T-61.5020 Luonnollisten kielten tilastollinen käsittely

Vastaukset 8, ke 21.3.2007, 12:15–14:00 — N-grammikielimallit
Versio 1.0

1. Alla on erään henkilön ja tilaston estimaatit sille, miten todennäköistä on, että alla annetut sanat seuraavat sanoja “tuntumaan jo”:

sana	tilasto trig	ihminen trig	ihminen lause
ja	0.00	0.00	0.00
hyvältä	1.00	0.18	0.40
kumisaapas	0.00	0.00	0.00
keväältä	0.00	0.23	0.50
ilman	0.00	0.05	0.05
päihtyneeltä	0.00	0.20	0.00
turhalta	0.00	0.23	0.05
koirineen	0.00	0.00	0.00
öljyiseltä	0.00	0.11	0.00
Turku	0.00	0.00	0.00

Taulukko 1: *Ihminen vs tilasto, trigrammiestimaatit*

Tarkemmalla tukimisilla huomataan, että taulukossa ihmisen antamat trigrammiestimaatit ovat jonkin verran pielessä ja tilastolliset pehmentämättömät trigrammiestimaatit aivan pielessä.

Tilastollisten estimaattien laskuun käytettiin n. 30 miljoonan sanan aineistoa. Tässä aineistossa ei yksikään annetuista taivutetuista trigrammeista esiintynyt kertaakaan. Trigrammit perusmuotoistamalla löydettiin 11 lausetta, joissa esiintyi “tuntua jo hyvä”. Estimaatti kaipaa siis selvästi tasoittamista, eikä senkään jälkeen ole kovin luotettava.

Myös esimerkki-ihmisen antamaa estimaattia voi epäillä, aivan mahdollisille lauseille on annettu nollatodennäköisyys, esim. “Kyllä alkaa tuntumaan jo kumisaapas jalassa”, lause joka voidaan tokaista vaikka pitkän vaelluksen päätteeksi. Toisaalta annetulla 2 desimaalin tarkkuudella estimaatit lienevät hyviä.

Kun testihenkilölle annettiin koko lause nähtäväksi, saatiin jo varsin laadukkaat estimaatit. Jotta tilastollisesti pystyttäisiin pääsemään samaan tulokseen, tarvitsisi mallin ymmärtää suomen kielen syntaksia (miten sanoja voidaan taivuttaa ja laittaa peräkkäin) sekä myös sanojen semanttista merkitystä (“helmikuu” on lopputalvea, melkein kevättä).

2. a) Suurimman uskottavuuden estimaatit voidaan laskea kaavasta

$$P(w_i|w_{i-1}, w_{i-2}, \dots) = \frac{C(w_i, w_{i-1}, w_{i-2}, \dots)}{C(w_{i-1}, w_{i-2}, \dots)},$$

missä funktio C kertoo opetusjoukossa nähtyjen näytteiden määrän.

Unigrammiestimaatissa unohdetaan riippuvuus kaikista edellisistä sanoista, bigrammiestimaatti riippuu vain edellisestä sanasta ja trigrammiestimaatissa käytetään historiana kahta edellistä sanaa.

Unigrammiestimaatit voidaan siis laskea

$$P(w_i) = \frac{C(w_i)}{C(0)},$$

missä $C(0)$ on opetusjoukon näytteiden lukumäärä. Estimaatit ovat siis samat kummallekin tehtävänannon historialle.

$$\begin{aligned}
P(\text{'olla'}) &= \frac{5}{98} = 0.051 \\
P(\text{'leuto'}) &= \frac{1}{98} = 0.001 \\
P(\text{'gorilla'}) &= \frac{0}{98} = 0.000
\end{aligned}$$

Bigrammiestimaatit saadaan ottamalla yksi sana historiasta käyttöön:

$$\begin{aligned}
P(w_i|w_{i-1}) &= \frac{C(w_i, w_{i-1})}{C(w_{i-1})} \\
P(\text{'olla'}|\text{'olla'}) &= \frac{0}{5} = 0.000 \\
P(\text{'leuto'}|\text{'olla'}) &= \frac{1}{5} = 0.200 \\
P(\text{'gorilla'}|\text{'olla'}) &= \frac{0}{5} = 0.000 \\
P(\text{'olla'}|\text{'vaikuttaa'}) &= \frac{0}{1} = 0.000 \\
P(\text{'leuto'}|\text{'vaikuttaa'}) &= \frac{0}{1} = 0.000 \\
P(\text{'gorilla'}|\text{'vaikuttaa'}) &= \frac{0}{1} = 0.000
\end{aligned}$$

Huomataan, että tämän mallin mielestä mikään sanayhdistelmä, jota se ei ole nähnyt, ei ole mahdollinen.

b) Laplacen estimaatissa aikaisemmin havaitsemattomille sanoille annetaan hieman todennäköisyyttä. Estimaatti vastaa prioria, että kaikki sanat ovat yhtä todennäköisiä. Käytännössä se lasketaan niin, että kuvitellaan että kaikkia sanoja on jo nähty kerran:

$$P(w_i|w_{i-1}, w_{i-2}, \dots) = \frac{C(w_i, w_{i-1}, w_{i-2}, \dots) + 1}{C(w_{i-1}, w_{i-2}, \dots) + N},$$

missä N on mallin sanaston koko.
Lasketaan siis estimaatit:

$$\begin{aligned}
P(\text{'olla'}) &= \frac{5+1}{98+64000} = 9.3\cdot 10^{-5} \\
P(\text{'leuto'}) &= \frac{1+1}{98+64000} = 3.1\cdot 10^{-5} \\
P(\text{'gorilla'}) &= \frac{0+1}{98+64000} = 1.6\cdot 10^{-5}
\end{aligned}$$

Bigrammeille:

$$\begin{aligned}
P(\text{'olla'}|\text{'olla'}) &= \frac{1}{5+64000} = 1.6\cdot 10^{-5} \\
P(\text{'leuto'}|\text{'olla'}) &= \frac{1+1}{5+64000} = 3.1\cdot 10^{-5} \\
P(\text{'gorilla'}|\text{'olla'}) &= \frac{1}{5+64000} = 1.6\cdot 10^{-5} \\
P(\text{'olla'}|\text{'vaikuttaa'}) &= \frac{1}{1+64000} = 1.6\cdot 10^{-5} \\
P(\text{'leuto'}|\text{'vaikuttaa'}) &= \frac{1}{1+64000} = 1.6\cdot 10^{-5} \\
P(\text{'gorilla'}|\text{'vaikuttaa'}) &= \frac{1}{1+64000} = 1.6\cdot 10^{-5}
\end{aligned}$$

Huomataan, että tässä priorioletus, että kaikki sanat ovat yhtä todennäköisiä ohjaa estimaatteja vahvasti, kaikki sanat ovat toisaan mallien mielestä lähes yhtä todennäköisiä.

c) Lidstonen estimaatissa voidaan säätää sitä, kuinka paljon uskotaan siihen, että sanat ovat yhtä todennäköisiä. Siinä kuvitellaan, että sanat on jo nähty λ kertaa ennen opetusaineistoa:

$$P(w_i|w_{i-1},w_{i-2},\ldots) = \frac{C(w_i,w_{i-1},w_{i-2},\ldots)+\lambda}{C(w_{i-1},w_{i-2},\ldots)+\lambda N},$$

Valitaan tässä $\lambda = 0.01$. Lasketaan estimaatit:

$$\begin{aligned}
P(\text{'olla'}) &= \frac{5+0.01}{98+0.01\cdot 64000} = 6.8\cdot 10^{-3} \\
P(\text{'leuto'}) &= \frac{1+0.01}{738} = 1.4\cdot 10^{-4} \\
P(\text{'gorilla'}) &= \frac{0.01}{738} = 1.4\cdot 10^{-5}
\end{aligned}$$

Bigrammeille:

$$\begin{aligned}
P(\text{'olla'}|\text{'olla'}) &= \frac{0.01}{645} = 1.6 \cdot 10^{-5} \\
P(\text{'leuto'}|\text{'olla'}) &= \frac{1 + 0.01}{645} = 1.6 \cdot 10^{-3} \\
P(\text{'gorilla'}|\text{'olla'}) &= \frac{0.01}{641} = 1.6 \cdot 10^{-5} \\
P(\text{'olla'}|\text{'vaikuttaa'}) &= \frac{0.01}{641} = 1.6 \cdot 10^{-5} \\
P(\text{'leuto'}|\text{'vaikuttaa'}) &= \frac{0.01}{641} = 1.6 \cdot 10^{-5} \\
P(\text{'gorilla'}|\text{'vaikuttaa'}) &= \frac{0.01}{641} = 1.6 \cdot 10^{-5}
\end{aligned}$$

Tässä tapauksessa opetusdata ohjaa selkeämmin estimaatteja. Sopiva λ voidaan valita laittamalla opetusjoukosta pieni osa sivuun ja testaamalla tällä sivuun laitetulla tekstillä, mikä λ toimii parhaiten.

3. Absoluuttisessa vähennyksessä havaituista n-grammien esiintymien määrästä vähennetään vakioarvo (tässä siis $D = 0.5$). Vähennys voidaan luonnollisesti tehdä vain silloin esiintymiä oli vähintään yksi. Kun vähennys tehdään bigrammijakauman estimaatteihin, jäljelle jää todennäköisyysmassaa, joka voidaan käyttää interpoloimiseen unigrammitodennäköisyyksien kanssa (tai vaihtoehtoisesti perääntymiseen). Bigrammimallille absoluuttisen vähennyksen ja interpoloinnin yhdistäminen antaa siis estimaatin:

$$P(w_i|w_{i-1}) = \frac{\max(C(w_i, w_{i-1}) - D, 0)}{C(w_{i-1})} + \gamma(w_{i-1})\frac{C(w_i)}{C(0)}$$

Tässä $\gamma(w_{i-1})$ on bigrammihistoriasta riippuva kerroin, joka normalisoi todennäköisyysjakauman summautumaan ykköseen. Sen arvo voidaan laskea siitä kuinka monta erilaista seuraajaa historialla opetusaineistossa oli:

$$\gamma(w_{i-1}) = |\{(w_{i-1}, w_i) : C(w_i, w_{i-1}) > 0\}| \cdot \frac{D}{C(w_{i-1})}$$

Aloitetaan laskemalla historioille “olla” ja “vaikuttaa” nämä interpolointikertoimet. Historialla “olla” on esiintynyt viisi erilaista seuraajaa ja se on yhteensä esiintynyt viisi kertaa. Historialla “vaikuttaa” on yksi esiintynyt seuraaja, ja esiintymiä myös yksi. Siispä:

$$\begin{aligned}
\gamma(\text{olla}) &= 5 \cdot \frac{0.5}{5} = 0.5 \\
\gamma(\text{vaikuttaa}) &= 1 \cdot \frac{0.5}{1} = 0.5
\end{aligned}$$

Nyt voidaan laskea interpoloidut bigrammitodennäköisyydet:

$$\begin{aligned}
P(\text{'olla'}|\text{'olla'}) &= \frac{0}{5} + 0.5\frac{5}{98} = 0.025 \\
P(\text{'leuto'}|\text{'olla'}) &= \frac{1}{5} + 0.5\frac{1}{98} = 0.205 \\
P(\text{'gorilla'}|\text{'olla'}) &= \frac{0}{5} + 0.5\frac{0}{98} = 0.0 \\
P(\text{'olla'}|\text{'vaikuttaa'}) &= \frac{0}{1} + 0.5\frac{5}{98} = 0.025 \\
P(\text{'leuto'}|\text{'vaikuttaa'}) &= \frac{0}{1} + 0.5\frac{1}{98} = 0.005 \\
P(\text{'gorilla'}|\text{'vaikuttaa'}) &= \frac{0}{1} + 0.5\frac{0}{98} = 0.0
\end{aligned}$$

Huomataan että menelmä ei tälläisenään anna yhtään todennäköisyyttä opetusaineiston ulkopuoliselle sanastolle (gorilla). Tämä voitaisiin korjata käyttämällä unigrammitodennäköisyyksille vielä jotain tasoitusmenetelmää. Mahdollista olisi myös jatkaa interpolointia niin sanoittuihin nollagrammeihin, mikä tarkoittaa tasajakauman kanssa interpolointia.

Interpolointi tai peräytyminen matalamman asteen n-grammeihin ei kuitenkaan aina ole täysin luotettavaa. Esimerkkinä voidaan tarkastella bigrammia "San Francisco". Koska paikannimi on hyvin tunnettu, estimaatti $P(\text{'Francisco'}|\text{'San'})$ on todennäköisesti luotettava. Eli jos edellinen sana on "San", ei tule ongelmia. Mutta entä jos edellinen sana ei ollut "San"? Tällöin Fanciscon todennäköisyyden pitäisi olla selvästi pienempi kuin silloin, kun tarkastellaan unigrammijakaumaa ilman tietoa edellisestä sanasta. Tähän ideaan pohjautuu Kneser-Ney -tasoitus, jossa alemman asteen n-grammien todennäköisyydet estimoidaan arvioimalla kuinka todennäköistä sanan on esiintyä *uudessa* kontekstissa. Tällä hetkellä parhaaksi todettu tasoitusmenetelmä on juuri modifioitu Kneser-Ney -interpolointi, joka käyttää näiden tyyppiestimaattien lisäksi absoluuttista vähennystä kolmella erikseen optimoitavalla vakiolla.

4. Muutetaan hämmentyneisyyden (perplexity) kaavaa niin, että voidaan suoraan käyttää log-todennäköisyyksiä:

$$\begin{aligned}
perp(w_1, w_2, \ldots, w_N) &= \prod_{i=0}^{N} P(w_i|w_{i-1}, \ldots, w_1)^{-\frac{1}{N}} \\
&= \prod_{i=0}^{N} 10^{-\frac{1}{N}\log(P(w_i|w_{i-1},\ldots,w_1))} \\
&= 10^{-\frac{1}{N}\sum_{i=0}^{N}\log(P(w_i|w_{i-1},\ldots,w_1))}
\end{aligned}$$

Lasketaan summa erikseen:

$$
\begin{aligned}
& \sum_{i=0}^{N} \log(P(w_i|w_{i-1},\ldots,w_1)) \\
= \; & \overbrace{-4.1763}^{\text{kielen}} \; \overbrace{-2.1276}^{\text{oppiminen}} \; \overbrace{-0.4656}^{\text{on}} \; \overbrace{-0.001-4.2492}^{\text{monimutkainen}} \; \overbrace{-0.8876}^{\text{ja}} \; \overbrace{+0.0495-4.1804}^{\text{huonosti}} \\
& \overbrace{-0.1415-0.1652-5.2195}^{\text{ymmärretty}} \\
= \; & -21.5644
\end{aligned}
$$

Niille sanoille, joille ei löytynyt trigrammimallia, jouduttiin käyttämään sekä perääntymiskerrointa että todennäköisyyttä. Jos myöskään bigrammimallia ei löytynyt, jouduttiin vielä kerran perääntymään.

Sijoitetaan vielä luvut hämmentyneisyyden lausekkeeseen

$$perp(w_1, w_2, \ldots, w_N) = 10^{\frac{-21.5644}{-7}} \approx 1200$$

Tulosta voi ajatella vaikka niin, että kielimalli vastaa sellaista kielimallia, joka joutuu valitsemaan 1 200 yhtä todennäköisen sanan väliltä (ei ihan eksaktisti paikkansapitävä väite).

Sana "tapahtumaketju" ei ollut 64 000 yleisimmän sanan joukossa ja ei sisältynyt siis kielimalliin. Kielimallin ohi meni siis $\frac{1}{8} \approx 13\%$ sanoista.