Fredrik Olsson, språkvetenskaplig databehandling – disputation 19 december 2008

Ny metod för att underlätta extraktion av information ur stora textmassor

I dagens stora flöde av databaserade texter är det viktigt att få fram system som underlättar sökandet av viss efterfrågad information. Är det möjligt att hitta information om t.ex. händelser i ett företag ur nyhetstexter; vem som lämnar vilken post, varför så skedde, till vilket företag och position som personen går etc? Fredrik Olsson tar i sin avhandling upp en ny metod för att underlätta uppmärkningen av förekomster av namn i databaserade textdokument.

Informationsextraktion handlar om att analysera text i syfte att identifiera och plocka ut information om fördefinierade typer av entiteter, händelser i vilka entiteterna är engagerade samt relationer mellan entiteter och händelser. Det handlar alltså om att få tillgång till strukturerad information från en till synes ostrukturerad informationskälla.

En av anledningarna till att informationsextraktion inte är tillgänglig för alla är att det krävs mycket arbete och tid för att anpassa ett system till att fungera för nya uppgifter i en ny textdomän. Ett system som hanterar ovanstående exempelscenario skulle förmodligen inte fungera alls om uppgiften ändrades till att identifiera interaktioner mellan proteiner beskrivna i biomedicinsk text.

Ett vedertaget sätt att angripa problemet med domänanpassning av system för informationsextraktion är att realisera dess komponenter med hjälp av maskininlärning, dvs. lärande datorprogram. Maskininlärning bygger i mångt och mycket på att det finns exempel att lära sig av. En komponent i ett extraktionssystem behöver se exempel på de fenomen det ska lära sig att identifiera, t.ex. entiteter och relationerna mellan dessa. Grunden till den här typen av maskininlärning är alltså
tillgången till stora mängder exempel. Dock finns det stora utmaningar i att ta fram bra exempel: det är mödosamt, tar tid och kräver en människa som känner domänen väl för att märka upp exempel i texter.

Att känna igen namn på t.ex. personer, företag och platser är grundläggande för informationsextraktion. Genom att känna igen namn kan vi också börja leta efter t.ex. relationerna, uttryckta i texten, mellan bärarna av namnen.

Fredrik Olsson beskriver i sin avhandling arbetet med att utveckla och utvärdera en metod, kallad BootMark, för att märka upp förekomster av namn i textdokument.
BootMark bidrar till att reducera den mängd dokument en mänsklig annoterare behöver märka upp för att träna en namnigenkännare med prestanda som är lika bra eller bättre än en namnigenkännare som är tränad på ett slumpmässigt urval
av dokument från samma korpus.

Avhandlingens titel: **Bootstrapping Named Entity Annotation by Means of Active Machine Learning. A method for creating corpora.**
Disputationen äger rum fredagen den 19 december 2008 kl. 13.15
Plats: Lilla hörsalen, Humanisten, Renströmsgatan 6
Opponent: Dr Miles Osborne, Edinburgh

Avhandlingen kan beställas från Institutionen för svenska språket, erik.falk@svenska.gu.se
För ytterligare information kontakta Fredrik Olsson, mobiltel. 0704 -15 54 10,
e-post: fredriko@sics.se
Hemsida: www.sics.se/people/fredriko
Avhandlingen finns även tillgänglig digitalt:
http://hum.gu.se/institutioner/svenska-spraket/publ/datal/