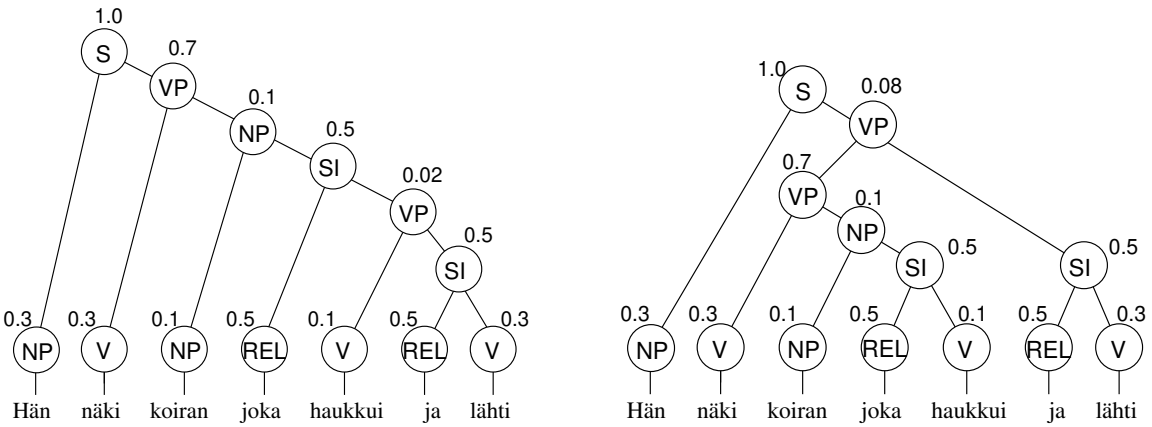
T-61.281 Luonnollisen kielen tilastollinen käsittely

Vastaukset 8, ti 18.3.2002, 16:15-18:00 Tilastolliset yhteydettömät kieliopit, Versio 1.0

1. Jäsennyspuun todennäköisyys lasketaan paloittelemalla se säännöstön yksiköiden kokoisiin paloihin ja kertomalla näiden palojen todennäköisyydet yhteen. Esimerkiksi muunnoksen ($NP \rightarrow$ hän) todennäköisyys voidaan suoraan poimia säännöstöstä ja on 0.3. Samoin esimerkiksi muunnos ($SI \rightarrow REL\ V$) on 0.5. Kuvaan 1 on poimittu taulukoidut todennäköisyydet.



Kuva 1: *Jäsennykset*

Nyt voimmekin laskea kummankin jäsennyksen todennäköisyydet kertomalla kaikki kuvasta löytyvät todennäköisyydet

$$\begin{aligned} P(\text{puu1}) &= 0.3 \cdot 0.3 \cdot 0.1 \cdot 0.5 \cdot 0.1 \cdot 0.5 \cdot 0.3 \cdot 0.5 \cdot 0.02 \cdot 0.5 \cdot 0.1 \cdot 0.7 \cdot 1.0 \\ &= 2.4 \cdot 10^{-8} \end{aligned}$$

$$\begin{aligned} P(\text{puu2}) &= 0.3 \cdot 0.3 \cdot 0.1 \cdot 0.5 \cdot 0.1 \cdot 0.5 \cdot 0.3 \cdot 0.5 \cdot 0.5 \cdot 0.1 \cdot 0.7 \cdot 0.08 \cdot 1.0 \\ &= 9.5 \cdot 10^{-8} \end{aligned}$$

Huomataan, että mallin mukaan toinen jäsennys on todennäköisempi. Ilman muuta tietoa tai oikein pilkutettua tekstiä lienee mahdotonta päätellä, kumpi jäsennys on oikea.

2. Sisäpuoli-algoritmi (inside algorithm) on hyvin samanlainen kuin eteenpäin algoritmi. Algoritmissa lasketaan puun todennäköisyyttä lähtemällä lehdistä ja kasaamalla koko ajan suurempia yksikköjä kunnes päästään juureen asti. Algoritmia läpikäydessä tullaan samalla kokeilleeksi kaikki mahdolliset jäsennykset.

   Merkitään $\beta_j(p, d)$ todennäköisyyttä, että puulla, joka kattaa sanat p:stä d:hen on juurena jäsennys j. Nyt siis voimme alustaa algoritmin $\beta_j(k, k)$ arvot lehdille. Koska

suurin osa lauseen sanoista voi olla kotoisin vain yhdestä ei-terminaalisymbolista, alustus on helppoa:

$$\begin{aligned}
\beta_{NP}(1,1) &= 0.3 && \text{Todennäköisyys, että ensimmäisen sanan tilana on NP ja havaitaan sana “hän”} \\
\beta_{V}(2,2) &= 0.29 && \text{Todennäköisyys, että toisen sanan ilana on V ja havaitaan sana “tunsi”} \\
\beta_{NP}(3,3) &= 0.15 && \text{3. sanan tila NP ja havainto “tuulen”} \\
\beta_{V}(3,3) &= 0.01 && \text{3. sanan tila V ja havainto “tuulen”} \\
\beta_{A}(4,4) &= 0.15 && \text{4. sanan tila A ja havainto “kalpeilla”} \\
\beta_{NP}(5,5) &= 0.15 && \text{5. sanan tila NP ja havainto “kasvoillaan”}
\end{aligned}$$

Kaikki muut todennäköisyydet ovat nollia. Alustuksesta voidaan edetä puun juureen summaamalla kaikkien todennäköisyyksin yli seuraavan kaavan mukaisesti:

$$\beta_j(p,q) = \sum_{r,s} \sum_{d=p}^{q-1} P(N^j \rightarrow N^r N^s)\beta_r(p,d)\beta_s(d+1,q)$$

Tässä siis summataan kaikkien mahdollisten sääntöjen (r,s) ja kaikkien mahdollisten jakojen (d:stä q:hun) yli. Koska yleensä sekä transitio-, että havaintosäännöt ovat melko harvoja, on suurin osa summan termeistä nollia.

Lasketaanpa sitten seuraavat arvot. Nyt kukin alipuulla koostuu kahdesta lapsesta ja juuresta:

$$\begin{aligned}
\beta_x(1,2) &= \sum_{y,z} \beta_y(1,1) \cdot \beta_z(2,2) \cdot P(x \rightarrow y\ z) = 0 && \text{Todennäköisyys,että sanat “hän”, “tunsi”} \\
& && \text{kattavan puun juuren sanaluokka olisi x} \\
\beta_{VP}(2,3) &= \beta_V(2,2) \cdot \beta_{NP}(3,3) \cdot P(VP \rightarrow V\ NP) && \text{Kaikki muut summan termit ovat nollia} \\
&= 0.29 \cdot 0.15 \cdot 0.7 = 0.030 \\
\beta_x(3,4) &= 0 \quad \forall x \\
\beta_{PP}(4,5) &= \beta_A(4,4)\beta_{NP}(5,5)P(PP \rightarrow A\ NP) \\
&= 0.15 \cdot 0.15 \cdot 1.0 = 0.023
\end{aligned}$$

Koskapa kieliopissa yhdellä solmulla olla vain kaksi lasta, saadaan seuraavalle kierrokselle todennäköisyydet summaamalla:

$$\begin{aligned}
\beta_S(1,3) &= \beta_{NP}(1,1)\beta_{VP}(2,3)P(S \rightarrow NP\ VP) + \sum_{x,y} \beta_x(1,2)\beta_y(3,3)P(S \rightarrow x\ y) \\
&= 0.3 \cdot 0.030 \cdot 1.0 + 0 = 0.0090 \\
\beta_x(2,4) &= 0 \quad \forall x \\
\beta_{NP}(3,5) &= \beta_{NP}(3,3)\beta_{PP}(4,5)P(NP \rightarrow NP\ PP) \\
&= 0.15 \cdot 0.023 \cdot 0.15 = 5.2 \cdot 10^{-4}
\end{aligned}$$

Vielä on pari kierrosta jäljellä:

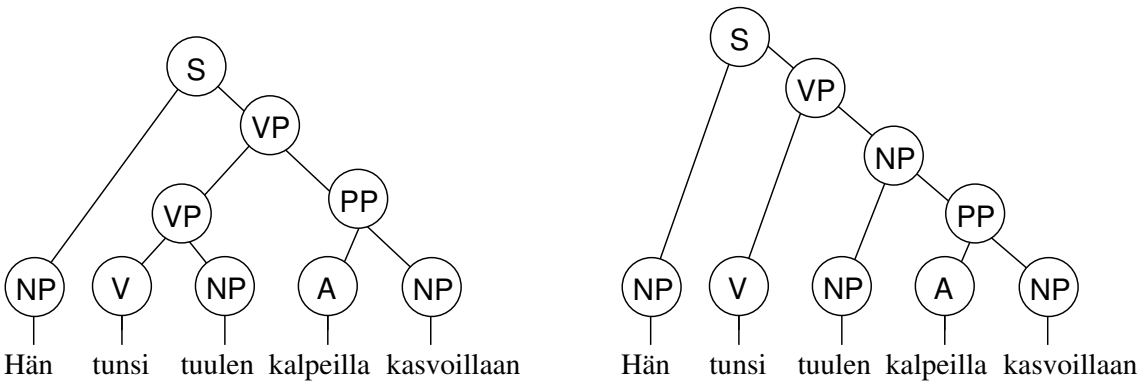
$$
\begin{aligned}
\beta_x(1,4) &= \sum_{y,z} \beta_y(1,1)\beta_z(2,4)P(x \to y\ z) + \sum_{y,z} \beta_y(1,2)\beta_z(3,4)P(x \to y\ z) \\
&\quad + \sum_{y,z} \beta_y(1,3)\beta_z(4,4)P(x \to y\ z) \\
&= 0 \\
\beta_{VP}(2,5) &= \beta_V(2,2) \cdot \beta_{NP}(3,5) \cdot P(VP \to V\ NP) \\
&\quad + \beta_{VP}(2,3) \cdot \beta_{PP}(4,5) \cdot P(VP \to V\ PP) \\
&= 0.29 \cdot 5,2 \cdot 10^{-4} \cdot 0.7 + 0.03 \cdot 0.023 \cdot 0.2 = 1.1 \cdot 10^{-4} + 1.4 \cdot 10^{-4} \\
&= 2.43 \cdot 10^{-4}
\end{aligned}
$$

Tässä kohtaa kannattaa huomata, että laskettaessa $\beta_{VP}(2,5)$, yhdistetään kahden erilaisen jäsennyksen todennäköisyydet. Jos nyt tehtäisi Viterbi-hakua, valittaisiin näistä kahdesta todennäköisempi ja merkattaisiin, kumpi jäsennys on parempi. Tässä tapauksessa siis on parempi jäsentää sanat "tunsi" ja "tuulen" verbilauseeksi ja sanat "kalpeilla kasvoillaan" liittolauseeksi. Toinen vaihtoehto olisi ollut jäsentää sana "tunsi" verbiksi ja "tuulen kalpeilla kasvoillaan" liitolauseeksi.

Lasketaanpa vielä homma loppuun ennen tarkempia pohdintoja.

$$
\begin{aligned}
\beta_S(1,5) &= \beta_{NP}(1,1) \cdot \beta_{VP}(2,5) \cdot P(S \to NP\ VP) \\
&= 0.3 * 2.43 \cdot 10^{-4} * 1.0 = 7.2 \cdot 10^{-5}
\end{aligned}
$$

Kuvaan 2 on piirretty mahdolliset jäsennykset. Mallin mukaan hieman todennäköisempi jäsennys on väärä, johtunee siitä että alkuperäinen malli oli hihasta ravistettu.



Kuva 2: *Mahdolliset jäsennykset. Vasen puu on hiukan todennäköisempi.*

3. Merkitään $q = 1 - p$. Olkoon $F(n)$ todennäköisyysmassa kaikissa n-syvyisissä tai matalammissa puissa (katso kuva 3). Nyt huomataan, että $F(n+1)$ voidaan ilmaista

F(1)	0	S
F(2)	q	S | w
F(3)	$q+pq^2$	S | w + S | w / \ S S | | w w
F(4)	$q+pF(3)^2$	S | w + S / \ F(3) F(3)
F(n+1)	$q+pF(n)^2$	S | w + S / \ F(n) F(n)

Kuva 3: *$F(n)$ eli kaikkiin n-syvyisiin tai matalampiin puihin uponnut todennäköisyysmassa.*

$F(n)$:n avulla. Joko $F(n+1)$ terminoituu suoraan todennäköisyydellä q tai sitten se jakautuu kahdeksi puuksi todennäköisyydellä p. Näiden puiden maksimisyvyys on n. Saadaan siis

$$F(n+1) = q + pF(n)^2 = 1 - p + pF(n)^2$$

Oletetaan, että sarja konvergoi johonkin arvoon (todistetaan myöhemmin). Tällaisessa tilanteessa $F(n+1) = F(n)$ eli

$$\begin{aligned} F(n+1) = 1 - p + pF(n)^2 &= F(n) \\ pF(n)^2 - F(n) + 1 - p &= 0 \\ px^2 - x + 1 - p &= 0 \end{aligned}$$

Viimeisellä rivillä on $F(n)$ merkitty lyhyemmin x. Toisen asteen yhtälön ratkaisukaavasta saadaan yhtälölle juuret

$$x = \begin{cases} 1 \\ \frac{1}{p} - 1 \end{cases}$$

Jotta voidaan todistaa, että funktio konvergoi, pitää todistaa, että se on kasvava. Tutkitaan erotusta $F(n+1) - F(n)$:

$$\begin{aligned} F(n+1) - F(n) &> 0 \\ 1 - p + pF(n)^2 - F(n) &> 0 \end{aligned}$$

Toisen asteen yhtälön ratkaisukaavan avulla saadaan, että funktio on kasvava jos:

$$\begin{aligned} 1 &: x > 1 \text{ ja } p < 0 \\ 2 &: x < 1 \text{ ja } p > 0 \end{aligned}$$

Näistä toinen alue on relevantti meille. Osoitetaan vielä, että funktio kasvaa asymptoottisesti kohti pienempää juurta ($\min(1, \frac{1}{p} - 1)$). Jos $F(n+1) > 1$, mitä ehtoja se asettaa $F(n)$:lle ?

$$\begin{aligned} F(n+1) = 1 - p + pF(n)^2 &> 1 \\ pF(n)^2 &> p \\ F(n) < -1 \quad \text{tai} \quad F(n) &> 1 \end{aligned}$$

Eli sarja voi saada yhtä suurempia arvoja vain, jos sarjan edellinen arvo oli jo suurempi kuin yksi. Nyt lähdemme liikkeelle nollasta $F(1) = 0$, joten yli yhden arvoja ei voida saada. Entäpä toinen juuri, mitä jos $F(n+1) > \frac{1}{p} - 1$? Merkitään jatkossa $F(n) = x$.

$$\begin{aligned} F(n+1) = 1 - p + px^2 &> \frac{1}{p} - 1 \\ x^2 &> \frac{1 - 2p + p^2}{p^2} \\ x^2 &> \left(\frac{1-p}{p}\right)^2 \end{aligned}$$

Saadaan ehdot

$1 : x > \dfrac{1}{p} - 1$ edellinen arvo oli jo suurempi kuin juuri

$2 : x < 1 - \dfrac{1}{p}$ F kasvava ja $F(2) = 1 - p < 1 - \dfrac{1}{p}$ kun $p < -1$ tai $p > 1$. Ehto ei täyty.

Myöskään tämän juuren ohi ei voi päästä. Sarjan summa on siis $\min(1, \frac{1}{p} - 1)$. Todennäköisyysjakauma on siis oikea todennäköisyysjakauma, kun $p \leq 0.5$. Kun $p > 0.5$, malli rupeaa generoimaan äärettömiä puita (missä generoidaan aina vaan uusia ei-terminaaleja) ja todennäköisyysmassaa katoaa näihin puihin.

Jos olisi ovelampi matemaatikko, tehtävän pystyisi ratkaisemaan muutamalla rivillä. Tarkastelemalla mallin generoimien ei-terminaalien määrän oletusarvoa verrattuna mallin generoimien terminaalien määrän oletusarvoon, päätynee samaan tulokseen.