T-61.281 Luonnollisten kielten tilastollinen käsittely

Vastaukset 2, ti 4.2.2003, 16:15-18:00 – Entropia, hämmentyneisyys, kontekstivapaa kieli, Versio 1.1

1. a) Sijoitetaan entropian kaavaan

$$H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

tehtävässä annetut arvot:

$$\begin{aligned} H(X) &= \frac{3}{32}\log_2\frac{32}{3} + \frac{3}{16}\log_2\frac{16}{3} + \frac{7}{32}\log_2\frac{32}{7} \\ &\quad + \frac{1}{8}\log_2 8 + \frac{1}{8}\log_2 8 + \frac{1}{4}\log_2 4 \\ &= 2.50 \text{ bittiä} \end{aligned}$$

b) Lähteen entropia, kun tiedetään, että edellinen symboli kuului joukkoon S on

$$H(X_i|X_{i-1} \in S) = \sum_{S=\{\text{'kissa','tuuli','kiipeilijä'}\}} p(s = S) H(V|s = S)$$

Tämän laskemiseksi meidän pitää osata laskea ehdollinen entropia $H(V|s)$. Jos $s =$"kissa", todennäköisyys, että sitä seuraa sana "naukaisi" on $\frac{2}{3}$ ja todennäköisyys, että sitä seuraa sana "katosi" on $\frac{1}{3}$. Kaikkien vaihtoehtojen yli summattuna todennäköisyydenhän pitää olla yksi, eli taulukossa annettuja todennäköisyyksiä joudutaan hieman skaalaamaan, tässä tapauksessa vakiolla $\frac{16}{3}$. Tällaisen lähteen entropiahan on

$$H(V|s = \text{'kissa'}) = \frac{2}{3}\log_2(\frac{3}{2}) + \frac{1}{3}\log_2(3)$$

Kun sijoitamme jokaista joukon S sanaa vastaavat todennäköisyydet, saamme

$$\begin{aligned} H(X_i|X_{i-1} \in S) &= \frac{3}{16}(\frac{2}{3}\log_2\frac{3}{2} + \frac{1}{3}\log_2 3) + \frac{3}{8}(\frac{1}{6}\log_2 6 + \frac{4}{6}\log_2\frac{6}{4} + \frac{1}{6}\log_2 6) \\ &\quad + \frac{7}{16}(\frac{1}{7}\log_2 7 + \frac{6}{7}\log_2\frac{7}{6}) \\ &= 0.90 \text{ bittiä} \end{aligned}$$

Huomataan, että kun tunnemme lähteen toiminnan paremmin, sen tuottamat sanat ovat vähemmän yllättäviä ja voimme koodata ne vähemmällä määrällä bittejä.

2. a) Kunkin alkeistapauksen todennäköisyys on $\frac{1}{30}$. Alkeistapauksia on 30. Sijoitetaan entropian kaavaan:

$$\begin{aligned} H(X) &= \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} \\ &= \sum_{i=1}^{30} \frac{1}{30}\log_2(30) \\ &= \log_2(30) \approx 4.9 \text{ bittiä} \end{aligned}$$

b) Sanan, jossa on vain yksi merkki, sanotaan vaikka joukon ensimmäinen merkki, todennäköisyys on

$$P(s = t_1) = \frac{1}{30} \cdot \frac{1}{30}$$

sillä ensimmäisen merkin pitää olla joukon ensimmäinen ja sitten pitää tulla sanaväli. Tällaisia sanoja on 29 kappaletta.

Vastaavasti, tietyn kahden merkin pituisen sanan todennäköisyys on

$$P(s = t_1, t_1) = \frac{1}{30} \cdot \frac{1}{30} \cdot \frac{1}{30}$$

Tällaisia sanoja on 29^2 kappaletta. Homma jatkuu samalla tavalla useammille sanoille.

Lasketaan tällaisen lähteen entropia:

$$\begin{aligned}
H(X) = & \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} \\
= & \ 29 * (\frac{1}{30})^2 \log_2(30^2) + 29^2 * (\frac{1}{30})^3 \log_2(30^3) + 29^3 * (\frac{1}{30})^4 \log_2(30^4) + \ldots \\
= & \ \frac{1}{29} \left(\left(\frac{29}{30}\right)^2 \cdot 2 \cdot \log_2(30) + \left(\frac{29}{30}\right)^3 \cdot 3 \cdot \log_2(30) + \left(\frac{29}{30}\right)^4 \cdot 4 \cdot \log_2(30) + \ldots \right) \\
= & \ \frac{\log_2(30)}{29} \left(-\frac{29}{30} + \sum_{i=0}^{\infty} i \cdot \left(\frac{29}{30}\right)^i \right)
\end{aligned}$$

Nyt tarvitaan suluissa olevan summan arvoa. Ratkaistaan annettu sarja seuraavasti:

$$\sum_{i=0}^{\infty} iq^i = q + 2q^2 + 3q^3 + 4q^4 + \ldots \tag{1}$$

Kerrotaan yhtälö q:lla.

$$q \sum_{i=0}^{\infty} iq^i = q^2 + 2q^3 + 3q^4 + 4q^5 + \ldots \tag{2}$$

Vähennetään yhtälö 2 puolittain yhtälöstä 1

$$(1-q) \sum_{i=0}^{\infty} iq^i = q + q^2 + q^3 + q^4 + \ldots \tag{3}$$

$$\sum_{i=0}^{\infty} iq^i = \frac{q + q^2 + q^3 + q^4 + \ldots}{1-q} \tag{4}$$

Kerrotaan yhtälö 4 vielä kerran q:lla

$$q \sum_{i=0}^{\infty} iq^i = \frac{q^2 + q^3 + q^4 + q^5 + \ldots}{1-q} \tag{5}$$

Nyt vähennetään yhtälöt 4 ja 5 toisistaan ja saadaan ratkaisu:

$$(1-q) \sum_{i=0}^{\infty} iq^i = \frac{q}{1-q} \tag{6}$$

$$\sum_{i=0}^{\infty} iq^i = \frac{q}{(1-q)^2} \tag{7}$$

Kun tämä hässäkkä sijoitetaan alkuperäiseen ongelmaan, saadaan

$$\begin{aligned}
& \frac{\log_2(30)}{29}\left(-\frac{29}{30} + \frac{\frac{29}{30}}{(1-\frac{29}{30})^2}\right) \\
= \; & \log_2(30)\left(30 - \frac{1}{30}\right) \\
= \; & 147 \text{ bittiä}
\end{aligned}$$

Ensi silmäyksellä tämä tulos saattaa tuntua hämmentävältä, eikä tuloksen pitäis olla sama kuin a)-kohdassa ? Pikainen likimainen matematiikka ehkä hälventää hieman epäluuloja: Sanan pituuden oletusarvo on 29, eli entropia per merkki on n. 147/29=5.0 bittiä. Korostettakoon vielä, että tämä viimeinen lasku on vain karkea approksimaatio.

On myös syy, miksi tulosten ei pitäisi olla aivan samat: Ensimmäinen lähde voi tuottaa sanan, jossa on kaksi välilyöntiä peräkkäin, kun taas toinen lähde ei voi annetun formuloinnin mukaan sitä tuottaa. Tästä johtuen pitäisi toisen lähteen entropia per merkki olla hieman alempi.

3. a) Merkitään mallin yksi antamaa hämmentyneisyyttä $Perp_1$, mallin 2 puolestaan $Perp_2$ ja niin edelleen.

$$\begin{aligned}
& Perp_1(\text{'kissa'}, \text{'menee'}, \text{'puuhun'}) \\
& = P_1(\text{sana}_1\text{='kissa'}, \text{sana}_2\text{='menee'}, \text{sana}_3\text{='puuhun'})^{-\frac{1}{3}} \\
& = (P_1(\text{sana='kissa'})P_1(\text{sana='menee'})P_1(\text{sana='puuhun'}))^{-\frac{1}{3}} \\
& = (0.1 \cdot 0.1 \cdot 0.1)^{-\frac{1}{3}} = 10
\end{aligned}$$

Malli 1 siis valitsee koko ajan keskimäärin kymmenestä eri sanasta. Tulos vaikuttaa oikealta. Entäpä malli 2 ?

$$\begin{aligned}
& Perp_2(\text{'kissa'}, \text{'menee'}, \text{'puuhun'}) \\
= \; & P_2(\text{sana}_1\text{=subjekti}, \text{sana}_2\text{=verbi}, \text{sana}_3\text{=kohde})^{-\frac{1}{3}} \\
= \; & (P_2(\text{sana=subjekti})P_2(\text{sana=verbi})P_2(\text{sana=kohde}))^{-\frac{1}{3}} \\
= \; & (0.33 \cdot 0.33 \cdot 0.33)^{-\frac{1}{3}} = 3
\end{aligned}$$

Malli 2 valitsee keskimäärin 3:sta eri vaihtoehdosta, tulos vaikuttaa järkevältä.

$$\begin{aligned}
& Perp_3(\text{'kissa'}, \text{'menee'}, \text{'puuhun'}) \\
= \; & P_3(\text{sana}_1\text{='kissa'}, \text{sana}_2\text{='menee'}, \text{sana}_3\text{='puuhun'})^{-\frac{1}{3}} \\
= \; & (P_3(\text{sana='kissa'}|\text{sana=ensimmäinen}) \\
& \cdot P_3(\text{sana='menee'} \mid \text{edellinen_sana} = \text{'kissa'}) \\
& \cdot P_3(\text{sana='puuhun'} \mid \text{edellinen_sana} = \text{'menee'}))^{-\frac{1}{3}} \\
= \; & (0.25 \cdot 0.33 \cdot 0.33)^{-\frac{1}{3}} = 3.32
\end{aligned}$$

Tämä malli valitsee siis keskimäärin 3.32 sanasta koko ajan.

Tämän esimerkin valossa kielimallit 1 ja 3 ovat vertailukelpoiset. Kielimalli 3 vaikuttaa näistä selvästi paremmalta. Kielimalli 2 ei voi verrata muihin, sillä se operoi selvästi pienemmällä symbolijoukolla. Selvempi esimerkki olisi ehkä kielimalli, jonka mielestä kaikki sanat kuuluvat ryhmään 1 ja tämän ryhmän todennäköisyys on siis 1. Tämä kielimalli siis hämmentyneisyyden mukaan täydellinen, sillä se ei ole yhtään yllättynyt mistään sanasta.

b) Tarkastellaanpa vielä toista testilausetta. Mallille 1

$$\begin{aligned}
& Perp_1(\text{'valas'}, \text{'on'}, \text{'kala'}, \text{'paitsi'}, \text{'ettei'}) \\
= \; & (P_1(\text{sana='valas'})P_1(\text{sana='on'})P_1(\text{sana='kala'}) \\
& \cdot P_1(\text{sana='paitsi'})P_1(\text{sana='ettei'}))^{-\frac{1}{5}} \\
= \; & (0.1 \cdot 0.1 \cdot 0.1 \cdot 0 \cdot 0)^{-\frac{1}{5}} \\
= \; & \frac{1}{0^{\frac{1}{5}}} = \infty
\end{aligned}$$

Huomataan, ettei hämmentyneisyyttä voida laskea, jos malli asettaa testijoukon sanalle todennäköisyyden nolla. Usein nämä sanat jätetään huomiotta ja saadaan siis

$$\begin{aligned}
& Perp_1(\text{'valas'}, \text{'on'}, \text{'kala'}) \\
= \; & (P_1(\text{sana='valas'})P_1(\text{sana='on'})P_1(\text{sana='kala'}))^{-\frac{1}{3}} = 10
\end{aligned}$$

Jotta tulos olisi mielekäs, on nyt myös ilmoitettava ohi kieliopin menneet sanat, tässä tapauksessa siis $\frac{2}{5} \cdot 100\% = 40\%$ sanoista ei osunut kielioppiin. Mallille 2 sadaan vastaavasti

$$\begin{aligned}
& Perp_2(\text{'valas'}, \text{'on'}) \\
= \; & (P_2(\text{sana=subjekti})P_2(\text{sana=verbi}))^{-\frac{1}{3}} \\
= \; & (0.33 \cdot 0.33)^{-\frac{1}{2}} = 3
\end{aligned}$$

Ohi kieliopin menee myös 60% sanoista.

Malliin kolme sopii vain kaksi ensimmäistä sanaa:

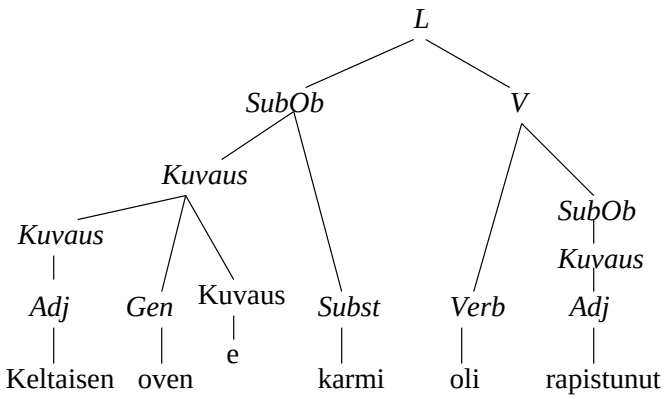
$$\begin{aligned}
& Perp_3(\text{'valas'}, \text{'on'}) \\
= \; & (P_3(\text{sana='valas'}|\text{sana=ensimmäinen}) \\
& \cdot P_3(\text{sana='on'} \mid \text{edellinen_sana} = \text{'valas'})))^{-\frac{1}{3}} \\
= \; & (0.25 \cdot 0.33)^{-\frac{1}{2}} = 3.5
\end{aligned}$$

Tässä siis 60% sanoista menee ohi kieliopin.

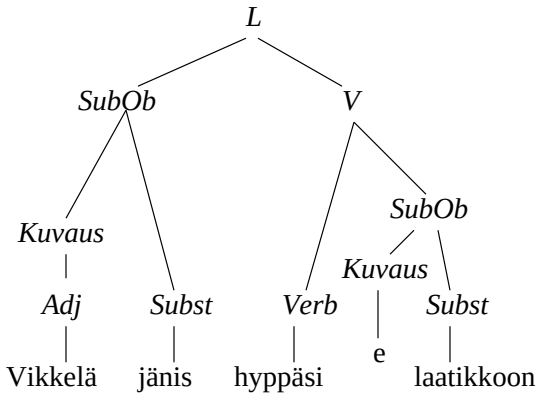
Ovatko b)-kohdan tulokset vertailukelpoisia ? Malli 2 voidaan diskata samoilla perusteilla kuin a)-kohdassakin. Malleja 1 ja 3 voidaan vertailla, kun otetaan myös huomioon ohi kieliopin menneet sanat. Malli 1 kattaa sanaston paremmin, mutta malli 3 antaa paremman hämmentyneisyyden. Usein kielimallin laatiminen on tasapainottelua näiden kahden ominaisuuden välillä.

Mikä siis on tarinan opetus ? Hämmentyneisyydellä voidaan verrata kahta kielimallia, jos tulokset lasketaan samalla tavalla ja myös ohi kieliopin menneitten sanojen osuus ilmoitetaan. Hämmentyneisyydellä voidaan myös ilmoittaa lähes millaisia tuloksia tahansa, jos laskuja on haluttu rukata johonkin suuntaan tai ohi sanaston menevää osuutta ei ilmoiteta. Kannattaa olla tarkkana ainakin eri lähteistä saatujen tulosten vertailussa.

4. Annetut lauseet on jäsennetty alhaalta ylöspäin. Kun säännöt eivät muuten sopineet, kokeiltiin, auttaisiko tyhjän symbolin "e" lisääminen jonnekin. Katso kuvat 1 ja 2. Järjetön lause, jonka kielioppi hyväksyy voisi olla: "Hyppäsi rapistunut oven." Kieliopilla ei ole säännöstöä, millä se pystyisi jäsentämään monimutkaisempia lauseita. Hylätty lause voisi olla: "Kieltolause ei onnistunut, kuten ei sivulausekaan."



Kuva 1: Jäsennys 1



Kuva 2: Jäsennys 2