En automatisk navnegjenkjenner for norsk, svensk og dansk

Janne Bondi Johannessen, Universitetet i Oslo
jannebj@mail.hf.uio.no

1. Generelt

I dette foredraget vil vi presentere et prosjekt som nettopp har kommet i gang, og som har fått NOK 1 000 000 fra NorFA til et nordisk nettverk. Prosjektet har to siktemål som henger nøye sammen. Det ene er å bygge nettverk mellom fire forsknings- og utviklingsmiljøer ved UiO (Tekstlaboratoriet), UiB (HIT-senteret), GU (Språkdata) og CST. Det andre er praktisk: å utvikle en automatisk navnegjenkjenner for norsk, svensk og dansk. (En automatisk navnegjenkjenner er et program som klarer å skille typer navn, som firmanavn, steds- og personnavn, fra hverandre). Produktet som utvikles, kan ha kommersiell interesse. Vi er samarbeider med Internett-firmaet FAST Search & Transfer, som vurderer å gå inn med studentstipendmidler.

For at det skal være mulig i et søke- eller informasjons-behandlingssystem å søke etter et navn med spesifisering av hva slags navnetype det er, må det utvikles særspråklige navnegjenkjennere. De tre deltagerspråkene har ulike navnetradisjoner, og samme navn kan gjerne være personnavn i ett land og steds- eller firmanavn i et annet. Videre vil navnenes språklige kontekst være ulik i de ulike språkene, slik at det må utvikles en separat navnegjenkjenner for hvert språk.

Vi ønsker altså å utvikle en automatisk navneentitetsgjenkjenner for norsk (bokmål og nynorsk), svensk og dansk. Mens en vanlig grammatisk tagger vanligvis vil kjenne igjen et ord som egennavn, er det en helt annen sak å kjenne igjen hva slags type egennavn det er snakk om i de enkelte tilfellene. Navnegjenkjenneren skal kunne ta en hvilken som helst ukjent tekst og bestemme for hvert egennavn hva slags navneentitet det dreier seg om: om det er et personnavn, et stedsnavn eller et firmanavn. Man skulle kanskje tro at en slik navnegjenkjenning ville være fort gjort om man bare hadde noen store navnelister, men i praksis er det ikke så lett. I Norge er det for eksempel vanlig at stedsnavn har gitt opphav til gårdsnavn, og så til etternavn: *Bondi* er både et stedsnavn og et personnavn. Og overalt i verden er det vanlig at personnavn brukes som firmanavn: Lefdal er både et firmanavn og et personnavn. Dessuten vil mange tekster, ikke minst fra aviser, ikke bare inneholde nasjonale navn, men navn fra hele verden.

De enkelte miljøene ved UiO, GU og CST kommer til å benytte ulike metoder for utvikling av navnegjenkjenneren, ikke bare fordi vi dermed kan sammenligne metodene og til slutt velge den eller de vi mener er best, men også fordi vi til dels har ulik programvare i utgangspunktet og ulik kompetanse i miljøene. I foredraget vil vi snakke om noen av metodene som er mulige å bruke, og hvilke vi vurderer eller har bestemt oss

for. Dette inkluderer f.eks. navnelister, statistiske tilnærminger, og kontekstuelle regler, eller en blanding av disse, slik det er beskrevet i Mikheev et al. (1998, 1999).

2. Hva finnes i dag for de enkelte språkene?

Det finnes i dag ikke noen skikkelig navnetypegjenkjenner for de tre skandinaviske språkene. For dansk og norsk finnes det ingen, mens det for svensk finnes en begrenset utgave (se nedenfor). Derimot finnes det et minimum av språklige verktøy som er nødvendige for å gjennomføre prosjektet.

Dansk

For dansk finnes det en Brill-tagger. Den er basert på transformasjonsregler snarere enn rent statistiske metoder. Den gjenkjenner ca. 96, 5 % av taggene korrekt, og plukker også ut egennavn (men over 10 % av feilene var feil ved gjenkjenning av egennavn) (Pedersen 2001). Det er denne taggeren som kommer til å bli brukt i prosjektet som et utgangspunkt.

Videre har det blitt utviklet en navnegjenkjenner for dansk til bruk for tekstresymeringssystemer, som spesielt legger vekt på å gjenfinne egennavn som sådanne, og i tillegg finne nominale syntagmer som koreferer med disse i tekstene (Nelson 2000).

Svensk

Innenfor EU-prosjektet AVENTINUS ble det utviklet en begrenset navnetypegjenkjenner av Dimitris Kokkinakis. Denne bruker ikke navnelister, og ikke statistikk, men lingvistisk kontekst. Mye av det som ble gjort i det prosjektet vil kunne brukes videre i det herværende prosjektet (se Kokkinakis et al 2001).

Det finnes flere grammatiske taggere som er aktuelle å bruke i prosjektet.

Norsk

Det finnes ingen navnetypegjenkjenner for norsk. Ved Universitetet i Oslo (Tekstlaboratoriet) finnes det en tagger som gjenkjenner egennavn (se Johannessen et al 2000). Leksikalsk funnrate (recall) er på 99,2%, mens presisjonen er på 96,8%. Denne vil bli brukt for den norske navnetypegjenkjenneren.

Referanser

Johannessen, J.B., K. Hagen og A. Anders Nøklestad. 2000. A Constraint-Based Tagger for Norwegian. I Lindberg, Carl-Erik og Steffen Nordahl Lund (red.): *17th Scandinavian Conference of Linguistics. Odense Working Papers in Language and Communication 19,* 31-48, University of Southern Denmark, Odense.

Kokkinakis, D., M. Gellerstam, Y. Cederholm, T. Rasmark. Språkdata Presentation/Discussion Paper. Foredrag presentert på første navnegjenkjennerseminar på Fefor, januar.

Mikheev, A., C. Grover og M. Moens. 1998. Description of the LTG system used for MUC-7. I *Seventh Message Understanding Conference (MUC-7): Proceedings of a conference held in Fairfax, Virginia*. http://www.muc.saic.com/proceedings/muc_7_toc.html

Mikheev, A., M. Moens og C. Grover. 1999. Named Entity Recognition without gazetteers. I *Proceedings of EACL 99, Ninth Conference of the European Chapter of the Association for Computational Linguistics*, s. 1-8.

Nelson, M. Propriumsyntagmer i tekstresumeringssystemer. Ph.d.-avhandling, Handelshøjskolen i København.

Pedersen, B. 2001. Danish Proper Nouns in the Brill Tagger. Foredrag presentert på første navnegjenkjennerseminar på Fefor, januar.