Latentin semanttisen tiedon käyttö tilastollisessa kielimallinnuksessa

Jukka Kuusisto
`jkuusist@cc.hut.fi`

31.10.2001

Suunnitelma

- Motivointi
- LSA
- Klusterointi LSA-avaruudessa
- Semanttinen luokittelu
- N-grammien ja LSA:n yhdistäminen

Motivointi

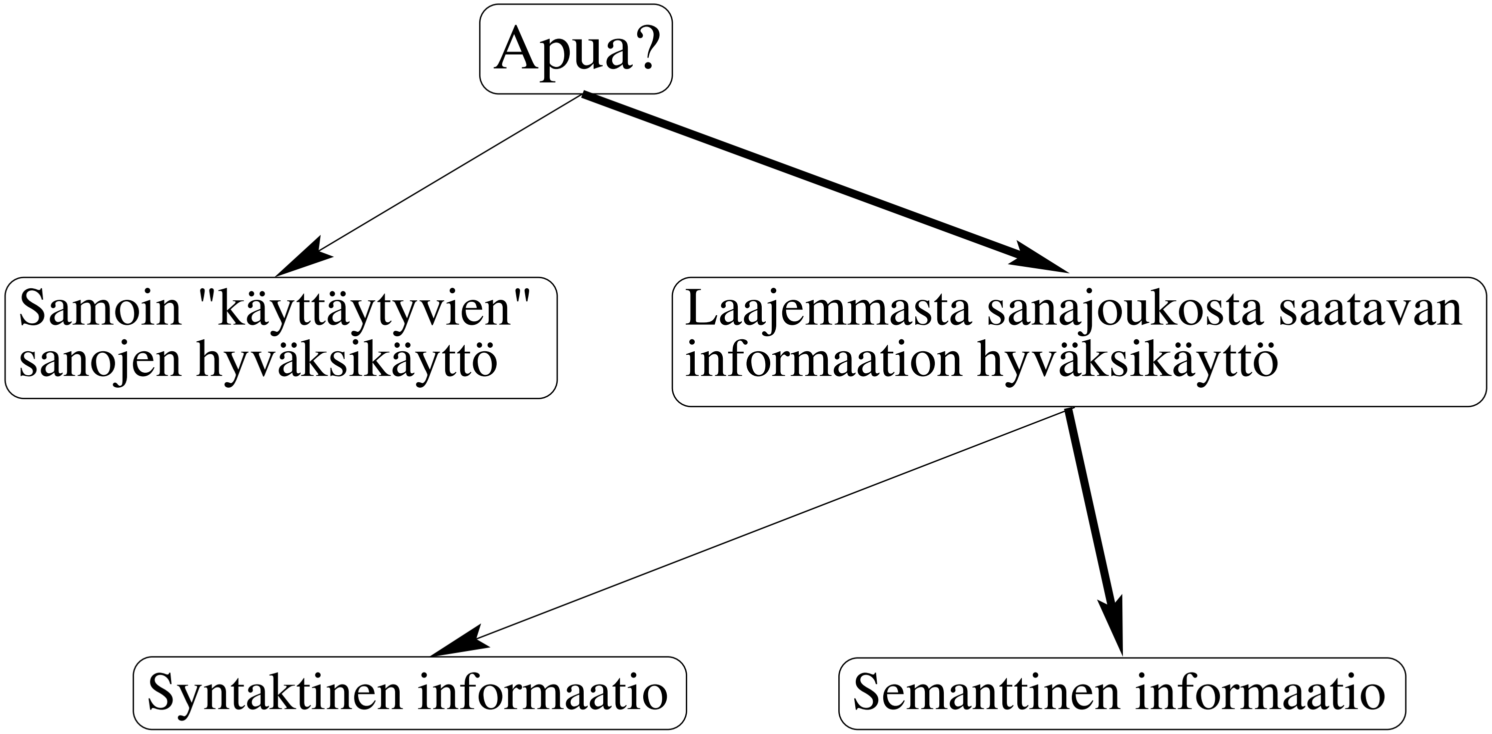
- tilastolliset n-grammit eivät riitä

- paikallisuusongelma
  - pieni $n \Rightarrow$ ennustusvoima pieni
  - suuri $n \Rightarrow$ parametriestimaatit epäluotettavampia

- käytännössä n on yleensä pieni

stocks fell sharply as a result of the announcement
stocks, as a result of the announcement, sharply fell

Paikallisuusongelman ratkaiseminen



Paikallisuusongelman ratkaisuun tähtäävät päätutkimussuunnat

Syntaktisen informaation käyttö

- jäsennetään tekstikappale pääsanoiksi ja käytetään n-grammia pääsanoille (**`[stocks]`** $\rightarrow$ **`[fell]`** $\rightarrow$ **`[sharply]`** $\rightarrow$ **`...`**)

- ongelmana riippuvuus jäsentimestä

Semanttisen informaation käyttö

- dokumentti = semanttisesti yhtenäinen (tagattu) joukko lauseita
- korreloivien sanaparien hyväksikäyttö
  - havaitaan "*stocks*" $\Rightarrow P($"*fell*"$)$++
- miten valita sanaparit?
  - $\Rightarrow$ LSA: kaikki sanakombinaatiot ovat mahdollisia sanapareja

LSA

- määritelmät:
  V on sanasto, jossa M kpl sanoja w_i
  T on opetustekstikorpus, jossa N dokumenttia d_j
  S on jatkuva vektoriavaruus

- LSA muodostaa kuvaukset
  $V \to S : \mathsf{LSA}(w_i) = \bar{u}_i$
  $T \to S : \mathsf{LSA}(d_j) = \bar{v}_j$

Piirteiden erottelu

- muodostetaan matriisi W, jonka alkiot kuvaavat sanojen esiintymistä dokumenteissa

- normalisoidaan dokumentin pituuden ja sanojen entropian suhteen

- matriisin W solu (i, j) saadaan kaavasta

$$w_{i,j} = (1 - \varepsilon_i)\frac{c_{i,j}}{n_j}, \tag{1}$$

  missä $c_{i,j}$ on w_i:n esiintymiskertojen lukumäärä d_j:ssä, n_j on d_j:n sanojen kokonaismäärä ja ε_i on sanan w_i normalisoitu entropia T:ssä

Piirteiden erottelu 2

- merkitään sanan w_i lukumäärää T:ssä t_i:llä
- normalisoitu entropia on tällöin

$$\varepsilon_i = -\frac{1}{\log N} \sum_{j=1}^{N} \frac{c_{i,j}}{t_i} \log \frac{c_{i,j}}{t_i} \tag{2}$$

  - ε_i lähellä ykköstä $\Rightarrow$ sana esiintyy monissa dokumenteissa
  - ε_i lähellä nollaa $\Rightarrow$ sana esiintyy harvoissa dokumenteissa
  - $\Rightarrow 1 - \varepsilon_i$ kuvaa sanan w_i indeksointivoimaa

Singulaariarvohajotelma (SVD)

- W:n ominaisuuksia:
  - W on kokoa $M \times N$ eli hyvin suuri
  - sanavektorit w_i ja dokumenttivektorit d_j ovat harvoja
  - sana- ja dokumenttivektorit ovat eri avaruuksissa

- ratkaisu: singulaariarvohajotelma

SVD 2

- singulaariarvohajotelma:

$$W \approx \widehat{W} = USV^T, \tag{3}$$

missä U on kokoa $M \times R$ oleva vasen singulaarimatriisi
S on kokoa $R \times R$ oleva singulaariarvojen diagonaalimatriisi
V on kokoa $N \times R$ oleva oikea singulaarimatriisi
R on hajotelman asteluku

Tulkintaa

- U:n ja V:n sarakkeet määrittävät ortonormaalin kannan u_i:den ja v_j:den virittämälle avaruudelle

- $\widehat{W}$ on paras rangia R oleva approksimaatio W:lle
  - R:n suuruus määrää poikkeaman

- S:ssä sanaa w_i edustaa sanavektori $\bar{u}_i = u_i S$

- S:ssä dokumenttia d_j edustaa dokumenttivektori $\bar{v}_j = v_j S$

Tulkintaa 2

- sanat ja dokumentit kuvataan siis piilossa olevien abstraktien konseptien lineaarisina kombinaatioina

- ideaalisesti $\hat{W}$ sisältää vain tärkeimmät rakenteelliset riippuvuudet W:stä

- semanttisesti toisiaan lähellä olevat sanat ja dokumentit ovat lähekkäin S:ssä

- sanojen ja dokumenttien klusterointi voidaan tehdä S:ssä

Sanojen klusterointi

- matriisi WW^T kuvastaa sanojen välistä yhtenäisyyttä

$$WW^T = US^2U^T, \tag{4}$$

joten sanojen läheisyyttä voidaan mitata $\bar{u}_i$:n ja $\bar{u}_j$:n välisen kulman kosinilla:

$$K(w_i, w_j) = \cos(u_i S, u_j S) = \frac{u_i S^2 u_j^T}{\| u_i S \| \| u_j S \|} \tag{5}$$

- tästä saadaan etäisyysmitta esim. (välille $[0, \pi]$):

$$D(w_i, w_j) = \arccos K(w_i, w_j) \tag{6}$$

Dokumenttien klusterointi

- vastaavasti matriisi $W^T W$ kuvastaa dokumenttien välistä yhtenäisyyttä

$$W^T W = V S^2 V^T, \tag{7}$$

joten dokumenttien läheisyyttä voidaan mitata $\bar{v}_i$:n ja $\bar{v}_j$:n välisen kulman kosinilla:

$$K(d_i, d_j) = \cos(v_i S, v_j S) = \frac{v_i S^2 v_j^T}{\| v_i S \| \| v_j S \|} \tag{8}$$

- kaavan (6) mukaista etäisyysmittaa voidaan käyttää tässäkin

LSA:n sovelluskohteita

- **semanttinen luokittelu**

- n-grammien ja LSA-kielimallin yhdistäminen

Semanttinen luokittelu

- halutaan löytää uudelle dokumentille sitä parhaiten vastaava aihe

- uutta dokumenttia $\tilde{d}_p$ nimitetään *pseudodokumentiksi*
  - oletetaan, että $\tilde{d}_p$ on “samanlainen” kuin muu korpus
  - toisin sanoen U ja S pidetään ennallaan

Semanttinen luokittelu 2

- singulaariarvohajotelmasta

$$\tilde{d}_p = US\tilde{v}_p^T \tag{9}$$

- nyt $\tilde{d}_p$:n kuvaus S:ssä on *pseudodokumenttivektori* $\tilde{\tilde{v}}_p$:

$$\tilde{\tilde{v}}_p = \tilde{v}_p S = \tilde{d}_p^T U \tag{10}$$

- $\tilde{d}_p$:n etäisyys dokumenttiklusterista D_c on $D(\tilde{d}_p, D_c)$ ja se voidaan laskea kaavasta (8)

Semanttinen päättely

- käsky- ja kontrollitehtävissä dokumenttiklusterit vastaavat tiettyjä tehtäviä

- klustereiden keskikohdat ovat ns. *semanttisia ankkureita*

- tuntematon sanajono tulkitaan siksi komennoksi, jonka semanttista ankkuria se on lähinnä

Semanttinen päättely 2

- ongelmia:
  - toisiaan lähellä olevat vektorit voivat tarkoittaa päinvastaista
  - pelkästään sanajärjestyksen muutos voi vaihtaa merkityksen
- ratkaisu: käytetään dokumenttien kuvaamiseen n:n sanan joukkoja yksittäisten sanojen sijaan

LSA:n sovelluskohteita

- semanttinen luokittelu
- **n-grammien ja LSA-kielimallin yhdistäminen**

N-grammit + LSA

- merkitään
  w_q = ennustettava sana
  H_{q-1} = w_q:n LSA-historia eli dokumentin alkuosa $\widehat{d}_{q-1}$

- LSA-kielimallin antaman todennäköisyys on

$$\Pr(w_q|H_{q-1}, S) = \Pr(w_q|\widehat{d}_{q-1}) \tag{11}$$

- määriteltäväksi jäävät siis pseudodokumentin esitysmuoto ja LSA-todennäköisyysmitta

Pseudodokumentin esitysmuoto

- kaavasta (10) saadaan

$$\tilde{\tilde{v}}_{q-1} = \tilde{v}_{q-1} S = \tilde{d}_{q-1}^{T} U, \tag{12}$$

  missä $\tilde{\tilde{v}}_{q-1}$ liikkuu LSA-avaruudessa q:n kasvaessa

- oletetaan, että hetkellä q havaitaan sana w_i

- lisäksi oletetaan, että $\varepsilon_i \approx \varepsilon_{i-1}$

Pseudodokumentin esitysmuoto 2

- kaavasta (1)

$$\tilde{d}_q = \frac{n_q - 1}{n_q}\tilde{d}_{q-1} + \frac{1-\varepsilon_i}{n_q}[0 \cdots 1 \cdots 0]^T, \tag{13}$$

missä vektorin i:s alkio on 1

- edelleen kaavasta (12)

$$\tilde{\tilde{v}}_q = \tilde{v}_q S = \frac{1}{n_q}[(n_q - 1)\tilde{\tilde{v}}_{q-1} + (1-\varepsilon_i)u_i] \tag{14}$$

eli $\tilde{\tilde{v}}_q$ voidaan päivittää suoraan LSA-avaruudessa

LSA-todennäköisyys

- $W(i, j)$ kuvaa sanan w_i ja dokumentin d_j yhteneväisyyttä
  - tämä alkio saadaan matriisin $US^{\frac{1}{2}}$ rivin i ja matriisin $VS^{\frac{1}{2}}$ rivin j välisestä pistetulosta $u_i S^{\frac{1}{2}} \cdot v_j S^{\frac{1}{2}}$

$$K(w_q, \tilde{d}_{q-1}) = \cos(u_q S^{\frac{1}{2}}, \tilde{v}_{q-1} S^{\frac{1}{2}}) = \frac{u_q S \tilde{v}_{q-1}^T}{\| u_q S^{\frac{1}{2}} \| \| \tilde{v}_{q-1} S^{\frac{1}{2}} \|} \quad (15)$$

- etäisyysmitta kuten aiemmin kaavan (6) avulla

- todennäköisyysmitta saadaan jakamalla todennäköisyysmassa etäisyyksien suhteessa

Yhdistäminen n-grammeihin

- intuitiivisesti $\Pr(w_q|\tilde{d}_{q-1})$ on suuri sisältösanoille ja pieni funktiosanoille
  - päinvastoin kuin tyypillisesti n-grammeilla

- kielimallin kokonaistodennäköisyyden johtamiseksi määritellään

$$\Pr(w_q|H_{q-1}^{(n+l)}) = \Pr(w_q|H_{q-1}^{(n)}, H_{q-1}^{(l)}) \quad (16)$$

$$= \frac{\Pr(w_q, H_{q-1}^{(l)}|H_{q-1}^{(n)})}{\sum_{w_i \in V} \Pr(w_i|H_{q-1}^{(n)}, H_{q-1}^{(l)})}, \quad (17)$$

missä (n) viittaa n-grammi-komponenttiin ja (l) LSA-komponenttiin

Yhdistäminen n-grammeihin 2

- laventamalla kaavan (17) nimittäjässä oleva termi ja olettamalla, että $\mathrm{Pr}(\tilde{d}_{q-1}|w_{q-1}\cdots w_{q-n+1}) = \mathrm{Pr}(\tilde{d}_{q-1}|w_{q-1})$ saadaan

$$\mathrm{Pr}(w_q|H_{q-1}^{(n+l)}) = \frac{\mathrm{Pr}(w_q|w_{q-1}\cdots w_{q-n+1})\mathrm{Pr}(\tilde{d}_{q-1}|w_q)}{\sum_{w_i \in V}\mathrm{Pr}(w_i|w_{q-1}\cdots w_{q-n+1})\mathrm{Pr}(\tilde{d}_{q-1}|w_i)} \qquad (18)$$

Yhdistäminen n-grammeihin 3

- Bayesin kaavan avulla saadaan näkyviin riippuvuus LSA-todennäköisyydestä:

$$\Pr(w_q|H_{q-1}^{(n+l)}) = \frac{\Pr(w_q|w_{q-1}\cdots w_{q-n+1})\frac{\Pr(w_q|\tilde{d}_{q-1})}{\Pr(w_q)}}{\sum_{w_i\in V}\Pr(w_i|w_{q-1}\cdots w_{q-n+1})\frac{\Pr(w_i|\tilde{d}_{q-1})}{\Pr(w_i)}} \quad (19)$$

Kontekstin laajuuden valinta

- n-grammin ja LSA:n antamia todennäköisyyksiä usein painotetaan jollain tavalla

- tunnistusvaiheessa dokumentti ei ole selkeästi määritelty

- käytetään eksponentiaalista hajoamista dokumenttihistorialle:

$$\tilde{\tilde{v}}_q = \frac{1}{n_q}[\lambda(n_q - 1)\tilde{\tilde{v}}_{q-1} + (1 - \varepsilon_i)u_i, \tag{20}$$

missä parametri λ, $0 < \lambda < 1$, riippuu aineiston yhtenäisyydestä

Tasoitus

- tasoituksessa käytetään apuna sanojen ja dokumenttien klusterointia LSA-avaruudessa

- sanojen tasoitus: sanaklustereita C_k käyttäen

$$\mathrm{Pr}(w_q|\tilde{d}_{q-1}) = \sum_{k=1}^{K} \mathrm{Pr}(w_q|C_k)\mathrm{Pr}(C_k|\tilde{d}_{q-1}), \tag{21}$$

jossa esiintyvät todennäköisyydet saadaan kaavoista (15) ja (5)

Dokumenttien tasoitus

- dokumenttiklustereita D_c käyttäen:

$$\mathsf{Pr}(w_q|\tilde{d}_{q-1}) = \sum_{c=1}^{C} \mathsf{Pr}(w_q|D_c)\mathsf{Pr}(D_c|\tilde{d}_{q-1}), \qquad (22)$$

jossa esiintyvät todennäköisyydet saadaan kaavoista (15) ja (8)

Yhdistetty tasoitus

$$\Pr(w_q|\tilde{d}_{q-1}) = \sum_{k=1}^{K} \sum_{c=1}^{C} \Pr(w_q|C_k, D_c)\Pr(C_k, D_c|\tilde{d}_{q-1}), \tag{23}$$

jota voidaan approksimoida kaavalla

$$\Pr(w_q|\tilde{d}_{q-1}) = \sum_{k=1}^{K} \sum_{c=1}^{C} \Pr(w_q|C_k)\Pr(C_k|D_c)\Pr(D_c|\tilde{d}_{q-1}), \tag{24}$$

jossa esiintyvät todennäköisyydet saadaan kaavoista (5), (11) ja (8)

- näistä eri tasoitusmenetelmistä saadan neljä erilaista malliperhettä hybrideille n-gram+LSA-kielimalleille

Yhteenveto

- LSA on herkkä opetusdatan suhteen:
  - esim. kirjoitustyyli; miten otetaan huomioon?

- puheentunnistusvirheet usein funktiosanoissa $\Rightarrow$ LSA ei auta:
  - syntaktisen informaation käyttö?

- tulevaisuus: yhdistetään
  - syntaktinen informaatio suuresta sanajoukosta
  - globaali semanttinen analyysi
  - käytännön tieto tehtävästä

Yhteenveto 2

- yhdistäminen akustiseen komponenttiin
- sovelluksen resurssit asettavat rajoja

Kotitehtävä

Osoitteesta http://www.hut.fi/~jkuusist/T-61.184/ löytyvät matriisit `W` ja `W_feat` sekä vektorit `newdoc` ja `newdoc_feat`. Matriisin `W` alkiot kuvaavat tiettyjen sanojen (matriisin rivit) esiintymisfrekvenssejä tietyissä dokumenteissa (matriisin sarakkeet). Näistä on laskettu valmiiksi dokumenttien pituuksien ja sanojen entropian huomioonottava piirrematriisi `W_feat`.

Vektori `newdoc` on pseudodokumentti ja vektori `newdoc_feat` siitä valmiiksi muodostettu pseudodokumenttivektori $\tilde{d}$. Mitä alkuperäistä dokumenttia pseudodokumentti on lähimpänä semanttisen luokittelun mielessä, jos latentin avaruuden ulottuvuus $R = 3$?

Vihje: Matlabin svd