

Tentamen: Taalverwerking en Informatie Ontsluiting (Deel I)

Docent: Dr. K. Sima'an
30 Maart 2006, 13:30–16:00
Let op: duur is 2.5 uur!!

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!

Vraag 1. Deze vraag gaat over waarschijnlijkheidsleer en statistiek:

- Geef *twee verschillende* formules (met uitleg) voor de waarschijnlijkheid $P(A | B)$ gebruikmakend van de volgende bekende waarschijnlijkheden: $P(A, B)$, $P(B | A)$, $P(A)$ en $P(B)$.
- A en B zijn twee independent (onafhankelijke) events in een gegeven event space. Geef een formule voor het berekenen van de kans $P(A \cap B)$ in termen $P(A)$, $P(B)$ en $P(A | B)$ (maak zelf de keuze uit deze waarschijnlijkheden).
- A_1 en A_2 vormen een partitie van de event B zodat $A_1 \cap A_2 = \emptyset$ en $A_1 \cup A_2 = B$. Geef een formule voor de waarschijnlijkheid $P(B)$ in termen van $P(A_1)$ en $P(A_2)$. Geef een bewijs van deze formule gebruikmakend van de axioma's van waarschijnlijkheidsleer (en verzamelingen theorie).
- Gegeven is de volgende tabel met de frequenties voor bigrammen over de alfabet $\{a, b, c\}$:

| | a | b | c |
|---|---|---|---|
| a | 1 | 0 | 1 |
| b | 2 | 2 | 1 |
| c | 0 | 5 | 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. Opgaven:

1. 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.
2. Pleeg een Add- λ Smoothing ($\lambda = 1$) op de frequenties uit deze tabel en bereken opnieuw de genoemde schattingen. Schrijf deze schattingen in dezelfde soort tabel.

Vraag 2. Deze vraag gaat over spellingscorrectie.

Zoals gewoonlijk, de notatie a_1^n staat voor de sequentie a_1, \dots, a_n .

- Stel u krijgt een text in het Nederlands waarin een sequentie van letters uit het alfabet (dus “[a-zA-Z]+”) is aanwezig die geen woord vormt in het Nederlands (een zogenaamd “typo”). Welke bronnen zou u raadplegen om een verzameling mogelijke correcties te maken van dat “typo”? (Wees kort en bondig!). Beschrijf tevens een u bekende methode die gebaseerd is op de aanname dat slechts 1 letter de oorzaak is van het typo.
- Stel voor dat C staat voor de verzameling mogelijke correcties van het “typo” w (w vormt geen woord in de taal), en dat de zin waarin w voorkomt in de text is

$$U = u_1, \dots, u_m, X, v_1, \dots, v_n$$

De volgende afleiding resulteert in een uitdrukking die uit twee delen bestaat

$$\arg \max_{X \in C} P(u_1^m, X, v_1^n | u_1^m, w, v_1^n) = \arg \max_{X \in C} P(u_1^m, w, v_1^n | u_1^m, X, v_1^n) P(u_1^m, X, v_1^n)$$

Welk deel staat voor het taakmodel (spellingsmodel) en hoe heet het andere deel?

- Geef een afleiding dat resulteert in een 1^{ste}-orde Markov approximatie van $P(u_1^m, X, v_1^n)$. Geef aan waar en welke onafhankelijkheids-aannames zijn gemaakt tijdens de afleiding.

Vraag 3. (40%) Deze vraag gaat over Markov taalmodellen over woord sequenties.

Zoals gewoonlijk, de notatie a_1^n staat voor de sequentie a_1, \dots, a_n .

Gegeven is een eindige vocabulair V van woorden. Een taalmodel (language model) over dit vocabulair kan gedefinieerd worden middels

$$P : V^+ \rightarrow [0, 1] \quad \sum_{s \in V^+} P(s) = 1$$

Dus, gegeven een zin w_1, \dots, w_n over V , wordt de waarschijnlijkheid van deze zin volgens dit model geschreven als $P(w_1, \dots, w_n)$.

A. Maak gebruik van regels uit de waarschijnlijkheidsleer om de term $P(w_1, \dots, w_n)$ te herschrijven naar een gelijkwaardig proces waarin de waarschijnlijkheid van ieder woord wordt geconditioneerd op de drie voorgaande woorden. *Geef een formule hiervoor!*. Welke Markov orde model is dit?

B. Hier is een mini-corpus over de vocabulair $V = \{a, b, c, \langle s \rangle, \langle /s \rangle\}$:

$\langle s \rangle a b a b a \langle /s \rangle$

$\langle s \rangle a c a c a c a \langle /s \rangle$

$\langle s \rangle a a a \langle /s \rangle$

- Welke waarschijnlijkheden moeten geschat worden in een 1^{ste}-orde Markov model getrained over dit corpus? (hint: je kan dit in een tabel weergeven).
- Wat zijn de *relatieve frequentie* schattingen van de waarschijnlijkheden genoemd in onderdeel (i) van deze vraag in het corpus?
- Wat is de formule voor de waarschijnlijkheid van de zin $\langle s \rangle a b a c b a a a \langle /s \rangle$ volgens de voorgaande schattingen (maak gebruik van de statistieken zoals in onderdeel ii van deze vraag)?

Vraag 4. Deze vraag gaat over Part-of-Speech tagging.

Gegeven een 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:

$$(Formule2) \quad \operatorname{argmax}_{t_1^n} P(t_1, \dots, t_n | w_1, \dots, w_n)$$

A. Geef (wees kort en bondig) ...

1. een definitie in eigen woorden van de term "ambiguiteit" m.b.t. POS-tagging.
2. een voorbeeld van een woord uit het Engels dat ambigue is m.b.t. POS-tagging.
3. voor iedere mogelijke POS-tag van dat woord een voorbeeld zin waarin het voorkomt.

B. Maak gebruik van de regel van Bayes om Formule 2 te herschrijven naar twee modellen: het taal-model en het lexicale-model. Welk deel staat voor het taalmodel, en welk deel voor het lexicale model?

C. Completeer nu de approximatie van het *lexicale-model* zodat elk woord slechts is afhankelijk van zijn eigen POS-tag.

Completeer ook de approximatie van het taal-model over POS-tag sequencies door gebruik te maken van een 2^{de}-orde Markov aanname.

D. 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 $P(w_1, \dots, w_n, t_1, \dots, t_n)$.

E. 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
```

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

1. Geef de relatieve frequentie schattingen van de volgende waarschijnlijkheden op grond van dit corpus:

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

2. Maak gebruik van de relevante schattingen die je hebt verkregen in onderdeel (E.1) 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^{ste}-orde Markov taal-model en van een *lexical-model* waarin de waarschijnlijkheid van ieder woord slechts op eigen POS-tag is ge-conditioneerd.