Een database van feiten

Taalwetenschappers werken aan computerprogramma's die automatisch feiten uit teksten filteren. Maar om dat goed te doen is nog heel wat wiskunde nodig.

Het eerste probleem voor een computer die teksten verwerkt, is dat één woord allerlei betekenissen kan hebben. Piek Vossen, hoogleraar Computationele Lexicologie aan de Vrije Universiteit, noemt *slag* als zijn favoriete voorbeeld: "Slag heeft achttien verschillende betekenissen. Je kunt je slag slaan, van slag zijn, een raar slag mensen hebben en zo nog vijftien. Als ik ze opnoem, dan ken je ze allemaal."

In een gemiddelde tekst heeft 60 tot 80% van de woorden verschillende betekenissen. Mensen zien tijdens het lezen vrijwel onmiddellijk de juiste betekenis, maar voor een computer is een tekst een enorme puzzel. Een artikel van duizend woorden waarin 70% van de woorden vijf mogelijke betekenissen heeft, geeft al ettelijke miljarden mogelijke combinaties.

De computer moet op de een of andere manier leren om bij elk woord in een tekst de juiste betekenis te kiezen. Het doel van Vossen is een computerprogramma dat nieuwsberichten leest en daaruit filtert welke feiten en meningen in het bericht staan. Zo wil hij een database van kennis opbouwen.

Lastig is dat twee teksten over hetzelfde onderwerp niet altijd dezelfde woorden gebruiken. Vossen: "Het ene nieuwsbericht heeft het over een aanslag, het andere over een aanval. In het ene staat dat Barack Obama er iets over zegt, in het andere staat dat president Obama dat doet." De computer moet bepalen in hoeverre gebruikte woorden op elkaar lijken. Hoe dichter de woorden bij elkaar staan in betekenis, hoe waarschijnlijker het is dat de artikelen over hetzelfde onderwerp gaan. Hierbij ontstaan enorme tabellen met informatie die met elkaar vergeleken moet worden.

Mate van gelijkheid

Om gelijkheid van woorden in kaart te brengen, gebruiken onderzoekers een netwerk dat aangeeft welke woorden vaak in dezelfde context voorkomen. Zo krijgen ze een soort stamboom waarin vergelijkbare woorden bij elkaar staan.

Bij het sorteren gebeuren soms onverwachte dingen. In Leuven maakten taalkundigen een netwerk van woorden met maar één betekenis. Ze gebruikten onder andere het woord *monitor* en hoopten dat hun model zou aantonen dat dit leek op het woord *beeldscherm*. Maar er dook een heel cluster van monitoren op in een heel ander deel van het netwerk. Toen bleek dat in Vlaanderen het woord *monitor* ook gebruikt wordt voor vrijwilligers die toezicht houden op spelende kinderen. Zo ontdekte het model automatisch de twee verschillende betekenissen.

Vossen is ervan overtuigd dat je de Wet van Zipf (zie pag. 79 van 'Talen temmen' door woordgebruik te tellen) kunt doortrekken naar betekenis. "Er zitten heel algemene wiskundige patronen achter de dingen die wij bestuderen."

Het is echter lastig om die patronen helder in beeld te krijgen uit een verzameling van losse teksten, omdat het vastleggen van woorden en betekenissen heel lastig blijkt. Neem het Nederlandse corpus SoNaR dat een steekproef van teksten uit allerlei genres bevat. Het bestaat in totaal uit 500 miljoen woorden. Groot, maar niet groot genoeg om de Nederlandse taal te vangen. Vossen: "We hadden een lijst van woorden met betekenissen die iedereen kent, het waren echt geen gekke dingen. Toch stond 28% van die betekenissen helemaal nergens in dat enorme corpus, terwijl we ze wel op internet konden vinden."

Uitzonderlijke formules

Op dit moment zitten de beste formules om de betekenis van een woord te kiezen in ongeveer 66% van de gevallen goed. Vossen ziet hoe in zijn vakgebied allerlei varianten van dezelfde formule rondgaan. "Wij zijn geen wiskundigen en formules vanuit het niets bedenken is lastig. Dus grasduinen we in de literatuur of iemand een soortgelijk probleem heeft opgelost met een wiskundig model, en dan passen we die oplossing een beetje aan. Maar dat is altijd een beetje riskant. In elke dataverzameling over taal staan zeldzame gevallen, en dan baseren mensen hun formule dus op woorden of betekenissen die maar één keer voorkomen."

Sommige statistici beweren dat hun formule enorm goed werkt, maar dan blijkt dat ze alleen goed scoren op een heel specifieke dataset. Vossen: "Bij een andere verzameling teksten blijft er vaak niet veel van over. Als ze getraind hebben op het NRC dan scoren ze daar 80%, maar bij de Telegraaf zijn ze dan slechts 60% correct. We zien dat het resultaat van al die verschillende formules sterk op elkaar begint te lijken als we ze op een grotere dataset gebruiken. Dus in essentie doen ze gewoon hetzelfde."

Vossen zoekt daarom ook samenwerking met bètawetenschappers die wel goed thuis zijn in formules. "Voor onze database van feiten willen we bijvoorbeeld ook een elegant model om bepaalde informatie in artikelen zwaarder mee te wegen als we twee teksten vergelijken. Daar ga ik binnenkort eens over praten met wiskundigen."