**Sequence Models**

for

Information Extraction, POS tagging,

Word Segmentation, Chunking, ...

CS224N

2007

(Some slides are mine; many slides are borrowed from Andrew McCallum and William Cohen’s IE Tutorial)

---

**Statistical sequence models for Information Extraction**

- There are several techniques for information extraction (template/wrapper learning, hand-coded rules) ...
- But statistical sequence models (Hidden Markov Models, MaxEnt markov models, CRFs) are good methods for sequence-based information extraction

- **Pros:**
  - Well-understood underlying statistical model
  - Can do some form of optimal inference along sequence
  - Portable, broad coverage, robust, good recall

- **Cons:**
  - Not necessarily as good for complex or multi-slot patterns
  - Only doing the entity mention labeling task (in general)

---

**Applying HMMs to IE**

(Leek 1997, Freitag and McCallum 2000)

- Multinomial HMMs are sequential version of naïve Bayes/LM.
- Document \(\Rightarrow\) generated by a stochastic process
- Observation \(\Rightarrow\) word
- State \(\Rightarrow\) "reason/explanation" for a given token
  - ‘Background’ state emits tokens like ‘the’, ‘said’, …
  - ‘Money’ state emits tokens like ‘million’, ‘euro’, …
  - ‘Person’ state emits tokens like ‘Tony’, ‘Prithi’, …
- Extraction: via the Viterbi (max likelihood parse) algorithm

---

**IE with Hidden Markov Models**

Given a sequence of observations:

Yesterday Pedro Domingos spoke this example sentence.

and a trained HMM:

Find the most likely state sequence: (Viterbi)

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

Person name: Pedro Domingos

---

**HMM for research papers: transitions A**

[Seymore et al., 99]

**HMM for research papers: emissions B**

[Seymore et al., 99]

Trained on 2 million words of BibTeX data from the Web
Freitag and McCallum (2000)
IE with HMMs details

- Partly fixed structure, partly hidden
  (constrained EM using remote supervision)
  - Class HMM (also used in comp. bio.)
- Parameter tying and shrinkage smoothing
  techniques
  - Better just to use a good unknown model?
- Structure learning of transition structure
  - Why not just plain EM?
- Results great on semi-structured data!
  - 92.9% token accuracy on paper/citations data
- Still rather modest on free form text

HMM IE results ($F_1$) on Freitag and McCallum
Acquisitions data

Other Sequence Modeling Tasks:
Chinese Word Segmentation
(also: Japanese, Thai, Ancient Greek, …)

- Basic units in written text are "characters".
- A sentence is a sequence of "characters", without explicit boundaries.
  - Meaningful units in written texts are "words"
  - Word meaning can differ greatly from characters
    - "and" vs. "monk"
  - But definition of "words" is debatable
    - Different segmentation standards defined by linguists
    - It's like whether you segment compounds (cf. German)

Other sequence modeling tasks

- Base noun phrase chunking
  - Small noun phrases are a useful unit for many applications of terminology extraction, web search
  - Mitsubishi has just announced a new 21.3-inch flat panel monitor for the Japanese market, and even though it offers two DVI ports and a UXGA resolution of 1,600 x 1,200, we're not sure how many folks will be willing to part with close to 200,000 yen
  - Sequence model marks segment start, end

Other sequence modeling tasks

- Topic/FAQ segmentation (question, answer)
- Part-of-speech tagging
- Musical sequences
- DNA sequences
- …
Results:
Train on ~500k words of news wire text.

- Sentence start-of-task

Task: Named Entity Extraction

- sentence end-of-task

HMM 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.
  \( NN \rightarrow \text{<NN, cap>}, \text{<NN, not cap>} \)
- Suffix features for unknown words

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

\[
\hat{P}(\text{tag} | \text{suffix}_t) = \lambda \hat{P}(\text{tag} | \text{suffix}_{t-1}) + \cdots + \lambda_i \hat{P}(\text{tag})
\]

HMM Example: “Nymble”

- Task: Named Entity Extraction
- [Bikel, et al 1998], [BBN “IdentFinder”]

Transition probabilities

Observation probabilities

\( P(s_t | s_{t-1}, o_{t-1}) \)

\( P(o_t | s_t, o_{t-1}) \)

Back-off to:

\( P(o_t | s_{t+1}) \)

\( P(o_t) \)

Results:

- Mixed: English 93%
- Upper: English 91%
- Mixed: Spanish 90%

Other examples of shrinkage for HMMs in IE: Freitag and McCallum ’99

What is a symbol?

- Ideally we would like to use many, arbitrary, overlapping features of words. Useful, but this is hard with HMMs

What's in a Name?

- Identity of word
  - Ends in “-ski”
  - Capitalized
  - Part of a noun phrase
  - In a list of city names
  - In bold font
  - Indented
  - In hyperlink anchor

Lots of learning systems are not confounded by multiple, non-independent features: decision trees, maxent models, neural nets, SVMs, ...

What is a symbol?

- Since 1997, probabilistic sequence approaches (BBN, NYU, then everyone) achieves state-of-the-art performance

- The purpose:
  - ... a lot of information is really associations between named entities.
  - ... for question answering, answers are usually named entities.
  - ... the same techniques apply to other slot-filling classifications.

Named Entity Extraction

- The task: find and classify names in text, for example:

- The European Commission [PER] said on Thursday it disagreed with German [LOC] advice.
- Only France [LOC] and Britain [LOC] backed Fischler [PER] ’s proposal.
- “What we have to be extremely careful of is how other countries are going to take Germany ’s lead”, Welsh National Farmers Union [ORG] ( NFU [LOC] ) chairman John Lloyd Jones [PER] said on BBC [ORG] radio.

- Other: a lot of information is really associations between named entities.
- ... for question answering, answers are usually named entities.
- ... the same techniques apply to other slot-filling classifications.

What is a symbol?

- A word token (for known words, seen more than \( k \) times)

<table>
<thead>
<tr>
<th>Word Feature</th>
<th>Example Text</th>
<th>Intention</th>
</tr>
</thead>
<tbody>
<tr>
<td>twoDigitYear</td>
<td>99</td>
<td>Two-digit year</td>
</tr>
<tr>
<td>fourDigitYear</td>
<td>1999</td>
<td>Four-digit year</td>
</tr>
<tr>
<td>containsDigitMultiple</td>
<td>99999</td>
<td>Product code</td>
</tr>
<tr>
<td>containsDigitAndDash</td>
<td>1999-99</td>
<td>Date</td>
</tr>
<tr>
<td>containsDigitAndDash</td>
<td>1999-99</td>
<td>Date</td>
</tr>
<tr>
<td>containsDigitAndPeriod</td>
<td>23,000.00</td>
<td>Monetary amount</td>
</tr>
<tr>
<td>containsDigitAndPeriod</td>
<td>23,000.00</td>
<td>Monetary amount, percentage</td>
</tr>
<tr>
<td>contain</td>
<td>45%99</td>
<td>Other number</td>
</tr>
<tr>
<td>contains</td>
<td>BBN</td>
<td>Organization</td>
</tr>
<tr>
<td>contains</td>
<td>M</td>
<td>Person name initial</td>
</tr>
<tr>
<td>firstWord</td>
<td>first word of sentence</td>
<td>No useful capitalization information</td>
</tr>
<tr>
<td>initCap</td>
<td>Sally</td>
<td>Capitalized word</td>
</tr>
<tr>
<td>lowerCase</td>
<td>car</td>
<td>Uncapitalized word</td>
</tr>
<tr>
<td>other</td>
<td>-</td>
<td>Punctuation marks, all other words</td>
</tr>
</tbody>
</table>
What is a symbol?

Idea: replace generative model in HMM with a maxent model, where state depends on observations

\[ \Pr(s_t | \mathbf{x}) = \cdots \]

Inference in Systems

- 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:
  - 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).
POS tagging: Ratnaparkhi’s MXPOST

- Sequential learning problem: predict POS tags of words.
- Uses MaxEnt model described above.
- Rich feature set.
- To smooth, discard features occurring < 10 times.

CMM 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>CMM (Ratn. 1996)</td>
<td>96.63</td>
<td>85.56</td>
</tr>
<tr>
<td>CMM (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>Without Smoothing</th>
<th>With Smoothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>96.54</td>
</tr>
<tr>
<td>Unknown Word Acc</td>
<td>85.20</td>
</tr>
</tbody>
</table>

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

Summary of POS 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 suffers 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.

CoNLL (2003) Named Entity Recognition task

Task: Predict semantic label of each word in text

<table>
<thead>
<tr>
<th>Foreign</th>
<th>Ministry</th>
<th>spokesman</th>
<th>Shen</th>
<th>Guofang</th>
<th>told</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>NNP</td>
<td>NN</td>
<td>NNP</td>
<td>NNP</td>
<td>NNP</td>
<td>NNP</td>
</tr>
<tr>
<td>I-NP</td>
<td>I-NP</td>
<td>I-NP</td>
<td>I-NP</td>
<td>I-NP</td>
<td>I-NP</td>
<td>I-NP</td>
</tr>
<tr>
<td>ORG</td>
<td>ORG</td>
<td>O</td>
<td>PER</td>
<td>PER</td>
<td>O</td>
<td>ORG</td>
</tr>
</tbody>
</table>

NER Results: Discriminative Model

- Increases from better features, a better classification model.
Sequence Tagging Without Sequence Information: POS tagging

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Token</th>
<th>Unknown</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertical</td>
<td>56,805</td>
<td>93.69%</td>
<td>82.61%</td>
<td>26.74%</td>
</tr>
<tr>
<td>3Words</td>
<td>239,767</td>
<td>96.57%</td>
<td>86.78%</td>
<td>48.27%</td>
</tr>
</tbody>
</table>

Using 3 words only works significantly better than using 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