Semesteroppgave og oblig.4, INF5820, H2008

Innleveringsfrist oblig 4, 12.11 kl 0900

Innleveringsfrist semesteroppgave, 15.12 kl 1200

Vi har nå kommet til det litt større prosjektet. Det vil både utgjøre semesteroppgaven som er grunnlag for karakteren — "Eksamen består av en semsteroppgave" står det på emnets hjemmeside — og det vil utgjøre innleveringsoppgave 4. I dette prosjektet skal du selv få velge hva du vil arbeide med. Du skal gjennomføre et språkteknologisk prosjekt relatert til emnets tema.

Overordnete krav til prosjektet

Den endelige innleveringen skal bestå av en prosjektrapport på rundt 10 sider, fortrinnsvis skrevet i LaTeX. Den skal ha form som en artikkel der du forteller om problemstilling, hva du har gjort, valg du har tatt underveis, resultater, referanser, og hva som ellers hører med. Denne skal leveres i papirform. Du skal også levere elektonisk relevante filer, som f.eks. eksperimentresultater og kode du har skrevet. Innleveringen vil bli bedømt på grunnlag av kreativitet og avgrensning av problemstilling; metoder som er valgt, designspørsmål og avgrensninger; gjennomføring og presisjon i gjennomføringen inkludert programmering som inngår; analyse av resultatene; og kvaliteten på selve rapporten og skrivingen. Om systemet du lager ikke gir så gode resultater som du ville tro, behøver ikke bety at det er et dårlig prosjekt om det ellers er godt gjennomført.

Prosjektplan

Som første del av prosjektet skal du lage en plan for prosjektet. Det er alltid nyttig å starte et prosjekt med en plan, så også her. Og det vil være et utgangspunkt for veiledning og for å motta råd: Er prosjektet gjennomførbart? Kan noe gjøres annerledes? Prosjektplanen vil tjene som innleveringsoppgave (oblig.) 4. Det vil si at du får ingen karakter på den og den vil ikke telle i karakteren til sluttprosjektet. Men det er obligatorisk å levere den inn. Hva bør planen inneholde?

- En foreløpig beskrivelse av problemet du vil arbeide med.
- En beskrivelse av hvordan du tenker deg å gå frem for å angripe problemet. Hva planlegger du å gjøre?

- En arbeidsplan. Hvor lang tid tar hver del? I hvilken rekkefølge vil du gjøre ting? (Husk å sette av tid til skrivingen! Husk at litt større eksperimenter kan ta tid å kjøre!)

Prosjektplanen bør være 1–2 sider.

Mulige prosjekt

Hva er et passende prosjekt? Det er opp til deg, men det er naturlig å ta utgangspunkt i det vi har sett på i forelesningene, og i de tidligere innleveringsoppgavene. Noen mulige retninger å gå i:

Ulike metoder for WSD Utgangspunkt i innleveringsoppgave 1. Men se på effekten av ulike metoder og parametre, f.eks. hvordan vil en kollokasjonsvektor gjøre det sammenliknet med en "bag of words"-vektor? Vektorens størrelse? Korpusets størrelse? Prøve med andre, større korpus enn det vi brukte. osv.

Mer realistisk WSD Vi brukte pseudo-word fordi vi ikke hadde "sense-tagged" korpus for norsk. Det kunne vært morsomt å gjøre ordentlig WSD. For å finne mer interessante ressurser for andre språk går det an å nøste herfra: http://ji.ehu.es/eneko/resources.html eller her:
http://www.d.umn.edu/~tpederse/ .

Ordlikhet Vi implementerte aldri ordlikhet (Jurafsky og Martin sek. 20.7). Det kunne vært morsomt å se hvordan det ble.

Sammendrag Lag et lite system for å lage sammendrag (J&M 23.3).

SMT i større stil SMT-ssytemet vi lagde i oblig. 3 var langt fra perfekt. En årsak var at treningsmaterialet var for lite. Et mulig treningsmateriale er Europarl-korpuset http://www.statmt.org/europarl/ . Det kunne også være interessant å se hvem som er best av Pharaoh og Moses: prøve med begge og evaluere forskjellene. Dessuten er det kanskje noen parametre det kan skrus på? Er det mulig å gjøre forbedringer? (Husk å legge inn tid til kjøringene.)

Noe helt annet som du virkelig har lyst til...