Semesteroppgave og oblig.4, INF5820, H2010

Innleveringsfrist oblig 4, 15.11

Innleveringsfrist semesteroppgave, 15.12

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 elektronisk og 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

- utforming av problemstilling og avgrensning
- valg av metoder for å løse problemet, eksperimentell design
- 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 det 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:

WSD

Gå videre fra oblig. 2 for WSD. Eksempler på mulige problemstillinger:

- I selve oppgaven sammenliknet vi effekten av å bruke 1- og 2-ords kontekster. Dette er et eksempel på en effekt en kan undersøke. (Men nå er den brukt opp.)
- En av dere utvidet prosjektet ved å se på størrelsen av trekkvektoren. Det ville også kunnet gi opphav til et prosjekt, hadde det ikke vært brukt.
- Vi brukte "bag of words" trekk. Hvordan vil en kollokasjonsvektor gjøre det til sammenlikning?
- Vi brukte "naïve Bayes", hvordan ville andre læringsalgoritmer gjort det, f.eks. eksempelbasert læring med "nearest neighbors"-metoden.
- Vi brukte en klassifikatorer som tildelte en av 6 klasser. Hva skjer hvis vi i stedet bruker 6 ulike klassifikatorer, en for hver klasse?
- Tilgangen på "sense-tagged" korpora er begrenset. Eksperimentér med "bootstrapping".
- osv.

Annen ordsemantikk

Litt avhengig av om du er interessert i dette eller har sett på noe som er relevant:

- Ordlikhet fra kontekst (J& M sec. 20.7)
- Ordlikhet fra thesaurus (J& M sec. 20.6)
- etc.

SMT

SMT-ssytemet vi lagde i oblig. 3 var langt fra perfekt. En årsak var at treningsmaterialet var for lite. Her er flere mulige veier å gå:

- Andre parallellkorpora som Europarl-korpuset http://www.statmt.org/europarl/ eller andre.
- Finnes det større treningskorpus for norsk, evt går det an å lage det?
- Hvilken effekt har max. fraselengde for oversettelsen?
- Hvordan får en parallellstilt oversatte tekster på setningsnivå?
- Vi brukte Pharaoh, hva med å skifte til Moses?
- I Moses-oppsettet er det muligheter for å gjøre optimaliseringer underveis, som å lære hvor mye vekting som skal til språkmodellen og hvor mye til oversettelsesmodellen.
- osv. Bare fantasien - og tiden - setter grenser (Husk å legge inn tid til kjøringene.)

Noe helt annet

som du virkelig har lyst til...

Gode råd

Listen over er en meny. Det er ikke meningen å gjøre alt! Dette er bare et lite prosjekt/eksperiment. Dere skal bruke omtrent 2 arbeidsuker. Da sier det seg selv at vi ikke forventer at dere skal finne opp noe revolusjonerende nytt, eller presetere bedre enn det som finnes i litteraturen. Nøkkelen til et godt resultat ligger i en god avgrensning.

Lykke til !
Jan Tore