Tentamen Natuurlijke Taalverwerking

Informatiekunde en Kunstmatige Intelligentie

28-6-2002

1. In een woordenlijst van de 60.000 meest frequente woorden in een corpus van 145 miljoen woorden, ontbreken toch nog 3,5% van de woorden die in een ongebruikt (ongezien) deel van het corpus voorkomen. Noem drie mogelijke redenen waarom de woordenlijst niet volledig is.

2. Welke taalkundige kennis is nodig om Nederlandse woorden (inclusief woorden die niet in een woordenlijst staan) automatisch te kunnen afbreken aan het eind van een woord.

3. Neem aan dat A de automaat is die wordt gedefinieerd door de reguliere expressie R en B de automaat voor S.

   (a) Wat wordt bedoeld met de concatenatie van R en S ($[R,S]$)? Hoe ziet de automaat voor de concatenatie van R en S eruit?

   (b) Wat wordt bedoeld met de disjunctie van R en S ($\{R,S\}$)? Hoe ziet de automaat voor de disjunctie van R en S eruit?

4. Wat is de compositie van twee transducers?

5. Geef twee constructies die erop duiden dat de grammatica van een taal als het Nederlands niet met reguliere expressies of finite state automaten beschreven kan worden.

6. Waarom moet je (om te bewijzen dat *de keeper mist de bal* een zin (`s`) is) een *definite clause grammar* (DCG) aanroepen met

   ```
   s([de,keeper,mist,de,bal],[])
   ```

   en niet met

   ```
   s([de,keeper,mist,de,bal],A)
   ```

   ?

7. Waarom is een DCG krachtiger dan een contextvrije grammatica (CFG)? Geef een voorbeeld van een constructie of fenomeen dat je gemakkelijker met een DCG dan met een CFG kunt beschrijven.

8. Gegeven is de volgende contextvrije grammatica G:

`s`	$\rightarrow$ `np vp`	`det`	$\rightarrow$ `de`
`np`	$\rightarrow$ `det n`	`n`	$\rightarrow$ `aanvallers`
`n`	$\rightarrow$ `a n`	`n`	$\rightarrow$ `goals`
`n`	$\rightarrow$ `n pp`	`n`	$\rightarrow$ `minuut`
`pp`	$\rightarrow$ `p np`	`tv`	$\rightarrow$ `scoren`
`vp`	$\rightarrow$ `iv`	`iv`	$\rightarrow$ `juichen`
`vp`	$\rightarrow$ `tv np`	`p`	$\rightarrow$ `in`
`vp`	$\rightarrow$ `vp pp`	`a`	$\rightarrow$ `laatste`

(a) Geef alle zindelen of *chart items* die een *bottom-up chart parser* op basis van deze grammatica zou produceren voor de zin *De aanvallers scoren de goals in de laatste minuut.*

(b) De grammatica E bevat naast alle regels van G ook de regel: `det` $\rightarrow \epsilon$. Welk effect heeft het toevoegen van zo'n regel op een *chart parser*? (Je mag ervan uitgaan dat de parser geen items toevoegt die identiek zijn aan items die reeds op de chart staan.)

(c) Hoe ziet een grammatica eruit die dezelfde zinnen herkent als grammatica E maar die niet de regel `det` $\rightarrow \epsilon$ bevat?

(d) Beschrijf wat een *shift-reduce conflict* is aan de hand van grammatica G en een voorbeeldzin.