5th SIPTA school on imprecise probability
16-20 July 2012, Pescara (Italy)

Robustly correcting mistakes made by OCR software

Jasper De Bock
University of Ghent (Belgium)
jasper.debock@ugent.be
(imprecise) state sequence estimation

A sequence of hidden state variables

A sequence of observable output variables
(imprecise) state sequence estimation

A sequence of hidden state variables

\[ X = \begin{cases} 
\text{sunny} & \text{or} \\
\text{cloudy} & \text{or} \\
\text{rainy} & \end{cases} \]

A sequence of observable output variables

\[ O = \begin{cases} 
\text{swimming} & \text{or} \\
\text{walking} & \text{or} \\
\text{watching TV} & \end{cases} \]
(imprecise) state sequence estimation

A sequence of hidden state variables

A sequence of observable output variables
(imprecise) state sequence estimation

Viterbi algorithm

\[ Q_1 (X_1) \quad Q_2 (X_2 | X_1) \quad Q_2 (X_3 | X_2) \]

\[ S_1 (O_1 | X_1) \quad S_2 (O_2 | X_2) \quad S_3 (O_3 | X_3) \]
(imprecise) state sequence estimation

\[ Q_1 (X_1) \quad Q_2 (X_2 | X_1) \quad Q_2 (X_3 | X_2) \]

\[ S_1 (O_1 | X_1) \quad S_2 (O_2 | X_2) \quad S_3 (O_3 | X_3) \]

EstiHMM algorithm
APPLICATIONS?
Applications of state sequence estimation

- Speech recognition
- Bio-informatics
  - Finding CpG-islands
  - Locating introns and exons
- Grammatical tagging
- OCR postprocessing
- ...

5th SIPTA school (2012) Jasper De Bock
OCR postprocessing

Optical character recognition software
OCR postprocessing

C → O → W

C → Q → W
OCR postprocessing

C → Q → W
OCR postprocessing

Viterbi

\[ Q_1 (X_1) \quad Q_2 (X_2 | X_1) \quad Q_2 (X_3 | X_2) \]

(usually) only one estimate

\[ C \quad Q \quad W \]

\[ S_1 (O_1 | X_1) \quad S_2 (O_2 | X_2) \quad S_3 (O_3 | X_3) \]
OCR postprocessing

EstiHMM

\( Q_1 (X_1) \quad Q_2 (X_2 | X_1) \quad Q_2 (X_3 | X_2) \)

(sometimes) multiple estimates

\( S_1 (O_1 | X_1) \quad S_2 (O_2 | X_2) \quad S_3 (O_3 | X_3) \)
Calculate relative frequencies in a (small) training set with known hidden states.
OCR postprocessing

\[ Q_1 (X_1) \quad Q_2 (X_2 | X_1) \quad Q_2 (X_3 | X_2) \]

Apply an IDM to a (small) training set with known hidden states

\[ S_1 (O_1 | X_1) \quad S_2 (O_2 | X_2) \quad S_3 (O_3 | X_3) \]
OCR postprocessing

La Divina Commedia

ORIGINAL WORDS IN THE BOOK

CORRESPONDING WORDS IN TEXT DOCUMENT

build an (imprecise) HMM

TRAINING SET

TESTING SET

TRAINING SET

TESTING SET

?
OCR postprocessing

La Divina Commedia

original

VITA

correctly read

digital

VITA

Solution Viterbi

VITA

Solution(s) EstiHMM-algoritme

VITA

5th SIPTA school (2012) Jasper De Bock
OCR postprocessing

La Divina Commedia

original

EH

correctly read

digital

EH

Solution Viterbi

EN

Solution(s) EstiHMM-algoritme

CH

EH

EN
OCR postprocessing

La Divina Commedia

Solution Viterbi

LO

Solution(s) EstiHMM-algoritme

LO

IO

5th SIPTA school (2012) Jasper De Bock
OCR postprocessing

La Divina Commedia

original

CHE

incorrectly read

digital

CNE

Solution Viterbi

ONE

Solution(s) EstiHMM-algoritme

CBE  CHE

CNE  CZE

ONE

5th SIPTA school (2012)
Jasper De Bock
Both algorithms are able to detect and correct errors
The EstiHMM algorithm (in this case) does not introduce errors in words that were already correct
EstiHMM sometimes returns multiple solutions and therefore (of course) includes the correct solution more often
If the EstiHMM algorithm gives a **single solution**, it will be identical to the solution given by the Viterbi algorithm.

EstiHMM giving a **single solution** serves as an indication that:
- the word we are applying it to does not contain errors
- the result returned by the **Viterbi algorithm is correct**
## OCR postprocessing

### La Divina Commedia

<table>
<thead>
<tr>
<th></th>
<th>total number</th>
<th>correct after OCR</th>
<th>wrong after OCR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EstiHMM (multiple solutions)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total number</td>
<td>45 (100%)</td>
<td>8 (17.8%)</td>
<td>37 (82.2%)</td>
</tr>
<tr>
<td>correct solution included</td>
<td>38 (84.4%)</td>
<td>8</td>
<td>30</td>
</tr>
<tr>
<td>correct solution not included</td>
<td>7 (15.6%)</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td><strong>Viterbi</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct solution</td>
<td>23 (51.1%)</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>wrong solution</td>
<td>22 (48.9%)</td>
<td>5</td>
<td>17</td>
</tr>
</tbody>
</table>

- EstiHMM giving **multiple solutions serves as an indication** that
  - the word we are applying it to does indeed contain **errors**
  - the result returned by the **Viterbi algorithm is less reliable**
- EstiHMM can be used to **robustify the precise result** given by the Viterbi algorithm
HOW CAN THIS BE USEFUL?
How can undecisiveness be useful?

- As a method of picking out the hard problems, which you then try to solve with more expensive or time-consuming methods (solve easy cases automatically and use experts only for the difficult ones!)

- If not deciding is a useful choice too, because making a wrong decision is dangerous or expensive (choosing between specific and general medication)
Thanks for your attention!