Nimisõnafraaside tuvastamine

Süntaksianalüsaator

17. märts 2006

Kaili Müürisep

Ülevaade

- Mis on NP?
- Kuidas teda tuvastatakse?
- Milleks seda tehakse?

NP

- NP traditsioonilises mõttes: fraas, mille põhjaks on nimisõna; võib koosneda teistest nimisõnafraasidest.
- Maksimaalne nimisõnafraas: NP koos kõikide oma laienditega.
- Baasnimisõnafraas: mitterekursiivne NP, mis ei sisalda teisi NPsid

[The survival] of [spinoff Cray Computer Corp] as [a fledging] in [the supercomputer business] appears to depend heavily on [the creativity] – and [longevity] – of [its chairman] and [chief designer], [Seymour Cray].

Chunk

- Mittelõikuvad tekstiregioonid

  [I] saw [a tall man] in [the park].

- Mitterekursiivsed
- Ei hõlma kogu lauset
- Ei kirjelda moodustajate struktuuri:

  [a tall man in [the park]

  [a tall man] in [the park]

- Ei ületa moodustajate piire

BaasNFde leidmise töö käik

- Teksti tükeldamine
- Morfoloogiline analüüs
- Morfosüntaktiline ühestamine
- *Chunking*

Peamine meetod

- Süsteem märgendab fraasi esimese sõna märgendiga B (beginning)
- Fraasi järgmised sõnad märgendiga I (in)
- Fraasist välja jäävad sõnad märgendiga O (out)

- Märgendamiseks kasutatakse:
  - Käsitsi kirjutatud reegleid
  - Masinõppimist

Milleks?

- Infootsisüsteemid kasutavad baasnf-e indekseerimisel
- Masintõlkimisel tõlkeühikuks
- Terminite tuvastamine
- QA-süsteemid
- Süntaksianalüsaatori preprotsessor

Miks mitte sügavam analüüs?

- Madal analüüs on lihtsam
- Veakindlam
- Kiirem

Reeglipõhine lähenemine - REChunk

- Regulaaravaldiste kogu ehk töötlus lõplike automaatidega
- Reeglid, mis leiavad chunke:

  `<DT>?<JJ>*<NN.?>`

  the/DT little/JJ cat/NN sat/VBD on/IN the/DT mat/NN

  [the/DT little/jj cat/NN] sat/VBD on/IN [the/DT mat/NN]
- Reeglid, mis leiavad chinke:

  `(<VB.?>|<IN>)+`

  [sat/VBD on/IN]

REChunk - 2

- Positiivne lähenemine:

```
(<DT>?<JJ>*<NN.?>) -->{\1}
```

- Välistav lähenemine:
  - Kogu tekst sulgudesse:

```
(<.*>*) --> {\1}
```

  - Chinkide abil tükeldamine

```
((<VB.?>|<IN>)+ --> }\1{
```

{the/DT little/jj cat/NN} sat/VBD on/IN {the/DT mat/NN}

Reeglipõhine lähenemine - NPTool

- Käsitsi kirjutatud reeglis, sarnased morf. Ühestamise reeglitele
- Osa reegleid lisavad märgendeid, teised eemaldavad
- @V – verbid
- @NH – nimisõnafraaside põhjad
- @>N – determnatiivid ja eestäiendid
- @N< - prepositsionifraasid järeltäiendina
- @CC – koordinatsioon
- @AH - muu

NPTool

the/@>N inlet/@>N and/@CC exhaust/@>N manifolds/@NH are/@V mounted/@V on/@AH opposite/@>N sides/@NH of/@N< the/@>N cylinder/@NH head/@V|@NH

Regulaaravaldistega võeti välja pikimad ja lühimad fraasid ja lasti inimesel valida, kumba ta tahab

Tänane praktikum

- Kirjutame reegleid, mis
- a) lisavad @B, @I ja @O märgendid sõnadele (mapping rules)
- c) ühestavad fraasipiirimärgendeid

Märgendite lisamise reeglid

- Sektsioon algab võtmesõnaga MAPPINGS
- Kahte tüüpi reegleid:

  MAP (@LABEL) TARGET (siht) IF (kontekstitingimused);

  - Lisab märgendi ja „paneb sõna lukku“, enam sellele sõnale märgendeid lisada ei saa.

  ADD (@LABEL) TARGET (siht) IF (kontesktitingimused);

  - Lubab märgendhaaval lisada

Näited

- MAP (@NEG) TARGET („ei“ V);
- MAP (@IMV @SUBJ @OBJ @PRD @INFN> @INFN< @ADVL) TARGET (V inf);

- ADD (@P>) TARGET (S gen) OR (P gen) OR (N gen) (*1 (.gen) BARRIER (FinV));