A simple method for citation metadata extraction using hidden Markov models

Erik Hetzner

(California Digital Library)

JCDL 2008

CALIFORNIA DIGITAL LIBRARY

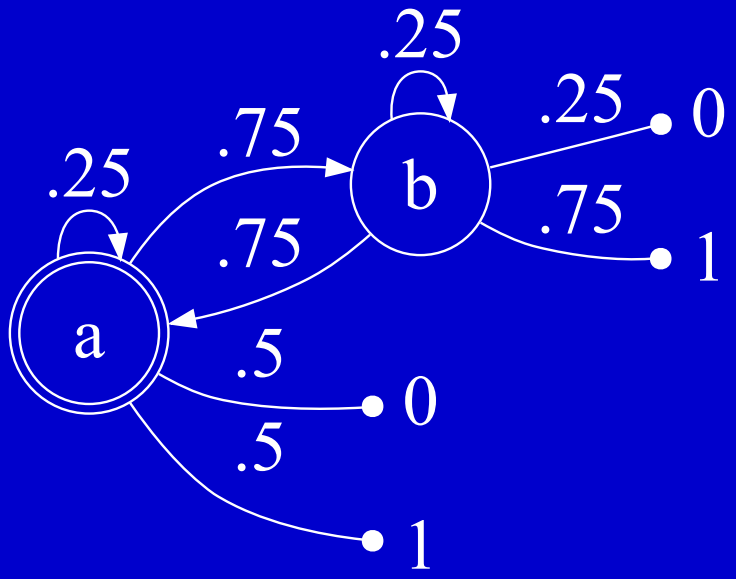
Advantages of our method

- 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 bellowing rainier grouching cozier flimsiest grizzly decorously jawboning tinier crookeder liberation sleeting heehawed puffin paisley daunt screenwriter ...

Alphabet of symbols: keywords

wAND wAPPEAR wCOMMUNICATIONS
wCONFERENCE wDE wDISSERTATION
wEDITOR wIN wINC wJOURNAL
wNOTICES wNUMBER wPAGES
wPHD wPRESS wPROCEEDINGS
wREPORT wSUBMITTED wTECHNICAL
wTHESIS wTRANSACTIONS
wUNIVERSITY wVAN wVOLUME

Alphabet of symbols: punctuation

pPERIOD pCOMMA pLEFTPAREN

pRIGHTPAREN pLEFTBRACKET

pRIGHTBRACKET pHYPEN pCOLON

pSEMICOLON pQUESTIONMARK

pMISC pAPOSTROPHE

pDOUBLEQUOTE pSINGLEQUOTE

Alphabet of symbols: word classes

wMONTH wSEASON

Alphabet of symbols: features

fINITIAL fTC fUPPER fLOWER

fNUMERAL4 fNUMERAL fMIXED

Tokens $\longrightarrow$ 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 $\longrightarrow$ symbols

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.

Tokens $\longrightarrow$ symbols

fTC, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.

Tokens $\longrightarrow$ symbols

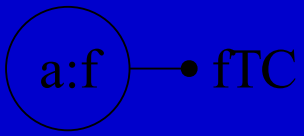
fTCpCOMMA Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.

Tokens $\longrightarrow$ 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																		

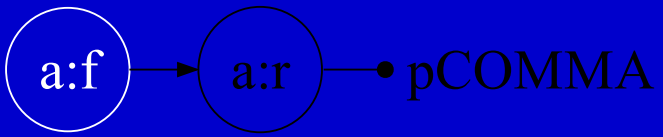
Label states

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.



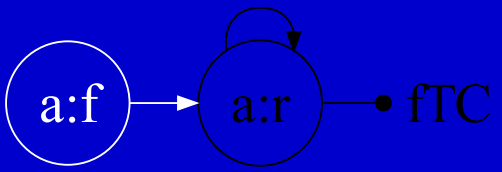
Label states

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.



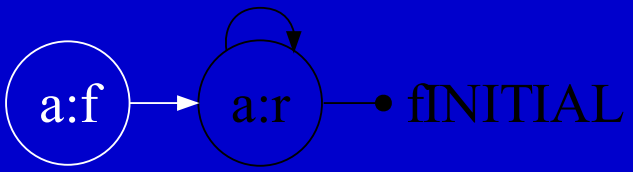
Label states

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.



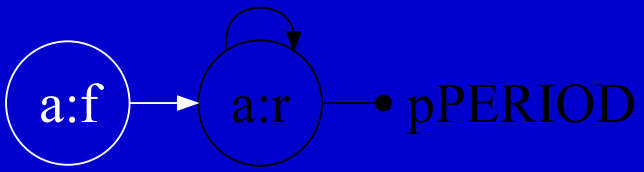
Label states

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.



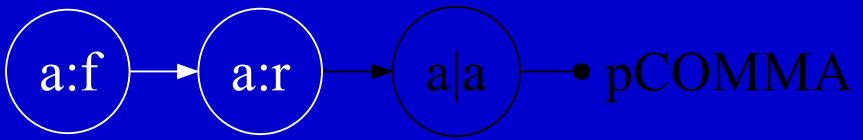
Label states

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.



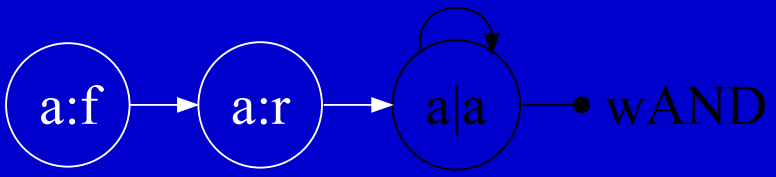
Separator states

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.



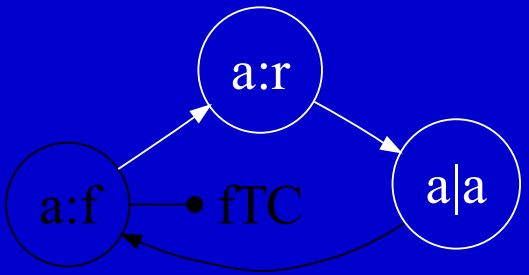
Separator states

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.



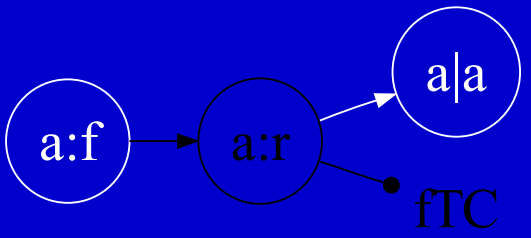
Label states

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.



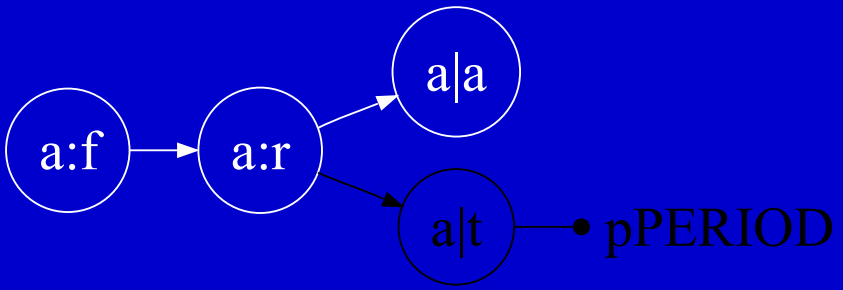
Label states

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.



Separator states

Friedman, Daniel P., and Matthias Felleisen. The Little Schemer. 4th Edition. Cambridge, Mass.: The MIT Press, 1995.



Results on the Cora dataset

token	.944
field	.892
whole instance	.613

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/

CALIFORNIA DIGITAL LIBRARY