T-61.5020 Luonnollisten kielten tilastollinen käsittely

Vastaukset 11, ke 18.4.2007, 12:15–14:00 — Puheentunnistus ja kielimallien evaluointi
Versio 1.0

1. Käytämme siis jälleen viterbi-algoritmia todennäköisimmän tilasekvenssin selvittämiseen kätketystä Markov-mallista. Eroja aikaisemman laskarin säätilamalliin on kolme: Emissiot tapahtuvat kaarissa eli tilojen välillä, mallissa on nollatransitiota ja lopetustila on määrätty.

   a) Alustetaan hila siten, että lähtötila on aina S_1. Merkitään ylös vain nollasta poikkeavat todennäköisyysarvot.

$$\delta_0(1) = 1$$

   Ensimmäinen havainto

   Aloitustilasta pääsee vain toiseen ja neljänteen tilaan, joten lasketaan niiden todennäköisyydet:

$$\begin{aligned}
\delta_1(2) &= a_{12}b_{12}(o_1) = 0.5 \cdot 10^{-1} = 5 \cdot 10^{-2} \\
\psi_1(2) &= 1 \\
\delta_1(4) &= a_{14}b_{14}(o_1) = 0.5 \cdot 10^{-3} = 5 \cdot 10^{-4} \\
\psi_1(4) &= 1
\end{aligned}$$

   Toinen havainto

   Toisesta tilasta päästään vain kolmanteen ja neljännestä vain viidenteen tilaan, joten vieläkään ei tarvita reittivalintoja.

$$\begin{aligned}
\delta_2(3) &= \delta_1(2)a_{23}b_{23}(o_2) = 5 \cdot 10^{-2} \cdot 1.0 \cdot 10^{-1} = 5 \cdot 10^{-3} \\
\psi_2(3) &= 2 \\
\delta_2(5) &= \delta_1(4)a_{45}b_{45}(o_2) = 5 \cdot 10^{-4} \cdot 1.0 \cdot 10^{-4} = 5 \cdot 10^{-8} \\
\psi_2(5) &= 4
\end{aligned}$$

   Tässä vaiheessa pitää kuitenkin huomata, että tiloista S_3 ja S_5 pääsee nollasiirtymällä aloitustilaan. Toisen havainnon jälkeen voidaan siis päätyä myös sinne. Lasketaan tilaan tulevista reiteistä todennäköisempi:

$$\begin{aligned}
\delta_2(1) &= \max(\delta_2(3)a_{31}\,,\, \delta_2(5)a_{51}) \\
&= \max(5 \cdot 10^{-3} \cdot 0.9\,,\, 5 \cdot 10^{-8} \cdot 1.0) \\
&= 4.5 \cdot 10^{-3} \\
\psi_2(1) &= 3
\end{aligned}$$

   Kolmas havainto

Nyt mahdollisia ovat siirtymät tilasta S_1 tilaan S_2 tai S_4 sekä tilasta S_3 tilaan S_4.

$$\begin{aligned}
\delta_3(2) &= \delta_2(1)a_{12}b_{12}(o_3) = 4.5 \cdot 10^{-3} \cdot 0.5 \cdot 10^{-3} = 2.25 \cdot 10^{-6} \\
\psi_3(2) &= 1 \\
\delta_3(4) &= \max(\delta_2(1)a_{14}b_{14}(o_3)\,,\, \delta_2(3)a_{34}b_{34}(o_3)) \\
&= \max(4.5 \cdot 10^{-3} \cdot 0.5 \cdot 10^{-2}\,,\, 5 \cdot 10^{-3} \cdot 0.1 \cdot 10^{-1}) \\
&= 5 \cdot 10^{-5} \\
\psi_3(4) &= 3
\end{aligned}$$

Neljäs havainto

Jälleen toisesta tilasta päästään vain kolmanteen ja neljännestä vain viidenteen tilaan.

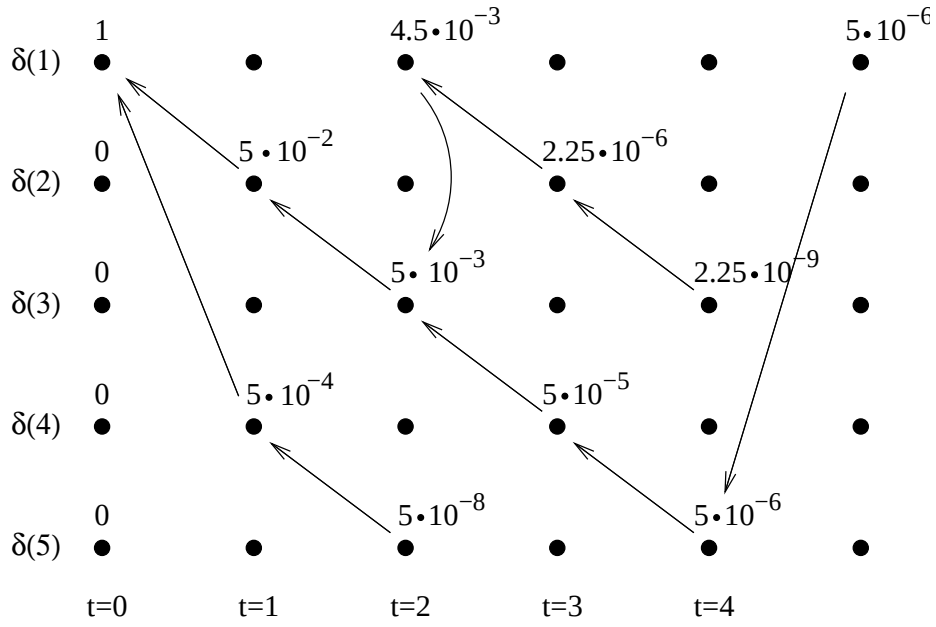
$$\begin{aligned}
\delta_4(3) &= \delta_3(2)a_{23}b_{23}(o_4) = 2.25 \cdot 10^{-6} \cdot 1.0 \cdot 10^{-3} = 2.25 \cdot 10^{-9} \\
\psi_4(3) &= 2 \\
\delta_4(5) &= \delta_3(4)a_{45}b_{45}(o_4) = 5 \cdot 10^{-5} \cdot 1.0 \cdot 10^{-1} = 5 \cdot 10^{-6} \\
\psi_4(5) &= 4
\end{aligned}$$

Lopetus

Lopuksi piti päästä takaisin tilaan S_1, mikä onnistuu nollasiirtymällä. Reitti-vaihtoehtoja on kaksi:

$$\begin{aligned}
\delta_4(1) &= \max(\delta_4(3)a_{31}\,,\, \delta_4(5)a_{51}) \\
&= \max(2.25 \cdot 10^{-9} \cdot 0.9\,,\, 5 \cdot 10^{-6} \cdot 1.0) \\
&= 5 \cdot 10^{-6} \\
\psi_4(1) &= 5
\end{aligned}$$

Laskettu hila on kuvassa 1. Palaamalla lopusta alkuun saadaan todennäköisin tilasekvenssi $S_1 \to S_2 \to S_3 \to S_4 \to S_5 \to S_1$. Tämä vastaa sanaa “jaon”.



Kuva 1: Viterbi-haun hila lopetustilaan saavuttaessa.

b) Nyt pitää ottaa huomioon myös kielimallin antamat todennäköisyydet. Lasketaan todennäköisyysarvot ehdollisina mahdollisille eri sanavaihtoehdoille w_j: $\delta_t(i, w_j)$. Kielimallitodennäköisyys kerrotaan mukaan aina kun sanan valinta tehdään. Kuljettaessa uudestaan aloitustilan kautta valinnat pitää ottaa huomioon käyttämällä bigrammitodennäköisyyksiä. Tämän jälkeen ne voidaan tyhjentää, koska kielimalli ei tarvitse pidempiä historioita.

Alustetaan hila samoin kuin a)-kohdassa. Tässä vaiheessa ei ole vielä sanavalintoja.

$$\delta_0(1, \_) = 1$$

Ensimmäinen havainto

Aloitustilasta pääsee toiseen ja neljänteen tilaan. Toinen tila voi päätyä sanaan "ja" tai "jaon", joten kumpikin vaihtoehto pitää laskea erikseen.

$$\begin{aligned}
\delta_1(2, \text{ja}) &= P(\text{ja}) a_{12} b_{12}(o_1) = 10^{-2} \cdot 0.5 \cdot 10^{-1} = 5 \cdot 10^{-4} \\
\psi_1(2, \text{ja}) &= 1 \\
\delta_1(2, \text{jaon}) &= P(\text{jaon}) a_{12} b_{12}(o_1) = 10^{-5} \cdot 0.5 \cdot 10^{-1} = 5 \cdot 10^{-7} \\
\psi_1(2, \text{jaon}) &= 1 \\
\delta_1(4, \text{on}) &= P(\text{on}) a_{14} b_{14}(o_1) = 10^{-2} \cdot 0.5 \cdot 10^{-3} = 5 \cdot 10^{-6} \\
\psi_1(4, \text{on}) &= 1
\end{aligned}$$

Toinen havainto

Toisesta tilasta päästään vain kolmanteen ja neljännestä vain viidenteen tilaan. Lisäksi kummastakin pääsee nollasiirtymällä ensimmäiseen tilaan. Tämä voidaan luonnollisesti tehdä vain sanoille jotka ovat käsitelty loppuun.

$$\begin{aligned}
\delta_2(3, \text{ja}) &= \delta_1(2, \text{ja}) a_{23} b_{23}(o_2) = 5 \cdot 10^{-4} \cdot 1.0 \cdot 10^{-1} = 5 \cdot 10^{-5} \\
\psi_2(3, \text{ja}) &= 2 \\
\delta_2(3, \text{jaon}) &= \delta_1(2, \text{jaon}) a_{23} b_{23}(o_2) = 5 \cdot 10^{-7} \cdot 1.0 \cdot 10^{-1} = 5 \cdot 10^{-8} \\
\psi_2(3, \text{jaon}) &= 2 \\
\delta_2(5, \text{on}) &= \delta_1(4, \text{on}) a_{45} b_{45}(o_2) = 5 \cdot 10^{-6} \cdot 1.0 \cdot 10^{-4} = 5 \cdot 10^{-10} \\
\psi_2(5, \text{on}) &= 4
\end{aligned}$$

$$\begin{aligned}
\delta_2(1, \text{ja}) &= \delta_2(3, \text{ja}) a_{31} = 5 \cdot 10^{-5} \cdot 0.9 = 4.5 \cdot 10^{-5} \\
\psi_2(1, \text{ja}) &= 3 \\
\delta_2(1, \text{on}) &= \delta_2(5, \text{on}) a_{51} = 5 \cdot 10^{-10} \cdot 1.0 = 5 \cdot 10^{-10} \\
\psi_2(1, \text{on}) &= 5
\end{aligned}$$

Kolmas havainto

Mahdollisia ovat siirtymät tilasta S_1 tilaan S_2 tai S_4 sekä tilasta S_3 tilaan S_4. Tilasta S_1 lähtevät siirtymät aloittavat uuden sanan, joten otamme kielimallitodennäköisyydet huomioon. Lisäksi pitää huomioida että tilassa S_1 oli kaksi eri sanavaihtoehtoa, joista nyt voidaan valita todennäköisempi.

$$\begin{aligned}
\delta_3(2,\text{ja}) &= \max(P(\text{ja}|\text{ja})\delta_2(1,\text{ja})\,,\, P(\text{ja}|\text{on})\delta_2(1,\text{on})) \cdot a_{12}b_{12}(o_3) \\
&= \max(10^{-4} \cdot 4.5 \cdot 10^{-5}\,,\, 10^{-2} \cdot 5 \cdot 10^{-10}) \cdot 0.5 \cdot 10^{-1} \\
&= 2.25 \cdot 10^{-10} \\
\psi_3(2,\text{ja}) &= 1 \\
\delta_3(4,\text{on}) &= \max(P(\text{on}|\text{ja})\delta_2(1,\text{ja})\,,\, P(\text{on}|\text{on})\delta_2(1,\text{on})) \cdot a_{14}b_{14}(o_3) \\
&= \max(10^{-2} \cdot 4.5 \cdot 10^{-5}\,,\, 10^{-4} \cdot 5 \cdot 10^{-10}) \cdot 0.5 \cdot 10^{-2} \\
&= 2.25 \cdot 10^{-9} \\
\psi_3(4,\text{on}) &= 1 \\
\delta_3(4,\text{jaon}) &= \delta_2(3,\text{jaon})a_{34}b_{34}(o_3) \\
&= 5 \cdot 10^{-8} \cdot 1.0 \cdot 10^{-1} = 5 \cdot 10^{-10} \\
\psi_3(4,\text{jaon}) &= 3
\end{aligned}$$

Neljäs havainto

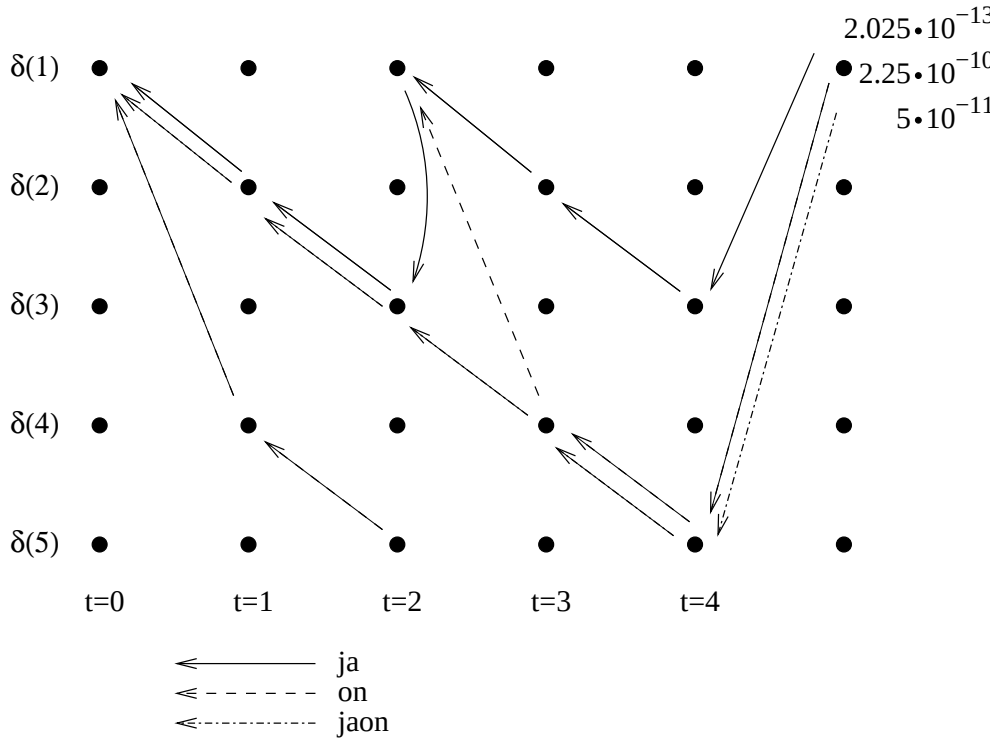
Toisesta tilasta päästään vain kolmanteen ja neljännestä vain viidenteen tilaan. Kummastakin päästään vielä nollasiirtymällä ensimmäiseen tilaan.

$$\begin{aligned}
\delta_4(3,\text{ja}) &= \delta_3(2,\text{ja})a_{23}b_{23}(o_4) = 2.25 \cdot 10^{-10} \cdot 1.0 \cdot 10^{-3} = 2.25 \cdot 10^{-13} \\
\psi_4(3,\text{ja}) &= 2 \\
\delta_4(5,\text{on}) &= \delta_3(4,\text{on})a_{45}b_{45}(o_4) = 2.25 \cdot 10^{-9} \cdot 1.0 \cdot 10^{-1} = 2.25 \cdot 10^{-10} \\
\psi_4(5,\text{on}) &= 4 \\
\delta_4(5,\text{jaon}) &= \delta_3(4,\text{jaon})a_{45}b_{45}(o_4) = 5 \cdot 10^{-10} \cdot 1.0 \cdot 10^{-1} = 5 \cdot 10^{-11} \\
\psi_4(5,\text{jaon}) &= 4
\end{aligned}$$

$$\begin{aligned}
\delta_4(1,\text{ja}) &= \delta_4(3,\text{ja})a_{31} = 0.9 \cdot 2.25 \cdot 10^{-10} = 2.025 \cdot 10^{-13} \\
\psi_4(1,\text{ja}) &= 3 \\
\delta_4(1,\text{on}) &= \delta_4(5,\text{on})a_{51} = 1.0 \cdot 2.25 \cdot 10^{-9} = 2.25 \cdot 10^{-10} \\
\psi_4(1,\text{on}) &= 5 \\
\delta_4(1,\text{jaon}) &= \delta_4(5,\text{jaon})a_{51} = 1.0 \cdot 5 \cdot 10^{-10} = 5 \cdot 10^{-11} \\
\psi_4(1,\text{jaon}) &= 5
\end{aligned}$$

Laskettu hila on kuvassa 2. Eri sanavaihtoehdot on piirretty kuvaan erilaisilla nuolilla. Kolmesta lopetustilaan päätyneestä polusta todennäköisin on $\delta_4(1,\text{on})$.

Palaamalla siitä taaksepäin saadaan todennäköisin tilasekvenssi $S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow S_1 \rightarrow S_4 \rightarrow S_5 \rightarrow S_1$. Nyt todennäköisimmäksi saatiin siis kahden sanan sekvenssi, "ja on".



Kuva 2: Hila lopetustilaan saavuttaessa. Eri sanavaihtoehdot on merkitty erilaisilla nuolilla.

2. Mallissa A on noin kolme kertaa vähemmän yksiköitä kuin mallissa B. Yksiköt ovat myös keskimäärin pienempiä, ja testidata sisältää enemmän mallin A yksiköitä. Tämän takia mallien yksikkökohtaisia entropialukuja ei voi verrata suoraan toisiinsa. Jos malli esimerkiksi käyttäisi yksiköinään kirjaimia, niin jokainen yksikkö olisi keskimäärin suhteellisen helppo ennustaa, mutta koko datan todennäköisyydestä ei luultavasti tulisi kovin hyvä.

   Sen sijaan voimme laskea malleille entropiat sanaa kohti. Risti-entropia testidatalle D voitiin laskea

$$H_M(D) = \frac{1}{n} \sum_{i=1}^{n} \log P_M(D_i) = \frac{1}{n} \log P_M(D). \tag{1}$$

   Jos yksiköiden määrän n sijaan jaamme datan uskottavuuden $P_M(D)$ logaritmin datan sanojen määrällä W_D, saamme normalisoidun entropian, joka ei riipu siitä minkälaisilla yksiköillä mallinnus tehdään:

$$H_M^W(D) = \frac{1}{W_D} \log P_M(D). \tag{2}$$

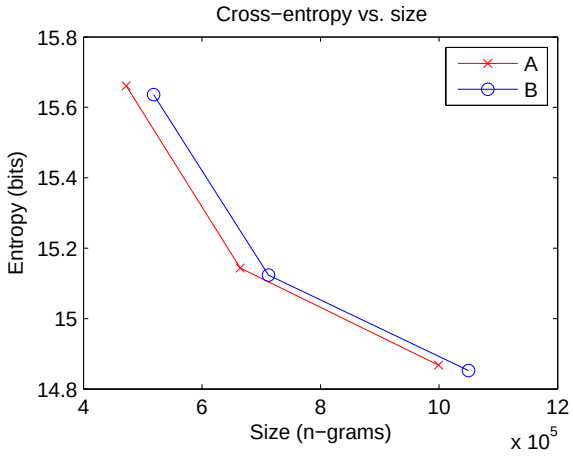
   Tiedämme luvut $H_M(D)$, n ja W_D, joiden avulla sanakohtaiseksi normalisoitu entropia voidaan kirjoittaa muodossa

$$H_M^W(D) = \frac{n}{W_D} \log H_M(D). \tag{3}$$

   Muutetaan annetut entropiat sanakohtaisiksi:

$$\begin{aligned}
H_{A1}^W(D) &= \frac{344\,960}{100\,000} \cdot 4.54 = 15.66 \\
H_{A2}^W(D) &= \frac{344\,960}{100\,000} \cdot 4.39 = 15.14 \\
H_{A3}^W(D) &= \frac{344\,960}{100\,000} \cdot 4.31 = 14.87 \\
H_{B1}^W(D) &= \frac{301\,271}{100\,000} \cdot 5.19 = 15.64 \\
H_{B2}^W(D) &= \frac{301\,271}{100\,000} \cdot 5.02 = 15.12 \\
H_{B3}^W(D) &= \frac{301\,271}{100\,000} \cdot 4.93 = 14.85
\end{aligned}$$

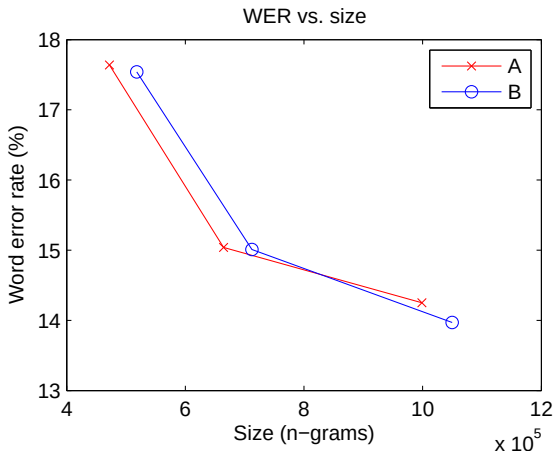
   Pilkonnan B entropiat ovat siis kaikissa kokoluokissa hieman parempi kuin pilkonnan A. Pitää kuitenkin ottaa huomioon myös mallien koot, jotka pilkonnalla B olivat kauttaaltaan suurempia. Tulosten vertailu on helpointa piirtämällä kuvaaja, jossa tulokset on esitetty koko-entropia -koordinaatistossa. Tämä on tehty kuvassa 3.



Kuva 3: Normalisoidut risti-entropiat.

Pilkonnan A mittauspisteitä yhdistävä murtoviiva on joka paikassa B:n viivasta katsottuna vasemmalla puolella. Pilkonta A näyttää siis antavan hieman parempia tuloksia kuin B suhteessa kielimallien kokoihin.

Katsotaan seuraavaksi tunnistustuloksia. Ne on laskettu suoraan sanoja kohti, joten normalisointeja ei tarvita. Mallin koko vs. sanavirhe -kuvaaja on esitetty kuvassa 4. Siitä nähdään, että tulokset isoilla ja pienillä malleilla menevät ristiin: A on parempi pienillä malleilla, mutta B ohittaa sen kun koot kasvavat yli 900 000 n-grammin.



Kuva 4: Sanavirheet.

Näyttää kohtuullisen selvältä, että pilkontaa A käyttävät mallit ovat parempia jos mallin koko on pieni. Suuremmilla malleilla tulokset ovat melko lähellä toisiaan. Lisäksi toiminnasta malleilla joiden koot ovat selvästi alle puoli miljoonaa, tai selvästi yli miljoonan, nämä tulokset eivät kerro mitään. Luotettavammat tulokset vaatisivat lisää mittauspisteitä sekä lukujen tilastollisten merkittävyyksien testaamista (esim. Wilcoxon signed-rank test).