Málvinnsla: Forritunarverkefni II - Mörkun texta

Háskólinn í Reykjavík - Tölvunarfræðideild

September 2007

1 Grunnmarkari (70%)

Í þessum hluta eigið þið að búa til *grunnmarkara* (e. base tagger) fyrir íslenskan texta, þ.e. markara sem velur ávallt líklegasta markið fyrir sérhvert orð. Þið þjálfið markarann ykkar með því að nota þjálfunarmálheildina *01TM.txt*. Þið ráðið því hvernig þið farið að þessu en þið gætuð gert þetta á eftirfarandi máta.

1. **Þjálfun**

   - Notið *uniq* og *sort* Linux (Cygwin) skipanir til að búa til skrána *01TM.freq*, sem sýnir hversu oft tiltekið orð er markað með tilteknu marki í *01TM.txt*. Dæmi um línur í *01TM.freq* er:

     ```
     19852 og c
     19770 , ,
     10497 að cn
      9024 í aþ
      6611 var sfg3eþ
      6320 hann fpken
      ...
      2959 af aþ
      ...
       192 af aa
      ...
     ```

   - Búið til Perl forrit, *buildLexicon.pl* sem les skrá á því sniði sem *01TM.freq* er á og skrifar út orðasafnsskrá, t.d. *01TM.lex* á eftirfarandi sniði:

     ```
     af 3151 aþ 2959 aa 192
     ```

```
afa 66 nkee 27 nkeþ 24 nkeo 15
afar 42 aa 41 nkfn 1
afarbáglega 1 aa 1
afatúni 1 nheþ 1
...
```

Sérhver lína sýnir hversu oft tiltekið orð kom fyrir ("af" kom fyrir 3151 sinni), og síðan mörk og tíðni fyrir viðkomandi orð (t.d. kom markið "aþ" fyrir 2959 sinnum fyrir á orðinu "að").

Athugið að orðasafnið sem út kemur inniheldur í raun meiri upplýsingar en þið þurfið. Fyrir þetta tiltekna verkefni (grunnmörkun) þá myndi duga ykkur að skila skrá á eftirfarandi sniði (þið megið gera það ef ykkur finnst hitt of strembið):

```
af aþ
afa nkee
afar aa
afarbáglega aa
afatúni nheþ
...
```

þ.e. skrá sem inniheldur eingöngu orð og líklegast markið (án tíðnitalna). Skráin með tíðniupplýsingum er hins vegar nytsamlegri ef t.d. búa þyrfti til markara byggðan á HMM.

2. **Mörkun á nýjum texta**. Skrifið forrit í Perl, *baseTagger.pl*, sem tekur inn orðasafn (sjá að ofan) og tilreiddan texta, þ.e. eitt orð í línu með auðri línu á milli setninga. Forritið skal skrifið út sérhvert orð ásamt líklegasta markinu. Ef orð er óþekkt þá skal skrifa út markið *nken-m* ef orðið byrjar á stórum staf, annars *nken*. Einnig skal skrifa út *<UNK>* í lok línunnar til að merkja að um óþekkt orð er að ræða.

   Dæmi um notkun: *perl baseTagger.pl 01TM.lex 01PM.txt 01PM.out*

   Dæmi um línur í *01PM.out*:

```
ég fp1en
stökk sfg3eþ
á aþ
eftir aþ
strætó nkeþ
og c
veifaði sfg3eþ
```

```
, ,
vagnstjórinn nken <UNK>
sá sfg3eþ
mig fp1eo
og c
stoppaði sfg3eþ
```

2 Hittni (30%)

Skrifið forrit í Perl, *accuracy.pl*, sem tekur inn eina skrá hvers línur eru á eftirfarandi sniði:

```
orð rétt_mark orð mark_úr_markara <UNK>
```

(<UNK> kemur aðeins fyrir þegar um óþekkt orð er að ræða).

Þið getið t.d. notað Linux *paste* til að sameina *01PM.txt* (*gold standard*) og *01PM.out* (úttakið úr grunnmarkaranum).

Úttakið (á skjá) á að vera hittni markarans fyrir þekkt orð, óþekkt orð og öll orð.

Dæmi um notkun: *perl accuracy.pl sameinud_skra*

Dæmi um úttak:

```
Number of tokens: 59169
Number of errors: 14419
Overall tagging accuracy: 75.63%
Tagging accuracy for known words: 81.53%
Number of unknown words: 4482
Unknown word ratio: 7.57%
Number of errors for unknown words: 4317
Tagging accuracy for unknown words: 3.68%
```

Hér er þægilegt að skrifa skel sem byrjar á því að sameina sameina *01PM.txt* og *01PM.out* í eina skrá og kallar svo á *accuracy.pl*.

3 Aukaverkefni - valfrjálst

Getið þið bætt við einföldum málfræðilegum reglum inn í *baseTagger.pl* í þeim tilgangi að hækka hittnina? Þið gætuð t.d. bætt við reglum til að bæta hittnina á óþekktum orðum og/eða á þekktum orðum.

Prófið ykkur áfram og notið *accuracy.pl* til að reikna út hittnina.

4 Skilafrestur

Öllum forritskóða (líka skeljakóða), ásamt úttaki úr *accuracy.pl* miðað við grunnmörkunina (og endurbætta mörkun) á 01PM.txt, skal skila í síðasta lagi föstudaginn 5. október, kl. 23:59.

Þið megið nota annað forritunarmál en Perl ef þið viljið.