T-61.281 Luonnollisen kielen tilastollinen käsittely

Harjoitus 7, ti 9.3.2004, 8:30-10:00 Sanaluokkien merkitseminen, Versio 1.0

1. a) Tee allaolevien taulukkojen avulla kätkettyihin Markov-ketjuihin perustuva sanaluokanmerkitsijä. Koska tässä tiedetään tarkalleen havainnot ja niitä vastaavat tilat, et tarvitse Baum-Welch –algoritmia, vaan vastaukseksi kelpuutetaan suurimman uskottavuuden estimaatit.

	Toinen sanaluokka					
Eka sanaluokka	AT	BEZ	IN	NN	VB	PISTE
AT	0	0	0	48636	0	19
BEZ	1973	0	426	187	0	38
IN	43322	0	1325	17314	0	185
NN	1067	3720	42470	11773	614	21392
VB	6072	42	4758	1476	129	1522
PISTE	8016	75	4656	1329	954	0

Taulukko 1: *Siirtymien lukumäärä*

	AT	BEZ	IN	NN	VB	PISTE
bear	0	0	0	10	43	0
is	0	10065	0	0	0	0
move	0	0	0	36	133	0
on	0	0	5484	0	0	0
president	0	0	0	382	0	0
progress	0	0	0	108	4	0
the	69016	0	0	0	0	0
.	0	0	0	0	0	48809

Taulukko 2: *Havaintojen lukumäärä*

   b) Laske seuraavien todennäköisyyksien suhde:
      * P(AT NN BEZ IN AT NN | The bear is on the move.)
      * P(AT NN BEZ IN AT VB | The bear is on the move.)
   c) Merkitse sanaluokat seuraavaan lauseeseen Viterbi-algoritmin avulla: "The bear is on the move."

   Tämä harjoitus vastaa suoraan kirjan tehtäviä 10.4-7.

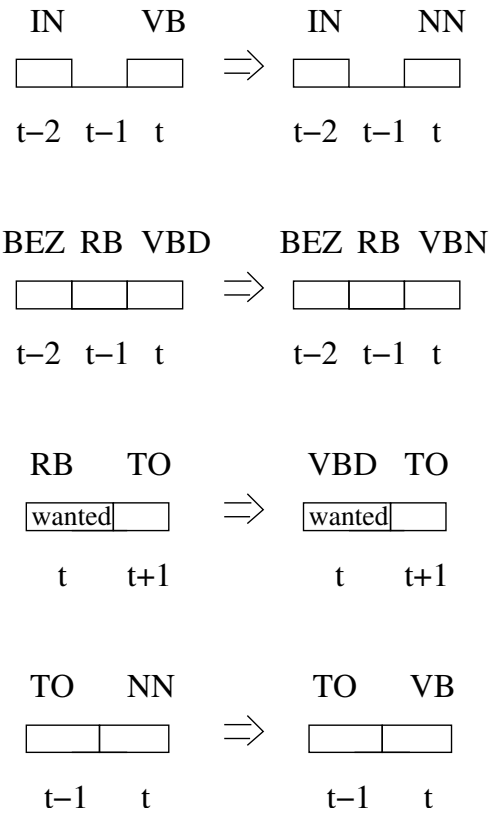
2. Olet töissä sanamuotomuuntajana kielitoimistossa. Pitkän sääntölitanian läpikäytyäsi sait aikaiseksi seuraavanlaiset merkinnät:

I)	PN	RB	TO	NN	IN	AT	VB
	I	wanted	to	look	inside	the	box

II)	PN	BEZ	RB	VBD	IN	AT	VB
	It	was	clearly	marked	on	the	board

III)	AT	NN	PN	BEZ	IN	VB	VB
	The	plane	he	is	on	will	crash

Iloisena lähes valmiista työstä lähdit lakisääteiselle kahvitauolle, mutta nyt sieltä palattuasi sinun pitää vielä soveltaa neljää seuraavaa sääntöä annettuihin lauseisiin:



Kuva 1: *Muunnossäännöt: palkit ja allaolevat indeksit kuvaavat kuinka kaukana sanojen on oltava toisistaan. Päällä olevat sanaluokkamerkinnät kertovat, mitkä sanaluokat pitää olla kyseessä että sääntö laukeaisi. Laatikon sisässä oleva teksti kertoo, mikä sana täytyy olla kyseessä, jotta sääntö laukeaisi. Merkinnät poikkeavat hieman kirjan merkinnöistä.*

Vaikka työtehtäviisi ei kuulukaan luova ajattelu, vilkaiset kuitenkin valmiit lauseet läpi. Kuinka kävi, ovatko tulokset hyviä vai huonoja ?

Harjoituksen tehtävät ovat englanniksi, sillä kirjan kuvaamat menetelmät toimivat järkevimmin suoraan englannin kielellä. Suomen kielessä pitäisi varmaankin pyrkiä ottamaan kielen erityispiirteet, kuten runsas taivutusmuotojen määrä hyötykäyttöön.