T-61.5020 Luonnollisen kielen tilastollinen käsittely

Vastaukset 5, ke 21.2.2007, 12:15–14:00 — Kollokaatiot
Versio 1.0

1. Lasketaan ensin tulokset sanaparille "valkoinen","talo" käsin:

   – Frekvenssimenetelmä: Bigrammeja "valkoinen","talo" oli 710 kappaletta.

   – Normalisoitu frekvenssimenetelmä: Sana "valkoinen" esiintyi 3665 kertaa ja sana "talo" 10 767 kertaa. Vertailuluvuksi saadaan $\frac{710}{3665 \cdot 10767} \approx 1.8 \cdot 10^{-5}$.

   Kaikkien sanojen tulokset frekvenssimenetelmälle on esitetty taulukossa 1 ja normalisoidulle frekvenssimenetelmälle taulukossa 2.

   Taulukko 1: *Frekvenssimenetelmän tulokset*

s_1	s_2	$C(s_1, s_2)$
ja	olla	7329
venäjä	presidentti	717
valkoinen	talo	710
kova	tuuli	279
aste	pakkanen	160
tuntematon	sotilas	154
sekä	myös	138
liukas	keli	106
hakea	työ	31
oppia	lukea	21
ottaa	onki	9
vihainen	mielenosoittaja	7
olla	ula	5
heittää	veivi	3
herne	nenä	3

   Huomataan, että jo "hihasta ravistetuilla" menetelmillä päästään kohtalaisiin tuloksiin.

2. Lasketaan käsin malliksi tulos jo tutulle kollokaatiolla "valkoinen", "talo". Keskiarvo:

$$\begin{aligned} Mean(\text{“valkoinen”}, \text{”talo”}) &= \frac{-1 \cdot 710 - 2 \cdot 2 + 1 + 2 \cdot 6}{710 + 2 + 1 + 6} \\ &\approx -0.975 \end{aligned}$$

Taulukko 2: *Normalisoidun frekvenssimenetelmän tulokset*

s_1	s_2	Normalisoitu frekvenssi$\cdot 10^{-8}$
liukas	keli	1981
aste	pakkanen	386
heittää	veivi	293
herne	nenä	268
valkoinen	talo	180
tuntematon	sotilas	163
vihainen	mielenosoittaja	68
kova	tuuli	35
ottaa	onki	21
venäjä	presidentti	10
oppia	lukea	8
hakea	työ	1
olla	ula	0
sekä	myös	0
ja	olla	0

Varianssi

$$
\begin{aligned}
&Var(\text{“valkoinen”}, \text{”talo”}) \\
&= \frac{(-1-(-0.975))^2 \cdot 710 + (-2-(-0.975))^2 \cdot 2 + (1-(-0.975))^2 \cdot 1 + (2-(-0.975))^2 \cdot 6}{710+2+1+6} \\
&\approx \quad 0.083
\end{aligned}
$$

Lopuille sanoille tulokset on annettu varianssin mukaan järjestettynä taulukossa 3.

Taulukkoa tarkastellessamme huomaamme, että menetelmä on löytänyt käytännössä kaikki kiinteät kollokaatiot, kuten valkoinen talo. Menetelmä ei pärjää hyvin harvalla aineistolla, esim. vihainen mielenosoittaja ei selvästikään ole kollokaatio, vaikka menetelmä sen toiseksi sijoittaakin.

Tarkasteluikkunan leveys vaikuttaa tietysti alueeseen, josta kollokaatioita etsitään. Jos aluetta kasvatetaan liian suureksi, rupeavat sanat esiintymään yhä useammin myös satunnaisesti yhdessä ja varianssi kasvaa suureksi. Liian pienellä ikkunalla ei pidempivaikutteisia kollokaatioita löydetä. Jos kollokaation toinen sana voi olla sekä referenssisanan edessä että takana, menetelmä tietysti hämääntyy täydellisesti.

3. Tilastollisissa testeissä tehdään nollahypoteesi siitä että sanapari on toisistaan riippumaton ($P(s_1, s_2) = P(s_1)P(s_2)$), ja tästä poikkevat havainnot ovat sattuman aikaansaamia. Testi antaa luotettavuustason hypoteesin paikkaansapitävyydelle. Yleensä tasoksi, jolla nollahypoteesia ei enää hyväksytä, valitaan enintään $0,05$.

Taulukko 3: *Pienimmän varianssin mukaan järjestetyt tulokset*

s_1	s_2	Keskiarvo	Varianssi
herne	nenä	-1.000	0.000
vihainen	mielenosoittaja	-1.000	0.000
tuntematon	sotilas	-1.025	0.025
valkoinen	talo	-0.975	0.083
ottaa	onki	-1.250	0.188
venäjä	presidentti	-1.128	0.472
kova	tuuli	-0.880	0.492
liukas	keli	-0.788	0.608
oppia	lukea	-0.606	1.087
heittää	veivi	-0.500	1.250
aste	pakkanen	-0.465	1.347
hakea	työ	-0.433	2.046
olla	ula	-0.250	2.438
sekä	myös	0.252	2.981
ja	olla	-0.083	3.635

T-testissä oletetaan että todennäköisyydet ovat normaalijakautuneita, ja tutkitaan eroaako havaintojoukon odotusarvo nollahypoteesin antaman jakauman odotusarvosta. Lasketaan siis t-arvot

$$t = \frac{\hat{x} - \mu}{\sqrt{\frac{s^2}{N}}},$$

missä $\hat{x}$ on näytejoukon keskiarvo, s^2 näytejoukon varianssi, N näytteiden lukumäärä ja μ jakauman keskiarvo. Tässä tapauksessa

$$\begin{aligned}
\mu &= P(s_1)P(s_2) = \frac{C(s_1)}{N}\frac{C(s_2)}{N} \\
\hat{x} &= \frac{C(s_1, s_2)}{N} = \hat{p} \\
s^2 &= p(1-p) = \hat{p}(1-\hat{p}) \approx \hat{p}
\end{aligned}$$

Esimerkiksi sanaparille "valkoinen talo" saadaan

$$t = \frac{\frac{710}{28181344} - \frac{2665 \cdot 10767}{28181344^2}}{\sqrt{\frac{710}{28181344^2}}} \approx 27$$

Jos t-testin tulos on yli 6.314, näyte on vedetty alle 5% todennäköisyydellä riippumattomasta jakaumasta. Valkoinen talo vaikuttaa siis kollokaatiolta. Taulukossa 4 on kaikkien sanojen tulokset. Huomataan, että viimeiset sanaparit saivat negatiivisia arvoja. Tämä johtuu siitä että ne esiintyvät vierekkäin harvemmin kuin nollahypoteesi antaa olettaa.

Taulukko 4: *t-testin tulokset*

s_1	s_2	t
valkoinen	talo	27
venäjä	presidentti	26
kova	tuuli	17
aste	pakkanen	13
tuntematon	sotilas	12
liukas	keli	10
oppia	lukea	4
hakea	työ	4
ottaa	onki	3
vihainen	mielenosoittaja	3
heittää	veivi	2
herne	nenä	2
olla	ula	0
sekä	myös	-9
ja	olla	-385

χ^2-testissä katsotaan annettujen sanojen esiintymistodennäköisyydet ja lasketaan niiden perusteella, kuinka monta kertaa sanojen pitäisi esiintyä yhdessä. Tätä lukua verrataan havaittuun lukuun ja jos nämä poikkeavat suuresti toisistaan, todetaan että sanojen pitää olla kollokaatioita.

Aloitetaan kasaamalla seuraavanlainen taulukko (taulukko 5):

Taulukko 5: χ^2-*testissä tarvittavia suureita.*

	w_1=valkoinen	$w_1 \neq$ valkoinen
w_2 =talo	710 (valkoinen talo)	$10767 - 710 = 10057$ (punainen talo)
$w_2 \neq$talo	$3665 - 710 = 2955$ (valkoinen mopo)	$28181344 - 710 - 10057 - 2955 = 28167622$ (punainen mopo)

Nämä arvot voidaan sijoittaa sitten kahden muuttujan χ^2-testin kaavaan:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

Luvut sijoittamalla saadaan siis:

$$\begin{aligned}\chi^2 &= \frac{28181344(710 \cdot 28167622 - 10057 \cdot 2955)^2}{(710 + 10057)(710 + 2955)(10056 + 28167622)(2955 + 28167622)} \\ &\approx 358771\end{aligned}$$

Jos χ^2-testin tulos on yli 3.843, näyte on vedetty alle 5% todennäköisyydellä riippumattomasta jakaumasta. Tässä siis valkoinen talo vaikuttaa kollokaatiolta. Kuitenkin kun katsomme taulukkoa 6, huomaamme että melkein kaikki sanat olisivat sen mukaan kollokaatioita. χ^2-testihän ei testaa sitä, ovatko sanat kollokaatioita, vaan sitä että ovatko sanat riippumattomia. Esimerkiksi sanapari "ja", "olla" on melko korkealla tuloksissa, sillä näiden kahden sanan välillä esiintyy negatiivinen korrelaatio: sanat esiintyvät harvemmin peräkkäin kuin niiden satunnaisuuden mukaan pitäisi. Tätä riippuvuutta ei voida tietysti pitää merkkinä siitä, että sanat olisivat kollokaatioita.

Taulukko 6: χ^2-*testin tulokset*

s_1	s_2	χ^2
liukas	keli	591591
valkoinen	talo	358771
aste	pakkanen	173726
tuntematon	sotilas	70409
ja	olla	29194
kova	tuuli	26644
venäjä	presidentti	18147
heittää	veivi	4120
herne	nenä	2258
vihainen	mielenosoittaja	1321
ottaa	onki	525
oppia	lukea	449
hakea	työ	47
sekä	myös	45
olla	ula	0

4. Yhteisinformaatio kertoo, kuinka paljon lisätietoa X:n havaitseminen antaa Y:stä. Jos X ja Y ovat riippumattomia, yhteisinformaatio on nolla. Lasketaan käsin malliksi tulos sanaparille "valkoinen","talo".

$$\begin{aligned}I(x,y) &= \log_2 \frac{P(X,Y)}{P(X)P(Y)} \\ &= \log_2 \frac{\frac{710}{28181344}}{\frac{3665}{28181344}\frac{10767}{28181344}} \\ &\approx 9.0\end{aligned}$$

Tulokset koko sanajoukolle on esitetty taulukossa 7.

Taulukko 7: *Yhteisinformaation mukaan järjestetyt tulokset*

s_1	s_2	MI
liukas	keli	12.4
aste	pakkanen	10.1
heittää	veivi	9.7
herne	nenä	9.6
valkoinen	talo	9.0
tuntematon	sotilas	8.8
vihainen	mielenosoittaja	7.6
kova	tuuli	6.6
ottaa	onki	5.9
venäjä	presidentti	4.8
oppia	lukea	4.5
hakea	työ	1.7
olla	ula	0.5
sekä	myös	-0.8
ja	olla	-2.5

Tulokset vaikuttavat hyviltä. Hieman kommenttia kirjan kritikkiin, että menetelmä erityisesti suosisi harvinaisia sanoja: Yksi tekijä joka tähän johtaa, on laskussa käytettyjen todennäköisyyksien estimointi — tässä käytetään maksimiuskottavuusestimaattoreita. Paremman tuloksen saa varmasti, jos asettaa sanapareille priorin, että ne ovat riippumattomia ja antaa datan sitten muokata tätä oletusta.

Yhteenvetona koko laskarista voisi sanoa vaikka seuraavaa: Heuristisilla menetelmillä (1. ja 2. tehtävä) voidaan päästä helpohkosti kohtalaisiin tuloksiin. Tehtävissä 3–4 sinänsä perustellut matemaattiset mallit mittaavat sanojen esiintymisen korrelaatiota, ei sitä, ovatko sanat kollokaatioita. Näillä menetelmillä voidaan silti saada hyviä tuloksia. Tilastomatematiikkaa on ehkä vaikeampi hahmottaa ja sitä käyttäessä on ymmärrettävä testin vaatimat oletukset. Todennäköisyyslaskuissa (4. tehtävä) nämä oletukset tuodaan eksplisiittisemmin esille. Todennäköisyyteen perustuvissa laskuissa joutuu myös harkitsemaan, miten tarvittavat todennäköisyydet approksimoidaan. Tässä on käytetty suurimman uskottavuuden estimaatteja (ML), jotka ovat ehkä liian herkkiä satunnaisvaihtelulle, kun näytteitä on suhteessa vähän. Parempana estimaattina voisi käyttää maksimi a posteriori (MAP) -estimaattia, jossa prioriuskomuksena olisi, että sanat eivät ole riippuvia. Tällöin malli väittäisi sanoja riippuviksi vasta kuin riittävä määrä dataa todistaa asian puolesta.