T-(538|725)-MALV, Málvinnsla
Málheildir og endanlegar stöðuvélar

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]Tölvunarfræðideild, Háskólinn í Reykjavík

Ágúst 2007

Outline

Outline

Málheild

- Málheild (e. corpus) er safn texta eða tals sem geymt er á tölvutæku formi.
- Innihald er oft sett saman eftir fyrirfram ákveðnum reglum.
- Kallað textasafn (e. text collection), frekar en málheild, ef um tilviljanakennt safn er að ræða.
- Mjög stórar málheildir (tugir milljóna orða) eru orðnar algengar í dag.

Tegundir málheilda

Efnisflokkar

- Sérstakir efnisflokkar, t.d. lög, vísindi, skáldsögur, dagblaðatexti, o.s.frv.
- Fjölbreyttir efnisflokkar:
  - Til að endurspegla málnotkun viðkomandi tungumáls.
  - "Balancing a corpus".
  - Dýrt að setja saman.

Skýringar

- Hrár texti án skýringa.
- Texti með skýringum/merkjum (e. annotation).

Tegundir málheilda

Efnisflokkar

- Sérstakir efnisflokkar, t.d. lög, vísindi, skáldsögur, dagblaðatexti, o.s.frv.
- Fjölbreyttir efnisflokkar:
  - Til að endurspegla málnotkun viðkomandi tungumáls.
  - "Balancing a corpus".
  - Dýrt að setja saman.

Skýringar

- Hrár texti án skýringa.
- Texti með skýringum/merkjum (e. annotation).

Málheildir með skýringum

Hvers konar skýringar?

- Sérhvert orð eða safn orða er merkt með málfræðilegu marki (e. tag).
- T.d. orðflokki, setningalið, merkingarflokki.
- Framkvæmt handvirkt og/eða með hjálp hugbúnaðar (e. semi-automatically).

Trjábanki (e. treebank)

- Málheild þar sem einhvers konar setningauppbygging er sýnd.
  - T.d. einstakir setningaliðir.
- Penn Treebank (University of Pennslyvania) er líklega þekktasti trjábankinn.

Málheildir með skýringum

Hvers konar skýringar?

- Sérhvert orð eða safn orða er merkt með málfræðilegu marki (e. tag).
- T.d. orðflokki, setningalið, merkingarflokki.
- Framkvæmt handvirkt og/eða með hjálp hugbúnaðar (e. semi-automatically).

Trjábanki (e. treebank)

- Málheild þar sem einhvers konar setningauppbygging er sýnd.
  - T.d. einstakir setningaliðir.
- Penn Treebank (University of Pennslyvania) er líklega þekktasti trjábankinn.

Dæmi um málheildir

Penn Treebank

- `http://www.cis.upenn.edu/~treebank/`
- Um 5 milljón orð.
- Markaður texti með forriti (markara).
- Textinn fenginn úr Wall Street Journal, 1989-1991.
  - `http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`
- Setningagreindur texti með forriti (þáttara).

Dæmi um mörk úr Penn Treebank

Markaður texti

- The/**DT** grand/**JJ** jury/**NN** commented/**VBD** on/**IN** a/**DT** number/**NN** of/**IN** other/**JJ** topics/**NNS** ./.
- DT=Determiner(Article)=Ákvæðisorð (Greinir)
- JJ=Adjective=Lýsingarorð
- NN=Noun=Nafnorð
- VBD=Verb, past tense=Sögn í þátíð
- IN=Preposition or subordinating conjunction=Forsetning eða aukatenging
- NNS=Noun, plural=Nafnorð í fleirtölu

Dæmi úr Penn Treebank

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

```
((S
    (NP-SBJ                                  // NP-SBJ=Noun phrase subject=Frumlag
        (NP (NNP Pierre) (NNP Vinken))       // NP=Noun phrase=Nafnliður
        (, ,)
        (ADJP
            (NP (CD 61) (NNS years))         // NNS=Noun, plural=Nafnorð, fleirtala
            (JJ old))                        // JJ=Adjective=Lýsingarorð
        (, ,))
    (VP (MD will)
        (VP (VB join)                        // VB=Verb, base form=Sögn í nafnhætti
        (NP (DT the (NN board))
            (PP-CLR (IN as)
                (NP (DT a) (JJ nonexecutive) (NN director) ))
            (NP-TMP (NNP Nov.) (CD 29) ))))
    (. .)
))
```

Dæmi um málheildir

British National Corpus (BNC)

- `http://www.natcorp.ox.ac.uk/`
- 100 milljón orð.
- “Balanced corpus”.
- Orðflokksmarkaður texti með forriti (markara).
  - `http://www.natcorp.ox.ac.uk/docs/bnc2postag_manual.htm`

Dæmi um málheildir

Íslensk orðtíðnibók

- $\approx$ 590 þús. orð (tókar)
- "Balanced corpus":
  - Íslensk skáldverk, þýdd skáldverk, ævisögur og endurminningar, fræðslutextar, barna- og unglingabækur.
- `http://www.lexis.hi.is/utg_ordabaekur.html`
- Markaður texti með forriti (eftir Stefán Briem) og leiðréttur handvirkt.

Dæmi úr Íslenskri orðtíðnibók

```
ég fp1en              // orð mark
stökk sfg1eþ          // Sjá skýringar á mörkum
á aa                  // í skjali undir ‘‘Annað efni’’
eftir aþ              // á vefsíðu námskeiðsins
stræto nkeþ
og c
veifaði sfg1eþ
, ,
vagnstjórinn nkeng
sá sfg3eþ
mig fp1eo
og c
stoppaði sfg3eþ
. .
```

Dæmi um málheildir

Stór íslensk málheild

- Er verið að setja saman hjá Orðabók Háskólans (Stofnun Árna Magnússonar í íslenskum fræðum).
- 900 textabútar, 25 milljón orð.
- `http://www.lexis.hi.is/malheild.htm`

Gagnsemi málheilda

- Gerða orðalista og orðabóka.
- Rannsóknir í málvísindum.
- Forsenda fyrir þróun ýmiss konar máltæknitóla, t.d.:
  - Markara
  - Þáttara (setningagreina)
  - Vélrænna þýðinga (sem nýta sér oft hliðstæðar (e. parallel) málheildir).

Outline

Endanleg stöðuvél (e. finite-state automaton)

- Vél sem samþykkir (e. accepts) eða hafnar (e. rejects) straumi af stöfum (þ.e. strengjum).
- Oft kallað “recognizer”.
- Getur einnig verið notuð sem “generator”, þ.e. vél sem býr til strengi.
- Mjög skilvirk m.t.t. hraða og minnisnotkunar.
- Hentug til notkunar í textaleit.
- Dæmi: Sjá Fig. 2.1 bls. 28 í kennslubók.

Endanleg stöðuvél

Stærðfræðileg skilgreining

Endanleg stöðuvél samanstendur af fimm hlutum $(Q, \Sigma, q_0, F, \delta)$:

1. Q er mengi af endanlegum stöðum, q_0, q_1 ... q_n.
2. Σ er endanlegt mengi af inntaksstöfum.
3. q_0 er upphafsstaða, $q_0 \in Q$.
4. F er mengi lokastaða, $F \in Q$.
5. δ er breytingafall (e. transition function) $Q \times \Sigma \rightarrow Q$. $\delta(q, i)$ skilar þeirri stöðu sem vélin fer í úr stöðu q miðað við inntak i.

Dæmi: $Q = \{q_0, q_1, q_2\}$, $\Sigma = \{a, b, c\}$, $F = \{q_2\}$, $\delta = \{\delta(q_0, a) = q_1, \delta(q_1, b) = q_1, \delta(q_1, c) = q_2\}$

Endanleg stöðuvél

Stærðfræðileg skilgreining

Endanleg stöðuvél samanstendur af fimm hlutum $(Q, \Sigma, q_0, F, \delta)$:

1. Q er mengi af endanlegum stöðum, q_0, q_1 ... q_n.
2. Σ er endanlegt mengi af inntaksstöfum.
3. q_0 er upphafsstaða, $q_0 \in Q$.
4. F er mengi lokastaða, $F \in Q$.
5. δ er breytingafall (e. transition function) $Q \times \Sigma \rightarrow Q$. $\delta(q, i)$ skilar þeirri stöðu sem vélin fer í úr stöðu q miðað við inntak i.

Dæmi: $Q = \{q_0, q_1, q_2\}$, $\Sigma = \{a, b, c\}$, $F = \{q_2\}$,
$\delta = \{\delta(q_0, a) = q_1, \delta(q_1, b) = q_1, \delta(q_1, c) = q_2\}$

Endanlegar stöðuvélar: Tvær tegundir

Löggeng stöðuvél - DFA (e. Deterministic Finite Automaton)

Aðeins ein leið út út hverri stöðu fyrir gefinn inntaksstaf.

Brigðgeng stöðuvél - NFA (e. Non-deterministic Finite Automaton)

- Fleiri en ein leið getur verið út úr hverri stöðu fyrir gefinn inntaksstaf.
- The path is not **determined** in advance.
- ϵ (tómi strengurinn) má vera inntaksstafur.
- Dæmi: Sjá Fig. 2.3 bls. 31.

Hægt að breyta NFA í DFA á vélrænan hátt.

Algrím til að líkja eftir DFA

- Inntak: strengur x sem endar á EOF. DFA, D, með upphafsstöðu s_o og mengi, F, af lokastöðum.
- Úttak: Svarið "yes" ef D samþykkir x, annars "no".

$s = s_o$
c = nextchar();
while (c <> EOF) {
    s = move(s, c);   // skilar þeirri stöðu sem farið er í úr s við inntak c
    c = nextchar();
}
if $s \in F$ then return "yes"
else return "no";

Aðgerðir á endanlegum stöðuvélum

Helstu aðgerðir

- Sammengi (e. union)
- Samtenging (e. concatenation)
- Endurtekning (e. iteration; “Kleene Closure”)

Sammengi

- Sammengi stöðuvéla A og B ber kennsl á (eða myndar) alla strengi í A og alla strengi í B.
- Táknað $A \cup B$.
- Nýja vélin búin til með því að búa til nýja upphafsstöðu með ϵ-legg yfir í bæði A og B (Sjá Fig. 2.7 bls. 34).

Aðgerðir á endanlegum stöðuvélum

Helstu aðgerðir

- Sammengi (e. union)
- Samtenging (e. concatenation)
- Endurtekning (e. iteration; "Kleene Closure")

Sammengi

- Sammengi stöðuvéla A og B ber kennsl á (eða myndar) alla strengi í A og alla strengi í B.
- Táknað $A \cup B$.
- Nýja vélin búin til með því að búa til nýja upphafsstöðu með ϵ-legg yfir í bæði A og B (Sjá Fig. 2.7 bls. 34).

Aðgerðir á endanlegum stöðuvélum

Samtenging

- Samtenging stöðuvéla A og B ber kennsl á (eða myndar) alla strengi sem eru samtenging tveggja strengja, sá fyrsti er samþykktur af A og sá seinni af B.
- Táknað AB.
- Nýja vélin búin til með því að tengja allar lokastöður úr A við upphafsstöðu í B með ϵ-legg (Sjá Fig. 2.8 bls. 34).

Aðgerðir á endanlegum stöðuvélum

Endurtekning

- "Closure" af stöðuvél A samþykkir samtengingar af hvaða fjölda sem er af strengjum úr A ásamt tóma strengnum ϵ.
- Táknað A^*. $A^* = \{\epsilon\} \cup A \cup AA \cup AAA \cup \ldots$
- Nýja vélin fengin með því að tengja lokastöðu í A við upphafsstöðu í A með ϵ-legg og bæta við nýrri upphafsstöðu (Sjá Fig. 2.9 bls. 34).

Aðgerðir á endanlegum stöðuvélum

Aðrar algengar aðgerðir

- **Sniðmengi** (e. intersection). Sniðmengi tveggja stöðuvéla A og B ber kennsl á alla strengi sem samþykktir eru af bæði A og B.
- **Mismunur** (e. difference). Mismunur tveggja stöðuvéla A og B ber kennsl á alla strengi sem samþykktir eru af A en ekki B.
- **"Uppfylling"** (e. complementation).
  - Σ^* táknar óendanlegt mengi allra mögulegra strengja úr inntaksstafrófinu Σ.
  - Uppfylling stöðuvélar A í Σ^* ber kennsl á alla strengi sem eru ekki samþykktir af A, þ.e. $\hat{A} = \Sigma^* - A$.

Aðgerðir til að auka skilvirkni

- ϵ-removal.
  - Breytir upphaflegri stöðuvél í jafngilda án ϵ-leggja.
- Determination.
  - Breytir NFA vél í DFA.
- Minimisation.
  - Býr til jafngilda vél sem inniheldur eins fáar stöður og mögulegt er.