16. Rask-ráðstefnan, 26. janúar 2002

Vélræn málfræðigreining með námfúsum markara

Eiríkur Rögnvaldsson
Auður Þórunn Rögnvaldsdóttir
Kristín Bjarnadóttir
Sigrún Helgadóttir

Mörkun texta

- Mörkun (e. *tagging*)
  - að merkja einingar í texta á kerfisbundinn hátt
    - bókstafi, orð, setningar; sérnöfn; erlend orð; o.s.frv.
- Orðflokksmörkun (e. *PoS tagging*)
  - *Gamla*<lo> *konan*<no> *mætti*<so> *þessum*<fn> *tveim*<to> *drengjum*<no> *í*<fs> *morgun*<no>
- Málfræðimörkun
  - kyn, tala, fall, persóna, háttur, tíð, stig, ákveðni

Mörkun og málfarsleiðrétting

- Málfarsleiðrétting er útilokuð án greiningar:
  - villur felast sjaldan í notkun óleyfilegra mynda
    - *föður**s*** í stað *föður*
    - *k**ey**ptu* í stað *k**au**ptu*
  - fremur í að nota réttar myndir á röngum stöðum
    - *Ég hitti syst**i**r þína* > *syst**u**r*
    - *vegna þeirrar tilhneiging**u*** > *tilhneiging**ar***
    - *fjöldi manna kom**u*** > *kom*
    - ***mér** langar* > ***mig** langar*

Stafsetning og vélrænar þýðingar

- Sama gildir um stafsetningarleiðréttingu
  - margar villur finnast aðeins með málgreiningu
    - *það er kominn morgu**n*** > *morgun**n***
    - *ég hitti Kristi**nn*** > *Kristi**n***
    - *hann er fari**n*** > *fari**nn***
- Vélrænar þýðingar krefjast málgreiningar
  - annars eru þær bara uppfletting í orðasafni
    - *hot spring river this book* (*hver á þessa bók*)

Málfræðileg mörkun

- Markari (e. *tagger*)
  - forrit sem markar (e. *tags*) texta
- Beygingarlýsing er forsenda málfræðimörkunar
  - hverju orði í textanum er flett upp í orðasafni
  - beygingarupplýsingar úr því færðar inn í textann
    - *í* fs.
    - *hesturinn* no. kk. et. nf. m.gr.
    - *fóruð* so. 2. pers. ft. þt. fh. gm.

Margræðni orðmynda

- Ekki fá öll orð ótvíræða greiningu í fyrstu lotu
  - þótt *í* sé ótvírætt orð er *á* það ekki:
    - *á* fs.
    - *á* so. 1./3. pers. et. nt. fh. gm. (*eiga*)
    - *á* no. kvk. et. nf./þf./þgf. (*á*)
    - *á* no. kvk. et. þf./þgf. (*ær*)
  - þótt *fóruð* sé ótvírætt orð er *fórum* það ekki:
    - *fórum* so. 1. pers. ft. þt. fh. gm. (*fara*)
    - *fórum* no. kvk. ft. þgf. (*fórur*)

Þrenns konar markarar

- Annað þrep í vinnslunni er nauðsynlegt
  - til að leysa úr slíkri tví- og margræðni
- Í því þrepi er önnur eða ein greiningin valin
  - en hinni eða hinum hafnað
- Markarar skiptast í:
  - reglumarkara (e. *rule-based taggers*)
  - tölfræðimarkara (e. *statistical/stochastic taggers*)
  - námfúsa markara (e. *transformation-based taggers*)

Reglumarkarar

- Reglumarkarar nota reglur um gerð setninga og setningarliða til að marka orðin
  - forsetning kemur t.d. sjaldan næst á undan sögn
    - því er líklegt að orðið *fórum* sé fremur nafnorð en sögn í sambandinu *í fórum mínum*
  - eignarfornafn sambeygist undanfarandi nafnorði
    - í sambandinu *hesta þinna* er *þinna* ótvírætt eignarfall og þannig sést að *hesta* er ef. en ekki þf.

Tölfræðimarkarar

- Tölfræðimarkarar nota tíðniupplýsingar til að velja líklegustu greininguna
  - *á* er mun oftar forsetning en nokkuð annað
    - yrði því rétt greint í setningunni *Ég er á leiðinni*
    - en ranglega greint sem fs. í setningunni *Ég á þetta*
  - *fórum* er mun oftar mynd af so. en af no.
    - yrði því rétt greint í setningunni *Við fórum heim*
    - en ranglega greint sem so. í sambandinu *í fórum mínum*

Munur á tegundum markara

- Báðar þessar tegundir hafa kosti og galla
  - sem þarf að vega og meta hverju sinni
- Fljótlegra er að koma upp tölfræðimörkurum
  - fjögur ársverk fóru t.d. í norskan reglumarkara
  - en þrír mannmánuðir í norskan tölfræðimarkara
- Reglumarkarar skila réttari greiningu
  - og ráða betur við margbrotna greiningu

Þjálfunarsafn

- Þjálfunarsafn (e. *training corpus*)
  - texti sem hefur verið greindur í höndunum
  - eftir sama kerfi og vélræna greiningin notar
- Þetta nýtist við gerð beggja tegunda markara
  - til að átta sig á mynstrum í textanum
    - sem hægt er að setja fram í regluformi
  - til að sjá tíðni mismunandi greininga sömu mynda
    - svo að hægt sé að velja líklegustu greininguna

Markari Brills

- Þekkt útfærsla kennd við Eric Brill
  - *Brill's tagger*, *Brill type tagger*
- Byggist á *transformation based learning*
  - markarinn er keyrður á þjálfunarsafn
  - þar sem hvert orð hefur a.m.k. tvo greiningarstrengi
  - í einræðum orðmyndum eru strengirnir samhljóða
  - fleirræð orðmynd fær fleiri greiningarstrengi
  - síðan þarf að velja rétta strenginn

Dæmi úr *Wall Street Journal Corpus*

- `wd(7799,a).`
- `tag(7799,'DT').`
- `tag('DT','DT',7799).`

- `wd(7800,good).`
- `tag(7800,'JJ').`
- `tag('JJ','JJ',7800).`

- `wd(7801,buy).`
- `tag(7801,'VB').`
- `tag('VB','NN',7801).`

- *wd* = orð
- *tag* = mark (greining)
- *7799, 7800, 7801* = hlaupandi númer
- *a* DT (greinir)
- *good* JJ (lo.)
- *buy* VB (so.)
- *buy* NN (no.)

Sniðmát og reglur

- *a good buy*
  - `tag:A>B <- tag:C@[-1].`
    - Breytið greiningarstreng *A* í greiningarstreng *B*
    - ef undanfarandi orð hefur greiningarstreng *C*
- *a bad taste*
  - `tag:VB>NN <- tag:JJ@[-1].`
    - Breytið greiningunni *so.* (VB) í greininguna *no.* (NN)
    - ef orðið á undan er *lo.* (JJ)

Reglusafnið prófað

- Markarinn kemur sér upp reglusafni
  - þegar hann er keyrður á þjálfunarsafnið
- Reglurnar eru mismargar og mismunandi
  - eftir stærð þjálfunarsafnsins
  - eftir því hversu oft hvert samband kemur fyrir
  - eftir því hvernig sniðmátin eru
- Rétt greint prófunarsafn (e. *test corpus*)
  - er nauðsynlegt til að meta gæði reglusafnsins

Giskari og markari

- Að fengnu reglusafni þarf að skrifa tvö forrit:
  - giskara (e. *unknown word guesser*)
    - til að greina orð sem finnast ekki í orðasafninu
    - eftir endingum, viðskeytum o.fl.
  - markarann sjálfan, sem
    - flettir upp í orðasafni með beygingarupplýsingum
    - skrifar mögulegar greiningar orða inn í textann
    - velur réttu greininguna í samræmi við reglusafnið

Verkefnið

- Að búa til málfræðimarkara fyrir íslensku
  - sem geti greint texta með a.m.k. 95% nákvæmni
- Byggt verður á *μ-tbl* eftir Torbjörn Lager
  - sem er útfærsla á markara Brills
- Þjálfunarsafn er til hjá Orðabók Háskólans
  - grunnskrár að *Íslenskri orðtíðnibók* (1991)
    - ritstjóri: Jörgen Pind
    - vélræn greining: Stefán Briem
    - handvirk greining: Friðrik Magnússon

Úr grunni *Íslenskrar orðtíðnibókar*

f p k e n	hann	hann
s f g 3 e þ o	átti	eiga
n h e o	afmæli	afmæli
a o	í	í
n k e o	dag	dagur
c	og	og
n k e n g	hvolpurinn	hvolpur
n k e n - m	Vaskur	Vaskur
s f g 3 e þ	var	vera
n v e n	afmælisgjöf	afmælisgjöf

Markaskrá og skipting safnsins

- Greiningin í *Íslenskri orðtíðnibók* er nákvæm
  - notuð er stór markaskrá (e. *tagset*)
  - alls 621 mismunandi greiningarstrengur
- Heildarsafnið er 500 þúsund orð og skiptist í
  - þjálfunarsafn:
    - verður allt að 450 þúsund orð – 48 þúsund í forkönnun
  - prófunarsafn:
    - verður allt að 50 þúsund orð – 12 þúsund í forkönnun

Öðrum greiningarstreng bætt við

- Hér hefur viðbótar-strengur verið keyrður inn í skrána

- `wd(38,'til').`
- `tag(38,'ae').`
- `tag('ae','ae',38).`

- `wd(39,'enda').`
- `tag(39,'c').`
- `tag('c','nkee',39).`

- *enda* fær hér viðbótar-strenginn ‘c’ (samtenging)
- Það er algengasta grein-ingin á *enda* í *Íslenskri orðtíðnibók* (189 dæmi)
- Samhengið sýnir þó að rétta greiningin á *enda* er hér ‘nkee’ (no. kk. et. ef.)
- Um þá greiningu eru hins vegar aðeins 7 dæmi í *Íslenskri orðtíðnibók*

Sniðmát fyrir íslensku

- `tag:A>B <- tag:C@[-1].`
- `tag:A>B <- tag:C@[1].`
- `tag:A>B <- tag:C@[-1,-2].`
- `tag:A>B <- tag:C@[-1,-2,-3].`
- `tag:A>B <- tag:C@[-1] & tag:D@[1].`
- `tag:A>B <- tag:C@[-1] & tag:D@[-2].`
- `tag:A>B <- tag:C@[-1] & tag:D@[-2] & tag:E@[-3].`
- `tag:A>B <- tag:C@[1,2].`
- `tag:A>B <- tag:C@[-1] & tag:D@[1,2].`
- `tag:A>B <- wd:C@[0].`
- `tag:A>B <- wd:C@[1].`
- `tag:A>B <- wd:C@[-1].`
- `tag:A>B <- wd:C@[0] & wd:D@[-1].`
- `tag:A>B <- wd:C@[0] & tag:D@[-1].`
- `tag:A>B <- wd:C@[0] & tag:D@[1].`
- `tag:A>B <- wd:C@[-1,-2].`
- `tag:A>B <- wd:C@[0] & wd:D@[-1] & wd:E@[-2].`

Fyrsta tilraun – stærsta markaskrá

- 79,5% í prófunarsafni fá ótvíræða greiningu
  - algengasta greiningin jafnframt sú rétta
- *μ-tbl* forritið keyrt þrisvar á þjálfunarsafnið
  - lærir nýjar reglur í hverri umferð
  - alls 609 reglur
- Síðan eru reglurnar keyrðar á prófunarsafnið
  - villum fækkar þá úr 2445 í 1026
  - þannig að 91,5% greiningarstrengja eru réttir

Önnur tilraun – fallstjórn tekin út

- 89,0% í prófunarsafni fá ótvíræða greiningu
  - algengasta greiningin jafnframt sú rétta
- *μ-tbl* forritið keyrt þrisvar á þjálfunarsafnið
  - lærir nýjar reglur í hverri umferð
  - alls 339 reglur
- Síðan eru reglurnar keyrðar á prófunarsafnið
  - villum fækkar þá úr 2445 í 613
  - þannig að 95,0% greiningarstrengja eru réttir

Niðurstöður úr forkönnun

- Árangurinn er að okkar mati mjög góður
  - tæplega verður komist hærra en í 98%
    - eftir það fer málfræðinga að greina á
- Nákvæm greining hefur bæði kosti og galla
  - gerir greiningu sumra orða erfiðari en ella
    - það er t.d. oft erfitt að greina fallstjórn so. og fs.
  - en auðveldar oft greiningu orða í umhverfinu
    - fallgreining er auðveldari ef fallstjórn er greind

Nokkrar íslenskar reglur

- `tag:sfg3eþ  >sfg1eþ  <- tag:fp1en@[-1,-2]`
- `tag:cn  >c  <- tag:svg3en@[1,2]`
- `tag:cn  >c  <- tag:svg3eþ@[1,2]`
- `tag:af  >fp1fn  <- wd:við@[0] & tag:sfg1fn@[1]`
- `tag:sfg3en  >sfg1en  <- tag:fp1en@[-1,-2]`
- `tag:cn  >c  <- tag:sfg3eþ@[1,2]`
- `tag:af  >fp1fn  <- wd:við@[0] & tag:sfg1fþ@[1]`
- `tag:sfg3eþ  >sfg1eþ  <- tag:fp1en@[1]`
- `tag:svg3eþ  >svg1eþ  <- tag:fp1en@[-1]`
- `tag:fpken  >fpkeo  <- tag:af@[-1]`
- `tag:cn  >c  <- tag:sfg3en@[1,2]`
- `tag:sfg3en  >sfg2en  <- tag:fp2en@[-1,-2]`
- `tag:ssg  >sþghen  <- wd:var@[-1,-2]`
- `tag:foheþ  >lheþsf  <- wd:einu@[0] & wd:í@[-1]`
- `tag:fahen  >faheo  <- tag:af@[-1]`
- `tag:af  >fp1fn  <- wd:við@[0] & tag:sfg1fn@[-1]`

Þgf. greint sem þf. á eftir fs.

- 10 dæmi um nve**o** sem ættu að vera nve**þ**
- 327: að þeir séu sleipir ***í*** **sögu** . það er eiginlega ekki
- 489: að fara í andaglas ***í*** **tölvu** . hvað þessir strákar þykjast
- 1372: ofan í sig , þessari **túpu** sem átti að liggja slétt
- 3130: sem hún fylgdist ***með*** hverri **hreyfingu** hetju sinnar . Alli gaf
- 4101: ætlaði að koma við ***í*** **sjoppu** og kaupa eitthvert sælgæti
- 5469: einar dyr voru ***á*** **framhlið** hans en þær voru sjaldan
- 5049: var farið að svima ***af*** **eftirvæntingu** . hún rétti út höndina
- 5883: einhverjum ástæðum var ömmu ***í*** **nöp** við þessa iðju . Jóra
- 6586: Stóri-Jón reyndi ***eftir*** bestu **getu** að segja frá ferðalaginu
- 10351: súkkulaði og kaffi inni ***í*** **stofu** . meira að segja Guðmundur

það greint sem nf. á eftir so.

- 9 dæmi um *fphe**n*** sem ættu að vera *fphe**o***
- 6662: átti ég að ***vita*** **það** muldraði lögregluþjónninn . má ég
- 8193: undir fótum sér , ***sá*** **það** , heyrði það , fékk
- 8199: heyrði það , ***fékk*** **það** í fangið þegar hann datt
- 9864: flugmaðurinn ***sagði*** mér **það** . þá lýgur hann því
- 10563: skotinn í henni , ***sérðu*** **það** ekki ? hvíslaði Kata spekingslega
- 10600: í bók að maður ***sjái*** **það** á augunum í fólki .
- 10873: jólafríið kemur . kennarinn ***segir*** **það** . ætlar hann að kenna
- 11426: já . amma mín ***kallaði*** **það** að krossa sig , en
- 11870: skapað líf . Jesús ***getur*** **það** . af því að hann

Vh. greindur sem fh. á eftir st.

- 7 dæmi um *sfg3eþ* sem ættu að vera *svg3eþ*
- 597: Grímsa . spurði ***hvort*** hann **vissi** hvað þetta ætti að þýða
- 906: ágúst 1741 ***án þess*** nokkur **hirti** um . Palli greip andann
- 2324: mömmu , ***sagði að*** hún **passaði** vel upp á þetta ,
- 3653: fer í . ***ef*** mamma **réði** væri ég algjört smábarn ,
- 5700: sárum hans ***á meðan*** hún **hjálpaði** ömmu að þrífa sjálfa sig
- 6134: var ***eins og*** Salómon Þór **áttaði** . sig hann tók á
- 6239: var ***engu líkara en*** hann **ætlaði** að detta um koll í

So. og lo. ranglega greind sem ao./fs.

- 4 dæmi um *af* sem ættu að vera *sfg3en*
- 5806: í átt að skrifpúltinu , **á** eftir að draga okkur öll
- 9338: fölnaði upp og hrópaði **á** nú að ráðast á mann
- 9497: afi ríkur . og hver **á** að passa köttinn fyrir Önnu
- 9817: þessu nútíma gargani . það **á** að vera hiti en miðstöðin
- 4 dæmi um *af* sem ættu að vera *lhensf*
- 1025: viss um að allt sé **rétt** sem stendur í gömlum bókum
- 4499: þetta allt í einu orðið **langt** , stundi Tóti og virtist
- 6530: hjálpar var það of . **seint** , hann litli bróðir og
- 8550: ömmu . en þó er **mikið** eftir . amma mín ég

Árangurinn má bæta með því að:

- Stækka þjálfunarsafnið
  - þannig fást fleiri og betri reglur
- Fjölga sniðmátum og endurbæta þau
  - e.t.v. leyfa þeim að skoða stærra umhverfi
- Einfalda greininguna
  - ekki er þó víst hvaða áhrif það hefur
- Lagfæra reglurnar eftir á
  - skoða villur og bæta reglum við handvirkt

Þökk fyrir áheyrnina

- Eiríkur Rögnvaldsson
  - eirikur@hi.is
- Auður Þórunn Rögnvaldsdóttir
  - audurro@hi.is
- Kristín Bjarnadóttir
  - kristinb@lexis.hi.is
- Sigrún Helgadóttir
  - sigrun.h@simnet.is