Foto: Erik Bouwer

Google intelligent?

Krijgen we ooit de beschikking over slimme zoektechnologie? Het antwoord is een onvoorwaardelijk ‘ja’. Sterker, die zoektechnologie is er al en zal de komende jaren steeds meer worden ingebouwd in de zoeksystemen die we kennen, zoals Google.

Door Henk Verbooy

Dit vertelt dr. Suzan Verberne, onderzoeker aan de Radboud Universiteit Nijmegen op het gebied van taaltechnologie, tijdens haar presentatie op het congres Kennis in Praktijk op 17 juni jongstleden. Wie denkt aan de combinatie ‘slim’ en ‘computer’, denkt onwillekeurig aan kunstmatige intelligentie. Slimme zoektechnologie zal dan ook wel met intelligentie zijn toegerust, is dan de snelle conclusie. Dat die conclusie wel erg snel getrokken is, blijkt uit een gesprek met haar tijdens de lunchpauze van het congres.

DRIE MANIEREN VAN ZOEKEN

In de jaren zestig, bij het begin van de kunstmatige intelligentie, werd bij het ontwikkelen van zoekmachines sterk geleund op de taalkunde, en dan met name de ideeën van Chomsky[4]. Sinds de jaren negentig, met de opkomst van internet, staat *web search* in het middelpunt van de belangstelling (denk aan Google). Daarbij wordt vooral gebruikgemaakt van kennisarme technieken: tellen en statistiek. Verberne beschrijft in haar lezing drie manieren van zoeken. Naast web search zijn dat *text mining* en *question answering*. Bij laatstgenoemde methode worden complete en zeer specifieke vragen geformuleerd. Het antwoord wordt – net als bij Google – in de context gegeven en bovendien kunnen meerdere ‘hits’ worden getoond, gesorteerd op relevantie. Probeer het zelf op www.let.rug.nl/~tiedeman/joost/ of www.lexxe.com/. Text

mining, de derde zoekmethode, vindt z'n toepassing vooral in specifieke domeinen, zoals patenten. Een goed voorbeeld is PHASAR (www.phasar.cs.ru.nl), een nieuwe zoekmachine die gebruikmaakt van taalanalyse en die interactief – dus samen met de gebruiker – de vraag preciseert op basis van informatie in de documenten in de database in combinatie met een thesaurus. PHASAR is nog in de prototypefase, maar levert al heel goede resultaten in een pilot met Medline, een database met bibliografische verwijzingen en abstracts van biomedische tijdschriftartikelen. Text mining (zie kader) kan de gebruiker veel tijd en moeite kosten, maar de precisie van het resultaat kan heel exact bepaald worden. Voor professionele onderzoekers is dat laatste het belangrijkst.

Wat betekent dat voor de KI in uw vakgebied?
In de taaltechnologie is KI verdrongen door data. In plaats van kennis worden nu grote hoeveelheden data gebruikt. Hoe krachtig dat is, blijkt uit het feit dat je met alleen maar data kunt vertalen. Zo gaat Google Translate op internet op zoek naar websites (pagina's) die vertalingen van elkaar zijn. Bijvoorbeeld vertalingen van literatuur. Aan de hand van stukjes parallelle tekst – woordopeenvolgingen die waarschijnlijk vertalingen van elkaar zijn – komt een vertaalde tekst tot stand. De vertaling die je daarmee automatisch kunt genereren, is lang niet zo goed als een menselijke vertaling, maar goed genoeg om een tekst te begrijpen die in een voor jou onbekende taal is geschreven. Google Translate werkt net zo goed als vertaalsystemen die wel taalkundige regels gebruiken.
Een ander voorbeeld is automatische documentenclassificatie, het clusteren van documenten. Dat gebeurt puur op basis van tellingen: er zijn documenten die vaak deze woorden gebruiken, en er zijn documenten die vaak die woorden gebruiken, dus dat zijn waarschijnlijk verschillende clusters.

Is intelligentie niet ook begrijpen en interpreteren?
Een mens is natuurlijk heel veel intelligenter dan een computer. Een computer is goed in het verwerken van heel veel data, een mens is heel goed in het leggen van associatieve verbanden.

Maken we daar dan niet een fout door het te hebben over intelligente systemen, terwijl ze helemaal niet intelligent zijn, en alleen maar heel veel data kunnen verwerken…
Maar voor het leggen van associatieve verbanden heb je heel veel data nodig en moet er dus heel veel data worden verwerkt. Neem de vraag 'Waarom is Engels de taal van de Verenigde Staten?' We stelden die vraag aan een computersysteem. Het antwoord, dat overigens correct was, begon met 'Na de Britse kolonisatie van Noord-Amerika…'. Mensen leggen meteen de link tussen Brits en Noord-Amerika, tussen Engels en Amerikanen, et cetera. Al die begrippen zijn in ons hoofd aan elkaar gelinkt. Maar het zijn geen echte synoniemen van elkaar. In de klassieke manier van taaltechnologie bedrijven, rust je een zoekmachine uit met een synoniemenlijst, maar dan heb je het nog steeds niet opgelost. In de moderne manier gebruik je data om intelligentie te leren. Net als mensen, die doen dat ook. Als we opgroeien horen we heel vaak het woord Amerikanen gebruikt worden in de context van Verenigde Staten. We leren dat daar een koppeling tussen is. Zonder dat we expliciet leren wat het verband is, namelijk dat Amerikanen in de Verenigde Staten wonen. En kijk nu eens naar Wikipedia: daarin zit heel veel data die door mensen is geschreven. Daar halen we de kennis uit die mensen erin gestopt hebben: woorden die samen voorkomen, maar ook artikelen die naar elkaar linken. Je kunt dus een systeem bouwen dat die kennis er uit haalt. Ik vind dat je dan wel van een intelligent systeem mag spreken.

Dus Google is intelligent?
Ik vind Google al in een aantal dingen heel slim en handig. Er kan veel meer mee dan vroeger. Mensen weten dat niet altijd. Google zou soms beter moeten herkennen wanneer je vraag een vraag is. Maar vaak doet Google dat al. Als ik zoek naar een trailer van een film, dan komt Google ook met YouTube-resultaten. Als ik google op 'restaurants Rotterdam' krijg ik ook een kaartje van Google Maps. Ik heb pas samen met collega's gekeken of je kunt voorspellen uit een query die uit twee woorden bestaat of iemand misschien 'how to' had bedoeld. Als mensen bijvoorbeeld zoeken naar papier-maché, dan is de kans groot dat ze bedoelen 'hoe maak ik papier-maché?'. Uit de statistiek kun je dat afleiden. Google wordt daarom steeds slimmer. Er werkt ook een groot aantal onderzoekers dat heel veel ontwikkelt. Maar ze zijn heel terughoudend het ook officieel aan te bieden. Ze zijn denk ik toch bang dat mensen zullen afhaken als het te complex wordt.

Kunstmatige intelligentie

Het KI-vakgebied werd officieel gedefinieerd in 1956 tijdens een conferentie op de campus van het Dartmouth College in de USA[1]. De verwachtingen waren hoog gespannen. Marvin Minsky[2], een van de belangrijkste initiatiefnemers, voorspelde dat KI binnen één generatie geen geheimen meer zou hebben en alle problemen waar de ontwikkelaars mee kampten opgelost zouden zijn. Maar het lag allemaal net iets lastiger; Minsky c.s. bleken te optimistisch, wat resulteerde in een afnemende belangstelling van de overheid – vooral de Britse en de Amerikaanse – met als gevolg steeds minder subsidie.
Van een brede belangstelling voor kunstmatige intelligentie (KI) is eigenlijk pas sprake sinds de jaren tachtig. In die tijd leek KI een commercieel succesvolle toepassing te hebben gevonden in de vorm van *expert systems*: computersystemen die menselijke kennis en expertise konden vastleggen. Ook het Japanse Fifth Generation Computer Systems project (FGCS)[3] droeg bij aan die hernieuwde belangstelling. FGCS was een publiek/privaat samenswerkingsinitiatief van het Japanse ministerie van Handel en Industrie en had als doel de ontwikkeling van een nieuwe generatie supercomputers en (daarmee) een nieuwe impuls voor KI. Het project was een mislukking of – afhankelijk wie je spreekt – z'n tijd te ver vooruit. Hoe dan ook, van de voorziene zesde generatie is het nooit gekomen. En omdat ook de expert systems niet die brede toepassing brachten waarop gehoopt werd, verdween KI weer uit het zicht. Desondanks is KI springlevend en wordt het vandaag de dag op verschillende terreinen toegepast: robotica, data mining, diagnostiek, spraakherkenning en logistiek zijn aansprekende voorbeelden.

En de betrouwbaarheid van de resultaten?
Onbewust weten we heel goed welke resultaten betrouwbaar zijn en welke niet. We weten dat we documenten terugkrijgen die de woorden bevatten die we hebben ingetypt, niets meer en niets minder. Daarnaast is het niet voor niets dat wanneer je noord-zuidlijn intikt NRC, gemeente Amsterdam en Wikipedia hoog scoren. Daar klikken mensen vaak op, omdat ze weten dat die bronnen betrouwbaar zijn. Dat is een vorm van *crowdsourcing*.

Mensen hebben genoeg gezond verstand om niet te zeggen 'het staat op Google dus het is waar?"
Ja. Mensen voelen heel goed aan wat wel en wat niet betrouwbare bronnen zijn. Natuurlijk, je kunt heel veel dingen vinden die niet waar zijn. Maar als je echt wilt weten of iets waar is, dan kom je daar wel achter.

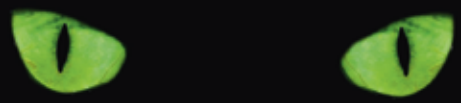
Text mining
Zoeken: in een grote database worden documenten gevonden die over het gezochte onderwerp gaan.
Analyseren: (delen van) de gevonden documenten worden door het systeem gecategoriseerd en geanalyseerd op relevante informatie.
Presenteren: gestructureerde teksten worden aan de gebruiker gepresenteerd. Specifieke informatie die gevraagd is wordt apart getoond.

Noten
[1] Bron: http://en.wikipedia.org/wiki/Artificial_intelligence
[2] Marvin Lee Minsky (1927), o.m. cognitiewetenschapper op het gebied van KI en mede-oprichter van het KI-laboratorium van het Massachusetts Institute of Technology.
[3] Bron: http://en.wikipedia.org/wiki/Fifth_generation_computer.
[4] Avram Noam Chomsky (1928). Amerikaans linguïst, filosoof, cognitiewetenschapper. Emeritus hoogleraar linguïstiek aan het Massachusetts Institute of Technology. Chomsky is de grondlegger van de generatieve grammatica; een model dat uitgaat van universeel aanwezige aangeboren structuren. Zie ook http://nl.wikipedia.org/wiki/Noam_Chomsky.

Industry Watch

Wie zijn de belangrijkste leveranciers en adviseurs voor kenniscentra, bibliotheken en andere kennisintensieve organisaties? Wat doen ze en wat leveren ze? U vindt ze op www.ikmagazine.nl in de rubriek Industry Watch. Een online overzicht doorzoekbaar op naam, hoofdactiviteit en/of productsoort.

Is uw organisatie actief in de wereld van kennis en informatie? Levert u technologie, advies of andere diensten? Zorg dat kennis- en informatiespecialisten u weten te vinden. Laat de informatie over uw bedrijf opnemen in Industry Watch. Basisregistratie is gratis. Neem contact op met moniquedejong@essentials-media.nl en vraag naar de mogelijkheden.



Industry Watch