DE COMPUTER BEGRIJPT ME

Piek Vossen lijkt zelf wel wat op een supercomputer. Hij is taalwetenschapper en programmeur; gedegen onderzoeker en efficiënt projectmanager. Hij werkt hard en succesvol aan computers die taal door en door begrijpen.

Eerst was er Deep Blue, de schaakcomputer die in 1997 won van schaakkampioen Gary Kasparov. In januari 2011 won computer Watson het kennisspel Jeopardy tegen twee topspelers. De supercomputer van IBM kan menselijke taal begrijpen en complexe vragen snel beantwoorden. Anna die op ikea.nl je vragen beantwoordt is er niets bij. *What 's next?*
Misschien wel dat je je hart kunt uitstorten tegen je computer. Een begripvolle avatar (m/v, stel in naar keuze) snapt niet alleen je woorden, maar ook je gezichtsuitdrukkingen. Hij interpreteert de klank en het volume van je stem en reageert bovendien zo dat hij je echt op weg helpt je problemen op te lossen. De computer heeft alle aspecten van menselijke communicatie onder de knie gekregen en alle mensen krijgen op een goedkope manier alle aandacht die ze nodig hebben. Interessant, aangezien ons zorgstelsel nu al uit zijn voegen barst.
Hoogleraar computationele lexicologie Piek Vossen (1960) onderzoekt bij de letterenfaculteit van de VU hoe computers taal kunnen begrijpen. Hij ziet nog niet zo snel een volwaardige psychotherapeut als standaardonderdeel van Windows, maar toch: "Op alle fronten wordt onderzoek gedaan, van het interpreteren door computers van gezichtsuitdrukkingen en stemklank tot woordkeus en context. Op een gegeven moment wordt dat naar elkaar toegebracht. Tot nu toe zie je vooral entertainment en franje, de Ikea-avatar is simpel. Maar als je de spierballen hebt van IBM én je beheerst de sociale aspecten van taal, kan de computer steeds menselijker communiceren."

BONUSKAART VAN ALBERT HEIJN

Wat Piek Vossen doet, kun je vergelijken met wat Albert Heijn via de Bonuskaart doet. De supermarkt registreert wat wij allemaal kopen, analyseert dat en ontdekt zo structuur in ons koopgedrag.

> 'Een Chinees praat heel anders over het milieu dan een Nederlander'

Data mining heet dat. Vossen doet aan *text mining*: structuur ontdekken in opgeschreven teksten. Hij legt het uit: "Taal is heel ambigu. Een tekst begrijpen is meer dan alleen woorden kennen. Een Chinees praat heel anders over bijvoorbeeld het milieu dan een Nederlander. In Nederland heeft het milieu een knuffelgehalte, in China niet. Als je een milieutekst vertaalt zoals Google Translate dat nu redelijk kan, ben je nog niet zo ver dat de Chinees en de Nederlander elkaar helemaal begrijpen." Taal bestaat uit woorden plus een context. Dat kan de intentie van de spreker zijn, of bijvoorbeeld de politieke, culturele of persoonlijke achtergrond van de spreker. Contexten herkennen en meenemen bij het interpreteren, daaraan werkt Vossen.
Neem bijvoorbeeld het wereldwijde klimaatprobleem. Dat lossen we gemakkelijker op als wetenschappers uit alle landen elkaar kunnen begrijpen. Vossen coördineert het EU-project Kyoto met als doel een internationaal wiki-portaal over milieu en ecologie. Het is een systeem waarmee bepaalde gemeenschappen, in dit geval klimaatwetenschappers, hun eigen terminologie en begrippen zo kunnen definiëren dat een computer die ook kan gebruiken. De computer haalt vervolgens de pure feiten uit de teksten. Je kunt dan in een oogopslag zien welke jaartallen genoemd worden voor het smelten van de gletsjers van de Himalaya, wie dat beweert en wat er beweerd wordt voor andere gletsjers in de wereld. Het systeem beheerst zeven talen en kan gemakkelijk worden uitgebreid naar andere talen. Als dit werkt, is het ook voor andere onderwerpen mogelijk zo'n systeem in te richten.

GESCHIEDENISRECORDER

Zo kreeg Vossen in mei bijna drie miljoen euro van de Europese Unie om een viertalige 'geschiedenisrecorder' te gaan ontwikkelen op het gebied van economisch nieuws. Samen met Italiaanse en Spaanse collega's en onder meer informatiedatabank LexisNexis gaat hij een computerprogramma ontwikkelen dat dagelijks het nieuws 'leest' en indexeert: het houdt precies bij wat wanneer waar gebeurd is en wie er bij betrokken is. Het maakt uit al die versnipperde nieuwsberichten één correct verhaal over het verleden. Het idee hierachter is dat zoeken in databanken tegenwoordig zinloos is geworden: ze bevatten te veel informatie. Wie bekijkt er meer dan de eerste tien hits van Google? We hebben dus meer geavanceerde hulp nodig. Een hulpprogramma moet niet alleen goed selecteren, maar ook interpreteren. Anders kunnen we zijn resultaat niet vertrouwen, laat staan er beleidsbeslissingen op baseren.
Dat interpreteren door een computerprogramma is juist de clou van Vossens werk.

VOORBEELD GESCHIEDENISRECORDER: VOLKSWAGEN & OVERNAME

Typ je in Google de zoekvraag 'Volkswagen take-over' dan krijg je duizenden hits met berichten door de jaren heen waarin Volkswagen is betrokken bij een overname. Ga je die resultaten goed bekijken, dan zie zie je dat de laatste jaren sprake is geweest van een overnameoorlog in de auto-industrie. Rond 2008/2009 vind je veel berichten dat Porsche van plan is om Volkswagen over te nemen. Ze kopen aandelen en er is speculatie op een overname. Vervolgens verandert de situatie binnen een jaar, mede door de bankencrisis uit 2008. Volkswagen koopt nu aandelen Porsche, management van Porsche wordt aangeklaagd door aandeelhouders en Volkswagen komt als winnaar uit de bus. Dit verhaal (en vele anderen) volgt niet uit de zoekresultaten van Google. Het kan pas worden opgebouwd door de berichten te interpreteren aan de hand van een hele reeks programma's. Voor ieder woord moet worden bepaald naar welke gebeurtenis het verwijst, wie er bij betrokken is (Porsche, Volkswagen, het management), wanneer het gebeurd is, of het ook echt gebeurd is. Vervolgens moeten die gebeurtenissen achter elkaar gezet worden tot een geschiedenis. Die programma's samen vormen dus uiteindelijk een geschiedenisrecorder.

De keus voor de focus op financieel-economisch nieuws is niet zomaar. In dit domein is veel informatie beschikbaar – dus een goede *test case* – en er ligt een grote beslisdruk bij allerlei professionals om er snel op te reageren. Zij zullen zo'n systeem met open armen ontvangen.

ALFA-BÈTA-KLIK

Hoe kon Piek Vossen eigenlijk zo'n belangrijke speler worden op zijn vakgebied? Uberhaupt is het al bijzonder dat iemand alfa en bèta tegelijk is. "Op de middelbare school was ik eerst een echt bètajongetje. Tot ik geïnteresseerd raakte in muziek en literatuur. Niet alleen lezen en luisteren maar ook schrijven en componeren.

> 'Op de middelbare school was ik eerst een echt bètajongetje'

Ik raakte ook gefascineerd door het systeem: wat gebeurt er precies als je teksten schrijft of muziek componeert? Hoe kom je van een idee tot woord of gevoel tot klank? Ik raakte ook sociaal bewogen, ik wilde schrijver worden. Ik ging Nederlands studeren. Toen ik daar kennismaakte met de mathematische generatieve grammatica, was daar de alfa-bèta-klik."
Vossen wilde weten hoe cognitieve processen in onze hoofden werken én hoe je die processen logisch kunt vastleggen. Dat is precies de kern van zijn werk. "Taalwetenschap is nu pas empirisch aan het werk. Nu pas analyseren we data en schrijven we computerprogramma's." Als hij een computerprogramma klaar heeft dat het interpreteren van ons overneemt, is zijn doel bereikt. En dan is er een volgende stap gezet op weg naar die psychotherapeut als standaardonderdeel van Windows.

SPINNENWEBBEN VAN ALLE WOORDEN

Als je eenmaal weet wat je wilt weten, hoe word je dan een succesvol wetenschapper? Een van Vossens succesfactoren is dat hij een model wist te ontwikkelen dat omarmd werd in de onderzoekswereld. Hij maakte in opdracht van de Europese Unie zogeheten wordnets in acht talen. Dat zijn spinnenwebben van alle woorden van een taal, voorzien van hun betekenissen en gegroepeerd in synoniemen en betekenisrelaties. Ze zijn toepasbaar in kunstmatige intelligentiesystemen, zoals die waar Vossen nu aan werkt. Tot 1996 was er alleen een Engelstalig wordnet. Vossen was de initiatiefnemer van de uitbreiding naar acht Europese talen. Dat was geen simpel kopieerwerk, want elke taal zit

> 'In het Spaans en Italiaans bijvoorbeeld, bestaat geen woord voor huisdier'

anders in elkaar en iedereen zit min of meer opgesloten in zijn eigen taal. Vossen: "Dat werk confronteerde me met fundamentele vragen. Wat is een woord en wat is een concept? Zijn alle denkbare samenstellingen woorden en concepten? Zijn er concepten zonder woorden? Welk woord is universeel? In het Spaans en Italiaans bijvoorbeeld, bestaat geen woord voor huisdier. Dat het Nederlands dat woord wel kent, zegt iets over de Nederlandse cultuur."
Toen de klus erop zat, was het wat betreft de EU klaar. "Dat zou het einde betekenen van het raamwerk van mensen en methoden waarmee we hadden gewerkt", zegt Vossen. "En dat wilde ik niet. Het is te belangrijk." Zonder dat raamwerk worden de 'spinnenwebben' voor de verschillende talen niet volgens dezelfde principes gebouwd en kunnen ze ook niet aan elkaar gekoppeld of met elkaar vergeleken worden. Daarom richtte hij met anderen in 2000 de Global WordNet Association op, waarvan hij nu een van de voorzitters is. De voornaamste doelen van deze wereldwijde non-profit organisatie zijn het ontwikkelen van wordnets in alle talen en die aan elkaar te koppelen. Dit project wordt ook wel de Global Wordnet Grid genoemd. Je kunt dan van iedere wordnet naar iedere ander wordnet springen via de betekenis van de woorden. Hoeveel talen in de wereld hebben eigenlijk een woord voor het concept huisdier? Het Italiaans bijvoorbeeld niet.

RISICO'S

Affiniteit met alfa én bèta, een omarmd model en een goed initiatief: nog steeds niet het complete recept van Vossens succes. Het is ook zijn manier van zoeken naar de essentie. "Ik kijk graag op een andere manier naar dingen. Ik sla graag stappen over, ik ga niet één voor één in vaste volgorde naar een eindpunt. Ik vergelijk het met de manier waarop kinderen dingen aanpakken. Toen onze kinderen klein waren, speelden we vaak het avonturenspel Myst. Je moest een puzzel oplossen en tot mijn verbazing waren de kinderen er beter in dan wij. Zij kijken niet conventioneel, maar durven stappen te zetten die niet voor de hand liggen. Ik probeer dat ook te doen. Daarmee neem ik risico's, maar als je methodologisch sterk bent, kun je er ver mee komen."

Tom Arends en Rianne Lindhout

>>

TROTS OP DE VU

Piek Vossen is blij dat hij op de VU terecht kwam. “Het samenwerken met anderen, ook van andere faculteiten, gaat hier heel goed. Ik heb elders wel meegemaakt dat collega’s vooral met rust gelaten wilden worden. Hier is de organisatie er voor jou; mensen werken niet langs elkaar heen. De sfeer is menselijk en prettig.”
Wel merkt de bètageoriënteerde taalwetenschapper dat hij binnen de letterenfaculteit een vreemde eend in de bijt is. “Als ik een grote hoeveelheid tekst wil onderzoeken, kan ik dat niet bij Letteren doen. Als een evangelist probeer ik mensen op de faculteit ervan te overtuigen dat we daarvoor faciliteiten nodig hebben, ook voor studenten.” Gelukkig is de exacte faculteit dichtbij: aan de overkant van het campusplein. Daar staat wél de supercomputer die Vossen nodig heeft voor zijn onderzoek. Het is voor hem dan ook prettig dat interdisciplinair onderzoek binnen de VU belangrijk is. Zogeheten IOZI’s, interdisciplinaire onderzoeksinstituten, hebben er een belangrijke status. Zo werkt Vossen binnen het Netwerkinstituut samen met onderzoekers van de faculteiten Exacte en Sociale wetenschappen.”