Information Extraction: Sequence Models, Information Extraction Tasks and Information Integration

CS 224N
2009
Sequence Inference

**Sequence Data**

**Local Level**
- Local Data
- Feature Extraction
- Label
- Features
- Maximum Entropy Models
- CG / L-BFGS
- Quadratic Penalties
- NLP Issues

**Sequence Level**
- Sequence Model
- Inference
- Sequence Inference
- NLP Issues
- Features
- Label
- Optimization
- Smoothing
- Classifier Type

**Sequence Model**
- Inference
- Features
- Label
MEMM inference in systems

- For a Conditional Markov Model (CMM) a.k.a. a Maximum Entropy Markov Model (MEMM), the classifier makes a single decision at a time, conditioned on evidence from observations and previous decisions.

- A larger space of sequences is explored via search

(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)

<table>
<thead>
<tr>
<th>Local Context</th>
<th>Decision Point</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT NNP VBD</td>
<td>0</td>
<td>W_0 22.6</td>
</tr>
<tr>
<td>The Dow fell</td>
<td>+1</td>
<td>W_{+1} %</td>
</tr>
<tr>
<td></td>
<td>-1</td>
<td>W_{-1} fell</td>
</tr>
<tr>
<td></td>
<td>-2</td>
<td>T_{-1} VBD</td>
</tr>
<tr>
<td></td>
<td>-3</td>
<td>T_{-1-T_{-2}} NNP-VBD</td>
</tr>
<tr>
<td></td>
<td>???</td>
<td>hasDigit? true</td>
</tr>
<tr>
<td></td>
<td>???</td>
<td>...</td>
</tr>
</tbody>
</table>
Two ways to Search: Beam Inference

- Beam inference:
  - At each position keep the top $k$ complete sequences.
  - Extend each sequence in each local way.
  - The extensions compete for the $k$ slots at the next position.

- Advantages:
  - Fast; and beam sizes of 3–5 are as good or almost as good as exact inference in many cases.
  - Easy to implement (no dynamic programming required).

- Disadvantage:
  - Inexact: the globally best sequence can fall off the beam.
Two ways to Search: Viterbi Inference

- **Viterbi inference:**
  - Dynamic programming or memoization.
  - Requires small window of state influence (e.g., past two states are relevant).

- **Advantage:**
  - Exact: the global best sequence is returned.

- **Disadvantage:**
  - Harder to implement long-distance state-state interactions (but beam inference tends not to successfully capture long-distance resurrection of sequences anyway).
Viterbi Inference: J&M Ch. 6

• I’m basically punting on this … read Ch. 6.
  – I’ll do dynamic programming for parsing

• It’s a small change from HMM Viterbi
  – From:

  \[ v_t(j) = \max_{i=1}^{N} v_{t-1}(i) P(s_j|s_i) P(o_t|s_j) \quad 1 \leq j \leq N, 1 < t \leq T \]

  – To:

  \[ v_t(j) = \max_{i=1}^{N} v_{t-1}(i) P(s_j|s_i, o_t) \quad 1 \leq j \leq N, 1 < t \leq T \]
Viterbi Inference: J&M Ch. 6

\[ v_1(2) = P(H|\text{start}, 3) \]

\[ v_1(1) = P(C|\text{start}, 3) \]

\[ v_2(2) = \max( P(H|H, 1) \cdot P(H|\text{start}, 3), \]
\[ P(H|C, 1) \cdot P(C|\text{start}, 3) ) \]

\[ v_2(1) = \max( P(C|H, 1) \cdot P(H|\text{start}, 3), \]
\[ P(C|C, 1) \cdot P(C|\text{start}, 3) ) \]

1. Start
2. H
3. C
4. End

\[ o_1 \quad o_2 \quad o_3 \]

\[ t \]
Part-of-speech tagging: HMM Tagging Models of Brants 2000

- Highly competitive with other state-of-the art models
- Trigram HMM with smoothed transition probabilities
- Capitalization feature becomes part of the state – each tag state is split into two e.g.
  \( NN \rightarrow \langle NN, \text{cap} \rangle, \langle NN, \text{not cap} \rangle \)
- Suffix features for unknown words

\[
P(w | \text{tag}) = P(\text{suffix} | \text{tag})(w | \text{suffix}) \\
\approx \hat{P}(\text{suffix})\tilde{P}(\text{tag} | \text{suffix}) / \hat{P}(\text{tag})
\]

\[
\tilde{P}(\text{tag} | \text{suffix}_n) = \lambda_1 \hat{P}(\text{tag} | \text{suffix}_n) + \lambda_2 \hat{P}(\text{tag} | \text{suffix}_{n-1}) + \ldots + \lambda_n \hat{P}(\text{tag})
\]
MEMM Tagging Models -II

• Ratnaparkhi (1996): local distributions are estimated using maximum entropy models
  – Previous two tags, current word, previous two words, next two words, suffix, prefix, hyphenation, and capitalization features for unknown words
• Toutanova et al. (2003)
  – Richer features, bidirectional inference, better smoothing, better unknown word handling

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall Accuracy</th>
<th>Unknown Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEMM (Ratn. 1996)</td>
<td>96.63</td>
<td>85.56</td>
</tr>
<tr>
<td>HMM (Brants 2000)</td>
<td>96.7</td>
<td>85.5</td>
</tr>
<tr>
<td>MEMM (T. et al 2003)</td>
<td>97.24</td>
<td>89.04</td>
</tr>
</tbody>
</table>
Smoothing: POS Tagging

• From (Toutanova et al., 2003):

<table>
<thead>
<tr>
<th></th>
<th>Overall Accuracy</th>
<th>Unknown Word Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Smoothing</td>
<td>96.54</td>
<td>85.20</td>
</tr>
<tr>
<td>With Smoothing</td>
<td>97.10</td>
<td>88.20</td>
</tr>
</tbody>
</table>

• Smoothing helps:
  – Softens distributions.
  – Pushes weight onto more explanatory features.
  – Allows many features to be dumped safely into the mix.
  – Speeds up convergence (if both are allowed to converge)!
Summary of POS Tagging

For POS tagging, the change from generative to discriminative model does not by itself result in great improvement.

One profits from discriminative models for specifying dependence on overlapping features of the observation such as spelling, suffix analysis, etc.

This additional power (of the CMM, CRF, Perceptron models) has been shown to result in improvements in accuracy.

A CMM allows integration of rich features of the observations, but can suffer from assuming independence from following observations; this effect can be relieved by moving to a CRF, but also by adding dependence on following words.

The higher accuracy of discriminative models comes at the price of much slower training.
CRFs [Lafferty, Pereira, and McCallum 2001]

- Another sequence model: Conditional Random Fields (CRFs)
- A whole-sequence conditional model rather than a chaining of local models.
  \[ P(c \mid d, \lambda) = \frac{\exp \sum \lambda_i f_i (c, d)}{\sum \exp \sum \lambda_i f_i (c', d)} \]
  - The space of \( c \)'s is now the space of sequences
    - But if the features \( f_i \) remain local, the conditional sequence likelihood can still be calculated exactly using dynamic programming
- Training is slower, but CRFs avoid causal-competition biases
- These (or a variant using a max margin criterion) are seen as the state-of-the-art these days, and fairly standardly used
NER Results: 
CoNLL (2003) Named Entity Recognition task

Task: Predict semantic label of each word in text

Foreign NNP I-NP ORG
Ministry NNP I-NP ORG
spokesman NN I-NP O
Shen NNP I-NP PER
Guofang NNP I-NP PER
told VBD I-VP O
Reuters NNP I-NP ORG

Standard evaluation is per entity, not per token
NER Results: Discriminative Model

- Increases from better features, a better classification model.
Sequence models? CoNLL 2003 NER shared task
Results on English Devset
CoNLL NER Results: CMM Order

![Bar chart showing NER results for CMM Order.](chart.png)
Sequence Tagging Without Sequence Information: POS tagging

Using 3 words works significantly better than using only the current word and the previous two or three tags instead! (Toutanova et al. 2003)
CoNLL NER: A real difference

- A difference of about 0.7% gives significance among good CoNLL results
- Here we get one!
- It was done with some Perl regular expressions
Biomedical NER Motivation

• The biomedical world has a huge body of information, which is growing rapidly.
  
  – MEDLINE, the primary research database serving the biomedical community, currently contains over 12 million abstracts, with 60,000 new abstracts appearing each month.

  – There is also an impressive number of biological databases containing information on genes, proteins, nucleotide and amino acid sequences, including GenBank, Swiss-Prot, and Fly-Base; each contains entries numbering from the thousands to the millions and are multiplying rapidly.
Motivation

• Currently, all of these resources are curated by hand by expert annotators at enormous expense.

• The information overload from the massive growth in the scientific literature has shown the necessity to automatically locate, organize and manage facts relating to experimental results.

• Natural Language Processing can aid researchers and curators of biomedical databases by automating these tasks.
Named Entity Recognition

- General NER vs. Biomedical NER

<PER> Christopher Manning </PER> is a professor at <ORG> Stanford University </ORG>, in <LOC> Palo Alto </LOC>.

<RNA> TAR </RNA> independent transactivation by <PROTEIN> Tat </PROTEIN> in cells derived from the <CELL> CNS </CELL> - a novel mechanism of <DNA> HIV-1 gene </DNA> regulation.
Why is this difficult?

• The list of biomedical entities is growing.
  – New genes and proteins are constantly being discovered, so explicitly enumerating and searching against a list of known entities is not scalable.
  – Part of the difficulty lies in identifying previously unseen entities based on contextual, orthographic, and other clues.

• Biomedical entities don’t have strict naming conventions.
  – Common English words such as *period*, *curved*, and *for* are used for gene names.
  – Entity names can be ambiguous. For example, in FlyBase, “clk” is the gene symbol for the “Clock” gene but it also is used as a synonym of the “period” gene.

• Biomedical entity names are ambiguous
  – Experts only agree on whether a word is even a gene or protein 69% of the time. (Krauthammer *et al.*, 2000)
  – Often systematic polysemies between gene, RNA, DNA, etc.
Interesting Features

- Word, and surrounding context
- Word Shapes
  - Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

<table>
<thead>
<tr>
<th>Varicella-zoster</th>
<th>Xx-xxx</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA</td>
<td>xXXX</td>
</tr>
<tr>
<td>CPA1</td>
<td>XXXd</td>
</tr>
</tbody>
</table>

- Character substrings
Features: What’s in a Name?

- oxa
  - drug: 18
  - company: 0
  - movie: 0
  - place: 0
  - person: 0

- : 708

- field
  - drug: 14
  - company: 0
  - movie: 8
  - place: 6
  - person: 0

Cotrimoxazole  Wethersfield

Alien Fury: Countdown to Invasion
Interesting Features

- Part-of-Speech tags
- Parsing information
- Searching the web for the word in a given context
  - \textit{X gene}, \textit{X mutation}, \textit{X antagonist}
- Gazetteer
  - list words whose classification is known
- Abbreviation extraction (Schwartz and Hearst, 2003)
  - Identify short and long forms when occurring together in text

\[ \text{... Zn finger homeodomain 2 (Zfh 2) ...} \]
Finkel et al. (2004) Results

• BioCreative task – Identify genes and proteins

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>81.3%</td>
<td>86.1%</td>
<td>83.6%</td>
</tr>
</tbody>
</table>

precision = \( \frac{tp}{tp + fp} \)

recall = \( \frac{tp}{tp + fn} \)

F1 = \( \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \)
Finkel et al. (2004) Results

- BioNLP task – Identify genes, proteins, DNA, RNA, and cell types

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>68.6%</td>
<td>71.6%</td>
<td>70.1%</td>
</tr>
</tbody>
</table>

$\text{precision} = \frac{tp}{tp + fp}$

$\text{recall} = \frac{tp}{tp + fn}$

$\text{F1} = \frac{2(\text{precision})(\text{recall})}{\text{precision} + \text{recall}}$
Quiz question!

• Answer in one sentence:

In discriminative sequence labeling tasks like NER, why do sequence models (that condition on other labels) often offer little value over using straight classifiers (which don’t condition on other labels)?
Information Extraction and Integration

Following slides from:

William Cohen
Andrew McCallum
Eugene Agichtein
Sunita Sarawagi
The Value of Text Data

- “Unstructured” text data is the primary source of human-generated information
  - Citeseer, comparison shopping, PIM systems, web search, data warehousing, scientific literature

- Managing and utilizing text: information extraction and integration

- Scalability: a bottleneck for deployment

- Relevance to data mining community
Example: A Solution

Job Seekers: Find your dream job!

- Check our Best Places to Find a Job January report.
- Open your FREE account and put your resume online.
- Search 24/7 with our FREE automatic JobHunters™.
- Research databases of over 50,000 employers.
- Get expert advice from our new Resource Center.
- Access salary surveys/calculators, relocation tools, networking opportunities, and training/testing tools.
- Use FlipDog.com to search jobs right from your desktop! Download Snippets today!

647,514
Job Opportunities from 63,641 Employers

Find a Job!

Employers
- Post Your Resume

Pigskin Places
- Health Care in NY: 2,776
- Health Care in MD: 1,056
- Sales in NY: 9,351
- Sales in MD: 875
- Computing in NY: 4,050
- Computing in MD: 4,414

Jobs for Sports Fans
- Head Football Coach
- Football Coach
- Asst Football Coach
- High School Football Coach
- Univ. Asst Football Coach

Showcase Jobs
MRI Management Recruiters of Charlotte North
We provide total staffing solutions in the areas of Human Resources, Compensation, Web-based HR self-service, and Customer Management Systems. Learn More

Looking for a Vice President of Academic Affairs to oversee planning, operation and evaluation of the college's academic programs. Learn More

Top 100 Web Sites™
PC Magazine, Nov. 2000

Top 10 Career Web Site
PC Magazine, Nov. 2000

Top 10 Job Site
WhizBang!
Job Openings:
Category = Food Services
Keyword = Baker
Location = Continental U.S.

<table>
<thead>
<tr>
<th>Web Jobs:</th>
<th>FlipDog technology has found these jobs on thousands of employer Web sites.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Food Pantry Workers</strong> at Lutheran Social Services</td>
<td>October 11, 2002</td>
</tr>
<tr>
<td><strong>Cooks</strong> at Lutheran Social Services</td>
<td>October 11, 2002</td>
</tr>
<tr>
<td><strong>Bakers Assistants</strong> at Fine Catering by Russell Morin</td>
<td>October 11, 2002</td>
</tr>
<tr>
<td><strong>Baker's Helper</strong> at Bird-in-Hand</td>
<td>October 11, 2002</td>
</tr>
<tr>
<td><strong>Assistant Baker</strong> at Gourmet To Go</td>
<td>October 11, 2002</td>
</tr>
<tr>
<td><strong>Host/Hostess</strong> at Sharis Restaurants</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td><strong>Cooks</strong> at Alta's Rastler Lodge</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td><strong>Line Attendant</strong> at Sun Valley Corporation</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td><strong>Food Service Worker II</strong> at Garden Grove Unified School District</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td><strong>Night Cook / Baker</strong> at SONOCO</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td><strong>Cooks/Prep Cooks</strong> at GrandView Lodge</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td><strong>Line Cook</strong> at Lone Mountain Ranch</td>
<td>October 10, 2002</td>
</tr>
<tr>
<td><strong>Production Baker</strong> at Whole Foods Market</td>
<td>October 08, 2002</td>
</tr>
<tr>
<td><strong>Cake Decorator/Baker</strong> at Mandalay Bay Hotel and Casino</td>
<td>October 08, 2002</td>
</tr>
<tr>
<td><strong>Shift Supervisors</strong> at Bruegger's Bagels</td>
<td>October 08, 2002</td>
</tr>
</tbody>
</table>
What is “Information Extraction”

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

<table>
<thead>
<tr>
<th>NAME</th>
<th>TITLE</th>
<th>ORGANIZATION</th>
</tr>
</thead>
</table>

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...
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...

<table>
<thead>
<tr>
<th>NAME</th>
<th>TITLE</th>
<th>ORGANIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates</td>
<td>CEO</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Bill Veghte</td>
<td>VP</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Richard Stallman</td>
<td>founder</td>
<td>Free Software Foundation</td>
</tr>
</tbody>
</table>
What is “Information Extraction”

As a family of techniques:

Information Extraction = segmentation + classification + clustering + association

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...
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
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

These two steps aka “named entity recognition”
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
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…
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."

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

Barto, Andrew G.  
(413) 545-2109  
burton@cs.umass.edu  
CS276

Professor.
Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.

Berger, Emery D.  
(413) 577-4211  
emery@cs.umass.edu  
CS344

Assistant Professor.

Brock, Oliver  
(413) 577-0384  
oli@cs.umass.edu  
CS246

Assistant Professor.

Clarke, Lori A.  
(413) 545-1328  
clarke@cs.umass.edu  
CS304

Professor.
Software verification, testing, and analysis; software architecture and design.

Cohen, Paul R.  
(413) 545-3538  
cohen@cs.umass.edu  
CS278

Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

Tables

<table>
<thead>
<tr>
<th>Cognitive Robotics</th>
<th>Logic Programming</th>
<th>Natural Language Generation</th>
<th>Complexity Analysis</th>
<th>Neural Networks</th>
<th>Games</th>
</tr>
</thead>
<tbody>
<tr>
<td>739: A Logical Account of Causal and Topological Maps</td>
<td>Emilio Remolina and Benjamin Ruppers</td>
<td>116: A System for Solving through Abduction and Deception</td>
<td>Marc Dworkin, Antonia Kakas, and Bert van Beek</td>
<td>758: Title Generation for Machine-Translated Documents</td>
<td>Rong Jin and Alexander G. Hauptmann</td>
</tr>
<tr>
<td>8:17: Let's go Nats! Complexity of Nested Circumscription and Abnormality Theory</td>
<td>Marco Ciardelli, Thomas Eiter, and Georg Gottlob</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>179: Knowledge Extraction and Comparison from Local Function Networks</td>
<td>Kenneth McGarry, Stefan Wermter, and John Moehly</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71: Iterative Widening Tricon</td>
<td>Cazemave</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Online-Execution of eColGol</th>
<th>Henrik Grosskreutz and Gerhard Lakemeyer</th>
</tr>
</thead>
<tbody>
<tr>
<td>349:</td>
<td>A Perspective on Knowledge Compilation</td>
</tr>
<tr>
<td>353:</td>
<td>Temporal Difference Learning Applied to a High Performance Game-Playing</td>
</tr>
</tbody>
</table>

| 258: Virtual-Execution for Constrained Formulations in Neural-Network Time-Series |
|-----------------------------|------------------------------|
| 131: | A Comparative Study of Logic Programs with Preference |
| 470: | Dealing with Dependencies between Content Planning and Surface Realization in a Pipeline Generation |
| 466: | A Perspective on Knowledge Compilation |
| 457: | Adaining Knowledge about the Environment |
| 456: | Adaining Knowledge about the Environment |
| 455: | Adaining Knowledge about the Environment |
Landscape of IE Tasks (2/4):

Intended Breadth of Coverage

Web site specific
- Formatting
  - Amazon.com Book Pages

Genre specific
- Layout
  - Resumes

Wide, non-specific
- Language
  - University Names
Landscape of IE Tasks (3/4): Complexity

E.g. word patterns:

<table>
<thead>
<tr>
<th>Closed set</th>
<th>Regular set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U.S. states</strong></td>
<td><strong>U.S. phone numbers</strong></td>
</tr>
<tr>
<td>He was born in Alabama…</td>
<td>Phone: (413) 545-1323</td>
</tr>
<tr>
<td>The big Wyoming sky…</td>
<td>The CALD main office can be reached at 412-268-1299</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Complex pattern</th>
<th>Ambiguous patterns, needing context and many sources of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U.S. postal addresses</strong></td>
<td>Person names</td>
</tr>
<tr>
<td>University of Arkansas</td>
<td>…was among the six houses sold by Hope Feldman that year.</td>
</tr>
<tr>
<td>P.O. Box 140</td>
<td></td>
</tr>
<tr>
<td>Hope, AR 71802</td>
<td>Pawel Opalinski, Software Engineer at WhizBang Labs.</td>
</tr>
<tr>
<td>Headquarters:</td>
<td></td>
</tr>
<tr>
<td>1128 Main Street, 4th Floor</td>
<td></td>
</tr>
<tr>
<td>Cincinnati, Ohio 45210</td>
<td></td>
</tr>
</tbody>
</table>
Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.
Broader View

Up to now we have been focused on ML methods for segmentation and classification.
# These are subordinate patterns
wordOrdinals="(?:first|second|third|fourth|fifth|sixth|seventh|eighth|ninth|tenth|eleventh|twelfth|thirteenth|fourteenth|fifteenth)";
my numberOrdinals="(?:\d?(?:1st|2nd|3rd|1th|2th|3th|4th|5th|6th|7th|8th|9th|0th))";
my $ordinals="(?:wordOrdinals|numberOrdinals)";
my $confTypes="(?:Conference|Workshop|Symposium)";
my $words="(?:[A-Z]\w+\s)";  # A word starting with a capital letter and ending with 0 or more spaces
my $confDescriptors="(?:international\s+|[A-Z]+\s)+";  # e.g. "International Conference ...' or the conference name for workshops (e.g. "VLDB Workshop ...")
my $connectors="(?:on|of)";
my $abbreviations="(?:\([A-Z]\w+\w+\W\s+\(\d\d\)+]?\))";  # Conference abbreviations like "(SIGMOD'06)"
# The actual pattern we search for.  A typical conference name this pattern will find is
# "3rd International Conference on Blah Blah Blah (ICBBB-05)"
my $fullNamePattern="((?:$ordinals\s+$words\s*|$confDescriptors)?$confTypes(?:\s+$connectors\s+.*?|\s+)?)$abbreviations?)?(?:\n|\r|\.|<)";

# Given a <dbworldMessage>, look for the conference pattern
lookForPattern($dbworldMessage, $fullNamePattern);

# In a given <file>, look for occurrences of <pattern>
# <pattern> is a regular expression
sub lookForPattern {
    my ($file,$pattern) = @_;
Machine Learning Methods

- Sequence models: HMMs, CMMs/MEMMs, CRFs
- Can work well when training data is easy to construct and is plentiful
- Can capture complex patterns that are hard to encode with hand-crafted rules
  - e.g., determine whether a review is positive or negative
  - extract long complex gene names
- Can be labor intensive to construct training data
  - Question: how much training data is sufficient?

The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300."
Broader View

Now touch on some other issues

3 Create ontology

Spider

Filter by relevance

IE

Tokenize

Segment

Classify

Associate

Cluster

Load DB

Database

1

2

4 Train extraction models

Document collection

Label training data

5 Data mine

Query, Search
Relation Extraction: Disease Outbreaks

- Extract structured relations from text

- May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly Ebola epidemic in Zaire, is finding itself hard pressed to cope with the crisis...

<table>
<thead>
<tr>
<th>Date</th>
<th>Disease Name</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 1995</td>
<td>Malaria</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>July 1995</td>
<td>Mad Cow Disease</td>
<td>U.K.</td>
</tr>
<tr>
<td>Feb. 1995</td>
<td>Pneumonia</td>
<td>U.S.</td>
</tr>
</tbody>
</table>

Information Extraction System (e.g., NYU’s Proteus)
Example: Protein Interactions

“We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex.”

CBF-A $\xleftrightarrow{\text{interact}}$ CBF-C

CBF-B associates CBF-A-CBF-C complex
Relation Extraction

• Typically requires Entity Tagging as preprocessing

• Knowledge Engineering
  – Rules defined over lexical items
    • “<company> located in <location>”
  – Rules defined over parsed text
    • “((Obj <company>) (Verb located) (*) (Subj <location>))”
  – Proteus, GATE, …

• Machine Learning-based
  – Supervised: Learn rules/patterns from examples
  – Partially-supervised: bootstrap from “seed” examples
    Agichtein & Gravano 2000, Etzioni et al., 2004, …
Example Extraction Rule [NYU Proteus]

For <company> appoints <person> <position>

(defpattern appoint
  "np-sem(C-company)? rn? sa? vg(C-appoint) np-sem(C-person) \',\'? to-be? np(C-position) to-succeed?:
  company-at=1.attributes, sa=3.span, lv=4.span, person-at=5.attribute:
  position-at=8.attributes |
...

(defun when-appoint (phrase-type)
  (let ((person-at (binding `person-at))
      (company-entity (entity-bound `company-at))
      (person-entity (essential-entity-bound `person-at `C-person))
      (position-entity (entity-bound `position-at))
      (predecessor-entity (entity-bound `predecessor-at))
      new-event)
    (not-an-antecedent position-entity)
    ;; if no company is specified for position, use agent
...

Example of Learned Extraction Patterns: Snowball [AG2000]
(1) Association as Binary Classification

Christos Faloutsos conferred with Ted Senator, the KDD 2003 General Chair.

Person-Role (Christos Faloutsos, KDD 2003 General Chair) → NO

Person-Role (Ted Senator, KDD 2003 General Chair) → YES

Do this with SVMs and tree kernels over parse trees.

[Zelenko et al, 2002]
(2) Association with Graphical Models

Capture arbitrary-distance dependencies among predictions.

Local language models contribute evidence to entity classification.

Random variable over the class of entity #1, e.g. over \{person, location, \ldots\}

Random variable over the class of relation between entity #2 and #1, e.g. over \{lives-in, is-boss-of, \ldots\}

Local language models contribute evidence to relation classification.

Dependencies between classes of entities and relations!

Inference with loopy belief propagation.

[Roth & Yih 2002]
Accuracy of Information Extraction

<table>
<thead>
<tr>
<th>Information Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>90-98%</td>
</tr>
<tr>
<td>Attributes</td>
<td>80%</td>
</tr>
<tr>
<td>Facts</td>
<td>60-70%</td>
</tr>
<tr>
<td>Events</td>
<td>50-60%</td>
</tr>
</tbody>
</table>

- Errors cascade (error in entity tag $\rightarrow$ error in relation extraction)
- This estimate is optimistic:
  - Holds for well-established tasks
  - Many specific/novel IE tasks exhibit much lower accuracy

[Feldman, ICML 2006 tutorial]
Broader View

Now touch on some other issues

When do two extracted strings refer to the same object?
Document 1: The Justice Department has officially ended its inquiry into the assassinations of John F. Kennedy and Martin Luther King Jr., finding "no persuasive evidence" to support conspiracy theories, according to department documents. The House Assassinations Committee concluded in 1978 that Kennedy was "probably" assassinated as the result of a conspiracy involving a second gunman, a finding that broke from the Warren Commission's belief that Lee Harvey Oswald acted alone in Dallas on Nov. 22, 1963.

Document 2: In 1953, Massachusetts Sen. John F. Kennedy married Jacqueline Lee Bouvier in Newport, R.I. In 1960, Democratic presidential candidate John F. Kennedy confronted the issue of his Roman Catholic faith by telling a Protestant group in Houston, "I do not speak for my church on public matters, and the church does not speak for me."


[From Li, Morie, & Roth, AI Magazine, 2005]
Important Problem

- Appears in numerous real-world contexts
- Plagues many applications
  - Citeseer, DBLife, AliBaba, Rexa, etc.
(2) Information Integration

[2001a, 2001b, 2003a]

Goal might be to **merge** results of two IE systems:

<table>
<thead>
<tr>
<th>Name:</th>
<th>Introduction to Computer Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number:</td>
<td>CS 101</td>
</tr>
<tr>
<td>Teacher:</td>
<td>M. A. Kludge</td>
</tr>
<tr>
<td>Time:</td>
<td>9-11am</td>
</tr>
<tr>
<td>Name:</td>
<td>Data Structures in Java</td>
</tr>
<tr>
<td>Room:</td>
<td>5032 Wean Hall</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Title:</th>
<th>Intro. to Comp. Sci.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num:</td>
<td>101</td>
</tr>
<tr>
<td>Dept:</td>
<td>Computer Science</td>
</tr>
<tr>
<td>Teacher:</td>
<td>Dr. Klütge</td>
</tr>
<tr>
<td>TA:</td>
<td>John Smith</td>
</tr>
<tr>
<td>Topic:</td>
<td>Java Programming</td>
</tr>
<tr>
<td>Start time:</td>
<td>9:10 AM</td>
</tr>
</tbody>
</table>