Teknillinen korkeakoulu, Tietotekniikan osasto
Informaatiotekniikan laboratorio
Timo Honkela, p. 050 384 1578

T-61.281 Luonnollisen kielen tilastollinen käsittely
tentti 8.1.2004

Kirjoita jokaiseen vastauspaperiisi seuraavat tiedot:

- nimi, opintokirjannumero, osasto (KIT-opiskelijat: myös yliopiston nimi)
- sanat: "T-61.281 Luonnollisen kielen tilastollinen käsittely, tentti 8.1.2004"

Esseetyyppisten kysymysten arvostelussa kiinnitetään huomiota vastauksen jäsentelyyn ja selkeyteen, tiiviyteen ja kattavuuteen.

1. Selitä lyhyesti seuraavat termit tai lyhenteet (1p/ termi):
   - LSI (latent semantic indexing)
   - kollokaatio
   - n-grammimalli
   - tekstinlinjaus (text alignment)
   - saanti (recall) tiedonhaussa
   - PCFG

2. Selitä kielimallien estimoinnin yleiset periaatteet ja esittele mallien estimoinnin perusmenetelmiä (esimerkiksi ristiinvalidointi). (6p)

3. a) Selitä lyhyesti, millä tavalla sanojen yhdessäesiintyminen voi antaa tietoa sanojen syntaktisista rooleista ja merkityksestä. Millä tavalla sanojen monitulkintaisuus vaikuttaa analyysiin? (3p)

   b) Taulukossa 1 on annettu eri sanojen esiintymä- ja yhteisesiintymäfrekvenssejä. Käytä yhteisinformaatiota sen mittaamiseen kuinka hyviä kollokaatioita sanaparit ovat ja anna paremmuusjärjestys (voit soveltaa suurimman uskottavuuden estimointia). (3p)

   Pisteittäinen yhteisinformaatio tapahtumien x ja y välillä: $\log \frac{P(x,y)}{P(x)P(y)}$

4. Selitä, mitä tarkoittavat Markov-ominaisuudet äärellinen horisontti ja aikariippumattomuus (4p), ja selitä, millä tavalla Markov-malleja voidaan esittää probabilistisina äärellisinä tila-automaatteina (2p).

5. Viterbi-algoritmi etsii tehokkaasti annettua havaintojonoa vastaavan todennäköisimmän tilajonon. Algoritmissa talletetaan jokaiseen hilapisteeseen siihenastinen todennäköisin polku:

$$\delta_j(t) = \max_{X_1 \ldots Xt-1} P(X_i \ldots X_{t-1}, o_1 \ldots o_{t-1}, X_t = j|\mu) \quad (1)$$

   Initialisointi perustuu tilojen alkutodennäköisyyksiin: $\delta_j(1) = \pi_j$.

   Induktioaskeleessa valitaan edellinen tila, josta tulee suurin jonon todennäköisyys:
   - talletetaan todennäköisyys: $\delta_j(t+1) = \max_{1 \le i \le N} \delta_i(t) a_{ij} b_{ijo_t}$
   - talletetaan osoitin ko. tilaan: $\phi_j(t+1) = \arg\max_{1 \le i \le N} \delta_i(t) a_{ij} b_{ijo_t}$

Merkinnällä a viitataan transitiotodennäköisyyksiin ja b:llä havaintotodennäköisyyksiin. Symboli μ tarkoittaa mallin parametrejä (a:t, b:t ja π:t), ja X tiloja.

Kehitä yksinkertainen esimerkki, jonka avulla esittelet algoritmin toimintaa. (6p).

s_1	s_2	$C(s_1)$	$C(s_2)$	$C(s_1, s_2)$
hakea	työ	11879	30364	37
herne	nenä	129	1119	3
aste	pakkanen	3729	343	69
heittää	veivi	9003	26	6
kova	tuuli	22900	4623	294
kuiva	keli	773	759	29

Taulukko 1: Sanojen esiintymistiheyksiä. $C(a)$ kertoo, kuinka monta kertaa tapahtuma a esiintyi testijoukossa. Aineistossa oli kaiken kaikkiaan 38 887 883 sanaa.