HMM-tagging

INF4820 – H2008

Jan Tore Lønning

Institutt for Informatikk
Universitetet i Oslo

30. september

Outline

1. Tagging
2. Bayes teorem
3. Skjulte Markov-modeller
4. HMM-tagging
5. Viterbi-algoritmen

Outline

1 Tagging
2 Bayes teorem
3 Skjulte Markov-modeller
4 HMM-tagging
5 Viterbi-algoritmen

Flertydighet

Example

```
"<fisker>"
      "fisk" subst appell mask ub fl @løs-np
      "fisker" subst appell mask ub ent @løs-np
      "fiske" subst appell nøyt ub fl @løs-np
      "fiske" verb pres tr1 i1 pa1 @fv
```

Example

Entydig i en sammenheng

- Hvor mange fisker fikk du?
- Ola fisker laks.
- Mange fisker fisker.

Hvorfor?

Nyttig for

- Grovparsing, “chunking”
- NER: (“named entity recogition”)
- Ordmeningsentydiggjøring (“Word sense disambiguation”, WSD)
- Lemmatisering, og dermed “information retrieval”, IR.
- Kjenner vi taggen til et ord kan vi si mer om ordene rundt:
- og hjelp til talegjenkjenning
- Talesyntese, ek. NO: *passasjer*
- lingvistisk forskning
- osv.

Hvordan?

1. Regelbaserte
   - Constraint Grammar:
     - EngCG
     - OSLO-Bergen-taggeren
     - ...
   - ...
2. Stokastiske (statistikkbaserte)
   - HMM-baserte
   - Maksimumentropibaserte modeller
   - ...
3. Transformasjonsbaserte

Oslo-Bergen-taggeren

http://omilia.uio.no:8050/cl/cgp/test.html

Example (Regler av typen)

- En setning skal ha et tensed verb.
- Hvis det står en artikkel umiddelbart til venstre for ordet og et verb til høyre, så er ikke dette et verb.

Prinsipper

- Sett inn alle mulige POS-tagger for alle ord.
- Bruk reglene til å fjerne tagger
- Ikke fjern alle taggene for et ord.
- Tillat resterende flertydigheter

Outline

1 Tagging

2 Bayes teorem

3 Skjulte Markov-modeller

4 HMM-tagging

5 Viterbi-algoritmen

Bayes teorem

- $$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \text{ betinget sannsynlighet}$$
- $P(A \cap B) = P(A \mid B)P(B)$ produktregelen
- $P(A \cap B) = P(B \mid A)P(A)$
- $P(A \mid B)P(B) = P(B \mid A)P(A)$
- $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \text{ Bayes' teorem}$$

Bayes teorem

Example (fra Wikipedia)

- 40% jenter, 60% gutter.
- Alle guttene bruker bukser.
- 50% av jentene bruker bukser.
- Hva er sannsynligheten for at en som går i bukser er jente?

Example (Løsning med Bayes)

- Sannsynlighet for at noen er jenter, $P(A) = 0,4$
- Sannsyn. for at en jente bruker bukser, $P(B \mid A) = 0,5$.
- Sannsyn. for å gå i bukser, $P(B) = P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A}) = 0,5 \times 0,4 + 1 \times 0,6 = 0,8$
- $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0,5 \times 0,4}{0,8} = 0,25$

Litt sjargong

- $P(A \mid B) = \frac{P(B|A)}{P(B)} P(A)$
- $P(A)$ prior sannsynlighet
  - I eksempelet: sannsynligheten for at det er en jente før vi vet hun går i bukser.
- $P(A \mid B)$ posterior sannsynlighet
  - I eksempelet: sannsynligheten for at det er en jente etter at vi har fått vite at vedkommende går i bukser.

Argmax-notasjon

Definition (Argmax)

$x_0 = \arg\max_x f(x)$ vil si at $f(y) \leq f(x_0)$ for alle y

Argmax og Bayes

$$\begin{aligned} x_0 &= \arg\max_x P(x \mid B) \\ &= \arg\max_x \frac{P(B \mid x)}{P(B)} P(x) \\ &= \arg\max_x P(B \mid x) P(x) \\ &= \arg\max_x [\log P(B \mid x) + \log P(x)] \end{aligned}$$

Outline

1 Tagging

2 Bayes teorem

3 Skjulte Markov-modeller

4 HMM-tagging

5 Viterbi-algoritmen

Markov-egenskapen

Markov-egenskapen

$$P(X_{n+1} = x \mid X_n = x_n, \ldots X_1 = x_1) = P(X_{n+1} = x \mid X_n = x_n)$$

- Neste tilstand avhenger bare av nåværende tilstand, ikke av fortidige.

- $$\begin{aligned} & P(X_1, X_2, \ldots, X_T) \\ = \; & P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2) \cdots P(X_T \mid X_1, \ldots X_{T-1}) \\ = \; & P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2) \cdots P(X_T \mid X_{T-1}) \\ = \; & P(X_1) \prod_{t=1}^{T-1} P(X_{t+1} \mid X_t) \end{aligned}$$

Markov-modell

- $a_{ij} = P(X_{t+1} = s_j \mid X_t = s_i)$
- Da vil $\sum_{j=n}^{N} a_{ij} = 1$ for alle i
- $\pi_i = P(X_1 = s_i)$, alternativt spesielle start og sluttilstander, s_0, s_F, og $a_{0i} = \pi_i$.

Sammenheng med FSA

- Dette svarer til en endeligtilstandsautomat (FSA) der
  - Kantene er merket med sannsynligheter, summen av sannsynligheter ut i fra en node er 1.
  - Inputalfabetet er lik navnet på tilstandene og en input svarer til tilsvarende gjennomløp av tilstander.
- Alternativ formulering
  - Deterministisk FSA med sannsynligheter på kantene, dvs.
  - ut i fra en node går det nøyaktig en kant for hver inputsymbol (sannsynligheten for denne kan være null)

Skjult Markov-modell

- Det vi observerer
- Det underliggende nivå (“årsakene”)

Skjult Markov-modell

Definition

- En endelig mengde tilstander, $Q = \{q_1, q_2, , \ldots, q_N\}$
- Et endelig signal alfabet, $V = \{V_1, v_2, , \ldots, v_V\}$
- En matrise av overgangssannsynligheter, $A = \{a_{ij} \mid 1 \leq i \leq N, 1 \leq j \leq N\}$
- En mengde emisjonssannsynligheter, $B = \{b_i(v_t) \mid 1 \leq i \leq N, v_t \in V\} \quad (b_{ij} = b_i(v_j))$
- Spesielle start og sluttilstander, q_0, q_F.
  - Disse har ikke emisjoner.
  - Det er overganger ut fra (men ikke inn i) q_0,
  - og inn i (men ikke ut fra) q_F

(Litt unøyaktig definisjon i J&M.)

Illustrasjoner

- Grafikk fra Blei (og Manning & Schütze)
- Eisners iskrem fra J& M

"Enkle" spørsmål:

- Spm. Gitt en sekvens av skjulte tilstander $q_{i_1} q_{i_2} \ldots q_{i_n}$, hva er sannsynligheten for en outputsekvens $v_{i_1} v_{i_2} \ldots v_{i_n}$?
- Svar $P(v_{i_1} v_{i_2} \ldots v_{i_n} \mid q_{i_1} q_{i_2} \ldots q_{i_n}) = \prod_{j=1}^{n} b_{i_j}(v_{i_j})$
- Spm. Gitt en outputsekvens $v_{i_1} v_{i_2} \ldots v_{i_n}$, hva er sannsynligheten for en sekvens av tilstander $q_{i_1} q_{i_2} \ldots q_{i_n}$?
- Svar $P(q_{i_1} q_{i_2} \ldots q_{i_n} \mid v_{i_1} v_{i_2} \ldots v_{i_n}) =$
$\frac{P(v_{i_1} v_{i_2} \ldots v_{i_n} | q_{i_1} q_{i_2} \ldots q_{i_n}) P(q_{i_1} q_{i_2} \ldots q_{i_n})}{P(v_{i_1} v_{i_2} \ldots v_{i_n})} = \frac{\prod_{j=1}^{n} b_{i_j}(v_{i_j}) \prod_{j=1}^{n} a_{i_{(j-1)} i_j}}{P(v_{i_1} v_{i_2} \ldots v_{i_n})}$

Mer kompliserte spørsmål

Gitt en bestemt HMM

1. Sannsynlighet ("likelihood"): Hva er sannsynligheten for en outputsekvens $v_{i_1} v_{i_2} \ldots v_{i_n}$?
2. Dekoding: Gitt en outputsekvens $v_{i_1} v_{i_2} \ldots v_{i_n}$, hva er den mest sannsynlige sekvensen av tilstander $q_{i_1} q_{i_2} \ldots q_{i_n}$?
3. Trening: Gitt en sekvens av observasjoner $v_{i_1} v_{i_2} \ldots v_{i_n}$ og en mengde tilstander, finn overgangs-, A, og emisjonssannsynlighetene B.

Outline

1 Tagging

2 Bayes teorem

3 Skjulte Markov-modeller

4 HMM-tagging

5 Viterbi-algoritmen

Stokastisk tagging

- Gitt en sekvens av ord $w_1 w_2 \ldots w_n$ (som vi vil skrive w_1^n)
- Hva er den mest sannsynlige taggsekvensen $t_1^n = t_1 t_2 \ldots t_n$?
-

$$\begin{aligned} \arg\max_{t_1^n} P(t_1^n \mid w_1^n) &= \arg\max_{t_1^n} \frac{P(w_1^n \mid t_1^n) P(t_1^n)}{P(w_1^n)} \\ &= \arg\max_{t_1^n} P(w_1^n \mid t_1^n) P(t_1^n) \end{aligned}$$

Ordsannsynlighetene

- Uttrykket vi ser på: $P(w_1^n \mid t_1^n)P(t_1^n)$
- $P(w_1^n \mid t_1^n) =$
  $P(w_n \mid w_1^{n-1} t_1^n)P(w_{n-1} \mid w_1^{n-2} t_1^n) \cdots P(w_1 \mid t_1^n)$
- Antar $P(w_i \mid w_{1,i} t_1^n) = P(w_i \mid t_i)$
- Altså (feilaktig) at et ord ikke avhenger av ordene før og etter, bare av sin tagg.
- $P(w_1^n \mid t_1^n) =$
  $P(w_n \mid t_n)P(w_{n-1} \mid t_{(n-1)}) \cdots P(w_1 \mid t_1) = \prod_{i=1}^{n} P(w_i \mid t_i)$

Tagg-sannsynlighetene

- Antar Markov-egenskapen $P(t_{(j+1)} \mid t_1^j) = P(t_{j+1} \mid t_j)$
- $$\begin{aligned} P(t_1^n) &= P(t_1)P(t_2 \mid t_1)P(t_3 \mid t_1^2)\cdots P(t_n \mid t_1^{(n-1)}) \\ &= P(t_1)P(t_2 \mid t_1)P(t_3 \mid t_2)\cdots P(t_n \mid t_{n-1}) \\ &= P(t_1)\prod_{i=1}^{n-1} P(t_{i+1} \mid t_i) \\ &= \prod_{i=1}^{n} P(t_i \mid t_{i-1}) \end{aligned}$$
- (Hvis $P(t_1 \mid t_0)$ er sannsynligheten for å starte i t_1)

Som gir

- $$\begin{aligned} P(w_1^n \mid t_1^n)P(t_1^n) &= \prod_{i=1}^{n} P(w_i \mid t_i) \prod_{i=1}^{n} P(t_i \mid t_{i-1}) \\ &= \prod_{i=1}^{n} P(w_i \mid t_i)P(t_i \mid t_{i-1}) \end{aligned}$$

- $$\begin{aligned} \hat{t}_1^n &= \arg\max_{t_1^n} P(t_1^n \mid w_1^n) \\ &= \arg\max_{t_1^n} P(w_1^n \mid t_1^n)P(t_1^n) \\ &= \arg\max_{t_1^n} \prod_{i=1}^{n} P(w_i \mid t_i)P(t_i \mid t_{i-1}) \end{aligned}$$

HMM-tagging

- Stokastisk tagging kan altså sees på som en HMM, der
- Ord er det vi observerer
- Tagger er de skjulte tilstandene
- Vi har gjort en del litt grove antagleser

Outline

1 Tagging

2 Bayes teorem

3 Skjulte Markov-modeller

4 HMM-tagging

5 Viterbi-algoritmen

Tilbake til de store spørsmål

Gitt en bestemt HMM

1. Sannsynlighet ("likelihood"): Hva er sannsynligheten for en outputsekvens $v_{i_1} v_{i_2} \ldots v_{i_n}$?
2. Dekoding: Gitt en outputsekvens $v_{i_1} v_{i_2} \ldots v_{i_n}$, hva er den mest sannsynlige sekvensen av tilstander $q_{i_1} q_{i_2} \ldots q_{i_n}$?
3. Trening: Gitt en sekvens av observasjoner $v_{i_1} v_{i_2} \ldots v_{i_n}$ og en mengde tilstander, finn overgangs-, A, og emisjonssannsynlighetene B.

Dekoding

- Finn $\hat{t}_1^n = \arg\max_{t_1^n} \prod_{i=1}^n P(w_i \mid t_i) P(t_i \mid t_{i-1})$
- Naiv algoritme: Regn ut $\prod_{i=1}^n P(w_i \mid t_i) P(t_i \mid t_{i-1})$ for alle mulige t_1^n og sammenlikn.
- Hvor mange slike t_1^n er det?
- m^n der m er antall tagger/tilstander
- Vi trenger smartere algoritme!

Egenskap

- Gitt ordsekvens (den ligger fast)
- Anta at vi har to tagg-sekvenser t_1^n og s_1^n.
- Anta at de er like fra en k, altså $t_j = s_j$ for $k \leq j \leq n$
- Da vil $\prod_{i=1}^{n} P(w_i \mid s_i)P(s_i \mid s_{i-1}) =$
  $\prod_{i=1}^{(k-1)} P(w_i \mid s_i)P(s_i \mid s_{i-1}) \prod_{i=k}^{n} P(w_i \mid s_i)P(s_i \mid s_{i-1}) =$
  $\prod_{i=1}^{(k-1)} P(w_i \mid s_i)P(s_i \mid s_{i-1}) \prod_{i=k}^{n} P(w_i \mid t_i)P(t_i \mid t_{i-1})$
- Altså det holder å sammenlikne
  $\prod_{i=1}^{(k-1)} P(w_i \mid s_i)P(s_i \mid s_{i-1})$ med
  $\prod_{i=1}^{(k-1)} P(w_i \mid t_i)P(t_i \mid t_{i-1})$ siden resten er identisk.

Hovedidé

- I stedet for $\prod_{i=1}^{n} P(w_i \mid t_i)P(t_i \mid t_{i-1})$ for først en sekvens t_1^n, deretter for neste sekvens s_1^n, osv.

1. Regn først ut $\prod_{i=1}^{1} P(w_i \mid t_i)P(t_i \mid t_{i-1})$ for alle sekvenser t_1^1 og ta vare på resultatet.
2. Deretter bruk dette til å regne ut $\prod_{i=1}^{2} P(w_i \mid t_i)P(t_i \mid t_{i-1})$ for alle sekvenser t_1^2 og ta vare på resultatet.
3. Og generelt på trinn k bruk $\prod_{i=1}^{(k-1)} P(w_i \mid t_i)P(t_i \mid t_{i-1})$ til å regne ut $\prod_{i=1}^{k} P(w_i \mid t_i)P(t_i \mid t_{i-1})$.
4. Men på hvert trinn, trenger vi ikke å se på alle mulige sekvenser, bare den beste sekvensen så langt for hver tilstand.
5. På hvert trinn k og hver tilstand t_j, ta vare på den beste tagsekvensen så langt.

Trellis

- Lengden av ordsekvensen: n, antall tilstander: m.
- "Trellis" (espealier) som har dimensjon $(m+2) \times n$, og
  - $v(i,j)$ sannsynligheten for tilstand q_i etter å ha sett w_j
  - $f(i,j)$ trellisruten vi kommer fra
- På trinn j, for hver q_i
  - Se på $v(l, j-1) \times a_{li}$ for alle l (der $1 \leq l \leq m$).
  - Velg den l som gir størst verdi, kall den k ($k = \arg\max_{t_1^n} v(l, j-1) \times a_{li}$)
  - La $f(i,j) = k$
  - $v(i,j) = v(k, j-1) \times a_{kj} \times b_i(w_j)$
- Må se spesielt på begynnelse og slutt: på gruppetimen!