Málvinnsla: Forritunarverkefni IV - Orðræðulíkan

Háskólinn í Reykjavík - Tölvunarfræðideild

Október 2007

1 Lýsing

Í þessu verkefni eigið þið að búa til einfalt orðræðulíkan fyrir (íslenskan) texta sem heldur utan um orðræðueiningar og nýtir líkanið til að leysa úr tilvísunum persónufornafna. Þið vinnið bara með nafnliði í textanum og gerið ráð fyrir því að þeir séu allir annað hvort ný vísun eða tilvísun á orðræðueiningar í líkaninu. Hér er dæmi um þetta:

- "Kisa sá stóran bíl, hann var rauður"
- Nafnliðir: "Kisa" (ný vísun), "stóran bíl" (ný vísun), "hann" (tilvísun, "stóran bíl" og "hann" er dæmi um samvísun)

1.1 Inntak

Það sem þið fáið í hendurnar er markaður texti sem þið byrjið á að þátta, annað hvort með IceParser eða þeim þáttara sem þið smíðuðuð sjálf í verkefni III. Þetta er dæmi um textann hér að ofan eftir að IceParser er búinn að þátta:

```
[NP Kisa nven NP]
[VP sá sfg3eþ VP]
[NP [AP stóran lkeosf AP] bíl nkeo NP]
, ,
[NP hann fpken NP]
[VPb var sfg3eþ VPb]
[AP rauður lkensf AP]
```

Þegar textinn er kominn á þáttað form, á að vera auðvelt að sjá þar alla nafnliði. Á þessu formi sendið þið textann inn í orðræðulíkanið, t.d. með skipun eins og:

```
discourseprocess parsedtext.txt output.txt
```

1.2 Aðferð

Þið getið notað það forritunarmál sem ykkur þykir þægilegast að vinna með. Orðræðulíkanið á að skoða hvern nafnlið fyrir sig í inntakinu, en þarf ekki að skoða annað. Þið getið t.d. notað reglulegar segðir til að ná út öllum nafnliðum. Ef nafnliður er ekki persónufornafn, skal búin til ný orðræðueining á lista yfir orðræðueiningar líkansins. Þessi listi á að vera svokallaður nýleikalisti (e. recency list) þar sem þið bætið alltaf nýjum einingum fremst.

Með hverri einingu skal geyma þrennt: Einkvæmt nafn, nafnliðinn sjálfan og einkenni. Einkvæma nafnið búið þið til sjálf og getið t.d. notað orð úr nafnliðnum að viðbættri raðtölu. Einkenni inniheldur gildi á nokkrum einkennandi breytum fyrir eininguna og má útfæra sem tengifylki (associative array, hash table, map). Breyturnar sem þið skulið setja þarna inn eru *kyn* og *tala* sem þið fáið úr mörkum nafnliðarins.

Þegar orðræðulíkanið fær nafnlið sem er persónufornafn í þriðju persónu, þá ætlum við að beita hér svokallaðri nýleikaaðferð við lausn tilvísunarinnar. Líkanið byrjar þá á því að bera einkenni persónufornafnis við einkenni fremstu einingarinnar á nýleikalistanum. Ef um samræmi er að ræða, skulið þið skrá það sem tilvísun á viðkomandi einingu. Ef samræmi er ekki við fyrstu eininguna, prófið þið næstu einingu á listanum og svo koll af kolli. Sé vísað á einingu sem er ekki fremst á nýleikalistanum, þá skal sú eining flutt fremst til að tákna að nýbúið sé að fjalla um hana aftur. Þannig eru fremstu einingarnar á listanum alltaf virkasta umfjöllunarefnið.

Séu persónufornöfnin í fyrstu eða annari persónu, þá getið þið búið til einingar á listanum sem tákna framleiðanda textans (sögumann) og viðtakanda (áheyrendur).

1.3 Úttak

Úttakið er tvíþætt. Fyrst skal ritaður listi af öllum orðræðueiningum sem eru í líkaninu eftir keyrslu, með þeirri nýjustu efst. Síðan er textinn allur skrifaður út, þar sem búið er að merkja alla nafnliði með nafni þeirrar orðræðueiningu sem vísað er í. Hær er dæmi um úttak:

```
OÐRÆÐUEININGAR (eftir nýleika):
BIL1, stóran bíl, {kyn:k, tala:e}
KISA1, kisa, {kyn:v, tala:e}

TEXTAÚTTAK:
[NP Kisa nven KISA1 NP]
[VP sá sfg3eþ VP]
```

```
[NP [AP stóran lkeosf AP] bíl nkeo BIL1 NP]
, ,
[NP hann fpken BIL1 NP]
[VPb var sfg3eþ VPb]
[AP rauður lkensf AP]
```

2 Úrvinnsla

Eins og fyrr segir, þá fáið þið markaðan texta sem þið beitið aðferðinni á. Skoðið úttakið ykkar vel og skilið tveimur handgerðum greiningarlistum yfir þá staði í textanum...

- ...þar sem líkanið lét persónufornafn benda á ranga orðræðueiningu
- ...þar sem líkanið bjó til tvær eða fleiri einingar fyrir sama hlutinn (og þar með, missti af samvísun)

Fyrir hver svona mistök, segið til um hvaða aukavitneskju líkanið hefði þurft að hafa til að geta leyst rétt úr (t.d. hvaða fleiri einkenni, málfræðileg eða merkingarleg, hefðu þurft að vera með orðræðueiningunum). Það er nóg að lýsa hverri tegund mistaka einu sinni. Takið fram hvort þið eruð að nota ykkar eigin þáttara eða IceParser til að undirbúa inntakið.

3 Skilafrestur

Forritskóða ásamt úttaki og úrvinnslu, skal skila í síðasta lagi föstudaginn 26. október, kl. 23:59.