Málvinnsla: Forritunarverkefni V - Lokaverkefni

Háskólinn í Reykjavík - Tölvunarfræðideild

Október 2007

1 Lýsing

Í þessu síðasta forritunarverkefni hafið þið nokkuð frjálsar hendur. Þið megið sjálf koma með hugmyndir að verkefnum en hér fyrir neðan eru nokkrar tillögur. Athugið að vegna tímaskorts er lögð áhersla á að búa til frumgerð (en ekki "fullkomna" afurð) af einhverju kerfi. Gert er ráð fyrir að kerfið byggi á einhverju af því sem við höfum farið yfir í námskeiðinu.

Þetta verkefni má vinna í tveggja manna hópum.

1.1 Málfræðileiðrétting

Þróið frumgerð af forriti sem les íslenskan texta og bendir á málfræðivillur í honum. Dæmi um villur sem leita skal að eru misræmi á milli frumlags og sagnar (t.d. "ég ert") og misræmi í nafnliðum (t.d. "þessi stóra strákur").

Forritið skal gefa notanda kost á að slá inn texta og forritið bendir síðan á þær villur sem finnast. Notið IceNLP til að marka og setningagreina textann.

1.2 Vélræn þýðing

Þróið frumgerð á forriti sem þýðir úr íslensku yfir í ensku með svokallaðri "shallow-transfer" aðferð. Markið og hlutaþáttið (e. shallow parse) með IceNLP og þýðið síðan setningalið fyrir setningalið. Í einhverjum tilvikum þurfið þið að endurraða setningaliðum í þýðingunni (t.d Maríu elskar Jón - John loves Mary).

Þýðingakerfi sem byggja á "transfer" aðferð þurfa á lemmum (uppflettimyndum orða) að halda en lemmari fylgir ekki með IceNLP. Búið því til orðasafn sem varpar orðmyndum (ekki lemmum) úr íslensku yfir í ensku.

1.3 Samsetning markara

Notið markarana þrjá *IceTagger*, *TnT* og *fnTBL*, til að búa til samsettan markara sem "kýs" um rétt mörk. Notið 01TM.txt sem þjálfunarmálheild fyrir *TnT* og *fnTBL* og prófið alla markarana á 01PM.txt.

Skrifið notendaviðmót sem býður notanda að slá inn texta til að marka með samsetta markaranum og birtir niðurstöðuna. Hver er hlutfallsleg bæting á samsettum markara miðað við þann einstaka markara sem nær hæstri nákvæmni?

1.4 Tölfræðilegt mállíkan

Búð til forrit sem getur myndað íslenskar orðarunur (setningar). Forritið byggir á þrístæðulíkani fyrir íslenskan texta.

Notið málheildina 01TM.txt (eingöngu orðin sjálf, ekki mörkin) við þjálfun. Notið síðan líkanið til að mynda orðarunur. Leyfið notandanum að slá inn fyrsta orðið en forritið velur síðan næstu n (stillanlegt af notanda) líklegustu orð með tillitil til mállíkansins.

1.5 Talmörkun texta

Markmiðið er að búa til forrit sem setur sjálfvirkt inn í texta sérstök ítónunarmörk sem myndu gefa talgervli til kynna hvar réttar áherslur eru útfrá einhverri þekkingu á samhengi. Einnig væri hægt að merkja tónfall setningahluta útfrá upplýsingaskipan (e. Information Structure). Aðferðin byggir á orðræðulíkani sem heldur utan um samhengi textans og nokkrum þekktum mynstrum sem lýsa tengslum raddar við orðræðu. Einnig er beitt einföldu algrími til að finna remu og þemu í texta.

Enn sem komið er höfum við ekki aðgang að íslenskum talgervli, þannig að ef þetta er gert á íslensku er nóg að skila talmörkuðum texta. Sé enska notuð, er hægt að nota talgervil sem úttak.

1.6 Náttúrulegt viðmót

Markmiðið er að búa til viðmót milli manns og tölvu sem byggir á eðlilegri íslensku. Notandi kerfisins myndi þá skrifa texta á venjulegu máli, t.d. spurningar, sem tölvan vinnur úr og svarar með viðeigandi texta, einnig á venjulegu máli.

Til að þetta sé raunhæft í stuttu verkefni, má takmarka samskiptin við mjög afmarkað svið og afmarkaðar setningagerðir. Hér gæti t.d. verið um fyrirspurnir í einfaldan gagnagrunn að ræða eins og "Hvert er símanúmer

Jóns?" eða brot úr textaævintýraleik: "T: Hvað viltu gera? N: Ég ætla að opna dyrnar. T: Þú opnar dyrnar og sérð fjársjóð."

2 Skilafrestur

Forritið skal sýnt í síðasta fyrirlestri, föstudaginn 9. nóvember.