T-61.281 Luonnollisten kielten tilastollinen käsittely

Vastaukset 7, ti 9.3.2004, 8:30-10:00 Sanaluokkien merkitseminen, Versio 1.0

1. a) Siirtymätodennäköisyyksien $P(t^k|t^j)$ suurimman uskottavuuden estimaatti on $\frac{C(t^j,t^k)}{C(t^j)}$. Siis esimerkiksi todennäköisyys

$$P(NN|AT) = \frac{48636}{48636 + 19} = 0.9996$$

Kun parametrien määrä on pieni ja dataa on paljon, suurimman uskottavuuden estimaatit ovat riittävän hyvät. Taulukkoon 1 on laskettu siirtymätodennäköisyys kaikkin sanaluokkien välille.

	Toinen sanaluokka					
Eka sanaluokka	AT	BEZ	IN	NN	VB	PISTE
AT	0	0	0	0.9996	0	0.0004
BEZ	0.7519	0	0.1623	0.0713	0	0.0145
IN	0.6971	0	0.0213	0.2786	0	0.0030
NN	0.0132	0.0459	0.5241	0.1453	0.0076	0.2640
VB	0.4337	0.0030	0.3399	0.1054	0.0092	0.1087
PISTE	0.5333	0.0050	0.3098	0.0884	0.0635	0

Taulukko 1: *Siirtymätodennäköisyydet*

Havaintotodennäköisyyksien $P(w^l|t^j)$ suurimman uskottavuuden estimaattori on $\frac{C(w^l,t^j)}{C(t^j)}$. Eli esimerkiksi todennäköisyys sille, että VB-tilassa olessa havaitaan sana bear on

$$P(\text{bear}|VB) = \frac{43}{43 + 133 + 4} = 0.2389$$

Taulukkoon 2 on merkitty pyydetyt havaintotodennäköisyydet.

	AT	BEZ	IN	NN	VB	PISTE
bear	0	0	0	0.0187	0.2389	0
is	0	1	0	0	0	0
move	0	0	0	0.0672	0.7389	0
on	0	0	1	0	0	0
president	0	0	0	0.7127	0	0
progress	0	0	0	0.2015	0.0222	0
the	1	0	0	0	0	0
.	0	0	0	0	0	1

Taulukko 2: *Havaintotodennäköisyydet*

b) Tehtävässä pyydettiin laskemaan tilojen $\mathbf{Q}_1 = \{\mathrm{AT}, NN, BEZ, IN, AT, NN\}$ ja $\mathbf{Q}_2 = \{\mathrm{AT}, NN, BEZ, IN, AT, VB\}$ todennäköisyyksien suhde kun tiedetään havainnot $\mathbf{O} = \{The, bear, is, on, the, move, .\}$. Merkitään juuri laskettuja mallin parametreja λ:lla.

$$\begin{aligned} P(\mathbf{Q}_i|\mathbf{O},\lambda) &= \frac{P(\mathbf{Q}_i,\mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda)} \\ &= \frac{P(\mathbf{Q}_i,\mathbf{O}|\lambda)}{\sum_{\text{kaikki } \mathbf{Q}_j} P(\mathbf{Q}_j,\mathbf{O}|\lambda)} \\ &= \frac{P(\mathbf{Q}_i|\lambda)P(\mathbf{O}|\mathbf{Q}_i,\lambda)}{\sum_{\text{kaikki } \mathbf{Q}_j} P(\mathbf{Q}_j|\lambda)P(\mathbf{O}|\mathbf{Q}_j,\lambda)} \end{aligned}$$

Yhtälöä on pyöritetty perustodennäköisyyskaavojen avulla. Huomataan, että yhtälön yläpuoli on helppo laskea: Tiedämme tilasekvenssin, joten $P(\mathbf{Q}_i|\lambda)$ ei tuota vaikeuksia kuten ei myöskään $P(\mathbf{O}|\mathbf{Q}_i,\lambda)$. Nimittäjän normalisointitermi on hankalampi, meidän pitäisi siis laskea yhteen kaikkien mahdollisten tilasekvenssejen todennäköisyydet kun tiedämme havainnot. Tämä voitaisiin laskea tehokkaasti eteenpäin-algoritmilla (katso edellinen laskari). Normalisointitermi on kuitenkin vakio kaikille sekvensseille ja koska kysymyksessä pyydettiin vain kahden sekvenssin todennäköisyyksien suhdetta, meidän ei tässä tarvitse siis ratkaista vakion arvoa ollenkaan.

Ratkaistaan todennäköisyyksien suhde:

$$\frac{P(\mathbf{Q}_1|\mathbf{O},\lambda)}{P(\mathbf{Q}_2|\mathbf{O},\lambda)} = \frac{P(\mathbf{Q}_1|\lambda)}{P(\mathbf{Q}_2|\lambda)} \cdot \frac{P(\mathbf{O}|\mathbf{Q}_1,\lambda)}{P(\mathbf{O}|\mathbf{Q}_2,\lambda)}$$

Ratkaistaan ensin ensimmäinen termi:

$$\begin{aligned} &\frac{P(\mathbf{Q}_1|\lambda)}{P(\mathbf{Q}_2|\lambda)} \\ = &\frac{P(\text{AT|PISTE})P(\text{NN|AT})P(\text{BEZ|NN})P(\text{IN|BEZ})P(\text{AT|IN})P(\text{NN|AT})P(\text{PISTE|NN})}{P(\text{AT|PISTE})P(\text{NN|AT})P(\text{BEZ|NN})P(\text{IN|BEZ})P(\text{AT|IN})P(\text{VB|AT})P(\text{PISTE|VB})} \\ = &\frac{P(\text{NN|AT})P(\text{PISTE|NN})}{P(\text{VB|AT})P(\text{PISTE|VB})} = \frac{0.9996 \cdot 0.2640}{0 \cdot 0.1087} = \infty \end{aligned}$$

Huomataan, että pelkästään siirtymätodennäköisyyksiä katselemalla sanaluokkasekvenssi $\mathbf{Q}_1$ on äärettömän paljon todennäköisempi. $\mathbf{Q}_2$:den todennäköisyyshän on nolla.

Katsellaan vielä havaintotodennäköisyyksiä, ettei sieltä tule mitään yllätyksiä

sekoittamaan edellisen kohdan perusteella tehtyjä päätelmiä:

$$\begin{aligned}
& \frac{P(\mathbf{O}|\mathbf{Q}_1,\lambda)}{P(\mathbf{O}|\mathbf{Q}_2,\lambda)} \\
= \; & \frac{P(\text{The}|\text{AT})P(\text{bear}|\text{NN})P(\text{is}|\text{BEZ})P(\text{on}|\text{IN})P(\text{the}|\text{AT})P(\text{move}|\text{NN})P(.|\text{PISTE})}{P(\text{The}|\text{AT})P(\text{bear}|\text{NN})P(\text{is}|\text{BEZ})P(\text{on}|\text{IN})P(\text{the}|\text{AT})P(\text{move}|\text{VB})P(.|\text{PISTE})} \\
= \; & \frac{P(\text{move}|\text{NN})}{P(\text{move}|\text{VB})} = \frac{0.0684}{0.7389} \approx 0.09
\end{aligned}$$

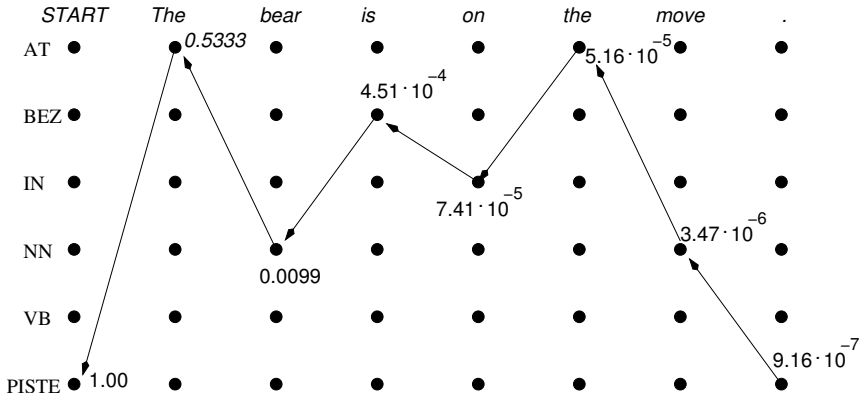
Koska $0.09 \cdot \infty = \infty$, mallimme mukaan on äärettömän paljon todennäköisempää, että sana *move* on nomini (NN) tässä lausessa.

c) Viterbi–haku on tarkemmin käsitelty edellisessä laskarissa. Alustetaan algorimi niin, että lähtötila on edellisen lauseen lopetuspiste: $\delta_{START}(PISTE) = 1$ ja $\delta_{START}(x) = 0$ kun $x \neq PISTE$.

Lasketaan tässä esimerkiksi muutaman tilan arvo. Koska sana *the* voi tulla vain tilasta AT (koska $P(the|AT) = 1$ ja $P(the|x) = 0$ kun $x \neq AT$), täytyy ensimmäisen tilan siis olla AT. Siihen kertyvä todennäköisyys on $\delta_{START}(PISTE) \cdot P(AT|PISTE) \cdot P(The|AT) = 1 \cdot 0.5333 \cdot 1$.

Sana *bear* voi tulla tiloista NN tai VB, mutta tilasta AT voidaan siirty vain tilaan NN. Seuraavan tilan täytyy siis olla NN. Siihen on kertynyt todennäköisyyttä $\sigma_{The}(AT)P(NN|AT)P(bear|NN) = 0.5333 \cdot 0.9996 \cdot 0.0187 = 0.0099$.

Samalla tavalla jatkamalla huomataan, että on ainoastaan yksi polku, joka johtaa alusta loppuun. Muissa kohdissa joko havaintomatriisin tai siirtymämatriisin nollat katkaisevat reitin. Paras polku on piirretty kuvaan 1. Siitä voidaan lukea, että paras tilasekvenssi $Q_{paras} = \{AT, NN, BEZ, IN, AT, NN, PISTE\}$.



Kuva 1: *Viterbi–haku. Matriisien harvuuden vuoksi (paljon nollia) ei hilaan tule juurikaan yli nollatodennäköisyydellä tapahtuvia siirtymiä. Siirtymiä, joiden todennäköisyys on nolla, ei ole piirretty.*

2. Käydään annetut säännöt läpi yksi kerrallaan. Ensimmäinen sääntö muuttaa VB:n NN:ksi, jos kaksi paikkaa sitä ennen on IN. Huomataan, että sääntö sopii joka lauseen viimeiseen sanaan. Pitää siis muuttaa

box: VB → NN

board: VB → NN

crash: VB → NN

Seuraava sääntö sopii vain toiseen lauseeseen. Muutetaan

marked: VBD → VBN

Kolmas sääntö sopii vain jos sana *wanted* on luokiteltu *RD*:ksi ja sitä seuraa TO. Tämä sopii ensimmäiseen lauseeseen ja muutetaan siis

wanted: RB → VBD

Viimeinen sääntö sopii vain ensimmäiseen lauseeseen:

look: NN → VB

Korjatut lauseet ovat siis:

I)	PN	VBD	TO	VB	IN	AT	NN
	I	wanted	to	look	inside	the	box
II)	PN	BEZ	RB	VBN	IN	AT	NN
	It	was	clearly	marked	on	the	board
III)	AT	NN	PN	BEZ	AT	VB	NN
	The	plane	he	is	on	will	crash

Kun katsellaan lopputulosta, huomataan että muuten kaikki sujui hyvin, mutta viimeisen lauseen viimeinen sana muutettiin turhaan verbistä nominiksi. Tämän turhan muutoksen voisi ohittaa muuttamalla ensimmäistä sääntöä niin, että ajanhetkellä $t-1$ pitää löytyä artikkeli (AT).