T-(538|725)-MALV, Málvinnsla
Hlutaþáttun

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]Tölvunarfræðideild, Háskólinn í Reykjavík

Október 2007

Outline

Outline

Full þáttun (e. deep/full parsing)

Markmið

- Að framkvæma djúpgreiningu.
- Að búa til "fullkomið" þáttunartré (e. parse tree).
- Tilteknar málfræðikenningar (líkön) notuð, eins og:
  - CFG – **C**ontext **F**ree **G**rammar
  - PCFG – **P**robabilistic **C**ontext **F**ree **G**rammar (Collins 1996; Charniak 1997)
  - HPSG – **H**ead-Driven **P**hrase **S**tructure **G**rammar (Pollard and Sag 1994).
  - LFG – **L**exical **F**unctional **G**rammar (Kaplan 1989).
  - DG – **D**ependency **G**rammar – (Tesnière 1966)

Full þáttun

Vandamál

- Erfitt og tímafrekt að búa til þáttara með gott “coverage”.
- Stærð lausnamengis getur vaxið með veldishraða.
  - Því oft reynir þáttarinn að búa til allar mögulegar greiningar.
- Þáttarinn getur líka stundum hafnað réttri greiningu á hluta setningar.
  - Ef hluturinn passar ekki inn í hið víðtæka þáttunartré.

Outline

Hlutaþáttun (e. shallow/partial parsing)

Markmið

- Að greina í setningarhluta án þess að krefjast þess að sérhver hluti passi inn í víðtæka þáttun (e. global parse).
- "to recover syntactic information efficiently and reliably from unrestricted text, by sacrificing completeness and depth of analysis" (Abney 1996).

Hvenær hentugt?

- Þegar djúpgreining er ekki nauðsynleg.
  - T.d. í upplýsingaheimt (e. information retrieval) og upplýsingaútdrætti (e. information extraction).
- Þegar skilvirkni er mjög mikilvæg.
- Þegar gæðum inntaks er ábótavant.

Hlutaþáttun (e. shallow/partial parsing)

Markmið

- Að greina í setningarhluta án þess að krefjast þess að sérhver hluti passi inn í víðtæka þáttun (e. global parse).
- "to recover syntactic information efficiently and reliably from unrestricted text, by sacrificing completeness and depth of analysis" (Abney 1996).

Hvenær hentugt?

- Þegar djúpgreining er ekki nauðsynleg.
  - T.d. í upplýsingaheimt (e. information retrieval) og upplýsingaútdrætti (e. information extraction).
- Þegar skilvirkni er mjög mikilvæg.
- Þegar gæðum inntaks er ábótavant.

Full þáttun vs. hlutaþáttun

- *Margir kysstu Maríu á skrifstofunni*
  - (Höskuldur Þráinsson (1999). Íslensk setningafræði)
- **Full þáttun**:
  - [S [NL Margir] [SL kysstu [NL Maríu [FL á [NL skrifstofunni]]]]]
  - [S [NL Margir] [SL kysstu [NL Maríu]] [FL á [NL skrifstofunni]]]
- **Hlutaþáttun**:
  - [NL Margir] [SL kysstu] [NL Maríu] [FL á [NL skrifstofunni]]

Outline

Fleiryrtar segðir (e. multiword expressions (MWE))

Skilgreining

- Röð tveggja eða fleiri orða sem haga sér eins og einn liður eða eining, t.d.:
  - Nöfn: persónur, fyrirtæki, stofnanir
  - Tímasegðir: tími, dagsetningar
  - Númeríska segðir: tölur og upphæðir.
  - Röð orða sem hafa sama hlutverk samtengingar, atviksorðs, lýsingarorðs eða forsetningar.
- Víðtækara en orðastæða.

Dæmi um fleiryrtar segðir

MWE_AdvP = fleiryrt segð sem hefur hlutverk **samtengingar**
MWE_PP = fleiryrt segð sem hefur hlutverk **forsetningar**
MWE_CP = fleiryrt segð sem hefur hlutverk **samtengingar**
MWE_AP = fleiryrt segð sem hefur hlutverk **lýsingarorðs**

```
[CP en c CP] [MWE_AdvP einhvern fokeo veginn nkeog MWE_AdvP]
[VP tengdist sfm3eþ VP] [NP það fphen NP]

[PP [MWE_PP uppi aa á aþ MWE_PP] [NP bakkanum nkeþg NP] PP]

[MWE_CP án ae þess fphee að cn MWE_CP] [VPi hafa sng VPi]
[NP nokkuð foheo NP] [PP fyrir aþ [NP stafni nkeþ NP] PP]

[MWE_AdvP við aa og c við aa MWE_AdvP] [VP gægðist sfm3eþ VP]
[NP hún fpven NP] [PP [MWE_PP yfir aa í ao MWE_PP] [NP bókina nveog NP] PP]

[CP og c CP] [NP [MWE_AP hvers fohee kyns nhee MWE_AP] líkamshirðingar nvee NP]
```

Fleiryrtar segðir

- Segðirnar á undan er yfirleitt hægt að leysa með orðalistum.
- Öðru máli gegnir um nöfn, tímasegðir, númerískar segðir.
  - Stundum þó notaðir sérstakir orðalistar fyrir nöfn, sem kallast **gazetteers**.
- Einnig þó nauðsynlegt að geta borið kennsl á segðir sem ekki eru "harðkóðaðar" í einhverjum lista.
- $\Rightarrow$ Reglulegar segðir!

Outline

Klumpar (e. chunks)

- Klumpur = Hópur orða (e. group of words)
- Munurinn á klumpi og setningarliði er sá að setningarliður getur innifalið hreiðraða (e. nested) liði af sömu tegund.
  - Sjá t.d. glæru nr. 21 í fyrirlestrinum "Samhengisfrjáls mállýsing og Prolog"
- Samhengisfrjáls mállýsing er notuð til að leyfa endurkvæmni í setningarliðum.
- Í skilgreiningu á klumpi er því ekki notuð nein endurkvæmni.
- Endanlegar stöðuvélar (reglulegar segðir) eru því nægjanlegar til að lýsa klumpum.

Lýsing á íslenskum nafnaklumpi

- Einföldum málið og gerum ráð fyrir að nafnaklumpur geti einungis innifalið:
  - atviksorð, lýsingarorð, nafnorð
  - t.d. “saga”
  - eða “skemmtileg saga”
  - eða “mjög skemmtileg saga”

Í JFlex:

```
%% Stöðuvél sem ber kennsl á einfalda nafnaklumpa
%public
%class NounChunk
%standalone
%unicode

%{
    String Open="[NP " ;
    String Close="NP] ";
%}

WhiteSpace = [ \t\f]
WordChar = [^\r\n\t\f ]
Word = {WordChar}+
WordSpaces = {Word}{WhiteSpace}+

Gender = [kvhx] /* k=masculine, v=feminine, h=neuter, x=unspec */
Number = [ef]   /* e=singular, f=plural */
Case = [noþe]   /* n=nominative, o=accusative, þ=dative, e=genitive */
```

Í JFlex (framhald):

```
AdverbTag = aa[me]?
AdjectiveTag = l{Gender}{Number}{Case}[sv][fme]
NounTag = n{Gender}{Number}{Case}[g\-]?[mös]?

Adverb = {WordSpaces}{AdverbTag}{WhiteSpace}+
Adjective = {WordSpaces}{AdjectiveTag}{WhiteSpace}+
Noun = {WordSpaces}{NounTag}{WhiteSpace}+

NounChunk = {Adverb}?{Adjective}?{Noun}

%%
{NounChunk} { System.out.print(Open + yytext() + Close);}
.           { System.out.print(yytext());}
```

Outline

Lagskiptir hlutaþáttarar

(e. cascading/incremental partial/shallow/finite-state parsers)

- Byggðir á mörgum stöðuferjöldum (e. finite-state transducers).
- Sérhvert stöðuferjald hefur ákveðið hlutverk, t.d. að:
  - Merkja MWE
  - Merkja sagnliði
  - Merkja nafnliði
  - Merkja forsetningarliði
  - o.s.frv.
- Inntakið er tilreiddur og markaður texti.
- Úttak úr einu ferjaldi er inntak í það næsta.
- Endanlegt úttak er upphaflegi textinn með setningafræðilegum upplýsingum (t.d. klumpum).

Lagskiptir hlutaþáttarar

Til fyrir ýmis tungumál

- spænsku (Molina et al. 1999)
- sænsku (Megyesi and Rydin 1999)
- þýsku (Müller 2004)
- frönsku (Aït-Mokhtar and Chanod 1997)
- íslensku (Hrafn Loftsson og Eiríkur Rögnvaldsson 2007)

Hraðvirkir

- Runa af stöðuferjöldum.

Lagskiptir hlutaþáttarar

Til fyrir ýmis tungumál

- spænsku (Molina et al. 1999)
- sænsku (Megyesi and Rydin 1999)
- þýsku (Müller 2004)
- frönsku (Aït-Mokhtar and Chanod 1997)
- íslensku (Hrafn Loftsson og Eiríkur Rögnvaldsson 2007)

Hraðvirkir

- Runa af stöðuferjöldum.

Þáttunarskema

Tilgangur

- Leiðbeiningar um hvers konar setningarliði og setningarfræðileg hlutverk eigi að merkja.
- Meginreglur um hvernig merkja skuli liði og setningafræðileg hlutverk.
- Málfræðiskilgreiningarmálheild (e. grammar definition corpus, GDC).
  - Safn dæmigerðra setninga sem hafa verið greindar (handvirkt) m.t.t. meginreglna og leiðbeininga.
  - Hlutverk er að gefa (einræð) svör við spurningum um hvernig greina eigi setningar í málinu.
  - Einnig hægt að nota til þess að hjálpa til við þróun þáttarans því hann þarf a.m.k. að geta greint setningar málheildarinnar á réttan hátt.

Outline

IceParser

- Hrafn Loftsson og Eiríkur Rögnvaldsson 2007
- Byggir á þáttunarskema:
  `http://nlp.ru.is/pdf/shallowAnnotation.pdf`
- Lagskiptur hlutaþáttari: `http://nlp.ru.is`
- Merkir setningarliði og setningafræðileg hlutverk.
  - Phrase/constituent structure module; 14 stöðuferjöld.
  - Syntactic functions module; 8 stöðuferjöld.

Hvernig er nákvæmnin mæld?

	Réttir liðir í *gold standard*	Rangir liðir
Myndaðir liðir af þáttara	A	B
Ekki myndaðir af þáttara	C	D

- **Nákvæmni** (e. precision):
  $P = \frac{A}{A \cup B} = \frac{\#\text{ réttra liða í úttaki þáttara}}{\text{heildarfjöldi liða í úttaki þáttara}}$
- **Griphlutfall** (e. recall): $R = \frac{A}{A \cup C} = \frac{\#\text{ réttra liða í úttaki þáttara}}{\text{heildarfjöldi liða í } \textit{gold standard}}$
- **F-measure** $= \frac{2*P*R}{P+R}$ (e. harmonic mean)

IceParser: Evaluation

Experimental setup

- A *gold standard* was constructed:
  - About 500 sentences randomly selected from the POS tagged *IFD* corpus.
  - Manually annotated with constituent structure and syntactic functions using the annotation scheme.
- The *Evalb* (Sekine & Collins, 1997) bracket scoring program used for automatic evaluation.
- The parser evaluated using correct POS tags and tags generated by *IceTagger*.
  - POS tagging accuracy was 91.1% (unknown word ratio 7.8%).

Results for the various phrase types

Phrase type	F-measure using correct POS tags	F-measure using *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%

Results for the various syntactic functions

Function type	F-measure using correct POS tags	F-measure using *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%
...			
All	84.3%	75.3%	100.0%