Sequence Models for Information Extraction, POS tagging, Word Segmentation, Chunking, ...

CS224N 2007
(Including some slides from Andrew McCallum and William Cohen)

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 ⇒ generated by a stochastic process
- Observation ⇒ word
- State ⇒ “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

HMM for research papers: emissions B

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.  
  已開發和尚在開發的資源
- Meaningful units in written texts are “words”  
  - Word meaning can differ greatly from characters  
  - “and” “still” “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

Sequence Model Chinese Word Segmenter
**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.
  \[ P(w|tage) = \sum \text{suffix} \cdot P(tage|\text{suffix}) \]
- Suffix features for unknown words

\[ P(tage|\text{suffix}) = \lambda_1 P(tage|\text{suffix}_{prev}) + \lambda_2 P(tage|\text{suffix}_{prevprev}) + \cdots \]

**Named Entity Extraction**

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

  ```
  The European Commission [EC] said on Thursday it disagreed with German [MISC] 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 (NFU) [ORG] chairman John Lloyd Jones [PER] said on BBC [ORG] radio.
  ```

- 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.

**What is a symbol?**

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

- Identity of word ends in "ski"
- is capitalized
- is part of a noun phrase
- is in a list of city names
- is under mode X in WordNet
- is in bold font
- is indented
- is 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?

identity of word
ends in “-ski”
is capitalized
is part of a noun phrase
is in a list of city names
is under node X in WordNet
is in bold font
is indented
is in hyperlink anchor

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

\[ \Pr(s_t \mid x_t) = \ldots \]

Inference in Systems

Sequence Level

Local Level

Classifier Type

Optimization

Smoothing

Sequence Data

Feature Extraction

Label

Maximum Entropy Models

Conjugate Gradient

Quadratic Penalties

NLP Issues

Beam Inference

Sequence Model

Inference

Best Sequence

• 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

Sequence Model

Inference

Best Sequence

• 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>

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.

Smoothing: POS Tagging

- From (Toutanova et al., 2003):

<table>
<thead>
<tr>
<th>Without Smoothing</th>
<th>Overall Accuracy</th>
<th>Unknown Word Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>96.54</td>
<td>85.20</td>
<td></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 (fairly) safely into the mix.
  - Speeds up convergence (if both are allowed to converge!)

CoNLL (2003) Named Entity Recognition task

Task: Predict semantic label of each word in text

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

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

<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)

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