Slovenski NLTK označevalnik

Niko Colnerič Nejc Banič

Fakulteta za Računalništvo in Informtiko
Univerza v Ljubljani

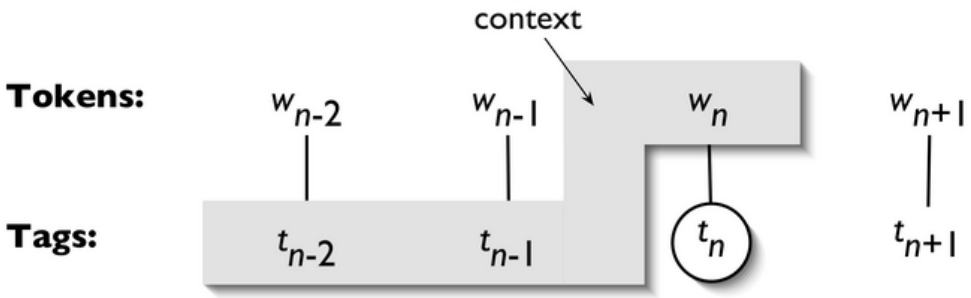
January 16, 2012

Motivacija

This is a short introduction to Beamer class.

Implementirani označevalniki

- **Trigram označevalnik**
  a priori najbolj verjetno oznako v kontekstu dolžine 3



- **Brillov označevalnik**
  ugane oznako vsako besede, nato popravi svoje napake tako, da uporabi seznam transformacijskih pravil
- označevalnik na podlagi **naivnega Bayes-ovega klasifikatorja**
  predpostavi pogojno neodvisnost atributov pri danem razredu

JOS korpus

Osnovna komponenta za izgradnjo slovenskega NLTK označevalnik je JOS korpus. Vsebuje zbirko različnih besedil s podrobno ročno preverjenimi jezikoslovnimi oznakami (npr. leme, oblikoskladenjske oznake itd.). Korpus je kompatibilen z MULTEXT-East V4 oblikoskladenjskimi specifikacijami (ang. *morphosyntactic descriptions (MSDs)*), ki definirajo kategorije (besedne vrste) za slovenski jezik, za vsako kategorijo pa tudi njene atribute in njihove vrednosti. Uporabila sva korpus *jos1M*, ki vsebuje 1 milijon besed z delno ročno preverjenimi lemami in oblikoskladenjskimi oznakami. Struktura JOS korpusa je zapisana v XML.

NLTK trainer

Postopek treniranja označevalnikov

1. Transformacija *.XML* korpusa v *.pos*
2. Treniranje z skripto iz NLTK-trainer
3. Ocenjevanje točnosti
4. Ocenjevanje hitrosti

Primer

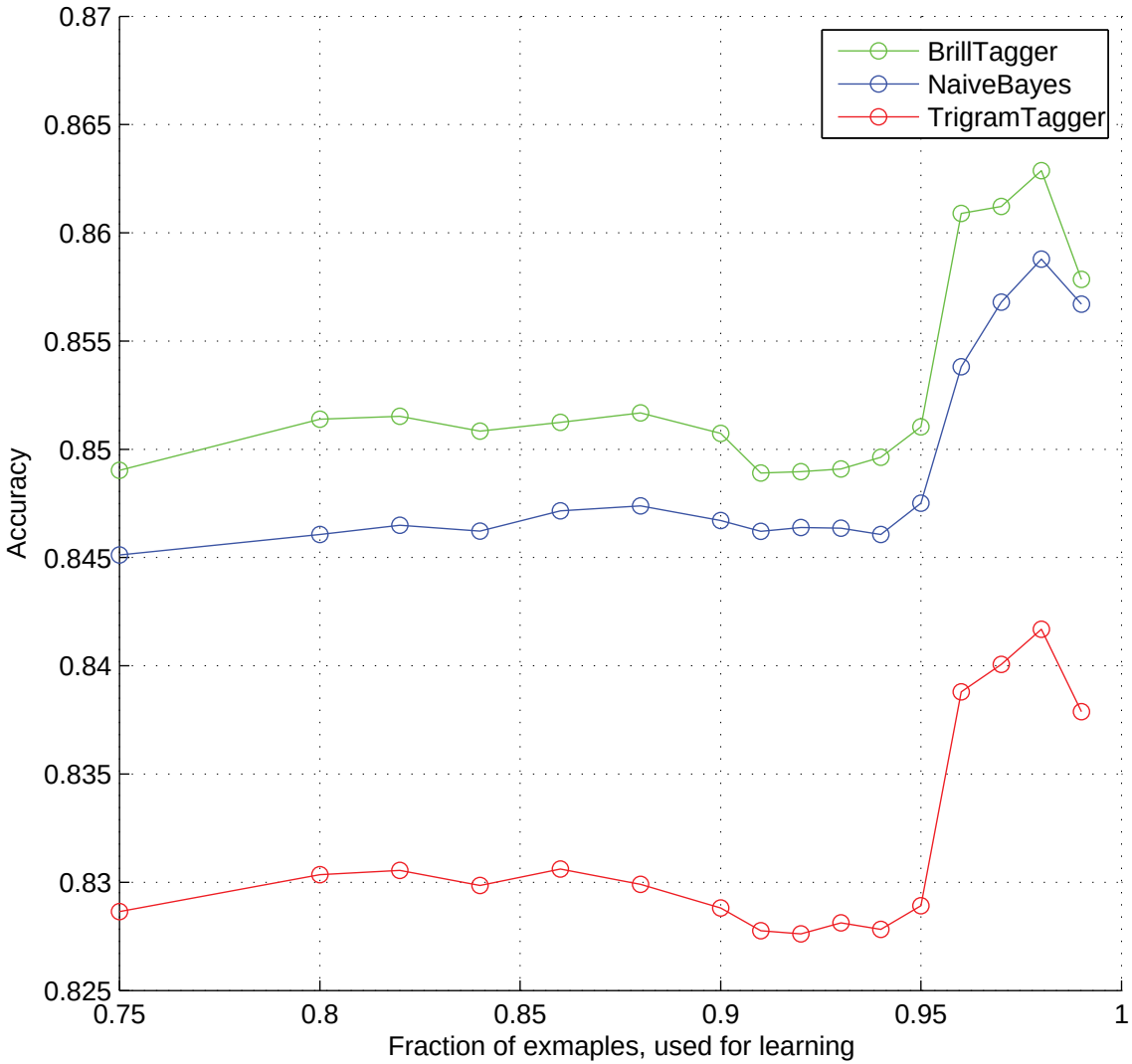
Lep je dan, vse diši že po pomladi

(Lep — PPNMEIN) - pridevnik
(je — GP-STE-N) - glagol
(dan — SOMEI) - samostalnik
(, — ,) - ni razlage
(vse — ZC-SEI) - zaimek
(diši — GGNSTE) - glagol
(že — L) - členek
(po — DM) - predlog
(pomladi — SOZEM) - samostalnik
(! — !) - ni razlage

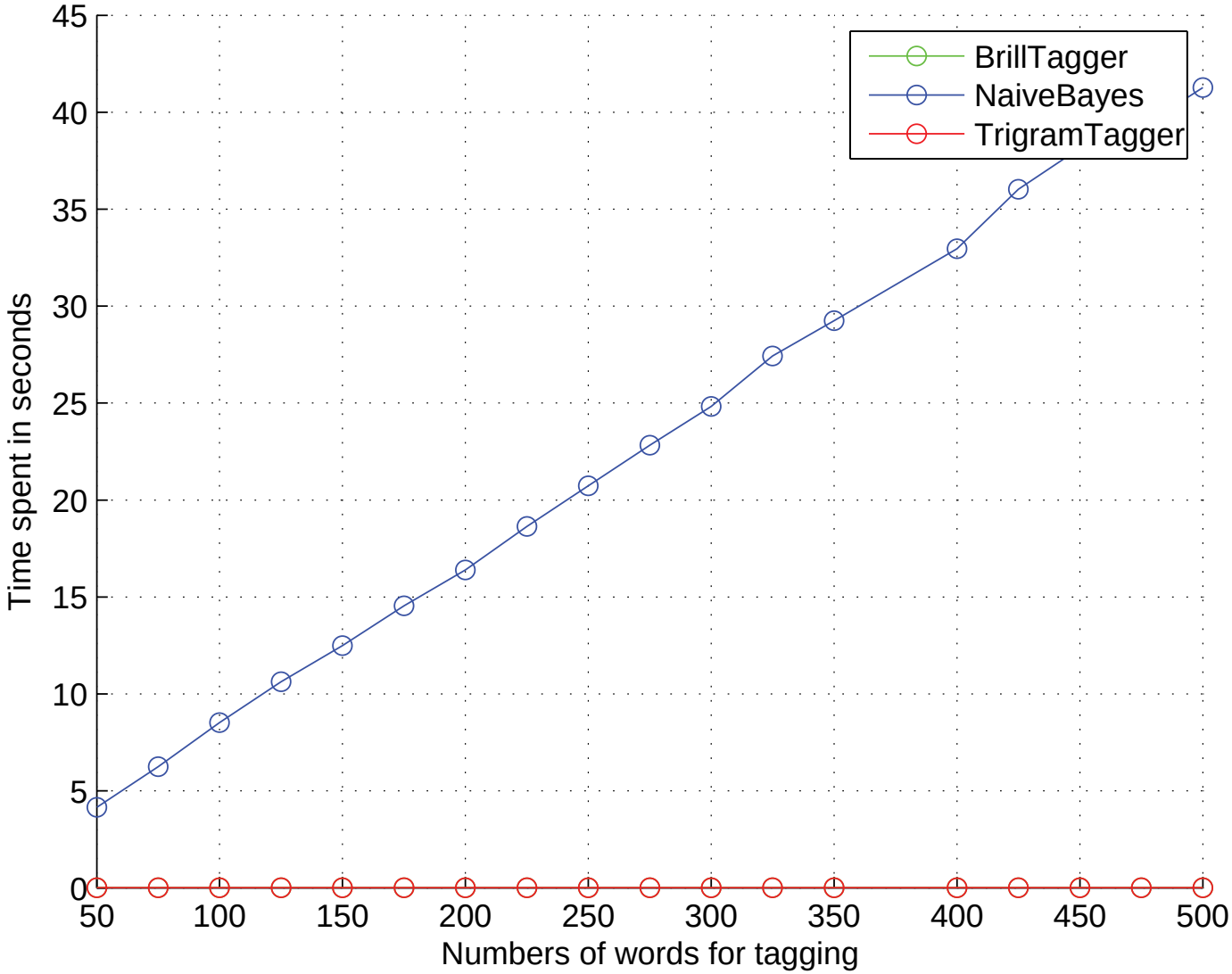
Celoten MSD za *Lep*

pridevnik	
vrsta	splošni
spol	moški
število	ednina
sklon	imenovalnik
živost	0
vid	0
oblika	0
oseba	0
nikalnost	0
stopnja	nedoločeno
določnost	ne
število_svojine	0
spol_svojine	0
naslonskost	0
zapis	0

Rezultati natančnost



Rezultati hitrost



Zaključek, izboljšave

Ne moreva trditi, da so uporabljeni označevalniki (Trigram, Brill in naivni Bayes) najhitrejši in najbolj natančni. Zato je ena možnost za izboljšavo uporaba drugih označevalnikov.
Če bi hotela izboljšati natančnost naših označevalnik, bi lahko v proces učenja uporabila tudi drugih korpuse.
Označevanje v slovenskem jeziku je praktično z uporabo knjižnice *Natural Language Toolkit* in projekta nltk-trainer. Evalvacija je pokazala, da sta tako Brill kot Trigram označila besede hitro, medtem ko je bil naivni Bayes-ov klasifikator veliko počasnejši. Razlike v natančnosti vseh treh označevalnikov praktično ni.