Comparing a Linguistic and a Stochastic Tagger.

Samanburður á tölfræðimarkara og reglumarkara.

Sagan

- Grein frá 1997
- Skv. greininni notuðu menn almennt tölfræði markara sem náðu almennt 95-97% nákvæmni.
- Vísa í reglumarkara og byggja sína vinnu á eldri reglumörkurum og fyrri reglumarkara EngCG og kynna EngCG-2 markarann

EngCG-2

- 3ja þrepa greining
  - Tilreiðing
  - Orðhlutafræðileg greining.
  - Einræðing(þó ekki að fullu)
- 180 margræðnimyndandi greiningar
- Hvert orð fær að meðaltali 1,7 -2,2 mismunandi greiningar
- Ekki er um að ræða fulla úrlausn margræðni.
- Eldri útgáfur með allt að 99,7% orða rétt greind

Gallar EngCG-2

- Ekki full úrlausn margræðni.
- Mikið hefur verið dregið í efa að um réttar niðurstöður sé að ræða m.a. Vegna þess að því hefur verið haldið fram að málfræðinga greini á um greiningu 3% orða og því tilgangslaust að koma með greiningu með nákvæmni yfir 97%.
- Því hefur einnig verið haldið fram að vegna þess að EngCG mörkin eru í eðli sínu ekki sértæk og því ætti að vera auðvelt að gera tölfræðimarkara sem að næði sambærilegum niðurstöðum.

En…

- Fyrri grein annars höfundar sýnir að hægt sé að ná næstum 100% samræmi meðal málfræðinga a.m.k. með EngCG markamenginu.
- Sýnt fram á í greininni að markamengið er álíka flókið fyrir tölfræðimarkara og önnur markamengi og að EngCG leysir mun betur úr margræðni en tölfræðimarkarinn.

Orðheildir

- Tölfræðimarkarinn var þjálfaður á 357 þúsund orðum úr Brown Corpus sem var fyrst greindur með EngCG markaranum og loks að fullu leyst úr margræðni og villum af sérfræðingi þar sem þörf var á.
- Prófunarmálheildin samanstóð af 55 þúsund orðum af fræðitexta. Sem var greindur en ekki leyst úr margræðni.

Orðheildir

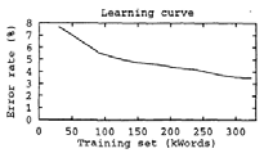
- Tveir sérfræðingar leystu svo úr margræðni og báru svo saman niðurstöður sem voru 99,3% eins og var í nær öllum tilfellum um að ræða handvömm frekar en skoðanaágreining í einungis 21 tilfelli var um að ræða ágreining um áherslur.
- Á þessum grundvelli var síðan útbúin ein rétt málheild.

Tölfræðimarkarinn

- Notar 3 stæður eins og oft hefur verið lýst í námskeiðinu.
- Reiknar 4 stærðir fyrir hvert orð og notar þær til að velja mörk.
- Útbúið var sér markamengi með 80 orðamörkum og 17 greinarmerkjum.

Prófanir tölfræðimarkara

- Tölfræðimarkarinn við þjálfun jók nákvæmni upp að 322 þúsund orðum en 35 þúsund orð voru tekin til hliðar.



Prófanir tölfræðimarkara

- Eftir það náði hann 3,51% nákvæmni og áttu áður óséð orð 1,08% af henni.
- Markarinn var þá þjálfaður á allri heildinni og prófaður á prófunarmálheildinni.
- Þá náði hann best 4,68% við fulla úrlausn margræðni.
- Þar eru um 2% vegna óséðra orða.

Samanburður á niðurstöðum

- Hlutfall milli villa markarana fer frá 8,6 við 1,026 mörk per orð í 28 við 1,070 mörk per orð.
- Hægt er að ná fækka villum tölfræðimarkarans niður í 0,15 % villur en þá eru líka yfir 15 mörk per orð.

Samanburður á niðurstöðum

Ambiguity (Tags/word)	Error rate (%) Statistical Tagger (δ)	(γ)	EngCG
1.000	4.72	4.68	
1.012		4.20	
1.025		3.75	
1.026		(3.72)	0.43
1.035		(3.48)	0.29
1.038		3.40	
1.048		(3.20)	0.15
1.051		3.14	
1.059		(2.99)	0.12
1.065		2.87	
1.070		(2.80)	0.10
1.078		2.69	
1.093		2.55	

Gagnrýni

- Um er að ræða of einfalt markamengi
- Nákvæmnin kemur til vegna margræðni.
- Notkun mannvera dregur úr heilindum prófananna þar sem þeir gætu undirbúið prófunarheildina fyrir markarann.

Svör

- Notaður var fullkominn tölfræðimarkari á sem getur skipt út margræðni fyrir villur.
- Notuð var orðheild sem var einrædd af tveimur sérfræðingum án þess að þeir hefðu aðgang að niðurstöðum markarana.

Niðurstöður

- Mikill munur var á gæðum niðurstaðna markarana við sömu margræðni.
- Niðurstöðurnar er ekki hægt að skýra með markamenginu, útskiptingu margræðni fyrir villur eða undirbúningi sérfræðingana.
- Heldur séu yfirburðir EngCG kerfisins einfaldlega skýring mismunar niðurstaðanna.

Spurningar

- ?

Minnkaða EngCG markamengið

Dæmi um mörkun

```
walk
   walk <SV> <SVO> V SUBJUNCTIVE VFIN
   walk <SV> <SVO> V IMP VFIN
   walk <SV> <SVO> V INF
   walk <SV> <SVO> V PRES -SG3 VFIN
   walk N NOM SG


walk V-SUBJUNCTIVE V-IMP V-INF
     V-PRES-BASE N-NOM-SG
```