

# Tentamen: Taalverwerking en Informatie-Ontsluiting

Opleiding: Kunstmatige Intelligentie, OWI Informatiewetenschappen

Docenten: K. Sima'an/R. Scha  
6 juli 2006, 13:30-16:30

## **Tussentoets instructies:**

Als u de tussentoets (Blok A) met voldoende gevolg hebt afgelegd (tentamencijfer  $\geq 6.0$ ), kunt u slechts de vragen over Blok B hieronder beantwoorden. U kunt de tussentoets **niet** herkansen als u reeds een tentamencijfer  $\geq 6.0$  hebt gehaald. **Vermeld bovenaan op vel 1 van uw tentamen-uitwerking of u de tussentoets gehaald hebt.**

## **Algemene instructies:**

Dit is *geen* open boek tentamen. Schrijf *duidelijk* en wees *kort en bondig*, maar geef uitleg waar nodig. Zorg ervoor dat uw naam, collegekaartnummer en email-adres staan op elk vel papier dat u inlevert. Mobiele telefoons zijn niet toegestaan en dienen tijdens het tentamen opgeborgen en uit het zicht te zijn.

VEEL SUCCES!

# 1 Blok A (Twee vragen)

Vraag 1. Deze vraag gaat over statistiek en taalmodellen.

- A. Gegeven is een eindige verzameling woorden (een vocabulair)  $V = \{a, b, c, d\}$ .
- (a) Geef een voorbeeld van een taal over  $V$
  - (b) Geef een voorbeeld van een taal-model over  $V$
- B.  $A$  en  $B$  zijn twee events waarvan bekend is dat

$$\text{Vergelijkingen : } P(B) = P(B|A) \quad P(A) = P(A|B)$$

- A1. Hoe wordt de verhouding tussen  $A$  en  $B$  genoemd in de waarschijnlijkheidsleer?
- A2. Maak gebruik van de vergelijkingen hierboven om een formule voor  $P(A, B)$  af te leiden (in termen van  $P(A)$  en  $P(B)$ ).
- C. Gegeven is de volgende tabel met de frequenties voor bigrammen over de alfabet  $\{a, b, c\}$ :

	a	b	c
a	2	0	2
b	4	4	2
c	0	10	0

De interpretatie van de tabel is als volgt: de horizontale symbolen  $a, b, c$  zijn opvolgers (tweede in de bigram) van de verticale symbolen. Dus, de event "b volgt c" (bigram  $\langle c, b \rangle$ ) komt 5 keer voor. Opdrachten:

- B1. Bereken de relatieve frequentie schattingen voor  $P(x | y)$  voor alle  $x$  en  $y$  uit de verzameling  $\{a, b, c\}$ . Schrijf deze schattingen in dezelfde soort tabel.
- B2. Bereken de relatieve frequentie schattingen voor  $P(x, y)$  voor alle  $x$  en  $y$  uit de verzameling  $\{a, b, c\}$ . Schrijf deze schattingen in dezelfde soort tabel.
- B3. Pleeg een Add- $\lambda$  Smoothing (wanneer  $\lambda = 1$ ) en bereken opnieuw de schattingen van  $P(x | y)$ . Schrijf deze schattingen in dezelfde soort tabel.
- D. Geef een formule voor een  $0^{de}$ -orde Markov model voor  $P(w_1, \dots, w_n | x)$ , waarbij  $w_1, \dots, w_n$  en  $x$  zijn allen woorden uit een eindige vocabulair.

**Vraag 2.** Deze vraag gaat over Part-of-Speech tagging.

Gegeven is de woord-sequentie  $w_1, \dots, w_n$  over een eindig vocabulair  $V$  en een eindige verzameling van Part-of-Speech (POS) tags  $T$ . Het probabilistische Markov POS-tagging model is bedoeld om de meest waarschijnlijke tag-sequentie  $t_1, \dots, t_n$  (waar  $t_i \in T$  voor alle  $1 \leq i \leq n$ ) te vinden voor deze zin als volgt:

$$(Formule1) \quad \operatorname{argmax}_{t_1^n} P(t_1, \dots, t_n | w_1, \dots, w_n)$$

- A. Geef een wiskundige formule die een approximatie vormt van Formule 1 zodat de waarschijnlijkheid van ieder paar  $\langle t_i, w_i \rangle$  geconditioneerd is op het voorgaande paar  $\langle t_{i-1}, w_{i-1} \rangle$  (neem aan dat de waarschijnlijkheid van het eerste paar geconditioneerd is op de start symbolen van de sequentie).
- B. Geef een formule voor het berekenen van de waarschijnlijkheid van een zin  $P(w_1, \dots, w_n)$  aan de hand van de waarschijnlijkheden van het POS tagging model  $P(w_1, \dots, w_n, t_1, \dots, t_n)$ .
- C. Hier is een tagged corpus

```
<s>/START John/NNP wants/VB a/DT pizza/NN ./DOT
<s>/START A/DT dog/NN ate/VB chocolate/NN ./DOT
<s>/START A/DT dog/NN ate/VB chocolate/NN ./DOT
```

Het spreekt voor zich dat de POS-tags zijn START, NNP, VB, NN, DOT.

- (a) Geef de relatieve frequentie schattingen van de volgende waarschijnlijkheden op grond van dit corpus:

- $P(\text{John} \mid \text{NNP})$ ,  $P(\text{pizza} \mid \text{NN})$ ,  $P(\text{pizza} \mid \text{NNP})$ ,  $P(\text{ate} \mid \text{VB})$   
 $P(. \mid \text{DOT})$ ,  $P(\langle s \rangle \mid \text{START})$
- $P(\text{NNP} \mid \text{START})$ ,  $P(\text{VB} \mid \text{NNP})$ ,  $P(\text{VB} \mid \text{NN})$ ,  $P(\text{DOT} \mid \text{NN})$ ,  
 $P(\text{NN} \mid \text{DT}, \text{VB})$

- (b) Maak gebruik van de relevante schattingen uit onderdeel (C.a) van deze vraag om de waarschijnlijkheid van de volgende “ge-tagged zin”

```
<s>/START John/NNP ate/VB pizza/NN ./DOT
```

te schatten aan de hand van een 1<sup>ste</sup>-orde Markov taal-model en van een *lexical-model* waarin de waarschijnlijkheid van ieder woord slechts op eigen POS-tag is ge-conditioneerd.

## 2 Blok B (Drie vragen)

### Vraag 3.

- a. Teken boomdiagrammen voor de syntactische structuur van de zinnen:

"Sommige mannen kijken naar een huis."

"Kijken alle mannen naar een huis?"

"Naar welk huis kijken alle mannen?"

- b. Specificeer een herschrijfgrammaticaatje (met een lexicon) dat (o.a.) de structuren genereert die u specificeerde in het antwoord op vraag a.
- c. Voeg semantische interpretatie-regels toe aan de herschrijfregels en het lexicon.
- d. Laat zien welke formule er op grond van deze regels afgeleid kan worden voor de zin:

"Alle mannen kijken naar een huis."

### Vraag 4.

- a. Noem 3 soorten taalhandelingen.
- b. Geef een voorbeeld van een indirecte taalhandeling.

### Vraag 5. Beschouw een document-collectie bestaande uit de documenten "Oranje wint WK", "Oranje verliest", "WK niet leuk".

- a. Bereken de cosinus-afstand van de query "Wie wint WK" tot elk van deze documenten.
- b. Maak de berekening opnieuw, maar geef nu gewichten aan de termen volgens de  $tf * idf$  maat.