Lexical Semantics

Regina Barzilay

MIT

July, 2005

Today: Semantic Similarity

It's not pinin,' it's passed on! This parrot is no more! It has ceased to be! It's expired and gone to meet its maker! This is a late parrot! It's a stiff! Bereft of life, it rests in peace! If you hadn't nailed him to the perch he would be pushing up the daisies! Its metabolical processes are of interest only to historians! It's hopped the twig! It's shuffled off this mortal coil! It's run down the curtain and joined the choir invisible! This.... is an EX-PARROT!

From Monty Python's "Dead Parrot" sketch: http://en.wikipedia.org/wiki/Dead_Parrot

Today: Semantic Similarity

This parrot is no more!
It has ceased to be!
It's expired and gone to meet its maker!
This is a late parrot!
This. . . is an EX-PARROT!

Motivation

Smoothing for statistical language models

- Two alternative guesses of speech recognizer:

  For breakfast, she ate durian.

  For breakfast, she ate Dorian.

- Our corpus contains neither "*ate durian*" nor "*ate Dorian*"

- But, our corpus contains "*ate orange*", "*ate banana*"

Motivation

Aid for Question-Answering and Information Retrieval

- Task: "Find documents about women astronauts"
- Problem: some documents use paraphrase of *astronaut*

> In the history of Soviet/Russian space exploration, there have only been three Russian women cosmonauts: Valentina Tereshkova, Svetlana Savitskaya, and Elena Kondakova.

Motivation

Exploration in language acquisition

- Miller&Charles: judgments of semantic similarity can be explained by the degree of *contextual interchangeability*
- Can we automatically predict which words human perceive as similar?

Computing Semantic Similarity

- Use human-created resources
- Acquire required knowledge from text

Lexicons and Semantic Nets

- Lexicons are word lists augmented with some subset of information
  - Parts-of-speech
  - Different word senses
  - Synonyms
- Semantic Nets
  - Links between terms (IS-A, Part-Of)

WordNet

- A big lexicon with properties of a semantic net
- Started as a language project by George Miller and Christiane Fellbaum at Princeton
- First became available in 1990

Category	Unique Forms	Number of Senses
Noun	114648	79689
Verb	11306	13508
Adjective	21436	18563
Adverb	4669	3664

Synset Example

1. water, H2O – (binary compound that occurs at room temperature as a clear colorless odorless tasteless liquid; freezes into ice below 0 degrees centigrade and boils above 100 degrees centigrade; widely used as a solvent)
2. body of water, water – (the part of the earth's surface covered with water (such as a river or lake or ocean); "they invaded our territorial waters"; "they were sitting by the water's edge")
3. water system, water supply, water – (facility that provides a source of water; "the town debated the purification of the water supply"; "first you have to cut off the water")
4. water – (once thought to be one of four elements composing the universe (Empedocles))
5. urine, piss, pee, piddle, weewee, water – (liquid excretory product; "there was blood in his urine"; "the child had to make water")
6. water – (a fluid necessary for the life of most animals and plants; "he asked for a drink of water")

WordNet Relations

- Original core relations:
  - Synonymy
  - Polysemy
  - Metonymy
  - Hyponymy/Hyperonymy
  - Meronymy
  - Antonymy
- New, useful addition for NLP:
  - Glosses
  - Links between derivationally and semantically related noun/verb pairs
  - Domain/topical terms
  - Groups of similar verbs

Synonymy

- Synonyms are different ways of expressing related concepts
  - Examples: *marriage, matrimony, union, wedlock*
- Synonyms are almost never truly substitutable:
  - Used in different contexts
  - Have different implications
    - This is a point of contention

Polysemy

- Most words have more than one sense
  - Homonymy: same word, unrelated meanings
    - *bank* (river)
    - *bank* (financial)
  - Polysemy: same word, related meanings
    - *Bob has ugly ears.*
    - *Alice has a good ear for jazz.*

Polysemy Information

POS	Monosemous	Polysemous
Noun	99524	15124
Verb	6256	5050
Adverb	16103	5333
Adjective	3901	768
Total	125784	26275

Metonymy

- Use one aspect of something to stand for the whole
  - Newscast: *The White House released new figures today.*
  - Waitperson: *The tofu sandwich spilled his drink.*

Hyponymy/Hyperonymy (ISA)

A is a hypernym of B if B is a type of A

A is a hyponym of B if A is a type of B

Example:

- bigamy (having two spouses at the same time)
- open marriage (a marriage in which each partner is free to enter into extraneous sexual relationships without guilt or jealousy from the other)
- cuckoldom (the state of a husband whose wife has committed adultery)
- polygamy (having more than one spouse at a time)
  - polyandry (having more than one husband at a time)
  - polygyny (having more than one wife at a time)

Meronymy

- Part-of relation
  - part-of (*beak, bird*)
  - part-of (*bark, tree*)
- Transitive conceptually but not lexically:
  - *The knob is a part of the door.*
  - *The door is a part of the house.*
  - *? The knob is a part of the house.*

Antonymy

- Lexical opposites
  - antonym (*large, small*)
  - antonym (*big, small*)
  - antonym (*big, little*)

Computing Semantic Similarity

Suppose you are given the following words. Your task is to group them according to how similar they are:

apple

banana

grapefruit

grape

man

woman

baby

infant

Using WordNet to Determine Similarity

apple

fruit

produce

. . .

banana

fruit

produce

. . .

man

male, male person

person, individual

organism

. . .

woman

female , female person

person, individual

organism

Similarity by Path Length

- Count the edges (is-a links) between two concepts and scale
- Leacock and Chodorow, 1998:

$$d(c_1, c_2) = -log \frac{length(c_1, c_2)}{2 * MaxDepth}$$

- Wu and Palmer, 1994:

$$d(c_1, c_2) = -log \frac{depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

Why use WordNet?

- Quality
  - Developed and maintained by researchers
- Habit
  - Many applications are currently using WordNet
- Available software
  - SenseRelate(Pedersen et al): `http://wn-similarity.sourceforge.com`

Similarity by Path Length

man

male, male person

person, individual

organism

. . .

woman

female , female person

person, individual

organism

baby

child, kid

offspring, progeny

relative, relation

person, individual

organism

. . .

Why not use WordNet?

- Incomplete (technical terms may be absent)
- The length of the paths are irregular across the hierarchies
- How to relate terms that are not in the same hierarchies?
  The “tennis problem”:
  - *Player*
  - *Racquet*
  - *Ball*
  - *Net*

Learning Similarity from Corpora

- You shall know a word by the company it keeps (Firth 1957)
- Key assumption: Words are similar if they occur in similar contexts

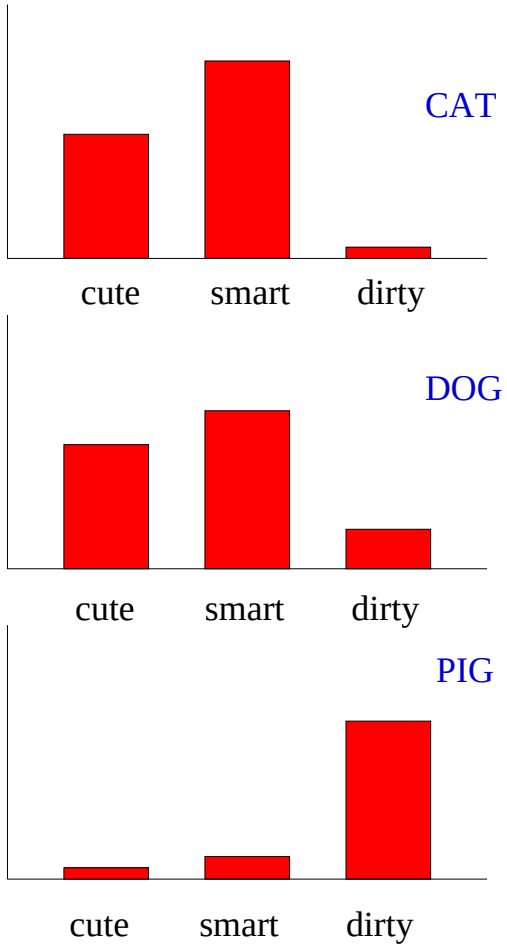
What is tizguino? (Nida, 1975)

A bottle of tizguino is on the table.

Tizguino makes you drunk.

We make tizguino out of corn.

Learning Similarity from Corpora



Learning Similarity from Corpora

- Define the properties one cares about, and be able to give numerical values for each property
- Create a vector of length n with the n numerical values for each item to be classified
- Viewing the n-dimensional vector as a point in an n-dimensional space cluster points that are near one another

Key Parameters

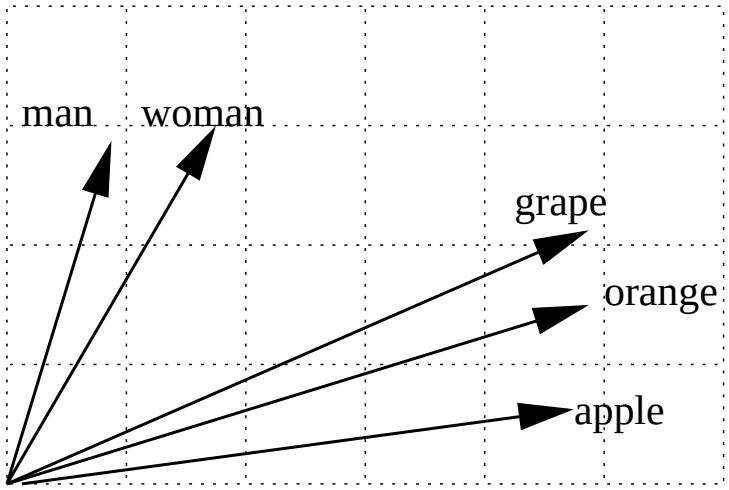
- The properties used in the vector
- The distance metric used to decide if two points are “close”
- The algorithm used to cluster

Example 1: Clustering by Next Word

Brown et al. (1992)

- $C(x)$ denotes the vector of properties of x ("context" of x)
- Assume alphabet of size K: $w^1, \ldots, w^K$
- $C(w^i) = \langle |w^1|, |w^2|, \ldots, |w^K| \rangle$, where $|w^j|$ followed $|w^i|$ in the corpus

Vector Space Model



Similarity Measure: Euclidean

Euclidean $|\vec{x}, \vec{y}| = |\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

	cosmonaut	astronaut	moon	car	truck
Soviet	1	0	0	1	1
American	0	1	0	1	1
spacewalking	1	1	0	0	0
red	0	0	0	1	1
full	0	0	1	0	0
old	0	0	0	1	1

$cos(cosm, astr) =$

$$\sqrt{(1-0)^2 + (0-1)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2}$$

Similarity Measure: Cosine

Each word is represented as a vector $\vec{x} = (x_1, x_2, \ldots, x_n)$

- Cosine $cos(\vec{x}, \vec{y}) = \frac{\vec{x} * \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x^2}\sqrt{\sum_{i=1}^{n} y^2}}$
  - Angle between two vectors
  - Ranges from 0 (cos(90)=0) to 1 (cos(0)=1)

Computing Similarity: Cosine

	cosmonaut	astronaut	moon	car	truck
Soviet	1	0	0	1	1
American	0	1	0	1	1
spacewalking	1	1	0	0	0
red	0	0	0	1	1
full	0	0	1	0	0
old	0	0	0	1	1

$$cos(cosm, astr) = \frac{1*0+0*1+1*1+0*0+0*0+0*0}{\sqrt{1^2+0^2+1^2+0^2+0^2+0^2}\sqrt{0^2+1^2+1^2+0^2+0^2+0^2}}$$

Term Weighting

Quantity	Symbol	Definition
term frequency	$tf_{i,j}$	# occurrences of w_i in d_j
document frequency	df_i	# documents that w_i occurs in

$$tf \times idf = \begin{cases} (1 + \log(tf_{i,j}))log\frac{N}{df_i} & \text{if } tf_{i,j} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Cosine vs. Euclidean

- Cosine applied to normalized vectors gives the same ranking of similarities as Euclidean distance does.
- Both metrics assume Euclidean space
  - Suboptimal for vectors of probabilities (0.0 and 0.1 *vs.* 0.9 and 1)

Mutual Information

- Definition: The mutual information $I(x;y)$ of two particular outcomes x and y is the amount of information one outcome gives us about another one
- $I(x;y) = (-\log P(x)) - (-logP(x|y)) = log\frac{P(x,y)}{P(x)P(y)}$

Example

$$I(pancake; syrup) = \log \frac{P(pancake,syrup)}{P(pancake)P(syrup)}$$

$$I(W_i = pancake; W_{i+1} = syrup) = \log \frac{P(W_i=pancake,W_{i+1}=syrup)}{P(W_i=pancake)P(W_{i+1}=syrup)}$$

- "pancake" and "syrup" have no relation to each other ($P(syrup|puncake) = P(syrup)$)

$$\begin{aligned} I(pancake, syrup) &= \log \frac{P(pancake, syrup)}{P(pancake)P(syrup)} \\ &= \log \frac{P(syrup|pancake)}{P(syrup)} \\ &= \log \frac{P(syrup)}{P(syrup)} = 0 \end{aligned}$$

Example(cont)

$$I(pancake; syrup) = \log \frac{P(pancake,syrup)}{P(pancake)P(syrup)}$$

- "pancake" and "syrup"are perfectly coordinated

$$\begin{aligned} I(pancake, syrup) &= \log \frac{P(pancake, syrup)}{P(pancake)P(syrup)} \\ &= \log \frac{P(pancake)}{P(pancake)P(syrup)} \\ &= \log \frac{1}{P(syrup)} \end{aligned}$$

Similarity for LM

Goal: find word clustering that decreases perplexity

$$
\begin{aligned}
H(L) &= -\frac{1}{N}\log P(w_1,\ldots,w_N) \\
&\approx \frac{-1}{N-1}\log\prod_{i=2}^{N} P(w_i|w_{i-1}) \\
&\approx \frac{-1}{N-1}\sum_{w^1 w^2} Count(w^1 w^2)\log P(w^2|W^1)
\end{aligned}
$$

Similarity for LM

Cluster-based generalization:

$$\begin{aligned} H(L,\pi) &\approx \frac{-1}{N-1} \sum_{w^1 w^2} Count(w^1 w^2) \log P(c_2|c_1) P(w^2|c_1) \\ &\approx H(w) - I(c_1, c_2) \end{aligned}$$

Average Mutual Information

- Definition: Average mutual information of the random variables X and Y, $I(X, Y)$ is the amount of information we get about X from knowing the value of Y, on the average.
- $I(X;Y) = \sum_{y=1}^{K} \sum_{x=1}^{K} P(w^x, w^y) I(w^x, w^y)$

Example: Syntax-Based Representation

- The vector $C(n)$ for a word n is the distribution of verbs for which it served as direct object
- $C(n) = P(v^1|n), P(v^2|n), \ldots, P(v^K|n)$
- Representation can be expanded to account for additional syntactic relations (subject, object, indirect-object, neutral)

Kullback Leibler Distance (Relative Entropy)

- Definition: The relative entropy $D(p||q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p
- $D(p||q) = \sum p(x) log \frac{p(x)}{q(x)}$
- Properties:
  - Non-negative
  - $D(p||q) = 0$ iff $p = q$
  - Not symmetric and doesn't satisfy triangle inequality

Representation

- Representation
  - Syntactic vs. Window-based
  - Context granularity
  - Alphabet size
  - Counts vs. Probability
- Distance
  - Vector-based vs. Probabilistic
  - Weighted vs. Unweighted

Problems with Corpus-based Similarity

- Low-frequency words skew the results
  - "breast-undergoing", "childhood-phychosis", "outflow-infundibulum"
- Semantic similarity does not imply synonymy
  - "large-small", "heavy-light", "shallow-coastal"
- Distributional information may not be sufficient for true semantic grouping

Not-so-semantic grouping

Method	Clinker
Direct Object	pollution increase failure
Next Word	addiction medalist inhalation Arabia growers
Adjective	full increase

State-of-the-art Methods

`http://www.cs.ualberta.ca/~lindek/demos/depsim.htm`

Closest words for *president*

leader 0.264431, minister 0.251936, vice president 0.238359, Clinton 0.238222, chairman 0.207511, government 0.206842, Governor 0.193404, official 0.191428, Premier 0.177853, Yeltsin 0.173577, member 0.173468, foreign minister 0.171829, Mayor 0.168488, head of state 0.167166, chief 0.164998, Ambassador 0.162118, Speaker 0.161698, General 0.159422, secretary 0.156158, chief executive 0.15158

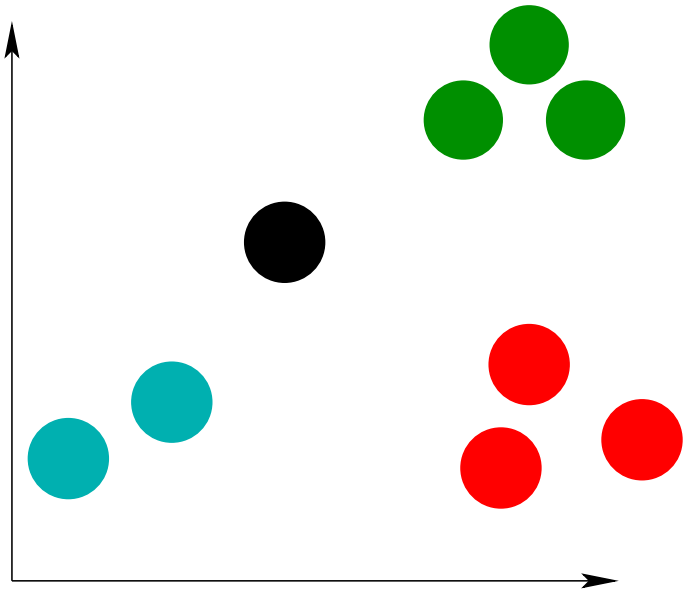
State-of-the-art Methods

Closest words for ?

anthropology 0.275881, sociology 0.247909, comparative literature 0.245912, computer science 0.220663, political science 0.219948, zoology 0.210283, biochemistry 0.197723, mechanical engineering 0.191549, biology 0.189167, criminology 0.178423, social science 0.176762, psychology 0.171797, astronomy 0.16531, neuroscience 0.163764, psychiatry 0.163098, geology 0.158567, archaeology 0.157911, mathematics 0.157138

Beyond Pairwise Similarity

- Clustering is "The art of finding groups in data"(Kaufmann and Rousseeu)
- Clustering algorithms divide a data set into homogeneous groups (clusters), based on their similarity under the given representation.



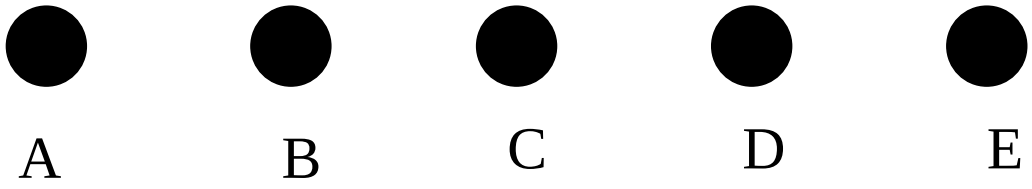
Hierarchical Clustering

Greedy, bottom-up version:

- Initialization: Create a separate cluster for each object
- Each iteration: Find two most similar clusters and merge them
- Termination: All the objects are in the same cluster

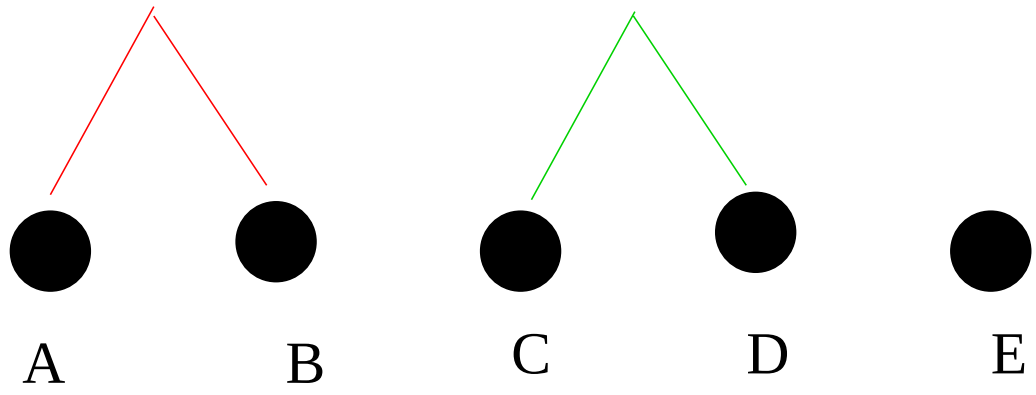
Agglomerative Clustering

	E	D	C	B
A	0.1	0.2	0.2	0.8
B	0.1	0.1	0.2	
C	0.0	0.7		
D	0.6			



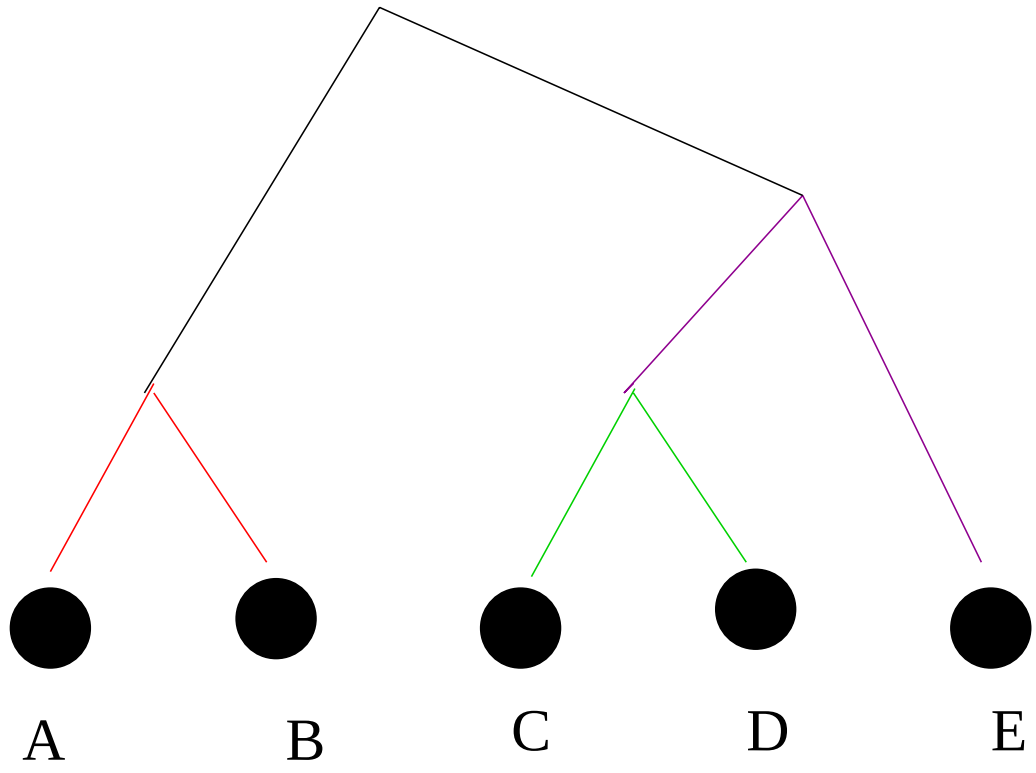
Agglomerative Clustering

	E	D	C	B
A	0.1	0.2	0.2	0.8
B	0.1	0.1	0.2	
C	0.0	0.7		
D	0.6			



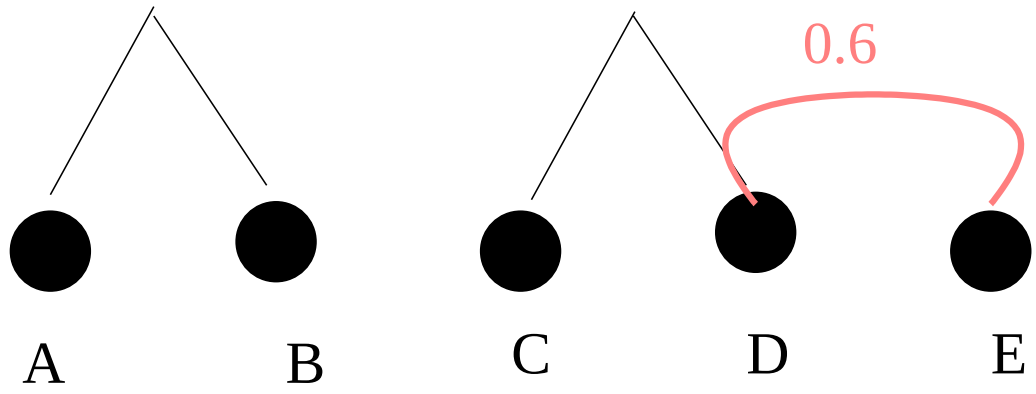
Agglomerative Clustering

	E	D	C	B
A	0.1	0.2	0.2	0.8
B	0.1	0.1	0.2	
C	0.0	0.7		
D	0.6			



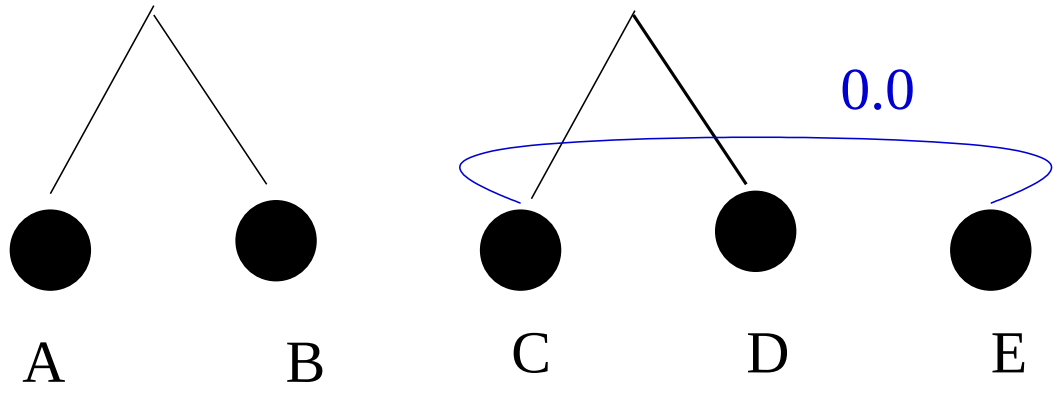
Clustering Function

	E	D	C	B
A	0.1	0.2	0.2	0.8
B	0.1	0.1	0.2	
C	0.0	0.7		
D	0.6			



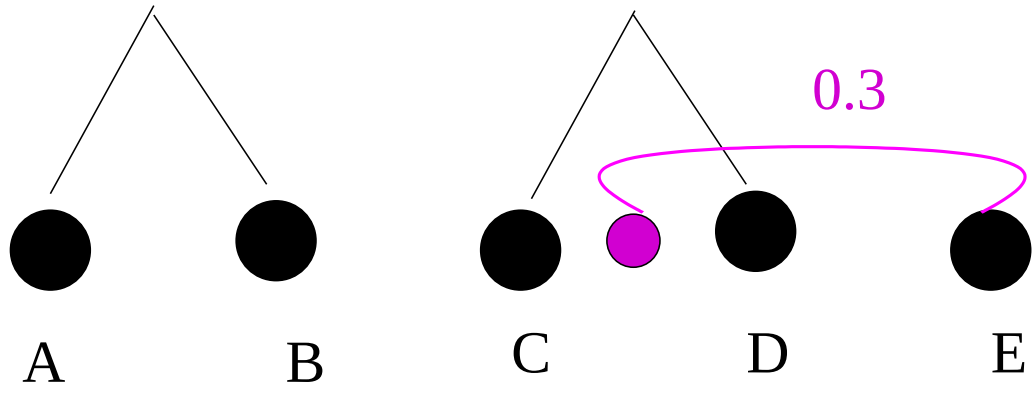
Clustering Function

	E	D	C	B
A	0.1	0.2	0.2	0.8
B	0.1	0.1	0.2	
C	0.0	0.7		
D	0.6			



Clustering Function

	E	D	C	B
A	0.1	0.2	0.2	0.8
B	0.1	0.1	0.2	
C	0.0	0.7		
D	0.6			



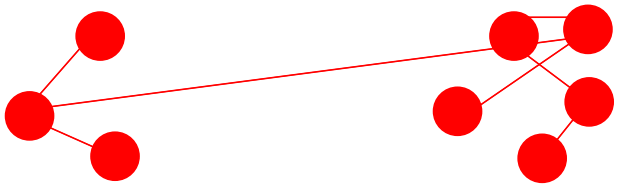
Clustering Function

- Single-link: Similarity of two most similar members
- Complete-link: Similarity of two least similar members
- Group-average: Average similarity between members

Single-Link Clustering

- Achieves Local Coherence
- Complexity $O(n^2)$
- Fails when clusters are not well separated

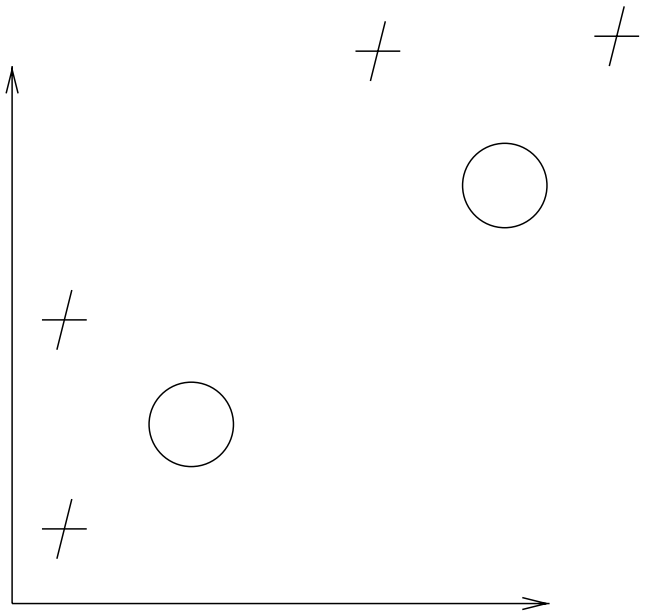
Complete-Link Clustering

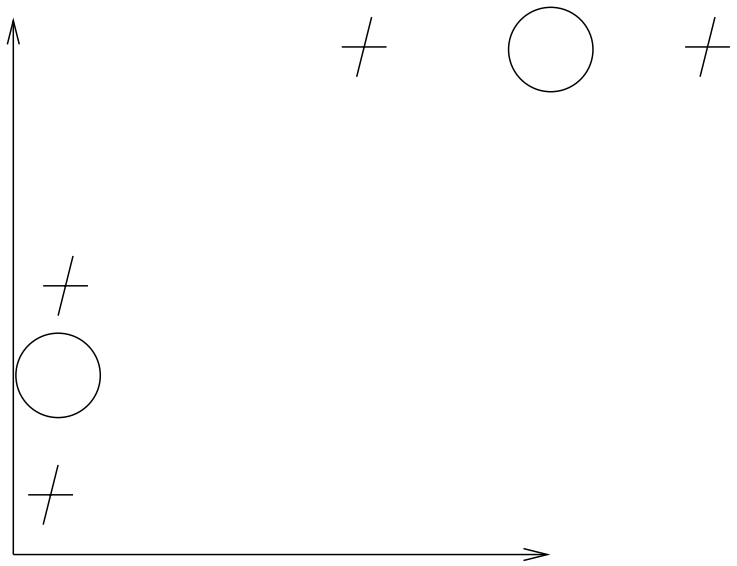


- Achieves Global Coherence
- Complexity $O(n^2 \log n)$
- Fails when clusters aren't spherical, or of uniform size

K-Means Algorithm: Example

Iterative, hard, flat clustering algorithm based on Euclidean distance





K-Means Algorithm

1. Choose k points at random as cluster centers
2. Assign each instance to its closest cluster center
3. Calculate the centroid (mean) for each cluster, use it as a new cluster center
4. Iterate (2-3) until the cluster centers don't change anymore

K-Means Algorithm: Hard EM

1. Guess initial parameters
2. Use model to make the best guess of c_i (E-step)
3. Use the new complete data to learn better model (M-step)
4. Iterate (2-3) until convergence

Evaluating Clustering Methods

- Perform task-based evaluation
- Test the resulting clusters intuitively, i.e., inspect them and see if they make sense. Not advisable.
- Have an expert generate clusters manually, and test the automatically generated ones against them.
- Test the clusters against a predefined classification if there is one

Comparing Clustering Methods

(Meila, 2002)

n	total # of points
n_k	# of points in cluster C_k
K	# of nonempty clusters
N_{11}	# of pairs that are in the same cluster under C and C'
N_{00}	# of pairs that are in the different clusters under C and C'
N_{10}	# of pairs that are in the the same cluster under C but not C'
N_{01}	# of pairs that are in the the same cluster under C' but not C

Comparing by Counting Pairs

- Wallace criteria

$$W_1(C, C') = \frac{N_{11}}{\sum_k n_k(n_k - 1)/2}$$

$$W_2(C, C') = \frac{N_{11}}{\sum_{k'} n_{k'}(n'_{k'} - 1)/2}$$

- Fowles-Mallows criterion

$$F(C, C') = \sqrt{W_1(C, C')W_2(C, C')}$$

Problems: ?

Comparing Clustering by Set Matching

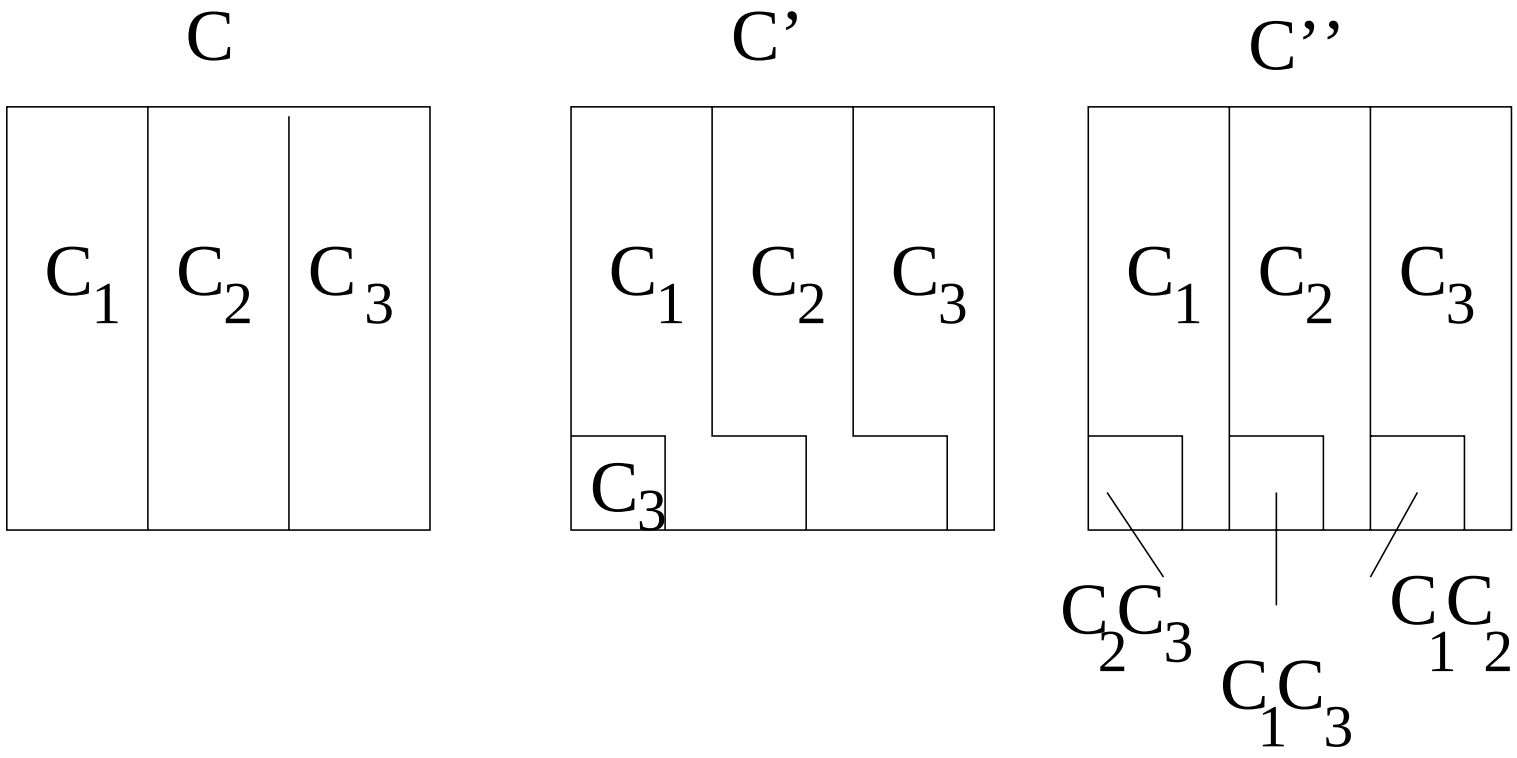
Contingency table M is a $K \times K$ matrix, whose kk' element is the number of points in the intersection of clusters C_k and $C'_{k'}$

$$L(C, C') = \frac{1}{K} \sum_{k} \max_{k'} \frac{2m_{kk'}}{n_k + n'_k}$$

Problems: ?

Comparing Clustering by Set Matching

$$L(C, C') = \frac{1}{K} \sum_{k} \max_{k'} \frac{2m_{kk'}}{n_k + n'_k}$$



Summary

- Lexicon-based Similarity Computation
  - WordNet relations
  - Path-based similarity
- Corpus-based Similarity Computation
  - Vector Space Model
  - Similarity Measures
  - Hierarchical Clustering