Antal van den Bosch, Co-professor vakgroep Induction of Linguistic Knowledge, Universiteit Tilburg
(antal.vdnbosch@uvt.nl)

Computers die praten en begrijpen waar je het over hebt zijn al tientallen jaren een typisch kenmerk van sciencefiction films. Zoals in zoveel andere terreinen van de kunstmatige intelligentie loopt de werkelijkheid daarbij achter; er zijn weliswaar belangrijke stappen gezet sinds de jaren '50, maar we weten allemaal dat de computer nog geen vlot meebabbelende conversatiepartner is. Wat is er in de laatste vijftig jaar dan wél gebeurd?

Computerlinguïstiek: een korte geschiedenis

De computerlinguïstiek, het vakgebied dat zich bezighoudt met de verwerking van natuurlijke taal door computers, is een menggebied van de linguïstiek (taalkunde), en de meer door informatica gedreven kunstmatige intelligentie. Het vakgebied ontstond in de jaren zestig[1] toen taalkundigen en informatici elkaar vonden. De invloedrijke taalkundige Noam Chomsky slaagde er in die jaren in om een formalisering van taal te vinden die streng minimalistisch en wiskundig was opgezet, zodat het voor de hand lag om te denken over computationele implementaties ervan.

Intussen ontwikkelde de computer zich als een steeds flexibeler inzetbare machine met allerlei nieuwe input- en outputmogelijkheden, en begonnen bedrijven zich te interesseren voor de automatische verwerking van taal en spraak door computers. Amerikaanse telecombedrijven als AT&T en kantoorautomatiseerders als Xerox, NEC en IBM richtten speciale onderzoekslaboratoria op om taal- en spraaktechnologie te ontwikkelen. Omdat deze bedrijven eerder geneigd waren om informatici in dienst te nemen dan (Chomskyaanse) taalkundigen ontstond er een wat eigenwijs vakgebied dat veel pragmatischer omging met taal, omdat er simpelweg systemen gebouwd moesten worden "die het deden".

In Nederland speelde Hugo Brandt Corstius een pioniersrol in het begin van de jaren '70 (zijn proefschrift uit 1970 was getiteld "Exercises in Computational Linguistics"). Aan het eind van de jaren '70 liepen er voldoende opgeleide computerlinguïsten in Nederland rond, onder andere opgeleid door Brandt Corstius, zodat de eerste vakgroepen en opleidingen gestart konden worden. De Universiteit van Tilburg had in 1987 de primeur met de afstudeerrichting *Taal en Informatica*; Groningen, Amsterdam, Utrecht, Twente en Nijmegen volgden.

De jaren '80 kunnen achteraf bezien worden als een gouden tijdperk voor de kunstmatige intelligentie, en ook voor de computerlinguïstiek en de taaltechnologie, ook in Nederland. Na de piek volgde echter een anticlimax. Veel beloftes konden niet worden waargemaakt; de band tussen universiteit en industrie werd weer losser, en tenslotte verdwenen ook de meeste bedrijfslaboratoria. Vooral dankzij de actieve bemoeienis van de nationale onderzoeksorganisatie NWO, die sinds begin jaren '90 een vijftal thematische programma's organiseerde rondom taal- en spraaktechnologie, is het vakgebied binnen de universitaire muren springlevend gebleven.

Betrekkelijk los van de taaltechnologie ontwikkelde de spraaktechnologie zich in de jaren '70 en '80 ook stormachtig, voortgedreven door grote investeringen van bedrijven. Gedreven door pragmatische doelen kwam men er in de jaren '80 achter dat probabilistische methoden uiterst effectief waren voor spraakherkenning. Tegelijkertijd ontwaakte de *machine learning* als een nieuw eigen vakgebied binnen de kunstmatige intelligentie, en begonnen computerlinguïsten, geïnspireerd door hun collega's uit de spraaktechnologie, deze automatisch lerende systemen ook toe te passen op taken in natuurlijke taalverwerking, zoals tekst-naar-spraakomzetting en automatisch ontleden. Dat ging zo snel en met een dusdanig groot succes dat binnen tien jaar tijd het vakgebied grotendeels is overgestapt op het gebruik van lerende systemen en probabilistische methoden voor de modellering van processen in de verwerking van natuurlijke taal.

"Induction of Linguistic Knowledge" in Tilburg

Binnen de Faculteit Communicatie en Cultuur van de Universiteit van Tilburg doet de vakgroep Taal- en Informatiewetenschappen sinds medio jaren '80 computerlinguïstisch onderzoek. Sinds 1990 wordt onderzoek gedaan naar de toepassing van lerende systemen in natuurlijke taalverwerking. De focus van dit deel van het onderzoek, geïnitieerd en uitgebouwd door prof. Walter Daelemans (nu hoogleraar aan de Universiteit Antwerpen), is het *memory-based learning*, een klasse lerende systemen die niet leren door een model te abstraheren uit voorbeelden, maar door de voorbeelden zelf onverkort te onthouden. Pas als het nodig is gaan deze zogenaamde *lazy learners* aan de slag; bij een nieuw binnenkomend voorbeeld dat moet worden verwerkt (geclassificeerd) zoekt een memory-based learner de meest gelijkende voorbeelden in zijn ge-

1. Een groot archief van het vakgebied is de ACL Anthology: http://acl.ldc.upenn.edu/

Een antwoord gevonden door de ROLAQUAD vraag-antwoordmodule op de vraag "Waardoor kan RSI ontstaan?" Het antwoord bevat het gevraagde concept (de aandoening RSI) en de gevraagde relatie ("veroorzaakt").

heugen, en extrapoleert de oplossing van de meest gelijkende voorbeelden naar het nieuwe geval. Deze extreem simpele leermethode blijkt goed te passen op natuurlijke taalverwerking, omdat abstractie helemaal niet goed is te rijmen met taal; het is vrijwel altijd beter om uitzonderingen en vreemde constructies in taal te onthouden, omdat ze altijd terugkomen. Wat voor een abstraherend lerend systeem ruis lijkt, is in werkelijkheid meestal een productieve uitzondering.

De onderzoeksgroep, door Daelemans de *Induction of Linguistic Knowledge* (ILK) onderzoeksgroep gedoopt, telt momenteel veertien medewerkers, waaronder vier promovendi, zes postdoc-onderzoekers en drie wetenschappelijk programmeurs. Er vindt hoofdzakelijk onderzoek plaats dat gefinancierd wordt uit nationale onderzoeksgelden, met name van het NWO en van SenterNovem[2]. Dit artikel belicht een aantal van deze lopende projecten die zich richten op toepassingen van taaltechnologie in uiteenlopende terreinen.

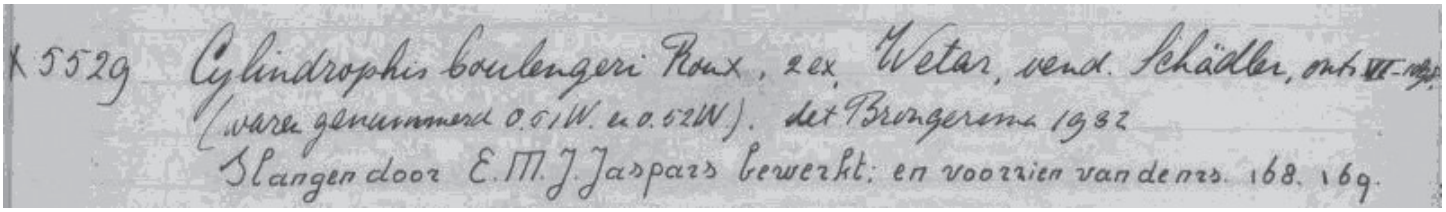
Hulp bij schrijven

Twee onderzoeksprojecten in de ILK onderzoeksgroep richten zich op automatische hulpmiddelen bij het schrijfproces. Sinds de computer de typmachine heeft vervangen schrijven we allemaal zelf, en schrijven we veel. De gereedschappen die ons ten dienste staan ontstijgen de typmachine eigenlijk niet eens zoveel; de meeste tekstverwerkers doen nog steeds alsof we op papier aan het typen zijn, en bieden niet veel meer nieuwigheid dan de deleteknop in plaats van het flesje tippex. Met de geavanceerde technologie die er bij is gekomen is men maar al te goed bekend: de spellingcorrector, die van de tien keer dat hij alarm slaat, er gemiddeld negen keer naast zit.

Het onderzoek van Martin Reynaert in het *D-Coi* project (Dutch Corpus Initiative) en daarvoor zijn promotieonderzoek, richt zich op de ontwikkeling van betere automatische spellingcorrectors. Reynaert specialiseerde zich tijdens zijn promotieonderzoek in het probleem van de *non-word error*, ofwel de "typfaut", die ten eerste gedetecteerd dient te worden, en vervolgens gecorrigeerd naar de bedoelde vorm. Momenteel wordt er in mijn NWO-Vici project *Implicit Linguistics* ook gewerkt aan de zogenaamde *confusible*, de typfout waarbij het ene woord ("too") verward wordt met een andere vorm ("to", of "two"). Om beide problemen op te lossen is het essentieel dat je naar de context, de linker- en rechterbuurwoorden van de vermeende fout kijkt. Met de *memory-based learning* aanpak blijkt dit goed te kunnen; zo is de groep hard op weg met een oplossing voor de bekende "dt"-fout.

In het door SenterNovem gefinancierde project *A Propos* ontwikkelt promovendus Toine Bogers oplossingen voor het probleem van het zoeken naar informatie tijdens het schrijfproces. Creatieve processen zoals het schrijven van teksten zijn effectiever wanneer ze niet teveel worden onderbroken door andersoortige taken, zoals bijvoorbeeld het googelen naar informatie. In het A Propos-project wordt een bestaande tool gebruikt, de IntelliGent agent van de in het project deelnemende bedrijven Search Expertise Centrum en IntelliGent B.V., die als achtergrondproces van het systeem meeleest met wat de gebruiker

2. SenterNovem is een agentschap van het Ministerie van Economische Zaken (Red.)



Een fragment uit een handgeschreven registerboek van Naturalis. Het fragment beschrijft een in Indonesie in 1898 gevonden slang, nog immer bewaard op sterk water in Leiden. In het MITCH project wordt een dergelijke beschrijving na digitalisering voorzien van metadata: diernaam, vinder, locatie, jaartal van determinatie, enzovoort.

typt. Voortdurend plukt de tool zoektermen uit de tekst die de schrijver net heeft geproduceerd, en stuurt deze door naar zoekmachines. Wanneer de tool relevante documenten gevonden denkt te hebben, bijvoorbeeld op het internet of op de harde schijf van de schrijver, verschijnt een pop-up in beeld met klikbare links naar de gevonden pagina's.

Bogers' werk concentreert zich op het verbeteren van de bestaande "Google"-achtige functionaliteit, door de tool automatisch uit te laten zoeken wie er in een werkgroep expertkennis bezit over specifieke onderwerpen. Zodra een nieuw document wordt geschreven dat aansluit bij de expertise van een van de collega's van de schrijver, wijst de tool op de relevante documenten van de collega-expert. Deze technologie, die de naam *recommender systems* draagt, bekend van de boekaanbevelingen van Amazon en de muzieksuggesties van Pandora, is hard op weg om een volgende generatie van zoekmachines te vormen.

Antwoorden op vragen

In het ROLAQUAD project (*Robust Language Understanding in Question-Answer Dialogues*) doen promovendus Sander Canisius en postdoc-onderzoeker Piroska Lendvai onderzoek naar methoden om antwoorden te vinden op vragen naar medische informatie. Het project doet dat in een groter nationaal verband, het NWO IMIX-programma, waarin verschillende taal- en spraaktechnologische onderzoeksgroepen modules bij elkaar brengen zoals een spraakherkenner, een spraaksynthesizer, en modules voor het plannen en managen van een dialoog. Het Tilburgse project levert een van de drie modules die de antwoorden bij gestelde vragen zoeken. Vragen kunnen ingetypt of ingesproken worden, wat betekent dat zelfs vragen over hetzelfde onderwerp nog op tientallen of honderden mogelijke manieren gesteld kunnen worden.

De methode die in ROLAQUAD is ontwikkeld zoekt antwoorden op vragen in medische encyclopedieën. ROLAQUAD maakt daarvoor gebruik van de overlap in woorden tussen de vraag en alle mogelijke antwoorden (passages in de encyclopedieën), maar berekent daarnaast ook een overlap in betekenis. Dat lijkt ambitieus, maar in een enigszins gesloten domein als "medische informatie" is er een redelijk beperkte lijst van objecten en concepten, en de relaties daartussen, die een groot deel van de medische kennis beschrijven: medicijnen, aandoeningen, lichaamsdelen, micro-organismen, en relaties als "veroorzaakt", "behandelt" en "voorkomt". De ROLAQUAD-module, die gebaseerd is op een *memory-based learner*, herkent deze domeinspecifieke concepten en relaties in een gestelde vraag, zoekt vervolgens het stuk tekst op dat de grootste overlap in concepten en relaties bevat, en geeft dan dit stuk tekst terug als antwoord.

Natuurhistorisch erfgoed

Onderzoekers van de ILK-groep zijn ook buiten de deur werkzaam. In 2005 begon het MITCH project, Mining Information from Texts in the Cultural Heritage, als onderdeel van het NWO CATCH programma, dat op dit moment in tien projecten zoals MITCH erfgoedinstellingen samen laat werken met universitaire onderzoeksgroepen op het gebied van het geavanceerd omgaan met digitaal erfgoed. In MITCH werkt de Tilburgse groep samen met Naturalis, het Nationaal Natuurhistorisch Museum in Leiden. In het onderzoekslab van Naturalis werken wetenschappelijk programmeur Steve Hunt, promovendus Marieke van Erp en postdoc-onderzoeker Caroline Sporleder aan wat wel de verborgen schat van Naturalis wordt genoemd: de vele honderden veldboeken en registerboeken die in detail beschrijven waar, door wie en in welke omstandigheden de miljoenen opgeslagen dieren en planten zijn gevonden. Zonder deze boeken is serieus onderzoek naar de collectie onmogelijk. Werken met de boeken zelf is berucht tijdrovend. Pas met gedigitaliseerde versies van boeken wordt het mogelijk om in korte tijd zittend achter de computer complexe onderzoeksvragen te beantwoorden, zoals: "Wat was de verspreiding van deze specifieke familie van gifkikkers in de Amazone, van 1800 tot nu, gemeten in periodes van 25 jaar?"

In het MITCH project worden de bestaande gedigitaliseerde boeken, die al handmatig zijn omgezet naar databases, automatisch opgeschoond (door middel van lerende

systemen) en voorzien van aanvullende informatie (metadata) die het mogelijk maakt om complexere zoekvragen te kunnen stellen over de data. Daarbij staat het principe voorop dat de menselijke experts het laatste woord hebben over veranderingen en verrijkingen. MITCH beoogt de experts van Naturalis, de collectiemanagers en de taxonomen, simpelweg veel tijdwinst te bezorgen.

Tot slot

Meer informatie over de ILK onderzoeksgroep is te vinden op de webpagina van de groep: *http://ilk.uvt.nl*. Op de website zijn publicaties te vinden, webdemo's van taaltechnologische modules, en software om zelf aan de slag te gaan met machine learning en taaltechnologie. In 2007 start aan de Universiteit van Tilburg de master-opleiding "Human Aspects of Information Technology", waarbinnen de onderzoeksgroep onderwijs aanbiedt over taal- en spraaktechnologie en de toepassing daarvan in informatie- en kennistechnologie. Een goed overzicht van de stand van zaken in de computerlinguïstiek in heel Nederland en België krijg je als je de jaarlijkse CLIN bijeenkomst (Computational Linguistics in The Netherlands, sinds 1990) bezoekt. De zeventiende CLIN wordt georganiseerd op 12 januari 2007 in Leuven:
http://www.ccl.kuleuven.be/CLIN17/. ø

Agenda

23 november 2006

[50 jaar AI]

Het Laboratorium voor Artificiële Intelligentie Brussel van de VU Brussel organiseert een feestje ter gelegenheid van de 50ste verjaardag van AI. Met onder andere Luc Steels (VU Brussel, België) en Tony Belpaeme (Universiteit Plymouth, Engeland).
http://arti.vub.ac.be/50jaarai/

25 november 2006

[Memen vs. Genen]

Studium Generale van de Universiteit Utrecht organiseert een symposium over mementheorie. Met onder andere Dr. Susas Blackmore (Universiteit Bristol) en Dr. Bas Haring (Universiteit Leiden)
http://www.sg.uu.nl/prog/2006b/memen.html

8 december 2006

[Symposium Verbeter de Wereld met AI]

De Connectie presenteert in samenwerking met de Vrije Universiteit van Amsterdam op 8 december 'Verbeter de Wereld met AI'.
http://www.connectie.org

19 januari 2007

[Mens & Machine]

Studiecentrum Soeterbeeck Ravenstein (RU Nijmegen) organiseert een seminar over de vrijheid en de vrees die techniek ons brengt. Met onder andere Dr. Gert-Jan Lokhorst (TU Delft) en Dr. Jan Vorstenbosch (Universiteit Utrecht).
http://www.ru.nl/actueel/agenda/seminars/2007/mens_en_machine/

Ook een agendapunt voor De Connectie?
Mail ons!
redactie@connectie.org