

INFORMATION EXTRACTION

David Kauchak
cs457
Fall 2011

www 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 cost 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

Google search results for "baker job opening". The results are scattered and difficult to navigate. Key results include:

- Mt. Baker School District**: You may also call 360.363.2075 for a voice message concerning our job listings. Our district applications may be downloaded from each job category site...
- CGI Job Opening**: Job Openings - Faculty & Other Researchers, Students, Journalists, Policy Makers ... Baker Institute for Animal Health, College of Veterinary Medicine ...
- Baker Hostetler - Staff Job Openings**: We business employee benefits employment intellectual property international legislative regulatory litigation private equity real estate tax accounting ...
- Baker & McKenna | Careers | Current Openings |**: We are always looking for talented, internationally minded people interested in building their career with a truly global law firm.
- Current Job Opening Search**: Click the results below to see all job openings ... Apprentice Baker, Assistant - Production, Architectural Drafting Intern, Architectural Project Leader ...
- Law Enforcement Job Submission**: Advertise Your Job Openings ... Add Baker, Human Resources Officer, Amtrak ... You can advertise your job openings to thousands of potential applicants at ...
- Links Check**: Chef Jobs: Find a Culinary Arts Job Nationwide Employer Opportunity ...

Annotations on the screenshot:

- Mt. Baker, the school district**: points to the first result.
- Baker Hostetler, the company**: points to the third result.
- Baker, a job opening**: points to the fourth result.

Timeless...

Bakery Jobs on CareerBuilder.com
www.careerbuilder.com/jobs/keyword/Bakery/...
Jobs 1 - 25 of 879 - Looking for Bakery Jobs? See currently available job openings on CareerBuilder.com. Browse the current listings and fill out job applications.

Baker Jobs, Employment | Indeed.com
www.indeed.com/q/Baker-jobs.html
Jobs 1 - 10 of 16047 - 16047 Baker Jobs available on Indeed.com. one search. all jobs. Jobs 1 - 10 of 16047 - 16047 Baker Jobs available on Indeed.com. one search. all jobs.

Job Openings - Baker University
www.baker.edu/jobs/...
If you are seeking employment in any of these areas, contact Baker University.

Baker LA Jobs on CareerBuilder.com
www.careerbuilder.com/jobs/Baker/...
Jobs 1 - 25 of 848 - Looking for Baker LA Jobs? See currently available job openings on CareerBuilder.com. Browse the current listings and fill out job ...

Down Under Bakery Pies: Job Openings at DUB Pies
www.dupies.com/jobs.php/...
Listing of job openings at DUB Pies. Down Under Bakery (DUB) Pies is looking for more staff - check out our list of vacancies.

Field Engineers | Geoscience | Jobs and Careers at Baker Hughes
jobs.bakerhughes.com/...
... Oil and Natural Gas? Baker Hughes has career information for you on these, more ... Search jobs, Baker Hughes Jobs ... Recent Job Openings, Competition ...

Corner Bakery Job Openings | Glassdoor
www.glassdoor.com/Job/Corner-Bakery-Job-Openings-E207310_P2...
45 Corner Bakery job openings. Search job openings, see if they fit - company salaries, reviews, and more posted by Corner Bakery employees.

Jobs - Baker University
www.baker.edu/jobs/...
See links at left for a complete list of Baker University job openings. It is the policy of Baker University to afford equal opportunity for all persons without distinction ...

A solution

FlipDog search results for "baker job opening". The interface is clean and organized. Key features include:

- Search Results Summary**: 647,514 jobs found.
- Filters**: Health Care in NY (2,203), Baker in NY (2,203), Computing in NY (2,203), Computing in MD (2,203).
- Job Seekers**: Find your dream job!
 - Check out Best Places to Work & Best Salaries 2012
 - Open your EBSE account and get your resume advice
 - Search 261 with our FREE resume jobfitting™
 - Research our database of 40,000+ employers
 - Get expert advice at our new Business Center
 - Access salary information, interview tips, interview questions, & interview advice tools
 - Use FlipDog to search jobs and view your desktop Download FlipDog today!
- Employers**: Products & Services
- Job Listings**:
 - Health Care in NY
 - Baker in NY
 - Computing in NY
 - Computing in MD
 - Links for requests from:
 - Health Care in NY
 - Baker in NY
 - Computing in NY
 - Computing in MD

Why is this better? How does it happen?

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

FlipDog search results for "baker job opening". The interface is clean and organized. Key features include:

- Search Results Summary**: 1 - 25 of 47 jobs shown below.
- Search filters**: Search tips, Show Jobs Posted: For all time periods
- View: Brief | Detailed**
- Web Jobs**: FlipDog technology has found these jobs on thousands of employer Web sites.
- Job Listings**:
 - Food Pantry Workers at Lutheran Social Services October 11, 2002 Archbold, OH
 - Cooks at Lutheran Social Services October 11, 2002 Archbold, OH
 - Bakers Assistants at Fine Catering by Russell Morin October 11, 2002 Atholton, MA
 - Baker's helper at Bird-in-Hand October 11, 2002 United States
 - Assistant Baker at Gourmet To Go October 11, 2002 Maryland Heights, MO
 - October 10, 2002 Raywonton, OR
 - October 10, 2002 Alta, UT
 - October 10, 2002 Hartsville, VT
 - October 10, 2002 Garden Grove, CA
 - October 10, 2002 Houma, LA
 - October 10, 2002 Miroso, MN
 - October 10, 2002 Big Sky, MT
 - October 09, 2002 Willowbrook, IL
 - October 09, 2002 Las Vegas, NV
 - October 08, 2002 Minneapolis, MN

And One more

David Kauch [show details](#) 12:41 PM (0 minutes ago) [Reply](#)

Let's meet at 185 E. 6th Street on Monday, May 18th. We can look at the new books and see what we think of them.

Dave

[Reply](#) [Forward](#)

Add to calendar
meet at 185 E. 6th Street
Wed May 18, 2011 - [add](#)

Ads

Distribute Press Releases
Send Your Press Release Direct To Journalists, Newsrooms and more!
[www.vocus.com](#)

Is Your Manuscript Ready?
Learn How To Get Your Manuscript Ready For Publication. Free Guide.
[www.xlibris.com](#)

Los Angeles Web Directory
Generate More Traffic For Your Website! See Our Website For Info, YourEasyDirectory.com/WebDirectory

More about...

[Meetings Conference »](#)
[Meeting Room Table »](#)
[AZOR Meeting 2011 »](#)
[Conference »](#)
[Paper »](#)
[Submit Manuscript »](#)
[Author Paper Submission »](#)

[About these links](#)

Information Extraction

Traditional definition: Recovering structured data from text

What are some of the sub-problems/challenges?

Management Team	Board Members
Board of Directors	<ul style="list-style-type: none"> Itzhak Fisher Chairman of Nielsen BuzzMetrics Thom Mastrelli Executive Vice President/Corporate Development, VNU Jonathan Carson CEO of Nielsen BuzzMetrics Mahendra Vora CEO and Owner, Vora Technology Park
Our Firm & WOMMA	<ul style="list-style-type: none"> Ori Levy President of Nielsen BuzzMetrics Israel Ron Schriener Senior Vice President and General Manager, Nielsen Ventures James O'Hara Senior Vice President and Chief Financial Officer, VNU's Media Measurement and Information Group
FAQ	
Contact Us	
Careers	

Information Extraction?

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

Management Team	Board Members
Board of Directors	<ul style="list-style-type: none"> Itzhak Fisher Chairman of Nielsen BuzzMetrics Thom Mastrelli Executive Vice President/Corporate Development, VNU Jonathan Carson CEO of Nielsen BuzzMetrics Mahendra Vora CEO and Owner, Vora Technology Park
Our Firm & WOMMA	<ul style="list-style-type: none"> Ori Levy President of Nielsen BuzzMetrics Israel Ron Schriener Senior Vice President and General Manager, Nielsen Ventures James O'Hara Senior Vice President and Chief Financial Officer, VNU's Media Measurement and Information Group
FAQ	
Contact Us	
Careers	

Information Extraction?


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

Management Team	Board Members
Board of Directors	<ul style="list-style-type: none"> Itzhak Fisher Chairman of Nielsen BuzzMetrics Thom Mastrelli Executive Vice President/Corporate Development, VNU Jonathan Carson CEO of Nielsen BuzzMetrics Mahendra Vora CEO and Owner, Vora Technology Park
Our Firm & WOMMA	<ul style="list-style-type: none"> Ori Levy President of Nielsen BuzzMetrics Israel Ron Schriener Senior Vice President and General Manager, Nielsen Ventures James O'Hara Senior Vice President and Chief Financial Officer, VNU's Media Measurement and Information Group
FAQ	
Contact Us	
Careers	

Information Extraction?

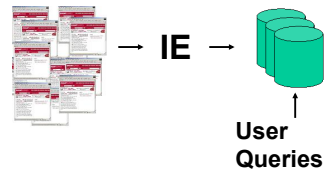
- Recovering structured data from text
 - Identifying fields (e.g. named entity recognition)
 - Understanding relations between fields (e.g. record association)
 - Normalization and deduplication

James O'Hara (I)

	Date of birth (location) 11 September 1927 Dublin, Ireland	<ul style="list-style-type: none"> • James O'Hara Senior Vice President and Chief Financial Officer, VNU's Media Measurement and Information Group
	Date of death (details) 3 December 1992 Glendale, California, USA	
Trivia Brother of Margaret O'Hara Sometimes Credited As: James L'burn / Jim O'Hara	Herkovic: Susan D. Whiting Douglas Darfield Paul J. Donato Sara Erickson Dave Harkness Jack Loftus	Jane is a member of the Nielsen senior leadership team and a senior member of the VNU MMI Finance team. She is based in New York and reports to both Susan Whiting, president and CEO of Nielsen Media Research, and Jim O'Hara , senior vice president and chief financial officer for VNU Media Measurement and Information.

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?

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 1987 to December 1995, **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.

From: David K. Kaelin-Lang
Date: Saturday, November 24, 2001 8:14 PM
To: otk@cs.ucsd.edu
Subject: AI seminar: David Kaelin-Lang Nov. 26th (Thu)

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

Title:

Booting for information extraction

Abstract:

In this talk I will examine Bootstrapped Wrapper Induction (BWI, Freitag & Kuhlmann) as an example 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 compare these results and show how the tests performed allow for a systematic analysis of how each of BWI's algorithmic components, particularly bootstrapping, contributes to its performance over comparable methods. I will also present a new metric, the F10 Ratio, which is a quantitative measure of the regularity of an extraction task, and

From: David E. Kaelin-Lang
Date: Saturday, November 24, 2001 6:16 PM
To: otk@cs.ucsd.edu
Subject: Faculty Research Seminar - Gary Cottrell - now (Thu)

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, Kaelin and I will be sending messages announcing the CSE 295 (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 may 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.

ZAGAT SURVEY

Washington, D.C./Baltimore Restaurants

REVIEW

Vote

In order to vote, you must be a registered, registered user or subscriber.

Sign up, register or subscribe here

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...

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

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

prefix contents suffix

What features would be useful?

Good Features for Information Extraction

begins-with-number	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 - <i>features of past & future</i> - last person name was female - next two words are "and Associates"	contains-question-mark
begins-with-ordinal		contains-question-word
begins-with-punctuation		ends-with-question-mark
begins-with-question-word		first-alpha-is-capitalized
begins-with-subject		indented
blank		indented-1-to-4
contains-alphanumeric		indented-5-to-10
contains-bracketed-number		more-than-one-third-space
contains-http		only-punctuation
contains-non-space		prev-is-blank
contains-number	prev-begins-with-ordinal	
contains-pipe	shorter-than-30	

Good Features for Information Extraction

Is Capitalized	Character n-gram classifier says string is a person name (80% accurate)	Word Features
Is Mixed Caps	In stopword list (the, of, their, etc)	<ul style="list-style-type: none"> lists of job titles Lists of prefixes Lists of suffixes 350 informative phrases
Is All Caps	In honorific list (Mr, Mrs, Dr, Sen, etc)	HTML/Formatting Features
Initial Cap	In person suffix list (Jr, Sr, PhD, etc)	<ul style="list-style-type: none"> {begin, end, in} x {, <i>, <u>, <NN>} x {lengths 1, 2, 3, 4, or longer} {begin, end} of line
Contains Digit	In name particle list (de, la, van, der, etc)	
All lowercase	In Census lastname list, segmented by P(name)	
Is Initial	In Census firstname list, segmented by P(name)	
Punctuation	In locations lists (states, cities, countries)	
Period	In company name list ("J. C. Penny")	
Comma	In list of company suffixes (Inc, & Associates, Foundation)	
Apostrophe		
Dash		
Preceded by HTML tag		

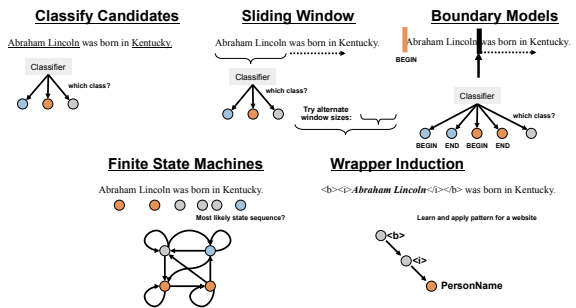
How can we pose this as a classification (or learning) problem?

... 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 ...

W_{pre} W_{c1} W_c W_{c2} W_{c3} W_{s1} W_{s2} W_{s3}
prefix contents suffix

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

Information Extraction by Sliding Window

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

W_{pre} W_{c1} W_c W_{c2} W_{c3} W_{s1} W_{s2} W_{s3}
prefix contents 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

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

E.g.
Looking for
seminar
location

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?

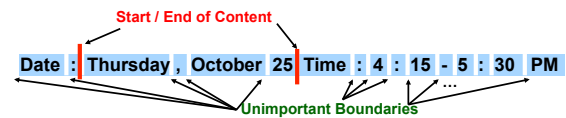


IE by Boundary Detection

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

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

ALL OTHERS NEGATIVE (0)

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

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.

$\langle \text{Date:} , [\text{CapitalizedWord}] \rangle$

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

Begin boundary detector:

Prefix	Suffix
<a href="	http
empty	">

End boundary detector:

text

match(es)?

Combining Detectors

Begin boundary detector:

Prefix	Suffix
<a href="	http
empty	">

End boundary detector:

text

↑
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

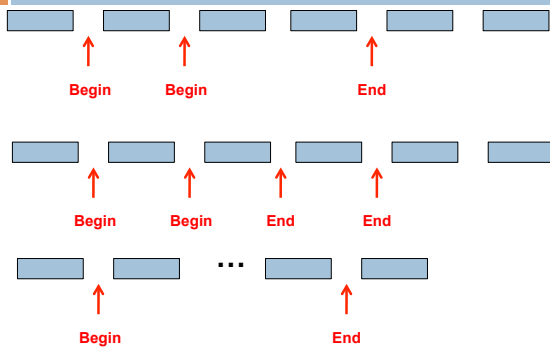
→ POSITIVE (1)

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

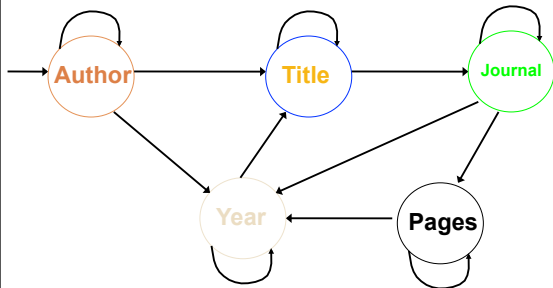
Predict a sequence of tags

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.

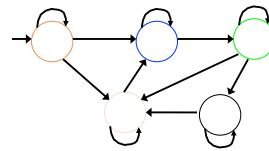
Ideas?

Hidden Markov Models (HMMs)



HMM: Model

- States: x_i
- State transitions: $P(x_i | x_j) = a[x_i | x_j]$
- Output probabilities: $P(o_i | x_i) = b[o_i | x_i]$



- 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:

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

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

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

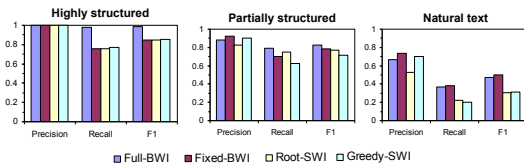
author author year title title title
Fahlman, S. E. (1991) The recurrent cascade

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

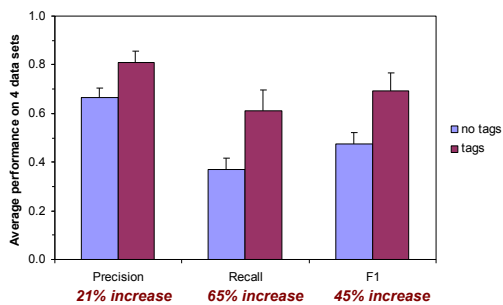
- Instead of altering methods, alter text
- Idea: Add limited grammatical information
 - Run shallow parser over text
 - Flatten parse tree and insert as tags

Example of Tagged Sentence:

Uba2p is located largely in the nucleus.

NP_SEG VP_SEG PP_SEG 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:

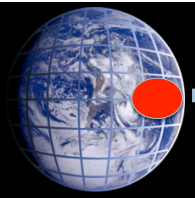


Grab a sample and label:



Approach 1: Old school style

Download the web:



Grab a sample and label:

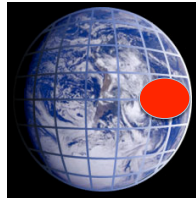


train model:



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

Seed set



author/title pairs



Google™



author/title occurrences in context

Bootstrapping

Seed set



author/title pairs



Google™



author/title occurrences in context



patterns



Bootstrapping

Seed set



author/title pairs



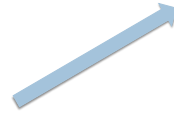
Google™



author/title occurrences in context



patterns



Brin, 1998

(Extracting patterns and relations from the world wide web)

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

Seed books

URL Pattern	Text Pattern
<code>www.sff.net/locus/c.*</code>	<code><LD>title by author (</code>
<code>dns.city-net.com/lanam/awards/hugos/1984.html</code>	<code><D>title</D> by author (</code>
<code>dolphin.upenn.edu/~cummins/texts/af-award.htm</code>	<code>author title (</code>

Patterns

Author	Title
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

New books

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

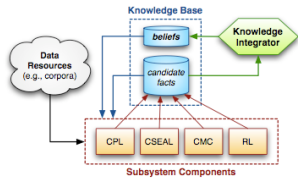
Final list

Henry James	The Europeans
Henry James	The Golden Boat
Henry James	The Portrait of a Lady
Henry James	The Turn of the Screw
Henry James	Turn of the Screw
Henry John Coker	Tracks of a Rolling Stone
Henry K. Howe	Landmarks in Christian History
Henry Kissel	Zephyr
Henry Lawson	In the Days When the World Was Wide
Henry Longfellow	The Song of Hiawatha
Henry Miller	Tropic of Cancer
Henry Potroski	Invention On Design
Henry Potroski	The Evolution of Useful Things
Henry Roth	Call It Sleep
Henry Summer Mainze	Ancient Law
Henry Turbeman, Lincoln, Phila	Characteristics of Literature
Henry Van Dyke	Days Off
Henry Van Dyke	The Blue Flower
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 Fairs of the Northwest
Herbert M. Mason, Jr	The Lafayette Bonafille
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. Kitzredge	Ingersoll: A Biographical Appreciation
Herman Hesse	Principles of Brain Functioning
Herman Hesse	Demian
Herman Hesse	Siddhartha
Herman Hesse	Siddhartha
Herman Melville	Bachelors, 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	Types: A Peep at Polynesian Life
Herman Weiss	Sunset Detectives
Herman Wouk	War and Remembrance
Herman Hesse	King's Last Summer
Herman Hesse	Knapf
Herman Hesse	Rosshalde
Herman Hesse	Strange News From Another Star
Heredotus	Historia
Heredotus	The Histories
Herschel Hobbs	The History of Herodotus
Herschel	Plato's Manua
Herschel	Foot Stage: Moon

NELL

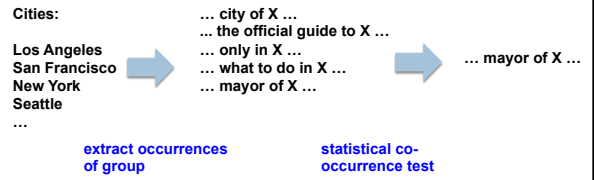
- NELL: Never-Ending Language Learning
 - <http://rtw.ml.cmu.edu/rtw/>
 - continuously crawls the web to grab new data
 - learns entities and relationships from this data
 - started with a seed set
 - uses learning techniques based on current data to learn new information

NELL



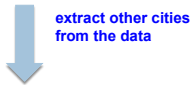
- 4 different approaches to learning relationships
- Combine these in the knowledge integrator
 - ▣ idea: using different approaches will avoid overfitting
- Initially was wholly unsupervised, now some human supervision
 - ▣ cookies are food => internet cookies are food => files are food

An example learner: coupled pattern learner (CPL)



CPL

... mayor of <CITY> ...

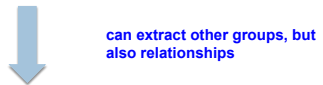


Albuquerque
Springfield
...

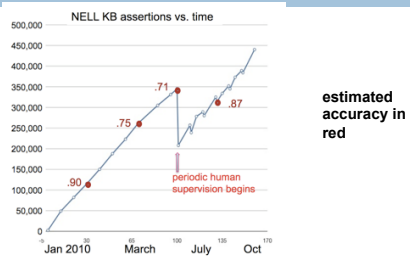
CPL

- Can also learn patterns with multiple groups

... X is the mayor of Y ...
... X plays for Y ...
... X is a player of Y ...



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