# T-61.5020 Luonnollisten kielten tilastollinen käsittely

Vastaukset 9, to 28.3.2007, 12:15–14:00 — Tilastollinen konekääntäminen
Versio 1.0

1. Etsitään todennäköisin käännös $\hat{e}$ ruotsinkieliselle lauseelle $r$:

$$\hat{e} = \underset{e}{\operatorname{argmax}} P(e|r) = \underset{e}{\operatorname{argmax}} P(e)P(r|e)$$

Käytetään kirjassa esitettyä mallia todennäköisyydelle $P(r|e)$:

$$P(r|e) = \frac{1}{Z} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} P(r_j|e_{a_j})$$

missä $m$ on alkuperäisen ruotsinkielisen lauseen pituus ja $l$ on käännetyn englanninkielisen lauseen pituus. Lasketaan kummallekin vaihtoehdolle:

$$P(r|e_1) = 1.0 \cdot 0.7 \cdot 0.9 \cdot 1.0 \cdot 1.0 \cdot 1.0 \cdot 0.1 = 0.063$$
$$P(r|e_2) = 1.0 \cdot 0.7 \cdot 1.0 \cdot 1.0 \cdot 1.0 \cdot 1.0 \cdot 1.0 \cdot 1.0 = 0.7$$

Tässä siis kokeillaan kaikkia käännössääntöjä jokaiselle ruotsinkielisen lauseen sanalle (huomioimatta sanajärjestystä). Koska säännöstö on hyvin harva, päästään näin yksinkertaiseen laskutoimitukseen.

Prioritodennäköisyys $P(e)$ saadaan kielimallista. Lasketaan se kummallekin lauseelle:

$$\begin{aligned} P(e_1) &= \prod_{i=1}^{l} P(w_i) = 0.18 \cdot 0.05 \cdot 0.01 \cdot 0.13 \cdot 0.1 \cdot 0.12 \cdot 0.02 = 2.8 \cdot 10^{-9} \\ P(e_2) &= 0.18 \cdot 0.07 \cdot 0.11 \cdot 0.21 \cdot 0.01 \cdot 0.13 \cdot 0.1 \cdot 0.01 = 3.8 \cdot 10^{-10} \end{aligned}$$

Kertomalla priori ja käännöstodennäköisyys huomataan, että jälkimmäinen käännös on todennäköisempi:

$$P(e_1)P(r|e_1) = 0.063 \cdot 2.8 \cdot 10^{-9} = 1.8 \cdot 10^{-10}$$
$$P(e_2)P(r|e_2) = 2.6 \cdot 10^{-10}$$

Huomattavaa on, että käännösmalli ei ota mitään kantaa sanajärjestykseen. Koska myöskään kielimalli (unigrammimalli) ei tätä tee, mallin mielestä sanajärjestyksellä ei ole mitään merkitystä. Jos mallilta kysytään todennäköisintä lausetta (ei testata eri vaihtoehtoja) huomataan, että todennäköisimpään lauseeseen ei voi tulla artikkeleja eikä sanaa “into”. Tämä johtuu siitä, että niiden lisääminen ei muuta käännöstodennäköisyyttä, mutta tiputtaa aina kielimallitodennäköisyyttä. Kielimalli siis suosii muutenkin lyhyempiä lauseita. Kasvattamalla kielimallin konteksti trigrammiksi, saisi ehkä artikkelit ja sanajärjestyksen paremmin kohdalleen.

Yleisesti tällä menetelmällä tarvitaan heuristiikkaa valitsemaan käännökset, joita tutkitaan. Kaikkien vaihtoehtojen läpikäynti on käytännössä mahdotonta.

2. Käytetään esimerkkisanaa $f$ = "tosiasia". Se on esiintynyt korpuksessa 989 kertaa. Normalisointia varten aineistosta pitää laskea erikseen kaikkien englannin sanojen esiintymismäärät.

**a-b)** Allaolevassa taulukossa on 20 suurinta arvoa saanutta englannin sanaa yhteisesiintymien frekvenssin sekä englannin sanan esiintymismäärällä normalisoidun frekvenssin mukaisesti. Nähdään että kumpikaan menetelmä ei annan toivottuja tuloksia. Ensimmäisessä ovat ongelmana hyvin yleiset sanat, joita esiintyy lähes jokaisessa lauseessa ja siten myös yhtäaikaa $f$:n kanssa. Jälkimmäisessä taas ongelmaksi muodostuvat hyvin harvinaiset sanat: Jos kerran esiintynyt sanaa sattuu esiintymään yhtä aikaa $f$:n kanssa, se saa suurimman mahdollisen arvon.

| $e$ | $C(e,f)$ |
|---|---|
| the | 2563 |
| of | 1128 |
| that | 1086 |
| is | 1040 |
| and | 829 |
| to | 823 |
| in | 726 |
| a | 716 |
| fact | 654 |
| it | 385 |
| this | 376 |
| we | 316 |
| are | 295 |
| not | 280 |
| for | 274 |
| have | 253 |
| be | 234 |
| which | 230 |
| on | 224 |
| has | 212 |

| $e$ | $\frac{C(e,f)}{C(e)}$ |
|---|---|
| winkler | 1.0000 |
| visarequired | 1.0000 |
| visaexempt | 1.0000 |
| veiling | 1.0000 |
| valuejudgment | 1.0000 |
| undisputable | 1.0000 |
| stayers | 1.0000 |
| semipermeable | 1.0000 |
| rulingout | 1.0000 |
| roentgen | 1.0000 |
| residuarity | 1.0000 |
| regionallevel | 1.0000 |
| redhaired | 1.0000 |
| poorlyfounded | 1.0000 |
| philippic | 1.0000 |
| pemelin | 1.0000 |
| paiania | 1.0000 |
| overcultivation | 1.0000 |
| outturns | 1.0000 |
| onesixth | 1.0000 |

**c)** Edellisissä menetelmissä oli ongelmana että ne eivät ottaneet huomioon käännöksen molempia suuntia: Jotta $e$ olisi todennäköinen käännös $f$:lle, $e$:n pitäisi esiintyä niissä lauseissa joissa $f$ esiintyi, sekä $f$:n pitäisi esiintyä niissä lauseissa joissa $e$ esiintyi. Tällöin todennäköisyyksien $P(e|f) = \frac{C(e,f)}{C(f)}$ sekä $P(f|e) = \frac{C(e,f)}{C(e)}$ pitäisi kummankin olla suuria. Kokeillaan seuraavana painotuksena näiden todennäköisyyksien tuloa.

Tulokset ovat seuraavan sivun vasemmanpuoleisessa taulukossa. Tällä kertaa löydettiin oikea käännös, ja myös toinen merkitykseltään läheinen sana "reality" on suhteellisen korkealla.

Kokeillaan vielä kollokaatiolaskarista tuttua $\chi^2$-testiä:

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

missä

$$\begin{aligned} O_{11} &= C(e,f) \\ O_{12} &= C(e,\neg f) = C(e) - C(e,f) \\ O_{21} &= C(\neg e,f) = C(f) - C(e,f) \\ O_{22} &= C(\neg e,\neg f) = N - C(e) - C(f) + C(e,f) \end{aligned}$$

ja $N$ aineiston kaikkien lauseiden määrä. Yli 3.843 arvon saaville sanoille tulokset tarkoittavat, että todennäköisyys sille että yhteisesiintymät olivat sattuman aikaansaamia, on alle 5%.

Suurimmat arvot saaneet sanat ovat oikeanpuoleisessa taulukossa. Testi näyttää toimivan erittäin hyvin: Ainoastaan "fact" ylitti luotettavuusrajan. Toisaalta jos haluisimme vaihtoehtoisia käännöksiä, kuten "reality", joku todennäköisyysarvon antava menetelmä olisi käyttökelpoisempi. Käytännössä käännöstodennäköisyydet etsitään iteratiivisesti EM-algoritmilla, jolloin rajoitetaan sitä että yksi englannin sana olisi monen suomenkielisen sanan käännös, mutta esitetyn kaltainen menetelmä voisi toimia todennäköisyyksien alustuksena.

| $e$ | $\log(\frac{C(e,f)}{C(e)} \cdot \frac{C(e,f)}{C(f)})$ |
|---|---:|
| fact | -3.9758 |
| the | -5.6159 |
| that | -5.8849 |
| is | -5.9086 |
| reality | -6.0057 |
| winkler | -6.2035 |
| of | -6.4963 |
| hedgehog | -6.6090 |
| a | -6.6577 |
| and | -6.8386 |
| visarequired | -6.8967 |
| visaexempt | -6.8967 |
| veiling | -6.8967 |
| valuejudgment | -6.8967 |
| undisputable | -6.8967 |
| stayers | -6.8967 |
| semipermeable | -6.8967 |
| rulingout | -6.8967 |
| roentgen | -6.8967 |
| residuarity | -6.8967 |

| $e$ | $\chi^2$ |
|---|---:|
| fact | 18.2155 |
| the | 3.5937 |
| that | 3.0070 |
| is | 2.8096 |
| of | 2.3485 |
| reality | 2.3166 |
| winkler | 2.0000 |
| hedgehog | 1.3323 |
| visarequired | 1.0000 |
| visaexempt | 1.0000 |
| veiling | 1.0000 |
| valuejudgment | 1.0000 |
| undisputable | 1.0000 |
| stayers | 1.0000 |
| semipermeable | 1.0000 |
| rulingout | 1.0000 |
| roentgen | 1.0000 |
| residuarity | 1.0000 |
| regionallevel | 1.0000 |
| redhaired | 1.0000 |