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

Decision Point | Features
---|---
-3 -2 -1 0 +1 | W₀ 22.6
| W₁ | %
| W₋₁ | fell
| T₋₁ | VBD
| T₋₁₋T₋₂ | NNP-VBD
| hasDigit? | true
| ... | ...

Local Context

<table>
<thead>
<tr>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>NNP</td>
<td>VBD</td>
<td>???</td>
<td>???</td>
</tr>
<tr>
<td>The</td>
<td>Dow</td>
<td>fell</td>
<td>22.6</td>
<td>%</td>
</tr>
</tbody>
</table>

(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)
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.
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 allow long-distance resurrection of sequences anyway).
HMM Part-of-speech Tagging Models -
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.
  \( \text{NN} \rightarrow <\text{NN},\text{cap}>,<\text{NN},\text{not cap}> \)
• Suffix features for unknown words

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

\[
\tilde{P}(\text{tag} | \text{suffix}_n) = \lambda_1\hat{P}(\text{tag} | \text{suffix}_n) + \lambda_2\hat{P}(\text{tag} | \text{suffix}_{n-1}) + ... + \lambda_n\hat{P}(\text{tag})
\]
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>HMM (Brants 2000)</td>
<td>96.7</td>
<td>85.5</td>
</tr>
<tr>
<td>MEMM (Ratn. 1996)</td>
<td>96.63</td>
<td>85.56</td>
</tr>
<tr>
<td>MEMM (T. et al 2003)</td>
<td>97.24</td>
<td>89.04</td>
</tr>
</tbody>
</table>
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.

\[
P(c \mid d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}
\]

• The space of \(c\)'s is now the space of sequences
  – But if the features \(f_i\) remain local, the conditional sequence likelihood can be calculated exactly using dynamic programming

• Training is slow, 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
Summary of Tagging

For 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. A CMM allows integration of rich features of the observations, but can suffer strongly from assuming independence from following observations; this effect can be relieved by adding dependence on following words. This additional power (of the CMM, CRF, Perceptron models) has been shown to result in improvements in accuracy. The higher accuracy of discriminative models comes at the price of much slower training.
Biomedical NER Motivation

• The biomedical field contains a large 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 polysemyies between gene, RNA, DNA, etc.
Maximum Entropy Markov Model

$$P(t \mid h) = \frac{\exp\left(\sum_{j=1}^{m} f_j(h,t)\lambda_j\right)}{\sum_{k=1}^{K} \exp\left(\sum_{j=1}^{m} f_j(h,t_k)\lambda_j\right)}$$
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>Word</th>
<th>Simplified Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varicella-zoster</td>
<td>Xx-xxx</td>
</tr>
<tr>
<td>mRNA</td>
<td>XXXX</td>
</tr>
<tr>
<td>CPA1</td>
<td>XXXd</td>
</tr>
</tbody>
</table>

- Character substrings

#Dpp# → #Dpp#, #Dpp, Dpp#, #Dp, Dpp, pp#, #D, Dp, pp, p#, D, p
Features: What’s in a Name?

- **oxa**
- **field**

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
  - $X$ gene, $X$ mutation, $X$ antagonist
- Gazetteer
  - list words whose classification is known
- Abbreviation extraction (Schwartz and Hearst, 2003)
  - Identify short and long forms when occurring together in text

... 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>

\[
\text{precision} = \frac{tp}{tp + fp}
\]

\[
\text{recall} = \frac{tp}{tp + fn}
\]

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

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

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

precision = tp / (tp + fp)
recall = tp / (tp + fn)
F1 = 2(precision)(recall) / (precision + recall)
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

• Managing and utilizing text: information extraction and integration

• Scalability: a bottleneck for deployment

• Relevance to data mining community
Example: A Solution
Job Title: Ice Cream Guru
Employer: foodscience.com
Job Category: Travel/Hospitality
Job Function: Food Services
Job Location: Upper Midwest
Contact Phone: 800-488-2611
Date Extracted: January 8, 2001
Source: www.foodscience.com/jobs_midwest.html
Other Company Jobs: foodscience.com-Job1
<table>
<thead>
<tr>
<th>Job Title</th>
<th>Company/Location</th>
<th>Date</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food Pantry Workers</td>
<td>Lutheran Social Services</td>
<td>October 11, 2002</td>
<td>Archbold, OH</td>
</tr>
<tr>
<td>Cooks</td>
<td>Lutheran Social Services</td>
<td>October 11, 2002</td>
<td>Archbold, OH</td>
</tr>
<tr>
<td>Bakers Assistants</td>
<td>Fine Catering by Russell Morin</td>
<td>October 11, 2002</td>
<td>Attleboro, MA</td>
</tr>
<tr>
<td>Baker's Helper</td>
<td>Bird-in-Hand</td>
<td>October 11, 2002</td>
<td>United States</td>
</tr>
<tr>
<td>Assistant Baker</td>
<td>Gourmet To Go</td>
<td>October 11, 2002</td>
<td>Maryland Heights, MO</td>
</tr>
<tr>
<td>Host/Hostess</td>
<td>Sharis Restaurants</td>
<td>October 10, 2002</td>
<td>Deaverton, OR</td>
</tr>
<tr>
<td>Cooks</td>
<td>Alta's Rustler Lodge</td>
<td>October 10, 2002</td>
<td>Alta, UT</td>
</tr>
<tr>
<td>Line Attendant</td>
<td>Sun Valley Corporation</td>
<td>October 10, 2002</td>
<td>Huntsville, UT</td>
</tr>
<tr>
<td>Food Service Worker II</td>
<td>Garden Grove Unified School District</td>
<td>October 10, 2002</td>
<td>Garden Grove, CA</td>
</tr>
<tr>
<td>Night Cook / Baker</td>
<td>SONOCO</td>
<td>October 10, 2002</td>
<td>Houma, LA</td>
</tr>
<tr>
<td>Cooks/Prep Cooks</td>
<td>GrandView Lodge</td>
<td>October 10, 2002</td>
<td>Nisswa, MN</td>
</tr>
<tr>
<td>Line Cook</td>
<td>Lone Mountain Ranch</td>
<td>October 10, 2002</td>
<td>Big Sky, MT</td>
</tr>
<tr>
<td>Production Baker</td>
<td>Whole Foods Market</td>
<td>October 08, 2002</td>
<td>Willowbrook, IL</td>
</tr>
<tr>
<td>Cake Decorator/Baker</td>
<td>Mandalay Bay Hotel and Casino</td>
<td>October 08, 2002</td>
<td>Las Vegas, NV</td>
</tr>
<tr>
<td>Shift Supervisors</td>
<td>Bruegger's Bagels</td>
<td>October 08, 2002</td>
<td>Minneapolis, MN</td>
</tr>
</tbody>
</table>
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...
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 + 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…
What is “Information Extraction”

As a family of techniques:

<table>
<thead>
<tr>
<th>Information Extraction =</th>
</tr>
</thead>
<tbody>
<tr>
<td>segmentation + classification + association + clustering</td>
</tr>
</tbody>
</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…

<table>
<thead>
<tr>
<th>Microsoft Corporation CEO Bill Gates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Gates</td>
</tr>
<tr>
<td>Microsoft Bill Veghte Microsoft VP</td>
</tr>
<tr>
<td>Richard Stallman Free Software Foundation</td>
</tr>
</tbody>
</table>
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.
Landscape of IE Tasks (1/4): Degree of Formatting

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  barto@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) 577-0334  oli@cs.umass.edu  CS246
Assistant Professor.

Clarke, Loni A.  (413) 545-1328  clarke@cs.umass.edu  CS304
Professor. Software verification, testing, and analysis; software architecture and design.

Cohen, Paul R.  (413) 545-3638  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>549: Online-Execution of ecGolog Plans Henrik Grosskreutz and Gerhard Lakemeyer</td>
<td>131: A Comparative Study of Logic Programs with Preference Torsten Schaub and Kevin</td>
<td>246: Dealing with Dependencies between Content Planning and Surface Realization in a Pipeline Generation</td>
<td>470: A Perspective on Knowledge Compilation Adrian Delaigle and Pierre Marquis</td>
<td>258: Violation-Guided Learning for Constrained Formulations in Neural-Network Time-Series</td>
<td>353: Temporal Difference Learning Applied to a High Performance Game-Playing</td>
</tr>
</tbody>
</table>
Landscape of IE Tasks (2/4):
Intended Breadth of Coverage

Web site specific
- Formatting
  - Amazon.com Book Pages

Genre specific
- Layout
  - Resumes
- Language
  - University Names

Wide, non-specific
- Amazon.com Book Pages
- Resumes
- 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><strong>Person names</strong></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>Pawel Opalinski, Software Engineer at WhizBang Labs.</td>
</tr>
<tr>
<td>Hope, AR 71802</td>
<td></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.

<table>
<thead>
<tr>
<th>Single entity</th>
<th>Binary relationship</th>
<th>N-ary record</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Person:</strong> Jack Welch</td>
<td><strong>Relation:</strong> Person-Title</td>
<td><strong>Relation:</strong> Succession</td>
</tr>
</tbody>
</table>
| **Person:** Jeffrey Immelt | **Person:** Jack Welch  
**Title:** CEO | **Company:** General Electric  
**Title:** CEO  
**Out:** Jack Welch  
**In:** Jeffrey Immelt |
| **Location:** Connecticut | **Relation:** Company-Location | **Company:** General Electric  
**Location:** Connecticut |

“Named entity” extraction
Broader View

Up to now we have been focused on 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="(?:\([A-Z]\w\w+\s{0,}\d{0,}\))"; # 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+$words*|$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

*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.*

- Can be labor intensive to construct training data
  - Question: how much training data is sufficient?
Broader View

Now touch on some other issues

3 Create ontology

Spider

Filter by relevance

IE

Segment
Classify
Associate
Cluster

1

Tokenize

2

Document collection

4 Train extraction models

Label training data

Load DB

Database

Query, Search

5 Data mine
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**."

\[
\begin{align*}
\text{CBF-A} & \overset{\text{interact}}{\longleftrightarrow} \text{complex} \overset{\text{interact}}{\rightarrow} \text{CBF-C} \\
\text{CBF-B} & \overset{\text{associates}}{\rightarrow} \text{CBF-A-CBF-C complex}
\end{align*}
\]
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
  – Learn rules/patterns from examples
    Dan Roth 2005, Cardie 2006, Mooney 2005, …
  – Partially-supervised: bootstrap from “seed” examples
    Agichtein & Gravano 2000, Etzioni et al., 2004, …

• Recently, hybrid models [Feldman2004, 2006]
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 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]
... This enzyme, UBC6, localizes to the endoplasmic reticulum, with the catalytic domain facing the cytosol. ...

Subcellular-localization (UBC6, endoplasmic reticulum)
(1) Association with Graphical Models

Capture arbitrary-distance dependencies among predictions.

Random variable over the class of entity #2, e.g. over \{person, location,\ldots\}

Local language models contribute evidence to entity classification.

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]
(1) Association with Graphical Models

Also capture long-distance dependencies among predictions.

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

Local language models contribute evidence to entity classification.

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 lower accuracy

[Feldman, ICML 2006 tutorial]
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.
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>Room:</td>
<td>5032 Wean Hall</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name:</th>
<th>Data Structures in Java</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title:</td>
<td>Intro. to Comp. Sci.</td>
</tr>
<tr>
<td>Num:</td>
<td>101</td>
</tr>
<tr>
<td>Dept:</td>
<td>Computer Science</td>
</tr>
<tr>
<td>Teacher:</td>
<td>Dr. Kludge</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>
