Oppgavesett 3, INF5820, H2008

Innleveringsfrist 22.10

Vi skal i dette settet arbeide med statistikkbasert maskinoversettelse (SMT). Vi vil bruke ferdig programvare. Det finnes flere fritt tilgjengelige standardprogrammer, som bl.a. kan finnes fra her: www.statm.org. Mange av programmene har rare navn fra Egypt. Det mest avanserte programmet heter Moses. Vi skal velge en tidligere og enklere versjon basert på Giza. Det er mange ting som skal virke sammen. Vi får hjelp hos kolleger i Danmark som har pakket det hele i en pakke kalt SMTlab, og som kan finnes her: http://www.id.cbs.dk/ dh/ngslt/fall06/exercise.html.

1 Oppgave

Last ned SMTlab. Installer det i følge manualen, og test det på et av de medfølgende korpusene. Innlevering: resultater av kjøringen.

2 Oppgave

Vi skal nå arbeide med vår egen tekst jhpstg.txt. For å unngå problemer med tegnsett fra sist har jeg lagt den i flere versjoner (se tabell 1). Filene ligger på mitt område.

Dette er samme norske tekst som på oppgavesett 2, men i tillegg har vi 2–3 oversettelser av teksten til engelsk. For å få et størst mulig materiale vil vi bruke alle tre oversettelsene. Vi kan gjøre det ved å legge de etter hverandre, noe slikt som "(norsk $\sim$ oversettelse1)+(norsk $\sim$ oversettelse2)+(del av norsk $\sim$ oversettelse3)". Gjør dette og kalle det norsk_engelsk.

navn	tegnsett	linjeslutt	
jhpstg.txt	utf-8	lf	(unix)
jhpstg_latin.txt	latin-1 (=iso-8859)	lf	(unix)
jhpstg_unicode_dos.txt	utf-8	cr+lf	(dos, windows)
jhpstg_latin_dos.txt	latin-1 (=iso-8859)	cr+lf	(dos, windows)

Table 1: Ulike versjoner av jhpstg.txt

3 Oppgave

Klargjør nå norsk_engelsk for bruk med SMTLab. Husk å skille ut et testkorpus. Innlevering: test- +treningskorpus.

4 Oppgave

Lag et oversettelsessystem fra norsk til engelsk.

5 Oppgave

Test systemet ditt på testkorpuset.

6 Oppgave

Se på resultatet. Hva gjør systemet bra? Hvor har det problemer? List opp minst 5 typer feil systemet gjør. Innlevering: Kjøringssresultat fra testkorpuset. Skriftlige svar på spørsmålene.

7 Oppgave

Se nå på demonstratorsiden for Logon (adresse får dere i e-post), og sammenlikn systemet ditt både med Logon-systemet og de andre systemene som kommer opp nederst under hver oversettelse. Hvordan kommer du ut i forhold til de andre? Hva gjør de bedre?

Sammenlikn både på eksempler fra testkorpuset ditt og på eksempelsetninger du konstruerer selv. Vær obs på at noen av de andre systemene, f.eks. Eriks, kan ha brukt deler av testkorpuset ditt som treningskorpus.

Innlevering: Skriftlig besvarelse.