Kõnekeele automaatne morfo-süntaktiline märgistamine

Mark Fišel (fishel@ut.ee)

22. märts 2007

Ettekanne sisu

- ülesandest
- morfo-süntaktiline märgistamine
- allikates kasutatud korpused
- kõnekeele märgistamine

Ülesanne

Morfo-süntaktiline märgistamine...

- PoS – leida iga teksti sõna puhul selle leksikaline kategooria (e. sõnaliik)
  - k.a. mitmesuste lahendamine
- morf – sama, aga iga märgis sisaldab infot ka käänete, teguviiside, ajade, isiku numbri jms/vms kohta

...kõnekeele jaoks

- kirjakeele jaoks – praktiliselt lahendatud ülesanne
- kõnekeele jaoks pole olemas nii palju resursse nagu kirjakeele puhul
- otseselt kirjakeele märgistajat ei saa kõnekeelele rakendada
  - teine ortograafia
  - prosoodia
  - ahhid-ohhid jm tundmatud sõnad

Morfo-süntaktiline märgistamine

Rakendatud meetodid

- HMM
- MBL
- TBL
- Bayes
- MaxEnt
- ...

Markovi peitmudel

- tõenäosuslik lõplik seisundiautomaat (PFSA)
- seisundite üleminekute kohta tehtud 1se (n-nda) astme Markovi eeldus

$$p_{trans}(s_i|s_{i-n}, s_{i-n+1}, \ldots, s_{i-1})$$

- lisaks on väljundi emiteerimine (mõlemad antud tõenäosustega)

$$p_{emit}(w_i|s_i)$$

Markovi peitmudel

- kolm ülesannet:
  1. arvutada mis on antud jada tõenäosus vastavalt antud automaadile
  2. antud jada puhul leida kõige tõenäolisem sellele vastav seisundite järjend
  3. leida seisundite üleminekute ja väljundi emiteerimise tõenäosuste hinnangud

HMM + PoS tagging

$$\mathbf{c} = c_1, \ldots, c_k$$

$$\mathbf{w} = w_1, \ldots, w_k$$

$$\arg\max_{\mathbf{c}} P(\mathbf{c}|\mathbf{w}) = \arg\max_{\mathbf{c}} \frac{P(\mathbf{w}|\mathbf{c}) \cdot P(\mathbf{c})}{P(\mathbf{w})}$$

$$P(\mathbf{w}) - const$$

leksikaline mudel: $P(\mathbf{w}|\mathbf{c}) = \prod_i P(w_i|c_i)$

konteksti mudel: $P(\mathbf{c}) = \prod_i P(c_i|c_{i-n}, \ldots, c_{i-1})$

Parameetride hindamine

- supervised learning:

$$P(w_i|c_i) = \frac{C(w_i, c_i)}{C(c_i)}$$

$$P(c_i|c_{i-n}, \ldots, c_{i-1}) = \frac{C(c_{i-n}, \ldots, c_{i-1}, c_i)}{C(c_{i-n}, \ldots, c_{i-1})}$$

- nõuab käsitsi märgistatud korpust
- leiab globaalse maksimumi

Parameetride hindamine

- unsupervised learning:
  - alustame mingite hinnangutega
  - parandame selle kasutades Baum-Welch (edasi-tagasi) algoritmi
- leiab lokaalse maksimumi

Allikates kasutatud korpused

Göteborgi kõnekeele korpus

- Gothenburg Spoken Language Corpus (GSLC)
- transkribeeritud audio ja video lindistused
  - ortograafia muudetud kõnekeele jaoks
- 1.2 mln sõna
- eesmärk – märgistada sõnaliikidega
- kokku 23/11 märgist

Hollandi kõnekeele korpus

- Spoken Dutch Corpus
- 10 mln sõna
- eesmärk – märgistada sõnaliikide ja morfoloogiaga
  - praegune lähenemine – märgiste süüsteem, mis sisaldab kõike vajaliku infot
  - kokku 313 märgist

GSLC märgistamine

GSLC märgistamine

Stockholm-Umeå Corpus (SUC):

- sisaldab kirjakeelt
- märgistatud sõnaliikidega
- + tagasiside (`fb`), jätkajad (`ocm`)
- – punktuatsioon (`dl`), võõrkeelne (`uo`)

Leksikaline mudel

- alge hinnang – SUC põhjal

$$\hat{P}(w|c) = \hat{P}_{SUC}(Std(Pros(w))|c)$$

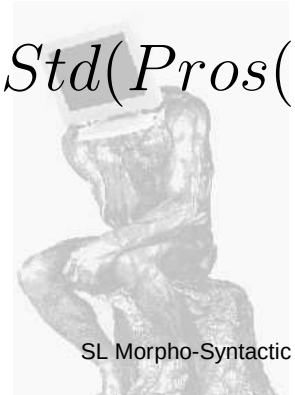
  - tundmatud sõnad:

$$\hat{P}_{SUC}(w_u|c) = \frac{1 - \sum_w C(w,c)}{C(c)} \text{ } (c \text{ – avatud sõnaliik})$$

Leksikaline mudel

- `fb` ja `ocm` väga sagedad kõnekeeles, nende jaoks pole SUC'is statistikat
- $Std(w)$ lisab mitmesust
  - å1 → att (to) VS. att (that)
  - ja{g} → jag (mina) VS. jag (self)
  - dom – nad/neid (kõnek.) VS. verdikt (kirjak.)
- SUC hinnangu välja lülitamise võimaluse vajadus

$$\hat{P}(w|c) = \hat{P}_{Man}(w|c)?\ \hat{P}_{Man}(Pros(w)|c)?$$
$$\hat{P}_{Man}(Std(Pros(w))|c)?\ \hat{P}_{SUC}(Std(Pros(w))|c$$



Leksikaline mudel

Ebaselged sõnad:

$$\hat{P}(w^*|c) = \sum_{w_i \in w^*} \hat{P}(w_i|c)$$

Leksikaline mudel

Lõpetamata sõnad:

$$\hat{P}(w^{+}|\texttt{ocm}) = \hat{P}_{SUC}(w_u|ocm)$$

$$\hat{P}(w^{+}|c) = 0, \text{ kui } c \neq \texttt{ocm}$$

Leksikaline mudel

Tundmatud sõnad:

kui w saab parsida numbrina:

$$\hat{P}(w|\texttt{rg}) = \hat{P}_{SUC}(w_u|\texttt{rg})$$

$$\hat{P}(w|c) = 0, \text{ kui } c \neq \texttt{rg}$$

muidu:

$$\hat{P}(w|c) = \hat{P}_{SUC}(w_u|c)$$

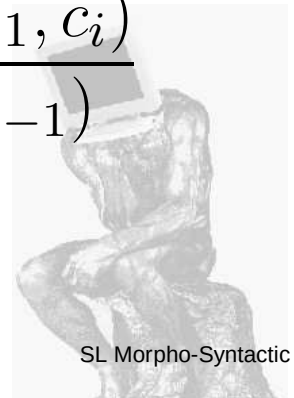
Konteksti mudel

- SUC mudel – treenitud SUC peal:

$$\hat{P}_{SUC}(c_i|c_{i-2}, c_{i-1}) = \frac{C_{SUC}(c_{i-2}, c_{i-1}, c_i)}{C_{SUC}(c_{i-2}, c_{i-1})}$$

- GSLC mudel:
  - GSLC'le rakendati uut leksikalist mudelit koos SUC konteksti mudeliga
  - selle peal treeniti uus GSLC mudel

$$\hat{P}_{GSLC}(c_i|c_{i-2}, c_{i-1}) = \frac{C_{GSLC}(c_{i-2}, c_{i-1}, c_i)}{C_{GSLC}(c_{i-2}, c_{i-1})}$$



Tulemused

- SUC konteksti mudel:
  - suurema märgiste hulgaga täpsus – $93.85\% \pm 0.48\%$
  - väiksema märgiste hulgaga täpsus – $96.16\% \pm 0.39\%$
- GSLC konteksti mudel:
  - suurema märgiste hulgaga täpsus – $95.29\% \pm 0.43\%$
  - väiksema märgiste hulgaga täpsus – $97.44\% \pm 0.32\%$
- kasutatud treenimishulkaga, SUC ja GSLC mudelite erinevus on oluline 0.99 tõenäosusega

SDC märgistamine

SDC märgistamine

- märgistati 3000 sõna käsitsi
- kolm inimmärgistajat, tulemus – "hääletamine"
- lemmatiseerimist testiti selle peal, kasutades 10-fold cross-validation meetodit

Otsene treenimine

- treeniti mitu PoS märgistajat käsitsi märgistatud korpuse peal
- tulemused on 80% ümbruses – väike hulk
- parim tulemus – 82.7%

Mäppinguga märgistamine

- võeti kasutusele mitu PoS märgistajat (treenitud teiste PoS süsteemide jaoks)
- defineeriti käsitsi mäppingut nende ja SDC PoS süsteemide vahel
- parim tulemus – 77.5%

Märgistajate kombineerimine

- 2. taseme otsustusprotsess
- masinõppe süsteemi sisendiks on sõna ja iga märgistaja väljund (omas PoS süsteemis)
- süsteemi väljundiks on SDC märgis
- süsteemiks oli TiMBL
- parim tulemus – 86.6%

Kokkuvõte

- kirjakeele PoS märgistamine on praktiliselt lahendatud küsimus
- kõnekeele PoS märgistamine on raskem kui kirjakeele puhul
- üldine lähenemine – võtta algsüsteemiks kirjakeele märgistajat (supervised), ning sobitada selle kõnekeele jaoks (unsupervised + käsitsi)

Aitäh tähelepanu eest!