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.

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:
    \[ v_t(j) = \max_{i=1}^N v_{t-1}(i) P(s_j | s_i, o_t) \quad 1 \leq j \leq N, 1 < t \leq T \]
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.\[
\text{NN} \rightarrow \langle \text{NN, cap} \rangle, \langle \text{NN, not cap} \rangle
\]
- Suffix features for unknown words

\[
P(w | \text{tag}) = \frac{P(\text{suffix}(w) | \text{tag}) P(\text{tag})}{\sum_{\text{tag}} P(\text{suffix}(w) | \text{tag}) P(\text{tag})}
\]
\[
\tilde{P}(\text{tag} | \text{suffix},...) = \lambda_1 \tilde{P}(\text{tag} | \text{suffix}) + \lambda_2 \tilde{P}(\text{tag} | \text{suffix},...) + \ldots + \lambda_n \tilde{P}(\text{tag})
\]

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 Word Accuracy</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:
  - Without Smoothing: 96.54/85.20
  - With Smoothing: 97.10/88.20

- 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.

\[
P(c | d, \lambda) = \frac{\exp \sum_{c'} \lambda(f(c,d))}{\sum_{c'} \sum_{d'} \exp \sum_{c'} \lambda(f(c',d'))}
\]

- The space of \(c’s\) is now the space of sequences
- But if the features \(f\) 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

CoNLL 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

Sequence Tagging Without Sequence Information: POS tagging

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 FlyBase; 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

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.

- Character substrings

Features: What’s in a Name?

- Drug
- Company
- Movie
- Person
- Place

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

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>

\[F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\]

- 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>

\[\text{precision} = \frac{tp}{tp + fp}, \quad \text{recall} = \frac{tp}{tp + fn}, \quad F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\]

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)?

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

Information Extraction and Integration

Following slides from:
William Cohen
Andrew McCallum
Eugene Agichtein
Sunita Sarawagi
Example: A Solution

Extracting Job Openings from the Web

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

What is “Information Extraction”

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

What is “Information Extraction”

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

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

"Fourteen months after opening our first retail store, our 21 stores are attracting over 180,000 visitors each week," said Steve Jobs, Apple’s CEO. "We are thrilled to have introduced the Mac in over 100 countries and feel this is a great example of Apple’s commitment to offering customers the best computer experience available."

"Fourteen months after opening our first retail store, our 21 stores are attracting over 180,000 visitors each week," said Steve Jobs, Apple’s CEO. "We are thrilled to have introduced the Mac in over 100 countries and feel this is a great example of Apple’s commitment to offering customers the best computer experience available."

The directory structure, link structure, formatting & layout of the Web is its own new grammar.
Landscape of IE Tasks (3/4):

Complexity

Closed set
U.S. states
He was born in Alabama.
The big Wyoming sky...

Regular set
U.S. phone numbers
Phone: (413) 545-1323

Complex pattern
U.S. postal addresses
University of Arkansas
P.O. Box 140
Hope, AR 71802

Ambiguous patterns, needing context and many sources of evidence
Person names
was among the six houses sold by Hope Feldman that year.

Headquarters:
1228 Main Street, 4th Floor
Cincinnati, Ohio 45210

Broader View

Create ontology
Segment Classify Associate Cluster
Train extraction models
Load DB

Machine Learning Methods

- Sequence models: HMMs, 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?

Landscape of IE Tasks (4/4): Single Field/Record

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
Person: Jack Welch
Location: Connecticut

Binary relationship
Relation: Person-Title
Person: Jack Welch Title: CEO
Relation: Company-Location
Company: General Electric Location: Connecticut

N-ary record
Relation: Succession
Company: General Electric Title: CEO Out: Jack Welch In: Jeffrey Immelt

“Named entity” extraction

Steps 1 & 2: Hand Coded Rule Example: Conference Name

- These are subordinate patterns
  - Word: Word, Ordinals: Ordinals
  - Conf Types: Conf Types
  - Conf Descriptors: Conf Descriptors
  - Connectors: Connectors
  - Abbreviations: Abbreviations

- The actual pattern we search for. A typical conference name this pattern will find is “3rd International Conference on Blah Blah Blah (ICBBB-05)”

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

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

Now touch on some other issues
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...

Information Extraction System (e.g., NYU’s Proteus)

<table>
<thead>
<tr>
<th>Date</th>
<th>Disease Name</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 95</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. 1996</td>
<td>Pneumonia</td>
<td>U.S.</td>
</tr>
<tr>
<td>May 1995</td>
<td>Ebola</td>
<td>Zaire</td>
</tr>
</tbody>
</table>

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."

Example Extraction Rule [NYU Proteus]

```plaintext
:: : For (company) appoints (person) (position)

(define-pattern appoint
  "pr-om(C=company)! re? ma? "gC=appoint) pr=om(C=person) "," "ha? ap(C=position) to (re? ma?)
  company-at(4.attribute, em=3, ep, 1r=4, ep, person-at(5.attribute, position-at(4.attribute) )

(define-pattern (person-to-type)
  (let ((person=entity (headg-person))
    (company-entity (tailg-company) )
    (person-entitl (entitl-person) )
    (position-entitl (entitl-position) )
    (associate-person entity-position)
    (new-event)
    (not-re-antecedent position-entitl)
  )
  if as company is specified for position, use agent)
```

Example of Learned Extraction Patterns: Snowball [AG2000]

1) Association as Binary Classification

| 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.

[Etzioni et al., 2004]
(2) Association with Graphical Models

Capture arbitrary-distance dependencies among predictions.

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

Local language models contribute evidence to entity classification.

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

Dependencies between classes of entities and relations!

Inference with loopy belief propagation.

---

Accuracy of Information Extraction

<table>
<thead>
<tr>
<th>Information Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>90-99%</td>
</tr>
<tr>
<td>Attributes</td>
<td>80%</td>
</tr>
<tr>
<td>Facts</td>
<td>60-75%</td>
</tr>
<tr>
<td>Events</td>
<td>50-69%</td>
</tr>
</tbody>
</table>

[Feldman, ICML 2008 tutorial]

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

---

Broader View

Create ontology

Now touch on some other issues

- Spill
- Tokenize
- Segment
- Classify
- Assign
- Extract
- IE
- Query
- Search
- Data mine

When do two extracted strings refer to the same object?

---

Extracted Entities: Resolving Duplicates

- When do two extracted strings refer to the same object?

---

Important Problem

- Appears in numerous real-world contexts
- Plagues many applications
  - Citeseer, DBLife, AliBaba, Rexa, etc.

---

(2) Information Integration

[Minton, Knoblock, et al 2001], [Duh, Domingos, Halevy 2001], [Richardson & Domingos 2003]

Goal might be to merge results of two IE systems:

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
<th>Teacher</th>
<th>Time</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to Computer Science</td>
<td>CS 101</td>
<td>M. A. Kludge</td>
<td>9-11am</td>
<td>5032 Wean Hall</td>
</tr>
<tr>
<td>Data Structures in Java</td>
<td></td>
<td>Dr. Kludge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intro to Comp. Sci.</td>
<td></td>
<td>John Smith</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic: Java Programming</td>
<td></td>
<td>Start time: 9-10 AM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>