CS544: Information Extraction  
Named Entity Recognition  

January 22, 2013  

Zornitsa Kozareva  
USC/ISI  
Marina del Rey, CA  
kozareva@isi.edu  
www.isi.edu/~kozareva
Extract Jobs from Multiple Sites

foodscience.com-Job2
JobTitle: Ice Cream Guru
Employer: foodscience.com
JobCategory: Travel/Hospitality
JobFunction: Food Services
JobLocation: Upper Midwest
Contact Phone: 800-488-2611
DateExtracted: January 8, 2001
Source: www.foodscience.com/jobs_midwest.html
OtherCompanyJobs: foodscience.com-Job1

If you dream of cold creamy chocolate or ooey-gooey cookie dough, there’s a great opportunity for you to maintain and expand this major corporation’s high-end ice cream business. You’ll be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or Dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship.
Contact: [email]
1-900-488-2611
What is “Information Extraction”?  

• Goal: identify specific pieces of information from the content of unstructured or semi-structured textual documents.

• Input:  
  – scenario of extraction (templates to be filled)  
  – document collection

• Output:  
  – a set of instantiated templates
Applications

• Apartment rental adds
• USC alert system
• Social event announcements
• Seminar announcements
• Conference call for papers
• Company information
• ...
# Google Squared

http://www.google.com/squared

<table>
<thead>
<tr>
<th>Item Name</th>
<th>Image</th>
<th>Description</th>
<th>Date Of Birth</th>
<th>Nationality</th>
<th>Date Of Death</th>
<th>Died</th>
</tr>
</thead>
<tbody>
<tr>
<td>Franz Boas</td>
<td><img src="image" alt="Franz Boas" /></td>
<td>Franz Boas (July 9, 1858 – December 21, 1942) was a German American anthropologist and a pioneer of modern anthropology.</td>
<td>July 9, 1858</td>
<td>American</td>
<td>December 21, 1942</td>
<td>22-Dec-1942</td>
</tr>
<tr>
<td>Edward Sapir</td>
<td><img src="image" alt="Edward Sapir" /></td>
<td>Edward Sapir (pronounced /seɪpr/) (January 26, 1884 – February 4, 1939) was a German-born American linguist.</td>
<td>January 26, 1884</td>
<td>American</td>
<td>February 4, 1939</td>
<td>4-Feb-1939</td>
</tr>
<tr>
<td>Leonard Bloomfield</td>
<td><img src="image" alt="Leonard Bloomfield" /></td>
<td>Leonard Bloomfield (April 1, 1887 – April 18, 1949) was an American linguist who led the development of structural linguistics in the United States.</td>
<td>April 1, 1887</td>
<td>American</td>
<td>April 18, 1949</td>
<td>April 18, 1949</td>
</tr>
<tr>
<td>Jonathan Edwards</td>
<td><img src="image" alt="Jonathan Edwards" /></td>
<td>Jonathan Edwards (May 20, 1878 – May 3, 1962) was an American linguist.</td>
<td>October 5, 1700</td>
<td>British</td>
<td>March 22, 1758</td>
<td>March 22, 1758</td>
</tr>
<tr>
<td>Benjamin Lee Whorf</td>
<td><img src="image" alt="Benjamin Lee Whorf" /></td>
<td>Benjamin Lee Whorf (April 24, 1897 – July 26, 1941) was an American linguist.</td>
<td>April 24, 1897</td>
<td>American</td>
<td>July 26, 1941</td>
<td>July 26, 1941</td>
</tr>
<tr>
<td>Donna Jo Napoli</td>
<td><img src="image" alt="Donna Jo Napoli" /></td>
<td>Donna Jo Napoli writes for all ages, from picture books through young adult books. Her awards include the Lowell Foundation Grant and the National Book Award.</td>
<td>02/28/1948</td>
<td>American</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joseph Greenberg</td>
<td><img src="image" alt="Joseph Greenberg" /></td>
<td>Joseph Harold Greenberg (May 28, 1915 – May 7, 2001) was a prominent and controversial American linguist, especially known for his work on American English.</td>
<td>May 28, 1915</td>
<td>American</td>
<td>May 7, 2001</td>
<td>May 7, 2001</td>
</tr>
<tr>
<td>Hans Kurath</td>
<td><img src="image" alt="Hans Kurath" /></td>
<td>Hans Kurath (13 December 1891–2 January 1992) was an American linguist of Austrian origin. He was full professor for English in the United States.</td>
<td>1891-12-13</td>
<td></td>
<td>2 possible values</td>
<td>4 possible values</td>
</tr>
<tr>
<td>William Labov</td>
<td><img src="image" alt="William Labov" /></td>
<td>An entry to appear in the Cambridge Encyclopedia of the Language Sciences, summarizing the work of a prominent linguist.</td>
<td>December 4, 1927</td>
<td></td>
<td></td>
<td>1 possible value</td>
</tr>
</tbody>
</table>
Sunday, February 24
The Academy Awards ceremony (The Oscars) 2013

News for oscars 2013

Oscars 2013: 5 facts about Emmanuelle Riva, Oldest Best Lead Actress nominee
OnTheRedCarpet.com - 17 hours ago
Out five facts about 'Amour's Emmanuelle Riva, 85, the oldest actress to be nominated for a Lead Actress Oscar.

Trailers For Oscars 2013 Best Picture Nominees - What's Your Pick?
OK! Magazine - 23 hours ago
OSCARS: Jennifer Lawrence And Jessica Chastain Stir Things Up In Tight Best Act...
Deadline.com - 1 day ago

85th Academy Awards Nominees, February 24th, 2013
oscar.go.com
Best Picture: Amour, Argo, Beasts of the Southern Wild, Django Unchained, Les Misérables, Life of Pi, Lincoln, Silver Linings Playbook, Zero Dark Thirty

Best Actor in a Leading Role: Bradley Cooper, Daniel Day-Lewis, Hugh Jackman, Joaquin Phoenix, Denzel Washington

Best Actress in a Leading Role: Jessica Chastain, Jennifer Lawrence, Emmanuelle Riva, Quvenzhané Wallis, Naomi Watts

Best Director: Amour (Michael Haneke), Beasts of the Southern Wild (Benh Zeitlin), Life of Pi (Ang Lee), Silver Linings Playbook (David O. Russell), Lincoln (Steven Spielberg)
Factual

http://www.factual.com/

<table>
<thead>
<tr>
<th>Video Game</th>
<th>Platform</th>
<th>Publisher</th>
<th>Developer</th>
<th>Genre</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 on 3 NHL Arcade</td>
<td>Xbox 360</td>
<td>EA Sports</td>
<td>EA Canada</td>
<td>Sports</td>
<td>Hockey</td>
</tr>
<tr>
<td>4x4 Evo 2</td>
<td>Xbox</td>
<td>2K Games</td>
<td>Terminal Reality, Inc.</td>
<td>Sports; Driving/Racing</td>
<td>Espionage</td>
</tr>
<tr>
<td>24: The Game</td>
<td>PlayStation 2</td>
<td>2K Games</td>
<td>SCE Studio Cambridge</td>
<td>Action</td>
<td>Action</td>
</tr>
<tr>
<td>25 to Life</td>
<td>PlayStation 2</td>
<td>Eidos Interactive</td>
<td>Avalanche Software LLC; Ritix</td>
<td>Action; Shooter</td>
<td>Crime</td>
</tr>
<tr>
<td>50 Cent: Bulletproof</td>
<td>Xbox</td>
<td>Vivendi Universal</td>
<td>Genuine Games Ltd.</td>
<td>Action; Shooter</td>
<td>Crime</td>
</tr>
<tr>
<td>1942: Joint Strike</td>
<td>Xbox 360</td>
<td>Backbone Entertainment</td>
<td></td>
<td>Coin-Op Classics</td>
<td></td>
</tr>
<tr>
<td>Abuse</td>
<td>PC</td>
<td>Alive Mediasoft; Bungie Studi</td>
<td>Crack dot Com</td>
<td>Action; Shooter</td>
<td>Sci-Fi; Horror</td>
</tr>
<tr>
<td>AC/DC Live: Rock Band</td>
<td>Xbox 360</td>
<td>Electronic Arts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ace Combat 5: The Unsung War</td>
<td>PlayStation 2</td>
<td>Namco</td>
<td>Namco Bandai Games Inc.</td>
<td>Action; Simulation; Flight Sim</td>
<td>Modern Military</td>
</tr>
<tr>
<td>Ace Combat X: Skies of Deception</td>
<td>PlayStation Portable</td>
<td>Namco Games</td>
<td>Namco Bandai Games Inc.</td>
<td>Action; Simulation; Flight Sim</td>
<td>Modern Military</td>
</tr>
<tr>
<td>Ace Combat Zero: The Belkan Waig</td>
<td>PlayStation 2</td>
<td>Namco</td>
<td>Namco Bandai Games Inc.</td>
<td>Action; Simulation; Flight Sim</td>
<td>Modern Military</td>
</tr>
<tr>
<td>Activision Anthology</td>
<td>PlayStation 2</td>
<td>Activision</td>
<td></td>
<td>Action; Strategy; Sports; Drivi</td>
<td></td>
</tr>
<tr>
<td>Act of War: Direct Action</td>
<td>PC</td>
<td>GFI Russia; Atari, Inc.</td>
<td>Eugen Systems</td>
<td>Strategy; Real-Time Strategy</td>
<td>Modern Military</td>
</tr>
<tr>
<td>Advance Guardian Heroes</td>
<td>Game Boy Advance</td>
<td>?UBI Soft</td>
<td>Treasure</td>
<td>Action; Fighting</td>
<td></td>
</tr>
<tr>
<td>Advance Wars: Days of Ruin</td>
<td>Nintendo DS</td>
<td>Nintendo</td>
<td>Intelligent Systems Co., Ltd.</td>
<td>Strategy</td>
<td>Post-Apocalyptic</td>
</tr>
<tr>
<td>Advance Wars: Dual Strike</td>
<td>Nintendo DS</td>
<td>Nintendo</td>
<td>Intelligent Systems Co., Ltd.</td>
<td>Strategy</td>
<td>Modern Military</td>
</tr>
<tr>
<td>Advent Rising</td>
<td>PC</td>
<td>Majesco Entertainment</td>
<td>GlyphX Games, LLC</td>
<td>Action; Shooter; Adventure</td>
<td>Sci-Fi</td>
</tr>
<tr>
<td>Advent Rising</td>
<td>Xbox</td>
<td>Majesco Entertainment</td>
<td>GlyphX Games, LLC</td>
<td>Action; Shooter; Adventure</td>
<td>Sci-Fi</td>
</tr>
<tr>
<td>Aegis Wing</td>
<td>Xbox 360</td>
<td>Microsoft Game Studios</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aeon Flux</td>
<td>Xbox</td>
<td>Majesco Sales Inc.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afro Samurai</td>
<td>Xbox 360</td>
<td>?Namco</td>
<td>Namco Bandai Games Inc.</td>
<td>Platformer; Action</td>
<td>Martial Arts; Ani</td>
</tr>
<tr>
<td>Age of Booty</td>
<td>Xbox 360</td>
<td>Certain Affinity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age of Empires III</td>
<td>PC</td>
<td>Microsoft Game Studios; Mac</td>
<td>Ensemble Studios</td>
<td>Strategy; Real-Time Strategy</td>
<td>Alternate Histori</td>
</tr>
<tr>
<td>Age of Empires: The Age of Kings</td>
<td>Nintendo DS</td>
<td>?Majesco Sales Inc.</td>
<td>Backbone Entertainment</td>
<td>Strategy</td>
<td>Fantaşy</td>
</tr>
</tbody>
</table>
Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.
IE with Single or Multiple Documents

- Single-document IE
  - extract facts from a specific document
  - cares what is reported in this particular story
  - usually only one opportunity to find a piece of information
  - must deal with the diversity of language

- Multiple-document IE
  - extract facts from a collection of documents (i.e. Web)
  - cares about the facts and does not care where the information comes from
  - has more opportunities to find the information
Genesis of IE

• DARPA funded IE in early to mid 1990’s

• Message Understanding Conference (MUC) was an annual competition where systems were evaluated

• Focused on extracting information from news articles, which are of interest to the intelligence community (CIA, NSA)
Message Understanding Conference

• Domains
    • Messages about naval operations
  – MUC-3 (1991) and MUC-4 (1992)
    • News articles about terrorist attacks
    • News articles about joint ventures and microelectronics
  – MUC-6 (1995)
    • News articles about management changes
  – MUC-7 (1997)
    • News articles about space vehicle and missile launches
Message Understanding Conference

• Types of information that must be extracted:

  – Named Entities
    • Person, Organization, Location names
  – Co-reference
    • Clinton <-> President Bill Clinton
  – Template element
    • Perpetrator, Target
  – Template relation
    • Incident
Today

• Named Entity Recognition

• Multi-class classification
  – Decision trees
Named Entity Recognition

Adam Smith works for IBM, London since February 2010.

- Identify mentions in text and classify them into a predefined set of categories of interest:
  - Person: Adam Smith
  - Organizations: IBM
  - Locations: London
  - Date: February 2010
United States presidential election of 2008, scheduled for Tuesday November 4, 2008, will be the 56th consecutive quadrennial United States presidential election and will select the President and the Vice President of the United States. The Republican Party has chosen John McCain, the senior United States Senator from Arizona as its nominee; the Democratic Party has chosen Barak Obama, the junior United States Senator from Illinois, as its nominee.
Types of Machine Learning

• Supervised Learning
  – labeled training examples with correct responses (targets) are provided
  – based on the training set, the algorithm *generalizes* to respond correctly to all possible inputs

• (Some) Methods:
  – Hidden Markov Models, k-Nearest Neighbors, Decision Trees, AdaBoost, SVM

• NLP Tasks:
  – Named Entity recognition, POS tagging, Parsing
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
Types of Machine Learning

• Semi-Supervised Learning
  – small percentage of labeled examples with correct responses are provided, the rest are unlabeled
  – label the unlabeled examples using the labeled ones, add the newly labeled data to the training data set

• Method:
  – Co-training, self-training, active learning

• NLP Tasks:
  – Named Entity Recognition, POS-tagging, Parsing
Multi-Class Classification (Example)

• Named Entity Recognition
  – person, organization, location, miscellaneous name

• Text Categorization by Topic
  – economy, sport, entertainment

• Weather Forecast
  – sunny, foggy, snowy, rainy

• Author Identification
Muti-Class Classification

• **Given**: some data items that belong to one of $N$ possible classes

• **Task**: train a classifier to predict the class for a new data item

• Geometrically: hard
Multi-class Classification
(Some) Multi-class Classification Algorithms

• Linear
  – Decision trees
  – Naïve Bayes

• Non Linear
  – K-nearest neighbors
  – Neural Networks
Linear class separators
(ex: Naïve Bayes)
Non Linear (ex: $k$ Nearest Neighbor)
Linear vs. Non Linear Algorithms

• **Linearly separable data**: if all the data points can be correctly classified by a linear decision boundary

• Linear
  – Advantages: simpler, less parameters
  – Disadvantages: high dimensional data is usually not linearly separable
  – Example: Perceptron

• Non linear
  – Advantages: more accurate
  – Disadvantages: more complicated, more parameters
  – Example: Kernel Methods
Things Students Enjoy Doing

✓ going to pub
✓ watching TV
✓ going to a party
✓ Studying

Build an algorithm that will let you decide what to do each evening without having to think about it every night?

• If you have an assignment due next day, you need to study
• If you feel lazy, then you don’t like going to the pub
• If there is no party, you cannot go to it
Decision Tree on How to Spend the Evening

- **Party?**
  - Yes: Go to Party
  - No:
    - **Deadline?**
      - Urgent: Study
      - Near: Watch TV
      - None: Go To Pub
    - Lazy?
      - Yes: Watch TV
      - No: Study
Decision Trees

• The classifier has a tree structure, where each node is either:
  – a **leaf** indicating the value of the target attribute (class) of examples
  – a **decision** specifying some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test

• An instance $x_p$ is classified by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance
Constructing Decision Trees

• Build a tree in a greedy manner starting at the root

• Choose the most informative feature at each step by computing the entropy \( H(p) = - \sum p_i \log_2 p_i \)

• Estimate how much the entropy of the whole training set would decrease if a particular feature is chosen for the next classification step

\[
\text{Gain}(S,F) = \text{Entropy}(S) - \sum_{f \in \text{values}(F)} \frac{|S_f|}{|S|} \text{Entropy}(S_f)
\]
## Walkthrough Example

<table>
<thead>
<tr>
<th>Set of Examples (S)</th>
<th>Feature (f1)</th>
<th>Feature (f2)</th>
<th>Feature (f3)</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>True</td>
</tr>
<tr>
<td>s2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>False</td>
</tr>
<tr>
<td>s3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>False</td>
</tr>
<tr>
<td>s4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>False</td>
</tr>
</tbody>
</table>

\[
\text{Entropy}(S) = -p_{\text{true}} \log_2 p_{\text{true}} - p_{\text{false}} \log_2 p_{\text{false}}
\]

\[
= -\frac{1}{4}\log_2 \frac{1}{4} - \frac{3}{4}\log_2 \frac{3}{4}
\]

\[
= 0.5 + 0.311 = 0.811
\]
### Walkthrough Example

<table>
<thead>
<tr>
<th>Set of Examples (S)</th>
<th>Feature (f1)</th>
<th>Feature (f2)</th>
<th>Feature (f3)</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>True</td>
</tr>
<tr>
<td>s2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>False</td>
</tr>
<tr>
<td>s3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>False</td>
</tr>
<tr>
<td>s4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>False</td>
</tr>
</tbody>
</table>

\[
\frac{|S_{f_1}|}{|S|} \text{Entropy}(S_{f_1}) = \frac{1}{4} \times \left( -\frac{0}{1} \log_2 \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1} \right) = 0
\]

\[
\frac{|S_{f_2}|}{|S|} \text{Entropy}(S_{f_2}) = \frac{2}{4} \times \left( -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) = \frac{1}{2}
\]

\[
\frac{|S_{f_3}|}{|S|} \text{Entropy}(S_{f_3}) = \frac{1}{4} \times \left( -\frac{0}{1} \log_2 \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1} \right) = 0
\]

\[
\text{Gain}(S, F) = \text{Entropy}(S) - \sum_{f \in \text{values}(F)} \frac{|S_f|}{|S|} \text{Entropy}(S_f)
\]

- **Gain** measures the reduction in entropy after splitting the dataset based on feature **F**.
- **Set of examples**
- **Possible feature**
- **# of members of S that have value f for feature F**
## Walkthrough Example

<table>
<thead>
<tr>
<th>Set of Examples (S)</th>
<th>Feature (f1)</th>
<th>Feature (f2)</th>
<th>Feature (f3)</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>True</td>
</tr>
<tr>
<td>s2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>False</td>
</tr>
<tr>
<td>s3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>False</td>
</tr>
<tr>
<td>s4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>False</td>
</tr>
</tbody>
</table>

\[
Gain(S, F) = 0.811 - (0 + 0.5 + 0) = 0.311
\]
Another Example

- List everything that you have done for the past few days to get a decent dataset

<table>
<thead>
<tr>
<th>Deadline?</th>
<th>Is there a party?</th>
<th>Lazy?</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urgent</td>
<td>Yes</td>
<td>Yes</td>
<td>Party</td>
</tr>
<tr>
<td>Urgent</td>
<td>No</td>
<td>Yes</td>
<td>Study</td>
</tr>
<tr>
<td>Near</td>
<td>Yes</td>
<td>Yes</td>
<td>Party</td>
</tr>
<tr>
<td>None</td>
<td>Yes</td>
<td>No</td>
<td>Party</td>
</tr>
<tr>
<td>None</td>
<td>No</td>
<td>Yes</td>
<td>Pub</td>
</tr>
<tr>
<td>None</td>
<td>Yes</td>
<td>No</td>
<td>Party</td>
</tr>
<tr>
<td>Near</td>
<td>No</td>
<td>No</td>
<td>Study</td>
</tr>
<tr>
<td>Near</td>
<td>No</td>
<td>Yes</td>
<td>TV</td>
</tr>
<tr>
<td>Near</td>
<td>Yes</td>
<td>Yes</td>
<td>Party</td>
</tr>
<tr>
<td>Urgent</td>
<td>No</td>
<td>No</td>
<td>Study</td>
</tr>
</tbody>
</table>
## Decision Trees

<table>
<thead>
<tr>
<th><strong>Pros</strong></th>
<th><strong>Cons</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>+ generate understandable rules</td>
<td>- error prone in multi-class classification</td>
</tr>
<tr>
<td>+ provide a clear indication of which features are most important for classification</td>
<td>and small number of training examples</td>
</tr>
<tr>
<td></td>
<td>- computationally expensive to train</td>
</tr>
<tr>
<td></td>
<td>(need to compare all possible splits; and also because of pruning)</td>
</tr>
</tbody>
</table>

\[O(N\log N)\] tree construction  
\[O(\log N)\] to return particular leaf