Süntaksianalüsaator

Sissejuhatus

Analüüsi etapid

- Teksti normaliseerimine – kooditabeli valik, ebameeldivate sümbolite eemaldamine, hüperlinkide eemaldamine, tabelite teisendamine või eemaldamine, sgml-märgendus
- Lausepiiride määramine
- Lause tükeldamine sõnadeks
- Morfoloogiline analüüs
- Süntaktiline analüüs
  - Morfosüntaktiline ühestamine
  - Pindsüntaktiline analüüs: fraasipiirid ja ühetasandiline funktsionaalne märgendus
  - Sügavam süntaktiline analüüs

Kasutatud kirjandus

- Syntactic wordclass tagging. Ed H. Van Halteren. Kluwer 1999
- Eesti keele formaalne grammatika. T. Roosmaa jt. Tartu 2001.

Morfosüntaktiline märgendamine

- Märgendid on kirjeldavad sümbolid, mis lisatakse teksti sõnadele käsitsi või automaatselt. (POS, muu morf. info, süntaktilised märgendid)
- Sõnaliikide märgendus on kõige laiemalt levinud, kuulub pigem süntaksisse kui morfoloogia alla:
  - Sõnaliigi valik sõltub 1) süntaktilisest distributsioonist; 2) süntaktilistest funktsioonist, mida sõna kannab 3) (semantikast)
- Sõnal võib olla 1 märgend või märgendite kombinatsioon

Näide

- the/AT Fulton/NP County/NP Grand/NP Jury/NP said/VBD Friday/NR an/AT investigation/NN of/IN Atlanta's/NP$ recent/JJ primary/NN election/NN produced/VBD no/AT evidence/NN that/CS any/DTI irregularities/NNS took/VBD place/NN ./. (Brown Corpus)
- Ühel sõnaliigil võib olla palju märgendeid:
  NN, NNS, NN$, NN'$, NP, NPS, NR (LOB)
- Mitu märgendit ühel sõnal: N SG NOM
- Sõnad on mitmesed:
  round – prep, adv, n, v, adj

Automaatne märgendamine

- Tükeldamine (lauseteks ja sõnadeks)
- Märgendite lisamine leksikoni põhjal (morfoloogiline analüüs, tundmatute sõnade märgendid)
- Mitmesuste eemaldamine e ühestamine

Ajaloost

- Lingvistiline lähenemine – reeglid määrab lingvist-grammatik. Sageli puhast lingvistilist reeglit pole, kasutatakse heuristilisi.
- Statistilised meetodid (data driven) – keele mudel genereeritakse automaatselt kasutades suuri märgendatud tekstihulki. Vajab siiski lingvistlise teadmisi, et valida sobiv märgendite süsteem ja eelmärgendada treeningkorpus.
- 1950-1960 – käsitsi kirjutatud reeglid, väike leksikon, tundmatute sõnade mõistataja, ikka tundmatud said üldise märgendi. Märgend eemaldati lokaalse konteksti abil (1-2 sõna mõlemale poole)
  - Verbi märgend eemaldada, kui eelnev on artikkel

  Alles jäi mitmene märgendus, mis ühestati käsitsi

  30 märgendit, korrektsus 90% (Klein ja Simmons 1963)

Ajaloost

- **1960-1970** Brown Corpus (1,1 milj. sõna, American English, 15 zhanri)
  TAGGIT – 71 märgendit, 3,300 reeglit, 5-sõnaline aken: 77% ühene, vigu pole teada
- **Hilised 70ndad** LOBi märgendamine
  CLAWS (Constituent-Likelihood Automatic Word Tagging System) – puhtalt statistiline
  - tokenizer
  - 7000 sõna leksikonis & oletaja; haruldastel variantidel erimärgend; 139 märgendit
  - Idioomide tuvastamine – käsitsi kirjutatud reeglid
  - CHAINPROBS treenitud 200000 sõnal, bigram
  - 96-97% korrektsus, kui ühene väljund. 99% kui 14% mitmene

Ajaloost (1980ndad)

- Fidditch (89), Hindle'i algoritm

  Kui reegel eksib treeningkorpusel:

  - Reegli usaldusväärsus langeb ja ta langeb rakendamise järjekorras
  - Õpitakse uus spetsiifiline reegel konkreetse sõna mustri põhjal
    - Uus reegel tõuseb või langeb järjekorras korpuse edasise töötlemise käigus
    - Kui liiga palju vigu, kustutatakse

Fidditch

- Algselt 350 käsitsi kirjutatud reeglit
- Treeniti 90% Browni korpusel
- Genereeriti 35000 reeglit
- Reeglites kasutatakse nii märgendeid kui ka 300 sõnavormi
- Korrektsus 97%

Ajaloost (1990ndad)

- HMM – 95-97%, kiired, võimaldab kasutada märgendamata tekste
- Brill tagger – 95-97%
- Närvivõrgud (Schmid) 96%
- Case-based taggers
- Kitsenduste grammatika (Karlsson, Voutilainen)
  - Inglise keel: 99% korrektne, 4-5% mitmene
  - Türgi keel: 97-99%, 1-2%
  - Norra keel: 99%, 4-5%
  - Portugali keel: 99%, 0-1
  - Eesti keel: 98%, 10-14%
  - Baski keel: 97%, 25%

Kitsenduste grammatika

- Koosneb käsitsi koostatud reeglitest kujul:
- Tee operatsioon X märgendiga Y kontekstis Z
- Ühestaja rakendab kõiki reegleid ükshaaval lause igale sõnale, tehes lause analüüsimisel mitu tsüklit
- Reeglite järjekord grammatikas ei ole määratud, kui reegleid saab grupeerida usaldusväärsuse järgi
- Kõigile sõnadele jääb alles vähemalt 1 analüüs. Mida ei suudeta analüüsida, jäävad mitmeseks. Eelistatakse mitmesust vigadele.

KG operatsioonid

- Valik: SELECT märgend IF ...
- Eemaldamine: REMOVE märgend IF ...

REMOVE (N NOM SG) (-1C (AUXMOD)) (0 INF));

Kontekstid

- Lokaalsed kontekstid

  -2 DET

  NOT 1 VFIN

  0 UPPERCASE
- Piiramata kontekstid: *1 NOUN
- Ettevaatlik režiim: 1C Verb
- Barjäärireeglid:

  *-2C RELPRON BARRIER VFIN

  *-2C RELPRON BARRIER VFIN LINK -1C NOUN

Näitereeglid

- SELECT (V) (-1 („<ei>“));
- REMOVE (P pers sg nom) (*-1(V)) BARRIER (Com) LINK (0 (ps1));

Mõtlemise koht

- Kaassõnad
  - Vali nimisõna kääne tagasõna rektsiooni abil
  - Sõna on tagasõna, kui eelnev nimisõna on õiges käändes
- Nimisõnad
  - Vali eelneva omadussõna kääne, kui järgnev ei ole nimisõna
- Omadussõnad
  - Vali omastav kääne, kui järgneb nimisõna 4 viimases käändes