T-61.5020 Luonnollisten kielten tilastollinen käsittely

Harjoitus 8, ke 21.3.2007, 12:15–14:00 — N-grammikielimallit
Versio 1.0

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 menetelmää
   b) ...Laplace-tasoitettuja estimaatteja
   c) ...Lidstone-tasoitettuja estimaatteja (*additive smoothing*) parametrilla $\lambda = 0.01$

   Käytä opetusaineistona tehtävän 1 lauseita.

3. Edellisessä tehtävässä laskettiin erilliset tasoitetut jakaumat unigrammimallin ja bigrammimallin todennäköisyyksille. Hyödyllisempää on kuitenkin yhdistää eri mittaisten n-grammien estimaatit yhdeksi perääntyväksi (back-off) tai interpoloiduksi malliksi. Esimerkiksi kun tarkasteltiin sanojen olla ja leuto todennäköisyyksiä kontekstissa "vaikuttaa", estimaateiksi tuli samat koska opetusaineistossa ei ollut kummallekaan bigrammille esiintymiä. Kuitenkin unigrammitodennäköisyyksistä tiedetään, että olla-verbi on huomattavasti yleisempi kuin sana leuto, ja täten voidaan ajatella sen olevan yleisempi myös ennen näkemättömässä kontekstissa.

   Laske siis tällä kertaa interpoloidun bigrammimallin antamat todennäköisyydet edellisen tehtävän esimerkkihistorioille. Käytä bigrammien tasoituksena absoluuttista vähennystä (*absolute discounting*) parametrilla $D = 0.5$.

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 1.

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 1: *Kielimalli 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.*

Ensimmäisen tehtävän alkuperäisen lauseen haettu sana oli "keväältä".