CS544: Named Entity Discrimination

January 31, 2013

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 Al
ged Killer Dad Denied Bond, Services Set for Two Little Girld. A judge had denied

Jerry Hobbs- Wikipedia, the free encyclopedia Dr. Jerry R. Hobbs (born 25 January 19
World Martial Arts Games Rank: 3rd Dan, Discipline: Ju-Jutsu Issued By: Jerry Hobbs
10th Dan Rank: 1st

Search Engine

with Named Entity Recognition
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.
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

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

**Bank**

**meaning 1:**
the slope beside a body of water

**meaning 2:**
depository financial institution

**Jerry Hobbs**

**group 1**

**group 2**

**group 3**
Importance of Name Ambiguity on the Web

• 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
He is seen as a national hero by those who live in Georgia.
Name Ambiguity in Data Bases

U.S. Census Bureau states 90,000 names are shared by 100,000,000 people
How can we solve this problem?
Types of Machine Learning

• Unsupervised Learning
  – correct responses (targets) are not provided
  – the algorithm identifies similarities between the inputs based on something in common

• Method:
  – Clustering

• NLP Tasks:
  – Named Entity Disambiguation, Text Categorization
Clustering

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

Clustering by

- Color
- Size
What is Clustering?

• Clustering is the process of grouping a set of objects into classes of similar objects, without the help of training examples
  – classification vs. clustering
Applications

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

- Define a feature vector to represent the data
  - often called a vector-space model
- Select Features
Feature Representation

• Bag-of-words: each term in a document is a feature of that document
Feature Representation

• TF: term frequency
  • definition: TF = t_{ij}
    – frequency of term i in document j
  • purpose: makes the frequent words for the document more important
Example

<table>
<thead>
<tr>
<th>Chapter1</th>
<th>#terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>@</td>
<td>46</td>
</tr>
<tr>
<td>@@</td>
<td>16</td>
</tr>
<tr>
<td>a</td>
<td>100</td>
</tr>
<tr>
<td>aardvark</td>
<td>1</td>
</tr>
<tr>
<td>ability</td>
<td>3</td>
</tr>
<tr>
<td>able</td>
<td>23</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>coefficient</td>
<td>1</td>
</tr>
<tr>
<td>clusters</td>
<td>10</td>
</tr>
<tr>
<td>with</td>
<td>19</td>
</tr>
<tr>
<td>words</td>
<td>34</td>
</tr>
<tr>
<td>the</td>
<td>200</td>
</tr>
<tr>
<td>zeros</td>
<td>1</td>
</tr>
</tbody>
</table>
## Example

<table>
<thead>
<tr>
<th>Chapter 1</th>
<th>#terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>@</td>
<td>46</td>
</tr>
<tr>
<td>@@</td>
<td>16</td>
</tr>
<tr>
<td>a</td>
<td>100</td>
</tr>
<tr>
<td>aardvark</td>
<td>1</td>
</tr>
<tr>
<td>ability</td>
<td>3</td>
</tr>
<tr>
<td>able</td>
<td>23</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>coefficient</td>
<td>1</td>
</tr>
<tr>
<td>clusters</td>
<td>10</td>
</tr>
<tr>
<td>with</td>
<td>19</td>
</tr>
<tr>
<td>words</td>
<td>34</td>
</tr>
<tr>
<td>the</td>
<td>200</td>
</tr>
<tr>
<td>zeros</td>
<td>1</td>
</tr>
</tbody>
</table>

Many low frequency words

Can we adjust tf?
Feature Representation

• **IDF**: inverted document frequency
  
  • definition: $$\text{IDF} = \log\left(\frac{N}{n_i}\right)$$
    
    – $$n_i$$: number of documents containing term $$i$$
    – $$N$$: total number of documents
  
  • purpose: makes rare words across documents more important
TF.IDF Term Weighting

• TF: term frequency
  • definition: TF = t_{ij}
  – frequency of term i in document j
  • purpose: makes the frequent words for the document more important

• IDF: inverted document frequency
  • definition: IDF = \log(N/n_i)
  – n_i : number of documents containing term i
  – N : total number of documents
  • purpose: makes rare words across documents more important

• TF.IDF (for term i in document j)
  • definition: t_{ij} \times \log(N/n_i)
Other Features

• Dependency Tree

- *find* has two features incremented by +1
  - *subj*: *John*
  - *obj*: *solution*

- *John* has one feature
  - *subj-of:* *find*
Clustering Process

- Quantifies the closeness between the feature vectors of two elements
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

• $D(A,B) = D(B,A)$  \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}

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

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

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

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

• D(A,B) ≤ D(A,C) + D(B,C) **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 Minkowski distance as

$$d(x, y) = \sqrt[p]{\sum_{i=1}^{m} |x_i - y_i|^p}$$

- Euclidean distance

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

- Manhattan distance

$$d(x, y) = \sum_{i=1}^{m} |x_i - y_i|$$
Example

Euclidean distance

\[ \sqrt{4^2 + 3^2} = 5 \]

Manhattan distance

\[ 4 + 3 = 7 \]

reminder buying milk from home to store
Edit Distance

- To measure the similarity between two objects, transform one of the objects into the other, and measure how much effort it took. The measure of effort becomes the distance measure.

The distance between Patty and Selma.

- Change dress color, 1 point
- Change earring shape, 1 point
- Change hair part, 1 point
- \( D(\text{Patty}, \text{Selma}) = 3 \)

The distance between Marge and Selma.

- Change dress color, 1 point
- Add earrings, 1 point
- Decrease height, 1 point
- Take up smoking, 1 point
- Lose weight, 1 point
- \( D(\text{Marge}, \text{Selma}) = 5 \)
Clustering Process

- The actual clustering algorithms
  - **Hierarchical algorithm**, which creates a hierarchical decomposition of the elements
  - **Partitioning algorithm**, which produces a single partitioning by optimizing some criterion
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
Example: Single-Link Method

Euclidean Distance

Distance Matrix
Partitioning Clustering

• Constructs a partition of \( n \) objects into a set of \( K \) clusters

• K-means algorithm:

1. Select \( k \) clusters arbitrarily.
2. Initialize cluster centers with those \( k \) clusters.
3. Do loop
   a) Partition by assigning or reassigning all data objects to their closest cluster center.
   b) Compute new cluster centers as mean value of the objects in each cluster.  
      \[ \mu_k = \frac{1}{c_k} \sum_{i \in c_k} x_i \]

Until no change in cluster center calculation
Example

Task:
Cluster the following objects into two clusters ($k=2$)

Use:
Manhattan distance

\[ d(x, y) = \sum_{i=1}^{m} |x_i - y_i| \]

Randomly initialize the clusters with the first two objects
C1={\{(0,0)\}}
C2={\{(0,1)\}}

Now:
2. Initialize cluster centers.
3a. Calculate the distance between each object and each cluster center, assigning the object to the closest cluster.
3b. Compute new cluster center for each cluster.
Clustering Process

- Evaluation of the produced clustering output
Clustering Evaluation

• Compare the clustering output with a gold standard (manually generated answer keys)

• Embed the clustering output in an application and using its evaluation measure
  • Example: search engine results
Putting Theory into Practice
(back to our NED Example)
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 ....
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 father.

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


<table>
<thead>
<tr>
<th></th>
<th>s1</th>
<th>s2</th>
<th>…</th>
<th>sn</th>
</tr>
</thead>
<tbody>
<tr>
<td>teach</td>
<td>2</td>
<td>0</td>
<td>…</td>
<td>7</td>
</tr>
<tr>
<td>kill</td>
<td>10</td>
<td>2</td>
<td>…</td>
<td>3</td>
</tr>
<tr>
<td>child</td>
<td>1</td>
<td>3</td>
<td>…</td>
<td>0</td>
</tr>
</tbody>
</table>
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 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>0</td>
<td>0</td>
</tr>
<tr>
<td>Cnt2:</td>
<td>1</td>
<td>0</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>
<td>0</td>
</tr>
<tr>
<td>Cnt4:</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
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>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></td>
<td>100</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>1000</td>
<td>0</td>
<td>0</td>
<td></td>
<td>100</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>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

- kill his
- prominent researcher
- criminal who

...  

- 8-year-old daughter

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

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

Frequency weights
Text Snippet Grouping

• group text snippets by similar meaning

• snippet similarity is calculated as \[ \text{sim}(\text{Cnt}_1, \text{Cnt}_2) = \sum_{i=1}^{n} w_{1i} \cdot w_{2i} \]

<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>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>

\[ \text{sim}(\text{Cnt}_1, \text{Cnt}_2) = (0\cdot1)+(1\cdot0)+(1\cdot0)+(0\cdot1)=0 \]
\[ \text{sim}(\text{Cnt}_1, \text{Cnt}_3) = (0\cdot0)+(1\cdot0)+(1\cdot0)+(0\cdot0)=0 \]
\[ \text{sim}(\text{Cnt}_1, \text{Cnt}_4) = (0\cdot1)+(1\cdot0)+(1\cdot0)+(0\cdot1)=0 \]
\[ \text{sim}(\text{Cnt}_2, \text{Cnt}_3) = (1\cdot0)+(0\cdot0)+(0\cdot0)+(0\cdot0)=0 \]
\[ \text{sim}(\text{Cnt}_2, \text{Cnt}_4) = (1\cdot1)+(0\cdot0)+(0\cdot0)+(1\cdot1)=2 \]
\[ \text{sim}(\text{Cnt}_3, \text{Cnt}_4) = (0\cdot1)+(0\cdot0)+(0\cdot0)+(0\cdot1)=0 \]
Final Output

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

• The clusters correspond to the different people sharing the same name
  – Cluster1: Jerry Hobbs the researcher
  – Cluster2: Jerry Hobbs the killer
  – Cluster3: Jerry Hobbs the singer
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?

- E-mail clustering by topic
- Organize documents into multiple categories like Google news
- Trace what two people talk on Twitter
- Sentiment analysis
- ...

59