Information Extraction

Introduction to Natural Language Processing CMPSCI 585, Spring 2004 University of Massachusetts Amherst



Andrew McCallum

Main Points

- Why IE?
- Components of the IE problem and solution
- Approaches to IE segmentation and classification
 - Sliding window
 - Finite state machines
- IE for the Web
- · Semi-supervised IE
- · Next time: relation extraction and coreference

File Edit View Go Communicator

🗶 Internet 🖆 Lookup 🖆 New&Cool 🥒 Netcaster

🛫 Bookmarks 🙏 Location: [http://www.camp.ca/cgi-shl/dbal.exe?template=/camp/d

N

• Optional class: CRFs for IE & coreference

Query to General-Purpose Search Engine: +camp +basketball "north carolina" "two weeks"

1 4 6 7 1 A 2 4 H A A 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Domain-Specific Search	Empty fields are ignored in the search.
The sector and the sector between the sector and th		Nilhous in Linktod States of America de very unut to
Bookinans & Ebcalon. arcovisca. com/cgi=bin/query/pg-qaki=nkaq=azi /	Engine	camp?
AltaVista found about 94867 Web pages for you. <u>Refine your search</u>		North Carolina 🔤
1. Parent Information - Camp High Rocks Amazon.com suggests		▶ What type of Camp do you want to go to?
rarent Guide – Lang High Hocks, Inis page is for Farents of Lang High Hocks campers, but portions can be useful for any parents of children that attend URL: higherocis.com/High21.htm		GENDER: Co-ed
Last modified 20-Jun-98 - page size 21K - in English [Translate]		ORGANIZATION: Doesn't Matter 🗆
2. [7/15/98] Tim Stevens: North Carolina players shine at ABCD basketball camp		STAY TYPE:
Ine N&U: Sports		Duay Camp
Last modified 22-Sep-98 - page size 14K - in English [Translate]		Residential Camp
3. Rasheed Wallace signs six-year. \$80 million contract		Especially Camp
Rasheed Wallace signs six-year, \$80 million contract. Copyright ⊚ 1997 Nando net Copyright ⊚ 1997 The Associat		Tours & Adventures
Press, For a complete roundup of IBI : weide ending head head head head head head head head		Outdoor Education
Last modified 17-Aug-98 - page size 5K - in English [<u>Translate</u>]		Conference Site
4 As WNRA makie Reid the right fit for Charlette (05/13/98, CHN)		Adult Carm
nbsp; Home Classifieds News Sports Arts Calendar. Chapel Hill News Real Estate Section Opinion Around		
Town, As WNBA rookie, Reid the right		How old are you?
Last modified 9-Aug-96 - page size 5K - in English [Translate.]		1
E May 21 1002 Negelation - St Androny's Church		How much do you want to spend? (per week)
A-Spire. St. Andrew's Presbyterian Church, Kitchener, Ontario. Vol. 6 No. 7. Friday, May 22,1998. In this Issue:		Doesn't Matter 🗆
Twenty-two new elders ordained Camp		Are you looking for a specialty camp?
Last modified 21-Jun-86 page size 20K - in English [Translate.]		Football
		Wind Surfing
b. vorm Carolina State: 1997 rai's Guide to ACC BASKET BALL The Players. ISHUA JABASS BENJAMIN. Herb Sendek sava he knows what it's like to edure "humbling. ego-bruisin		Basketball
times as a basketball player. The		Canceing -
URL: www.nando.net/SportServer/FansGuideIncsulplayers.html		Aquatic
Zast mouneu 27 - Aug-30 - Page size 411 - 11 chigan [11316402]		How long do you want to go for?





Example: The Problem



Example: A Solution



Extracting Job Openings from the Web



s I.S.		• Employers • Support 🔺
ice.	Fetch Your Next Job Here" Return to Results Modify Search New	Resource Center Search
Servi	Degrees Online Online & Project Mgt. Click here to e-mail your resume to 10 of Head Hunters with. ResumeZapper.com Online & Project Mgt. ResumeZapper.com	DOD's by the setting of the setting
ti e g	1 - 25 of 47 jobs shown below	1 2 Next >
	Search these results for: Search tips	Show Jobs Posted: For all time periods
L L Č	View: Brief <u>Detailed</u> Web Jobs: FlipDog technology bas found these jobs on thousands of employer	Web sites.
	Food Pantry Workers at Lutheran Social Services	October 11, 2002 Archbold, OH
<pre>2</pre> 2	Cooks at Lutheran Social Services	October 11, 2002 Archbold, OH
ti 9 g (Bakers Assistants at Fine Catering by Russell Morin	October 11, 2002 Attleboro, MA
	Baker's Helper at Bird-in-Hand	October 11, 2002 United States
o e a	Assistant Baker at Gourmet To Go	October 11, 2002 Maryland Heights, MO
	Host/Hostess at Sharis Restaurants	October 10, 2002 Beaverton, OR
	Cooks at Alta's Rustler Lodge	October 10, 2002 Alta, UT
	Line Attendant at Sun Valley Coporation	October 10, 2002 Huntsville, UT
	Food Service Worker II at Garden Grove Unified School District	October 10, 2002 Garden Grove, CA
	Night Cook / Baker at SONOCO	October 10, 2002 Houma, LA
	Cooks/Prep Cooks at GrandView Lodge	October 10, 2002 <u>Nisswa, MN</u>
	Line Cook at Lone Mountain Ranch	October 10, 2002 Big Sky, MT
	Production Baker at Whole Foods Market	October 08, 2002 Willowbrook, IL
	Cake Decorator/Baker at Mandalay Bay Hotel and Casino	October 08, 2002 Las Vegas, NV
	Shift Supervisors at Brueggers Bagels	October 08, 2002 Minneapolis, MN

Job Openings:

Data Mining the Extracted Job Information



What is "Information Extraction"



What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the opensource concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"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...



What is "Information Extraction"

As a family Information Extraction = of techniques: segmentation + classification + clustering + association October 14, 2002, 4:00 a.m. PT For years, Microsoft Corporation CEO Bill **Microsoft Corporation** Gates railed against the economic philosophy of open-source software with Orwellian fervor, CEO denouncing its communal licensing as a "cancer" that stifled technological innovation. Bill Gates Microsoft Today, Microsoft claims to "love" the opensource concept, by which software code is Gates made public to encourage improvement and Microsoft development by outside programmers. Gates **Bill Veghte** himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Microsoft Windows operating system--to select VP customers. **Richard Stallman** "We can be open source. We love the concept of shared source," said Bill Veghte, a founder Microsoft VP. "That's a super-important shift Free Software Foundation for us in terms of code access." Richard Stallman, founder of the Free Software Foundation, countered saying...

What is "Information Extraction"

As a family of techniques:

Information Extraction =

segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the opensource concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"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...

Microsoft Corporation CEO **Bill Gates**

Microsoft Gates Microsoft **Bill Veghte** Microsoft VP

Richard Stallman founder **Free Software Foundation**

What is "Information Extraction"

As a family Information Extraction = of techniques: segmentation + classification + association + clustering October 14, 2002, 4:00 a.m. PT For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation. Today, Microsoft claims to "love" the opensource concept, by which software code is made public to encourage improvement and development by outside programmers. Gates

himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"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..

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is "Information Extraction"

As a family of techniques: Information Extraction = segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the opensource concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"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 in Context



IE History

Pre-Web

- · Mostly news articles
 - De Jong's FRUMP [1982]
 - · Hand-built system to fill Schank-style "scripts" from news wire
 - Message Understanding Conference (MUC) DARPA ['87-'95]. TIPSTER ['92-'96]
- Most early work dominated by hand-built models
 - E.g. SRI's FASTUS, hand-built FSMs.
 - But by 1990's, some machine learning: Lehnert, Cardie, Grishman and then HMMs: Elkan [Leek '97], BBN [Bikel et al '98]

Web

- AAAI '94 Spring Symposium on "Software Agents"
 - Much discussion of ML applied to Web. Maes, Mitchell, Etzioni.
- Tom Mitchell's WebKB, '96
 - Build KB's from the Web.
- Wrapper Induction
 - Initially hand-build, then ML: [Soderland '96], [Kushmeric '97],...

What makes IE from the Web Different?

Less grammar, but more formatting & linking

Newswire

new grammar.



Landscape of IE Tasks (1/4): **Pattern Feature Domain**

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Non-grammatical snippets, rich formatting & links

Barto, A	ndrew G.	(413) 545-2109	barto@cs.umass.edu	CS276
Prof Con mot cont	essor. nputational neuroscience or control, artificial neu rol, motor development.	e, reinforcement le ral networks, adap	arning, adaptive tive and learning	<u>a</u> ()
Berger,	Emery D.	(413) 577-4211	emery@cs.umass.edu	CS344
Assi	stant Professor.			a ()
Brock, C	Diver	(413) 577-033	4 oli@cs.umass.edu	CS244
Assi	stant Professor.			a
Clarke,	Lori A.	(413) 545-1328	clarke@cs.umass.edu	CS30
Prof Soft and	essor. ware verification, testin design.	g, and analysis; so	ftware architecture	a
Cohen,	Paul R.	(413) 545-3638	cohen@cs.umass.edu	CS27
Prof Plan intel	essor. ning, simulation, natur: ligent data analysis, into	al language, agent- elligent user interf	based systems, aces.	<u>a</u>

Dr. Steven Dr. Minton is Association the founder Intelligence Minton was project lead Institute. A (Carnegie M Principal In- taught at Sta Frank Huyl Mr. Huybrec	An, a	n ns			
8:30 - 9:30 AM	Invited Talk: P Joseph Y. Halpe	lausibility Measures	: A General App	roach for Repres	enting Uncerta
9:30 - 10:00 AM	Coffee Break				
9:30 - 10:00 AM 10:00 - 11:30 AM	Coffee Break Technical Paper	r Sessions:			
9:30 - 10:00 AM 10:00 - 11:30 AM Cognitive Robotics	Coffee Break Technical Paper Logic Programming	r Sessions: Natural Language Generation	Complexity Analysis	Neural Networks	Games

Grammatical sentences

and some formatting & links

Press

Dr. Steven Minton - Founder/CTO

and Bert

Landscape of IE Tasks (3/4): **Pattern Complexity**

E.g. word patterns:

Closed set	Regular set
U.S. states	U.S. phone numbers
He was born in <u>Alabama</u>	Phone: (413) 545-1323
The big Wyoming sky	The CALD main office can reached at <u>412-268-1299</u>
Complex pattern U.S. postal addresses	Ambiguous patterns, needing context and many sources of evide
University of Arkansas <u>P.O. Box 140</u> <u>Hope, AR 71802</u>	Person names was among the six house sold by <u>Hope Feldman</u> that
Headquarters: <u>1128 Main Street, 4th Floor</u> Cincinnati, Ohio 45210	Pawel Opalinski, Software Engineer at WhizBang Labs

be

nce

es year.

Landscape of IE Tasks (2/4): **Pattern Scope**

Web site specific	Genre specific	Wide, non-specific		
Formatting Amazon.com Book Pages	Layout Resumes	Language University Names		
ALL AND ALL AN	Jason D. M. Rennie Mesekusets (1974) MT AL JA NEA 733 MT AL JA NEA 733 MT JA	8:30 - 9:30 AM Invited Talk: Plausibility Measures: A General Approv 2:30 - 10:00 AM Conference on the Invited Invited International Conference Income 10:00 - 11:30 AM Technical Paper Sensions: Cognitive Logic Natural Language Complexity Natural Natures		
Chine Learning Chine Learning Search Search Search Search Search Search NEW Super Saver Shipping FREE of Learning Investigation (Search)	Assume there exists the advanced autyris of data for the purposes of clearification. Statistic torrection is to the advanced autyris of data for the purposes of clearification. Statistic torrection is the advanced of the statistic torrection is advanced autor L Douglas Baker Hinto Advance Autor Hinto Advance Hinto Advanc	737: A Logical 116: ASystem 738: Title 417: Lark 50: 11 Account of Caular Poblem Generation forn Nats: E and Topological Solving, Machine Translated Complexity of Omplexity of Omplexity of Omplexity of Omplexity of Caular Solving, Maps Remotine M National Solving, Maps Remotine M National Solving, And Braymann Manor multiple Machine Translation Manormality K Anomal Solas, Macrosci Caula, and Ever from Macrosci Caula,		
yoort mestore a mey nom Pert	Circle Price Priceogr. Pril (103) (11) 481-868 Circle Price (11) 481-868	Dr. Steven Minton - Founder(CTO Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a		
eat Buy the book who + + + + + + + + + + + + + + + + + + +	PD. Complet Source, progress MS. Complet Source, progress Technical Weinstry of Refin Borin, Genzer Distange Fidery (2023) Biniversity of Michigan American Complete Sources, progressing, 2049 8.5. E. Complete Ediptically, Dama Caril Laux, 1905	project leader at USC's Information Sciences Institute, A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at INASA Arnes and taught at Stanford, UC Berkeley and USC.		
+ Buy both now1 * Buy both now1	Eponiance Carrogle Heliko Usiversity Idef presset Inn carrently parsing my disentitive meanch: a hierarchical probabilitie used for prevent placesing my disentitive meanch: a hierarchical probabilitie used for prevent placesing my disentitive users the disentitive data of the Toget Totel Structure Contract	Frank Huybrechts - COO Mr. Huybrechts has over 20 years of		

Landscape of IE Tasks (4/4): **Pattern Combinations**

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity	Binary relationship	N-ary record
Person: Jack Welch	Relation: Person-Title Person: Jack Welch	Relation: Succession Company: General Electric
Person: Jeffrey Immelt	<i>Title:</i> CEO	<i>Title:</i> CEO <i>Out:</i> Jack Welsh <i>In:</i> Jeffrey Immelt
Location: Connecticut	Relation: Company-Location Company: General Electric Location: Connecticut	

"Named entity" extraction

Evaluation of Single Entity Extraction

TRUTH:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

PRED:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.



State of the Art Performance

- Named entity recognition
 - Person, Location, Organization, ...
 - F1 in high 80's or low- to mid-90's
- · Binary relation extraction
 - Contained-in (Location1, Location2) Member-of (Person1, Organization1)
 - F1 in 60's or 70's or 80's
- Wrapper induction
 - Extremely accurate performance obtainable
 - Human effort (~30min) required on each site

Landscape of IE Techniques (1/1): Models



Sliding Windows

Extraction by Sliding Window

GRAND CHALLENGES FOR MACHINE LEARNING Jaime Carbonell School of Computer Science Carnegie Mellon University

E.g. Looking for seminar location

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

Extraction by Sliding Window



E.g.

seminar

location

CMU UseNet Seminar Announcement

Extraction by Sliding Window

Extraction by Sliding Window



A "Naïve Bayes" Sliding Window Model

[Freitag 1997]

<mark>00 : p</mark>	om Place :	Wean Hall	Rm 5409	Speaker :	Sebastian	Thrun
W _{t-m}	W _{t-1}	W _t	W_{t+n}	W_{t+n+1}		W_{t+n+m}
	prefix	conte	ents		suffix	
P("Wean H	Hall Rm 5409'	' = LOCATIO	ON) =			
				Ä		
Prior probabilit of start position	y Prior probabi n of length	lity Proba prefix	ability words	Probability contents words	Prol s suffi	oability k words
Try all start posi	itions and reasor	able lengths	Estimate th counts fror	iese probabilitie n labeled trainir	es by (smoothe ng data.	d)

If P("Wean Hall Rm 5409" = LOCATION) is above some threshold, extract it.

Other examples of sliding window: [Baluja et al 2000] (decision tree over individual words & their context)

"Naïve Bayes" Sliding Window Results

Domain: CMU UseNet Seminar Announcements

GRAND CHALLENGES FOR MACHINE LEARNING	
Jaime Carbonell School of Computer Science Carnegie Mellon University 3:30 pm 7500 Wean Hall	<u>Field</u> Person N
achine learning has evolved from obscurity n the 1970s into a vibrant and popular iscipline in artificial intelligence during he 1980s and 1990s. As a result of its uccess and growth, machine learning is volving into a collection of related isciplines: inductive concept acquisition, nalytic learning in problem solving (e.g. nalogy, explanation-based learning), earning theory (e.g. PAC learning), genetic lgorithms, connectionist learning, hybrid	Location: Start Time
ystems, and so on.	

F1 ame: 30% 61% 98% e:

SRV: a realistic sliding-window-classifier **IE system**

[Frietag AAAI '98]

- What windows to consider?
 - all windows containing as many tokens as the shortest example, but **no more** tokens than the longest example
- How to represent a classifier? It might:
 - Restrict the length of window;
 - Restrict the vocabulary or formatting used before/after/inside window;
 - Restrict the **relative order** of tokens:
 - Etc…

<title>Course Information for CS213</title> <h1>CS 213 C++ Programming</h1>

"A token followed by a 3-char numeric token just after the title"

SRV: a rule-learner for sliding-window classification

Top-down rule learning:

let RULES = ;;

```
while (there are uncovered positive examples) {
  // construct a rule R to add to RULES
  let R be a rule covering all examples;
  while (R covers too many negative examples) {
     let C = argmax<sub>c</sub> VALUE( R, R Æ C, uncoveredExamples)
        over some set of candidate conditions C
     let R = R Æ C:
  }
  let RULES = RULES [ {R};
}
```

SRV: a rule-learner for sliding-window classification

Search metric: SRV algorithm greedily adds conditions to maximize "information gain" of *R*

VALUE(R,R',Data) = IData|*p ($p \log p - p' \log p'$) where p(p') is fraction of data covered by R(R')

To prevent overfitting:

rules are built on 2/3 of data, then their false positive rate is estimated with a Dirichlet on the 1/3 holdout set.

Candidate conditions: ...

SRV: a rule-learner for sliding-window classification

- Primitive predicates used by SRV:
 - token(X,W), allLowerCase(W), numerical(W), …
 - nextToken(W,U), previousToken(W,V)
- HTML-specific predicates:
 - inTitleTag(W), inH1Tag(W), inEmTag(W),...
 - emphasized(W) = "inEmTag(W) or inBTag(W) or …"
 - tableNextCol(W,U) = "U is some token in the column after the column W is in"
 - tablePreviousCol(W,V), tableRowHeader(W,T),...

Learning "first-order" rules

- A sample "zero-th" order rule set: (tok1InTitle Æ :tok1StartsPara Æ tok2triple)
 Ç (prevtok2EqCourse Æ prevtok1EqNumber) Ç ...
- First-order "rules" can be learned the same way—with additional search to find best "condition" phrase(X) Ã firstToken(X,A), :startPara(A), nextToken(A,B), triple(B) phrase(X) Ã firstToken(X,A), prevToken(A,C), eq(C,'number'), prevToken(C,D), eq(D,'course')
- Semantics: " $p(X) \tilde{A} q(X), r(X,Y), s(Y)$ " = "{X : 9 Y : q(X) Æ r(X,Y) Æ s(Y)}"

SRV: a rule-learner for sliding-window classification

- Non-primitive "conditions" used by SRV:
 - $every(+X, \underline{f}, \underline{c}) = 8 W2X : f(W) = c$
 - variables tagged "+" must be used in earlier conditions
 - underlined values will be replaced by constants, e.g., "every(X, isCapitalized, true)"
 - $some(+X, W, < \underline{f_1, ..., f_k} >, \underline{g}, \underline{c}) = \mathbf{g} W: g(f_k(...(f_1(W)...)) = c$
 - e.g., some(X, W, [prevTok,prevTok],inTitle,false)
 - set of "paths" <f₁,...,f_k> considered grows over time.
 - tokenLength(+X, <u>relop</u>, <u>c</u>):
 - position(+W,direction,relop, c):
 - e.g., tokenLength(X,>,4), position(W,fromEnd,<,2)

Utility of non-primitive conditions in greedy rule search

Greedy search for first-order rules is hard because useful conditions can give no immediate benefit:

phrase(X) **Ã** token(X,A), prevToken(A,B),inTitle(B), nextToken(A,C), tripleton(C)

<title>Course Information for CS213</title> <h1>CS 213 C++ Programming</h1>	"A token followed by a 3-char numeric token just after the title"
courseNumber(X) Ã tokenLength(X,=,2),	Non-primitive conditions

every(X, in litle, false), some(X, A, <previousToken>, inTitle, true), some(X, B, <>. tripleton, true)

make greedy search easier



Rapier: an alternative approach

[Califf & Mooney, AAAI '99]

A bottom-up rule learner:

initialize RULES to be one rule per example;

repeat {

randomly pick N pairs of rules (R_i, R_i) ;

let $\{G_1, G_N\}$ be the consistent pairwise generalizations;

let G* = argmin_G COST(G,RULES);

let RULES = RULES [{G*} – {R': G* ¶ R'}

where COST(G,RULES) = size of RULES- {R': G ¶ R'} and "G¶ R" means every example matching G matches R

Rapier: an alternative approach

- Combines top-down and bottom-up learning
 - Bottom-up to find common restrictions on content
 - Top-down greedy addition of restrictions on context
- Use of part-of-speech and semantic features (from WORDNET).
- Special "pattern-language" based on sequences of tokens, each of which satisfies one of a set of given constraints
 - < <tok2{'ate', 'hit'},POS2{'vb'}>, <tok2{'the'}>, <POS2{'nn'>>

Rapier: results – precision/recall



Rapier – results vs. SRV

System	stime		etime		loc		speaker	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
RAPIER	93.9	92.9	95.8	94.6	91.0	60.5	80.9	39.4
RAP-WT	96.5	95.3	94.9	94.4	91.0	61.5	79.0	40.0
RAP-W	96.5	95.9	96.8	96.6	90.0	54.8	76.9	29.1
NAIBAY	98.2	98.2	49.5	95.7	57.3	58.8	34.5	25.6
SRV	98.6	98.4	67.3	92.6	74.5	70.1	54.4	58.4
WHISK	86.2	100.0	85.0	87.2	83.6	55.4	52.6	11.1
WH-PR	96.2	100.0	89.5	87.2	93.8	36.1	0.0	0.0

Rule-learning approaches to slidingwindow classification: Summary

- SRV, Rapier, and WHISK [Soderland KDD '97]
 - Representations for classifiers allow restriction of the relationships between tokens, etc
 - Representations are carefully chosen subsets of even more powerful representations based on logic programming (ILP and Prolog)
 - Use of these "heavyweight" representations is complicated, but seems to pay off in results
- Can simpler representations for classifiers work?

BWI: Learning to detect boundaries

[Freitag & Kushmerick, AAAI 2000]

- Another formulation: learn three probabilistic classifiers:
 - START(i) = Prob(position i starts a field)
 - END(j) = Prob(position j ends a field)
 - LEN(k) = Prob(an extracted field has length k)
- Then score a possible extraction (i,j) by START(i) * END(j) * LEN(j-i)
- *LEN(k)* is estimated from a histogram

BWI: Learning to detect boundaries

- BWI uses **boosting** to find "detectors" for *START* and *END*
- Each weak detector has a *BEFORE* and *AFTER* pattern (on tokens before/after position *i*).
- Each "pattern" is a sequence of tokens and/or wildcards like: anyAlphabeticToken, anyToken, anyUpperCaseLetter, anyNumber, ...
- Weak learner for "patterns" uses greedy search (+ lookahead) to repeatedly extend a pair of empty *BEFORE,AFTER* patterns

Problems with Sliding Windows and Boundary Finders

- Decisions in neighboring parts of the input are made independently from each other.
 - Naïve Bayes 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, and their pairing happens as a separate step.

BWI: Learning to detect boundaries



Finite State Machines

Hidden Markov Models

HMMs are the standard sequence modeling tool in genomics, music, speech, NLP, ...



Observation (emission) probabilities: $P(o_t|s_t)$ Usually a multinomial over atomic, fixed alphabet Training:

Maximize probability of training observations (w/ prior)

IE with Hidden Markov Models

Given a sequence of observations:



Any words said to be generated by the designated "person name" state extract as a person name:

Person name: Lawrence Saul

HMM Example: "Nymble" [Bikel, et al 1998], [BBN "IdentiFinder"] **Task: Named Entity Extraction** Transition Observation Perso probabilities probabilities end-of- $P(o_t | s_t, s_{t-1})$ $P(s_t | s_{t-1}, o_{t-1})$ sentence start-ofsentence Org or $P(o_t | s_t, o_{t-1})$ (Five other name classes) Back-off to: Back-off to: $P(s_t | s_{t-1})$ $P(o_t | s_t)$ Other $P(s_t)$ $P(o_t)$ Train on 450k words of news wire text. **Results:** Case Language F1. Mixed English 93% English 91% Upper 90% Mixed Spanish

Other examples of shrinkage for HMMs in IE: [Freitag and McCallum '99]

HMMs for IE: A richer model, with backoff

Simple HMM structure for IE

HMMs for IE: Augmented finite-state structures with linear interpolation

- 4 state types:
 - Background (generates words not of interest),
 - Target (generates words to be extracted),
 - **P**refix (generates typical words preceding target)
 - Suffix (words typically following target)



- Properties:
 - Extracts one type of target (e.g. target = person name), we will build one model for each extracted type.
 - Models different Markov-order n-grams for different predicted state contexts.
 - even thought there are multiple states for "Background", state-path given labels is unambiguous. Therefore model parameters can all be computed using counts from labeled training data

More rich prefix and suffix structures

- In order to represent more context, add more state structure to prefix, target and suffix.
- But now overfitting becomes more of a problem.



Figure 1: Two example HMM structures. Circle nodes represent non-target states; hexagon nodes represent target states.

Linear interpolation across states

uniform

- Is defined in terms of some hierarchy that represents the expected similarity between parameter estimates, with the estimates at the leaves
- Shrinkage based parameter estimate in a leaf of the hierarchy is a linear interpolation of the estimates in all distributions from the leaf to ist root context

prefix

- Shrinkage smoothes the distribution of a state towards that of states that are more data-rich
- It uses a linear combination of probabilities

suffix

Evaluation of linear interpolation

· Data set of seminar announcements.

	speaker	location	stime	etime
None	0.513	0.735	0.991	0.814
Uniform	0.614	0.776	0.991	0.933
Global	0.711	0.839	0.991	0.595
Hier.	0.672	0.850	0.987	0.584

Table 4: Effect on F1 performance of different shrinkage configurations on four seminar announcement fields, given a topology with a window size of four and four parallel length-differentiated target paths.

IE with HMMs: Learning Finite State Structure

Information Extraction from Research Papers

References

Leslie Pack Kaelbling, Michael L. Littman and Andrew W. Moore. Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research, pages 237-285, May 1996.

Headers

Submitted 9/95; published 5/96

LPK@CS.BROWN.EDU

MLITTMAN@CS.BROWN.EDU

Reinforcement Learning: A Survey

Leslie Pack Kaelbling. <u>Michael L. Littman</u> Computer Science Department, Boz 1910, Brown University Providence, RI 02912-1910 USA

nal of Artificial Intelligence Research 4 (1996) 237-285

Andrew W. Moore Smith Hall 221, Carnegic Mellon University, 5000 Forbes Avenue Filtiburgh, FA 15213 USA

Abstract

This paper surveys the field of reinforcement learning from a computer-science perspective. It is written to be accessible to researchers familiar with machine learning. Both the historical basis of the field and a broad sciencition of current works are summarized. Reinforcement learning is the problem faced by an agent that learns behavior through the land and error interactions with an dynamic environment. The work coefficient lear has a of the word "reinforcement." The paper discusses central lanes of reinforcement learning, including trading of exploration and exploitation, exhibiling the foundations of the field via Maxiev decision theory, hearning from delayed reinforcement, enasting empirical models to accelerate learning, making use of generalization and hierarchy, and coping with hidden state. It concludes with a survey of some implemented systems and an assessment of the practical utility of current methods for reinforcement, and.

1. Introduction

Reinforcement learning dates back to the early days of cybernetics and work in statistics,

Information Extraction with HMMs



[Seymore & McCallum '99]

Importance of HMM Topology

- Certain structures better capture the observed phenomena in the prefix, target and suffix sequences
- Building structures by hand does not scale to large corpora
- Human intuitions don't always correspond to structures that make the best use of HMM potential

Structure Learning

Two approaches

- Bayesian Model Merging Neighbor-Merging V-Merging
- Stochastic Optimization Hill Climbing in the possible structure space by spiltting states and gauging performance on a validation set

Bayesian Model Merging

Maximally Spesific Model



Bayesian Model Merging

 Iterates merging states until an optimal tradeoff between fit to the data and model size has been reached

 $\mathsf{P}(\mathsf{M} \mid \mathsf{D}) \thicksim \mathsf{P}(\mathsf{D} \mid \mathsf{M}) \: \mathsf{P}(\mathsf{M})$

M = Model D = Data





P(D | M) can be calculated with the Forward algorithmP(M) model prior can be formulated to reflect a preference for smaller models

HMM Emissions



HMM Information Extraction Results

Per-word error rate	Headers	References
One state/class Labeled data only	0.095	
Model Merging Labeled data only	0.087 <i>(8% b</i> e	tter)
One state/class +BibTeX data	0.076 (20% b	etter)
Model Merging +BibTeX	0.071 (25% b	<i>etter)</i> 0.066

Stochastic Optimization

- Start with a simple model
- Perform hill-climbing in the space of possible structures
- Make several runs and take the average to avoid local optima



State Operations

- Lengthen a prefix
- Split a prefix
- · Lengthen a suffix
- Split a suffix
- Lengthen a target string
- Split a target string
- Add a background state



LearnStructure Algorithm

procedure LearnStructure(LabeledSet, Ops) ValidSet $\leftarrow 1/3$ of LabeledSet TrainSet ← LabeledSet – ValidSet CurModel \leftarrow the simple model Keepers ← {CurModel} $I \leftarrow v$ while I < 20 and CurModel has fewer than 25 states Candidates $\leftarrow \{M | M \in op(CurModel) \land op \in Ops\}$ for $M \in Candidates$ $score(M) \leftarrow average of 3 runs trained on$ TrainSet and scored for F1 on ValidSet CurModel $\leftarrow M \in$ Candidates with highest score Keepers \leftarrow Keepers \cup {CurModel} $I \leftarrow I + 1$ for $M \in \mathsf{Keepers}$ $score(M) \leftarrow average F1$ from 3-fold cross-validation on LabeledSet return $M \in$ Keepers with highest score

Part of Example Learned Structure

Locations

Speakers



Accuracy of Automatically-Learned Structures

	speaker	location	acquired	dlramt	title	company	conf	deadline	Average
Grown HMM	76.9	87.5	41.3	54.4	58.3	65.4	27.2	46.5	57.2
vs. SRV	+19.8	+16.0	+1.1	-1.6	—	—	—	—	+8.8
vs. Rapier	+23.9	+14.8	+12.5	+15.1	-11.7	+24.9	—	—	+13.3
vs. Simple HMM	+24.3	+5.6	+14.3	+5.6	+5.7	+11.1	+15.7	+6.7	+11.1
vs. Complex HMM	-2.1	+6.7	+7.5	-0.3	-0.3	+19.1	+0.0	-6.8	+3.0

Table 2:	Difference in	1 F1	performance	between	the	HMM	using	a learned	structure	and	other	methods.	The $+$
numbers	indicate how	mucl	h better our	Grown H	MM	did th	an the	alternativ	re method.				

Limitations of HMM/CRF models

- HMM/CRF models have a linear structure
- Web documents have a **hierarchical** structure
 - Are we suffering by not modeling this structure more explicitly?
- How can one learn a hierarchical extraction model?
 - Coming up: STALKER, a hierarchical wrapperlearner
 - But first: how do we train wrapper-learners?

Tree-based Models

• Extracting from one web site

- Use *site-specific* formatting information: e.g., "the JobTitle is a boldfaced paragraph in column 2"
- For large well-structured sites, like parsing a formal language

Extracting from many web sites:

- Need general solutions to entity extraction, grouping into records, etc.
- Primarily use content information
- Must deal with a wide range of ways that users present data.
- Analogous to parsing natural language

• Problems are complementary:

- Site-dependent learning can collect training data for a siteindependent learner
- Site-dependent learning can boost accuracy of a site-independent learner on selected key sites









STALKER: Hierarchical boundary finding

[Muslea, Minton & Knoblock 99]

- Main idea:
 - To train a hierarchical extractor, pose a series of learning problems, one for each node in the hierarchy
 - At each stage, extraction is simplified by knowing about the "context."



Neural Information Proce	ssing Systems 2002 (NIPS*2002) Main Page - Microsoft Internet Explorer		1
Ele Edit View Favorites	Tools Help		
🌏 Back 👻 🕤 👻 🛃 🔮	🏠 🔎 Search 👷 Favorites 🜒 Media 🛛 😥 🖓 🐨 💌 🕅 💥 🚳	Address Links »	•
Google - nips-2002	💌 💏 Search Web 🛛 😨 Search Site 🛛 📢 Page Info 👻 🔂 Up 🔹 🔗 Highlight 🛛 🖏 r	nips 2002	
		nere.]
(5	EEOPE-null AETER-(< li >))		
	SEPORE-IIII, AFTER-(I,,,))	_	
Tutorial 🔃	picsClick for abstracts and bios.		
	Michael Kearns, University of Pennsylvania:		
	Computational Game Theory		
	Sebastian Seung, Howard Hughes Medical		Ш.
	Institute and MIT: Neural Integrators		Į.
	<u>Yair Weiss</u> , Hebrew University, <u>Jianbo Shi</u> ,		
	Carnegie Mellon University, and Serge Belongie.		
	UC San Diego: Eigenvector Methods for		
	Clustering and Image Segmentation		
	Dishard M Korn LIC Barkelay and International		
	Computer Science Institute Mathematical		
	Computer Science Institute: Mathematical,		
	Statistical and Algorithmic Challenges from		
	Genomics and Molecular Biology		
	Martin Cooke, University of Sheffield:		
	Computational Auditory Scene Analysis in		
	Comparation and Addition y Coonto Analysis in	ACHTEC - VistualDecku	

• Incural throumation Proce	ssing Systems 2002 (MPS*2002) Main Page - Microsoft Internet Explorer	
ile Edit View Favorites	Iools Help	
	👔 🔑 Search 👷 Favurites 😋 media 🤕 🔝 👻 🔤 🚮 💆 🥨 🖉 🚺	Aguress Links
30034C + mps-2002		ips 2002
	Categories	* Online NIPS
		papers from
(BEFORE=(:), AFTER=null)	previous years?
Tutorial To	pies Click for abstracts and bios	<u>here</u> .
1000100100	predention for <u>paratitional as the prede</u> t	
	Michael Kellrns, University of Pennsylvania:	
	Computational Game Theory	Important
		Dates
	Sebastian Seund, Howard Hughes Medical	° Tutorials: Dec 9.
	Institute and MIT: Neural Integrators	2002
		* Conference: <u>Dec</u>
	Yair Weiss, Hebrew University, Jianbo Shi,	<u>10, 11, 12 (until</u>
	Carnegie Mellon University, and Serge Belongie,	<u>noon).</u> 2002
	UC San Diego: Eigenvector Methods for	*Workshops: <u>Dec</u>
	Clustering and Image Segmentation	12 (evening
		reception), 13, 14,
	Dishard M. Kam, U.O. Dashalar, and betamatismai	ZUUZ * Dantar
	Richard IVI. Karp, UC Berkeley and International Operational States Colored and International	Poster
	Computer Science Institute: Mathematical,	10 11 2002
	Statistical and Algorithmic Challenges from	* Demonstrations
	Genomics and Molecular Biology	Dec 10, 11, 2002
		* Camera-ready
	Martin Cooke. University of Sheffield:	papers due:
	Computational Auditory Scene Analysis in	January 10, 2003
		ACHITEC - VistualDack -



🚈 Neural Information Processing Systems 2002 (NIP5*2002) Main Page - Microsoft Internet Explorer	<u>_ </u> ×
Elle Edit Yiew Favorites Iools Help	
😋 Back 🔻 🕤 👻 😰 🐔 🔎 Search 👷 Favorites 😵 Media 🤣 😥 🔹 💹 👻 🖾 🖏	Address Links »
Google 🗸 nips-2002 💽 💏 Search Web 😻 Search Site 🚺 Page Info 🔻 🛅 Up 🔹 🖉 Highlight 🛛 👸 r	vips 2002
Categories Tutorial TopicsClick for <u>abstracts and bios</u> .	■ Online NIPS papers from previous years? <u>here</u> .
Michael Kearns, University of Pennsylvania: Computational Game Theory	Important Dates
(BEFORE=(:), AFTER=null)	* Tutorials: <u>Dec 9,</u> 2002
Yair Weiss Hebrew University, <u>Jianbo Shi</u> , Carnegie Melon University, and <u>Serge Belongie</u>	<u>10, 11, 12 (until</u> <u>noon)</u> , 2002
UC San Diego: Eigenvector Methods for Clustering and Image Segmentation	* Workshops: <u>Dec</u> <u>12 (evening</u> reception), 13, 14,
Richard M. Karp, UC Berkeley and International Computer Science Institute: Mathematical, Statistical and Algorithmic Challenges from Genomics and Molecular Biology	2002 * Poster presentations: <u>Dec</u> <u>10, 11</u> , 2002 * Demonstrations: <u>Dec 10, 11</u> , 2002 * Camera-ready
<u>Martin Cooke</u> , University of Sheffield: Computational Auditory Scene Analysis in	papers due: January 10, 2003

Stalker: hierarchical decomposition of two web sites



Stalker: summary and results

- · Rule format:
 - "landmark automata" format for rules
 - E.g.: <a>W. Cohen CMU: Web IE
 - STALKER: BEGIN = SkipTo(<, /, a, >), SkipTo(:)
- Top-down rule learning algorithm
 - Carefully chosen ordering between types of rule specializations
- Very fast learning: e.g. 8 examples vs. 274
- · A lesson: we often control the IE training data!

Learning Formatting Patterns "On the Fly": "Scoped Learning"



Formatting is regular on each site, but there are too many different sites to wrap. Can we get the best of both worlds?

Scoped Learning Generative Model

- 1. For each of the D documents:
 - a) Generate the multinomial formatting feature parameters ϕ from $p(\phi | \alpha)$
- 2. For each of the N words in the document:
 - a) Generate the *n*th category c_n from $p(c_n)$.
 - b) Generate the *n*th word (global feature) from $p(w_n | c_n, \theta)$
 - c) Generate the *n*th formatting feature (local feature) from $p(f_n | c_{n'} \phi)$

$$p(\phi, \mathbf{c}, \mathbf{w}, \mathbf{f}) = p_{\alpha}(\phi) \prod_{n=1}^{N} p(c_n) p_{\theta}(w_n | c_n) p(f_n | c_n, \phi)$$



Inference

Given a new web page, we would like to classify each word resulting in \bm{c} = {c₁, c₂,..., c_n}

 $p(\mathbf{c}|\mathbf{w},\mathbf{f}) = rac{\int \prod_{n=1}^{N} p(w_n|c_n) p(f_n|c_n,\phi) p(c_n) p(\phi) d\phi}{\int \prod_{n=1}^{N} \sum_{c_n} p(w_n|c_n) p(f_n|c_n,\phi) p(c_n) p(\phi) d\phi}$

This is not feasible to compute because of the integral and sum in the denominator. We experimented with two approximations:

- MAP point estimate of ϕ

- Variational inference

🚰 Work at the Y! - Microsoft Internet Explorer prov	rided by WhizBang! Labs	_ 🗆 🗙
Eile Edit View Favorites Tools Help		*
← Back → → → 🎱 🖉 🚮 🕅 🐼 Search 💿 Fav	rorites 🎯 History 🔹 🗃 🐨 = 📃	
Address 2. Documents and Settings (mccallum) Deskto	p\scoped\jobopening-global.html	▼ Links ≫
Family Services Director	Creative, exception and entry working with people's Seeking director for program development, Injunemation and advantisation in Matt poses a Bachelor's Degree in Recreation, Family Studies or related field. Strong interpersonal and organizational salib a musit. Excellent leneths: Service resumes to Java Rion, Dr of Camping and Family Services, North Sukurban YMCA, 2705 Techny Road, Northbrook, L. 80062.	<u>*</u>
Massage Therapist - Male	The North Suburban YMCA is seeking a certified massage therapist to work part time in our men's program center. Flexible hours, y membership, on-ste child care available if needed. Please contact <u>Harian Stritchko by email</u> or call at 847-272- 7250.	
Starbucks Server	Early day, evening and weekend shifts available for in-house cafe serving the Starbuck's product line. An exciting opportunity and membership is included! Contact Sarah Tucker at 847-272-7250 x.213.	
Teacher for ChildCare Center	Part-time 2-6 pm, Monday through Friday. Minimum requirements are 60 college credit hours in Early childhood or Education or similar subject. At least one year experience working with 2-5 year olds. Contact Helen at (847) 272-7250 x222 and fax resume to (847) 272-7587.	
Art Coordinator	Creative? Enjoy working with children? the North suburban Y is looking for an art coordinator for the summer. Call Jane at (847)272-7250 for more information.	
Teachers	Seeking part-time early childhoot tachers for summer or all year. 2-3 mornings per week from 9am-11:15am. Free child care on-site while you work. Free YMCA membership. Callege dagere required in education or related field. Fick up an application at the front deak or call Caryn Shuhman, Child Development Coordinator at (47) 727-7250 x232.	
Group Exercise Personal Training	Interested individuals with proper certification may contact <u>Myleen Signorini</u> at (847) 272-7250 × 217	
Cutomer Service Rep	OVERGUALLIFED APPLY HEREI Hone your stills by working in a friendly environment. The front desk is looking for part time start for work flexible shifts for early weekday mornings, day and evening shifts. Benefits include VMCA membership and babysting during your shift. Please contact <u>Sanah Tucker or Cherry Steward</u> (847) 272-2720 x 213.	
Lifeguards and Swim Instructors	Love to swim? Love kids? Put the two together and make a difference. The North Suburban YMCA is looking for qualified and experienced swim instructors and	_
A Done		My Computer

Global Extractor: Precision = 46%, Recall = 75%

MAP Point Estimate

If we approximate ϕ with a point estimate, ϕ , then the integral disappears and c decouples. We can then label each word with:

$$\hat{c}_n = \arg\max_{c_n} p(w_n | c_n) p(f_n | c_n, \hat{\phi}) p(c_n)$$

A natural point estimate is the posterior mode: a maximum likelihood estimate for the local parameters given the document in question:

$$\hat{\phi} = rg\max_{\phi} p(\phi | \mathbf{f}, \mathbf{w})$$

E-step:

$$p^{(t+1)}(c_n|w_n, f_n; \phi) \propto p^{(t)}(f_n|c_n; \phi) p(w_n|c_n) p(c_n)$$

M-step:

$$\hat{\phi}_{c,f} = p^{(t+1)}(f|c;\phi) \propto \sum_{\{n:c_n=c,f_n=f\}} p^{(t)}(c_n|f_n,w_n)$$

Work at the Y! - Microsoft Internet Explorer prov	ided by WhizBang! Labs	
<u>Eile Edit View Favorites Iools H</u> elp		*
🗢 Back 🔹 🤿 🗸 🙆 Search 🔅 Fav	orites 🎯 History 🛛 🔁 🗃 🖬 + 📃	
Address 🥙 C:\Documents and Settings\mccallum\Desktop	\scoped\jobopening-localglobal.html	▼ Links ≫
Family Services Director	Creative, energetic and enjoy working with people? Seeking director for program development, implementation and administration. Nust possess a Bachedr's Degree in Recretation, Frankly Studies or releader tields. Strong intergers sonal and organizational skills a must. Excellent benefits. Send resumes to Jane Kim, Dr of Camping and Family Services, North Suburban YMCA, 2705 Techny Road, Northbrook, L 60052.	×
Massage Therapist - Male	The North Suburban YMCA is seeking a certified massage therapist to work part time in our men's program center. Flexible hours, y membership, on-ste child care available if needed. Please contact <u>Hartan Strichko by email</u> or call at 847-272- 7250.	
Starbucks Server	Early day, evening and weekend shifts available for in-house cafe serving the Starbuck's product line. An exciting opportunity and membership is included! Contact Sarah Tucker at 847-272-7250 x.213.	
Teacher for ChildCare Center	Part-time 2-6 pm, Monday through Friday. Minimum requirements are 60 college credit hours in Early childhood or Education or similar subject. At least one year experience working with 2-5 gen clids. Contact Helen at (847) 272-7250 x222 and fax resume to (847) 272-7587.	
Art Coordinator	Creative? Enjoy working with children? the North suburban Y is looking for an art coordinator for the summer. Call Jane at (847)272-7250 for more information.	
Teachers	Seeking part-time early childhood teachers for stummer or all year. 2-3 mornings per week from Sam 11:15m, Three child care on-tale while you work. Free YMCA membership, College degree required in education or related hild. Pick up an application at the front desk or call Caryn Shutman, Child Development Coordinator at (847) 272-2732 v323.	
Group Exercise Personal Training	Interested individuals with proper certification may contact $\underline{\text{Myleen Signorini}}$ at (847) $272\text{-}7250\times217$	
Cutomer Service Rep	OVERGUALHEED APPLY URBEIT Hore your stills by working in a trianuly environment. The front desk is looking for part time start to work flowled shifts for early weekledy morning, day and evening shifts. Benefits include YMCA member high and bahysitting during your shift. Presse context <u>Sarah Tucker or Chern (Savard 1697) 227-225</u> , 213.	
Lifeguards and Swim Instructors	Love to swim? Love kids? Put the two together and make a difference. The North Suburban YMCA is looking for qualified and experienced swim instructors and	I
Done		My Computer

Scoped Learning Extractor: Precision = 58%, Recall = 75% <u>A Error = -22%</u>

Broader View



Now touch on some other issues

(3) Automatically Inducing an Ontology

[Riloff, '95]

Two inputs:



(2)

Heuristic "interesting" meta-patterns.

Linguistic Pattern	Example
1. <subject> active-verb</subject>	<pre><pre>perpetrator> bombed</pre></pre>
2. <subject> active-verb direct-object³</subject>	<pre><pre>claimed responsibility</pre></pre>
3. <subject> passive-verb</subject>	<victim> was <u>murdered</u></victim>
4. <subject> verb infinitive</subject>	<pre><pre>perpetrator> attempted to <u>kill</u></pre></pre>
5. <subject> auxiliary noun</subject>	<victim> was <u>victim</u></victim>
6. active-verb <direct-object></direct-object>	bombed <target></target>
7. passive-verb <direct-object>⁴</direct-object>	killed <victim></victim>
8. infinitive <direct-object></direct-object>	to <u>kill</u> <victim></victim>
9. verb infinitive <direct-object></direct-object>	threatened to <u>attack</u> <target></target>
10. gerund <direct-object></direct-object>	killing <victim></victim>
11. noun auxiliary <direct-object></direct-object>	fatality was <victim></victim>
12. noun preposition <noun-phrase></noun-phrase>	<u>bomb</u> against <target></target>
13. active-verb preposition <noun-phrase></noun-phrase>	killed with <instrument></instrument>
14. passive-verb preposition <noun-phrase></noun-phrase>	was <u>aimed</u> at <target></target>
15. infinitive preposition <noun-phrase>³</noun-phrase>	to fire at <victim></victim>

(3) Automatically Inducing an Ontology



Broader View

Now touch on some other issues



1

(4) Training IE Models using Unlabeled Data

[Collins & Singer, 1999]

...says Mr. Cooper, a vice president of ...

NNP NNP appositive phrase, head=president

Use two independent sets of features:

Contents: full-string=*Mr._Cooper*, contains(*Mr.*), contains(*Cooper*) Context: context-type=*appositive*, appositive-head=*president*

on

on

on

1. Start with just seven rules: and ~1M sentences of NYTimes

full-string=New_York	\ Location
fill-string=California	\ Location
full-string=U.S.	\ Location
contains(Mr.)	\ Person
contains(Incorporated)	\ Organizatio
full-string=Microsoft	\ Organizatio
full-string=I.B.M.	\ Organizatio

- 2. Alternately train & label using each feature set.
- 3. Obtain 83% accuracy at finding person, location, organization & other in appositives and prepositional phrases!

See also [Brin 1998], [Riloff & Jones 1999]

Broader View

Now touch on some other issues



(5) Data Mining: Working with IE Data

- Some special properties of IE data:
 - It is based on extracted text
 - It is "dirty", (missing extraneous facts, improperly normalized entity names, etc.
 - May need cleaning before use
- What operations can be done on dirty, unnormalized databases?
 - Query it directly with a language that has "soft joins" across similar, but not identical keys. [Cohen 1998]
 - Construct features for learners [Cohen 2000]
 - Infer a "best" underlying clean database [Cohen, Kautz, MacAllester, KDD2000]

(5) Data Mining: Mutually supportive IE and Data Mining [Nahm & Mooney, 2000]

Extract a large database

Learn rules to predict the value of each field from the other fields. Use these rules to increase the accuracy of IE.

Example DB record

title: Senior DBMS Consultant

application: SQL Server, Oracle

required years of experience: 3

desired years of experience: 5

language: Powerbuilder, Progress, C, C++, Visual Basic

area: Electronic Commerce, Customer Service

Filled Job Template

salary: Up to \$55K

platform: UNIX, NT

required degree: BS

state: TX

city: Dallas

country: US

Sample Learned Rules

platform:AIX & !application:Sybase & application:DB2 \application:Lotus Notes

language:C++ & language:C & application:Corba & title=SoftwareEngineer \ platform:Windows

language:HTML & platform:WindowsNT & application:ActiveServerPages \ area:Database

Language:Java & area:ActiveX & area:Graphics \ area:Web