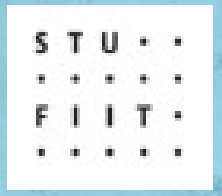
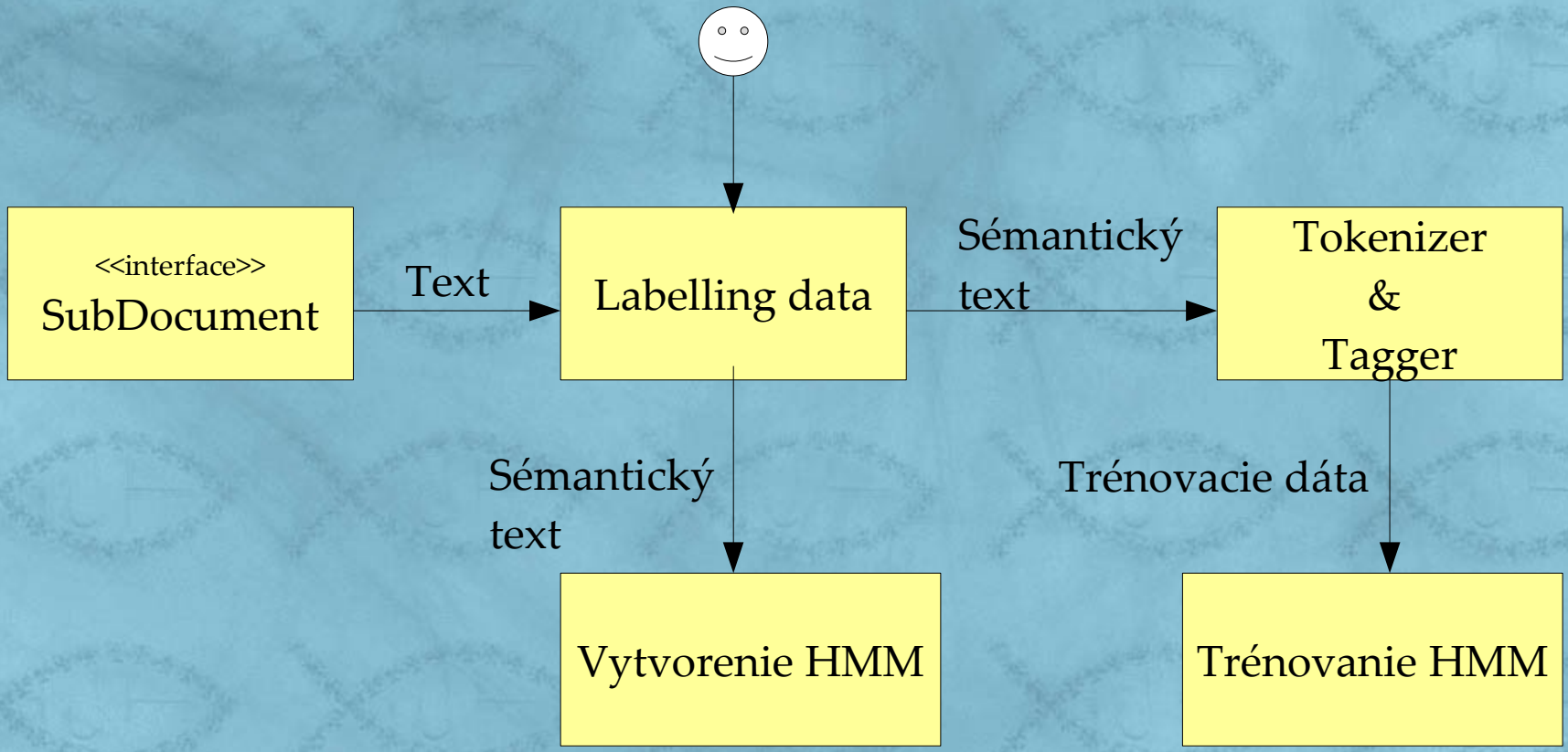
Bibliographic information wrapping

Miroslav Legéň



Motivácia

- framework na učenie sa vzorov z dokumentov
- Polo-automatické wrappovanie
- HMM
- Wrappovanie nie len html



- Apache Lucene
- Jahmm

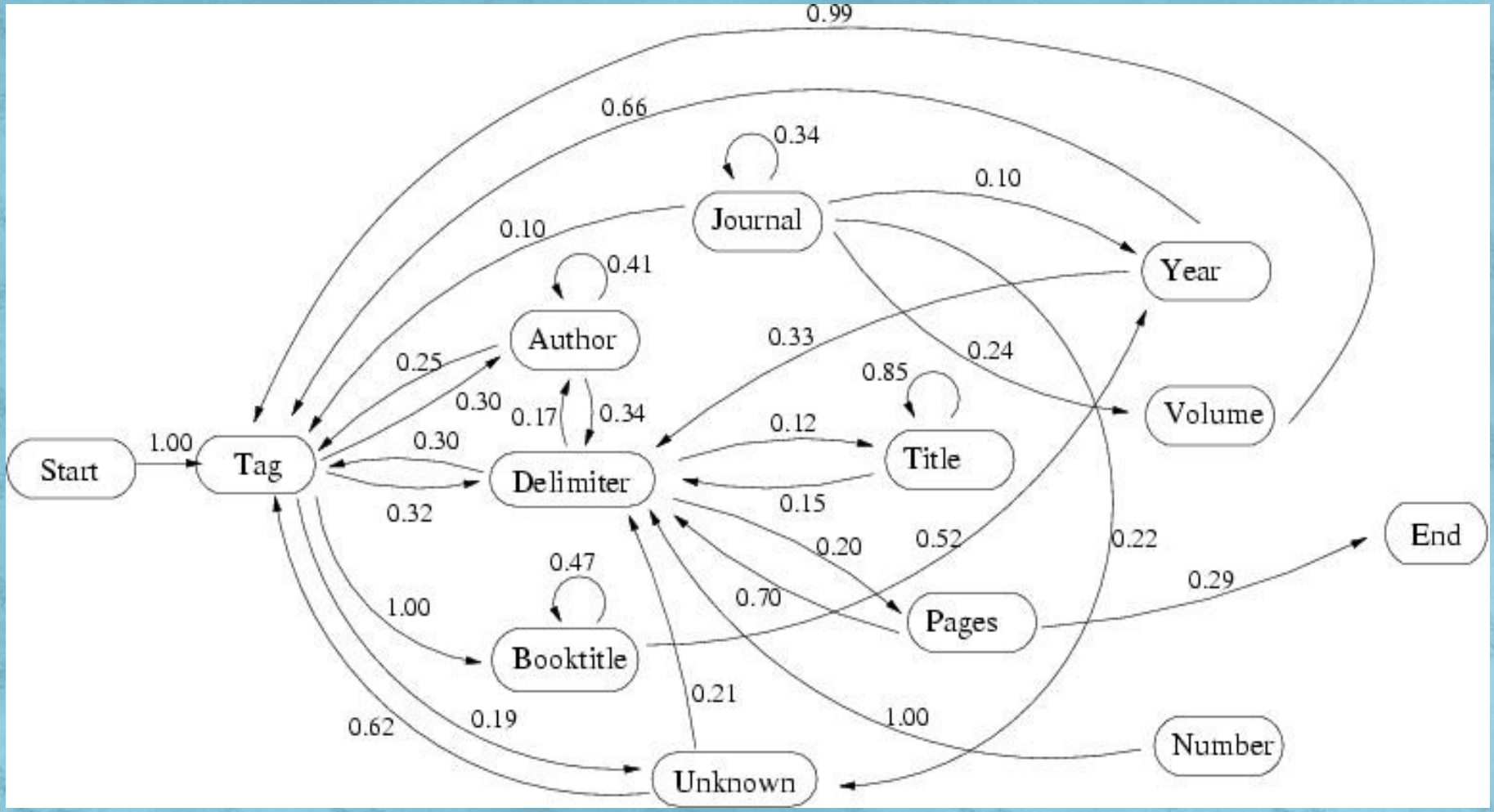
HMM – Hidden markov model

- Pravdepodobnostný konečný automat
- Videteľné stavy sú dané pravdepodobnostnou f()
- Efektívny nástroj pre spracovanie textu
- Štruktúra:
  - A - prechody
  - B - pozorovania
  - ¶ - počiatočné pravdepodobnosti

Použitie HMM

- Vytvárame novú štruktúru
- Stavy = labely + start + end + delimiter + unknown
- Výstupné symboly = tokeny
- Trénujeme model – Baum-Welch algoritmus
- Automatické značkovanie – viterbiho algoritmus

Related work



Obr.: HMM pre BibTex by Junfei Geng

Referencie

- Rabiner L. R. *A tutorial on Hidden Markov Models and Selected Applications in Speech recognition*
- Geng J. *Automatic extraction and integration of bibliographic information on the web using hidden markov models*
- Seymore K., McCallum A., Rosenfeld R. *Learning Hidden Markov Model Structure for Information Extraction*

Ďakujem za pozornosť