


Hlutaþáttun íslensks texta

Hrafn Loftsson og Eiríkur Rögnvaldsson

Rask-ráðstefna

27. janúar 2007

Setningafræðileg þáttun

- Greining setningarliða
  - nafnliður, sagnliður, forsetningarliður ...
- Greining setningafræðilegra hlutverka
  - frumlag, andlag, sagnfylling ...
- Mismunandi ítarleg þáttun:
  - full þáttun (full/deep parsing)
    - heildargreining – allir möguleikar sýndir
  - hlutaþáttun (partial/shallow parsing)
    - greining í einstaka liði og setningarhlutverk

Tegundir þáttunar

- Full þáttun (full parsing; deep parsing)
  - þar sem búið er til fullkomið **þáttunartré** (parse tree) fyrir sérhverja setningu
  - oft margir möguleikar
- Hlutaþáttun (partial parsing; shallow parsing)
  - þar sem setningar eru greindar í setningarhluta
  - án þess að krefjast þess að sérhver hluti passi inn í víðtæka þáttun (e. global parse)


Hlutaþáttun

- Í mörgum tilvikum er nægjanlegt að greina setningar í setningarhluta eða setningarliði
  - án þess að krefjast þess að liðirnir passi inn í víðtækt þáttunartré
- Þetta getur átt við á ýmsum sviðum
  - **upplýsingaútdrætti** (e. information extraction)
  - eða **textaútdrætti** (e. text summarization)
  - þar sem greining setningarliða er mikilvægari en full þáttun

Mismunandi þáttun setningar

- Full þáttun – mismunandi greiningar:
  - $[_S\ [_{NL}$ Margir$]\ [_{SL}$ hittu $[_{NL}$ Maríu $[_{FL}$ á $[_{NL}$ skrifstofunni$]]]]]$
    - eða:
  - $[_S\ [_{NL}$ Margir$]\ [_{SL}$ hittu $[_{NL}$ Maríu$]]\ [_{FL}$ á $[_{NL}$ skrifstofunni$]]]$
- Hlutaþáttun – ein greining:
  - $\{_{FRL}\ [_{NL}$ Margir$]\}\ [_{SL}$ hittu$]\ \{_{ANDL}\ [_{NL}$ Maríu$]\}\ [_{FL}$ á $[_{NL}$ skrifstofunni$]]$
- Setningarliðirnir ekki felldir saman í eitt tré

Kostir hlutaþáttunar

- Full þáttun
  - nákvæmari og sýnir alla möguleika, en:
  - frek á tíma og reiknigetu
  - viðkvæm fyrir villum í inntaki
- Hlutaþáttun
  - sýnir ekki formgerðina eins nákvæmlega, en:
  - skilar greiningu þrátt fyrir villur í inntaki
  - hentar því vel t.d. fyrir texta á netinu

Þáttunarskema

- [Gerð liða]
  - NP(s) - nafnliður
  - AP(s) - lýsingarorðsliður
  - AdvP - atviksliður
  - PP - forsetningarliður
  - CP - aðaltenging
  - SCP - aukatenging
  - VP(i/b/s/p/g) - sagnliður
  - InjP - upphrópun
  - MWEx - orðasamband
- {*Hlutverk}
  - *SUBJ - frumlag
  - *OBJ - andlag
  - *OBJAP - andlag m. lo.
  - *OBJNOM - nefnifallsandlag
  - *IOBJ - óbeint andlag
  - *COMP - sagnfylling
  - *QUAL - eignarfallseinkunn
  - *TIMEX - tímaaukafall
  - *X > - tengist so. á eftir
  - *X < - tengist so. á undan

Útfærsla

- „Incremental finite-state parser“
  - stigvaxandi þáttari byggður á endanlegum stöðuaðferðum
- Röð af stöðuferjöldum (finite-state transducers)
  - þar sem sérhvert stöðuferjald:
    - ber kennsl á tiltekið mynstur í inntaki
    - skrifar greiningarupplýsingar inn í inntakstextann
    - skilar breyttum texta út, tilbúnum til meðhöndlunar fyrir næsta stöðuferjald

Stöðuferjöldin

- Stöðuferjöldin skiptast í tvo flokka:
- Ferjöld sem greina setningarliði
  - atviksliði, lýsingarorðsliði, nafnliði, forsetningarliði, sagnliði, o.s.frv.
- Ferjöld sem greina setningafræðileg hlutverk
  - frumlög, andlög, sagnfyllingar, eignarfallseinkunnir

Setningarliðir

- Hönnunarforsendur:
  - Reynt að nýta beygingarleg einkenni sem minnst þegar setningarliðir eru greindir
    - orðflokkur og röð orða látin stýra greiningu
  - Hægt væri að nýta samræmi í kyni, tölu og falli til að stýra greiningu á nafnliðum
    - en þar með væri dregið úr gagnsemi setninga-greiningarinnar fyrir málfræðileiðréttingu

Röð ferjalda, 1

- Keyrð í tiltekinni röð – einfaldir liðir fyrst
  - 1. atviksliðir
    - var sfg3eþ **[AdvP** mjög aa **AdvP]** gott lhensf félagslíf nhen
  - 2. lýsingarorðsliðir
    - var sfg3eþ **[AP** [AdvP mjög aa AdvP] gott lhensf **AP]** félagslíf nhen

Röð ferjalda, 2

- 3. nafnliðir
  - var sfg3eþ **[NP** [AP [AdvP mjög aa AdvP] gott lhensf AP] félagslíf nhen **NP]**
- 4. sagnliðir
  - **[VPb** var sfg3eþ **VPb]** [NP [AP [AdvP mjög aa AdvP] gott lhensf AP] félagslíf nhen NP]
- 5. forsetningarliðir
  - **[PP** af aþ [NP þessum fakfþ [AP stöðugu lkfþvf AP] ósigrum nkfþ mínum fekfþ NP] **PP]**

Fleiri ferjöld

- Ýmis orðasambönd eru merkt sérstaklega
  - ef þau haga sér eins og eitt orð
    - [PP **[MWE_PP út aa um ao MWE_PP]** [NP gluggann nkeog NP] PP]
    - **[MWE_AdvP allt fohen í aþ einu lheþsf MWE_AdvP]**
    - [PP í aþ [NP **[MWE_AP neins fokee konar nkee MWE_AP]** samfloti nheþ NP] PP]

Setningafræðileg hlutverk

- Síðan koma nokkur stöðuferjöld sem merkja setningafræðileg hlutverk
  - þau nýta sér setningarliðamerkingar og fallamerkingar undanfarandi ferjalda
    - en ekki aðrar upplýsingar úr mörkuninni
- Aukafallsfrumlög eru merkt sérstaklega
  - gerður var sérstakur listi um sagnir sem taka aukafallsfrumlög

Eignarföll og frumlög

- Eitt stöðuferjald merkir eignarfallseinkunnir
  - [NPa [APa síðustu lveove AP] nóttina nveog NP]
    {*QUAL [NPg okkar fp1fe NP] *QUAL}
- Eitt stöðuferjald merkir frumlög
  - **{*SUBJ> [NPn ég fp1en NP] *SUBJ>}**
    [VP tók sfg1eþ VP] [NPa ákvörðun nveo NP]
  - [NPa hvað fsheo NP] [VP á sfg1en VP]
    {*SUBJ< [NPn ég fp1en] *SUBJ<}
    [VPi að cn segja sng VPi] ? ?

Andlög og sagnfyllingar

- Eitt ferjald merkir **andlög** og **sagnfyllingar**
  - nýtir sér liðamerkingar, frumlagsmerkingar og föll
    - {*SUBJ> [NPn ég fp1en NP] *SUBJ>} [VP veitti sfg1eþ VP] **{*IOBJ< [NPd því fpheþ NP] *IOBJ<}** **{*OBJ< [NPa athygli nveo NP] *OBJ<}**
    - {*SUBJ> [NPn ég fp1en NP] *SUBJ>}

      [VPb er sfg1en VPb]

      {*COMP< [APn viss lkensf AP] *COMP<}

Endanleg útkoma

{*SUBJ> [NP augnaráðið nheng NP] *SUBJ>}
[VP negldist sfm3eþ VP]
[PP við ao [NP [AP gráa lkeovf AP] jakkann nkeog NP] PP]
[SCP sem ct SCP]
{*SUBJ> [NP hann fpken NP] *SUBJ>}
[VPb var sfg3eþ VPb]
[VPi að cn klæða sng VPi]
{*OBJ< [NP sig fpkeo NP] *OBJ<}
[PP úr aþ PP]
[CP og c CP]
[VPi hengja sng VPi]
[PP [MWE_PP inn aa í ao MWE_PP] [NP skáp nkeo NP] PP]

Mat á frammistöðu þáttarans

- Búið var til prófunarsafn
  - 509 setningar úr grunni *Íslenskrar orðtíðnibókar*
  - valdar tilviljanakennt
- Þetta safn var greint í höndunum
  - í samræmi við þáttunarskemað
- Sú greining myndar *gold standard*
  - sem greining þáttarans er borin saman við

Setningarliðir

- Nákvæmni í greiningu setningarliða
  - bæði miðað við rétt mörk úr *Íslenskri orðtíðnibók* og mörk úr *IceTagger*

Phrase type	F-measure correct tags	F-measure *IceTagger*	Freq. in test data
AdvP	91.8%	85.1%	8.2%
AP	95.1%	86.3%	8.1%
APs	87.0%	68.6%	0.5%
NP	96.8%	93.0%	37.6%
NPs	80.4%	74.3%	1.5%
PP	96.7%	91.3%	13.0%
VPx	99.2%	93.8%	19.3%
CP	100.0%	99.6%	5.7%
SCP	99.6%	97.6%	3.4%
InjP	100.0%	96.3%	0.2%
MWE	96.9%	92.6%	2.5%
All	96.7%	91.9%	100.0%

Setningafræðileg hlutverk

- Nákvæmni í greiningu setninga-fræðilegra hlutverka
  - bæði miðað við rétt mörk úr *Íslenskri orðtíðnibók* og mörk úr *IceTagger*

Function type	F-measure correct tags	F-measure *IceTagger*	Freq. in test data
SUBJ	68.2%	47.6%	4.7%
SUBJ>	92.7%	89.4%	30.3%
SUBJ<	83.7%	75.1%	12.3%
OBJ	0.0%	0.0%	0.2%
OBJ>	43.5%	20.0%	0.8%
OBJ<	90.2%	78.2%	19.7%
OBJAP>	71.4%	57.2%	0.2%
OBJAP<	75.0%	46.2%	0.4%
OBJNOM<	30.8%	16.7%	0.6%
IOBJ<	73.3%	51.9%	0.9%
COMP	56.9%	40.0%	2.8%
COMP>	91.3%	91.3%	1.3%
COMP<	75.1%	70.0%	12.7%
QUAL	87.7%	77.9%	10.4%
TIMEX	74.7%	55.9%	2.7%
All	84.3%	75.3%	100.0%

Röng greining liða, 1

- Ranglega greindur atviksliður:
  - [PP um [NP það NP] PP] [VP vissi VP] [NP stelpan NP] **[AdvP ekki þá AdvP]**
    - hér er *ekki* setningaratviksorð en stendur ekki með tíðaratviksorðinu *þá*
- Ranglega greindur lýsingarorðsliður:
  - [CP og CP] [VP tóku VP] [NP **[AP [AdvP fram AdvP] eigin AP]** dósir NP]
    - hér er *fram* sagnarögn en stendur ekki með *eigin*

Röng greining liða, 2

- Ranglega greindur samsettur nafnliður:
  - [AP sterkur AP] [VPb var VPb] **[NPs [NP hann NP] [CP og CP] [NP íþróttamaður NP] NPs]** [AP ágætur AP]
    - hér standa *hann* og *íþróttamaður* í sama falli og aðaltenging á milli, og eru því greindir sem samsettur nafnliður
    - á hinn bóginn er lo. *ágætur* ekki greint sem hluti nafnliðar af því að það stendur á eftir no.

Röng greining hlutverka

- Ófullkomin greining frumlags
  - [VPb er VPb] [AdvP ekki AdvP] [VPi að koma VPi] {*SUBJ [NP matur NP] *SUBJ}?
    - hér truflar liður milli sagnar og frumlags greininguna
- Röng liðgreining > röng hlutverksgreining
  - {*OBJ< [NP [AP [AdvP fram AdvP] eigin AP] dósir NP] *OBJ<}
    - *fram* greint sem hluti andlags af því að það var greint sem hluti lýsingarorðsliðar

Niðurstöður

- Niðurstöðurnar eru góðar
  - 96,7% nákvæmni í greiningu setningarliða
  - 84,3% nákvæmni í greiningu hlutverka
- Árangurinn mætti bæta með því að
  - nýta beygingarlegar upplýsingar meira
  - byggja meira á ýmiss konar orðalistum
  - endurbæta stöðuferjöldin og fjölga þeim