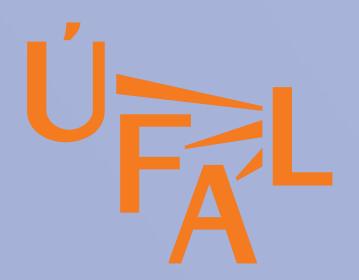


# COMPOST Dutch The Best POS Tagger for Dutch Spoken Language



Jan Raab and Eduard Bejček
Institute of Formal and Applied Linguistics
Charles University in Prague, Czech Republic
{raab,bejcek}@ufal.mff.cuni.cz

CLIN 19
Groningen
Netherlands
January 22, 2009

### **COMPOST**

×

- ~# COMmon POS Tagger
- ~# based on Averaged Perceptron (Collins, 2002)
- ~# Linux and Windows platform
- ~# languages: English, Czech, Slovak ...and Dutch!

# The task



- ~# for every word assign one tag
- ~# in fact, there are two steps:
  - morphological analysis
     (for every word form offer all possible tags)
  - disambiguation
     (chose one tag from possible tags)

# Algorithm



- # data based morphological analysis (for every word all seen tags, for OOV 11 most frequent tags)
- ~# Averaged Perceptron (Collins, 2002)
  - based on HMM
  - main parameter: set of features
- ~# easy to (re)train

# Data



- ~# CGN, Dutch part
- ~# randomly split into:
  - train (5 824 127 words)
  - devel test (642 448 words)
  - eval test (573 873 words)
- ~# tuned on train and dtest
- ~# finally retrained on train+dtest, measured on etest

## Results



- ~# previous best: Bosch et al., 2006: 97,1%
- ~# COMPOST on eval test: 97,27%
- ~# over 1% error reduction (direct comparison on same data not possible)
- ~# speed about 100k words per minute
- ~# easy-to-use application

# Download



- ~# http://ufal.mff.cuni.cz/compost/dutch
- ~# free for academic use
- ~# compiled for Linux and Windows
- ~# online version soon (no need to install anything)

### References



- # van den Bosch, A., Schuurman, I., Vandeghinste, V.: *Transferring PoS-tagging and lemmatization tools from spoken to written Dutch corpus development.* LREC2006.
- # Collins, M.: Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. Proceedings of EMNLP 2002.
- # Hajič, J.: *Disambiguation of Rich Inflection*. Karolinum, Praha, 2004.

en sommige gaan we misschien wel gebruiken dus dan uh dan moet je ze tuurlijk niet afdekken. VG(neven) ADJ(nom,basis,met-e,mv-n) WW(pv,tgw,mv) VNW(pers,pron,nomin,red,1,mv) BW() BW() WW(inf,vrij,zonder) BW() BW() TSW() BW() WW(pv,tgw,ev) VNW(pers,pron,nomin,red,2v,ev) VNW( pers,pron,stan,red,3,mv) BW() BW() WW(inf,vrij,zonder) LET()

afsluiters plafonddozen. N(soort,mv,basis) N(soort,mv,basis) LET()

zijn dat standaard afsluiters plafonddozen. WW(pv,tgw,mv) VNW(aanw,pron,stan,vol,3o,ev) ADJ(prenom,basis,zonder) N(soort,mv,basis) soort,mv,basis) LET()

zijn dat standaardmaat? WW(pv,tgw,mv) VNW(aanw,pron,stan,vol,3o,ev) N(soort,ev,basis,zijd,stan) LET()

nee. TSW() LET()



This work has been supported by grants 1ET100300517 and 1ET101120503 of Grant Agency of the Academy of Science of the Czech Republic and project 201/05/H014 of the Czech Science Foundation.