Information Extraction from the World Wide Web

CSE 454

Based on Slides by

William W. Cohen
Carnegie Mellon University
Andrew McCallum
University of Massachusetts Amherst

From KDD 2003

Administrivia

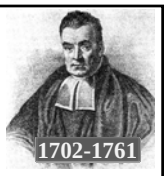
- Homework
  - Due today
- Projects
  - Proposals due today
  - Group meetings (30min) next week.
  - Please email me with times that work

Open Slots – week of 4/21

	Tues	Wed	Thurs	Fri
9:00				XXXX
9:30				XXXX
10:00				
10:30		XXXXX		
11:00		XXXXX		XXXX
11:30		XXXXX		XXXX
12:00				XXXX
12:30		XXXXX		XXXX
1:00		XXXXX		XXXX
1:30		XXXXX	XXXXX	XXXX
	XXXXX	XXXXX	XXXXX	XXXX
4:30		XXXXX	XXXXX	XXXX
5:00		XXXXX	XXXXX	XXXX

Quick Review

Bayes Theorem



$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

Bayesian Categorization

- Let set of categories be $\{c_1, c_2, \ldots c_n\}$
- Let E be description of an instance.
- Determine category of E by determining for each c_i

$$P(c_i \mid E) = \frac{P(c_i)P(E \mid c_i)}{P(E)}$$

- P(E) can be determined since categories are complete and disjoint.

$$\sum_{i=1}^{n} P(c_i \mid E) = \sum_{i=1}^{n} \frac{P(c_i)P(E \mid c_i)}{P(E)} = 1$$

$$P(E) = \sum_{i=1}^{n} P(c_i)P(E \mid c_i)$$

Naïve Bayesian Motivation

- Problem: Too many possible instances (exponential in *m*) to estimate all P(*E* | c_i)
- If we assume features of an instance are independent given the category (c_i) (*conditionally independent*).

$$P(E \mid c_i) = P(e_1 \wedge e_2 \wedge \cdots \wedge e_m \mid c_i) = \prod_{j=1}^{m} P(e_j \mid c_i)$$

- Therefore, we then only need to know P(e_j | c_i) for each feature and category.

Information Extraction

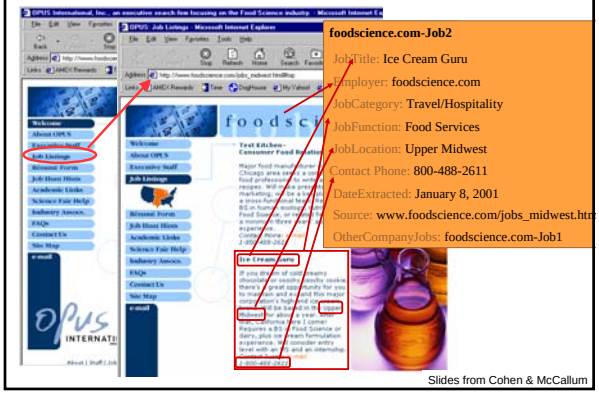
Example: The Problem

Slides from Cohen & McCallum

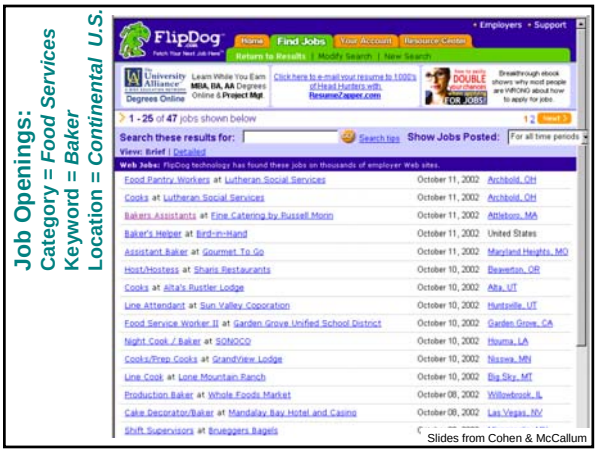
Example: A Solution

Slides from Cohen & McCallum

Extracting Job Openings from the Web



Slides from Cohen & McCallum



Slides from Cohen & McCallum

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 open-source 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…

NAME	TITLE	ORGANIZATION

Slides from Cohen & McCallum

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 open-source 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

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Slides from Cohen & McCallum

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 open-source 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

Slides from Cohen & McCallum

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 open-source 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

Slides from Cohen & McCallum

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 open-source 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

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

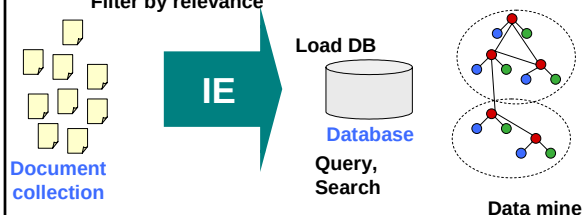
Slides from Cohen & McCallum

IE in Context

Create ontology

Spider

Filter by relevance



Train extraction models

Label training data

Slides from Cohen & McCallum

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
  - First by hand, then ML: [Doorenbos '96], [Soderland '96], [Kushmerick '97],…

Slides from Cohen & McCallum

What makes IE from the Web Different?

Less grammar, but more formatting & linking

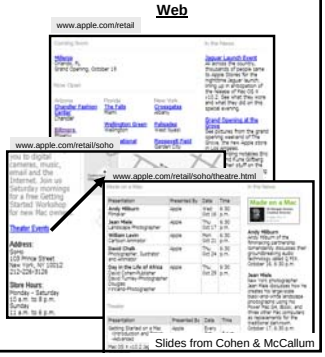
Newswire

Apple to Open Its First Retail Store in New York City

MACWORLD EXPO, NEW YORK--July 17, 2002--Apple's first retail store in New York City will open in Manhattan's SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple's largest retail store to date and is a stunning example of Apple's commitment to offering customers the world's best computer shopping experience.

"Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week," said Steve Jobs, Apple's CEO. "We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles."

Web



The directory structure, link structure, formatting & layout of the Web is its own new grammar.

Slides from Cohen & McCallum

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

Text paragraphs without formatting

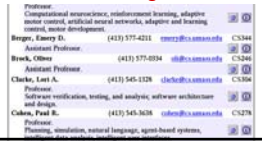
Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in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

Grammatical sentences and some formatting & links

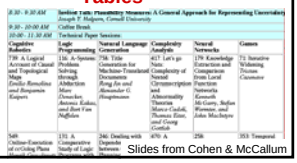
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

Non-grammatical snippets, rich formatting & links



Tables



Slides from Cohen & McCallum

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

Web site specific

Formatting

Amazon Book Pages

Genre specific

Layout

Resumes

Wide, non-specific

Language

University Names

Slides from Cohen & McCallum

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

E.g. word patterns:

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

Slides from Cohen & McCallum

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 *Title:* CEO	*Relation:* Succession *Company:* General Elec *Title:* CEO *Out:* Jack Welsh *In:* Jeffrey Imme
Person: Jeffrey Immelt		
Location: Connecticut	*Relation:* Company-I *Company:* General Electric *Location:* Connecticut	

"Named entity" extraction

Slides from Cohen & McCallum

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.

$$\text{Precision} = \frac{\#\text{ correctly predicted segments}}{\#\text{ predicted segments}} = \frac{2}{6}$$

$$\text{Recall} = \frac{\#\text{ correctly predicted segments}}{\#\text{ true segments}} = \frac{2}{4}$$

$$F1 = \text{Harmonic mean of Prec. + Recall} = \frac{1}{((1/P) + (1/R))/2}$$

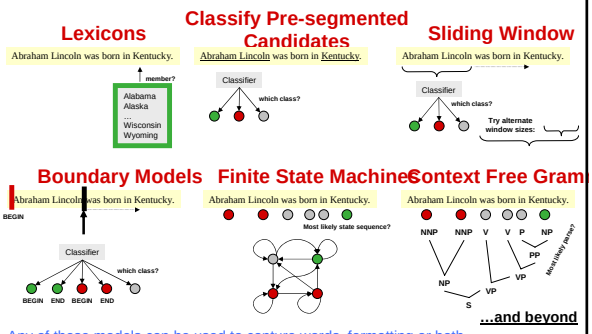
Slides from Cohen & McCallum

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

Slides from Cohen & McCallum

Landscape of IE Techniques (1/1): Models



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

Slides from Cohen & McCallum

Landscape: Focus of this Tutorial

Pattern complexity	closed set	regular	**complex**	**ambiguous**
Pattern feature domain	**words**	**words + formatting**	**formatting**	
Pattern scope	**site-specific**	genre-specific	**general**	
Pattern combinations	**entity**	binary	n-ary	
Models	lexicon	regex	**window**	**boundary**	**FSM**	**CFG**

Slides from Cohen & McC

Sliding Windows

Slides from Cohen & McCallum

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

Slides from Cohen & McCallum

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

Slides from Cohen & McCallum

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

Slides from Cohen & McCallum

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

Slides from Cohen & McCallum

A “Naïve Bayes” Sliding Window Model

[Freitag 1997]

… **00 : pm 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 | contents | suffix

Estimate Pr(LOCATION|window) using Bayes rule

Try all “reasonable” windows (vary length, position)

Assume independence for length, prefix words, suffix words, content words

Estimate from data quantities like: Pr(“Place” in prefix|LOCATION)

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)

Slides from Cohen & McCallum

“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

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.
```

Field	F1
Person Name:	30%
Location:	61%
Start Time:	98%

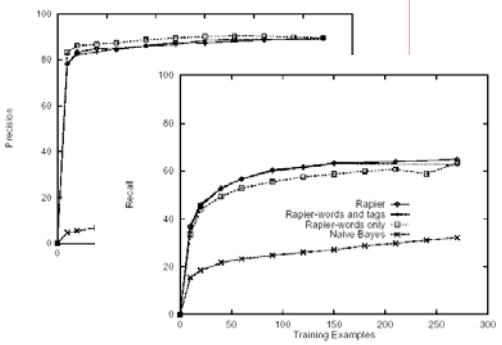
Slides from Cohen & McCallum

Realistic sliding-window-classifier IE

- 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.
- Learning Method
  - SRV: Top-Down Rule Learning *[Frietag AAAI '98]*
  - Rapier: Bottom-Up *[Califf & Mooney, AAAI '99]*

Slides from Cohen & McCallum

Rapier: results – precision/recall



Slides from Cohen & McCallum

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

Slides from Cohen & McCallum

Rule-learning approaches to sliding-window 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?

Slides from Cohen & McCallum

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

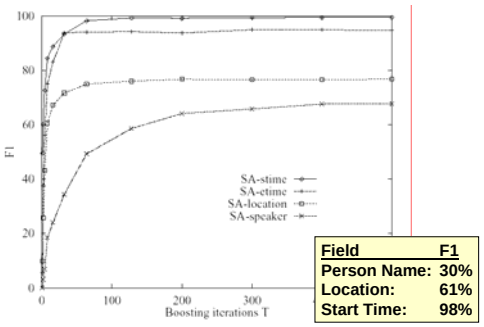
Slides from Cohen & McCallum

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, anyNumber, …
- Weak learner for "patterns" uses greedy search (+ lookahead) to repeatedly extend a pair of empty *BEFORE,AFTER* patterns

Slides from Cohen & McCallum

BWI: Learning to detect boundaries



Field	F1
Person Name:	30%
Location:	61%
Start Time:	98%

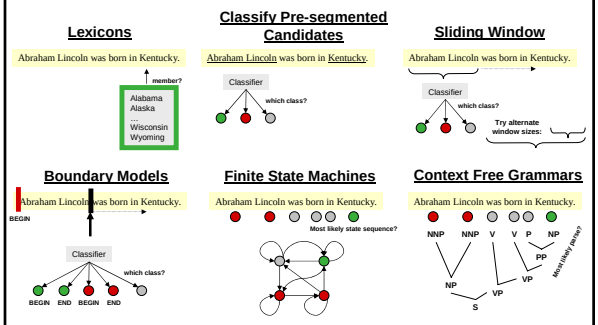
Slides from Cohen & McCallum

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.

Slides from Cohen & McCallum

Landscape of IE Techniques (1/1): Models



Each model can capture words, formatting, or both

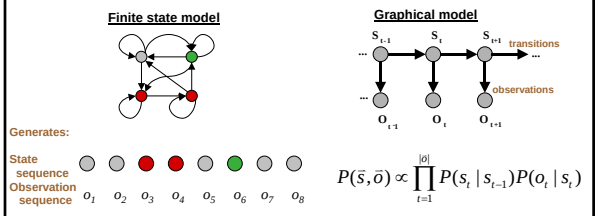
Slides from Cohen & McCallum

Finite State Machines

Slides from Cohen & McCallum

Hidden Markov Models (HMMs)

standard sequence model in genomics, speech, NLP, …



$$P(\vec{s},\vec{o}) \propto \prod_{t=1}^{|o|} P(s_t \mid s_{t-1}) P(o_t \mid s_t)$$

Parameters: for all states $S=\{s_1,s_2,\ldots\}$

Start state probabilities: $P(s_t)$

Transition probabilities: $P(s_t|s_{t-1})$

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

Training:

Maximize probability of training observations

Slides from Cohen & McCallum

Example: The Dishonest Casino

A casino has two dice:

- Fair die
  P(1) = P(2) = P(3) = P(5) = P(6) = 1/6
- Loaded die
  P(1) = P(2) = P(3) = P(5) = 1/10
  P(6) = 1/2

Casino player switches back-&-forth between fair and loaded die once every 20 turns

Game:

1. You bet $1
2. You roll (always with a fair die)
3. Casino player rolls (maybe with fair die, maybe with loaded die)
4. Highest number wins $2

Slides from Serafim Batzoglou

Question # 1 – Evaluation

GIVEN

A sequence of rolls by the casino player

1245526462146146136136661664661636616366163…

QUESTION

How likely is this sequence, given our model of how the casino works?

This is the **EVALUATION** problem in HMMs

Slides from Serafim Batzoglou

Question # 2 – Decoding

GIVEN

A sequence of rolls by the casino player

1245526462146146136136661664661636616366163…

QUESTION

What portion of the sequence was generated with the fair die, and what portion with the loaded die?

This is the **DECODING** question in HMMs

Slides from Serafim Batzoglou

Question # 3 – Learning

GIVEN

A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616516…

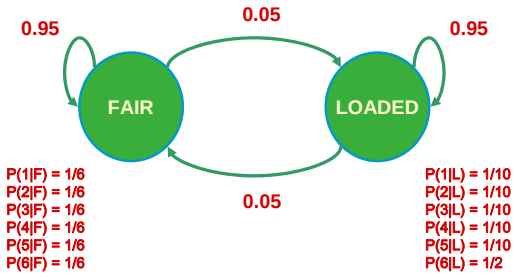
QUESTION

How "loaded" is the loaded die? How "fair" is the fair die? How often does the casino player change from fair to loaded, and back?

This is the **LEARNING** question in HMMs

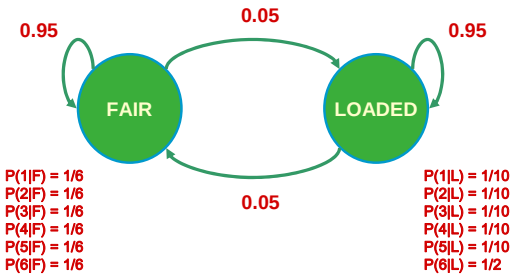
Slides from Serafim Batzoglou

The dishonest casino model

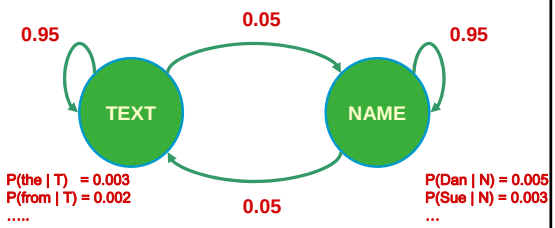


Slides from Serafim Batzoglou

What's this have to do with Info Extraction?



What's this have to do with Info Extraction?



IE Resources

- Data
  - RISE, `http://www.isi.edu/~muslea/RISE/index.html`
  - Linguistic Data Consortium (LDC)
    - Penn Treebank, Named Entities, Relations, etc.
  - `http://www.biostat.wisc.edu/~craven/ie`
  - `http://www.cs.umass.edu/~mccallum/data`
- Code
  - TextPro, `http://www.ai.sri.com/~appelt/TextPro`
  - MALLET, `http://www.cs.umass.edu/~mccallum/mallet`
  - SecondString, `http://secondstring.sourceforge.net/`
- Both
  - `http://www.cis.upenn.edu/~adwait/penntools.html`

Slides from Cohen & McCallum

References

- [Bikel et al 1997] Bikel, D.; Miller, S.; Schwartz, R.; and Weischedel, R. Nymble: a high-performance learning name-finder. In *Proceedings of ANLP'97*, p194-201.
- [Califf & Mooney 1999], Califf, M.E.; Mooney, R.: Relational Learning of Pattern-Match Rules for Information Extraction, in *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (AAAI-99).
- [Cohen, Hurst, Jensen, 2002] Cohen, W.; Hurst, M.; Jensen, L.: A flexible learning system for wrapping tables and lists in HTML documents. *Proceedings of The Eleventh International World Wide Web Conference* (WWW-2002)
- [Cohen, Kautz, McAllester 2000] Cohen, W; Kautz, H.; McAllester, D.: Hardening soft information sources. *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining* (KDD-2000).
- [Cohen, 1998] Cohen, W.: Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity, in *Proceedings of ACM SIGMOD-98*.
- [Cohen, 2000a] Cohen, W.: Data Integration using Similarity Joins and a Word-based Information Representation Language, *ACM Transactions on Information Systems*, 18(3).
- [Cohen, 2000b] Cohen, W. Automatically Extracting Features for Concept Learning from the Web, *Machine Learning: Proceedings of the Seventeenth International Conference (ML-2000).*
- [Collins & Singer 1999] Collins, M.; and Singer, Y. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.
- [De Jong 1982] De Jong, G. An Overview of the FRUMP System. In: Lehnert, W. & Ringle, M. H. (eds), *Strategies for Natural Language Processing*. Larence Erlbaum, 1982, 149-176.
- [Freitag 98] Freitag, D: Information extraction from HTML: application of a general machine learning approach, *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (AAAI-98).
- [Freitag, 1999], Freitag, D. *Machine Learning for Information Extraction in Informal Domains.* Ph.D. dissertation, Carnegie Mellon University.
- [Freitag 2000], Freitag, D: Machine Learning for Information Extraction in Informal Domains, *Machine Learning* 39(2/3): 99-101 (2000).
- Freitag & Kushmerick, 1999] Freitag, D; Kushmerick, D.: Boosted Wrapper Induction. *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (AAAI-99)
- [Freitag & McCallum 1999] Freitag, D. and McCallum, A. Information extraction using HMMs and shrinakge. In Proceedings AAAI-99 Workshop on Machine Learning for Information Extraction. AAAI Technical Report WS-99-11.
- [Kushmerick, 2000] Kushmerick, N: Wrapper Induction: efficiency and expressiveness, *Artificial Intelligence*, 118(pp 15-68).
- [Lafferty, McCallum & Pereira 2001] Lafferty, J.; McCallum, A.; and Pereira, F., Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *In Proceedings of ICML-2001.*
- [Leek 1997] Leek, T. R. *Information extraction using hidden Markov models.* Master's thesis. UC San Diego.
- [McCallum, Freitag & Pereira 2000] McCallum, A.; Freitag, D.; and Pereira. F., Maximum entropy Markov models for information extraction and segmentation, In *Proceedings of ICML-2000*
- [Miller et al 2000] Miller, S.; Fox, H.; Ramshaw, L.; Weischedel, R. A Novel Use of Statistical Parsing to Extract Information from Text. *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*. p. 226 - 233.

Slides from Cohen & McCallum