T-61.281 Luonnollisen kielen tilastollinen käsittely

Harjoitus 5, ti 24.2.2004, 8:30-10:00 N-grammikielimallit, Versio 1.1

1. Tämä tehtävä kannattanee tehdä järjestyksessä, kurkkimatta seuraaviin kohtiin, jolloin estimaatteihisi ei vaikuta kuin siihen mennessä kertynyt tieto.

   a) Tehtävänäsi on arvioida sanoja "tuntumaan jo" seuraavan sanan todennäköisyydet. Mahdollisia jatkosanoja ovat

   - ja
   - hyvältä
   - kumisaapas
   - keväältä
   - ilman
   - päihtyneeltä
   - turhalta
   - koirineen
   - öljyiseltä
   - Turku

   Vertaa antamiasi estimaatteja kielioppiaineistosta laskettuihin (tasoittamattomiin) estimaatteihin. Kumpi antoi paremman todennäköisyyden oikealle sanalle ?

   b) Tiedät koko lauseen alun, joka on "Leuto sää ja soidinmenonsa aloittaneet tiaiset ovat saaneet helmikuun tuntumaan jo X". Arvioi nyt tämän kontekstin perusteella uudestaan annettujen sanojen todennäköisyydet.

   c) Mitä erilaisia tietoja kielioppimalli tarvitsisi, jotta se pystyisi vastaamaan ihmisen suorituskykyyn b)-kohdassa ?

   Alkuperäisessä lauseessa ollut sana löytyy paperin kääntöpuolelta.

2. Mallin sanaston koko on 64 000 sanan perusmuotoa. Tiedetään että sanahistoria on (1) "vuosi joka olla" ja (2) "tämä tehtävä vaikuttaa". Arvioi todennäköisyydet sille, että seuraava sana on "olla", "leuto" tai "gorilla". Arvioi uni- ja bigrammitodennäköisyydet käyttäen

   a) ...suurimman uskottavuuden menetelmällä
   b) ...Laplace-tasoitettuja estimaatteja.
   c) ...Lidstone-tasoitettuja estimaatteja.

   Käytä opetusaineistona tehtävän 1 lauseita.

r	N_r
1	22717431
2	1062078
3	250977
4	94156
5	47650
6	29346
7	19267
8	13449
9	9899
10	7462
⋮	
7722	1

Taulukko 1: *Trigrammien frekvenssien frekvenssit.*

3. Good-Turing tasoituksen laskemisessa voi olla enemmän mutkia matkassa, kuin ensi näkemältä luulee. Laske siis Good-Turing estimaatit 30 miljoonan sanan aineistosta lasketuille arvoille, jotka löytyvät osoitteesta http://www.cis.hut.fi/Opinnot/T-61.281/Laskarit03 /h5/data3.txt. Mallin on oletettu tuntevan 64 000 yleisintä sanaa, muut sanat on kasattu sanaksi "UNK", joka edustaa kaikkia sanaston ulkopuolisia sanoja. Taulukon ensimmäisessä sarakkeessa r on annettu trigrammin esiintymisfrekvenssi. Toisessa sarakkeessa N_r taasen annetaan, kuinka monta trigrammia aineistosta löytyy, jotka ovat esiintyneet r kertaa. Voit myös laskea estimaatit pienemmällä aineistolla, Taulukkoon 1 on valittu joitain arvoja mainitusta datatiedostosta:

4. Laske annetun kielimallin hämmentyneisyys (perplexity) lauseelle "Kielen oppiminen on monimutkainen ja huonosti ymmärretty tapahtumaketju."

   Perääntyvän (back-off) kielimallin todennäköisyydet voidaan laskea seuraavasti:

$$P(w_3|w_2,w_1) = \begin{cases} T(w_1,w_2,w_3) & \text{jos löytyy trigrammi } w_1,w_2,w_3 \\ bo(w_1,w_2)P(w_3|w_2) & \text{jos löytyy bigrammi } w_1,w_2 \\ P(w_3|w_2) & \text{muutoin} \end{cases}$$

$$P(w_2|w_1) = \begin{cases} T(w_1,w_2) & \text{jos löytyy bigrammi } w_1,w_2 \\ bo(w_1)T(w_2) & \text{muuten} \end{cases}$$

   Funktioiden T ja bo arvot voidaan löytää taulukosta 2.

Ensimmäisen tehtävän alkuperäisen lauseen haettu sana oli "keväältä".

n-grammi	log(T)	log(bo)
kielen	-4.1763	-0.2917
kielen oppiminen	-2.1276	-0.0526
kielen oppiminen on	-0.4656	
oppiminen on	-0.5889	-0.001
on monimutkainen	-4.2492	-0.0697
on monimutkainen ja	-0.8876	
monimutkainen ja	-0.8660	0.0495
ja huonosti	-4.1804	-0.1415
huonosti	-4.2513	-0.1652
ymmärretty	-5.2195	-0.0870

Taulukko 2: *Kielimallli on opetettu n. 30 miljoonan sanan lähdemateriaalista 64 000 yleisimmälle sanalle. Estimaatit on laskettu Good-Turing tasoituksella ja Katz-perääntymisellä. Taulukkoon on poimittu tehtävän kannalta relevantit arvot.*