solution

- Bertrand and Russell
  - http://www.naclo.cs.cmu.edu/problems2014/N2014-HS.pdf

Reading Comprehension

build a system that read paragraph and answer questions

Mars Polar Lander - Where Are You?

(January 18, 2000) After more than a month of searching for a signal from NASA's Mars Polar Lander, mission controllers have lost hope of finding it. The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life ever existed on Mars. Polar Lander was to have touched down December 3 for a 90-day mission. It was to land near Mars' south pole. The lander was last heard from minutes before beginning its descent. *The last effort to communicate with the three-legged lander ended with frustration at 8 a.m Monday.* "We didn't see anything," said Richard Cook, the spacecraft's project manager at NASA's Jet Propulsion Laboratory. The failed mission to the Red Planet cost the American government more than $200 million dollars. Now, space agency scientists and engineers will try to find out what could have gone wrong. They do not want to make the same mistakes in the next mission.
(sources: CBC "For Kids" web page, Associated Press, CBC News Online, CBC Radio news, NASA)

1. When did the mission controllers lose hope of communicating with the lander?
   Answer: *8AM, Monday Jan. 17, 2000*
2. Who is the Polar Lander's project manager?
3. Where on Mars was the spacecraft supposed to touch down?
4. What did the Mars Global Surveyor do?
5. What was the mission of the Mars Polar Lander?

Pranav Anand, Eric Breck, Brianne Brown, Marc Light, Gideon Mann, Ellen Riloff, Mats Rooth, Michael Thelen. 2000. Fun with Reading Comprehension

Text Understanding

There are four bungalows in our cul-de-sac. They are made from these materials: straw, wood, brick and glass.

Mrs. Scott's bungalow is somewhere to the left of the wooden one and the third one along is brick. Mrs. Umbrella owns a straw bungalow and Mr. Tinsley does not live at either end, but lives somewhere to the right of the glass bungalow. Mr. Wilshaw lives in the fourth bungalow, whilst the first bungalow is not made from straw.

Who lives where, and what is their bungalow made from?

http://www.brainbashers.com/showpuzzles.asp?puzzle=ZSOP

Word Sense Disambiguation

- "The thieves took off with 100 gold **bars**".
  - Did they steal 100 drinking establishments?
  - Or 100 measures of a song?

polysemy.

Important for machine translation.

Determine which sense a word is from the dictionary, given the context.

Word Sense Disambiguation

Bar=Noun
- S: (n) barroom, bar, saloon, ginmill, taproom (a room or establishment where alcoholic drinks are served over a counter) "he drowned his sorrows in whiskey at the bar"
- S: (n) bar (a counter where you can obtain food or drink) "he bought a hot dog and a coke at the bar"
- S: (n) bar (a rigid piece of metal or wood; usually used as a fastening or obstruction or weapon) "there were bars in the windows to prevent escape"
- S: (n) measure, bar (musical notation for a repeating pattern of musical beats) "the orchestra omitted the last twelve bars of the song"
- S: (n) bar (an obstruction (usually metal) placed at the top of a goal) "it was an excellent kick but the ball hit the bar"
- S: (n) prevention, bar (the act of preventing) "there was no bar against leaving"; "money was allocated to study the cause and prevention of influenza"
- S: (n) bar ((meteorology) a unit of pressure equal to a million dynes per square centimeter) "unfortunately some writers have used bar for one dyne per square centimeter"
- S: (n) bar (a submerged (or partly submerged) ridge in a river or along a shore) "the boat ran aground on a submerged bar in the river"
- S: (n) legal profession, bar, legal community (the body of individuals qualified to practice law in a particular jurisdiction) "he was admitted to the bar in New Jersey"
- S: (n) stripe, streak, bar (a narrow marking of a different color or texture from the background) "a green toad with small black stripes or bars"; "may the Stars and Stripes forever wave"
- S: (n) cake, bar (a block of solid substance (such as soap or wax)) "a bar of chocolate"
- S: (n) Browning automatic rifle, BAR (a portable .30 caliber automatic rifle operated by gas pressure and fed by cartridges from a magazine; used by United States troops in World War I and in World War II and in the Korean War)
- S: (n) bar (a horizontal rod that serves as a support for gymnasts as they perform exercises)
- S: (n) bar (a heating element in an electric fire) "an electric fire with three bars"
- S: (n) bar ((law) a railing that encloses the part of the courtroom where the judges and lawyers sit and the case is tried) "spectators were not allowed past the bar"

Bar=Verb
- S: (v) bar, debar, exclude (prevent from entering; keep out) "He was barred from membership in the club"
- S: (v) barricade, block, blockade, stop, block off, block up, bar (render unsuitable for passage) "block the way"; "barricade the streets"; "stop the busy road"
- S: (v) banish, relegate, bar (expel, as if by official decree) "he was banished from his own country"
- S: (v) bar (secure with, or as if with, bars) "He barred the door"

WSD is Important for Translation

- Paul plays soccer
  - Paul joue au football
- Paul plays the guitar
  - Paul joue de la guitare
- "wall" in German
  - die Chinesische Mauer (The Great Wall of China)
  - (otherwise Wand)
- "wall" in Spanish
  - pared, muro, muralla

example "play" can transform differently

Named Entity Recognition

Wolff, currently a journalist in Argentina, played with Del Bosque in the final years of the seventies in Real Madrid.

```
Wolff B-PERSON
,  O
currently  O
a  O
journalist  O
in  O
Argentina  B-LOCATION
,  O
played  O
with  O
Del  B-PER
Bosque  I-PER
in  O
the  O
final  O
years  O
of  O
the  O
seventies  O
in  O
Real  B-ORG
Madrid  I-ORG
.  O
```

output of named entity recognition.

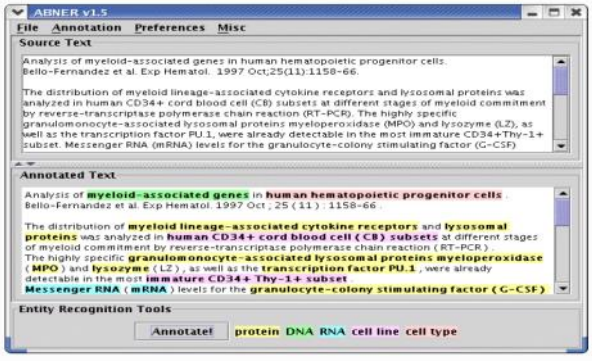
B. begin.
I end.

- http://cogcomp.cs.illinois.edu/page/demo_view/NER
- http://nlp.stanford.edu:8080/ner/

To be covered later in this course.

Named Entity Recognition

use this in a different domain



http://pages.cs.wisc.edu/~bsettles/abner

Semantic Role Labeling

- [A0 He] [AM-MOD would] [AM-NEG n't] [V **accept**] [A1 anything of value] from [A2 those he was writing about] .
- **V:** verb
  A0: acceptor
  A1: thing accepted
  A2: accepted-from
  A3: attribute
  AM-MOD: modal
  AM-NEG: negation
- http://cogcomp.cs.illinois.edu/page/demo_view/SRL

accept
A0: acceptor
A1: the accepted

Coreference Resolution

Identify phrases that refer to the same entity.

- Barack Obama visited China. The US president met with his Chinese counterpart.
- Cynthia went to see her aunt at the hospital. She was scheduled for surgery on Monday.
- Because he was sick, Michael stayed home on Friday.

Anaphoric: the name of entity appears first

aphora: the entity name appear after it has been referred to

Ellipsis, Parallelism, and Underspecification

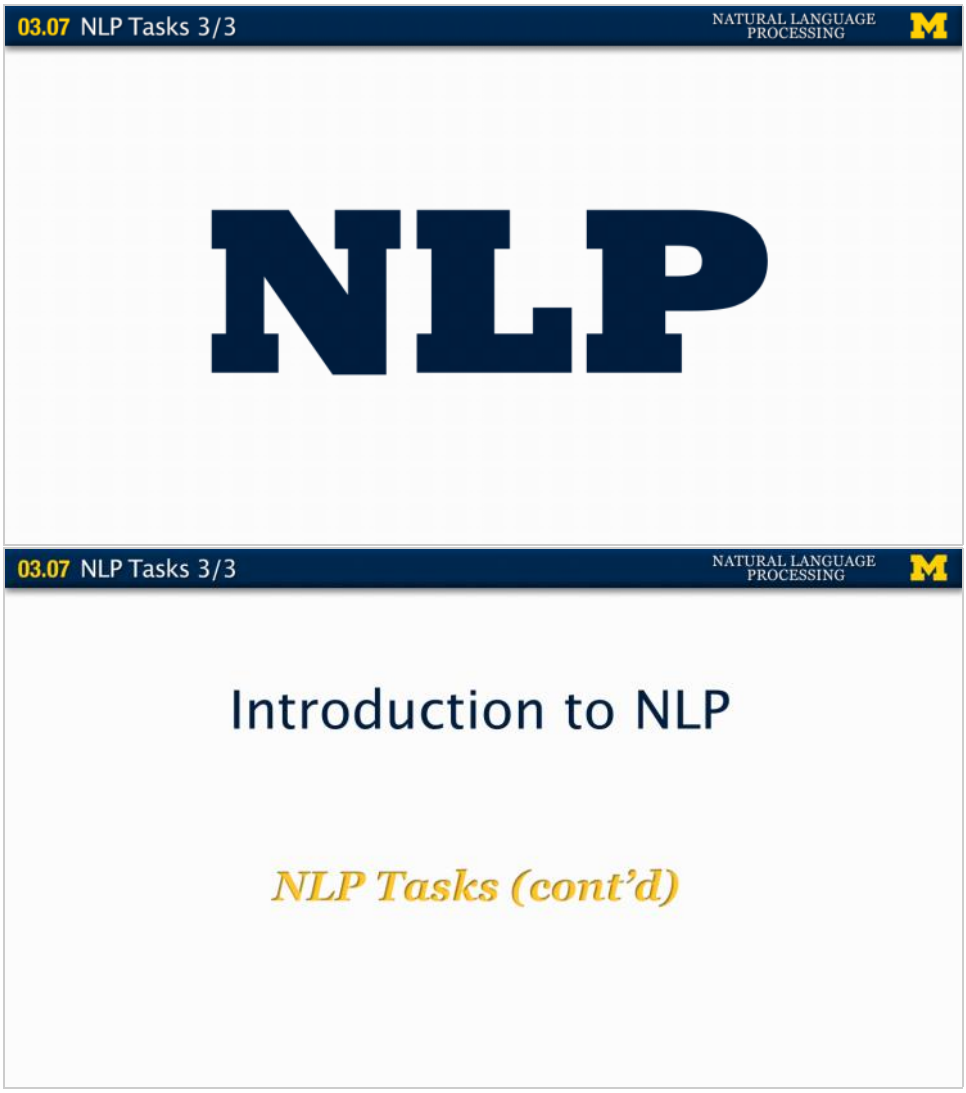
- Chen speaks Chinese. I don't. [speak Chinese]
- Santa gave Mary a book and Johnny a toy.

Santa gave

parallel structure
=> need to infer thing


03.06 NLP Tasks 2/3
NATURAL LANGUAGE PROCESSING
NLP

3.7


03.07 NLP Tasks 3/3
NATURAL LANGUAGE PROCESSING
NLP
03.07 NLP Tasks 3/3
NATURAL LANGUAGE PROCESSING
Introduction to NLP
NLP Tasks (cont'd)

Question Answering

- **"The antagonist of Stevenson's Treasure Island." (Who is Long John Silver?)**
- http://blog.reddit.com/2011/02/ibm-watson-research-team-answers-your.html
- "Watson is powered by 10 racks of IBM Power 750 servers running Linux, and uses 15 terabytes of RAM, 2,880 processor cores and is capable of operating at 80 teraflops. Watson was written in mostly Java but also significant chunks of code are written C++ and Prolog, all components are deployed and integrated using UIMA."

Jeopardy Questions

- From the competition between the IBM Watson system and two human champions (Ken Jennings and Brad Rutter)
- Sample questions:
  - On December 8, 2008 this national newspaper raised its newsstand price by 25 cents to $1 : *USA Today*
  - In 2010 this former first lady published the memoir "Spoken From the Heart" : *Laura Bush*"
  - This person is appointed by a testator to carry out the directions & requests in his will : *Executor*"
  - Familiarity is said to breed this, from the Latin for "Despise" : *Contempt*"
  - As of 2010, Croatia & Macedonia are candidates but this is the only former Yugoslav republic in the EU : *Slovenia*
  - The ancient "Lion of Nimrud" went missing from this city's national museum in 2003 (along with a lot of other stuff) : *Baghdad*
  - It's just a bloody nose! You don't have this hereditary disorder once endemic to European royalty : *Haemophilia*
  - It's Michelangelo's fresco on the wall of the Sistine Chapel, Depicting the saved and the damned : *The Last Judgement*
  - She "Died in the church and was buried along with her name. Nobody came" : *Eleanor Rigby*
  - It's a 4-letter term for a summit; the first 3 letters mean a type of simian : *Apex*
  - A camel is a horse designed by this : *Committee*
- Watson's answers: 66 correct and 9 incorrect (e.g., the one in the category "US Cities" about a city with two airports named after a World War II hero and a World War II battle)
- Watson's two day winning streak was $77,147. Ken Jennings ended with $24,000 and Brad Rutter with $21,600. Watson donated $500,000 to both World Vision and World Community Grid charities from the $1,000,000 prize.
- http://www.quora.com/What-questions-were-asked-in-the-Jeopardy-episode-involving-Watson

Sentiment Analysis

Sentiment Analysis

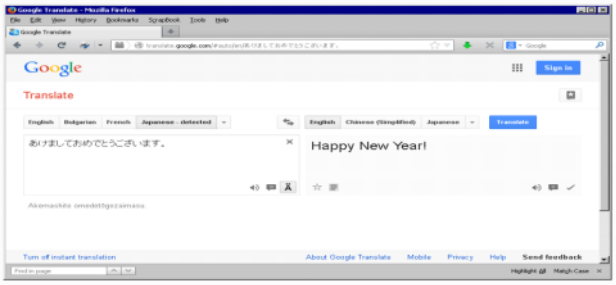
- "I like the camera because I can edit images so easily, exactly as I do my iPad. I have found that its difficult to frame a picture when there isn't a zoom function as with the iPad. With this camera I can adjust my images by cropping as I did with my iPad but better yet, this camera has a built in zoom. A stretch or pinch of the fingers bring in the subject closer or back out again. With this iPhone I can also, as I dido with my iPad, enhance, crop, rotate, red eye reduce, and set a range of tints. I am also quite impressed with the quality of the images. Pretty darn good especially better than I expected for low light situations where I can use the built-in flash! Quite frankly I was quite surprised with these built in features. I also hope too experiment with and learn what HDR photography is. It's built into this iPhone and can be activated by a the touch of an icon. "

- http://www.epinions.com/review/apple_iphone_5c_latest_model_16gb_graphite_unlocked_smartphone/content_640679317124

Machine Translation

One of the hottest topic in NLP

- あけましておめでとうございます。
- Happy New Year!



Machine Translation

Most popular open source system.

- Moses
- www.statmt.org

Elephants are social animals. They live with their families, give hugs and call each other by using their trunks as trumpets. They also might know how to help each other.

In a recent elephant study by researchers from the United States and Thailand, pairs of giant animals learned to work together to get some ears of corn. Other animals, especially some primates, are already known to work together to complete tasks, but now elephants have joined the club. Perhaps the finding is not too surprising: Scientists suspect that elephants, with their big brains and survival savvy, may be among the smartest animals on the planet.

Joshua Plotnik, who worked on the study, told Science News that the animals didn't just learn a trick. Instead, the ways the elephants behaved show that they understand how working together brings benefits to everyone involved. Plotnik is a comparative psychologist now at the University of Cambridge in England. Psychology is the study of behaviors and mental processes, and comparative psychologists study how animals other than humans behave.

problem

Les éléphants sont des animaux sociaux. Ils vivent avec leur famille, faire des câlins et appeler les uns les autres en utilisant leurs troncs trompettes. Ils pourraient également savoir comment aider les uns les autres.

Dans une étude récente d'éléphants par des chercheurs des États-Unis et la Thaïlande, des paires d'animaux géants ont appris à travailler ensemble pour obtenir des épis de maïs. D'autres animaux, en particulier des primates, sont déjà connus pour travailler ensemble pour accomplir des tâches, mais maintenant, les éléphants ont rejoint le club. Peut-être le résultat n'est pas trop surprenant: Les scientifiques soupçonnent que les éléphants, avec leurs gros cerveaux et de bon sens de survie, peut-être parmi les plus intelligents des animaux sur la planète.

Joshua Plotnick, qui a travaillé sur l'étude, dit Nouvelles de la Science que les animaux n'ont pas seulement appris un truc. Au lieu de cela, les moyens les éléphants se comportent montrent qu'ils comprennent comment travailler ensemble apporte des avantages à toutes les personnes impliquées. Plotnik est un psychologue comparative maintenant à l'Université de Cambridge en Angleterre. La psychologie est l'étude des comportements et des processus mentaux, et étude comparative des psychologues comment les animaux autres que les humains se comportent.

study that compare different studies

https://student.societyforscience.org/article/theres-no-i-elephant

Elephants are social animals. They live with their families, give hugs and call each other by using their trunks as trumpets. They also might know how to help each other.

In a recent elephant study by researchers from the United States and Thailand, pairs of giant animals learned to work together to get some ears of corn. Other animals, especially some primates, are already known to work together to complete tasks, but now elephants have joined the club. Perhaps the finding is not too surprising: Scientists suspect that elephants, with their big brains and survival savvy, may be among the smartest animals on the planet.

Joshua Plotnik, who worked on the study, told Science News that the animals didn't just learn a trick. Instead, the ways the elephants behaved show that they understand how working together brings benefits to everyone involved. Plotnik is a comparative psychologist now at the University of Cambridge in England. Psychology is the study of behaviors and mental processes, and comparative psychologists study how animals other than humans behave.

Les éléphants sont des animaux sociaux. Ils **vivent** avec leur famille, **faire** des câlins et **appeler** les uns les autres en utilisant leurs troncs trompettes. Ils pourraient également savoir comment aider les uns les autres.

Dans une étude récente d'éléphants par des chercheurs des États-Unis et la Thaïlande, des paires d'animaux géants ont appris à travailler ensemble pour obtenir des épis de maïs. D'autres animaux, en particulier des primates, sont déjà connus pour travailler ensemble pour accomplir des tâches, mais maintenant, les éléphants ont rejoint le club. Peut-être le résultat n'est pas trop surprenant: Les scientifiques soupçonnent que **les éléphants**, avec leurs gros cerveaux et de bon sens de survie, **peut-être** parmi les plus intelligents des animaux sur la planète.

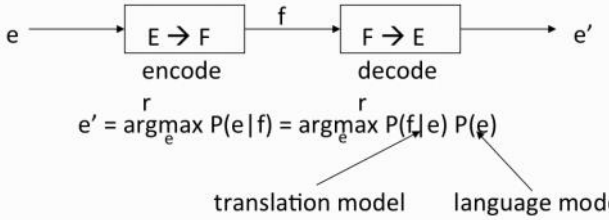
Joshua Plotnick, qui a travaillé sur l'étude, **dit Nouvelles de la Science** que les animaux n'ont pas seulement appris un truc. Au lieu de cela, les moyens les éléphants se comportent montrent qu'ils comprennent comment travailler ensemble apporte des avantages à toutes **les personnes** impliquées. Plotnik est un psychologue **comparative** maintenant à l'Université de Cambridge en Angleterre. La psychologie est l'étude des comportements et des processus mentaux, et **étude comparative des psychologues** comment les animaux autres que les humains se comportent.

Problem

03.07 NLP Tasks 3/3 NATURAL LANGUAGE PROCESSING

Machine Translation

- Noisy channel model ("Chinese Whispers")



grammatical in English
faithful to the original language

03.07 NLP Tasks 3/3 NATURAL LANGUAGE PROCESSING

Machine Translation

- IBM Method

Text Summarization

Provide summary of the original text.

Health Benefits

- Eating a diet rich in vegetables and fruits as part of an overall healthy diet may reduce risk for heart disease, including heart attack and stroke.
- Eating a diet rich in some vegetables and fruits as part of an overall healthy diet may protect against certain types of cancers.
- Diets rich in foods containing fiber, such as some vegetables and fruits, may reduce the risk of heart disease, obesity, and type 2 diabetes.
- Eating vegetables and fruits rich in potassium as part of an overall healthy diet may lower blood pressure, and may also reduce the risk of developing kidney stones and help to decrease bone loss.
- Eating foods such as vegetables that are lower in calories per cup instead of some other higher-calorie food may be useful in helping to lower calorie intake.

Nutrients

- Most vegetables are naturally low in fat and calories. None have cholesterol. (Sauces or seasonings may add fat, calories, or cholesterol.)
- Vegetables are important sources of many nutrients, including potassium, dietary fiber, folate (folic acid), vitamin A, and vitamin C.
- Diets rich in potassium may help to maintain healthy blood pressure. Vegetable sources of potassium include sweet potatoes, white potatoes, white beans, tomato products (paste, sauce, and juice), beet greens, soybeans, lima beans, spinach, lentils, and kidney beans.
- Dietary fiber from vegetables, as part of an overall healthy diet, helps reduce blood cholesterol levels and may lower risk of heart disease. Fiber is important for proper bowel function. It helps reduce constipation and diverticulosis. Fiber-containing foods such as vegetables help provide a feeling of fullness with fewer calories.
- Folate (folic acid) helps the body form red blood cells. Women of childbearing age who may become pregnant should consume adequate folate from foods, and in addition 400 mcg of synthetic folic acid from fortified foods or supplements. This reduces the risk of neural tube defects, spina bifida, and anencephaly during fetal development.

Consensus and differences of the input text.

03.07 NLP Tasks 3/3 NATURAL LANGUAGE PROCESSING

Summary

Eating vegetables is healthy.

not produced by some existing system.

03.07 NLP Tasks 3/3 NATURAL LANGUAGE PROCESSING

☆ UMich. NewsInEssence.

Text to Speech

http://www2.research.att.com/~ttsweb/tts/demo.php

Text to Speech

- www.ivona.com

Entailment and Paraphrasing

Answer questions about paraphrasing.

ID	TEXT	HYPOTHESIS	TASK	VALUE
1586	*The Republic of Yemen is an Arab, Islamic and independent sovereign state whose integrity is inviolable, and no part of which may be ceded.*	*The national language of Yemen is Arabic.*	QA	True
1076	*Most Americans are familiar with the Food Guide Pyramid but a lot of people don't understand how to use it and the government claims that the proof is that two out of three Americans are fat.*	*Two out of three Americans are fat.*	RC	True
1667	*Regan attended a ceremony in Washington to commemorate the landings in Normandy.*	*Washington is located in Normandy.*	IE	False
2016	*Google files for its long awaited IPO.*	*Google goes public.*	IR	True
2097	*The economy created 228,000 new jobs after a disappointing 112,000 in June.*	*The economy created 228,000 jobs after disappointing the 112,000 of June.*	MT	False
893	*The first settlements on the site of Jakarta were established at the mouth of the Ciliwung, perhaps as early as the 5th century AD.*	*The first settlements on the site of Jakarta were established as early as the 5th century AD.*	CD	True
1960	*Bush returned to the White House late Saturday while his running mate was off campaigning in the West.*	*Bush left the White House.*	PP	False
586	*The two suspects belong to the 30th Street gang, which became embroiled in one of the most notorious recent crimes in Mexico: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.*	*Cardinal Juan Jesus Posadas Ocampo died in 1993.*	QA	True

Table 1. Examples of Text-Hypothesis pairs

Ido Dagan, Oren Glickman and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge

Discourse Analysis

- Anaphoric relations:

> *1. Mary helped Peter get out of the car. He thanked her.*
> *2. Mary helped the other passenger out of the car.*
> *The man had asked her for help because of his foot injury.*

Tom appeared on the sidewalk with a bucket of whitewash and a long-handled brush. He surveyed the fence, and all gladness left him and a deep melancholy settled down upon his spirit. (Tom Sawyer)

03.07 NLP Tasks 3/3 NATURAL LANGUAGE PROCESSING

Dialogue Systems

Dialogue System

- I would like to make a reservation at Sorrento.
- For when?
- 8 pm Friday night.
- We only have availability for 7 pm and 10 pm.
- Sorry, these don't work for me.

03.07 NLP Tasks 3/3 NATURAL LANGUAGE PROCESSING

Other Applications

- Spelling Correction
- Web search
- Natural language interfaces to databases
- Parsing job postings
- Summarizing medical records
- Information extraction for databases ex: translate natural language into a SQL query.
- Social network extraction from text
- Alignment of text w/ other signal (time series)
- Essay grading
- Generating weather reports, sports reports, and news stories

NLP

edit distance : apple & pair