Administrative

- Projects
  - evaluation
- writeup
  - reread the writeup section in the final project handout
  - NOT a report of what happened
- common format
  - abstract high-level summary of the paper, including problem, approach, results and take-home message from paper (1 paragraph)
  - introduction: what is the problem, why do we care (cite related papers)
  - algorithm/approach: what is your approach
  - experiments/results: evaluation metric, experimental setup, results and analysis of the results
  - conclusion: a paragraph of two wrapping up

Administrative

- Quiz 4
  - keep up with book reading
- Search
  - uninformed search
    - BFS
    - DFS
  - uniform-cost search
  - depth-limited search
  - iterative deepening search
- informed search
  - greedy-first search
  - A* search
  - completeness, optimality
  - heuristics
  - admissibility
  - graph search vs. tree search

Administrative

- Quiz 4 continued
  - machine translation
    - noisy channel model
  - MT problems
    - preprocessing
    - translation modeling
    - phrase-based model
    - decoding/search
    - parameter evaluation
    - evaluation (BLEU)
Administrivia

- Information retrieval
  - challenges
  - inverted index
  - boolean vs. ranked query
    - tf-idf query
  - phrases/proximity queries
  - pagerank
- Information extraction (Today’s material)

Administrative

- Talk by Joe 12:30 in MBH 505 tomorrow (Wednesday)

Simplification evaluation

<table>
<thead>
<tr>
<th></th>
<th>Grammatical</th>
<th>Meaning</th>
<th>Simplicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>4.7</td>
<td>4.07</td>
<td>2.9</td>
</tr>
<tr>
<td>Phrase</td>
<td>4.5</td>
<td>4.23</td>
<td>2</td>
</tr>
<tr>
<td>Simple Wiki</td>
<td>4.2</td>
<td>3.73</td>
<td>2.73</td>
</tr>
</tbody>
</table>

Reasonable correlation: 0.5-0.75

Head-to-head

<table>
<thead>
<tr>
<th></th>
<th>Grammatical</th>
<th>Meaning</th>
<th>Simplicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>4.7</td>
<td>4.1</td>
<td>2.4</td>
</tr>
<tr>
<td>Phrase</td>
<td>4.38</td>
<td>4</td>
<td>1.94</td>
</tr>
</tbody>
</table>

Individual

More labeling?
A problem

- Mt. Baker, the school district
- Baker Hostetler, the company
- Baker, a job opening

Timeless...

- Baker Jobs on CareerBuilder.com
- Baker, a job opening on CareerBuilder.com

A solution

- Why is this better?
- How does it happen?

Job Openings:
- Category = Food Services
- Keyword = Baker
- Location = Continental U.S.
Extracting Job Openings from the Web

Title: Ice Cream Guru
Description: If you dream of cold cream...
Contact: susan@foodscience.com
Category: Travel/Hospitality
Function: Food Services

Another Problem

Often structured information in text

Another Problem
And One more

Information Extraction

Traditional definition: Recovering structured data from text

What are some of the sub-problems/challenges?

- Recovering structured data from text
  - Identifying fields (e.g., named entity recognition)

- Understanding relations between fields (e.g., record association)
Information Extraction?

- Recovering structured data from text
  - Identifying fields (e.g., named entity recognition)
  - Understanding relations between fields (e.g., record association)
  - Normalization and deduplication

Information extraction

- Input: Text Document
  - Various sources: web, e-mail, journals, ...
- Output: Relevant fragments of text and relations possibly to be processed later in some automated way

Not all documents are created equal...

- Varying regularity in document collections
- Natural or unstructured
  - Little obvious structural information
- Partially structured
  - Contain some canonical formatting
- Highly structured
  - Often, automatically generated

Examples?

BACKGROUND: The most challenging aspect of revision hip surgery is the management of bone loss. A reliable and valid measure of bone loss is important since it will aid in future studies of hip revisions and in preoperative planning. We developed a measure of femoral and acetabular bone loss associated with failed total hip arthroplasty. The purpose of the present study was to measure the reliability and the intraoperative validity of this measure and to determine how it may be useful in preoperative planning. METHODS: From July 1997 to December 1998, forty-five consecutive patients with a failed hip prosthesis in need of revision surgery were prospectively followed. Three general orthopaedic surgeons were taught the radiographic classification system, and two of them classified standardized preoperative anteroposterior and lateral hip radiographs with use of the system. Interobserver testing was carried out in a blinded fashion. These results were then compared with the intraoperative findings of the third surgeon, who was blinded to the preoperative ratings. Kappa statistics (unweighted and weighted) were used to assess correlation. Interobserver reliability was assessed by examining the agreement between the two preoperative raters. Prognostic validity was assessed by examining the agreement between the assessment by either Rater 1 or Rater 2 and the intraoperative assessment (reference standard).

RESULTS: With regard to the assessments of both the femur and the acetabulum, there was significant agreement (p < 0.0001) between the preoperative raters (reliability), with weighted kappa values of >0.75. There was also significant agreement (p < 0.0001) between each rater’s assessment and the intraoperative assessment (validity) of both the femur and the acetabulum, with weighted kappa values of >0.75. CONCLUSION: With use of the newly developed classification system, preoperative radiographs are reliable and valid for assessment of the severity of bone loss that will be found intraoperatively.
Partially Structured:
Seminar Announcements

Extract time, location, speaker, etc.

Highly Structured:
Zagat’s Reviews

Extract restaurant, location, cost, etc.

Information extraction approaches

For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. “We can be open source. We love the concept of shared source,” said Bill Veghte, a Microsoft VP. “That’s a super-important shift for us in terms of code access.”

Richard Stallman, founder of the Free Software Foundation, countered saying...

IE Posed as a Machine Learning Task

- Training data: documents marked up with ground truth
- Extract features around words/information
- Pose as a classification problem

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates</td>
<td>CEO</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Bill Veghte</td>
<td>VP</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Richard Stallman</td>
<td>Founder</td>
<td>Free Soft</td>
</tr>
</tbody>
</table>

How can we do this? Can we utilize any tools/approaches we’ve seen so far?

What features would be useful?
Good Features for Information Extraction

Example word features:
- identity of word
- is in all caps
- ends in "ski"
- is part of a noun phrase
- is in a list of city names
- is under node X in WordNet or Cyc
- is in bold font
- is in hyperlink anchor
- features of past & future
- last person name was female
- next two words are "and Associates"

contains-question-mark
contains-question-word
ends-with-question-word
first-alpha-is-capitalized
indented
indented-1-to-4
indented-5-to-10
more-than-one-third-space
only-punctuation
prev-is-blank
prev-begins-with-ordinal
shorter-than-30

Is Capitalized
Is Mixed Caps
Is All Caps
Initial Cap
Contains Digit
All lowercase
Is Initial
Punctuation
Period
Comma
Apostrophe
Dash

html/Formatting Features
- {begin, end, in} x {<b>, <i>, <a>, <hN>} x {lengths 1, 2, 3, 4, or longer}
- {begin, end} of line

Lots of possible techniques

How can we pose this as a classification (or learning) problem?

Data  | Label  | train a predictive model
0     | 0      | classifier
0     | 0      |
1     | 1      |
1     | 0      |
0     | 0      |
Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g., analogy, explanation-based learning), learning theory (e.g., PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.
Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g., analogy, explanation-based learning), learning theory (e.g., PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

- Standard supervised learning setting
  - Positive instances: Windows with real label
  - Negative instances: All other windows
  - Features based on candidate, prefix and suffix
Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g., analogy, explanation-based learning), learning theory (e.g., PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.
### IE by Boundary Detection

**Input:** Linear Sequence of Tokens

**Date:** Thursday, October 25  
**Time:** 4:15 - 5:30 PM

How can we pose this as a machine learning problem?

<table>
<thead>
<tr>
<th>Data</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>train a predictive model</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

### Learning: IE as Classification

Learn **TWO** binary classifiers, one for the beginning and one for the end

- **Begin**
  - **Date:** Thursday, October 25  
  - **Time:** 4:15 - 5:30 PM

- **End**
  - **Date:** Thursday, October 25  
  - **Time:** 4:15 - 5:30 PM

- **ALL OTHERS NEGATIVE (0)**


def \text{Begin}(i) =
\begin{cases}
1 & \text{if } i \text{ begins a field} \\
0 & \text{otherwise}
\end{cases}

### One approach: Boundary Detectors

A “**Boundary Detectors**” is a pair of token sequences \(p, s\)

- A detector matches a boundary if \(p\) matches text before boundary and \(s\) matches text after boundary
- Detectors can contain wildcards, e.g. “capitalized word”, “number”, etc.

\[
<\text{Date} , \{\text{CapitalizedWord}\}>
\]

**Date:** Thursday, October 25

Would this boundary detector match anywhere?
One approach: Boundary Detectors

A "Boundary Detector" is a pair of token sequences \(p, s\):
- A detector matches a boundary if \(p\) matches text before boundary and \(s\) matches text after boundary.
- Detectors can contain wildcards, e.g. "capitalized word", "number", etc.

\(<\text{Date: } [\text{CapitalizedWord}])\>

Date: Thursday, October 25

Combining Detectors

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>(&lt;a \text{href=&quot;}})</td>
<td>http</td>
</tr>
<tr>
<td>empty</td>
<td>&quot;&gt;</td>
</tr>
</tbody>
</table>

\(<\text{http://www.cs.pomona.edu})\>

match(es)?

Combining Detectors

Learning: IE as Classification

Learn TWO binary classifiers, one for the beginning and one for the end.

\(<\text{Begin}) \text{ \text{\&}} \text{End})\>

Date: Thursday, October 25 Time: 4:15 - 5:30 PM

Say we learn \text{Begin} and \text{End}, will this be enough? Any improvements? Any ambiguities?
Some concerns

Learning to detect boundaries

- Learn three probabilistic classifiers:
  - \( \text{Begin}(i) \) is the probability position \( i \) starts a field
  - \( \text{End}(j) \) is the probability position \( j \) ends a field
  - \( \text{Len}(k) \) is the probability of an extracted field having length \( k \)

- Score a possible extraction \((i,j)\) by:
  - \( \text{Begin}(i) \times \text{End}(j) \times \text{Len}(j-i) \)

- \( \text{Len}(k) \) is estimated from histogram data

- \( \text{Begin}(i) \) and \( \text{End}(j) \) may combine multiple boundary detectors!

---

Problems with Sliding Windows and Boundary Finders

- Decisions in neighboring parts of the input are made independently from each other.
- Sliding Window may predict a "seminar end time" before the "seminar start time".
- It is possible for two overlapping windows to both be above threshold.
- In a Boundary-Finding system, left boundaries are laid down independently from right boundaries

---

Modeling the sequential nature of data: citation parsing


What patterns do you see here?
Ideas?
Some sequential patterns
- Authors come first
- Title comes before journal
- Page numbers come near the end
- All types of things generally contain multiple words

Predict a sequence of tags


Ideas?

Hidden Markov Models (HMMs)

HMM: Model
- States: $x_i$
- State transitions: $P(x_i | x_j) = a[x_i | x_j]$
- Output probabilities: $P(o_i | x_j) = b[o_i | x_j]$
- Markov independence assumption
HMMs: Performing Extraction

- Given output words:
  - fahlman s e 1991 the recurrent cascade correlation learning architecture nips 3 190 205
- Find state sequence that maximizes:
  \[ \prod_{i} a[x_{i} | x_{i-1}] H[a_{i} | x_{i}] \]
  - State transition
  - Output probabilities
- Lots of possible state sequences to test (5**14**)

IE Evaluation

- precision
  - of those we identified, how many were correct?
- recall
  - what fraction of the correct ones did we identify?
- F1
  - blend of precision and recall

IE Evaluation

Ground truth

Fahlman, S. E. (1991) The recurrent cascade

System

Fahlman, S. E. (1991) The recurrent cascade

How should we calculate precision?

5/6? 2/3? something else?
Data regularity is important!

- As the regularity decreases, so does the performance

Improving task regularity

- Instead of altering methods, alter text
- Idea: Add limited grammatical information
  - Run shallow parser over text
  - Flatten parse tree and insert as tags

Example of Tagged Sentence:

Uba2p is located largely in the nucleus.

Tagging Results on Natural Domain

Average performance on 4 data sets

Bootstrapping

Problem: Extract (author, title) pairs from the web
Approach 1: Old school style

Download the web:

Grab a sample and label:

train model:

classifier

Approach 1: Old school style

Download the web:

Grab a sample and label:

train model:

classifier

run model on web and get titles/authors
Approach 1: Old school style

Problems? Better ideas?

Bootstrapping

Seed set
author/title pairs

Google

author/title occurrences in context

patterns

Bootstrapping

Seed set
author/title pairs

Google

author/title occurrences in context

patterns
Brin, 1998
(Extracting patterns and relations from the world wide web)

<table>
<thead>
<tr>
<th>Seed books</th>
<th>1st iteration</th>
<th>2nd iteration</th>
<th>3rd iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unique pairs</td>
<td>Occurrences</td>
<td>patterns</td>
</tr>
<tr>
<td>H. D. Hammond</td>
<td>5</td>
<td>199</td>
<td>3</td>
</tr>
<tr>
<td>The Death, The Sick and Other Ghosts</td>
<td></td>
<td>3972</td>
<td>105</td>
</tr>
<tr>
<td>R. G. Wells</td>
<td></td>
<td>9938</td>
<td>346</td>
</tr>
<tr>
<td>The First Man in the Moon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R. G. Wells</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Time Machine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R. G. Wells</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>When the Shaper Sleeps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M. M. Rameau</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Golem</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Seed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H. P. Lovelock &amp; August Derleth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Author of the Abyss</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R. P. Lovelock</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Case of Charles Dexter Dried</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R. P. Lovelock</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Dodo That Came to Bermond and Other Stories</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Experiments

<table>
<thead>
<tr>
<th>NELL: Never-Ending Language Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ <a href="http://rtw.ml.cmu.edu/rtw/">http://rtw.ml.cmu.edu/rtw/</a></td>
</tr>
<tr>
<td>▶ continuously crawls the web to grab new data</td>
</tr>
<tr>
<td>▶ learns entities and relationships from this data</td>
</tr>
<tr>
<td>▶ started with a seed set</td>
</tr>
<tr>
<td>▶ uses learning techniques based on current data to learn new information</td>
</tr>
</tbody>
</table>

NELL

- Never-Ending Language Learning
- http://rtw.ml.cmu.edu/rtw/
- continuously crawls the web to grab new data
- learns entities and relationships from this data
- started with a seed set
- uses learning techniques based on current data to learn new information
4 different approaches to learning relationships
- Combine these in the knowledge integrator
- Idea: using different approaches will avoid overfitting
- Initially was wholly unsupervised, now some human supervision
- Cookies are food => internet cookies are food => files are food

An example learner: coupled pattern learner (CPL)

Cities:
- Los Angeles
- San Francisco
- New York
- Seattle
- ... city of X ...
- ... the official guide to X ...
- ... only in X ...
- ... what to do in X ...
- ... mayor of X ...
- ... mayor of X ...

extract occurrences of group
statistical co-occurrence test

CPL

... mayor of <CITY> ...

extract other cities from the data
Albuquerque
Springfield

CPL

- Can also learn patterns with multiple groups
- X is the mayor of Y ...
- X plays for Y ...
- X is a player of Y ...
- can extract other groups, but also relationships
  Antonio Villaraigosa
  mayor of Los Angeles
NELL performance

For more details: http://rtw.ml.cmu.edu/papers/carlson-aaai10.pdf

NELL

- The good:
  - Continuously learns
  - Uses the web (a huge data source)
  - Learns generic relationships
  - Combines multiple approaches for noise reduction

- The bad:
  - Makes mistakes (overall accuracy still may be problematic for real world use)
  - Does require some human intervention
  - Still many general phenomena won't be captured