Hidden Markov Models for Information Extraction

Speaker:
Bibek Paudel

Tutor:
Maximilian Dylla

Saarland University, Germany
May 26, 2011
Motivation

- Location -> Russisches Haus der Wissenschaft und Kultur
Motivation

- **Location** -> Russisches Haus der Wissenschaft und Kultur
- **Speaker** -> Prof. Barbara Liskov
Motivation

- **Location** -> Russisches Haus der Wissenschaft und Kultur
- **Speaker** -> Prof. Barbara Liskov
- Document usually contains much irrelevant text (sparse)
Motivation

- Location -> Russisches Haus der Wissenschaft und Kultur
- Speaker -> Prof. Barbara Liskov
- Document usually contains much irrelevant text (sparse)
- Find relations
Motivation

- **Location** -> *Russisches Haus der Wissenschaft und Kultur*
- **Speaker** -> *Prof. Barbara Liskov*
- Document usually contains much irrelevant text (sparse)
- Find relations
- Populate database slots with phrases from documents
Agenda

1 Motivation

2 Hidden Markov Models
   - Introduction
   - The Task

3 Example

4 Problems

5 Procedure

6 Results

7 Conclusion
Finite State Machines
Finite State Machines
A generative process
Yes, Albert Einstein was born in Ulm.
Introduction

- Finite State Machines
- A generative process
- Next state depends only on current state
Introduction

Finite State Machines

A generative process

Next state depends only on current state

Given some text, recover the states that generated the text

Parameter: for all states $S = \{s_1, s_2, \ldots\}$

Start state probabilities: $P(s_1)$

Transition probabilities: $P(s_t | s_{t-1})$ Usually a multinomial over atomic, fixed alphabet

Observation (emission) probabilities: $P(o_t | s_t)$

Training:

Maximize probability of training observations (w/ prior)
Introduction

- Finite State Machines
- A generative process
- Next state depends only on current state
- Given some text, recover the states that generated the text
Why HMM

- Reasons for using HMM
The Task

Given:
a sequence of observations: Yes, Albert Einstein was born in Ulm.
a trained HMM
The Task

Given:
- a sequence of observations: Yes, Albert Einstein was born in Ulm.
- a trained HMM

- Find the most likely state sequence
The Task

Given:
- a sequence of observations: Yes, Albert Einstein was born in Ulm.
- a trained HMM

- Find the most likely state sequence
- Any words generated by the 'red state' are 'names'
The Task

Given:
- a sequence of observations: Yes, Albert Einstein was born in Ulm.
- a trained HMM

- Find the most likely state sequence
- Any words generated by the 'red state' are 'names'
- Viterbi Algorithm
An Example

Prepare a matrix of conditional probabilities (pairwise)

Define Observation Sequence

\[ O = \{ S_3, S_3, S_1, S_1, S_1, S_3, S_2, S_3 \} \]

Find the probability

\[ P(O | \text{Model}) \]

First order truncation by Markov property gives:

\[ P(S_3) \prod P(q_t | q_{t-1}) \]
An Example

Prepare a matrix of conditional probabilities (pairwise)

$$\begin{bmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{bmatrix}$$
An Example

- Prepare a matrix of conditional probabilities (pairwise)

\[
\begin{bmatrix}
0.4 & 0.3 & 0.3 \\
0.2 & 0.6 & 0.2 \\
0.1 & 0.1 & 0.8
\end{bmatrix}
\]
An Example

- Prepare a matrix of conditional probabilities (pairwise)
- Define Observation Sequence $O = \{S_3, S_3, S_1, S_1, S_1, S_2, S_3\}$
An Example

- Prepare a matrix of conditional probabilities (pairwise)
- Define Observation Sequence $O = \{S_3, S_3, S_1, S_1, S_1, S_2, S_3\}$
- Find the probability $P(O \mid Model)$
An Example

- Prepare a matrix of conditional probabilities (pairwise)
- Define Observation Sequence $O = \{S_3, S_3, S_1, S_1, S_1, S_2, S_3\}$
- Find the probability $P(O \mid Model)$
- $P(O \mid Model) = P(S_3, S_3, S_1, S_1, S_1, S_2, S_3 \mid Model)$
An Example

- Prepare a matrix of conditional probabilities (pairwise)
- Define Observation Sequence $O = \{S_3, S_3, S_1, S_1, S_1, S_2, S_3\}$
- Find the probability $P(O \mid Model)$
  
  $P(O \mid Model) = P(S_3, S_3, S_1, S_1, S_1, S_2, S_3 \mid Model)$

- First order truncation by Markov property gives:
  
  $P(S_3) \prod P(q_t \mid q_{t-1})$
An Example

- Prepare a matrix of conditional probabilities (pairwise)
- Define Observation Sequence \( O = \{ S_3, S_3, S_1, S_1, S_1, S_2, S_3 \} \)
- Find the probability \( P(O | Model) \)
- \( P(O | Model) = P(S_3, S_3, S_1, S_1, S_1, S_2, S_3 | Model) \)
- First order truncation by Markov property gives:
  \[
P(S_3) \prod P(q_t | q_{t-1})
  \]
  \[
P(S_3)P(S_3 | S_3)P(S_1 | S_3)P(S_1 | S_1)P(S_1 | S_1)P(S_2 | S_1)P(S_3 | S_2)
  \]
Prepare a matrix of conditional probabilities (pairwise)

Define Observation Sequence

\[ O = \{ S_3, S_3, S_1, S_1, S_1, S_2, S_3 \} \]

Find the probability

\[ P(O | \text{Model}) = \prod P(q_t | q_{t-1}) \]

First order truncation by Markov property gives:

\[ P(S_3) \]
Agenda

1 Motivation

2 Hidden Markov Models
   - Introduction
   - The Task

3 Example

4 Problems

5 Procedure

6 Results

7 Conclusion
Problems

For an observation sequence $O = O_1...O_T$, the three canonical HMM problems are:

- Evaluation Problem
  1. Given a model, find probability of Observation Sequence
  2. Forward-Backward
For an observation sequence $O = O_1...O_T$, the three canonical HMM problems are:

- **Evaluation Problem**
- **Inference Problem**
  1. Given the Observation Sequence $O = O_1...O_T$ and the model $\lambda$, find the most likely state sequence
  2. Choose an optimal state sequence $Q = q_1q_2...q_T$
Problems

For an observation sequence $O = O_1 \ldots O_T$, the three canonical HMM problems are:

- Evaluation Problem
- Inference Problem
- Learning Problem

1. How to model parameters in order to maximize probability of Observation Sequence?
2. We have to produce the actual model, the matrix
Agenda

1 Motivation

2 Hidden Markov Models
   • Introduction
   • The Task

3 Example

4 Problems

5 Procedure

6 Results

7 Conclusion
Procedure: HowTo

Model $\lambda = (A, B, \pi)$ and observation sequence $O$

- Target, Prefix, Suffix and Background
Procedure: HowTo

Model $\lambda = (A, B, \pi)$ and observation sequence $O$
- Target, Prefix, Suffix and Background
- Lengthen, Split and Add
Procedure: HowTo

Model \( \lambda = (A, B, \pi) \) and observation sequence \( O \)

- Target, Prefix, Suffix and Background
- Lengthen, Split and Add
- Shrinkage
Model $\lambda = (A, B, \pi)$ and observation sequence $O$

- Target, Prefix, Suffix and Background
- Lengthen, Split and Add
- Shrinkage
- Use dynamic programming to find the most likely state sequence
Procedure: HowTo

Model $\lambda = (A, B, \pi)$ and observation sequence $O$

- Target, Prefix, Suffix and Background
- Lengthen, Split and Add
- Shrinkage
- Use dynamic programming to find the most likely state sequence
- The Viterbi Algorithm

Figure: Viterbi
Procedure: The Details

- One HMM per class of information
Procedure: The Details

- One HMM per class of information
- Train on Labelled data
Procedure: The Details

- One HMM per class of information
- Train on Labelled data
- Background and Target States
1 Motivation

2 Hidden Markov Models
   - Introduction
   - The Task

3 Example

4 Problems

5 Procedure

6 Results

7 Conclusion
## Results

<table>
<thead>
<tr>
<th></th>
<th>speaker</th>
<th>location</th>
<th>acquired</th>
<th>dlramt</th>
<th>title</th>
<th>company</th>
<th>conf</th>
<th>deadline</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grown HMM</td>
<td>76.9</td>
<td>87.5</td>
<td>41.3</td>
<td>54.4</td>
<td>58.3</td>
<td>65.4</td>
<td>27.2</td>
<td>46.5</td>
<td>57.2</td>
</tr>
<tr>
<td>vs. SRV</td>
<td>+19.8</td>
<td>+16.0</td>
<td>+1.1</td>
<td>-1.6</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>+8.8</td>
</tr>
<tr>
<td>vs. Rapier</td>
<td>+23.9</td>
<td>+14.8</td>
<td>+12.5</td>
<td>+15.1</td>
<td>-11.7</td>
<td>+24.9</td>
<td>—</td>
<td>—</td>
<td>+13.3</td>
</tr>
<tr>
<td>vs. Simple HMM</td>
<td>+24.3</td>
<td>+5.6</td>
<td>+14.3</td>
<td>+5.6</td>
<td>+5.7</td>
<td>+11.1</td>
<td>+15.7</td>
<td>+6.7</td>
<td>+11.1</td>
</tr>
<tr>
<td>vs. Complex HMM</td>
<td>-2.1</td>
<td>+6.7</td>
<td>+7.5</td>
<td>-0.3</td>
<td>-0.3</td>
<td>+19.1</td>
<td>+0.0</td>
<td>-6.8</td>
<td>+3.0</td>
</tr>
</tbody>
</table>

**Figure:** Results
Conclusion

- Automatic generation of HMMs for IE
- Very effective and popular
- Better alternatives exist today (eg: CRF)
1. Information Extraction with HMM Structures Learned by Stochastic Optimization. Dayne Freitag and Andrew McCallum. AAAI’00.