A simple method for citation metadata extraction using hidden Markov models

Erik Hetzner

(California Digital Library)

JCDL 2008
Advantages of our method

- Good performance on homogeneous citations.
- Reasonable performance on heterogeneous citations.
- Extractor can be implemented in a few pages of code.
Improving HMM performance

- Reduce the size of the alphabet by mapping words to a smaller set of symbols.
- Use two states for each label: first & rest.
- Use ‘separator states’, one for each possible transition between labels.
Hidden Markov models
Alphabet of symbols: words?

exorcised throed deposed roil vaporized rattletrap mocking prohibit sleetier effectual tweeter decremented atrophied nearby captor earn oboe ticked inoculate algorithmic extremist inherited burping silenced harassment doctrinaire emptiest tarting freewheeled parqueting gentlewoman optimal dashboard taskmaster acceptance mucky prototyping virtual recapture perpetrate junking rewrote goody cooperated mottling yahoo gridiron successfully bumper siphoned witchcraft jettison capering grouchier disallowed eyeballing medic sullen certitude tearier parlor becoming morphological cognomen saddening apprenticed signpost lignite wishing boldface postage audibility jingoistic lousy reacted rivulet arboreal primping eddy belatedly necessity ordinance retrogressed perverting sponging neutralizer deadlier inferential easel aptly trapeze circumlocution descanted caressing redeemable entice thunderstruck lectured postmarking twanged belowing rainier grouching cozier flimsiest grizzly decorously jawboning tinier crookeder liberation sleeting heehawed puffin paisley daunt screenwriter ...
Alphabet of symbols: keywords

AND APPEAR COMMUNICATIONS
CONFERENCE DE DISSERTATION
EDITOR IN INC JOURNAL
NOTICES NUMBER PAGES
PHD PRESS PROCEEDINGS
REPORT SUBMITTED TECHNICAL
THEESIS TRANSACTIONS
UNIVERSITY VAN VOLUME
Alphabet of symbols: punctuation

pPERIOD pCOMMA pLEFTPAREN
pRIGHTPAREN pLEFTBRACKET
pRIGHTBRACKET pHYPEN pCOLON
pSEMICOLON pQUESTIONMARK
pMISC pAPOSTROPHE
pDOUBLEQUOTE pSINGLEQUOTE
Alphabet of symbols: word classes

\texttt{wMONTH} \quad \texttt{wSEASON}
Alphabet of symbols: features

fINITIAL    fTC    fUPPER    fLOWER
fNUMERAL4    fNUMERAL    fMIXED
Tokens → symbols

1 $^[aA][nN][dD]$ → wAND
2 $^[Jj]an(uary)$ → cMONTH
3 \. $→ pPERIOD
4 , $ → pCOMMA
5 $^[A-Z]$ → fINITIAL
6 $^[A-Z][A-Z]+$ → fUPPER
...
Tokens → symbols

Tokens $\rightarrow$ symbols

Tokens $\rightarrow$ symbols

Tokens → symbols

fTC pCOMMA fTC fINITIAL pPERIOD pCOMMA

wAND fTC fTC pPERIOD wTHE fTC wTC pPERIOD

fMIXED wEDITION pPERIOD fTC pCOMMA

fTC pPERIOD pCOLON wTHE fUPPER fTC pCOMMA

fNUMERAL4 pPERIOD
Label states

Label states

Label states

Label states


![Diagram with nodes a:f, a:r, and a label fINITIAL]
Label states

Separator states


\[
\begin{tikzpicture}
  \node (a) at (0,0) [circle,draw] {a:f};
  \node (b) at (1,0) [circle,draw] {a:r};
  \node (c) at (2,0) [circle,draw] {a|a};
  \node (d) at (3,0) [circle,draw] {pCOMMA};
  \draw[->] (a) to (b);
  \draw[->] (b) to (c);
\end{tikzpicture}
\]
Separator states


a:f → a:r → a|a → wAND
Label states

Label states

Separator states

Results on the Cora dataset

<table>
<thead>
<tr>
<th>Type</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>token</td>
<td>.944</td>
</tr>
<tr>
<td>field</td>
<td>.892</td>
</tr>
<tr>
<td>whole instance</td>
<td>.613</td>
</tr>
</tbody>
</table>
Improving HMM performance

- Reduce the size of the alphabet by mapping words to a smaller set of symbols.
- Use two states for each label: first & rest.
- Use ‘separator states’, one for each possible transition between labels.
Erik Hetzner

erik.hetzner@ucop.edu

http://purl.net/net/egh/hmm_cite_parser/