Vélræn þáttun fornískenslensku: Samanburður aðferða

Lykilorð: þáttun, trjábanki, fornískenslenska, setningafræði

Efnisyfirlit

1 Leiðbeiningar

Þær skrár sem vísað er til hér eru aðgengilegar hér:
https://github.com/antonkarl/icecorpus/tree/master/parsald/comparison

1.1 Um tölfræðiþáttarann

Berkeley-þáttarinn var notaður til að búa til tól sem þáttar íslenskan texta. Þessi íslenski þáttari byggir að miklu leyti á verkefni sem var unnið árið 2014, en þar var einmitt búinn til þáttari sem var þjálfaður á IcePaHC. Þar og hér var fylgt hugmyndum Petrovs o.fl. (2012) en þar með var markamengin umtalsvert einfaldað. Að auki var liðgerð (heiti liða) einfölduð.

1.2 Grammar búið til

Til þess að búa til grammar þarf að hafa nýjustu útgáfuna af BerkeleyParser.jar (þegar þetta er skrifað er það BerkeleyParser1-7.jar). Því næst er farið í skipanaglugga (command-glugga). Hér er training.psd skjal með þeim trjám sem á að æfa á. ice.gr er svo grammar-fællinn sem verður búinn til.

(1) java -cp BerkeleyParser-1.7.jar edu.berkeley.nlp.PCFGLA.GrammarTrainer -path simpleicepahc7selectedtextswithouttextitexti.psd -out ice.gr -treebank SINGLEFILE

Þegar textarnir voru valdir var öðrum hverjum texta hent út. Textarnir sem ice.gr var æft á eru: 1150.homiliubok, 1210.thorlakur, 1250.thetubrot, 1270.gragas, 1300.alexander, 1325.arni, 1350.finnbogi, 1400.gunnar, 1400.viglundur, 1450.ectorssaga, 1450.vilhjalmur, 1480.jarlmann, 1525.georgius, 1540.ntjohn, 1611.okur, 1630.gerhard, 1659.pislarsaga, 1675.armann, 1675.modars, 1720.vidalin, 1745.klim, 1791.jonsteingrims, 1835.jonasedli, 1859.hugvekjur, 1882.torfhildur, 1888.grimur, 1902.fossar, 1908.ofurefli, 1985.margsaga, 2008.mamma

Textarnir sem síðan eru notaðir til að athuga hve góður þáttarinn ice.gr er eru: 1150.firstgrammar, 1210.jartein, 1250.sturlunga, 1260.jomsvikingar, 1275.morkin, 1310.grettir, 1350.bandamennM, 1350.marta, 1400.gunnar2, 1450.bandamenn, 1450.judit, 1475.aevintyri, 1525.erasmus, 1540.ntacts, 1593.eintal, 1628.olafuregils, 1650.illugi, 1661.indiafari, 1675.magnus, 1680.skalholt, 1725.biskupasogur, 1790.fimmbraedra, 1830.hellismenn, 1850.piltur, 1861.orrusta, 1883.voggur, 1888.vordraumur, 1907.leysing, 1920.arin, 1985.sagan, 2008.ofsi

1.3 Texti þáttaður

Byrjað er á því að breyta trjábankanum (ath. hér er átt við útgefinn trjábanka, IcePaHC 0.9 en ekki nýjustu gerð hans á github), sem inniheldur þáttaða texta, í óþáttaðar textaskrár, í þeirri röð sem hér er sýnt:

(2) a. ./psd-to-pos fæll.psd [gefur .pos fæl]
b. python3 rm-lemmata.py fæll.pos > fæll.changed.psd
c. ./psd-to-pos-post-rmlemmata fæll.changed.psd [gefur .pos fæl]
d. python3 rm-lemmata-posteverything.py fæll.changed.pos > fæll.changed.tok

Textaskrárnar sem verða hér til eru geymdar í möppunni tokFiles á github.

Því næst eru textaskrárnar sem búnar eru til með þessum hætti þáttaðar með tölfræðiþáttaranum. Það er gert með eftirfarandi skipun (hér er fæll.changed.tok inntaksskráin og fæll.parsed úttakið, þ.e. skrá með textanum þáttuðum):

(3) java -jar BerkeleyParser-1.7.jar -gr ice.gr -inputFile fæll.changed.tok -outputFile fæll.parsed

.parsed-skrárnar er að finna í möppunni IceGrammarParsedFiles á github.

Næsta skref er að láta IceTagger marka textann. Hér þarf að beita „svindli“ vegna þess að markarinn lítur á svigamerki sem mörk sem þarf að marka rétt eins og heiti á liðum og orðflokkum (svo sem CP eða NOUN). Svindlið felst í því að breyta skilgreiningum í encodemarkup.py og decodemarkup.py.

(4) ./txt2ipsd.sh fæll.changed [hér á ekki að skrifa fæll.changed.parsed]

Þessi skrá tekur .tok-skrá og skilar henni markaðri. txt2ipsd.sh vinnur með encodemarkup.py og decodemarkup.py. Fyrst fær txt2ipsd.sh skrána encodemarkup.py til að breyta t.d. svigamerkinu “(” í orðið “danced” alls staðar. Því næst fær txt2ipsd IceTagger til að marka textann í skránni. Hér markar IceTagger öll orð, en líka t.d. svigamerki. Þess vegna er gott að breyta “(” í t.d. “danced” vegna þess að IceTagger markar “danced” sem “danced e” (erlent orð). Þegar markarinn hefur lokið sér af fær txt2ipsd.sh skrána decodemarkup.py til að breyta t.d. “danced e” alls staðar í “(”. Því næst fær txt2ipsd.sh skrána tag-word.py til að breyta röð orða og marka, þannig að mark fari næst á undan viðkomandi orði en ekki næst á eftir (eins og IceTagger gerir). Þegar því er lokið virkjar txt2ipsd.sh skrána decodemarkup2.py sem eyðir öllum bilum á eftir “(” og bilum á undan “)”. Með þessu mæti er skráin gerð læsileg fyrir CorpusSearch.

skráin txt2ipsd.sh gefur psd-skrár sem er að finna í möppunni ParsedComparisonFiles á github. Þetta er samanburðarmálheildin sem vísað er til hér fyrir neðan.

2 Samanburður IcePaHC og ice.gr

2.1 Bein og óbein andlög

Hér höfum við áhuga á því að athuga hversu áreiðanlegar niðurstöður nýi þáttarinn gefur okkur fyrir röð beinna og óbeinna andlaga. Í ómarkaðri orðaröð fer óbeint andlag á undan beinu andlagi en það er ekki alltaf svo.

Til að byrja með er rétt að athuga talningu á beinum og óbeinum andlögum. Í þeim skrám sem til skoðunar eru í IcePaHC eru bein andlög í aðal- og aukasetningum (leitarfyrirspurn: IP-MAT*|IP-SUB* idoms NP-OB1*) 17318 talsins og óbein andlög í aðal- og aukasetningum (leitarfyrirspurn: IP-MAT*|IP-SUB* idoms NP-OB2*) 2716. Í samanburðarmálheildinni eru bein andlög (leitarfyrirspurn: IP idoms NP-OB1) 16487 talsins og óbein andlög (leitarfyrirspurn: IP idoms NP-OB2) 2475 talsins.

Næsta skref er að athuga hve oft beint óbeint andlag sagnar fer á undan beinu andlagi, og öfugt, hve oft beint andlag sagnar fer á undan óbeinu andlagi hennar. Leitarfyrirspurnirnar eru sýndar hér fyrir neðan.

(5) IcePaHC: óbeint andlag á undan beinu

```
IP-MAT*|IP-SUB* idoms NP-OB1*
AND IP-MAT*|IP-SUB* idoms NP-OB2*
AND NP-OB2* precedes NP-OB1*
```

(6) IcePaHC: beint andlag á undan óbeinu

```
IP-MAT*|IP-SUB* idoms NP-OB1*
AND IP-MAT*|IP-SUB* idoms NP-OB2*
AND NP-OB1* precedes NP-OB2*
```

(7) Samanburðarmálheild: óbeint andlag á undan beinu

```
IP idoms NP-OB1
AND IP idoms NP-OB2
AND NP-OB2 precedes NP-OB1
```

(8) Samanburðarmálheild: beint andlag á undan óbeinu

```
IP idoms NP-OB1
AND IP idoms NP-OB2
AND NP-OB1 precedes NP-OB2
```

Í IcePaHC fer beint andlag á undan óbeinu 450 sinnum en óbeint andlag á undan beinu 1315 sinnum. Í samanburðarmálheildinni fer beint andlag á undan óbeinu 101 sinni en óbeint andlag 1149 sinnum á undan beinu andlagi.

2.2 Innri gerð nafnliða

Viðbúið er að næsta athugun komi nokkuð illa út vegna þess að hér reynir ekki einungis á þáttarann heldur einnig markarann (IceTagger). Leitað er að nafnliðum sem innihalda lýsingarorð í frumstigi og samnafn (nafnorð sem er ekki sérnafn) og innbyrðis röð þeirra athuguð. Ómörkuð röð í nútímamáli er lýsingarorð á undan nafnorði en auðvitað eru fjölmörg dæmi um hina röðina.

(9) IcePaHC/Samanburðarmálheild: nafnorð fer næst á undan lýsingarorði

```
NP* idoms N-*|NS-*
AND NP* idoms ADJ-*
AND N-*|NS-* iprecedes ADJ-*
```

(10) IcePaHC/Samanburðarmálheild: lýsingarorð fer næst á undan nafnorði (samnafni)

```
NP* idoms N-*|NS-*
AND NP* idoms ADJ-*
AND ADJ-* iprecedes N-*|NS-*
```

Í IcePaHC eru 5643 dæmi um að lýsingarorðið fari næst á undan nafnorðinu (ómarkaða röðin) en einungis 276 dæmi um að nafnorðið fari næst á undan lýsingarorðinu. Í samanburðarmálheildinni eru 5568 dæmi um að lýsingarorðið fari næst á undan nafnorðinu en 567 dæmi um að nafnorðið fari næst á undan lýsingarorðinu.

2.3 Frásagnarumröðun

Frásagnarumröðun hefur rannsökuð nokkuð í íslensku (), þar á meðal með hjálp IcePaHC (). Hér er um að ræða dæmi á borð við *Fór hann svo heim*, þar sem yfirleitt væri sagt *Hann fór svo heim*; í frásagnarumröðun hefst setning sem sagt á persónubeygðri sögn

(11) IcePaHC: frásagnarumröðun, aðalsetning hefst á sögn og fer næst á undan frumlagi

```
IP-MAT|IP-MAT-SPE idomsfirst finite_verb
AND IP-MAT|IP-MAT-SPE idoms NP-SBJ*
AND finite_verb iprecedes NP-SBJ*
AND NP-SBJ* idoms !\**
AND IP-MAT|IP-MAT-SPE isRoot
```

(12) IcePaHC: aðalsetning hefst á frumlagi og fer næst á undan sögn

```
IP-MAT|IP-MAT-SPE idomsfirst NP-SBJ*
AND IP-MAT|IP-MAT-SPE idoms finite_verb
AND NP-SBJ* iprecedes finite_verb
AND NP-SBJ* idoms !\**
AND IP-MAT|IP-MAT-SPE isRoot
```

(13) Samanburðarmálheild: frásagnarumröðun, aðalsetning hefst á sögn og fer næst á undan frumlagi

```
IP idomsfirst finite_verb
AND IP idoms NP-SBJ*
AND finite_verb iprecedes NP-SBJ*
AND NP-SBJ* idoms !\**
AND IP isRoot
```

(14) Samanburðarmálheild: aðalsetning hefst á frumlagi og fer næst á undan sögn

```
IP idomsfirst NP-SBJ*
AND IP idoms finite_verb
AND NP-SBJ* iprecedes finite_verb
AND NP-SBJ* idoms !\**
AND IP isRoot
```

Dæmi um frásagnarumröðun í IcePaHC sem finnast með þessu móti eru 2144 talsins (hér þarf setningin að hefjast á persónubeygðri sögn). Dæmi þar sem aðalsetning hefst á sýnilegu frumlagi og fer næst á undan aðalsögn eru 7950. Samanburðarmálheildin gefur 7935 dæmi um hina hefðbundnu röð frumlag-persónubeygð sögn í aðalsetningum en 2111 dæmi um frásagnarumröðun.[1] Dæmi um að sýnilegt frumlag fari á undan persónubeygðri sögn eru aftur á móti 7375 talsins (hér þarf setningin að hefjast á frumlagi).

3 Umræða og næstu skref

Niðurstöðurnar nokkuð góðar, sérstaklega hvað varðar frásagnarumröðun. Ef við vitum hvaða setningagerðir er hægt að kanna með þessu móti getur þetta gagnast í 1) verkefnum þar sem gögnin eru of viðamikil til að hægt sé að handþátta þau (big data, sbr. verkefnið Gríðarstór stafrænn textagrunnur), 2) verkefnum þar sem ekki er peningur/tími til að handþátta gögnin. Hér mætti nefna 19. aldar verkefnið. Til er safn bréfa frá 19. öld. Erfitt er að skoða setningafræði þeirra án þess að þátta hann. Hins vegar býr verkefnið ekki yfir nauðsynlegri kunnáttu til að þátta bréfin handvirkt. Með því að nota þá aðferð sem hér hefur verið kynnt má fá tiltölulega áreiðanlegar niðurstöður fyrir bréfasafnið fyrir tilteknar setningagerðir. Mikilvægt er þó að gera frekari tilraunir með hvaða setningagerðir virka vel og hverja illa.

Möguleg atriði til að athuga í framhaldsrannsóknum:

- Orðaröð — VO, OV
- Forsetningarstrand
- Kjarnafærsla í aukasetningum
- Ákveðnihamlan
- Ópersónulega háttarsagnagerðin

[1] Hér hefur finite_verb verið skilgreint fyrirfram sem: *MDD*|MDP*|*HVP*|*HVD*|*DOP*|*DOD*|*BEP*|*BED

Heimildaskrá

Petrov, Slav, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. *LREC*.