

# CSE842: Natural Language Processing

## Lecture 16: Word Similarity

3/16/2009

CSE842, Spring 2009, MSU

1

# Entropy

- Formula

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

- Notation

- $\mathbf{X}$ : A discrete set of symbols
- $X$ : A random variable over  $\mathbf{X}$
- $p(x)$ : The probability mass function of  $X$

3/16/2009

CSE842, Spring 2009, MSU

2

# Explanation of Entropy

- The average uncertainty of a single random variable.
- The average amount of information in a random variable.
- The average number of bits required to transmit the outcome of the random variable.

3/16/2009

CSE842, Spring 2009, MSU

3

# Mutual Information

- Formula

$$I(X; Y) = H(X) - H(X | Y) = H(Y) - H(Y | X)$$

- Interpretation

- The reduction in uncertainty of one random variable due to knowing about another
- The amount of information one random variable contains about another

- Characteristics

- Symmetric
- Non-negative
- Zero iff  $X, Y$  are independent

3/16/2009

CSE842, Spring 2009, MSU

4

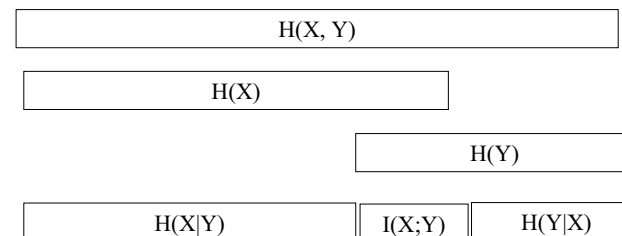
## Mutual Information

$$\begin{aligned}
 I(X;Y) &= H(X) - H(X|Y) \\
 &= H(X) + H(Y) - H(X,Y) \\
 &= \sum_x p(x) \log \frac{1}{p(x)} + \sum_y p(y) \log \frac{1}{p(y)} + \sum_{x,y} p(x,y) \log p(x,y) \\
 &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
 \end{aligned}$$

Pointwise mutual information (PMI)

$$I(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

## Relationship



## K-L divergence

- Formular

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = E_p \left( \log \frac{p(x)}{q(x)} \right)$$

- Interpretation

- A measure of how different two probability distributions are
- The average number of bits that are wasted by encoding events from a distribution  $p$  with a code based on a not-quite-right distribution  $q$

- Characteristics

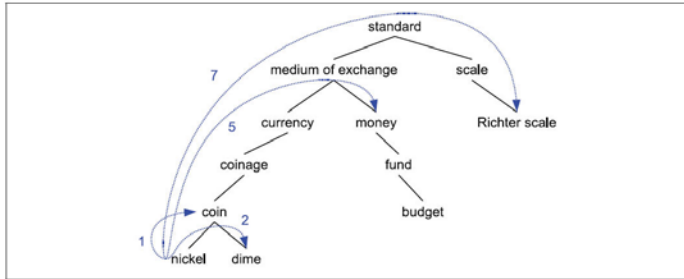
- Non-negative:  $D(p \parallel q) = 0$  iff  $p = q$
- Non-symmetric:  $D(p \parallel q) \neq D(q \parallel p)$

## Word Similarity

- Synonyms: binary relation
- Word similarity or semantic distance: looser metric
- Word similarity and word relatedness
  - “doctor” and “hospital” are closely related, but not similar
  - Similarity is a subcase of relatedness
- Many applications in information retrieval, question answering, language modeling, etc.
- Algorithms to measure word similarity
  - Thesaurus methods
  - Distributional methods

# Similarity based on Path Length

From some thesaurus:



$$\text{Sim}_{\text{path}}(c1,c2) = -\log \text{pathlen}(c1,c2)$$

# Similarity based on Information Content

Resnik (1995): Similarity between two words is measured by their common information; the more two words have in common, the more similar they are.

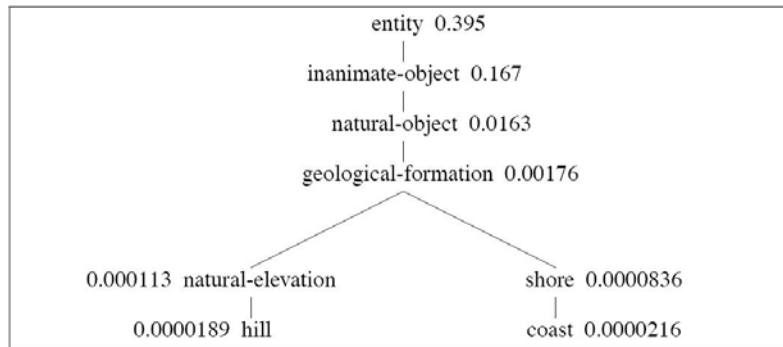
common amount of information is measured by the information content of the lowest common subsumer of the two words (LCS)

$$\text{Sim}_{\text{resnik}}(c1,c2) = -\log P(\text{LCS}(c1,c2))$$

$$P(c) = \frac{\sum_{w \in \text{words}(c)} \text{count}(w)}{N}$$

# Similarity based on Information Content

Probability of P(c) given a fragment of the WordNet hierarchy



# Similarity based on Information Content

- Lin(1998): Take into consideration of both commonality and difference

$$\text{sim}_{\text{Lin}}(c_1, c_2) = \frac{2 \times \log P(\text{LCS}(c_1, c_2))}{\log P(c_1) + \log P(c_2)}$$

- Jiang-Cornrath Distance:

$$\text{Dist}_{\text{JC}}(c_1, c_2) = 2 \times \log P(\text{LCS}(c_1, c_2)) - (\log P(c_1) + \log P(c_2))$$

# Distributional Methods

- Limitations of thesaurus methods:
  - Often lack words, especially new or domain-specific words
  - Only work if rich hyponymy knowledge is present in the thesaurus
  - Difficult to compare words in different hierarchies (e.g., nouns with verbs).
- Distributional methods: the meaning of a word is related to the distribution of words around it.
  - How to represent the distribution of words? -> word co-occurrence vectors
  - How to measure similarity based on the vector representation?

# Word Co-occurrence Vectors

$$\vec{w} = (f_1, f_2, \dots, f_N)$$

How the co-occurrence terms are define? (bag-of-words? Words with grammatical relations to the target words?)

An example for the word “cell” from Lin(1998)

	<i>subj-of</i> , absorb	<i>subj-of</i> , adapt	<i>subj-of</i> , behave	...	<i>pobj-of</i> , inside	<i>pobj-of</i> , into	...	<i>mmod-of</i> , abnormality	<i>mmod-of</i> , anemia	<i>mmod-of</i> , architecture	...	<i>obj-of</i> , attack	<i>obj-of</i> , call	<i>obj-of</i> , come from	<i>obj-of</i> , decorate	...	<i>nmod</i> , bacteria	<i>nmod</i> , body	<i>nmod</i> , bone marrow
cell	1	1	1		16	30		3	8	1		6	11	3	2		3	2	2

# Word Co-occurrence Vectors

$$\vec{w} = (f_1, f_2, \dots, f_N)$$

How these terms are weighted? (Binary? Frequency? Mutual information? )

- A probabilistic measure

$$assoc_{prob}(w, f) = P(f | w) = \frac{count(f, w)}{count(w)}$$

- Better measurement: mutual information

$$assoc_{PMI}(w, f) = \log \frac{P(w, f)}{P(w)P(f)}$$

# Word Co-occurrence Vectors

$$\vec{w} = (f_1, f_2, \dots, f_N)$$

An example for the word “cell” from Lin(1998)

	<i>subj-of</i> , absorb	<i>subj-of</i> , adapt	<i>subj-of</i> , behave	...	<i>pobj-of</i> , inside	<i>pobj-of</i> , into	...	<i>mmod-of</i> , abnormality	<i>mmod-of</i> , anemia	<i>mmod-of</i> , architecture	...	<i>obj-of</i> , attack	<i>obj-of</i> , call	<i>obj-of</i> , come from	<i>obj-of</i> , decorate	...	<i>nmod</i> , bacteria	<i>nmod</i> , body	<i>nmod</i> , bone marrow
cell	1	1	1		16	30		3	8	1		6	11	3	2		3	2	2

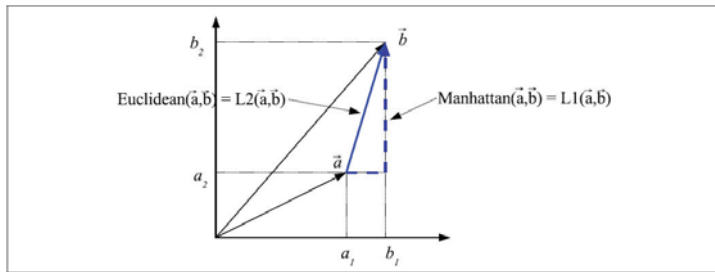
Weight of the feature: could be association between each target word and a given feature f:

$$assoc_{PMI}(w, f) = \log \frac{P(w, f)}{P(w)P(f)}$$

## Distance between Two Vectors

How to measure the distance between two vectors?

$$L_1 \text{ norm: } L_1(\vec{x}, \vec{y}) = \sum_{i=1}^N |x_i - y_i| \quad L_2 \text{ norm: } L_2(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}$$



## Approaches based on IR

Vector-space model

$$sim_{\cosine}(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}| |\vec{w}|} = \frac{\sum_{i=1}^N v_i \times w_i}{\sqrt{\sum_{i=1}^N v_i^2} \sqrt{\sum_{i=1}^N w_i^2}}$$

Other related metrics:

$$sim_{Jaccard}(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^N \min(v_i, w_i)}{\sum_{i=1}^N \max(v_i, w_i)}, \quad sim_{Dice}(\vec{v}, \vec{w}) = \frac{2 \times \sum_{i=1}^N \min(v_i, w_i)}{\sum_{i=1}^N (v_i + w_i)}$$

## Information Theoretical Models

Jenson-Shannon divergence : based on conditional probability association:  $P(f|w)$

$$sim_{JS}(\vec{v} || \vec{w}) = D(\vec{v} || \frac{\vec{v} + \vec{w}}{2}) + D(\vec{w} || \frac{\vec{v} + \vec{w}}{2})$$

Where:  $D(\cdot, \cdot)$  is KL Divergence.

## Evaluating Word Similarity

$S$ : the set of words that are defined as similar in the thesaurus: e.g., being in the same synset, or share the same hyperny, etc.

$S'$  be the set of words that are classified as similar by some algorithm

$$precision = \frac{|S \cap S'|}{|S'|}$$

$$recall = \frac{|S \cap S'|}{|S|}$$

# Computational Discourse

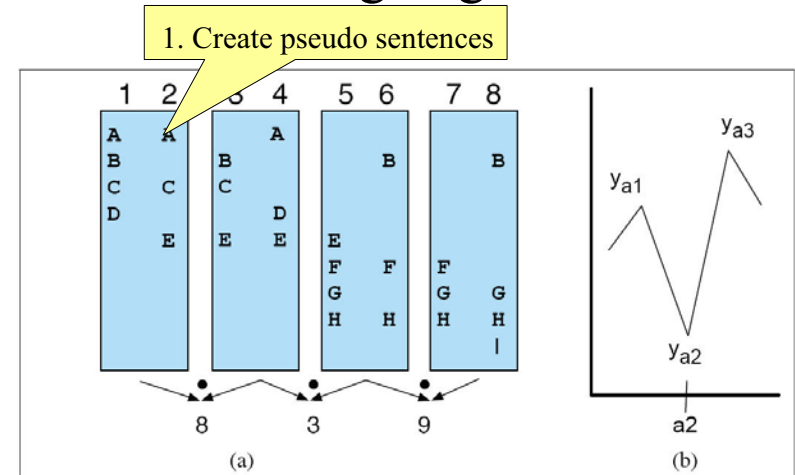
# Pragmatics and Discourse

- Pragmatics: context dependent meaning
- Discourse: anything longer than a single utterance or sentence, a group of sentences
  - Monologue
  - Dialogue:
    - May be multi-party
    - May be human-machine
- Topics
  - Discourse segmentation
  - Text coherence
  - Reference resolution
  - Coreference resolution

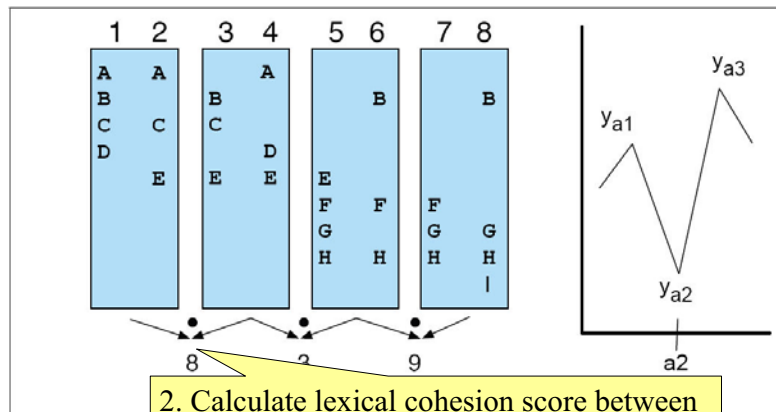
# Discourse Segmentation

- Separate a document into a linear sequence of subtopics
- Unsupervised approaches:
  - Cohesion-based approach
  - Cohesion: the use of certain linguistic devices to link or tie together textual units (e.g., repetition of words, hyponyms, synonyms, etc.)
  - Approaches: TextTiling, clustering.
- Supervised approaches:
  - Train classifiers based on annotated data

# TextTiling Algorithm

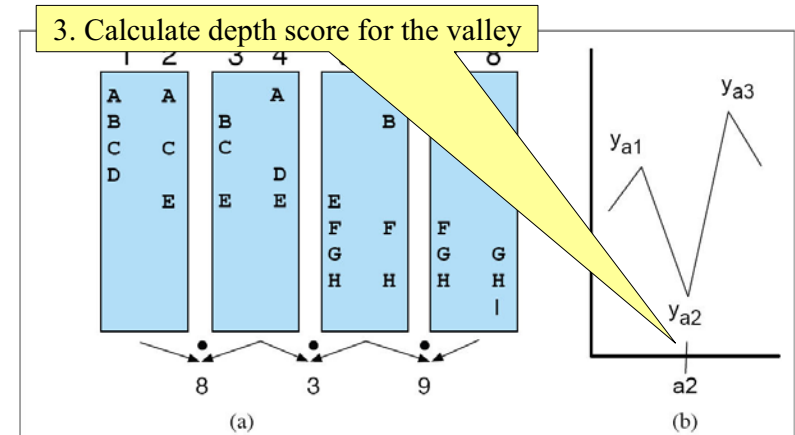


# TextTiling Algorithm

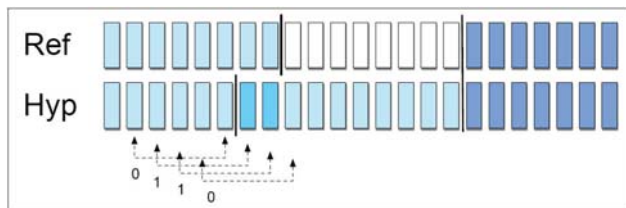


2. Calculate lexical cohesion score between sentences based on similarity of text before and after the gap

# TextTiling Algorithm



# DS Evaluation: WindowDiff



$$WindowDiff(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| \neq 0)$$

Where  $b(i,j)$  is the number of boundaries between positions  $i$  and  $j$  in a text

# Text Coherence

- (a) John hid Bill's car keys. He was drunk.
- (b) \*John hid Bill's car keys. He likes spinach.

Coherence: the meaning relations between two textual units.

Example of relations (Hobbs'79)

**Result:** The Tin Woodman was caught in the rain. His joints rusted.

**Explanation:** John hid Bill's car keys. He was drunk.

**Parallel:** The Scarecrow wanted some brains. The Tin woodman wanted a heart.

**Elaboration:** Dorothy was from Kansas. She lived in the midst of the great Kansas prairies.

**Occasion:** Dorothy picked up the oil-can. She oiled the Tin Woodman's joints

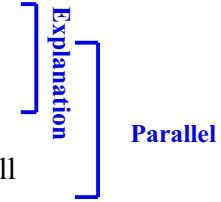
# Discourse Structure

- S1: John went to the bank to deposit his paycheck
- S2: He then took a train to Bill's car dealership.
- S3: He needed to buy a car
- S4: The company he works for now isn't near any public transportation
- S5: He also wanted to talk to Bill about their softball league.



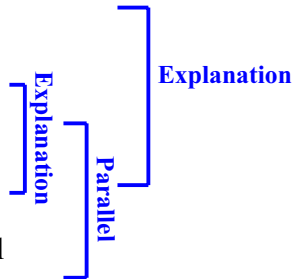
# Discourse Structure

- S1: John went to the bank to deposit his paycheck
- S2: He then took a train to Bill's car dealership.
- S3: He needed to buy a car
- S4: The company he works for now isn't near any public transportation
- S5: He also wanted to talk to Bill about their softball league.



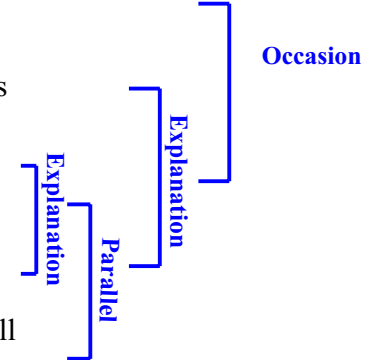
# Discourse Structure

- S1: John went to the bank to deposit his paycheck
- S2: He then took a train to Bill's car dealership.
- S3: He needed to buy a car
- S4: The company he works for now isn't near any public transportation
- S5: He also wanted to talk to Bill about their softball league.



# Discourse Structure

- S1: John went to the bank to deposit his paycheck
- S2: He then took a train to Bill's car dealership.
- S3: He needed to buy a car
- S4: The company he works for now isn't near any public transportation
- S5: He also wanted to talk to Bill about their softball league.



# Text Coherence

S1: John went to the bank to deposit his paycheck

S2: he then took a train to Bill's car dealership

S3: He needed to buy a car.

S4: The company he works for now isn't near any public transportation.

S5: He also wanted to talk to Bill about their softball league.

