Taaltechnologie voor mensen met communicatieve beperkingen: een optie?

Emiel Krahmer, Erwin Marsi en Lilian Beijer

Inleiding

Voor een grote groep Nederlanders is het lezen van "gewone" teksten niet eenvoudig, omdat ze om uiteenlopende redenen beperkt zijn in hun leesbegrip. Moderne taaltechnologische toepassingen zouden voor deze groep een uitkomst kunnen bieden, bijvoorbeeld door teksten automatisch te vereenvoudigen en meer toegankelijk te maken. Toch zijn tot dusver maar weinig van dit soort toepassingen ontwikkeld. Hoe komt dat? Een mogelijke verklaring is dat mensen uit de taal- en spraakpathologie en uit de taaltechnologie elkaar doorgaans niet zo vaak tegenkomen, en bovendien een andere professionele taal spreken. Voor iemand uit de wereld van de taal- en spraakpathologie vraagt het enige inspanning om al die taaltechnologische uiteenzettingen te doorgronden, en vervolgens nog een tandje extra om de brug te slaan tussen deze relatief onbekende maar boeiende materie enerzijds en de realiteit van mensen met communicatieve beperkingen anderzijds. De taaltechnoloog heeft het omgekeerde probleem; want hoewel het begrijpen van de taaltechnologische ontwikkelingen meestal niet al teveel problemen oplevert, heeft de taaltechnoloog weinig tot geen zicht op de specifieke behoeftes en wensen van mensen met communicatieproblemen. In dit artikel, dat een direct resultaat is van de TST (Taal- en SpraakTechnologie)-themadag "De Gebruiker Centraal" (Rotterdam, 30 november 2006), proberen we een brug te slaan tussen deze beide perspectieven, door eerst stil te staan bij de specifieke doelgroep van mensen met communicatieve beperkingen, en vervolgens te kijken wat de taaltechnologie voor die groep te bieden heeft.

Communicatieve beperkingen en behoeftes

Mensen kunnen om uiteenlopende redenen te maken hebben met communicatieve beperkingen, en de aard en ernst van die beperkingen kunnen bovendien behoorlijk verschillen van persoon tot persoon. Een specifieke groep hierbij zijn de mensen met communicatieve beperkingen als gevolg van verworven neurologisch letsel. Voorbeelden van verworven neurologisch letsel zijn een beroerte of traumatisch hersenletsel als gevolg van een ongeval. Tot de mogelijke gevolgen van deze lesies behoren fatische stoornissen (verminderd taalbegrip en/of verminderde taalproductie) en andere cognitieve stoornissen zoals verminderde aandacht en geheugenstoornissen. De beperkingen op communicatief vlak van deze patiënten kunnen zich uiten in onder andere een verminderd leesbegrip. Met

name lange teksten met veel complexe zinnen, minder frequent voorkomende woorden en abstracte betekenissen kunnen voor deze groep patiënten problemen opleveren. Concreet betekent dit dat deze patiënten beperkt zijn in bijvoorbeeld het lezen van krantenartikelen, tijdschriften en brieven van instanties. Deze beperkingen kunnen leiden tot sociaal isolement en afhankelijkheid van anderen. Gebruik van software met behulp waarvan teksten vereenvoudigd kunnen worden of waarmee oefenmateriaal gegenereerd kan worden ten behoeve van een verbeterd leesbegrip kan bijdragen aan het in communicatief opzicht optimaal functioneren. Het vereenvoudigen en samenvatten van teksten is overigens niet alleen voor deze groep relevant, het zou ook een belangrijk ondersteunend communicatiemiddel voor bijvoorbeeld dyslectische leerlingen kunnen zijn.

Wat heeft taaltechnologie te bieden?

Automatisch samenvatten is een algemene toepassing die al geruime tijd in de belangstelling van taaltechnologen staat, en er zijn inmiddels redelijk betrouwbare methodes om op basis van allerlei features automatisch te bepalen wat de belangrijkste zinnen in een tekst zijn. Het vereenvoudigen van zinnen kan gezien worden als een soort extreme samenvattingstaak, waarbij de vereenvoudigde zin als een samenvatting gezien kan worden van de oorspronkelijke zin. Toch zijn dit geen technieken die met hun huidige kwaliteit direct ingezet kunnen worden voor de hierboven beschreven doelgroep. De belangrijkste beperking is dat het achter elkaar plakken van de belangrijkste zinnen uit een tekst niet meteen een vlot leesbare samenvatting oplevert; de opbouw en structuur van de tekst wordt immers doorbroken. Een andere complicatie, zeker bij zogenaamde multi-document samenvattingen, is dat er veel overlap kan zitten tussen zinnen. Het heeft weinig zin om twee zinnen in de samenvatting op te nemen die beiden weliswaar belangrijke informatie bevatten, maar tegelijkertijd ook in grote mate overlappen. Het feit dat zinnen in hun betekenis kunnen overlappen is een algemeen probleem voor allerhande taaltechnologische toepassingen, en is precies het onderwerp van het DAESO (Detecting And Exploiting Semantic Overlap) project, dat gefinancierd wordt door Stevin en loopt van 2006 tot 2009.

Het DAESO Stevin project

Vergelijk de volgende twee zinnen eens (beide een antwoord op de vraag “Wat is een beroerte?”):

> *(1) Wordt bloedvoorziening naar de hersenen plotseling ergens onderbroken, spreken we van een beroerte.*

(2) Een beroerte of CVA (cerebrovasculair accident) is een plotse onderbreking van de bloedstroom naar een bepaald deel van de hersenen.

Deze twee zinnen vertonen een sterke semantische overlap, hoewel er maar twee inhoudswoorden zijn die letterlijk in de beide zinnen voorkomen ("beroerte" en "hersenen"). Daarnaast bevatten de zinnen ook woorden en frases die semantisch gerelateerd maar niet identiek zijn, zoals "bloedvoorziening" in (1) en "bloedstroom" in (2). Hoewel er het een en ander bekend is over semantische relaties (zoals synoniemen) tussen woorden, is er veel minder bekend over semantische overlap op frase en zinsnivo, en dit is specifiek waar het DAESO-project zich op richt. Er wordt in dit project aan tools gewerkt die deze overlap kunnen detecteren, vertrekkend vanuit twee gerelateerde zinnen, zoals zinnen (1) en (2) hierboven. De zinnen worden eerst automatisch syntactisch geanalyseerd, bijvoorbeeld met behulp van de Alpino parser die ontwikkeld is aan de Universiteit van Groningen. Op basis van deze analyses wordt vervolgens automatisch gekeken welke woorden en frases uit de ene zin gerelateerd zijn aan woorden en frases uit de andere zin, en deze worden "opgelijnd" (aligned). Niet alleen worden semantisch overlappende frases opgelijnd, ook de aard van de overlap wordt automatisch vastgesteld; "een beroerte" in zin (1) is bijvoorbeeld minder specifiek dan "een beroerte of CVA (cerebrovasculair accident)" in zin (2). Op basis van deze analyses kan vervolgens geprobeerd worden om een nieuwe zin te produceren die de gedeelde informatie uit beide zinnen bevat ("zinsfusie"). Aangezien we nu ook informatie hebben over de specifieke semantische relaties tussen de frases in de respectievelijke zinnen, zou het mogelijk kunnen zijn om automatisch een vereenvoudigde versie van de beide zinnen produceren: "Een beroerte is een plotselinge onderbreking van de bloedvoorziening van de hersenen." Op deze manier hopen we tevens de kwaliteit van multi-document samenvattingen te kunnen verbeteren. Door belangrijke zinnen uit verschillende brondocumenten te fuseren, komt er minder redundantie in de samenvatting. Daarnaast kan overbodige of gedetailleerde informatie uit zinnen gefilterd worden ("zinscompressie, of -vereenvoudiging"). Zo proberen we de informativiteit en leesbaarheid van automatisch geproduceerde samenvattingen te kunnen verbeteren. De aanpak die we hierboven hebben beschreven gaat uit van *paren* gerelateerde zinnen, maar dat is geen essentiële beperking. Op basis van de verkregen inzichten is het ook mogelijk om bijvoorbeeld een enkele zin te vereenvoudigen, door bepaalde woorden of syntactische constructies te vervangen door eenvoudigere varianten.

Een belangrijke complicatie is dat er wel resources zijn voor het detecteren van semantische overlap op woord-nivo (zoals WordNet of de binnen Stevin ontwikkelde Cornetto database), maar er weinig beschikbaar was voor semantische overlap op zinsnivo, zeker voor het Nederlands. Vandaar dat het eerste jaar van het DAESO project in het teken stond van het aanleggen van een corpus dat hiervoor gebruikt kan worden. Dit corpus bestaat uit paren zinnen uit verschillende tekstgenres en met een verschillende mate van semantische overlap.

Aan de ene kant van het spectrum zitten bijvoorbeeld autocue-ondertitel paren van het NOS journaal. De ondertiteling is vaak een kortere versie van de zinnen die nieuwslezers van de autocue lezen, en dit is een nuttige bron van informatie voor zinsvereenvoudiging. Het andere uiterste zijn verschillende nieuwsberichten die dezelfde gebeurtenis beschrijven. Hier is de mate van overlap veel meer wisselend, wat met name informatief is voor het ontwikkelen van multi-document summarizers.

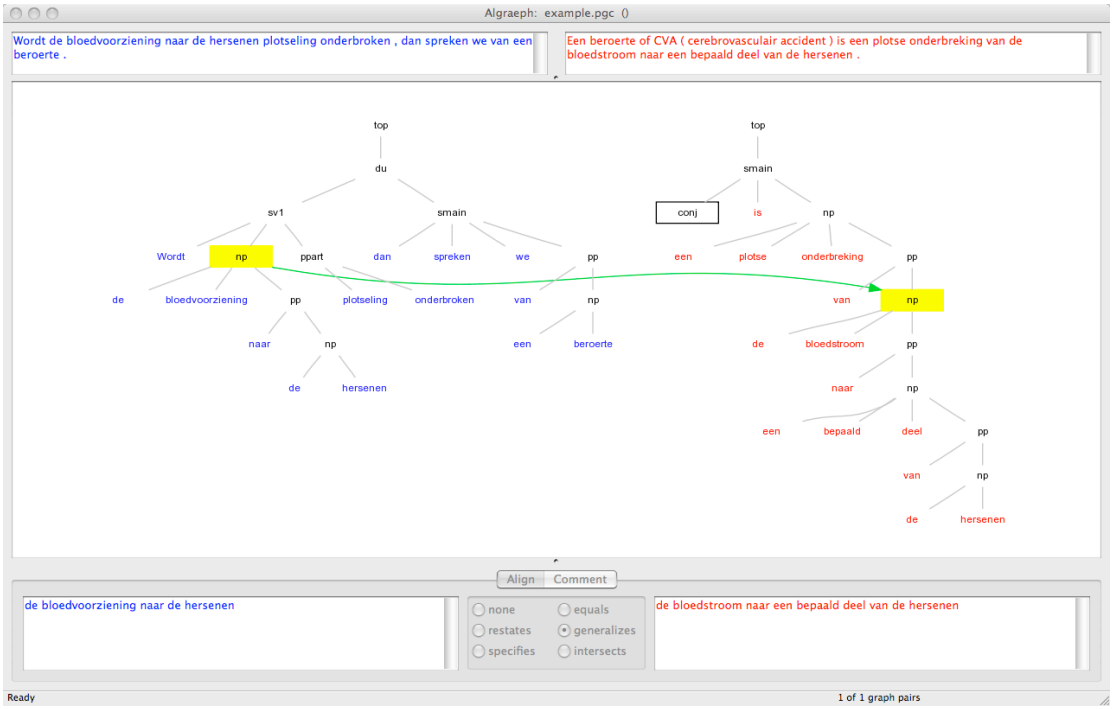
Toepassingen

De DAESO- tools zijn op dit moment nog volop in ontwikkeling, en het is op dit moment nog onduidelijk in welke mate deze tools succesvol zullen zijn. Maar ze hebben duidelijk de potentie om ingezet te worden bij toepassingen voor mensen met communicatieve beperkingen. Naast min of meer standaard-toepassingen als automatisch samenvatten, zouden zinsfusie en zinscompressie ook gebruikt kunnen worden om zinnen en teksten te vereenvoudigen zodat deze meer toegankelijk worden. Daarnaast zouden inzichten over semantische overlap ingezet kunnen worden om moeilijke syntactische constructies om te zetten in eenvoudigere, en om abstracte woorden of frases te vervangen door meer concrete varianten. Te denken valt ook aan het automatisch berekenen van de leesbaarheid van een tekst (met behulp van formules uit de taalbeheersingswereld), en het automatisch reviseren van teksten die met behulp van bovenstaande technieken als te moeilijk zijn beoordeeld. Ook is het denkbaar om trainingsmateriaal te ontwikkelen voor deze groep, bijvoorbeeld door automatisch parafrases te genereren voor een zin, om zo zinsbetekenissen beter te kunnen illustreren in trainingsmateriaal.

Conclusie

Software zoals hierboven beschreven zou vele voordelen kunnen bieden. Hulpmiddelen die het lezen van de krant of van brieven van instanties vereenvoudigen, kunnen bijdragen aan een grotere zelfstandigheid en een volwaardiger maatschappelijke participatie van mensen met communicatieve beperkingen. Uiteraard kunnen verbeterde en meer actuele oefeningen ook helpen bij het bereiken van deze doeleinden. Al met al lijken er dus allerhande mogelijkheden te zijn om taaltechnologie in te zetten voor mensen met communicatieve beperkingen. Reden te meer voor taal- en spraakpathologen en taaltechnologen om elkaar eens wat vaker te zien, en elkaars professionele taal beter te leren spreken. Wellicht kunnen uit deze contacten projecten voortvloeien die de toepassing en evaluatie van taal- en spraaktechnologische tools voor communicatief beperkte doelgroepen beogen.

Meer informatie over Daeso, zie: http://daeso.uvt.nl/.



Een screenshot van de Algraeph annotatietool die wordt gebuikt om de syntactische structuren van gelijksoortige zinnen handmatig op te lijnen. In het bovenste paar tekstboxen staan de twee invoerzinnen. In het midden staan de syntactische structuren voor deze twee zinnen. De rechthoek rondom "conj" geeft aan dat dit deel van de syntactische structuur momenteel is "ingeklapt". De door de annotator geselecteerde knopen zijn geel gemarkeerd; de groene pijl geeft aan dat deze twee knopen zijn opgelijnd. De met deze knopen corresponderende woordreeksen staan in het onderste paar textboxen. De semantische relatie tussen deze woordreeksen - zoals weergegeven in het middelste paneel van "radiobuttons" - is "generalizes", omdat de linker woordreeks een generalisatie is van de rechter woordreeks. Een volledige oplijning van twee zinnen bestaat uit een oplijning van alle gelijksoortige knopen en hun semantische relaties.