Sprogteknologi skal gøre fremtidens søgemaskiner intelligente

Hvad hjælper det, at den nødvendige viden er tilgængelig på internettet, hvis vi ikke kan trække den ned på vores pc-skærme, når vi har brug for den? Sprogteknologien skal revolutionere fremtidens søgemaskiner ved at lære dem sprog, så de kan forstå, hvad vi søger, og besvare vores spørgsmål. Ikke kun med henvisninger til X dokumenter, men også med fuldgyldige svar formet i naturlige sætninger.

Af journalist Maria Bernbom

Derfor er nutidens søgemaskiner uintelligente

”21 dokumenter matcher din søgning…” Fantastisk. To-tre klik til, og du har fundet den oplysning, du søger. For fem år siden, ville oplysningen have kostet dig mindst tre telefonbogsopslag, to telefonopringninger og en halv time af din kostbare tid. Anno 2002 når du dit mål på 48 sekunder. Hvis du er heldig.

For i takt med at indholdet på nettet vokser, bliver vi i vores jagt på umiddelbart simple svar ofte bombarderet med så mange henvisninger til dokumenter, at overblikket forsvinder, og frustrationerne begynder at melde sig.

For det, vi leder efter på nettet, er jo sjældent henvisninger til dokumenter. Det er svar på specifikke og relativt simple spørgsmål som

- ”Hvordan kommer vi af med husstandens storskrald?”
- ”Hvornår har filmen *De to tårne* premiere i Danmark?”
- ”Min virksomhed skal have installeret kildevandsbeholdere. Hvem er den billigste udbyder?”

Men selvom den viden, det kræver at besvare disse spørgsmål, er tilgængelig på nettet, er end ikke de bedste søgemaskiner ”intelligente” nok til at finde dem frem, bearbejde dem og præsentere dem for os i den hensigtsmæssige form: ”De to tårne har premiere i Danmark onsdag den 18. december 2002”. For maskinerne forstår ikke, hvad det *egentlig* er, vi spørger om.

Og det er her, sprogteknologien kommer ind i billedet.

Nettet skal lære sprog

Sprogteknologien skal få nettet til at forstå os – på vores naturlige sprog. Den skal lære søgemaskinerne, at vi med et spørgsmål som ”Hvilken bilforhandler i Nordjylland sælger

Jaguarer?" er interesserede i at vide noget om, hvor i et bestemt geografisk område, man kan købe en bil af mærket Jaguar. Men at vi hverken er interesserede i forhandlere i Storkøbenhavn eller i dyrerigets jaguarer.

Fremtidens søgemaskiner skal med andre ord kunne sprog. De skal forstå betydningsmulighederne i ord og sætninger. Og derfor forskes der allerede ivrigt i udviklingen af mere intelligente søgemaskiner på universiteter og i erhvervslivet verden over. I Danmark for eksempel med projekterne:

- MOSES – På vej mod et klogere net **[Internt LINK]**
- OntoQuery – Fordi thiamin også er et vitamin **[Internt LINK]**
- Ankiro – Når erhvervslivet træner søgemaskiner og chatrobotter **[nternt LINK]**

MOSES – På vej mod et klogere net

- Det nuværende web er "en samling tekster", som mennesker kan læse og bruge. Vi ser fremtidens web som "et semantisk web", der ikke bare kan læses, men som også selv kan finde de svar, vi søger. Og vel at mærke uanset hvilket sprog, vi søger på, fortæller seniorforsker Patrizia Paggio fra Center for Sprogteknologi i København, der sammen med universiteterne i Rom og København og en række softwarefirmaer i Italien, Frankrig og Holland står bag det europæiske MOSES-projekt. Et projekt, der først og fremmest vil undersøge to ting:

- Hvordan kan vi "mærke" de tekster, vi lægger på nettet, så søgemaskinerne har lettere ved at forstå, hvad de indeholder? **[Internt LINK til "** Nøgleord skal opsummere webteksters indhold**]**
- Hvordan får vi søgemaskinerne til at søge mere intelligent på nettet? **[Internt LINK til** "Søgemaskinerne skal forstå sprog**"]**

Nøgleord skal opsummere webteksters indhold
Ligesom biblioteker markerer hver bog med en række stikord, der fortæller, hvad bogen handler om, skal teksterne på fremtidens net også opmærkes. Altså forsynes med nogle nøgleord, som gør det muligt for søgemaskinerne at vurdere, om det enkelte dokument er relevant for en given forespørgsel.

- Det kræver selvfølgelig, at den, der lægger teksten på nettet, opmærker den med en række nøgleord. Men det burde ikke medføre det store besvær, vurderer Patrizia Paggio. - Ligesom vi i dag skal definere nogle grafiske virkemidler såsom skrifttype og - størrelse, når vi lægger en tekst op på nettet, vil vi i fremtiden blot skulle tilføje nogle stikord om, hvad vores tekst handler om, forklarer hun.

Den lidt ambitiøse tanke er da også, at opmærkningen en dag vil kunne foregå delvis automatisk.
- Med tiden vil "maskineriet" sandsynligvis blive så intelligent, at det selv kan foreslå nogle nøgleord at opmærke vores tekster med, så vi kan nøjes med at godkende dens forslag, lyder seniorforskerens fremtidsvision.

Søgemaskinerne skal forstå sprog
Skal forudsigelsen om automatisk opmærkning holde stik, er det dog langt fra tilstrækkeligt, at en søgemaskine kan genkende en række nøgleord i en tekst . Den skal også vide, at nogle ord,

der staves ens, betyder noget forskelligt, og at nogle vidt forskelligt stavede ord, kan have samme eller nært beslægtede betydninger.

- Tag for eksempel vores forsøg. De vil koncentrere sig om opmærkningen af dele af Københavns og Roms universiteters hjemmesider, forklarer Patrizia Paggio. - Her kan en studerende for eksempel tænkes at spørge søgemaskinen: ”Hvilke kurser udbydes i middelalderkunst i forårssemestret 2002?”

Når søgemaskinen leder efter svaret på det spørgsmål, kan den måske ikke umiddelbart finde nogen ”kurser” i middelalderkunst. Men så kan den måske finde et ”foredrag”, et ”seminar”, en ”workshop” eller en ”studiegruppe”. Ord, der alle har næsten samme betydning som et kursus, og som den studerende derfor kan tænkes at være interesseret i oplysninger om.

- Det kræver, at nogen har fortalt maskinen, at ordene har samme betydningsindhold, forklarer Patrizia Paggio. - Og derfor er det vigtigt for det semantiske web, at vi får skabt nogle fornuftige ontologier. Altså begrebssystemer, der relaterer ord og begreber til hinanden, så søgemaskinerne kan vide, at både et kursus og et seminar har noget med undervisning at gøre.

OntoQuery – Fordi thiamin også er et vitamin

Et sådant begrebssystem arbejder forskerne på det danske OntoQuery-projekt med at udvikle.

De har udarbejdet en prototype på en ontologi-baseret søgemaskine [LINK til http://www.ontoquery.dk], der kan bruges ved søgning om ernæringsrelaterede emner i Den Store Danske Encyklopædi.

- Vores søgemaskine skal kunne genkende samme indhold i forskellig forklædning, forklarer seniorforsker Bolette Sandford Pedersen fra Center for Sprogteknologi, der er en af partnerne i projektet. - For eksempel skal maskinen vide, at der er en vis relation mellem følgende begreber:

- Mangel på vitamin
- Vitaminmangel
- Mangel på thiamin
- Mangel på næring

Forskerne skal med andre ord forklare maskinen, hvordan sprog er bygget op. For mens nutidens søgemaskiner traditionelt arbejder ud fra:

- Genkendelse af mønstre (bogstaver og tal i bestemte kombinationer)
- Statistik (hvilke sider, der er bedst besøgt og linket mest til, og hvor og hvor ofte vores søgeord er nævnt i dokumentet)

skal fremtidens søgemaskiner også vide:

- Hvordan begreber relaterer til hinanden
- Om begreberne er fuldgyldige synonymer eller under- eller overbegreber til hinanden

- Især for lægfolk er det vigtigt, at maskinerne er i stand til at genkende begreber trods deres forskellige indpakninger, understreger Bolette Sandford Pedersen. - Eksperter kan jo ofte

spørge relativt præcist, mens lægfolk typisk spørger mere generelt. De ved for eksempel ikke nok om vitaminer til at spørge specifikt på et bestemt vitamin, men spørger i stedet generelt: ”Er der b-vitaminer i kornprodukter?” Og der skal maskinerne være intelligente nok til at vide, at svaret på det spørgsmål godt kan gemme sig i dokumenter om ”Thiamin i fødevarer”.

Kunstig intelligens
Og ”intelligens” er ordet. Eller rettere ”kunstig intelligens”.

- Da man i 50erne og 60erne fik øjnene op for muligheden for at skabe kunstig intelligens, ofrede amerikanerne store summer på forskning og udvikling. Men de opdagede, at sprog var en uendelig mere kompleks størrelse, end de havde forestillet sig, og derfor gik udviklingen i sig selv igen, fortæller ph.d. Troels Andreasen fra Datalogi ved Roskilde Universitetscenter, der er leder af Laboratoriet for Intelligente Systemer og projektleder på OntoQuery-projektet.

- Men nu er maskinernes regnekraft blevet så stor, at det giver helt nye muligheder for at udføre de komplekse beregninger, som er nødvendige for, at maskinerne kan lære at forstå, hvad det er, vi vil, forklarer Troels Andreasen.

Regnekraften gør forskellen
Den enorme regnekraft gør det blandt andet muligt at benytte statistiske metoder til at afgøre, hvordan og i hvilke sammenhænge, hvilke ord bliver brugt på nettet. Ligesom den gør det muligt at udstyre maskinerne med viden – for eksempel ontologier - der kan få dem til at ræsonnere sig frem til en bedre forståelse af, hvad vi mener.

- Maskinerne vil nok aldrig blive i stand til at forstå alt, hvad vi siger til dem. Men nye metoder, der udnytter den store regnekraft, kan hjælpe os et langt stykke ved at bearbejde store tekstmængder for os, forklarer Troels Andreasen. - Og når der så yderligere kobles nogle gode ontologier på, er vi nået langt, lyder hans forsigtige vurdering. – Men vi har ingen ambitioner om, at der en dag skal være én stor ontologi, som omfatter alle begreber og al viden i verden, understreger forskeren, der snarere forestiller sig en masse ontologier, der hver især knytter sig til specifikke fagområder, men som samlet vil tilføre søgning på nettet helt nye perspektiver.

Ankiro: Når erhvervslivet træner søgemaskiner og chatrobotter
---

At udviklingen af intelligente søgemaskiner også har kommercielle perspektiver vidner ikke alene store søgemaskiner som Google og Yahoos interesse for ontologierne om. Også hos de mindre softwareudviklere er der ”noget i gryden”.

Hos softwarefirmaet Ankiro på Frederiksberg satte jagten på intelligensen ind allerede i 1998. Og selvom virksomheden - som store dele af IT-branchen i øvrigt - oplevede godt med bølgegang i sluthalvfemserne, kan den i dag 14-mand store virksomhed fremvise adskillige beviser på, at maskiner *kan* oplæres. Ikke mindst i form af de søgemaskiner, den har udviklet til store virksomheder som Jubii, Kommunernes Landsforening, KPMG og Dataløn.

Ikke ord, men begreber
- Vi startede med at købe Politikens Ordbøgers retskrivnings- og synonymordbøger. På den måde fik vi mere end 60.000 ord og samtlige af deres bøjninger til rådighed, forklarer firmaets direktør, Bo Vincents.

- Men hvis søgningen skal være intelligent, skal den jo ikke søge på *ordet* økonomi, men på *begrebet* økonomi. Den skal omfatte alle bøjninger af begrebet, alle synonymer, alle over- og underbegreber. Og alle deres synonymer med under- og overbegreber og så fremdeles. Derfor indgår de også i vores søgninger. Om end med forskellig vægt, forklarer Bo Vincents.

Udgangspunkt i brugeren
Hans ord illustrerer tydeligt kompleksiteten i at arbejde med sprog.

- Fordi sprog er individuelt, og brugerne ofte kalder ting noget andet end deres officielle betegnelse, må vi tage udgangspunkt i brugerens terminologi, når vi oplærer maskinerne, forklarer direktøren.
- Søger en bruger for eksempel oplysninger om befordringsgodtgørelse på Dataløns hjemmeside [eksternt LINK: www.dataloen.dk], vil han sikkert skrive "kørselsfradrag" i søgefeltet, selvom det ikke er fagtermen. Og der skal maskinen være i stand til at lave koblingen fra kørselsfradrag til befordringsgodtgørelse.

Hertil kommer, at mange brugere enten staver dårligt eller skriver ukoncentreret. Og derfor har Ankiro udstyret sin søgemaskine med en fonetisk stavekontrol, der kan opfange og udligne slå- og stavefejl. Såsom at æ'et i "kørsælsfradrag" skal erstattes med et e, eller at j'et i "ferjepenge" skal erstattes med et i.

Chatrobotter servicerer de besøgende
Den teknologi kommer også brugerne til gode i Ankiros chatrobotter. For eksempel i chatrobotten *Rosa*, der på Socialdemokratiets hjemmeside [eksternt LINK: www.socialdemokratiet.dk] besvarer spørgsmål om partiets politik. Selv hvis de skrives så hjælpeløst som sætningen "Vad er jæres miliøpolitik?"

Udover ordbøger og ontologier har Rosa og Ankiros andre chatrobotter brug for en såkaldt *parser*. En komponent i sætningsanalysen, som ved hjælp af sin viden om strukturer og regler for sproget, kan finde svar på det, brugerne spørger om, og præsentere svarene i naturlige sætninger.

- Parseren trænes i, hvilke kombinationer af for eksempel emner, sætningstyper, og stemninger, der skal føre til hvilke svar. Og på den måde kan dialogrobotten give svar på tiltale – også når brugeren bliver lidt vel kærlig eller grov i tonen, forklarer Bo Vincents.

Naturligt sprog bliver et krav
En af de store udfordringer, Ankiro-folkene arbejder med, er at koble søgemaskinerne og chatrobotterne sammen, så brugerne frit kan vælge, om de vil serviceres i et naturligt sprog eller ved at skrive søgeord i et felt, når de søger oplysninger på nettet.

- Lige nu kan det måske være svært at se perspektiverne i at kunne kommunikere med søgemaskinerne på naturligt sprog, erkender Bo Vincents. - Men tanken bliver relevant, når den danske taleteknologi bliver brugbar. På det tidspunkt vil vi *tale* til computeren, og da vil det være unaturligt for os at kommunikere via andet end naturligt sprog, forudsiger han.

[Slut på artikel]

Om kilderne

Troels Andreasen er ph.d. i datalogi og ansat som lektor ved Datalogi på Roskilde Universitetscenter, hvor han ud over at undervise forsker i kunstig intelligens, databaser, informationssøgning og ontologier.
Homepage: www.ruc.dk/~troels
Mail: troels@ruc.dk

Patrizia Paggio er ph.d. i datalingvistik og ansat som seniorforsker hos Center for Sprogteknologi. Hendes seneste forskningsinteresser omfatter blandt andet natursprogsbehandling i søgning, dialogsystemer og brug af sprog og håndbevægelser i computergrænseflader.
Homepage: www.cst.dk
Mail: patrizia@cst.dk

Bolette Sandford Pedersen er ph.d. i datalingvistik og seniorforsker hos Center for Sprogteknologi, hvor hun forsker i leksikalsk semantik og ontologier.
Homepage: www.cst.dk
Mail: bolette@cst.dk

Bo Vincents er direktør i softwarevirksomheden Ankiro, der ved hjælp af informationsteknologi og sprogteknologi udvikler søgemaskiner, chatrobotter og andre produkter, der kan effektivisere vores kommunikation, informationssøgning og videndeling.
Homepage: www.ankiro.dk
Mail: bov@ankiro.com

Om OntoQuery-projektet

OntyQuery er et dansk, tværfagligt projekt med deltagelse af CST (Center for Sprogteknologi), DTU (Danmarks Tekniske Universitet), HHK (Handelshøjskolen i København), RUC (Roskilde Universitetscenter) og SDU (Syddansk Universitet).
Homepage: www.ontoquery.dk
Mail: het.id@cbs.dk

Om skribenten

Maria Bernbom er cand.ling.merc. og journalist. Hun er direktør i Kommunikationsbureauet Reflekt, hvor hun ved siden af sine tekstforfatnings-, rådgivnings- og undervisningsopgaver skriver artikler om blandt andet sprog og kommunikation.
Homepage: www.reflekt.dk
Mail: maria@reflekt.dk