Enterprise Language Processing

Het is al een tijdje een cliché om te zeggen dat we dreigen te verdrinken in de oceaan van bits die over het Internet gestuurd wordt. Gelukkig bestaat bijna de helft van die bits uit ongevaarlijk vermaak (audio en video). Het overgrote deel van de rest van die bits, zoals het verkeer op Facebook of Twitter, is ook alleen relevant voor een klein aantal privépersonen. Maar dat neemt niet weg dat er ook berichten omgaan die wel degelijk relevant zijn voor overheden en bedrijven. De politie wil graag meteen meer weten van tweets over bommen in de vertrekhal van Schiphol voordat ze door ongeruste burgers gewaarschuwd wordt. Bedrijven willen meteen weten dat er in blogs de draak gestoken wordt met een nieuw product. Overheden en bedrijven kunnen er belang bij hebben dat ze meteen weten dat er ineens een boel te doen is over een specifiek onderwerp, bijvoorbeeld geruchten over nieuwe producten waar een bedrijf mee zal komen of voorgenomen fusies. Het gaat om minuscule speldjes in een enorme hooiberg, maar wel speldjes die grote gevolgen kunnen hebben als ze niet op tijd ontdekt worden. Het behoeft geen betoog dat het vinden van die speldjes enorm bemoeilijkt wordt doordat berichten in heel veel verschillende talen geformuleerd kunnen zijn.

Lou Boves
CLST, Radboud Universiteit Nijmegen

Enterprise Language Processing zou een onderzoeks- en ontwikkelingsprogramma moeten worden dat voortbouwt op STEVIN en dat de nodige hulpmiddelen oplevert voor de beheersing van de informatiezondvloed, specifiek met het oog op geschreven en gesproken documenten in het Nederlands. Het lange-termijn doel van dat programma is het creëren van automaten die talige documenten kunnen 'begrijpen', zodat ze kunnen beslissen welke documenten voor de organisatie waarvoor ze 'werken' relevant zijn. Ze moeten de informatie in die documenten zo kunnen presenteren dat beslissers in de kortst mogelijke tijd verantwoorde keuzes kunnen maken en beslissingen nemen.

Het automatisch begrijpen van teksten in een natuurlijke taal is een heilige graal in een aantal onderzoeksgebieden, waaronder in ieder geval de informatica, de kunstmatige intelligentie, de filosofie en de taalwetenschap. Dat maakt *Enterprise Language Processing* tot een multidisciplinaire onderneming.

Waar staan we in 2012?

De modale *DIXIT*-lezer zal wellicht geen behoefte hebben aan een overzicht van wat er op dit moment mogelijk is bij automatisch verwerken van natuurlijke taal. Een paar verwijzingen volstaan. Zoals de soms akelige precisie waarmee Google advertenties op je scherm zet die gerelateerd zijn aan wat je net in een e-mail geschreven, of op een webpagina gelezen hebt. Of het feit dat IBM's Watson computer nationale kampioenen in het spelletje *Jeopardy* verslaat. En, wat gesproken taal betreft, de prestaties van toepassingen zoals *Siri* en haar broertjes en zusjes in andere mobiele platformen.

De modale *DIXIT*-lezer weet ook dat er nog heel veel niet kan. We weten nog steeds niet hoe we op een zoekvraag kunnen reageren met een feitelijk antwoord, in plaats van met een eindeloze rij documenten waar het antwoord misschien in staat. We weten nog steeds niet hoe we alle dreigtweets automatisch kunnen herkennen, en hoe we snel alle internetfora kunnen vinden waar een product of een idee belachelijk gemaakt of juist aangeprezen wordt. En we weten nog steeds niet hoe we van de computer een effectieve tutor kunnen maken, die behulpzaam kan reageren op halve vragen en antwoorden van een leerling.

Het kan wel interessant zijn om na te gaan hoe de huidige kennis en technologie tot stand gekomen is. In juli 2012 heeft er in Amerika een discussie gewoed over de vraag of het 'Internet' het resultaat is van onder-zoek dat grotendeels betaald is door de overheid, of dat het ontstaan is door investeringen van commerciële bedrijven. Wie naar de geschiedenis kijkt, kan niet anders dan instemmen met de conclusie van het recente rapport Continuing Innovation in Information Technology [1] van de Amerikaanse National Research Council, opgesteld door een commissie die voornamelijk bestond uit vertegenwoordigers van de grote IT-bedrijven: met publiek geld gefinancierd onderzoek ligt aan de basis van zo ongeveer alle informatietechnologie die we hebben, maar het waren de bedrijven die de **toepassingen** van de basistechnologie en de basiskennis ontwikkeld hebben. Jammer genoeg is er in Nederland geen Topsector 'Informatie- en Communicatietechnologie' die een soortgelijke publiek-private samenwerking kan stimuleren.

Hoe zouden we verder moeten gaan?

Als er één trend is die in het oog springt in het wereldwijde onderzoek op het veld van *Enterprise Language Processing*, dan is dat de toepassing van zelflerende algoritmen die gebruik maken van een hele boel verschillende soorten bronnen: van de documenten die via het Internet toegankelijk zijn tot gegevens over de keuzes en voorkeuren van afzonderlijke individuen. Er is onderzoek nodig naar nieuwe algoritmen die met een minimale vorm van supervisie kunnen leren van hun omgang met (voornamelijk talige, maar ook niet-talige) gegevens. Een fundamentele vraag is hierbij tot op welke hoogte een lichaamsloze machine talige informatie kan 'begrijpen' door te zoeken naar verbanden tussen woorden in teksten en niet-talige gegevens over de 'echte wereld'. Waarschijnlijk is het probleem nog lastiger als de woorden 'geraden' moeten worden wanneer het om gesproken documenten gaat.

Communicatie is per definitie sociaal. Kennis wordt gezien als informatie die met andere leden van een groep gedeeld wordt, en die binnen een groep op dezelfde manier geïnterpreteerd wordt. We hebben het stadium bereikt waar dit sociale karakter van taalcommunicatie niet meer genegeerd kan worden. Dat heeft geleid tot nieuwe methoden zoals Crowd Sourcing en Social Information Processing, die voor de ontwikkeling van *Enterprise Language Processing* onmisbaar zijn. Maar de inzet van die methoden vereist ook nog een boel onderzoek: hoe kun je mensen stimuleren om hun kennis en ervaring ter beschikking te stellen, zonder dat hun economische belangen en privacy in het geding komen, en hoe kun je de kwaliteit van de bijdragen van onbekenden bewaken? Juist doordat taal zonder context geen betekenis heeft zal het fundamentele onderzoek altijd ingebed moeten zijn in de ontwikkeling van concrete toepassingen, door concrete bedrijven, voor concrete gebruikers. Er is dus hoe dan ook een vorm van samenwerking nodig tussen universiteiten, hogescholen, overheidsorganisaties en bedrijven.

[1)] *http://www.nap.edu/catalog.php?record_id=13427*

Veel Europese talen bedreigd met digitale uitsterving

Minstens 21 Europese talen lopen grote risico's om het digitale tijdperk niet te overleven. Daarvoor waarschuwen vooraanstaande taaltechnologische experts uit heel Europa in een nieuwe studie. Op 26 september, de Europese dag van de talen, is een reeks van 30 'witboeken' oftewel taalrapporten gepresenteerd, die per taal de risico's inzichtelijk maken. De studie is uitgevoerd door META-NET, een Europees excellentienetwerk met 60 onderzoeksinstellingen in 34 landen. Ook voor het Nederlands is een dergelijk taalwitboek geschreven.

Jan Odijk
UiL-OTS, Universiteit van Utrecht

Ruim 200 experts hebben voor 30 van de ongeveer 80 Europese talen vastgesteld in hoeverre zij digitaal worden ondersteund met taaltechnologie. De conclusie luidt dat de digitale ondersteuning voor 21 van de 30 talen 'niet-bestaand' is of op zijn best 'zwak'. Bekende voorbeelden van taaltechnologische toepassingen zijn programma's voor spellings- en grammaticacontrole, interactieve persoonlijke assistenten op smartphones (zoals Siri op de iPhone), gesproken telefoonmenu's, automatische vertaalsystemen, zoekmachines op het web, en de stemmen in autonavigatiesystemen.

Slechte taaltechnologische voorzieningen

Voor iedere taal is de taaltechnologische ondersteuning op vier verschillende gebieden vastgesteld: automatisch vertalen, spraakinteractie, tekstanalyse en de beschikbaarheid van taalbronnen. Verschillende talen, bijvoorbeeld IJslands, Lets, Litouws en Maltees krijgen de laagste score op alle gebieden. In totaal scoren 21 talen slecht op minimaal één gebied. Opmerkelijk is dat geen enkele taal de categorie 'excellente ondersteuning' krijgt. Alleen het Engels wordt beschouwd als een taal met 'goede ondersteuning', gevolgd door talen zoals het Nederlands, Frans, Duits, Italiaans en Spaans met 'beperkte ondersteuning'.

Blijvende inspanningen nodig voor ondersteuning van het Nederlands

De situatie van het Nederlands geeft naar mijn mening aanleiding tot voorzichtig optimisme. Dat er voor het Nederlands 'beperkte