Taggers

PA154 Language Modeling (6.2)

Pavel Rychlý

pary@fi.muni.cz

Statistical Tagger

- using Viterbi algorithm to find the most probable sequence of tags
- sometimes even greedy search works
- the hard part is to find probabilities

TreeTagger

- Helmut Schmid, Stuttgart 1994
- originally developed and evaluated on English, later also German
- disambiguation of proper nouns (named entities) and regular words
- smoothing with Equivalence Classes
  - words with the same set of possible tags
- tag is atomic, no attributes or categories
- probabilities: decision trees
- Vitterbi algorithm

TreeTagger - decision tree

- *house* in “The big house” is
  - *NN* with probability 0.7
  - *ADJ* with probability 0.1

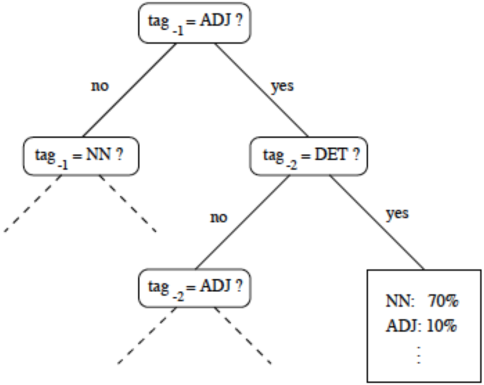
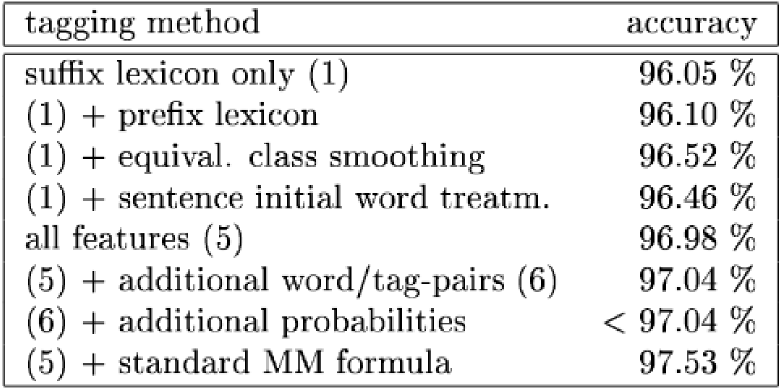


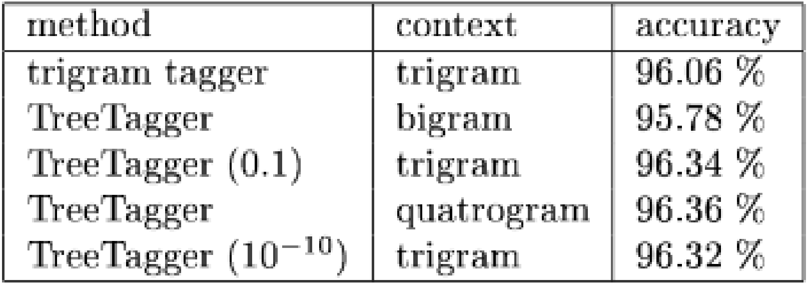
Figure 1: A sample decision tree

TreeTagger – results



tagging method	accuracy
suffix lexicon only (1)	96.05 %
(1) + prefix lexicon	96.10 %
(1) + equival. class smoothing	96.52 %
(1) + sentence initial word treatm.	96.46 %
all features (5)	96.98 %
(5) + additional word/tag-pairs (6)	97.04 %
(6) + additional probabilities	< 97.04 %
(5) + standard MM formula	97.53 %

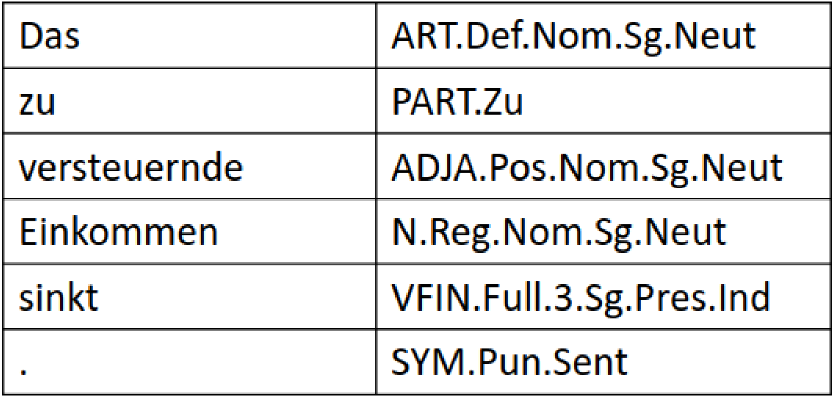
TreeTagger – results



method	context	accuracy
trigram tagger	trigram	96.06 %
TreeTagger	bigram	95.78 %
TreeTagger (0.1)	trigram	96.34 %
TreeTagger	quatrogram	96.36 %
TreeTagger (10^{-10})	trigram	96.32 %

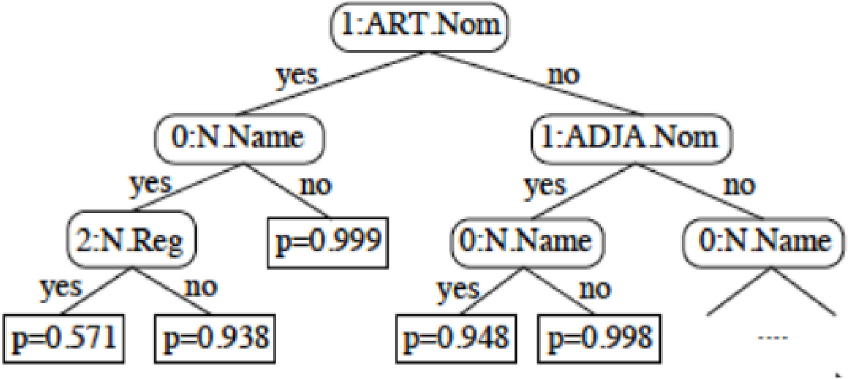
RFTagger

- Helmut Schmid, Florian Laws, Stuttgart 2008
- non-atomic tags

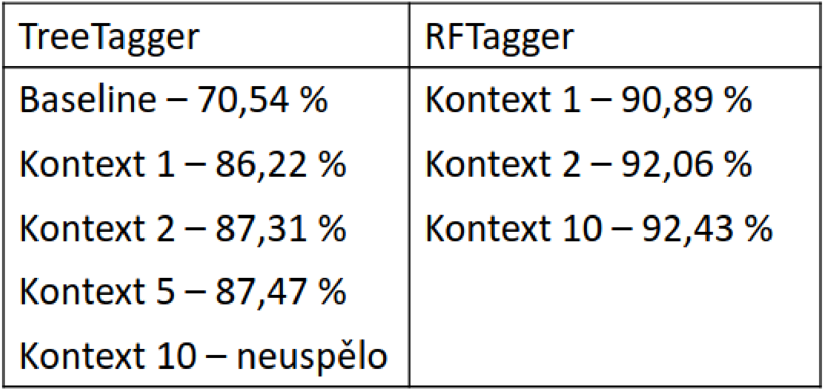


Das	ART.Def.Nom.Sg.Neut
zu	PART.Zu
versteuernde	ADJA.Pos.Nom.Sg.Neut
Einkommen	N.Reg.Nom.Sg.Neut
sinkt	VFIN.Full.3.Sg.Pres.Ind
.	SYM.Pun.Sent

RFTagger – decision tree



RFTagger – results



TreeTagger	RFTagger
Baseline – 70,54 %	Kontext 1 – 90,89 %
Kontext 1 – 86,22 %	Kontext 2 – 92,06 %
Kontext 2 – 87,31 %	Kontext 10 – 92,43 %
Kontext 5 – 87,47 %	
Kontext 10 – neuspělo	