Exercise set 3

INF 5830, H2009, 14.10

Vi skal se på glatting av en trigram HMM-tagger. Vi skal regne for hånd, dvs. det er lov å bruke kalkulator, men vi skal ikke skrive et større program som løser problemet. Vi har 10 forskjellige tagger: a, b, c, …, j. Vi skal bare se på P(X | ab), eller mer formelt P(T_i =x | T_{i-1}=b, T_{i-2}=a) der x kan ta verdiene a, b, c, …, j.

Trigramantall									
aba	abb	abc	abd	abe	abf	abg	abh	abi	abj
0	0	10	5	0	0	0	0	0	0

Bigramantall									
ba	bb	bc	bd	be	bf	bg	bh	bi	bj
5	10	15	50	40	0	0	0	0	0

Unigramantall									
a	b	c	d	e	f	g	h	i	j
75	125	90	80	60	200	145	35	20	90

Oppgave A

Lag en linjær interpolering for P(X | ab). Du kan la $\lambda_1 = \lambda_2 = \lambda_3$.

Oppgave B

Både for bigram og trigam har flere av rutene verdien 0. Dette vil vi gjøre noe med. Regn ut uglattete P(X | ab). Glatt dem deretter. I denne oppgaven kan vi bruke Laplace-glatting: ”legg til en”.

Oppgave C

Gjør det samme for bigram.

Oppgave D

Bruk disse glattede sannsynlighetene fra punkt B og C og prøv å beregne ”Katz backoff” for P(X|ab). Se J&M seksj 4.7. Mens de andre punktene er rett frem, er dette en litt større utfordring.

Oppgave E

Sammenlikn resultatet fra oppgave D med resultatet fra oppgave A.

Vi vil se på løsning av innleveringsoppgaven.

Jeg anbefaler stadig alle til å gjøre de oppgavene fra Nivres webkurs som jeg har foreslått tidligere. Jeg svarer på spørsmål fra disse oppgavene og kan gjennom dem hvis noen har prøvd og står fast.