Att automatiskt känna igen och kategorisera namn i svenska texter

Dimitrios Kokkinakis, forskare, svenska språket

> INTERVJU MED DÖDEN. För att bearbeta smärtan har fadern fått sonen att ge sin cancer ett namn, som är Alban. Sedan intervjuar han tumören Alban om varför han finns till. [...]

Inledning

Ämnet för den här uppsatsen är namnigenkänning. Namnigenkänning är en specialiserad stödteknologi inom datalingvistik som går ut på att automatiskt känna igen och förse namn och namnliknande uttryck i löpande text med etiketter, som t.ex. person eller plats. Det system som beskrivs här nedan har utvecklats inom ett skandinaviskt samarbetsprojekt som heter ”Nomen Nescio” (NN). Inom NN-projektet har man utvecklat en kunskapsplattform innehållande verktyg för namnigenkänning. Verktygen består av språkliga resurser, t.ex. namnlistor och algoritmiska metoder. Algoritmiska metoder är datorprogram, som kan hantera namn i svenska, danska och norska. Mer information om nätverket finns på följande webbadress: http://g3.spraakdata.gu.se/nn/. Denna uppsats handlar om det svenska systemet. De citat som återges i uppsatsen är hämtade ur olika svenska dagstidningar.

> Duderhof, Gubanits, Ingris, Järvisaaari, Kallivieri, Kattila, Keltos, Kolpana, Koporje, Koprina, Kosemkina, Liissilä, Moloskovits, Narva, Pavlovsk, Ropsja, Serebetta, Skvorits, Slavänka, Soikina, Spanko, Tyris, Valkeasaari – alla dessa protestantiska församlingar med rötter i svenskt 1600-tal [...]

Olika slags namn

Namn eller namnentiteter (NE) kan bestå av *enkla egennamn*: ”Svensson”, *sammansatta egennamn*: ”Anna-Karin”, *ordgrupp*: ”New York Rangers”, initialord av begynnelsebokstäverna i de ord som ingår i ordgruppen: ”TT”, *kombination av egennamn och siffror*: ”Boeing 747”, *namn bestående av icke-egennamn*: ”Tre Kronor” i kontexten ”Tre Kronor besegrade Finland i VM”, *utländska uttryck*: filmen ”A Friend of the Deceased”. Vidare kan ett namn bestå av versaler och siffror, ”JAS-39”, av versaler och gemener ”GlaxoSmithKline”, endast av siffror som i mobiltelefoniföretaget ”3”, siffror och gemener som i kapitalbolaget ”3i”, gemener, versaler och siffror, ”G-77-mötet” eller andra konventioner som t.ex. ”’Star Trek’-serien”. Förutom alla stavnings- och kombinationsmöjligheter försvåras namnigenkänning av det faktum att det finns många tvetydigheter som kan uppstå när man på automatisk väg med hjälp av dator försöker bestämma om ett ord överhuvudtaget är ett namn och sedan om det refererar till en person, en plats, en organisation, ett objekt av något slag eller en händelse.

Beroende på kontexten kan till exempel ”Hans” eller ”Inga” vara ett egennamn eller ett pronomen; medan egennamnet ”Victoria” kan syfta till en uppsjö av olika slags namnkategorier, bl.a.:

- personnamn (t.ex. kronprinsessan Victoria)
- idrottsplats (t.ex. OS-arenan ”Victoria”)
- teaternamn (t.ex. film på Victoria)
- namn på ett kafé (t.ex. Café Victoria)
- teaterpjäs (t.ex. sextimmarsdramat Victoria)
- symaskin (t.ex. symaskinen Singer Victoria)

> Nu vill jag rekommendera The Ballistic Brothers ”Perfect Day” (cd-spår, ”Give ‘Em Enough Dope 3”), Universal Jones ”Phoenix Rising” (tolva), Billy Bragg ”William Bloke” (cd), Death in Vegas ”Rocco” (kommande tolva), Victor Lewis-Smith ”Inside The Magic Rectangle” (bok), David Holmes ”My Mate Paul” (kommande tolva), Buccaneer ”Skettel Concerto” (cd-spår, ”Greensleeves Sampler 14”), The Chemical Brothers featuring Noel Gallagher ”Setting

Sun” (kommande singel), Joe Queenan ”The Unkindest Cut” (bok) och Nik Cohn ”Need” (bok).

Namn mot resten av ordförrådet

Att kunna känna igen namn och vidare kategorisera olika namnvarianter är en svår men viktig stödteknologi i många datalingvistiska tillämpningar, som t.ex. informationsutvinning, maskinöversättning och textsammanfattning. Ett- och flerordiga egennamn samt flerordsuttryck som fungerar som namn skiljer sig från andra ord och fraser på olika sätt. Först och främst är det ett faktum att egennamn inte brukar finnas med i vanliga ordböcker. Namnens stavningsmöjligheter kan variera en hel del även inom samma text, namnet ”John Fitzgerald Kennedy” kan påträffas i texter med variantformerna: ”J.F.K.”, ”JFK”, ”J.F. Kennedy”, ”Kennedy”, ”Kennedy, J.F.” eller ”Kennedy, John Fitzgerald”. Namn sträcker sig längre än vad traditionell grammatik kallar för egennamn eftersom namn kan ha en komplex inre struktur, speciellt i de fall där vanliga substantiv, prepositioner och adjektiv är en del av namnet, t.ex. medieföretaget ”Film i Väst” och fastighetsföretaget ”Akademiska Hus Syd”.

Även ordböcker som försöker täcka egennamn blir snabbt otillräckliga och inaktuella på grund av att nya namn ständigt skapas eller lånas in i språket. Detta betyder däremot inte att sådana speciella namnordböcker är onödiga i igenkänningsarbetet, utan ett viktigt komplement i sammanhanget, t.ex. för ortsnamn (se metodavsnitt nedan).

Namn kan ha en viss betydelse, men de behöver inte ha det, vilket annars är en viktig förutsättning för alla andra ord som tillhör andra ordklasser som t.ex. verb och substantiv. Ett typiskt egennamn anses endast vara en etikett på en unik referent. Faktum är att på det semantiska planet kan namnens betydelse i stor utsträckning bero på talarens personliga erfarenheter och talsituationen, vilket innebär att ett namn inte behöver ha en konstant betydelse mellan alla språkanvändare. Kryptiska beteckningar kan vara exempel på sådana namn som t.ex. [supernova] ”SN 1987 A”, [viruset] ”H5N1” eller [flygplanet] ”F117A”.

Västindiska barer som torde vara värda en omväg: Soggy Dollar Bar (Painkillerns ursprungsort) och Foxy’s, Jost Van Dyke, båda Brittis-

ka jungfruöarna. Dune Preserve, Anguilla. Kontiki och Karibuni, St Martin. Bomba's Shack, Tortola, Brittiska jungfruöarna. Admiral's Inn och O J's, Antigua. Les Tortues och Nilce's Bar, Guadeloupe. Firefly, Mustique. Gong on the Beach, St Kitts.

Metod för namnigenkänning

Hur man identifierar ett egennamn är inte självklart. Språkspecifika detaljerade specifikationer i uppgiftsbeskrivningen har tagits fram i en serie av s.k. MUC-konferenser (Message Understanding Conferences), men dessa har dock varit anpassade till de specifika teman som dessa konferenser handlade om, t.ex. raketuppskjutningar. Enligt de specifikationer som har tagits fram, ingår det i namnigenkänningsuppgiften att identifiera och korrekt kategorisera följande: *personnamn* och *organisationsnamn* samt olika typer av *plats- och ortnamn*, men också information av icke-namnkaraktär, t.ex. *datum-, tids- och andra numeriska uttryck*. Däremot behandlades inte namn på produkter, händelser, kometer, planeter, filmer, fartyg, böcker, husdjur och mycket mer som kan påträffas i olika typer av texter på MUC-konferenserna. Inom NN-projektet har man vidare utvecklat dessa specifikationer så att de kan omfatta ytterligare namnkategorier (se nästa avsnitt).

En bra utgångspunkt för igenkänningsarbete är att använda en kombination av namnlistor, kontextuell information samt skrivkonventioner (t.ex. inleds många egennamn i svenska med en versal, vilket är en viktig ledtråd i igenkänningen). Alla dessa källor kan dock vara problematiska. Versaler är till exempel inte alltid en säker ledtråd, och det kan vara svårt att avgöra hur långt namnet sträcker sig. I kontexten ”Ett fotografi med titeln Galna turister visar en gatumarknad i Brasilien” är ”Galna turister visar en gatumarknad i Brasilien” ett namn där versaler inte spelar en avgörande roll för klassificeringen av det aktuella namnet. Eller är det bara ”Galna turister” och ”Brasilien” som är namnen i sammanhanget? Ibland krävs det mycket mer kontext för att kunna avgöra namngränserna. Namnlistor kan innehålla egennamn som, beroende av kontexten, i vissa fall kan vara allt annat än egennamn, t.ex. ”Inga”, ”Fred”, ”Hans”, ”Axel”, ”Visa” och ”Klara”. Följaktligen kan sådana ord skapa mer problem än

lösningar om man tar för givet att eftersom de börjar med versal kan dessa ord endast förekomma som egennamn.

> Barnfilm och britter årets ljuspunkter; Big Daddy, Deep Blue Sea, Never Been Kissed, The Ninth Gate, Holy Smoke, Runaway Bride, Austin Powers – The Spy Who Shagged Me, Fight Club, Best Laid Plans, Blair Witch Project, Unlucky Monkey och Drive Me Crazy. [SvD2000]

Namnkategorier och systemets struktur

Det svenska NE-systemet kan känna igen namn som kan tillhöra sex olika huvudtyper. Det finns ytterligare två grupper, en grupp som behandlar numeriska uttryck och en som identifierar tidsuttryck. Huvudtyperna är: *lokaliteter, personer, organisationer, händelser, produkter* och *andliga produkter.* Inom varje huvudkategori finns det ett antal förfinade underkategorier. Inom organisationskategorin finns t.ex. idrottsrelaterade och politiska organisationer. Inom händelsekategorin finns det namn på religiösa och kulturella händelser medan det inom produktkategorin finns en grupp för bil-, båt- och flygmärken och en grupp för läkemedel.
Det svenska systemet består av fyra delar. Dessa är:

1) **en uppsättning grammatikregler för varje namnkategori**, där reglernas huvudkomponent är ett antal nyckelord, t.ex. titelled för personer (”chefen”, ”ministern”), ledtrådsord för organisationer (”AB”, ”Banken”, ”Ltd”), typiska verb för varje namnkategori; t.ex. ett personnamn följer ofta eller föregår ofta vissa verb: ”säger”, ”erkände” eller ”medverkade”, medan typiska avledningssuffix för lokaliteter är ”xxxbukten”, ”xxxstaden”, ”xxxborg”, ”xxxland”. Alla dessa ledtrådar brukar finnas i närheten av egennamn eller vara en del av namnet och därmed ge stöd i igenkänningsprocessen;

2) **listor på flerordsnamn**, t.ex. ”Amerikanska Jungfruöarna” och ”Palma de Mallorca”, mestadels extraherade från dokument på Internetplatser;

3) **namnlistor av enkla namn** t.ex. ”Kalle”, ”Sverige” och ”Volvo”;

4) **ett datorprogram som utnyttjar den information som har identifierats med hjälp av de föregående stegen** för att kunna markera namn i en text där kontexten är osäker eller inte innehåller säkra ledtrådar. Till exempel, om man antar att man har textfragmenten: *LGP Telecom redovisar en vinst […] LGP bedömer att dess nisch [...]* och att systemet med hjälp av flerordsnamnskomponenten har identifierat ”LGP Telecom” som en organisation och sedan påträffas ”LGP” inom samma text, där ”LGP” är ett ord som varken finns i namnlistorna eller har identifierats av grammatikreglerna, då bedömer detta program att ”LGP” med stor sannolikhet också är en benämning på en organisation, troligtvis samma som ”LGP Telecom” inom samma text. Följaktligen läggs en lämplig markering till i texten: [*LGP Telecom=organisation] redovisar en vinst […] [LGP=organisation] bedömer att dess nisch [...]*.

> Unika för Sverige på den nationella listan är annars Annika, Appell, Bellona, British Queen, Early Puritan, Eloge, Evergood, Frieslander, Jätte-Bintje, King Edward, Maria, Minerva, Ofelia, Ovatio, Premiere, Prevalent, Producent och Provia. Vilken som är bäst är en smaksak, men flera på listan är mycket svåra att få tag på - odlingen är blygsam.

Testgränssnitt

Det svenska systemets funktionalitet kan testa på nätet via NN:s hemsida (http://g3.spraakdata.gu.se/nn/). Man kan välja mellan att skriva in och testa egna texter, att välja från en meny med lagrade texter, eller helt enkelt hämta tidningstexter från svenska nyhetssidor. Genom att sedan klicka på en knapp som heter *Submit Selections* bearbetas indata och efter 5–7 sekunder får man resultatet i ett nytt föster. Varje namn som har identifierats av systemet, och beroende på vilken namnkategori namnet tillhör är det nu markerat med olika färger. Det finns ytterligare valmöjligheter. Man kan t.ex. utvärdera huruvida systemet har eller inte har lyckats att identifiera och förse olika namn med de rätta etiketterna i en text. Beroende på vad användaren av systemet markerar som rätt, delvis rätt eller fel, och antal namn som systemet inte lyckades identifiera, får man en procentsiffra mellan 0 och 100 %, där 100 % står för ett perfekt resulat.

Sammanfattning

Denna uppsats har gett en översiktlig beskrivning av ett system för namnigenkänning för svenska. Systemet har utvecklats inom ”Nomen Nescio”-projektet. Projektets syfte var dels att knyta samman olika forskningsmiljöer i tre nordiska länder, Sverige, Norge och Danmark, dels att utveckla en resursbas av kunskap, standarder och konkreta verktyg för namnigenkänning i och för de nordiska språken.

Namnigenkänning är en icke-trivial verksamhet som innefattar olika delmoment, allt från att kunna bestämma att ett eller flera ord verkligen är ett namn eller en namngrupp till att kunna avgöra vilken sorts namnkategori det/de tillhör. Resultatet av vårt arbete kan testas via Internet där användaren har möjlighet att själv avgöra hur bra eller dåligt systemet är. Systemet kommer att finnas tillgängligt för utprövning och stöd för forskning under de närmaste åren.