INFORMATION EXTRACTION

David Kauchak
cs457
Fall 2011

some content adapted from:
http://www.cs.cmu.edu/~knigam/15-505/ie-lecture.ppt

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 cast search
      - depth limited search
      - iterative deepening search
    - informed search:
      - greedy-first search
      - A* search
    - completeness, optimality
    - heuristics
      - admissibility
    - graph search vs. tree search

Administrivia

- 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)
- http://www.cs.middlebury.edu/~schar/dept/seminars/cs-seminar-2011-12-7-Redmon.pdf

Simplification evaluation

Head-to-head

	Grammatical	Meaning	Simplicity
Syntax	4.7	4.07	2.9
Phrase	4.5	4.23	2
Simple Wiki	4.5	3.73	2.73

Reasonable correlation: 0.5-0.75

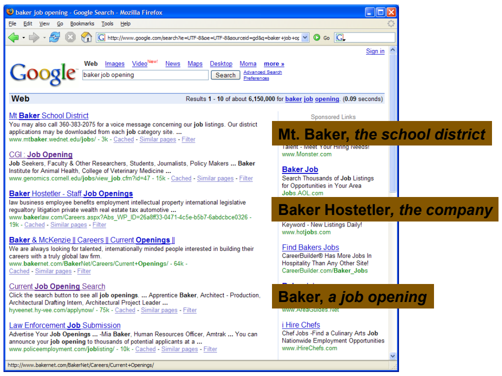
Individual

	Grammatical	Meaning	Simplicity
Syntax	4.7	4.1	2.4
Phrase	4.38	4	1.94

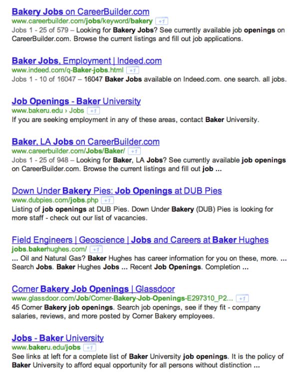
Simplification evaluation

- More labeling?

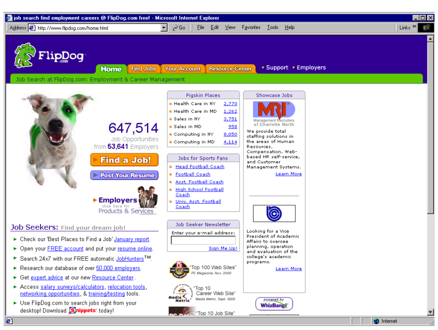
A problem



Timeless…



A solution

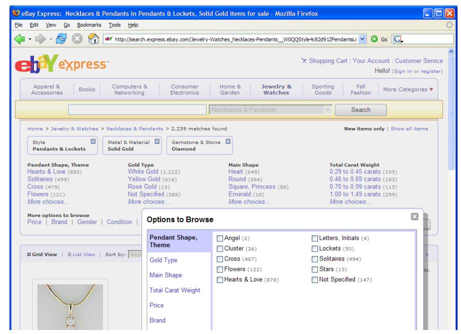


Why is this better? **How does it happen?**

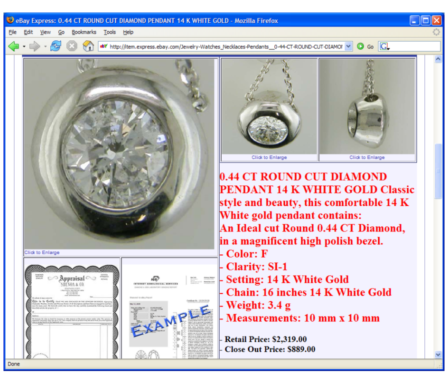
Job Openings:
Category = *Food Services*
Keyword = *Baker*
Location = *Continental U.S.*

Extracting Job Openings from the Web

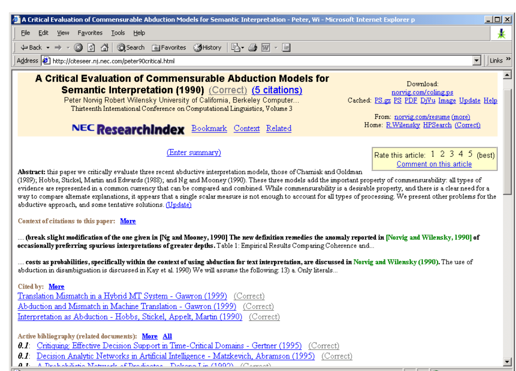
Another Problem



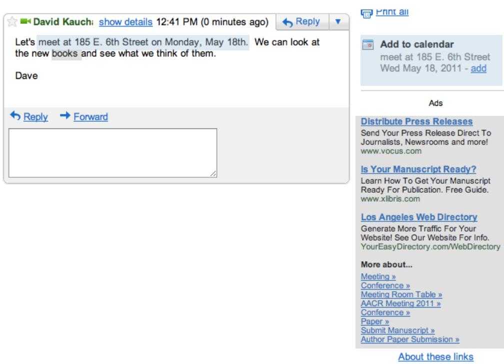
Often structured information in text



Another Problem



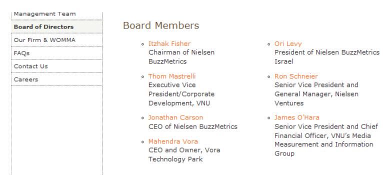
And One more



Information Extraction

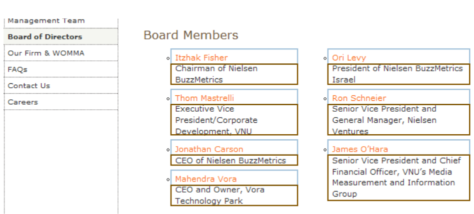
Traditional definition: **Recovering structured data from text**

What are some of the sub-problems/challenges?



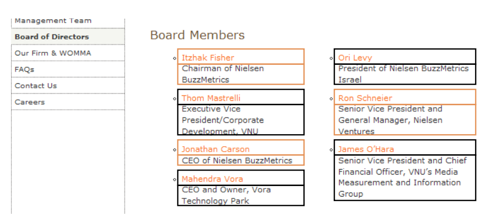
Information Extraction?

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



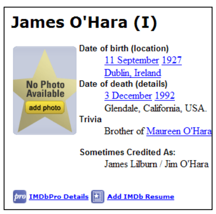
Information Extraction?

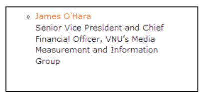
- Recovering structured data from text
  - Identifying fields (e.g. named entity recognition)
  - Understanding relations between fields (e.g. record association)

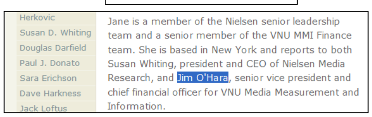


Information Extraction?

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

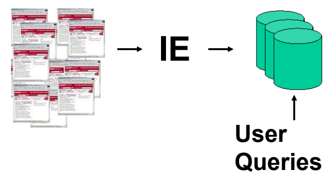




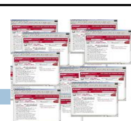


Information extraction

- 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 <u>document collections</u>
- Natural or unstructured
  - Little obvious structural information
- Partially structured
  - Contain some canonical formatting
- Highly structured
  - Often, automatically generated

Examples?

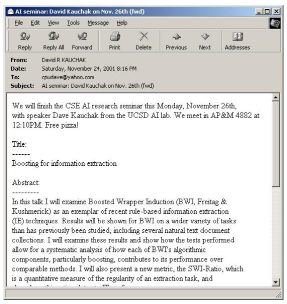
Natural Text: MEDLINE Journal Abstracts

Extract number of subjects, type of study, conditions, etc.

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. CONCLUSIONS: 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.



AI seminar: David Kauchak on Nov. 26th (fwd)

From: David R KAUCHAK
Date: Saturday, November 24, 2001 8:16 PM
To: cpudave@yahoo.com
Subject: AI seminar: David Kauchak on Nov. 26th (fwd)

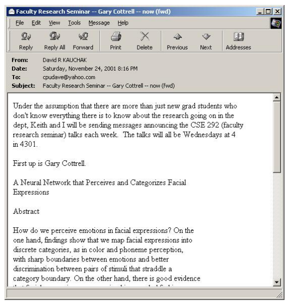
We will finish the CSE AI research seminar this Monday, November 26th, with speaker Dave Kauchak from the UCSD AI lab. We meet in AP&M 4882 at 12:10PM. Free pizza!

Title:

Boosting for information extraction

Abstract:

In this talk I will examine Boosted Wrapper Induction (BWI, Freitag & Kushmerick) as an exemplar of recent rule-based information extraction (IE) techniques. Results will be shown for BWI on a wider variety of tasks than has previously been studied, including several natural text document collections. I will examine these results and show how the tests performed allow for a systematic analysis of how each of BWI's algorithmic components, particularly boosting, contributes to its performance over comparable methods. I will also present a new metric, the SWI-Ratio, which is a quantitative measure of the regularity of an extraction task, and



Faculty Research Seminar -- Gary Cottrell -- now (fwd)

From: David R KAUCHAK
Date: Saturday, November 24, 2001 8:16 PM
To: cpudave@yahoo.com
Subject: Faculty Research Seminar -- Gary Cottrell -- now (fwd)

Under the assumption that there are more than just new grad students who don't know everything there is to know about the research going on in the dept, Keith and I will be sending messages announcing the CSE 292 (faculty research seminar) talks each week. The talks will all be Wednesdays at 4 in 4301.

First up is Gary Cottrell.

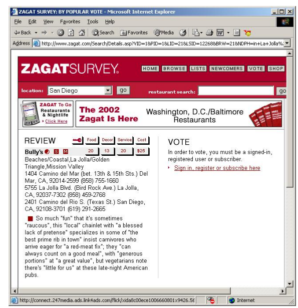
A Neural Network that Perceives and Categorizes Facial Expressions

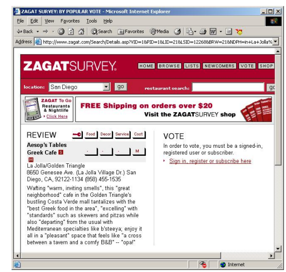
Abstract

How do we perceive emotions in facial expressions? On the one hand, findings show that we map facial expressions into discrete categories, as in color and phoneme perception, with sharp boundaries between emotions and better discrimination between pairs of stimuli that straddle a category boundary. On the other hand, there is good evidence

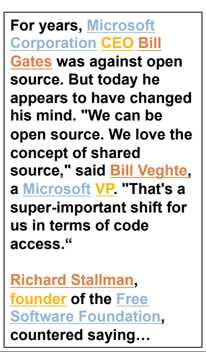
Highly Structured: Zagat's Reviews

Extract restaurant, location, cost, etc.





Information extraction approaches

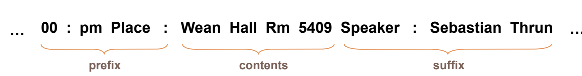


Name	Title	Organization
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	Founder	Free Soft..

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

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



What features would be useful?

Good Features for Information Extraction

- begins-with-number
- begins-with-ordinal
- begins-with-punctuation
- begins-with-question-word
- begins-with-subject
- blank
- contains-alphanum
- contains-bracketed-number
- contains-http
- contains-non-space
- contains-number
- contains-pipe

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-mark
- 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

Good Features for Information Extraction

- Is Capitalized
- Is Mixed Caps
- Is All Caps
- Initial Cap
- Contains Digit
- All lowercase
- Is Initial
- Punctuation
- Period
- Comma
- Apostrophe
- Dash
- Preceded by HTML tag

- Character n-gram classifier says string is a person name (80% accurate)
- In stopword list (the, of, their, etc)
- In honorific list (Mr, Mrs, Dr, Sen, etc)
- In person suffix list (Jr, Sr, PhD, etc)
- In name particle list (de, la, van, der, etc)
- In Census lastname list; segmented by P(name)
- In Census firstname list; segmented by P(name)
- In locations lists (states, cities, countries)
- In company name list ("J. C. Penny")
- In list of company suffixes (Inc, & Associates, Foundation)

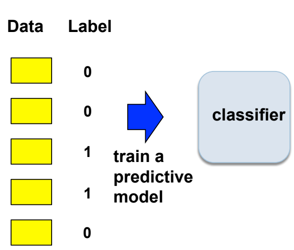
Word Features
- lists of job titles,
- Lists of prefixes
- Lists of suffixes
- 350 informative phrases

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

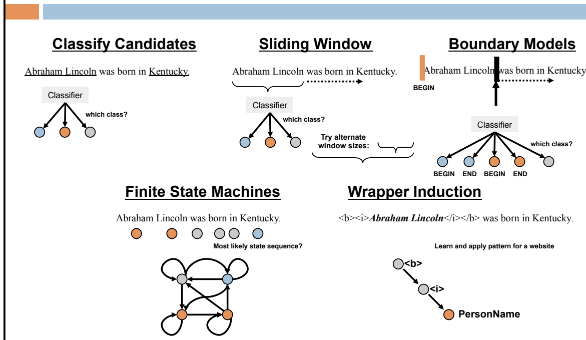
How can we pose this as a classification (or learning) problem?

… 00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun …

prefix — contents — suffix



Lots of possible techniques



Any of these models can be used to capture words, formatting or both.

Information Extraction by Sliding Window

**E.g.
Looking for seminar location**

```
GRAND CHALLENGES FOR MACHINE LEARNING

         Jaime Carbonell
     School of Computer Science
    Carnegie Mellon University

             3:30 pm
          7500 Wean Hall

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.
```

CMU UseNet Seminar Announcement

Information Extraction by Sliding Window

**E.g.
Looking for seminar location**

```
GRAND CHALLENGES FOR MACHINE LEARNING

         Jaime Carbonell
     School of Computer Science
    Carnegie Mellon University

             3:30 pm
          7500 Wean Hall

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.
```

CMU UseNet Seminar Announcement

Information Extraction by Sliding Window

**E.g.
Looking for seminar location**

```
GRAND CHALLENGES FOR MACHINE LEARNING

         Jaime Carbonell
     School of Computer Science
    Carnegie Mellon University

             3:30 pm
          7500 Wean Hall

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.
```

CMU UseNet Seminar Announcement

Information Extraction by Sliding Window

**E.g.
Looking for seminar location**

```
GRAND CHALLENGES FOR MACHINE LEARNING

         Jaime Carbonell
     School of Computer Science
    Carnegie Mellon University

             3:30 pm
          7500 Wean Hall

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.
```

CMU UseNet Seminar Announcement

Information Extraction by Sliding Window

E.g. Looking for seminar location

```
GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell
School of Computer Science
Carnegie Mellon University

3:30 pm
7500 Wean Hall

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.
```

CMU UseNet Seminar Announcement

Information Extraction by Sliding Window

… **00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun** …

$\underbrace{w_{t-m} \ \cdots \ w_{t-1}}_{\text{prefix}} \quad \underbrace{w_t \ \cdots \ w_{t+n}}_{\text{contents}} \quad \underbrace{w_{t+n+1} \ \cdots \ w_{t+n+m}}_{\text{suffix}}$

- Standard supervised learning setting
  - Positive instances?
  - Negative instances?

Information Extraction by Sliding Window

… **00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun** …

$\underbrace{w_{t-m} \ \cdots \ w_{t-1}}_{\text{prefix}} \quad \underbrace{w_t \ \cdots \ w_{t+n}}_{\text{contents}} \quad \underbrace{w_{t+n+1} \ \cdots \ w_{t+n+m}}_{\text{suffix}}$

- Standard supervised learning setting
  - Positive instances: Windows with real label
  - Negative instances: All other windows
  - Features based on candidate, prefix and suffix

IE by Boundary Detection

E.g. Looking for seminar location

```
GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell
School of Computer Science
Carnegie Mellon University

3:30 pm
7500 Wean Hall

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.
```

CMU UseNet Seminar Announcement

IE by Boundary Detection

E.g. Looking for seminar location

```
     GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
        School of Computer Science
       Carnegie Mellon University

               3:30 pm
            7500 Wean Hall

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.
```

CMU UseNet Seminar Announcement

IE by Boundary Detection

E.g. Looking for seminar location

```
     GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
        School of Computer Science
       Carnegie Mellon University

               3:30 pm
            7500 Wean Hall

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.
```

CMU UseNet Seminar Announcement

IE by Boundary Detection

E.g. Looking for seminar location

```
     GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
        School of Computer Science
       Carnegie Mellon University

               3:30 pm
            7500 Wean Hall

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.
```

CMU UseNet Seminar Announcement

IE by Boundary Detection

E.g. Looking for seminar location

```
     GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
        School of Computer Science
       Carnegie Mellon University

               3:30 pm
            7500 Wean Hall

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.
```

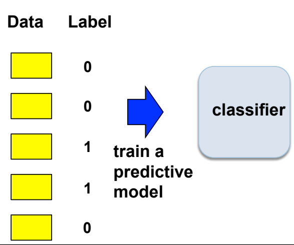
CMU UseNet Seminar Announcement

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?

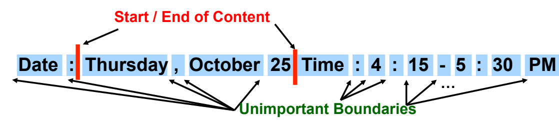


IE by Boundary Detection

Input: Linear Sequence of Tokens

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

Method: Identify start and end Token Boundaries



Output: Tokens Between Identified Start / End Boundaries

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

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

POSITIVE (1)

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

ALL OTHERS NEGATIVE (0)

`Begin(i)=` 1 if *i* begins a field
0 otherwise

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.

<Date: , [*CapitalizedWord*]>

Date: Thursday, October 25

Would this boundary detector match anywhere?

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.

<Date: , [*CapitalizedWord*]>

Date: Thursday, October 25

Combining Detectors

	Prefix	Suffix
Begin **boundary detector:**	`<a href="`	`http`
End **boundary detector:**	*empty*	`">`

`text<b><a href="http://www.cs.pomona.edu">`

match(es)?

Combining Detectors

	Prefix	Suffix
Begin **boundary detector:**	`<a href="`	`http`
End **boundary detector:**	*empty*	`">`

`text<b><a href="http://www.cs.pomona.edu">`

Begin End

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

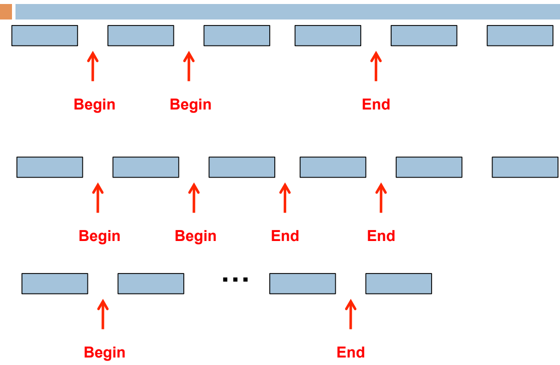
POSITIVE (1)

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

ALL OTHERS NEGATIVE (0)

Say we learn Begin and End, will this be enough? Any improvements? Any ambiguities?

Some concerns



Learning to detect boundaries

- Learn **three** probabilistic classifiers:
  - *Begin(i)* = probability position *i* starts a field
  - *End(j)* = probability position *j* ends a field
  - *Len(k)* = probability an extracted field has length k
- Score a possible extraction *(i,j)* by
  *Begin(i) * End(j) * Len(j-i)*
- *Len(k)* is estimated from a histogram data
- *Begin(i)* and *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

- Fahlman, Scott & Lebiere, Christian (1989). The cascade-correlation learning architecture. Advances in Neural Information Processing Systems, pp. 524-532.
- Fahlman, S.E. and Lebiere, C., "The Cascade Correlation Learning Architecture," Neural Information Processing Systems, pp. 524-532, 1990.
- Fahlman, S. E. (1991) The recurrent cascade-correlation learning architecture. NIPS 3, 190-205.

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

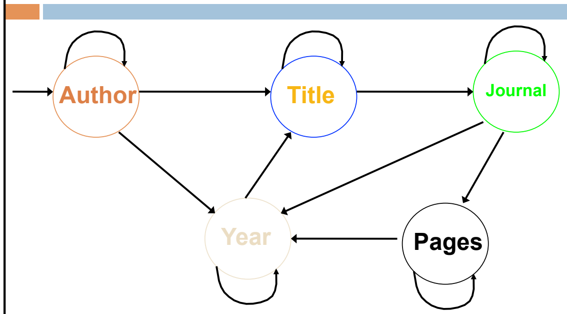
Predict a sequence of tags

author author year title title title
Fahlman, S. E. (1991) The recurrent cascade

title title title journal pages
correlation learning architecture. NIPS 3, 190-205.

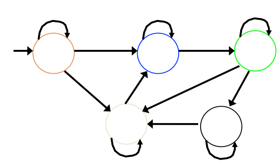
Ideas?

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:

$$\prod_i a[x_i \mid x_{i-1}]\, b[o_i \mid x_i]$$

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

author author year title title title

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

System

author pages year title title title

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

How should we calculate precision?

IE Evaluation

Ground truth

author author year title title title

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

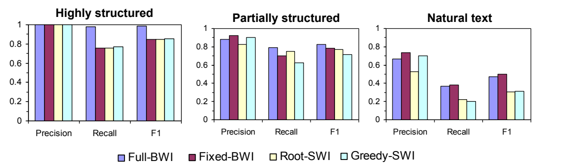
System

author pages year title title title

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

5/6? 2/3? something else?

Data regularity is important!



- As the regularity decreases, so does the performance

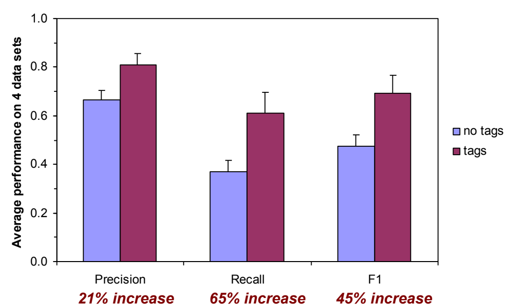
Improving task regularity

- 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:

$\underbrace{\text{Uba2p}}_{\text{NP\_SEG}}\ \underbrace{\text{is located largely}}_{\text{VP\_SEG}}\ \underbrace{\text{in}}_{\text{PP\_SEG}}\ \underbrace{\text{the nucleus}}_{\text{NP\_SEG}}.$

Tagging Results on Natural Domain



Bootstrapping

Problem: Extract (author, title) pairs from the web

Abraham Lincoln by James Russell Lowell
Action Front by Boyd Cable
Several short stories based on real events in WWI that try to give a sense of what it was like for the people on the front lines.
Adventure by Jack London
Adventure of Wisteria Lodge, The by Arthur Conan Doyle
Adventure of the Bruce-Partington Plans, The by Arthur Conan Doyle
Adventure of the Cardboard Box, The by Arthur Conan Doyle
Adventure of the Devil's Foot, The by Arthur Conan Doyle
Adventure of the Dying Detective, The by Arthur Conan Doyle
Adventure of the Red Circle, The by Arthur Conan Doyle
Adventures of Colonel Daniel Boone, The by John Filson

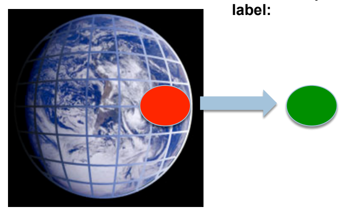
Approach 1: Old school style

Download the web:

Approach 1: Old school style

Download the web:

Grab a sample and label:

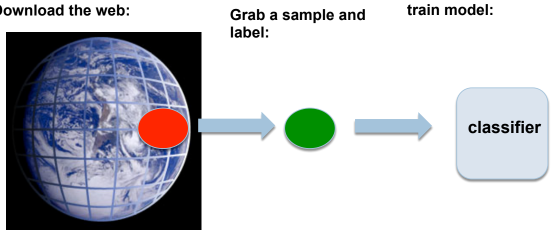


Approach 1: Old school style

Download the web:

Grab a sample and label:

train model:

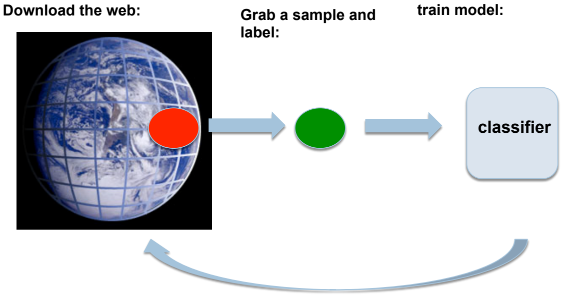


Approach 1: Old school style

Download the web:

Grab a sample and label:

train model:

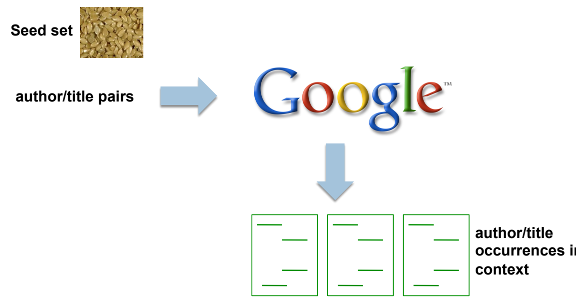


run model on web and get titles/authors

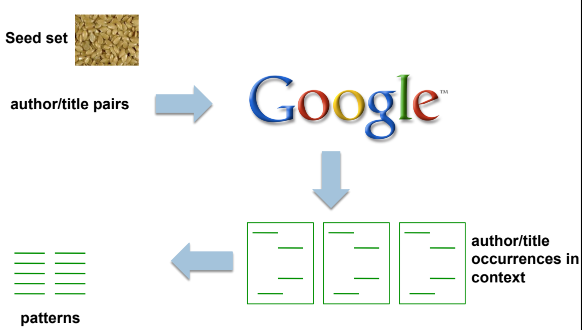
Approach 1: Old school style

Problems? Better ideas?

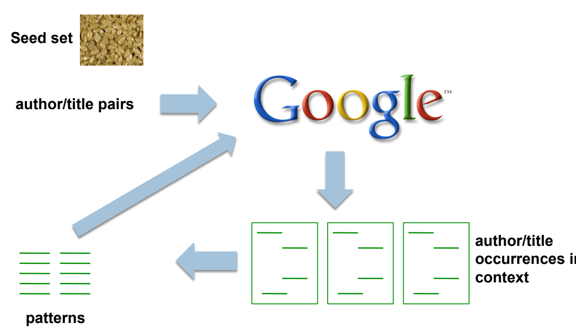
Bootstrapping



Bootstrapping



Bootstrapping



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

Seed books

Isaac Asimov	The Robots of Dawn
David Brin[3]	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

Patterns

<table>
<tr><th>URL Pattern</th><th>Text Pattern</th></tr>
<tr><td>www.sff.net/locus/c.*</td><td><LI><B>title</B> by author (</td></tr>
<tr><td>dns.city-net.com/Ĩmann/awards/hugos/1984.html</td><td><i>title</i> by author (</td></tr>
<tr><td>dolphin.upenn.edu/ďcummins/texts/sf-award.htm</td><td>author || title || (</td></tr>
</table>

New books

H. D. Everett	The Death-Mask and Other Ghosts
H. G. Wells	First Men in the Moon
H. G. Wells	Science Fiction: Volume 2
H. G. Wells	The First Men in the Moon
H. G. Wells	The Invisible Man
H. G. Wells	The Island of Dr. Moreau
H. G. Wells	The Science Fiction Volume 1
H. G. Wells	The Shape of Things to Come: The Ultimate Revolution
H. G. Wells	The Time Machine
H. G. Wells	The War of the Worlds
H. G. Wells	When the Sleeper Wakes
H. M. Hoover	Journey Through the Empty
H. P. Lovecraft & August Derleth	The Lurker at the Threshold
H. P. Lovecraft	At the Mountains of Madness and Other Tales of Terror
H. P. Lovecraft	The Case of Charles Dexter Ward
H. P. Lovecraft	The Doom That Came to Sarnath and Other Stories

Experiments

	1^{st} iteration	2^{nd} iteration	3^{rd} iteration
Unique (author, title) pairs	5	4047	9127
Occurrences	199	3972	9938
patterns	3	105	346
Result: unique pairs	4047	9127	15257

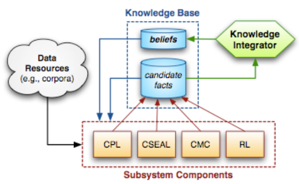
Final list

Henry James	The Europeans
Henry James	The Golden Bowl
Henry James	The Portrait of a Lady
Henry James	The Turn of the Screw
Henry James	Turn of the Screw
Henry John Coke	Tracks of a Rolling Stone
Henry K. Rowe	Landmarks in Christian History
Henry Kisor	Zephyr
Henry Lawson	In the Days When the World Was Wide
Henry Longfellow	The Song of Hiawatha
Henry Miller	Tropic of Cancer
Henry Petroski	Invention On Design
Henry Petroski	The Evolution of Useful Things
Henry Roth	Call It Sleep
Henry Sumner Maine	Ancient Law
Henry Tuckerman, Lindsay, Phila	Characteristics of Literature
Henry Van Dyke	The Blue Flower
Henry Van Dyke, Scrib	Days Off
Henry Van Loon	Life and Times of Pieter Stuyvesant
Henry Wadsworth Longfellow	Paul Revere's Ride
Henry Wadsworth Longfellow	Evangeline
Henry Wadsworth Longfellow	The Song of Hiawatha
Herbert Donald	Lincoln
Herbert M. Hart	Old Forts of the Northwest
Herbert M. Mason, Jr	The Lafayette Escadrille
Herbert R. Lottman	Jules Verne: An Exploratory Biography
Herbert Spencer	The Man Versus the State
Herman Daly	For the Common Good
Herman Daly	Valuing the Earth
Herman E. Kittredge	Ingersoll: A Biographical Appreciation
Herman Haken	Principles of Brain Functioning
Herman Hesse	Demian
Herman Hesse	Siddhartha
Herman Hesse	Sidharta
Herman Melville	Bartleby, the Scrivener
Herman Melville	Billy Budd
Herman Melville	Billy Budd
Herman Melville	Moby Dick
Herman Melville	The Confidence Man
Herman Melville	The Encantadas, or Enchanted Isles
Herman Melville	Typee: A Peep at Polynesian Life
Herman Weiss	Sunset Detectives
Herman Wouk	War And Remembrance
Hermann Hesse	Klingsor's Last Summer
Hermann Hesse	Knulp
Hermann Hesse	Rosshalde
Hermann Hesse	Strange News From Another Star
Herodotus	Histories
Herodotus	The Histories
Herodotus	The History of Herodotus
Herschel Hobbs	Pastor's Manual
Hetschel	First Stage: Moon

NELL

- 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

NELL



- 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

Albequerque
Springfield
…

CPL

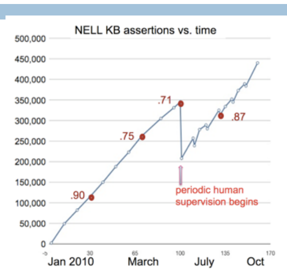
- Can also learn patterns with multiple groups

… 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



estimated accuracy in red

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