Ordlikhet

INF5820 – H2008

Jan Tore Lønning

Institutt for Informatikk
Universitetet i Oslo

17. september

Outline

Glatting

Motivasjon

- Et endelig antall observasjoner er ikke helt representative.
- Det at vi ikke har sett noe utelukker ikke at det finnes.
- Ønsker ikke å multiplisere med 0 (ingen informasjon)
- — og i hvert fall ikke dividere med 0 (tull).

Eksempel

- $$P(s_i) = \frac{\text{count}(s_i, w)}{\text{count}(w)} \text{ for } i = 1, \ldots n$$
- $$\text{Her vil } \sum_{i=1}^{n} P(s_i) = 1$$
- Ønsker ingen $P(s_i) = 0$
- Vi antar at vi har sett alt en gang mer enn vi har:

$$P(s_i) = \frac{\text{count}(s_i, w) + 1}{\text{count}(w) + n}$$

- Egenskapen beholdes $\sum_{i=1}^{n} P(s_i) = 1$ (en sannsynlighet)

Kommentarer

- Å legge til 1 kan bli for mye hvis det er mange usette klasser.
- Anta at $\vec{f}$ er som i boka, men har 20 plasser.
- Da blir det 2^{20}, dvs. mer enn en million forskjellige vektorer.
- Hvis vi nå ville telle count($\vec{f}, w$) og w har noen hundre forekomster, vil det bli alt for mye sannsynlighet på det vi ikke har sett.
- Vi kan legge til mindre enn 1.
- Det finnes mange mer avanserte metoder for glatting.

Trekksannsynligheter

- Litt uklart i boka hva f_j er f.eks. i (20.8).
- Mer presist: det er et vokabular $v_1, v_2, \ldots v_m$.
- En kontekstvektor $\vec{f}$ representerer en kontekst c ved at $f_j = 1$ hvis v_j er tilstede i c, for $j = 1, \ldots m$
- $f_j = 0$ ellers
- count(f_j, s) teller altså opp de forekomstene av s der v_j er i konteksten for $f_j = 1$
- og de forekomstene av s der v_j ikke er i konteksten for $f_j = 0$.
- $$\text{Glatting: } P(f_j \mid s) = \frac{\text{count}(f_j, s) + 1}{\text{count}(s) + 2}$$

Beslutningslisteklassifikator

- Også basert på “supervised” korpuslæring.
- En annen klassifikator
- En rekke tester
- Metoder for å lære denne

Ordboksbaserte metoder

- Lesk-algoritmen
- Hovedidé: Sammenlikn kontkest med det du finner i en definisjonsordbok
- Enkleste form: Sammenlikn ordene i konteksten med orden i definisjonene for ordet vi skal dismbiguere.
- Mer avansert: Sammnlikn definisjonene til ordene i konteksten med definsjonene for ordet som skal disambigueres.

Flere metoder

- Bruk thesaurus, f.eks. WordNet og relasjoner i det.
- Oversettelse:
  - Tospråklig ordbok
  - Korpus av oversettelser
  - Helge Dyvik's speilmetode

Bootstrapping

- Når vi har lite treningsdata
  1. Start med noe treningsdata
  2. Lag en klassifikator fra dette
  3. Bruk dette til å klassifisere et større korpus
  4. Plukk ut de beslutningene du er mest sikker på
  5. Utvid treningsdata med de beslutningene du er mest sikker på
  6. Gjenta

WSD — mange muligheter

Valg

1. Kunnskapskilde: oppmerket enspråklig korpus (supervised methods), umerket enspråklig korpus (unsupervised methods), oversettelseskorpus, enspråklig ordbok, tospråklig ordbok
2. Hvilken informasjon vi vil trekke ut av kilden
3. Hvordan vi vil trekke informasjon ut av kilden
4. Hvordan vi vil bruke informasjonen

- Maskinlæring: 3 + 4

Praksis

Kombinasjon av ulike metoder

Ordlikhet (“similarity”)

Ordlikhet vs. WSD

- Ett ord flere meninger
  - WSD
- Flere ord (nesten) samme mening
  - Ordlikhet, “word similarity”
- Men: WSD går på en spesifikk forekomst av et ord.
- Ordlikhet går på mønster over alle forekomster, en leksikalsk egenskap

Likhet — “similarity”

- (Nesten) synonymer
- Men også hyponymer eller søsken i et hyponym-hierarki, eller mer generelt (nære) slektninger.
- De fleste metoder vil finne ord med liknende egenskaper men svært forskjellig innhold: *vått–tørt, stor–liten, blå–rød*

Thesaurusbaserte metoder

- Kilde: en thesaurus, f.eks. Word Net
- Et uttrykk for nærhet i thesaurusen som vi vil bruke
- Algoritmer/maskinlæringsteknikker for å finne hvilke ord som står i nærhetsrelasjonen
- Boka nevner 5 forskjellige algoritmer

Distribusjonbaserte metoder

Se nærmere på disse fordi

- Liknende teknikker som i WSD
- og til IR
- Krever bare et (umerket) korpus

Hovedidé:

- “You shall know a word by the company it keeps!” (Firth 1957)
- Liknende ord opptrer i liknende kontekster.
- Jfr. WSD: samme mening opptrer i liknende kontekster.

Noen typer valg

- Hvordan definerer vi kontekst (jfr. WSD)?
- Hvordan bruker vi kontekstene?
  - Hvordan vekter vi kontekstvektorene?
  - Hvordan beregner vi avstanden mellom kontekstvektorene?

Samforekomstvektor

Valg

- Hvor stort vindu? 2 ord? 1000 ord?
- Hvilke ord i vinduet?
  - Alle ord unntatt stoppord, eller
  - Nøkkelord (jfr. bokas WSD)?
    - → Går ikke siden vi ikke vet hvilke ord som hører sammen.
  - Ord som står i bestemte relasjoner til ordene vi ser på

Samforekomstvektor

Grammatiske relasjoner

I discovered dried tangerines.

Ord	funksjon	til
discover	subject	I
tangerine	object-of	discover
dried	adj-mod-of	tangerine
I	subj-of	discover
tangerine	adj-mod	dried

Relasjoner av ulike typer

- Grammatiske — avhenger delvis av grammatisk modell
- Semantiske

Kontekstassosiasjon

- En feature $f = (r, w')$, eksempel: (obj-of *discover*)

$$\text{Skulle tro at } \text{assoc}_{\text{prob}}(w, f) = P(f \mid w) = \frac{\text{count}(f, w)}{\text{count}(w)}$$

- Slik vi gjorde med WSD.
- Ikke galt, men andre virker bedre

Pointwise mutual information

- $$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$
- Denne er 1 hvis de er uavhengige, > 1 hvis det forekommer ofte sammen
- $$\text{assoc}_{\text{PMI}}(w,f) = \log_2 \frac{P(w,f)}{P(w)P(f)}$$

Kontekstassosiasjon

- Andre mål:
  - Lin association measure
  - t-test
- Vi skal ikke/trenger ikke lære alle i detaljer:
  - Skjønne det generelle bildet
  - Trenger å skjønne noe i detalj
  - Bruker tildels pakker, andres programmer
  - Være klar over at det kan være mer som skjer enn det vi har lært
  - Være i stand til å gå i detalj når vi trenger det

Vektorlikhet

Hva skal vi se på?

- Differanse, eller
- Skalarprodukt?

Differanse

- Lengden av avstandsvektoren
- $$\text{Manhatten: } \sum_{i=1}^{N} |x_i - y_i|$$
- $$\text{Euklidsk: } \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

Produkt

- $$\text{sim}_{\text{dot-product}}(\vec{v}, \vec{w}) = \vec{v} \cdot \vec{w} = \sum_{i=1}^{N} v_i \times w_i$$
- $$\text{sim}_{\text{cosine}}(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}|\,|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2 \sum_{i=1}^{N} w_i^2}}$$
- =1 hvis de peker i samme retning