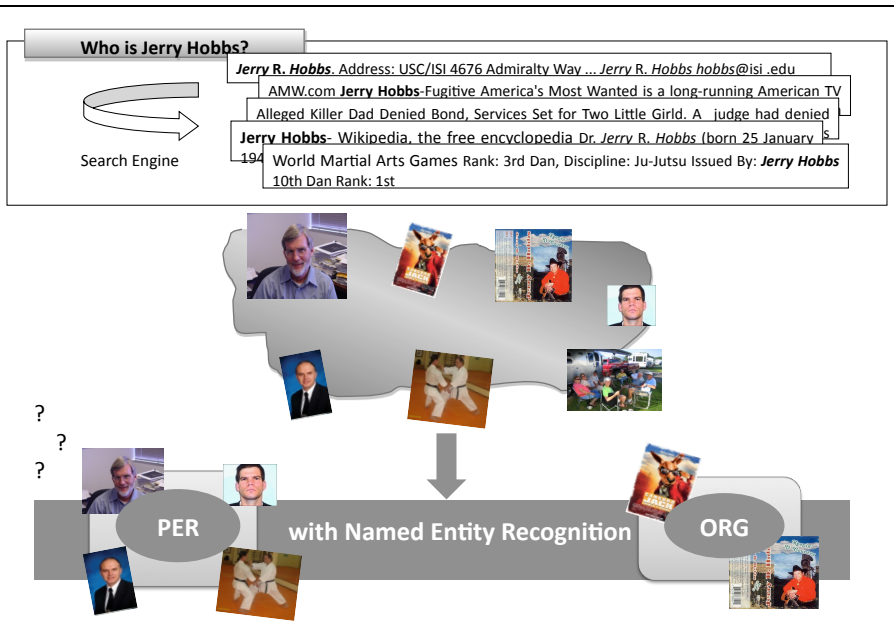
CS544: Named Entity Discrimination

March 30, 2010

Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva



1

NE Recognition vs. NE Discrimination

- NE Recognition = detection & classification of entity mentiones into a predefined set of categories.

  ⇒ achieves only a partial disambiguation of names

- NE Discrimination = finding the actual entity denoted by a particula name occurrence in text.

? ? ?

2

Ideally, we want

Dr. *Jerry* R. *Hobbs* is a prominent researcher in the fields of computational linguistic

Jerry R. *Hobbs*. Address: USC/ISI 4676 Admiralty Way .

Jerry Hobbs is the rage-filled, domestic-abusing career criminal who killed his

Jerry Hobbs, who is accused of killing his 8-year-old daughter and her best

Jerry Hobbs, a fifth generation farmer, Wayne Cryts finished harvesting his crop

3

Problem Formulation

- A *text snippet* is a small fragment of text that contains from one to three sentences
- Input:
  - N *text snippets* that mention a particular proper name (it can be person, organization or location)
- Output:
  - K clusters, where each cluster has *text snippets* that are similar to each other and different from the *snippets* in the rest of the clusters

4

Input

- Dr. ***Jerry* R. *Hobbs*** (born 25 January 1942) is a prominent researcher in the fields of computational linguistics, discourse analysis, and artificial
- ***Jerry Hobbs*** is the rage-filled, domestic-abusing career criminal who killed his 8-year-old daughter and her 9-year-old friend, with scarcely **...**
- ***Jerry Hobbs***, Author. A fifth generation farmer, Wayne Cryts finished harvesting his crop in the fall of 1980 and hauled more than 32000 bushels of soybeans **...**
- ***Jerry Hobbs***, who is accused of killing his 8-year-old daughter and her best **...** On Wednesday, a judge denied bail for ***Jerry Hobbs***, 34, **...**
- Fugitives | ***Jerry Hobbs*** - Brief - Father Denied Bail Awaits Trial For Children s Murders Jerry Branton Hobbs accused of the stabbing deaths **...**
- ***Jerry* R. *Hobbs***. Address: USC/ISI 4676 Admiralty Way **...** ***Jerry* R. *Hobbs*** *hobbs*@isi.edu. USC/ISI, 4676 Admiralty Way, Marina del Rey, CA 90292

5

Output

- **Cluster 1:**
  - Dr. ***Jerry* R. *Hobbs*** (born 25 January 1942) is a prominent researcher in the fields of computational linguistics, discourse analysis, and artificial
  - ***Jerry* R. *Hobbs***. Address: USC/ISI 4676 Admiralty Way **... *Jerry* R. *Hobbs*** *hobbs*@isi.edu. USC/ISI, 4676 Admiralty Way, Marina del Rey, CA 90292

- **Cluster 2:**
  - ***Jerry Hobbs*** is the rage-filled, domestic-abusing career criminal who killed his 8-year-old daughter and her 9-year-old friend, with scarcely **...**
  - ***Jerry Hobbs***, who is accused of killing his 8-year-old daughter and her best **...** On Wednesday, a judge denied bail for ***Jerry Hobbs***, 34, **…**
  - Fugitives | ***Jerry Hobbs*** - Brief - Father Denied Bail Awaits Trial For Children s Murders Jerry Branton Hobbs accused of the stabbing deaths **...**

- **Cluster 3:**
  - Fugitives | ***Jerry Hobbs*** - Brief - Father Denied Bail Awaits Trial For Children s Murders Jerry Branton Hobbs accused of the stabbing deaths **...**

6

Disambiguation vs. Discrimination

Disambiguation

- **the total number of senses is known**
- **the meaning of each sense is known**
- **the order is based on the frequency**

bank

meaning 1: 

the slope beside a body of water

meaning 2: 

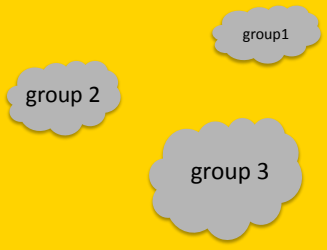
depository financial institution

…

Discrimination

- **the total number of senses is unknown**
- **the meaning of each sense is unknown**
- **no specific mapping of cluster& sense**

bank



7

On the Web …

- Nobody knows how many senses (meanings) are there for a given person name
- It is impossible to estimate and trace the most frequent sense
  - the task is time consuming and tedious for humans
  - new Web pages constantly appear
  - old Web pages might be deleted over time

8

Importance of NE Discrimination

- Queries about NEs constitute significant portion of Web queries:
  - 11-17% contain person name*
  - 4% are about a person name*
- Ideally, search results should be clustered such that each cluster corresponds to the same individual
  - faster fact extraction
  - more accurate information retrieval

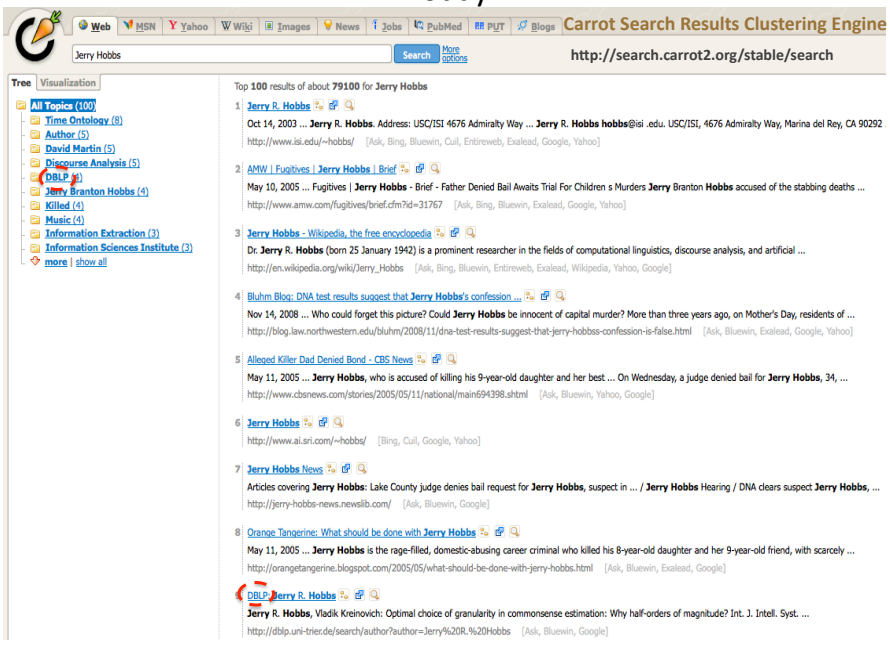
* study by Javier Artiles, 2009

9

Today


WIKIPEDIA
The Free Encyclopedia
navigation
Main page
Contents
Featured content
Current events
Random article
search
Go
Search
interaction
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Donate to Wikipedia
Help
toolbox
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Cite this page
languages
article
discussion
edit this page
history
Madonna
From Wikipedia, the free encyclopedia
Not all proper names are listed in Wikipedia
Madonna (Italian: My Lady) may refer to:
Christianity
Mary (mother of Jesus), from which other uses generally derive
Madonna (art), a portrait of Mary
Madonna (Edvard Munch), a painting by Edvard Munch
Entertainer
Madonna (entertainer), the American singer-songwiter-producer and actress
Madonna (album), the entertainer's self-titled first album
Madonna (video compilation), a music video collection
Other uses in entertainment
Madonna (...And You Will Know Us by the Trail of Dead album)
Madonna (studio), a Japanese adult video (AV) company based in Tokyo
The "Madonna" was a type of bob cut in the U.S. in the twenties
This disambiguation page lists articles associated with the same title.
If an internal link led you here, you may wish to change the link to point directly to the intended article.
Categories: Disambiguation pages

Today


Web
MSN
Yahoo
Wiki
Images
News
Jobs
PubMed
PUT
Blogs
Carrot Search Results Clustering Engine
Jerry Hobbs
Search
More options
http://search.carrot2.org/stable/search
Tree
Visualization
All Topics (100)
Time Ontology (8)
Author (5)
David Martin (5)
Discourse Analysis (5)
DBLP (4)
Jerry Branton Hobbs (4)
Killed (4)
Music (4)
Information Extraction (3)
Information Sciences Institute (3)
more | show all
Top 100 results of about 79100 for Jerry Hobbs
1 Jerry R. Hobbs
Oct 14, 2003 ... Jerry R. Hobbs. Address: USC/ISI 4676 Admiralty Way ... Jerry R. Hobbs hobbs@isi .edu. USC/ISI, 4676 Admiralty Way, Marina del Rey, CA 90292
http://www.isi.edu/~hobbs/ [Ask, Bing, Bluewin, Cuil, Entireweb, Exalead, Google, Yahoo]
2 AMW | Fugitives | Jerry Hobbs | Brief
May 10, 2005 ... Fugitives | Jerry Hobbs - Brief - Father Denied Bail Awaits Trial For Children s Murders Jerry Branton Hobbs accused of the stabbing deaths ...
http://www.amw.com/fugitives/brief.cfm?id=31767 [Ask, Bing, Bluewin, Exalead, Google, Yahoo]
3 Jerry Hobbs - Wikipedia, the free encyclopedia
Dr. Jerry R. Hobbs (born 25 January 1942) is a prominent researcher in the fields of computational linguistics, discourse analysis, and artificial ...
http://en.wikipedia.org/wiki/Jerry_Hobbs [Ask, Bing, Bluewin, Entireweb, Exalead, Wikipedia, Yahoo, Google]
4 Bluhm Blog: DNA test results suggest that Jerry Hobbs's confession ...
Nov 14, 2008 ... Who could forget this picture? Could Jerry Hobbs be innocent of capital murder? More than three years ago, on Mother's Day, residents of ...
http://blog.law.northwestern.edu/bluhm/2008/11/dna-test-results-suggest-that-jerry-hobbss-confession-is-false.html [Ask, Bluewin, Exalead, Google, Yahoo]
5 Alleged Killer Dad Denied Bond - CBS News
May 11, 2005 ... Jerry Hobbs, who is accused of killing his 9-year-old daughter and her best ... On Wednesday, a judge denied bail for Jerry Hobbs, 34, ...
http://www.cbsnews.com/stories/2005/05/11/national/main694398.shtml [Ask, Bluewin, Yahoo, Google]
6 Jerry Hobbs
http://www.ai.sri.com/~hobbs/ [Bing, Cuil, Google, Yahoo]
7 Jerry Hobbs News
Articles covering Jerry Hobbs: Lake County judge denies bail request for Jerry Hobbs, suspect in ... / Jerry Hobbs Hearing / DNA clears suspect Jerry Hobbs, ...
http://jerry-hobbs-news.newslib.com/ [Ask, Bluewin, Google]
8 Orange Tangerine: What should be done with Jerry Hobbs
May 11, 2005 ... Jerry Hobbs is the rage-filled, domestic-abusing career criminal who killed his 8-year-old daughter and her 9-year-old friend, with scarcely ...
http://orangetangerine.blogspot.com/2005/05/what-should-be-done-with-jerry-hobbs.html [Ask, Bluewin, Exalead, Google]
DBLP: Jerry R. Hobbs
Jerry R. Hobbs, Vladik Kreinovich: Optimal choice of granularity in commonsense estimation: Why half-orders of magnitude? Int. J. Intell. Syst. ...
http://dblp.uni-trier.de/search/author?author=Jerry%20R.%20Hobbs [Ask, Bluewin, Google]
11

Today

Carrot Search

http://search.carrot-search.com/carrot2-webapp/search

Jerry Hobbs

Tree | Visualization

- All Topics (100)
- Jerry R. Hobbs (25)
  - Discourse Analysis (5)
  - David (5)
  - Feng (4)
  - Semantic (4)
  - DBLP (4)
  - Scientific Commons (3)
  - Srini Narayanan (3)
  - Science (3)
  - Microsoft Academic Search (2)
  - Formal Theories of the Commonsense World (2)
  - more | show all
- John Hobbs (9)
- Murder (9)
- David Martin (5)
- Denied Bond (4)
- Music (4)
- Videos (4)
- Laura Hobbs (3)
- Facebook (3)
- 8-year-old Daughter (3)
- LinkedIn (3)
- Public (4)
- Lake County Judge (3)
- Media (3)
- Roy Jerry Hobbs (2)
- Own Daughter (2)
- Mathematics Genealogy Project (2)
- Daml-time Mailing List Archive (2)
- Information Extraction (2)
- Information Sciences Institute (2)
- Coherence (2)
- Abduction (2)
- Finite-state (2)
- References (2)
- Write Fiction (2)
- Design (2)
- Interpretation (2)
- Other topics (14)

Cluster **Jerry R. Hobbs** with **25** documents (search for more like this)

1 Jerry R. Hobbs
Oct 14, 2003 ... **Jerry** R. **Hobbs**. Address: USC/ISI 4676 Admiralty Way ... **Jerry** R. **Hobbs hobbs**@isi .edu. USC/ISI, 4676 Admiralty Way, Marina del Rey, CA 902
http://www.isi.edu/~hobbs/ [Ask, Bing, Bluewin, Cuil, Entireweb, Exalead, Google, Yahoo]

3 Jerry Hobbs - Wikipedia, the free encyclopedia
Dr. **Jerry** R. **Hobbs** (born 25 January 1942) is a prominent researcher in the fields of computational linguistics, discourse analysis, and artificial ...
http://en.wikipedia.org/wiki/Jerry_Hobbs [Ask, Bing, Bluewin, Entireweb, Exalead, Wikipedia, Yahoo, Google]

11 DBLP: Jerry R. Hobbs
Jerry R. **Hobbs**, Vladik Kreinovich: Optimal choice of granularity in commonsense estimation: Why half-orders of magnitude? Int. J. Intell. Syst. ...
http://dblp.uni-trier.de/search/author?author=Jerry%20R.%20Hobbs [Ask, Bluewin, Google]

14 Publications in Discourse Analysis
Hobbs, Jerry R., and Rutu Mulkar-Mehta, 2008. ``Using Abduction for Video-Text Coreference", Proceedings, BOEMIE Workshop, Koblenz, Germany, December 20
http://www.isi.edu/~hobbs/discourse-references/discourse-references.html [Ask, Bluewin]

15 Feng Pan's Homepage - Feng Pan
Feng Pan and **Jerry** R. **Hobbs**. 2006. "Temporal Arithmetic Mixing Months and ... **Jerry** R. **Hobbs** and Feng Pan. 2004. "An Ontology of Time for the Semantic We
http://fengpan.net/ [Ask, Bluewin, Google]

18 Time Ontology in OWL [Jerry R. Hobbs, Feng Pan]
Jerry R. **Hobbs**, Feng Pan. **Jerry** R. **Hobbs**, Feng Pan, Time Ontology in OWL, World Wide Web Consortium, Working Draft WD-owl-time-20060927, September 20
http://dret.net/biblio/reference/owltime [Ask, Cuil]

22 Guest Speaker: Jerry R. Hobbs
Bio: Dr. **Jerry** R. **Hobbs** is a prominent researcher in the fields of computational linguistics, discourse analysis, and artificial intelligence. ...
http://www.cs.umn.edu/medlang/speakers/hobbs.html [Ask, Bluewin]

23 AI Center :: People
SRI International's Artificial Intelligence Center (AIC) is one of the world's ... Pan, Feng and **Hobbs**, **Jerry** R. Temporal Aggregates in OWL-Time, in Proceedings, ...
http://www.ai.sri.com/people/hobbs [Bing, Yahoo]

29 DBLP: Jerry R. Hobbs
Coauthor Index
http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/h/Hobbs:Jerry_R=.html [Bing]

34 Scientific Commons: Jerry R. Hobbs

12

Today

INTELIUS people search

Jerry Hobbs

People Search | Advanced »

Find hidden pictures and profiles of all your friends. Click Here

Intelius Search Results | Public Records Search | Email Search by Name | Background Check

http://search.intelius.com/

U.S. Census Bureau states 90,000 names are shared by 100,000,000 people

More details for **Jerry Hobbs**: Find Email Hidden Profiles Address History

Jerry Hobbs male, 68 years old
Yale University, menlo park, abduction, SRI International, Information Sciences Institute, Knowledge representation... Jerry R. Hobbs
Dr. Jerry R. Hobbs (born 25 January 1942) is a prominent researcher in the fields of computational linguistics, discourse analysis, and artificial...

Jerry male, 58 years old (Marina del Rey, California, United States)
single, Aquarius

Jerry Hobbs female, 20 years old (Scottsboro, Alabama, United States)
single, high school, Pisces, White / Caucasian, A few extra pounds, straight
'Remove Profile'.

Jerry male, 35 years old (Homestead, Florida, United States)
single, security officer, gay, virgo, Grad / professional school, White / Caucasian
I am easy going I like to drive fast meet friendly people and hang-out I am looking for friends i can hangout with and go places with so let me know.

Jerry Hobbs male, 42 years old (Las Vegas, Nevada, United States)
single, gemini, slim, Grad / professional school, White / Caucasian, From Las Vegas
Born to race!

Jerry Hobbs male, 26 years old (Grove City, Pennsylvania, United States)
single, leo

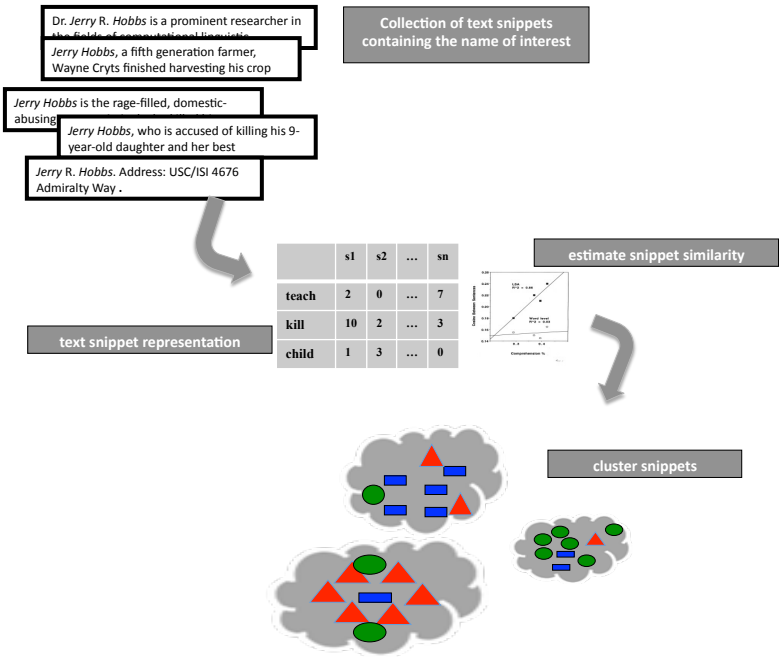
Jerry Hobbs male (Orlando, Florida, United States)

Jerry Hobbs male, 23 years old (United States)
gemini, in a relationship, straight, likes Not a big fan of them (book), likes look at interests music, likes only adult swim (TV show)
i am a high person (that is about my personality). not in to homosexual males i will cus fags out. lezbians are f***ing amsome.

Jerry Hobbs female (United Kingdom)
voluptuous, Sagittarius, White / Caucasian, in a relationship, straight, From BASILDON
Hi well my name is jerry, im 23 and live in good old basildon, essex. i have a partner of 2 years and a little boy who is 8 months old, im on here...

Jerry Hobbs male (Bloomington, Indiana, United States)
single, Libra

13



14

Text Snippet Representation

- The context of each snippet is represented by a vector with *k* dimensions
- Each dimension indicates whether a particular feature occurred in the context
  - the value can be binary, a frequency count etc.
- The features capture the characteristics of the context to be clustered
- Intuitively, vectors/contexts that share the same features will be similar to each other

15

Contexts (input text snippets)

- Cnt1: Dr. ***Jerry* R. *Hobbs*** (born 25 January 1942) is a prominent researcher in the fields of computational linguistics, discourse analysis, and artificial
- Cnt2: ***Jerry Hobbs*** is the rage-filled, domestic-abusing career criminal who killed his 8-year-old daughter and her 9-year-old friend, with scarcely **...**
- Cnt3: ***Jerry Hobbs***, Author. A fifth generation farmer, Wayne Cryts finished harvesting his crop in the fall of 1980 and hauled more than 32000 bushels of soybeans **...**
- Cnt4: ***Jerry Hobbs***, who is accused of killing his 9-year-old daughter and her best **...** On Wednesday, a judge denied bail for ***Jerry Hobbs***, 34, **...**

16

Text Snippet Features (1)

- Unigram – a single word that occurs more than a given number of times

binary values

	kill	artificial	researcher	...	daughter
Cnt1:	0	1	1		0
Cnt2:	1	0	0		1
Cnt3:	0	0	0		0
Cnt4:	1	0	0		1

17

Text Snippet Features (1)

- Unigram – a single word that occurs more than a given number of times

- kill 1000
- artificial 500
- researcher 200

…

- daughter 100

frequency estimated from corpus

frequency values

	kill	artificial	researcher	…	daughter
Cnt1:	0	500	200		0
Cnt2:	1000	0	0		100
Cnt3:	0	0	0		0
Cnt4:	1000	0	0		100

18

Text Snippet Features (2)

- Bigram– an ordered pair of words that occur together more often than expected by chance

binary values

	kill his	prominent researcher	criminal who	…	8-year-old daughter
Cnt1:	0	1	0		0
Cnt2:	1	0	1		1
Cnt3:	0	0	0		0
Cnt4:	1	0	0		1

19

Text Snippet Features (2)

- Bigram– an ordered pair of words that occur together more often than expected by chance

- kill his 21.2
- prominent researcher 102.9
- criminal who 68.5

…

- 8-year-old daughter 35.9

$-\log P(w_1 \mid w_0)$, log-likelihood scores based on frequency estimated from corpus

frequency weights

	kill his	prominent researcher	criminal who	…	8-year-old daughter
Cnt1:	0	102.9	0		0
Cnt2:	21.2	0	68.5		35.9
Cnt3:	0	0	0		0
Cnt4:	21.2	0	0		35.9

20

Underlying Premise*

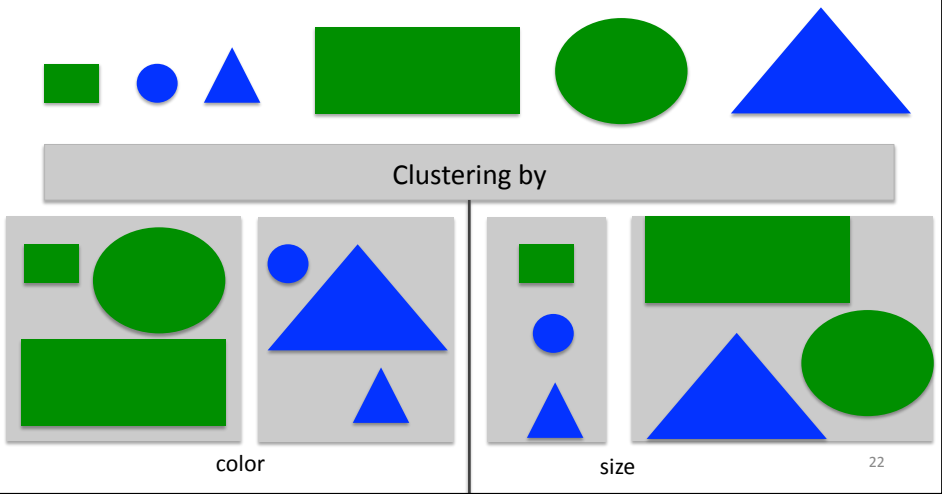
- You shall know a word by the company it keeps
  - Firth, 1957 (Studies in Linguistic Analysis)
- Meanings of words are determined by their distributional patterns (Distributional Hypothesis)
  - Harris, 1968 (Mathematical Structures of Language)
- Words that occur in similar contexts will have similar meanings (Strong Contextual Hypothesis)
  - Miller and Charles, 1991 (Language and Cognitive Processes)

* This slide is adapted from a tutorial of Ted Pedersen

21

Clustering

- Clustering is the process of grouping a set of objects into classes of similar objects



22

Text Snippet Clustering

- group text snippets by similar meaning
- snippet similarity is calculated as $sim(Cnt_1, Cnt_2) = \sum_{i=1}^{n} w_{1i} * w_{2i}$

	kill	artificial	researcher	daughter
Cnt1:	0	1	1	0
Cnt2:	1	0	0	1
Cnt3:	0	0	0	0
Cnt4:	1	0	0	1

sim(Cnt1,Cnt2)=(0*1)+(1*0)+(1*0)+(0*1)=0

sim(Cnt1,Cnt3)=(0*0)+(1*0)+(1*0)+(0*0)=0

sim(Cnt1,Cnt4)=(0*1)+(1*0)+(1*0)+(0*1)=0

sim(Cnt2,Cnt3)=(1*0)+(0*0)+(0*0)+(0*0)=0

sim(Cnt2,Cnt4)=(1*1)+(0*0)+(0*0)+(1*1)=2

sim(Cnt3,Cnt4)=(0*1)+(0*0)+(0*0)+(0*1)=0

23

Hierarchical Clustering

Agglomerative or bottom-up

- begin with each element as a separate cluster
- merge clusters into successively large cluster
- repeat until one cluster is left

Divisive or top-down

- begin with all elements in a whole cluster
- divide clusters into successively smaller cluster
- repeat until all elements are in singleton clusters

24

Cluster Proximity Estimate

- Single-Link
  - Nearest Neighbor: the closest members
- Complete-Link
  - Furthest Neighbor: the furthest members
- Average-Link
  - Average of all cross cluster pairs
- Centroid
  - Centers of gravity

25

Partitioning Clustering

- Constructs a partition of *n* objects into a set of K clusters
- K-means algorithm:

Input: Desired number of clusters, *k*

Initialize: the *k* cluster centers (random if necessary)

Iterate:

1. Decide the class memberships of the N objects by assigning them to the nearest cluster centroids (mean)
2. Re-estimate the *k* clusters, by assuming the membership found above are correct

$$\vec{\mu}_k = \frac{1}{c_k} \sum_{i \in C_k} \vec{x}_i$$

Terminate:

If none of the N objects changed membership in the last iteration, exit

26

Final Output

- A set of clusters containing a certain number of *text snippets, i.e.* small text fragments
- For each cluster assign cluster labels:
  - top 10 most significant unigrams/bigrams of each cluster act as a descriptive label
  - top 10 most unique unigrams/bigrams for each cluster act as discriminating label

27

Cluster Evaluation

- Internal criterion
  - intra-class high similarity
  - inter-class low similarity
  - the quality depends on the object representation and the similarity measure used
- External criterion (clustering quality)
  - measure the ability to discover the named entity groups in the gold standard data
  - asses the clustering with respect to ground truth

28

Web People Search Challenge

- The first challenge was organized in 2007
- WePS focuses on person and organization name disambiguation of Web pages
- For each ambiguous name, the system must return the documents and the attributes which are relevant for the different senses of the name
- There is an upcoming challenge on 1st of July 2010
- More information at: http://nlp.uned.es/weps/

29

Name Discrimination Demo

- SenseClusters by Ted Pedersen
  http://marimba.d.umn.edu/cgi-bin/SC-cgi/index.cgi

- The software can be used for:
  - proper name discrimination
  - word sense discrimination
  - e-mail clustering
  - synonym finding

30

Thoughts on

- What else we can use machine learning
  - e-mail classification (spam vs. non-spam)
  - product reviews (useful vs. non-useful)
  - emotion classification of text (anger vs. happiness vs. joy vs. disgust)
  - …
- What else we can use clustering for
  - e-mail, document organization by similar topics
  - grouping flickr images based on similar label tags
  - generating adds for similar documents
  - …

31