Aihesekoitemallit

Lauri Seitsonen

8.11.2001

Lauri.Seitsonen@hut.fi

Rukmini M. Iyer and Mari Ostendorf, Modeling Long Distance Dependence in Language: Topic Mixtures Versus Dynamic Cache Models, IEEE Trans. speech and audio processing, 7:1, January 1999

Sisällys

- *Aihesekoitemallin idea*
- *Käytäntöä*
- *Kokeellisia tuloksia*
- *Yhteenveto*

Miksi?

- *n-gram -mallit hukkaavat pidemmälle kuin n ulottuvat riippuvuudet*
- *Idea: käytetään yhden mallin sijasta useaa mallia (”aihemallia”), jotka edustavat tarkemmin tiettyä dokumenttityyppiä, ja etsitään sopivat painokertoimet eri malleille*

Aihesekoitemallin idea

- *m:n aiheen sekoitemalli, peruskaava:*

$$P(w_i, \ldots, w_T) = \sum_{k=1}^{m} \lambda_k \left[\prod_{i=1}^{T+1} P_k(w_i \mid w_{i-1}) \right]$$

- *Perusproseduuri:*
  - klusteroi data haluttuun määrään klustereita
  - tee n-gram -mallit klustereista (tasoitus)
  - estimoi sekoitteiden painokertoimet λ_k
- *Parannuksia:*
  - aiheen tasoitus
  - dynaamiset mallit

Klusterointi

- *Klusterointialgoritmi:*
  - Lähtötilassa C* klusteria, tavoite C
  - Yhdistetään ne kaksi klusteria A_i ja A_j , joilla suurin samankaltaisuusmitta S_{ij}
  - Toistetaan kunnes jäljellä C klusteria
- *Samankaltaisuusmitta:*

$$S_{ij} = \sum_{w \in A_i \cap A_j} \frac{N_{ij}}{\left|A^w\right| \left|A_i\right| \left|A_j\right|}$$

$\left|A_i\right|$ sanojen (luokkien) lkm klusterissa

$\left|A^w\right|$ sanan w sisältävien klusterien lkm

$$N_{ij} = \sqrt{\frac{N_i + N_j}{N_i N_j}}$$

N_i artikkelien lkm klusterissa i

Parametrien estimointi

- *Lasketaan aluksi n-gram aihemallit klusteroinnin perusteella*
- *Viritetään mallit EM-algoritmilla:*
  - E-askel: lasketaan todennäköisyydet millä opetuslauseet kuuluvat m aiheeseen
  - M-askel: lasketaan uudelleen n-gram -todennäköisyydet, mukana back-off

Parametrien estimointi

- *E-askel:* $\hat{z}_{ij}^{(p)} = \dfrac{P_j^{(p)}(y_i)\lambda_j^{(p)}}{\sum_{k=1}^{m} P_k^{(p)}(y_i)\lambda_k^{(p)}}$
- *M-askel:* $P_j(w_c \mid w_b) = (1-\phi_b^j)P_j^{ML}(w_c \mid w_b) + \phi_b^j P_j(w_c)$

$$\phi_b^j = \frac{\sum_q \frac{\sum_{i=1}^{N} n_{bq}^i \hat{z}_{ij}^{(p)}}{\sum_{i=1}^{N} n_{bq}^i}}{\sum_q \sum_{i=1}^{N} n_{bq}^i \hat{z}_{ij}^{(p)} + \sum_q \frac{\sum_{i=1}^{N} n_{bq}^i \hat{z}_{ij}^{(p)}}{\sum_{i=1}^{N} n_{bq}^i}}$$

n_{bc}^i bigrammin $\langle w_b, w_c \rangle$ lkm lauseessa i

n_c^i unigrammin w_c lkm lauseessa i

N opetuslauseiden lkm

$$P_j^{ML}(w_c \mid w_b) = \frac{\sum_{i=1}^{N} \hat{z}_{ij}^{(p)} n_{bc}^i}{\sum_q \sum_{i=1}^{N} n_{bq}^i \hat{z}_{ij}^{(p)}} \qquad P_j(w_c) = \frac{\sum_{i=1}^{N} \hat{z}_{ij}^{(p)} n_c^i}{\sum_q \sum_{i=1}^{N} n_q^i \hat{z}_{ij}^{(p)}}$$

Mallin tasoitus

- *Ongelmia:*
  - alioppiminen
    - interpoloidaan aihemallit yleisen koko datalla opetetun n-gram -mallin kanssa
  - aiheisiin kuulumattomat lauseet
    - lisätään aihemalleihin yleinen malli

$$P(w_i,\ldots,w_T)=\sum_{k=1}^{m,G}\lambda_k\prod_{i=1}^{T+1}\left[\alpha_k P_k(w_i\mid w_{i-1})+(1-\alpha_k)P_G(w_i\mid w_{i-1})\right]$$

- *Estimoidaan α_k ja λ_k*

Estimoidaan α_k ja λ_k

- *Piilossa pidetyn datajoukon lauseet klusteroidaan todennäköisimmän aiheen mukaan*
- *Alussa painot tasan, estimoidaan ensin α_k ja sitten λ_k*

$$\alpha_k^{new} = \frac{1}{\sum_{l=1}^{N_k} T_l} \sum_{l=1}^{N_k} \sum_{i=1}^{T_l} \frac{\alpha_k^{old} P_k(w_i^l \mid w_{i-1}^l)}{\alpha_k^{old} P_k(w_i^l \mid w_{i-1}^l) + (1-\alpha_k^{old}) P_G(w_i^l \mid w_{i-1}^l)}$$

N_k lauseiden lkm klusterissa i

T_l sanojen lkm lauseessa l

$$\lambda_k^{new} = \frac{1}{N} \sum_{i=1}^{N} \frac{\lambda_k^{old} P_k(w_1, \cdots, w_{T_i})}{\sum_{j=1,\cdots,m,G} \lambda_j^{old} P_j(w_1, \cdots, w_{T_i})}$$

N lauseiden lkm datajoukossa

Dynaamisuus

- *Opitaan uusista havaituista lauseista, pidetään dokumentin sisällä välimuistia havaituista n-grammeista*
- *Välimuistit kaikille aiheille, myös yleiselle mallille*
  - sanojen frekvenssejä lisätään aihemalleihin sen mukaan miten todennäköisesti havaittu lause kuuluu kyseiseen aiheeseen (eli lisätään osia frekvenssistä)
  - aiheen välimuistimalli estimoidaan kertyneiden lukumäärien mukaan

$$P(w_i,\ldots,w_T)=\sum_{k=1}^{m,G}\lambda_k\prod_{i=1}^{T+1}\left[(1-\mu)P_k^s(w_i\mid w_{i-1})+\mu P_k^c(w_i\mid w_{i-1})\right]$$

P_k^s tasoitettu staattinen malli P_k^c välimuistimalli

Tuloksia

- *Opetusaineisto: North American Business (NAB) news (n. 230 miljoonaa sanaa)*
  - miljoona sanaa sivuun lausetason ja n-gram -tason painojen estimoimiseen
- *Sanasto: 46K*
- *Testiaineisto: 1994 ARPA kehitys- ja evaluointitestijoukot*
- *Käytössä ohjattu adaptoituminen (oikea kirjoitusasu annettu)*

Tuloksia

- *Testijoukon perpleksiteetti*

Test	Adaptation	Trigram model	5-component mixture model
Dev	No	211	165
Dev	Yes	171	141
Eval	No	210	175
Eval	Yes	175	145

- *Testijoukon WER*

Test	Adaptation	Trigram model	5-component mixture model
Dev	No	10,5 %	10,2 %
Dev	Yes	10,3 %	10,2 %
Eval	No	11,5 %	11,0 %
Eval	Yes	11,1 %	10,8 %

Tuloksia 2

- *Opetusaineisto: Switchboard korpus (1500 keskustelua, n. 2,1 miljoonaa sanaa)*
  - 10 000 sanaa kymmenestä keskustelusta sivuun sekoitepainojen estimoimiseksi
- *Sanasto 10K*
- *Testiaineisto: BBN:n sisäinen (seitsemästä keskustelusta koostuva)*
- *Käytössä staattinen malli*
- *Tulokset testijoukolle*

LM	Perplexity	WER
SWBD baseline	118	41,1 %
6-mixture	112	40,6 %

Yhteenveto

- *Lausetason n-gram -sekoitemallilla saavutettavissa parempia tuloksia kuin pelkällä n-gram -mallilla*
  - staattinen malli lauseiden sisällä oleville riippuvuuksille
  - dynaaminen malli myös artikkelin sisällä oleville riippuvuuksille
- *Saattaa kuitenkin kärsiä helposti datan vähyydestä*