CS544: Named Entity Discrimination

January 27, 2011

Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva
Who is Jerry Hobbs?

Jerry R. Hobbs. Address: USC/ISI 4676 Admiralty Way ... Jerry R. Hobbs hobbs@isi.edu

AMW.com Jerry Hobbs-Fugitive America's Most Wanted is a long-running American TV
Alleged Killer Dad Denied Bond, Services Set for Two Little Girls. A judge had denied

Jerry Hobbs- Wikipedia, the free encyclopedia Dr. Jerry R. Hobbs (born 25 January 1941)
World Martial Arts Games Rank: 3rd Dan, Discipline: Ju-Jutsu Issued By: Jerry Hobbs
10th Dan Rank: 1st
NE Recognition vs. NE Discrimination

• NE Recognition = detection & classification of entity mentions into a predefined set of categories.

⇒ achieves only a partial disambiguation of names

• NE Discrimination = finding the actual entity denoted by a particular name occurrence in text.
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 ... www.isi.edu/~hobbs/ - Cached - Similar

AMW | Fugitives | Jerry Hobbs | Brief
May 10, 2005 ... Fugitives | Jerry Hobbs | Brief - Father Denied Bail Awaits Trial For Children's Murders Jerry Brandon Hobbs accused of stabbing the 4-year-old boys to death. www.amw.com/fugitives/brief.cfm?id=31767 - Cached - Similar

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 intelligence. en.wikipedia.org/wiki/Jerry_Hobbs - Cached - Similar

Image results for Jerry Hobbs - Report images

Bluhm Blog: DNA test results suggest that Jerry Hobbs's confession is false... 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 Chicago got a shock when the FBI announced it had... blogs.northwestern.edu/.../dna-test-results-suggest-that-jerry-hobbs-confession-is-false.html - Cached - Similar

Alleged Killer Dad Denied Bond - CBS News
May 11, 2005 ... Jerry Hobbs, who is accused of murdering his 9-year-old daughter and her best friend, on Wednesday, an judge denied bail for Jerry Hobbs, 34, ... www.cbsnews.com/stories/2005/05/11/...main694398.shtml - Cached - Similar

Jerry Hobbs
www.ai.sri.com/~hobbs/ - Similar

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, ... Jerry-hobbs-news.newslib.com/ - Cached - Similar

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 a thought... orangetangerine.blogspot.com/.../what-should-be-done-with-jerry-hobbs.html - Cached - Similar

Wayne Cryts - One Man With Courage
Jerry Hobbs, Author. A fifth generation farmer, Wayne Cryts finished harvesting his crop the night of the murder. Wayne Cryts finished harvesting his crop in the morning. A man named William holds a book of soybeans ... www.waynecryts.com/ - Cached - Similar

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. ... www.cs.umn.edu/medlang/speakers/hobbs.html - Cached

Ideally, we want

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

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

Jerry Hobbs is the rage-filled, domestic-abusing career criminal who killed his 8-year-old daughter and her best friend... Jerry Hobbs, who is accused of killing his 8-year-old daughter and her best friend...

Jerry Hobbs, a fifth generation farmer, Wayne Cryts finished harvesting his crop... Wayne Cryts finished harvesting his crop in the morning. A man named William holds a book of soybeans...
U.S. Census Bureau states 90,000 names are shared by 100,000,000 people
The first debate was in Oxford.
The first debate was in Oxford. A draw for Obama would be considered a win.
He is seen as a national hero by those who live in Georgia after the independence from the USSR.
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
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
Why is it called “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

**Jerry Hobbs**

- group 1
- group 2
- group 3
Today

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

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

1. Jerry R. Hobbs

2. Jerry Hobbs - Wikipedia, the free encyclopedia

3. DBLP: Jerry R. Hobbs

4. Publications in Discourse Analysis

5. Feng Pan's Homophone - Feng Pan

6. Time Ontology in OWL

7. 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/mcl/lang/speakers/hobbs.html [Ask, Bluewin]

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

9. DMLP: Jerry R. Hobbs
Clustering
Unsupervised Learning

• Labels must be automatically discovered

• Method: clustering

• Tasks which can be resolved: Named Entity Disambiguation, any type of Text Classification task
Clustering

- Are there any “groups” in the data?
- What is each group?
- How many?
- How to identify them?
What is Clustering?

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

• A common and important task that finds many applications in Science, Engineering among others
  – group genes that perform the same function
  – group individuals that has similar political view
  – identify similar objects from pictures
  – categorize documents of similar topics
  – disambiguate named entities (in our case)
Issues in Clustering

• What is a natural grouping among these objects?
• What makes objects “related”?
  – "similarity/distance” metric
• Representation for objects
  – through vector space
• How many clusters?
  - fixed a-priory
  - completely data driven
  - avoid “trivial” clusters - too large or small

• Clustering Algorithms
  – Partition algorithms
  – Hierarchical algorithms
What is the natural grouping among these objects?
What is the natural grouping among these objects?

Clustering by

- Color
- Size
What is Similarity?

• Hard to define, but we know it when we see it.
• Easier to think in terms of the distance between vectors
Properties of distance measure

- $D(A,B) = D(B,A)$  \textit{Symmetry}

- $D(A,A) = 0$  \textit{Constancy of Self-Similarity}

- $D(A,B) = 0$ if $A = B$  \textit{Positivity Separation}

- $D(A,B) \leq D(A,C) + D(B,C)$  \textit{Triangular Inequality}
Properties of distance measure

• $D(A,B) = D(B,A)$  \textit{Symmetry}

Otherwise you could claim that “Alex looks like Bob, but Bob looks nothing like Alex”

• $D(A,A) = 0$  \textit{Constancy of Self-Similarity}

• $D(A,B) = 0$ if $A = B$  \textit{Positivity Separation}

• $D(A,B) \leq D(A,C) + D(B,C)$  \textit{Triangular Inequality}
Properties of distance measure

- \( D(A,B) = D(B,A) \) \hspace{1cm} \textit{Symmetry}  
  Otherwise you could claim that “Alex looks like Bob, but Bob looks nothing like Alex”

- \( D(A,A) = 0 \) \hspace{1cm} \textit{Constancy of Self-Similarity}  
  Otherwise you could claim that “Alex looks more like Bob, than Bob does”

- \( D(A,B) = 0 \) if \( A = B \) \hspace{1cm} \textit{Positivity Separation}

- \( D(A,B) \leq D(A,C) + D(B,C) \) \hspace{1cm} \textit{Triangular Inequality}
Properties of distance measure

• \( D(A,B) = D(B,A) \) \textit{Symmetry}
  Otherwise you could claim that “Alex looks like Bob, but Bob looks nothing like Alex”

• \( D(A,A) = 0 \) \textit{Constancy of Self-Similarity}
  Otherwise you could claim that “Alex looks more like Bob, than Bob does”

• \( D(A,B) = 0 \) if \( A= B \) \textit{Positivity Separation}
  Otherwise there are objects in your world that are different, but you cannot tell apart

• \( D(A,B) \leq D(A,C) + D(B,C) \) \textit{Triangular Inequality}
Properties of distance measure

• $D(A, B) = D(B, A)$  \textit{Symmetry}
  Otherwise you could claim that “Alex looks like Bob, but Bob looks nothing like Alex”

• $D(A, A) = 0$  \textit{Constancy of Self-Similarity}
  Otherwise you could claim that “Alex looks more like Bob, than Bob does”

• $D(A, B) = 0$ if $A = B$  \textit{Positivity Separation}
  Otherwise there are objects in your world that are different, but you cannot tell apart

• $D(A, B) \leq D(A, C) + D(B, C)$  \textit{Triangular Inequality}
  Otherwise you could claim that “Alex is very like Bob and Alex is very like Carl, but Bob is very unlike Carl”
Distance measures

• Given two objects \( x \) and \( y \) both with \( n \) values

\[
x = (x_1, x_2, \ldots, x_n)
\]
\[
y = (y_1, y_2, \ldots, y_n)
\]

calculate the Euclidean distance as

\[
d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}
\]
Bottom-up Clustering

- Begin with each element in a separate cluster
- Merge clusters into successively large cluster
- Repeat until one cluster is left

Recommended reading:
Chapter 14 on Clustering from the book of Manning & Schütze
Top-down Clustering

- Begin with all elements in a whole cluster
- Divide clusters into successively smaller cluster
- Repeat until all elements are in singleton clusters

Recommended reading:
Chapter 14 on Clustering from the book of Manning & Schütze
Cluster Proximity Estimate

• Single-Link
  – Nearest Neighbor: the closes members

• Complete-Link
  – Furthest Neighbor: the furthest members

• Centroid
  – Centers of gravity
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
Back to NE Discrimination
Dr. Jerry R. Hobbs (born 25 January 1942) is a prominent researcher in the fields of computational linguistics, discourse analysis, and artificial intelligence.

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 ...
Problem Formulation

• 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
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
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:**
  – *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 ....
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
Dr. Jerry R. Hobbs is a prominent researcher in the fields of computational linguistics.

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

Jerry Hobbs is the rage-filled, domestic-abusing.

Jerry Hobbs, who is accused of killing his 9-year-old daughter and her best.


teach | s1 | s2 | ... | sn
--- | --- | --- | --- | ---
teach | 2 | 0 | ... | 7
kill | 10 | 2 | ... | 3
child | 1 | 3 | ... | 0

Snippet similarity

Cluster snippets
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, 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
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, ...
Text Snippet Features (1)

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

<table>
<thead>
<tr>
<th></th>
<th>kill</th>
<th>artificial</th>
<th>researcher</th>
<th></th>
<th>daughter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cnt1:</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Cnt2:</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Cnt3:</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Cnt4:</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
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
Text Snippet Features (2)

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

<table>
<thead>
<tr>
<th></th>
<th>kill his</th>
<th>prominent researcher</th>
<th>criminal who</th>
<th>...</th>
<th>8-year-old daughter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cnt1:</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Cnt2:</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Cnt3:</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Cnt4:</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
**Text Snippet Features (2)**

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

<table>
<thead>
<tr>
<th></th>
<th>kill his</th>
<th>prominent researcher</th>
<th>criminal who</th>
<th>...</th>
<th>8-year-old daughter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cnt1:</strong></td>
<td>0</td>
<td>102.9</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td><strong>Cnt2:</strong></td>
<td>21.2</td>
<td>0</td>
<td>68.5</td>
<td></td>
<td>35.9</td>
</tr>
<tr>
<td><strong>Cnt3:</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Cnt4:</strong></td>
<td>21.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>35.9</td>
</tr>
</tbody>
</table>

\[-\log P(w_1 | w_0), \text{ log-likelihood scores based on frequency estimated from corpus}\]

**frequency weights**
Text Snippet Clustering

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

<table>
<thead>
<tr>
<th></th>
<th>kill</th>
<th>artificial</th>
<th>researcher</th>
<th>daughter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cnt1:</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>sim(Cnt1, Cnt2) = (0<em>1)+(1</em>0)+(1<em>0)+(0</em>1) = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cnt2:</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sim(Cnt1, Cnt3) = (0<em>0)+(1</em>0)+(1<em>0)+(0</em>0) = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sim(Cnt1, Cnt4) = (0<em>1)+(1</em>0)+(1<em>0)+(0</em>1) = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cnt3:</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sim(Cnt2, Cnt3) = (1<em>0)+(0</em>0)+(0<em>0)+(0</em>0) = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sim(Cnt2, Cnt4) = (1<em>1)+(0</em>0)+(0<em>0)+(1</em>1) = 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sim(Cnt3, Cnt4) = (0<em>1)+(0</em>0)+(0<em>0)+(0</em>1) = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cnt4:</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sim(Cnt3, Cnt4) = (0<em>1)+(0</em>0)+(0<em>0)+(0</em>1) = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Final Output

• Each cluster consists of a certain number of text snippets, i.e. small text fragments

• For each cluster assign cluster labels:
  – top 10 most significant bigrams of each cluster act as a descriptive label
  – top 10 most unique bigrams for each cluster act as discriminating label
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 for clustering quality
  – measure the ability to discover the named entity groups in the gold standard data
  – assess the clustering with respect to ground truth
What other features can one use?
Ideas for more features if we use structured text like Wikipedia
Ideas for more features if we use structured text like Wikipedia
Ideas for more features if we use structured text like Wikipedia
Ideas for more features if we use structured text like Wikipedia
Ideas for more features if we use structured text like Wikipedia
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

• Last such challenge was on 1st of July 2010

• More information at:  http://nlp.uned.es/weps/
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
What would you do with clustering?
What would you do with clustering?

- E-mail spam classification
- Organize the documents into multiple categories like Google news
- Trace what two people would talk about
- Sentiment analysis
- Tweet trend identification
- Tweet topic detection
- ...