T-(538|725)-MALV, Málvinnsla
Mörkun - með reglum

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]Tölvunarfræðideild, Háskólinn í Reykjavík

September 2007

Outline

1 Mörkun

2 Nákvæmni í mörkun

3 Tegundir markara

4 Málfræðilegir reglumarkarar

5 Markari sem lærir reglur

Outline

1 Mörkun

2 Nákvæmni í mörkun

3 Tegundir markara

4 Málfræðilegir reglumarkarar

5 Markari sem lærir reglur

Hvað er mörkun (e. POS tagging)?

Skilgreining

- Að **marka** (e. to tag) sérhvert orð í texta með **marki** (e. tag) (greiningarstreng).
- Markið sýnir orðflokk og beygingarleg einkenni.

Af hverju vandamál?

- Sum orð eru margræð (e. ambiguous).
- Uppfletting á orði í orðasafni eða niðurstaða úr orðhlutafræðilegri greiningu $\Rightarrow$ fleiri en eitt mark (greining).
- Talað um markara sem "disambiguator".
- Markari framkvæmir *einræðingu*.

Hvað er mörkun (e. POS tagging)?

Skilgreining

- Að **marka** (e. to tag) sérhvert orð í texta með **marki** (e. tag) (greiningarstreng).
- Markið sýnir orðflokk og beygingarleg einkenni.

Af hverju vandamál?

- Sum orð eru margræð (e. ambiguous).
- Uppfletting á orði í orðasafni eða niðurstaða úr orðhlutafræðilegri greiningu $\Rightarrow$ fleiri en eitt mark (greining).
- Talað um markara sem "disambiguator".
- Markari framkvæmir *einræðingu*.

Mörkun

Gagnsemi mörkunar

- Markið gefur mikilvægar upplýsingar um orðið og umhverfi þess.
  - "You shall know a word by the company it keeps" (Firth, 1957)
  - Kyn, tala og fall lýsingarorðs gefur t.d. til kynna sambærilega þætti eftirfarandi nafnorðs.
- Hjálpar til við talgervingu.
  - OBject (nafnorð) vs. obJECT (sögn)
- Grunnur að málfræðileiðréttingu, vélrænum þýðingum, setningagreiningu.
- Gerð markaðra málheilda.

Mörkun

Markamengi

- **Markamengi** (e. tag set) er mengi allra greiningarstrengja (marka).
- Skilgreind eru ólík markamengi fyrir mismunandi tungumál.
- Einnig ólík markamengi fyrir mismunandi verkefni innan sama tungumáls.
- **Íslenska**: *Íslensk orðtíðnibók* – 660 mörk.
- **Enska**: *Penn Tree Bank*: 45 mörk, *Brown Corpus*: 87 mörk.
- **Sænska**: *Parole*: 139 mörk.
- **Tékkneska**: Um 1000 mörk.

Mörkun

Full einræðing (e. full disambiguation)

- Sérhverju orði er úthlutað aðeins einu marki.
- Algengast en ...
- ...stundum getur markari ekki framkvæmt fulla einræðingu.

Dæmi um mörkun íslensku

"Gamli maðurinn borðar kalda súpu með mjög góðri lyst"

gamli	lkenvf
maðurinn	nkeng
borðar	sfg3en_sfg2en
kalda	lveosf_lkfosf_lkeovf_lkeþvf_lkeevf_lvenvf_ lhenvf_lheovf_lheþvf_lheevf
súpu	nveo_nveþ_nvee
með	aþ_aa
mjög	aa
góðri	lveþsf
lyst	nven_nveo_nveþ

Dæmi um mörkun íslensku - Einræðing

"Gamli maðurinn borðar kalda súpu með mjög góðri lyst"

gamli	lkenvf
maðurinn	nkeng
borðar	sfg3en
kalda	lveosf
súpu	nveo
með	aþ
mjög	aa
góðri	lveþsf
lyst	nveþ

Mörkun

Grunnmörkun

- Í rannsókn fyrir ensku og frönsku: 50-60% tóka hafa eitt mark, 15-25% hafa aðeins tvö mörk.
- Með því að marka alltaf með algengasta markinu næst meiri en 75% nákvæmni.
- Kallað *grunnmörkun* (e. baseline tagging).
- Charniak (1993) hefur náð meira en 90% nákvæmni með grunnmörkun fyrir ensku.
- Athugið að undirliggjandi markamengi hefur mikil áhrif.

Grunnmörkun

Í íslenskri orðtíðnibók:

- Ómargræðar orðmyndir: 84,16%
- Margræðar orðmyndir: 15,84%
- Margræðar orðmyndir með 2 mörk: 11,07%
- Margræðar orðmyndir með 3 mörk: 2,96%
- Margræðar orðmyndir með 4 mörk: 0,97%

Hvaða orð eru margræð?

- Yfirleitt algengu orðin, smáorðin.

Grunnmörkun

Í íslenskri orðtíðnibók:

- Ómargræðar orðmyndir: 84,16%
- Margræðar orðmyndir: 15,84%
- Margræðar orðmyndir með 2 mörk: 11,07%
- Margræðar orðmyndir með 3 mörk: 2,96%
- Margræðar orðmyndir með 4 mörk: 0,97%

Hvaða orð eru margræð?

- Yfirleitt algengu orðin, smáorðin.

Grunnmörkun

Algeng orð og mörk þeirra í Orðtíðnibókinni

```
33181 .  .
22176 og    c
22083 ,     ,
21011 að    cn_c_aþ_aa
15319 í     aþ_ao_aa
12450 á     aþ_ao_sfg1en_sfg3en_aa_nven_nveo_nveþ_au
8040  hann  fpken_fpkeo
7905  var   sfg3eþ_sfg1eþ_lkensf
7676  sem   ct_c_aa_sfg1en
6357  er    sfg3en_sfg1en_ct_c
```

Outline

1 Mörkun

2 Nákvæmni í mörkun

3 Tegundir markara

4 Málfræðilegir reglumarkarar

5 Markari sem lærir reglur

Mælikvarðar á nákvæmni

Full einræðing

$$\text{hittni (e. accuracy)} = \frac{\#\text{ rétt markaðra tóka}}{\text{heildarfjöldi tóka}} \tag{1}$$

Ekki full einræðing

$$\text{nákvæmni (e. precision)} = \frac{\#\text{ réttra marka í úttaki markara}}{\text{heildarfjöldi marka í úttaki markara}} \tag{2}$$

$$\text{griphlutfall (e. recall)} = \frac{\#\text{ réttra marka í úttaki markara}}{\text{heildarfjöldi réttra marka}} \tag{3}$$

$$\text{margræðnihlutfall} = \frac{\text{heildarfjöldi marka í úttaki markara}}{\text{heildarfjöldi tóka}} \tag{4}$$

Mælikvarðar á nákvæmni

Full einræðing

$$\text{hittni (e. accuracy)} = \frac{\#\text{ rétt markaðra tóka}}{\text{heildarfjöldi tóka}} \tag{1}$$

Ekki full einræðing

$$\text{nákvæmni (e. precision)} = \frac{\#\text{ réttra marka í úttaki markara}}{\text{heildarfjöldi marka í úttaki markara}} \tag{2}$$

$$\text{griphlutfall (e. recall)} = \frac{\#\text{ réttra marka í úttaki markara}}{\text{heildarfjöldi réttra marka}} \tag{3}$$

$$\text{margræðnihlutfall} = \frac{\text{heildarfjöldi marka í úttaki markara}}{\text{heildarfjöldi tóka}} \tag{4}$$

Mælikvarðar á nákvæmni

Dæmi: 100 tókar

- Markari beitir fullri einræðingu og markar 95 tóka rétt. $\Rightarrow$ *hittni* $= 95/100 = 95\%$
- Markari beitir ekki fullri einræðingu og skilar af sér 105 mörkum; 95 þeirra eru rétt.
  - $\Rightarrow$ *nákvæmni* $= 95/105 = 90,5\%$
  - $\Rightarrow$ *griphlutfall* $= 95/100 = 95,0\%$
  - $\Rightarrow$ *margræðnihlutfall* $= 105/100 = 1,05$

  Athugið að þegar um full einræðingu er að ræða þá er *hittni*=*nákvæmni*=*griphlutfall* og *margræðnihlutfall*=1,0

Nákvæmni í mörkun

Hvað getur haft áhrif á nákvæmni?

- Tegund markara – gæði mállíkans.
- Stærð markamengis.
- Hlutfall óþekktra orða.
  - Möguleg mörk óþekktra orða eru ekki þekkt!
  - Giskara (e. unknown word guesser) þarf til.
- Stærð þjálfunarmálheildar.
- Tegund prófunarmálheildar.

Nákvæmni í mörkun

- Enska:
  - 96,7% (Brants, 2000)
  - 2,9% hlutfall óþekktra orða.
  - Þjálfunarmálheild: 1.000.000 orð.
  - Markamengi: 45 mörk (Penn Tree Bank).
- Sænska:
  - 93,6% (Megyesi, 2002)
  - 15% hlutfall óþekktra orða.
  - Þjálfunarmálheild: 100.000 orð.
  - Markamengi: 139 mörk.
- Íslenska:
  - 91,5% (Loftsson, 2006)
  - 6,8% hlutfall óþekktra orða.
  - Þróunarmálheild: 59.000 orð.
  - Markamengi: 660 mörk.

Outline

1 Mörkun

2 Nákvæmni í mörkun

3 Tegundir markara

4 Málfræðilegir reglumarkarar

5 Markari sem lærir reglur

Tegundir markara

Reglur vs. tölfræði

- Reglur skoða nánasta umhverfi orðs og eyða eða breyta tilteknu marki.
- Reglur geta verið handskrifaðar eða “lærðar” á vélrænan hátt úr markaðri málheild.
- Tölfræðiaðferðir eru notaðar til að úthluta orðum í setningu líklegustu markarununa.
- Tölfræðiaðferðir nota tíðniupplýsingar (t.d. um n-stæður) sem eru “lærðar” á vélrænan hátt úr markaðri málheild.

Tegundir markara

Málfræðilegir reglumarkarar (e. linguistic rule-based taggers)

- Byggja á handgerðum málfræðilegum reglum.
- Eingöngu hægt að nota þá til að marka tiltekið tungumál með tilteknu markamengi.

Gagnamarkarar (e. data-driven taggers)

- Óháðir tungumáli og markamengjum.
- Nota fyrirfram markaðar málheildir til að safna upplýsingum á vélrænan hátt sem síðar eru notaðar við einræðingu á nýjum texta.
- Upplýsingarnar geta t.d. verið í formi tölfræði eða reglna.

Tegundir markara

Málfræðilegir reglumarkarar (e. linguistic rule-based taggers)

- Byggja á handgerðum málfræðilegum reglum.
- Eingöngu hægt að nota þá til að marka tiltekið tungumál með tilteknu markamengi.

Gagnamarkarar (e. data-driven taggers)

- Óháðir tungumáli og markamengjum.
- Nota fyrirfram markaðar málheildir til að safna upplýsingum á vélrænan hátt sem síðar eru notaðar við einræðingu á nýjum texta.
- Upplýsingarnar geta t.d. verið í formi tölfræði eða reglna.

Outline

1 Mörkun

2 Nákvæmni í mörkun

3 Tegundir markara

4 Málfræðilegir reglumarkarar

5 Markari sem lærir reglur

Málfræðilegir reglumarkarar

Dæmigerð virkni

1. Sérhverju orði úthlutað öllum mögulegum greiningarstrengjum.
   - Flett upp í orðasafni og/eða orðhlutafræðilegur greinir eða giskari notaður.
2. Einræðing með reglum
   - Óviðeigandi mörkum eytt m.t.t. samhengis (e. reductionist approach)

Málfræðilegir reglumarkarar

Dæmigerð virkni

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.

Málfræðilegir reglumarkarar

Constraint Grammar Framework (Fred Karlsson 1990)

- Orðhlutafræðilegur greinir (sem byggir á tveggja laga kerfi) skilar öllum mögulegum greiningum fyrir hvert orð.
- Reglur (e. constraints) skrifaðar sem eyða mörkum m.t.t. samhengis.
- Reglur oft í þúsundum, t.d. EngCG-2 með 3,600 reglur.
- Tímafrekt í þróun en nákvæmni há. Framkvæmir þó ekki fulla einræðingu fyrir öll orð.
- Samuelsson and Voutilainen (1997):
  - Griphlutfall: 99,6%
  - Margræðnihlutfall: 1,02.

Málfræðilegir reglumarkarar

IceTagger - Hrafn Loftsson

- Giskari fyrir óþekkt orð: *IceMorphy*
- Staðbundnar (e. local) reglur
  - Um 175 reglur.
  - Fjarlægja tiltekið mark í tilteknu umhverfi.
  - Staðbundna umhverfið er 5 orð.
- Víðværar (e. global) reglur
  - Leitaraðferðir (e. heuristics)
  - Giska á setningafræðilegt hlutverk orða (frumlag, sögn, andlag).
  - Merkja forsetningaliði.
  - Nota ofangreint til að þvinga fram beygingarlegt samræmi milli orða þar sem við á.

Dæmi um virkni víðværu reglnanna

ég/fp1en fór/sfg3eþ_sfg1eþ svartar/lvfosf_lvfnsf
götur/nvfo_nvfn í/aþ_ao vesturátt/nveo_nveþ

- "í vesturátt" merkt sem forsetningaliður
- "fór" merkt sem aðalsögn
- "ég" merkt sem frumlag og 3. persónu markið fjarlægt úr "fór"
- "svartar" og "götur" merkt sem andlag og nefnifallið fjarlægt
  - "fór" krefst andlags í þolfalli
- forsetningarmarkið *aþ* markið fjarlægt úr "í" því "fór-í" stýrir þolfalli
- *nveþ* markið fjarlægt úr "vesturátt"

ég/fp1en fór/sfg1eþ svartar/lvfosf götur/nvfo í/ao vesturátt/nveo

IceTagger

Full einræðing

- Líklegasta markið valið ef orð er ennþá margrætt eftir beitingu staðværra og víðværra reglna.
- *IceTagger* er þá sambland af málfræðilegum reglumarkara og grunnmarkara.

Prófun

- `http://nlp.ru.is` og veljið *IceNLP*.

IceTagger

Full einræðing

- Líklegasta markið valið ef orð er ennþá margrætt eftir beitingu staðværra og víðværra reglna.
- *IceTagger* er þá sambland af málfræðilegum reglumarkara og grunnmarkara.

Prófun

- `http://nlp.ru.is` og veljið *IceNLP*.

Outline

1 Mörkun

2 Nákvæmni í mörkun

3 Tegundir markara

4 Málfræðilegir reglumarkarar

5 Markari sem lærir reglur

Markari sem lærir reglur

Brill's tagger (Eric Brill 1992)

- Gagnamarkari.
- Lærir reglur í þjálfun sem breyta einu marki í annað.
  - $\Rightarrow$ "Transformation-based learning"
- Orðasafn búið til úr þjálfunarmálheild.
  - Markið með hæstu tíðni (líklegasta markið) fyrir sérhvert orð merkt sérstaklega.

Brill's tagger

Virkni

- Úthlutar fyrst sérhverju orði líklegast markinu (grunnmörkun).
- Beitir síðan lista af reglum (umbreytingum; e. transformations) til að breyta mörkuninni.
- Reglum er beitt í ákveðinni röð og sérhverri umbreytingu er beitt á textann frá vinstri til hægri.
- Dæmi fyrir ensku:
  - "The can rusted"
  - Með líklegasta marki: The/**art** can/**modal** rusted/**verb**.
  - Regla: *Change the tag from modal to noun if the previous word is an article.*
  - Útkoma: The/**art** can/**noun** rusted/**verb**.

Brill's tagger

Hvernig lærast reglurnar?

- Reglur byggja á sniðmátum (e. templates).
- Sniðmátin takmarka þær reglur sem geta orðið til.
- Dæmi um sniðmát:

```
alter(A, B, prevtag(C))        Change A to B if preceding tag is C.
alter(A, B, nextbigram(C,D))   Change A to B if next bigram tag is C D.
```

- Fyrir ensku notaði Brill 11 sniðmát sem leiddu af sér um 500 reglur, nægjanlegar til að ná um 97% hittni.

Brill's tagger - Þjálfunaralgrímið

St.	Operation	Input	Output
1.	Base tagging	Corpus	Corpus(1)
2.	Compare POS of each word in *Gold standard* and Corpus(i)	*Gold standard* Corpus(i)	List of errors
3.	For each error, instantiate the rule templates to correct the error	List of errors	List of tentative rules
4.	For each rule, compute on Corpus(i) # of good transf. - # of bad transf.	Corpus(i) Tentative rules	Scored tentative rules
5.	Select the rule that has the greatest error reduction and append it to the ordered list of tranformations	Tentative rules	Rule(i)
6.	Apply Rule(i) to Corpus(i)	Corpus(i) Rule(i)	Corpus(i+1)
7.	If number of errors $< \delta$ exit else go to step 2		

Brill's tagger

Óþekkt orð

- Markar fyrst óþekkt orð sem sérnafn ef orðið byrjar á stórum staf.
- Markar fyrst öll önnur óþekkt orð sem nafnorð.
- Beitir síðan sérstökum sniðmátum (sjá bls. 155) til að búa til reglur sem breyta marki á óþekktu orði úr X í Y.