

Tentamen: Natuurlijke Taalverwerking

Docent: Dr. K. Sima'an
29 October 2004, 9.30–12.00

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. (15%) Deze vraag gaat over reguliere uitdrukkingen. Construeer een reguliere uitdrukking die matcht met

1. tenminste één **a** gevolgd door een willekeurig aantal **b**'s en eindigend in één of meer **c**'s
2. één of meer keer achtereenvolgende woorden uit de verzameling {**ja, nee**}
3. vijf willekeurige getallen, inclusief "newline"
4. de taal $\{a^n \mid n \geq 3\}$
5. alle getallen die bestaan uit 2 cijfers, gevolgd door een punt, dan 6 cijfers na het punt, vervolgens een positieve/negatieve exponent met twee cijfers. Voorbeeld: $27.000000e-19$.

Vraag 2. (15%) 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”?
- Stel voor dat C staat voor de verzameling mogelijke correcties van het “typo” w , en dat de zin waarin w voorkomt in de text is $U = u_1^m, w, v_1^n$. Geef in het kort, in eigen woorden, een uitleg van de uitdrukking in Formule 1 hieronder:

$$(Formule 1) \quad \arg \max_{X \in C} P(u_1^m, X, v_1^n \mid u_1^m, w, v_1^n)$$

Hoe correspondeert dit met het “Noisy-Channel Model”?

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

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

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. (30%) 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. Gegeven een verzameling zinnen $Z \subseteq V^+$.
Schrijf een wiskundige formule dat staat voor de "meest waarschijnlijke zin in Z "?
- B. 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 twee voorgaande woorden.
- C. Gegeven is een corpus van zinnen (sequenties van woorden over V). Wat zijn de formules om ieder van de waarschijnlijkheden (i) $P(w_n | w_{n-1})$ en (ii) $P(w_{n-1}, w_n)$ te schatten middels relatieve frequentie?
- D. Wat is de 2^{de}-orde Markov approximatie van $P(w_1, \dots, w_n)$? Geef een formule hiervoor.!
- E. Hier is een mini-corpus over de vocabulair $V = \{a, b, c, \langle s \rangle, \langle /s \rangle\}$:

$$\begin{aligned} &\langle s \rangle a b a b a \langle /s \rangle \\ &\langle s \rangle a c a c a c a \langle /s \rangle \\ &\langle s \rangle a a a \langle /s \rangle \end{aligned}$$

- i. Welke waarschijnlijkheden moeten geschat worden in een 1^{ste}-orde Markov model getrained over dit corpus? (hint: je kan dit in een tabel weergeven, maar dat hoeft niet per se).
- ii. Welke n-grammen zijn gemoeid in onderdeel (i) van deze vraag in het corpus en wat zijn hun frequenties in het corpus?
- iii. Wat zijn de *relatieve frequentie* schattingen van de waarschijnlijkheden genoemd in onderdeel (i) van deze vraag in het corpus?
- iiii. 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 twee voorgaande schattingen (maak gebruik van de statistieken zoals in onderdeel iii van deze vraag)?

Vraag 4. (40%) 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 ...

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.

D. Completeer ook de approximatie van het taal-model over POS-tag sequencies door gebruik te maken van de volgende Markov aannames: (i) 0-orde en (ii) 2^{de}-orde.

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:

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

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.