CS544: Named Entity Discrimination

March 30, 2010

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

Who is Jerry Hobbs?

<table>
<thead>
<tr>
<th>Jerry R. Hobbs</th>
<th>Address: USC/ISI 4676 Admiralty Way</th>
<th>Jerry R. Hobbs <a href="mailto:hobbs@isi.edu">hobbs@isi.edu</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>AMW.com Jerry Hobbs</td>
<td>Fugitive: America's Most Wanted is a long-running American TV show</td>
<td></td>
</tr>
<tr>
<td>Alleged Killer Dad Denied Bond, Services Set for Two Little Girls A judge had denied</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jerry Hobbs</td>
<td>Wikipedia, the free encyclopedia Dr. Jerry R. Hobbs (born 25 January 1942) World Martial Arts Games Rank: 3rd Dan, Discipline: Ju-Jutsu Issued By: Jerry Hobbs 10th Dan Rank: 1st</td>
<td></td>
</tr>
</tbody>
</table>

with Named Entity Recognition

PER

ORG
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.

Ideally, we want

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

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

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

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

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 intelligence. 
  – 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 friend. On Wednesday, a judge denied bail for Jerry Hobbs, 34, ... 
  – Fugitives | Jerry Hobbs - Brief - Father Denied Bail Awaits Trial For Children Murders Jerry Branton Hobbs accused of the stabbing deaths ...

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

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

group 1

group 2

group 3
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

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 Artilles, 2009
Today

Madonna

From Wikipedia, the free encyclopedia

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 (Edward Munch), a painting by Edward Munch

Entertainer

- Madonna (entertainer), the American singer-songwriter-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 bocce 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

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

Today

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, 60 years old

Jerry Hobbs: male, 60 years old (Martha's Vineyard, Massachusetts, United States)

Jerry Hobbs: female, 26 years old (Kissimmee, Florida, United States)

Jerry Hobbs: female, 18 years old

Jerry Hobbs: male, 50 years old (Las Vegas, Nevada, United States)

Jerry Hobbs: male, 40 years old (Orlando, Florida, United States)

Jerry Hobbs: male, 29 years old (United States)

Jerry Hobbs: male, 23 years old (United Kingdom)

Jerry Hobbs: male, 18 years old

Jerry Hobbs: male, 18 years old

Jerry Hobbs: male, 19 years old

Jerry Hobbs: male, 19 years old (United States)
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.
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 intelligence.

• 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 values

<table>
<thead>
<tr>
<th>Cnt1:</th>
<th>kill</th>
<th>artificial</th>
<th>researcher</th>
<th>...</th>
<th>daughter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>500</td>
<td>200</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Cnt2:</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Cnt3:</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cnt4:</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Text Snippet Features (2)

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

binary values

<table>
<thead>
<tr>
<th>Cnt1:</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></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</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>0</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
  
  - kill his 21.2
  - prominent researcher 102.9
  - criminal who 68.5
  - ... 
  - 8-year-old daughter 35.9

  \(-\log P(w_1 | w_2)\), log-likelihood scores based on frequency estimated from corpus

<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>102.9</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Cnt2:</td>
<td>21.2</td>
<td>0</td>
<td>68.5</td>
<td></td>
<td>35.9</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>21.2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>35.9</td>
</tr>
</tbody>
</table>

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
Clustering

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

Text Snippet Clustering

- group text snippets by similar meaning
- snippet similarity is calculated as \( sim(Cnt_{i}, Cnt_{j}) = \sum_{w_{i}} w_{i} * w_{j} \)

<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>Cnt2:</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cnt3:</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cnt4:</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
sim(Cnt_1, Cnt_2) = (0*1) + (1*0) + (1*0) + (0*1) = 0
sim(Cnt_1, Cnt_3) = (0*0) + (1*0) + (1*0) + (0*0) = 0
sim(Cnt_1, Cnt_4) = (0*1) + (1*0) + (1*0) + (0*1) = 0
sim(Cnt_2, Cnt_3) = (1*0) + (0*0) + (0*0) + (0*0) = 0
sim(Cnt_2, Cnt_4) = (1*1) + (0*0) + (0*0) + (1*1) = 2
sim(Cnt_3, Cnt_4) = (0*1) + (0*0) + (0*0) + (0*1) = 0
\]
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

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
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

\[
\bar{\mu}_k = \frac{1}{c_k} \sum_{x_i \in C_k} x_i
\]

Terminate:
If none of the \( N \) objects changed membership in the last iteration, exit

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
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
  - assess the clustering with respect to ground truth

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 1\textsuperscript{st} of July 2010
Name Discrimination Demo

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

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
  - ...