Statistilised meetodid loomuliku keele süntaktilisel analüüsil

Süntaksiteooriad ja -mudelid 2005/06

Kaili Müürisep

ATI

18. mai 2006

1. Süntaktiline mitmesus

2. Morfoloogiline ühestamine

Statistilised meetodid loomuliku keele süntaktilisel analüüsil

Süntaksiteooriad ja -mudelid 2005/06

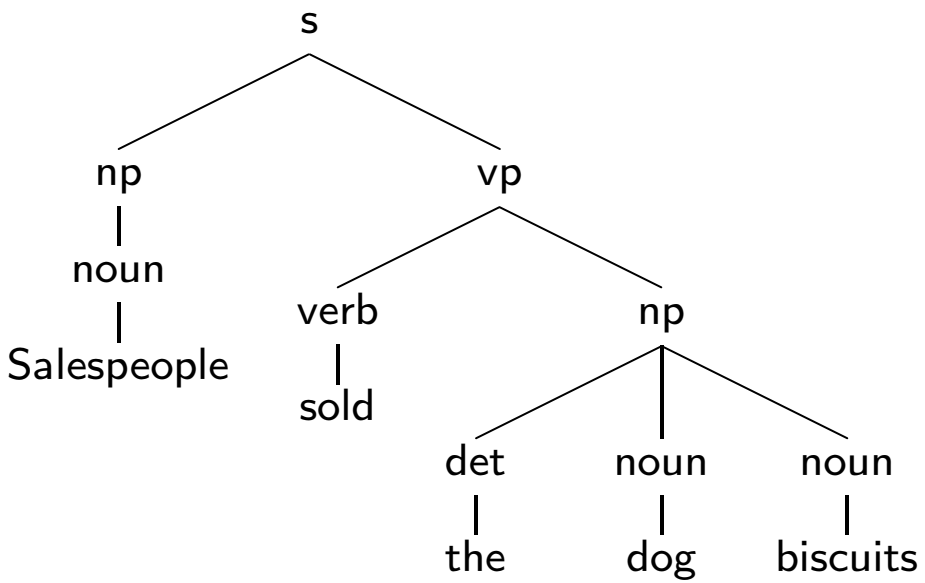
Kaili Müürisep

ATI

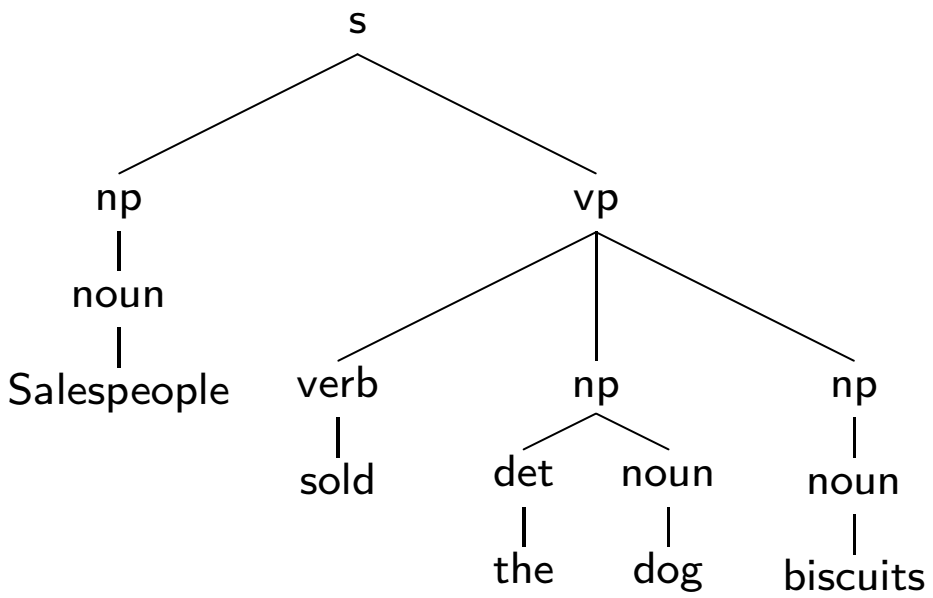
18. mai 2006

Süntaksipuud

(1)

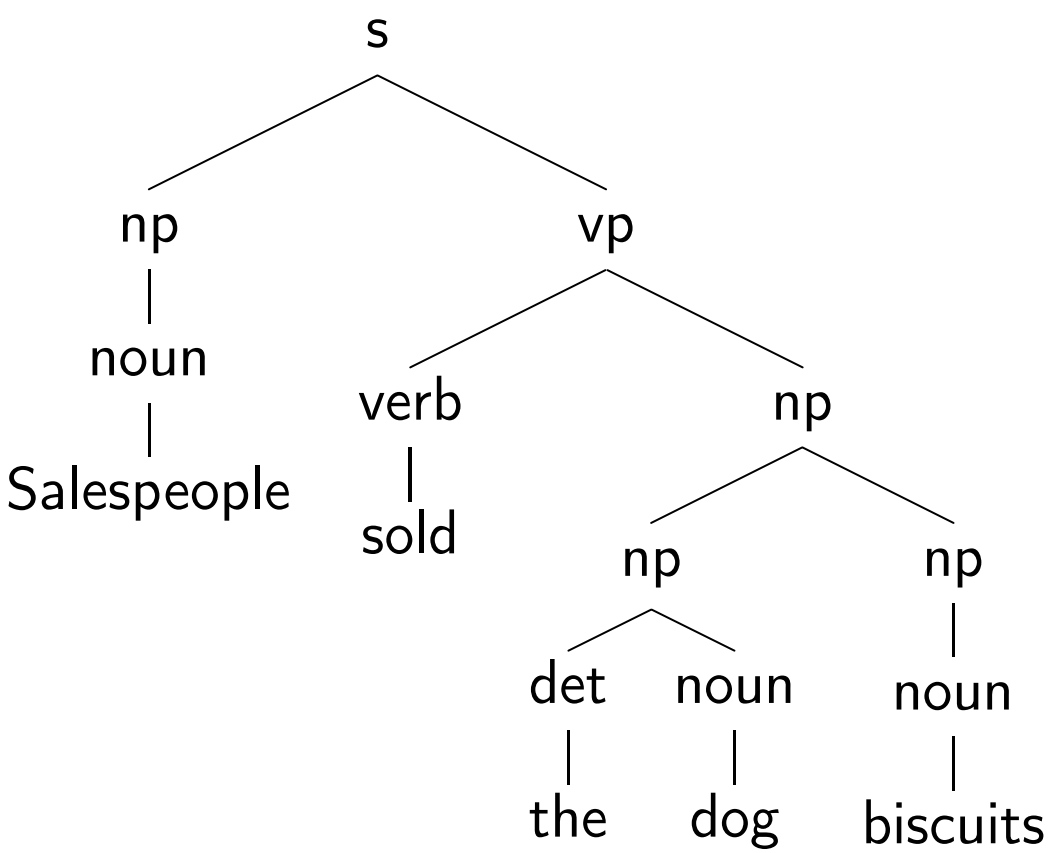


(2)



Veel üks puu

(3)



Probleemist

(4)

The	can	will	rust
det	modal	**modal**	noun
	noun	noun	**verb**
	verb	verb	

Inglise keeles 30-150 POS-märgendit.

Probleemist

(5)

The	can	will	rust
det	modal	**modal**	noun
	noun	noun	**verb**
	verb	verb	

Inglise keeles 30-150 POS-märgendit.

Baseline

Kui on 300000-sõnaline märgendatud korpus. Leida iga sõna jaoks selle kõige sagedasem märgend. Võtta uus tekst ja omistada igale sõnale see märgend. Tundmatud sõnad saavad nime märgendi. Mis võiks olla sellise märgendaja korrektsus?

Probleemist

(6)

The	can	will	rust
det	modal	**modal**	noun
	noun	noun	**verb**
	verb	verb	

Inglise keeles 30-150 POS-märgendit.

Baseline

Kui on 300000-sõnaline märgendatud korpus. Leida iga sõna jaoks selle kõige sagedasem märgend. Võtta uus tekst ja omistada igale sõnale see märgend. Tundmatud sõnad saavad nime märgendi. Mis võiks olla sellise märgendaja korrektsus?

Vastus

90%

Brilli ühestaja ehk tranformatsioonipõhine õppimine

- Leksikaalne märgendaja - lisab kõige tõenäolisema märgendi
- Tundmatute sõnade mõistataja
- Kontekstipõhine märgendaja

Brilli märgendaja näide

Algsed laused

1. Chapman /np killed /vbn John /np Lennon /np
2. John /np Lennon /np was /bedz shot /vbd by /by Chapman /np
3. He /pps witnessed /vbd Lennon /np killed /vbn by /by Chapman /np

Brilli märgendaja näide

Algsed laused

1. Chapman /np killed /vbn John /np Lennon /np
2. John /np Lennon /np was /bedz shot /vbd by /by Chapman /np
3. He /pps witnessed /vbd Lennon /np killed /vbn by /by Chapman /np

Reeglid

1. vbn vbd PREVTAG np
2. vbd vbn NEXTTAG by

Brilli märgendaja näide

Algsed laused

1. Chapman /np killed /vbn John /np Lennon /np
2. John /np Lennon /np was /bedz shot /vbd by /by Chapman /np
3. He /pps witnessed /vbd Lennon /np killed /vbn by /by Chapman /np

Reeglid

1. vbn vbd PREVTAG np
2. vbd vbn NEXTTAG by

Analüüsitud laused -reegel 1

1. Chapman/np killed/vbd John/np Lennon/np
2. John/np Lennon/np was/bedz shot/vbd by/by Chapman/np
3. He/pps witnessed/vbd Lennon/np killed/vbd by/by Chapman/np

Brilli märgendaja näide

Algsed laused

1. Chapman /np killed /vbn John /np Lennon /np
2. John /np Lennon /np was /bedz shot /vbd by /by Chapman /np
3. He /pps witnessed /vbd Lennon /np killed /vbn by /by Chapman /np

Reeglid

1. vbn vbd PREVTAG np
2. vbd vbn NEXTTAG by

Analüüsitud laused -reegel 2

1. Chapman/np killed/vbd John/np Lennon/np
2. John/np Lennon/np was/bedz shot/vbn by/by Chapman/np
3. He/pps witnessed/vbd Lennon/np killed/vbn by/by Chapman/np

Reeglite liigid

- A B PREVTAG C
- A B PREV1OR2OR3TAG C
- A B PREV1OR2TAG C
- A B NEXTTAG C
- A B NEXT1OR2TAG C
- A B SURROUNDTAG C D
- A B NEXTBIGRAM C D
- A B PREVBIGRAM C D

t $\longrightarrow$ t' in context C
Keerukus RKn, kus R - reeglite arv, K - konteksti pikkus, n -sõnade arv
Reeglite teisendamisel automaadiks saavutati keerukus, mis sõltub ainult n-st

TBL algoritm

1. Märgenda iga sõna tema kõige sagedama POS-märgendiga
2. for k= 1, 2, ...
   1. Leia kõik võimalikud transformatsioonid, mis võivad korpuses aset leida
   2. omista t_k -le reegel, mis põhjustab suurimat vigade vähenemist
   3. rakenda t_k korpusel
   4. lõpeta, kui lõpetamiskriteerium on täidetud
3. Väljasta t_1, t_2, ... t_k

Algoritmi rakendamine

Muster: X $\rightarrow$ Y kui järgmine sõna on Z

(7) TRAIN: The/Det race/Noun ends/Verb in/Prep a/Det defeat/Noun
GUESS: The/Det race/Verb ends/Noun in/Prep a/Det defeat/Verb

Algoritmi rakendamine

Muster: X $\rightarrow$ Y kui järgmine sõna on Z

(8) TRAIN: The/Det race/Noun ends/Verb in/Prep a/Det defeat/Noun
GUESS: The/Det race/Verb ends/Noun in/Prep a/Det defeat/Verb

Det $\rightarrow$ Y kui järgmine sõna on 'race' -=1 Verb $\rightarrow$ Noun kui järgmine sõna on ends +=1

Algoritmi probleemid

- ahne algortm - leitakse hetkel parim lahendus
- transformatsioone rakendatakse vasakult paremale

- vasakult paremale AAAA $\rightarrow$ ABAB
- paremalt vasakule AAAA $\rightarrow$ ABBB

Sellegipoolest korrektsus 96-98%, kui pole tundmatuid sõnu.
Kui suurendada treeningkorpust 64.000 sõnalt 640.000 sõnale, kasvab korrektsus 0.4%

5 popimat reeglit

1. NN $\rightarrow$ VB , kui eelmine märgend on TO
2. VBP $\rightarrow$ VB , kui üks 3 eelmisest märgendist on MD
3. VBP $\rightarrow$ VB , kui üks 2 eelmisest märgendist on MD
4. VB $\rightarrow$ NN , kui üks 2 eelmisest märgendist on DT
5. VBD $\rightarrow$ VBN , kui üks 3 eelmisest märgendist on VBZ

N parimat märgendit

n-best wordclass tagging
Vahel on vaja, et märgendus oleks pigem korrektne kui ühene.
Kui erinevused tõenäosuste vahel on väga väikesed, väljastada kõik ligilähedase tõenäosusega märgendid.
Kuidas see võiks välja näha Brilli ühestajas?

N parimat märgendit

n-best wordclass tagging
Vahel on vaja, et märgendus oleks pigem korrektne kui ühene.
Kui erinevused tõenäosuste vahel on väga väikesed, väljastada kõik ligilähedase tõenäosusega märgendid.
Kuidas see võiks välja näha Brilli ühestajas?

Märgendite lisamise reeglid

'yen' esineb ingliskeelsetes tekstides kord ainsuses nimisõnana, kord mitmuses.
Lisa märgend SingularNoun kui vaadeldav sõna on 'yen' ja eelnev sõna on 'the'.

Juhendamata õppimine

Paljud (pooled) sõnad on ühesed - õpi nende põhjal.

(9) The/AT can/NN is/BEZ open

can on NN, kui AT $_B$EZ kontekstis on enamik üheseid sõnu NN-märgendiga

Kasutatakse teistsugust lähenemist:

- Algselt lisatakse kõik leksikonis olevad märgendid
- Reeglid on kujul:
  X_1 X_2 ... $X_n \rightarrow X_i$ kui
  - eelmine sõna on W
  - järgmine sõna on W
  - eelmine märgend on T
  - järgmine märgend on T

Korrektsus 95-96%

Tagasi algse algoritmi juurde

Notatsioonist:

- tõenäosus, et i-nda sõna märgend on t: $p(t \mid w_i)$
- leida t, mille korral tõenäosus suurem: $\arg \max_t p(t|w_i)$
- leida märgendite jada $t_{1,n}$, mille korral tõenäosus suurim: $\arg\max_{t_{1,n}} \prod^{n}_{i=1} p(t_i|w_i)$
- $\arg\max_{t_{1,n}} \prod_i p(t_i|t_{i-1})p(w_i|t_i)$

Korrektsus 91%

Peidetud Markovi mudel

HMM on lõplik automaat, milles olekute siiretel on tõenäosused ja mis väljastab sõnade järjendi.
Eelmisele valemile vastab HMM, mille olekuteks on märgendid, tõenäosus minekuks ühest olekust teise on $p(t_i|p_{i-1})$ ja olekus t_i stringi väljastamisel on $p(w_i|t_i)$.
Kui on antud sõnade järjend, leida olekute järjend, mida masin läbib, et väljastada see sõnade järjend, et tõenäosus oleks suurim.
Kust need tõenäosused sinna tulid?

HMM-i treenimine

Baum-Welchi algoritm

Kui me teame algseid väärtusi, siis me saame leida mudeli olekute kombinatsioonide sagedused ja nende põhjal tõenäosused ümber hinnata.

1. Määra algsed tõenäosused
2. Rakenda BW algoritmi algsetel andmetel ja arvuta uued tõenäosused.
3. korda sammu 2 kuni lõpetamise kriteerium on täidetud (tõenäosused muutusid vähem kui mingi lävi vms)

Dekodeerimine

- Leida antud sõnajärjendi tõenäosus
- Leida parim tee sellise sõnajärjendini

Dekodeerimine

- Leida antud sõnajärjendi tõenäosus
- Leida parim tee sellise sõnajärjendini

Esimene ülesanne

Peame leidma igal ajahetkel t iga oleku S_i kohta tõenäosuse, et ajahetkel t on mudel oleksus S_i, arvestades samuti väljundit)
Tõenäosus = $\sum$ kõigi teede tõenäosused
Peame leidma eelmise sammu tõenäosused, siirdetõenäosused ja väljundsõna väljastamise tõenäosuse.

Dekodeerimine

- Leida antud sõnajärjendi tõenäosus
- Leida parim tee sellise sõnajärjendini

Teine ülesanne ehk dekodeerimine

Leida kõige suurema tõenäosusega olekute jada.
Leidub mitu erinevat teed, mis väljastavad sama tulemuse.
Järjestikuline läbivaatamine - eksponentsiaalne keerukus.
Viterbi algoritm:
Kui me teame kõiki optimaalseid teid, mis viivad kõikidesse olekutesse ajahetkel t-1, on lihtne leida parimat teed ajahetkel t.
Tuleb vaadata nende teede laiendusi, mis väljastavad õige väljundi ja jätta meelde parim iga oleku jaoks.
Alustades ajahetkest 1 konstrueeritakse laiendid järjestikuliselt.
Parim tee, mille abil jõuti lõppolekusse leitakse nii, et jäetakse meelde viimane oleks, millest tuldi ja sel teel saab rekursiivselt olekute järjendi kätte.