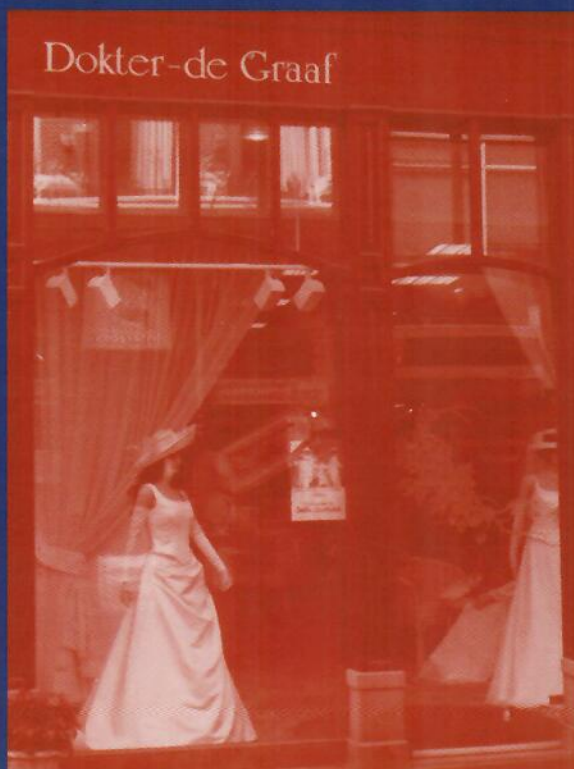


DIXIT

TIJDSCHRIFT OVER TOEGEPASTE TAAL- EN SPRAAKTECHNOLOGIE

Genetische algoritmen

Spraakherkenning bij PCM



‘De kunst van
intelligente
naamsinterpretatie’

'Di•gi•taal'

**Taalsoftware op
elke werkplek**

**ook voor
intranet**



van Dale

IN BEDRIJF

www.vandale.nl/kantoren

INHOUD

De kunst van intelligente naamsinterpretatie

Steeds meer bedrijven erkennen het belang van datakwaliteit. Goede kwaliteit van de klantgegevens is een essentiële voorwaarde voor klantgerichte bedrijfsvoering en de implementatie van CRM geworden. Correcte verwerking van namen van bedrijven en personen vereist echter diepe kennis van zaken. Wat betekenen de verschillende naamselementen, uit welke onderdelen bestaan ze, en hoe hangen deze onderdelen met elkaar samen. Kortom: identificatiesoftware moet in staat zijn om tenaamstellingen te interpreteren.

4

Spraakherkenning bij PCM

In 2003 en 2004 heeft TeleCats voor een aantal uitgevers van landelijke dagbladen een service ontwikkeld waarmee incidentele klachten automatisch worden afgehandeld. Bellers worden geïdentificeerd via hun ingesproken postcode waarna het systeem controleert of op deze postcode de afgelopen tijd niet al eerder een klacht is binnengekomen. Is dit niet het geval, dan handelt de computer de klacht af. Een interview met Renate van der Toorn over de voor- en nadelen van dit systeem.

8

Organisatie, let op uw woorden!

Organisaties zijn zich steeds bewuster van hun imago. Ze doen er alles aan om dat imago positief 'uit te stralen'. Een nieuw terrein dat daarbij aandacht krijgt, is de geschreven taal van de organisatie. Taaltechnologie helpt een handje. Ruim baan voor de 'elektronische eindredacteur'.

13

Rubrieken

Column	7
Team voor Taal 2005	10
Uitgelicht	11
Uit de boeken	16
NOTaS-nieuws	17

Names of mass destruction

Niemand zal nut & noodzaak van taal- en spraaktechnologie betwisten – om maar even in Betuwelijnjargon te spreken – maar op de een of andere manier lukt het maar niet om bovengemiddelde aandacht te krijgen. De 'populaire' vakpers vermeldt plichtmatig de zoveelste update van een of andere applicatie, of publiceert het zoveelste jubelverhaal waarin een bedrijf vertelt hoe succesvol zijn systeem wel niet is.

Misschien is de huidige alomtegenwoordige nadruk op efficiency – in bedrijfsmatige termen wel te verstaan – er de oorzaak van dat TST in de publiciteit het rudimentaire niveau maar niet ontstijgt. TST is veel meer dan een spellingchecker of een IVR-systeem. Met andere woorden: het belang en de mogelijkheden van TST worden ondergewaardeerd. Niet dat kostenbeheersing door bijvoorbeeld het gebruik van een IVR-systeem *onbelangrijk* is, maar alleen hier de nadruk op leggen doet TST tekort.

Met behulp van TST worden op zeer efficiënte wijze tot nu toe ontoegankelijke archieven blootgelegd. Ook is het sinds enige tijd mogelijk om het materiaal van televisiezenders te doorzoeken zonder dat elke band afgedraaid hoeft te worden. Bovendien zal, als de voorkeuren niet bedriegen, op korte termijn het semantisch web de kinderschoenen zijn ontgroeid, en kunnen de miljarden webpagina's worden gefilterd op inhoud en relevantie. Niet dat we door deze technologische ontwikkelingen het kostenbeheersingparadijs eindelijk kunnen betreden, wel dat toegang wordt verkregen tot informatie, tot *kennis*. En informatie en kennis is nog altijd de toegangspoort naar de toekomst. En, inderdaad, naar efficiency.

Maar TST kan ook een belangrijk wapen zijn om maatschappelijk ongemak te bestrijden. Vorig jaar was er sprake van enig tumult omdat ambulances en brandweerwagens naar verkeerde adressen uitrukten vanwege bestandsvervuiling. Onnodig leed, omdat bestanden met behulp van TST eenvoudig zijn te schonen. Met behulp van dezelfde technieken kan bijvoorbeeld uitkeringsfraude worden aangepakt, of er juist voor worden gezorgd dat iemand de uitkering krijgt waar hij recht op heeft. Allemaal een kwestie van de juiste *match* die door de inzet TST kan worden bewerkstelligd.

En zelfs het meest actuele maatschappelijke ongemak, de huidige permanente terreurdreiging, heeft baat bij de mogelijkheden van TST. Je kunt voor elk gebouw een zwaarbewapende soldaat posteren of, iets rigouzeuzer, een compleet leger inzetten om waar ook ter wereld mogelijke terroristen op te jagen en uit te schakelen, maar dit zal een kat-en-muis-spel blijken zolang je de voedingsbodem van die muis niet wegneemt: geld! Het is bepaald niet moeilijk voor 'professionele' terroristen om het geld daar te krijgen waar het nodig is, omdat de systemen achter de wereldwijde geldstromen de *match* niet maken tussen Usama bin Laden, Osama bin Laden en Oesama bin Laden. Of Al-Kaida en Al-Qaida. Om maar een paar willekeurige namen te noemen. TST maakt het mogelijk om deze naamnuances te filteren en te duiden, zodat de namen achter de wapens grijpbare mensen worden. En dat is in elk geval wel zo efficiënt.

Pierre Pieterse
Hoofdredacteur DIXIT

De kunst van intelligente naamsinterpretatie

Kennnis van relatiezaken?



Steeds meer bedrijven erkennen het belang van datakwaliteit. Goede kwaliteit van de klantgegevens is een essentiële voorwaarde voor klantgerichte bedrijfsvoering en de implementatie van Customer Relationship Management (CRM) geworden. Om het beoogde kwaliteitsniveau te bereiken, worden onder de noemer identificatiesoftware allerlei tools ingezet voor zaken als invoercontrole, fraudebestrijding en het koppelen en ontdebellen van bestanden.

| HOLGER WANDT |

Correcte verwerking van namen van bedrijven en personen vereist echter diepere kennis van zaken. We moeten weten wat de verschillende naamselementen betekenen, uit welke onderdelen ze bestaan en hoe deze onderdelen met elkaar samenhangen. Kort en goed, identificatiesoftware moet in staat zijn om tenaamstellingen te interpreteren.

Intelligente interpretatie van persoonsnamen en bedrijfsnamen is een uitdaging van formaat. Hoe weten we dat *Harry Jansen Internationaal Vervoersbedr. BV* en *H. Jansen Int. Transp. Ond. BV* waarschijnlijk twee verschillende tenaamstellingen zijn voor hetzelfde bedrijf? Hoe stellen we vast dat *Robert* een voornaam is in *Robert van London* en een achternaam in *Robert & Van London NV*? Om bovenstaande vragen te beantwoorden, maakt een ieder gebruik van zijn of haar eigen kennis van taal en

cultuur en de bijbehorende regels en context. Voor correcte geautomatiseerde interpretatie ('wat is wat in een naam?') is het noodzakelijk deze natuurlijke taalverwerking na te bootsen. Hoe gaat zo iets nou in zijn werk?

WOORDENBOEK

Om te beginnen is het noodzakelijk een landspecifiek woordenboek samen te stellen, waarin alle relevante elementen die in namen kunnen voorkomen, worden vastgelegd. Daarbij is het zaak om niet alleen de betekenis van een bepaald woord mee te geven, maar tevens aandacht te besteden aan de volgende aspecten.

Een woord moet worden opgenomen in al zijn betekenissen. *Abraham* is een voornaam, maar ook een achternaam. *Albert Heijn* is een valide combinatie van voornaam en achternaam, maar toch vooral bekend als bedrijfsnaam van de supermarktketen.

Er moet rekening worden gehouden met de waarschijnlijkheid van een woord in een bepaalde betekenis.

Amsterdam is een achternaam, maar de waarschijnlijkheid dat dit woord als geografische aanduiding voorkomt, is vele malen groter.

De specifieke attributen van de naamselementen moeten worden opgenomen. Elk element heeft een combinatie van eigenschappen zoals afkorting, meervoud, het wel of niet voorkomen in een samenstelling, acroniem, et cetera. De relatie met andere kenniselementen moet worden beschreven. Het woord *ond* heeft geen zelfstandige betekenis, maar het kan voorkomen als afkorting van de woorden *onderwijs*, *onderzoek*, *onderhoud*, *onderneming* en *onderling(e)*.

De generieke en specifieke regels voor afbreking, interpunctie, afkorting en kapitalisatie moeten worden vastgesteld.

DUBBELZINNIGHEID

Wanneer het woordenboek eenmaal is gevuld, moeten we constateren dat we er nog niet zijn. Naast de dubbelzinnigheid van de woorden die we al hebben vastgelegd, hebben we bij namen namelijk ook te maken met de dubbelzinnigheid die we niet kunnen vastleggen. Elk woord of elke woordgroep kan immers een bedrijfsnaam of een onderdeel van een bedrijfsnaam zijn (zie figuur 1). Elke reeks letters kan een reeks voorletters of een fantasiewoord zijn. En dat valt niet in een woordenboek vast te leggen.

Hiervoor moet er rekening worden gehouden met zaken als open lettergrepen, uitgangen en uitspreekbaarheid van letterreeksen (de reeks *MRCP* is door zijn onuitspreek-

baarheid waarschijnlijk een reeks voorletters dan een fantasiewoord).

Het is dus als het ware de vorm van de woorden die eveneens een rol speelt bij de interpretatie.

CONTEXT

Uiteraard moeten we voor een het vaststellen van de juiste betekenis van een woord ook kijken naar de context waarin het woord staat.

J. Bakker, arts

J Arts, bakker

Bakker Arts & Crafts

Het woord *arts* heeft in het bovenstaande voorbeeld drie betekenissen, die door contextanalyse kunnen worden gededuceerd. Hiervoor is wel nodig dat het woord in al zijn betekenissen bekend is en dat er kennis is over de structuur van tenaamstellingen. Deze structuur is bij namen minder eenduidig dan bijvoorbeeld de syntactische regels in natuurlijke taal. Hoewel er aantal 'regels' zijn, die ook vaak worden gevolgd (een beroep komt na de familienaam en een rechtsvorm staat aan het begin of het eind van de tenaamstelling), zijn er geen strikte afspraken over de structuur van tenaamstellingen. *Jansen BV Internationale Betonbouw Tilburg* is een even valide naam als *Internationale Betonbouw Jansen Tilburg BV*.

GRAMMATICA

Om de structuur of het gebrek aan structuur te gebruiken voor contextanalyse, moet geautomatiseerde interpretatie in staat worden gesteld om associaties te maken. Deze associaties berusten op statistische gegevens. Door het inlezen van vele miljoenen tenaamstellingen ontstaat inzicht in specifieke woordcombinaties en patronen. Op deze manier zien we 'herkenbare' combinaties als *restaurant* en *jaargetijden*, maar ook verrassende zaken als *gemeenschappelijk* en *bezit*.

Deze associaties, tezamen met het woordenboek, kennis over woordvorm en contextanalyse, zijn de pijlers voor het samenstellen van een grammatica voor naamsinterpretatie.

Deze grammatica bepaalt vervolgens wat de meest waarschijnlijke betekenis van een woord in een bepaalde naam is. Dit gebeurt door toekenning van begripsscores, contextscores en scores van alle andere hierboven beschreven facetten van interpretatie.

Op deze manier kan er bijvoorbeeld onderscheid worden gemaakt tussen de persoonstenaamstelling *J.K.H. Jongman* en de bedrijfstenaamstelling *Q.R.Z. Amsterdam*.

NAAM	CLASSIFICATIE
Siemens	Familienaam
Jasmijn Bloemen & Planten	Voornaam
Restaurant Het Volle Leven	Naamwoordgroep
Croissanterie A la Gare	Voorzetselgroep
Solystic	Fantasienaam
Bakkerij Karel Aalbers	Persoonsnaam
Make me Happy Clothing	Imperatief als naam?
'Het' Humoristisch Elektronisch Tijdschrift	Lidwoord of acroniem?

Figuur 1. Elk woord of elke woordgroep kan een bedrijfsnaam zijn

INTERPRETATIE IN DE PRAKTIJK

Human Inference levert intelligente oplossingen voor verbetering en borging van datakwaliteit. De *in-house* ontwikkelde interpretatiemethodiek is de basis voor alle producten, die hiervoor worden ingezet.

Een regelmatig terugkerend probleem met betrekking tot datakwaliteit in bedrijfsprocessen, is de vervuiling van databases. Bij het invoeren van namen zorgen tikfouten, afkortingen, het invoeren van de handelsnaam in plaats van de formele naam, het wijzigen van de volgorde van de verschillende naamsdelen, associaties ten gevolge van namen die verkeerd zijn begrepen of verstaan, homoniemen en dergelijke voor een berg identificatie-ellende.

Onderstaand voorbeeld is illustratief voor de hieruit volgende verschillen in notatie van hetzelfde bedrijf.

Evertzen & Zoon Hoveniersond. v.o.f. Amsterdam Z.O.
VOF Tuinbouw Everse en Zn. A'dam

Traditionele identificatiemethoden, die geen gebruik maken van kennisgestuurde interpretatie, gaan bij het zoeken naar en het ontdebellen van dit soort tenaamstellingen de mist in.

Het moge duidelijk zijn dat de hier beschreven interpretatie, waarbij mathematische en statistische methoden worden verrijkt met landspecifieke kennis van het namenuniversum, nogal arbeidsintensief is. In onze multiculturele samenleving is de verzameling van voornamen en achternamen (Henk de Vries versus Samantha Hafizoglu) immers continu in beweging en ook in het bedrijfsleven speelt internationalisatie een steeds grotere rol (naast de traditionele transportonderneming kennen we tegenwoordig ook bedrijven voor im- en export logistics). Dat betekent, dat de kennis voortdurend moet worden bijgewerkt om de interpretatiemethodiek de mogelijkheid te geven op de juiste manier te 'argumenteren'. Alleen dan is er sprake van werkelijke kennis van relatiezaken.

Holger Wandt is principal advisor bij Human Inference

New

The first ever British English electronic spellchecker
with the unrivalled authority of Oxford,
fully integrated in Microsoft® Word

Oxford Spellchecker & Dictionary

Up to date and comprehensive: checks your spelling against more than 340,000 words from the *Oxford Dictionary of English*, including proper names, and technical and specialist terms

Fully integrates with Microsoft® Office 2000, XP, or 2003 — replacing the Microsoft® standard spellchecker

Dictionary link to the *Oxford Dictionary of English* allows you to check definitions, and to look up words independently

Definitive Oxford spellings: user settings can standardize alternative (including American) spellings to the preferred Oxford spelling, and warn about commonly confused words.

Wilt u Oxford Spellchecker & Dictionary bestellen of wilt u meer informatie?
Bezoek www.polderland.nl of neem contact met ons op via info@polderland.nl of 024-3522866.

OXFORD
UNIVERSITY PRESS

Language & Speech Technology

polderland

Genetische algoritmen

'Krant toch geen commissaris bij uitgever PCM' kopt NRC Handelsblad op 7 augustus 2004. Als je dat leest, sta je even raar te kijken. Verderop in het artikel blijkt dat het om ene heer Krant gaat, ex-directeur van zakenbank Kempen & Co. Hij was adviseur van Apax bij de overname van PCM en werd als beloning door de koper voorgedragen als commissaris. Naast het toucheren van een vorstelijke commissie van 10 miljoen euro dat hem uiteindelijk de das om deed. Maar dat terzijde.

Bij dit soort koppen denk ik meteen en bepaald niet onwillekeurig aan hoe een automatische vertaling hier de mist in zou gaan: 'Newspaper not member of the PCM Publishers's supervisory board', of iets dergelijks. Zeker een contextafhankelijke vertaling, want 'krant' heeft alles met uitgeverij en (dus) PCM te maken. Is dat erg? Als een mens ook op zijn verkeerde been gezet wordt, mag dat een machine ook overkomen. Of toch niet?

Vroeger hadden mensen slechts een naam – iets dat wat wij nu als voornaam kennen – met eventueel een toevoeging om verwarring te voorkomen: Jaap van Sjalff. Of een beroep – Bob de Bouwer – of een referentie naar waar men woonde, bijvoorbeeld Gerrit Groot Koerskamp. En als Gerrit verhuisde, werd zijn naam ook anders, zeg Gerrit van de Sloot. Of een nieuwe inwoner in een stad of dorp werd genoemd naar de plek, plaats of streek waar hij vandaan kwam, zoals Van Bergen Henegouwen. Het kwam dus voor dat een zoon een geheel andere naam kreeg of koos dan zijn vader had; vanwege zijn beroep, levensomstandigheden of gedrag. Napoleon voerde de burgerlijke stand in en verordonneerde dat iedereen een vaste familienaam moest aannemen. (Per decreet van 18 augustus 1811 kreeg men een jaar de tijd om zich in zijn of haar gemeente zich bij het Register van Naamsaanneming te laten inschrijven.)

Als bezetter werden Napoleon's voorschriften niet door iedereen even serieus genomen. Veel mensen vonden het maar onzin en zeiden maar wat bij de registratie: Van 't Zelfde, Onétogniet, Lachniet, enzovoort.

Onder de titel *Modern koken in Nederland* verschijnt een reeks kookboeken met streekgebonden gerechten. In het boek over de Veluwe staat kok Eric van Veluwen centraal. Hij werkt met lokale producten: vis uit de IJssel, vlees van boerderijen van de Veluwe, en natuurlijk de lokale moestuin. Gait L. Berk schreef de boeken *Ons bomenland* en *Zwerven door ons bomenland*. En *Vogelleven in Duin en Veld* is van de hand van dr. Charles E. Raven. Bob Flowerdew is een bekende Engelse tv-tuinier en boekenschrijver. En wat dacht u van Jonnie Boer? Hij is weliswaar sterrenkok, maar dweept met de lokale agrariërs die hem het Giethoorns lam leveren. Iets van een boer zit er dus zeker wel in.

Op de middelbare school heette mijn gymleraar Kuitenbrouwer. Toeval? Volgens mij wordt het hoog tijd voor een onderzoek om na te gaan of mensen met een bepaalde familienaam overwegend ook in die sector werkzaam zijn dan wel die belangstelling of affiniteit hebben. Het moet toch genetisch zijn bepaald. Als betovergrootvader Koppedraaijer heette, zal een groot deel van de nazaten vandaag de dag wel in een kippenslachterij werken. Of niet? En als dat in kaart is gebracht, kan met recht een genetisch algoritme worden toegevoegd aan de vertaalsoftware: dan blijft de Jos Brouwer, voortrekker van PINT (Vereniging Promotie INformatie Traditioneel Bier) en ex-hoofdredacteur van het bierblad met dezelfde naam, Brouwer en wordt het niet in *birraio* vertaald.

Als betovergrootvader Koppedraaijer heette, zal een groot deel van de nazaten vandaag de dag wel in een kippenslachterij werken.

Norbert Mergen is één van de oprichters van Human Inference. Sinds kort is hij teruggetreden als directeur-eigenaar.

Dialogspecificaties bepalen het succes van spraakherkenning

Spraakherkenning bij PCM

In 2003 en 2004 heeft TeleCats voor een aantal uitgevers van landelijke dagbladen (*De Telegraaf* en *PCM*) een service ontwikkeld waarmee incidentele klachten automatisch worden afgehandeld. Bellers worden geïdentificeerd via hun ingesproken postcode waarna het systeem controleert of op deze postcode de afgelopen tijd niet al eerder een klacht is binnengekomen. Is dit niet het geval, dan handelt de computer de klacht af. Blijkt er binnen een kort tijdsbestek vaker een klacht te zijn geweest, dan wordt de beller doorgeschakeld met een agent van de uitgever.

| WIM LUIMES |

Uitgevers van kranten zijn zeer bezorgd over het aantal abonnees en willen geen klanten voor het hoofd stoten door gebruik te maken van anonieme, kille systemen. Toch zijn ook zij er van overtuigd dat alles met agents afhandelen op den duur geen haalbare kaart meer is. Daarom is voorzichtig begonnen met de introductie van een automatische 'abonneedienst' waarbij erg veel aandacht is besteed aan de wijze waarop klanten worden behandeld. PCM startte begin 2004 met deze wijze van klachtenafhandeling voor één titel, te weten het *Algemeen Dagblad*. Doordat de meeste mensen tevreden waren, werden ook de bezorgklachten van de overige uitgaven van

PCM via de door TeleCats gebouwde applicatie afgehandeld. Het dagblad *De Telegraaf* heeft ook deze applicatie inmiddels ook draaien. Het feit dat nu ongeveer alle grote Nederlandse dagbladen deze TeleCats-applicatie gaan gebruiken, was aanleiding voor een interview met Renate van der Toorn – manager van de *call centers* van de landelijke dagbladen van PCM – over dit soort applicaties in de gevoelige wereld van de dagbladuitgevers.

Hoe bevalt de huidige spraakgestuurde applicatie?

Erg goed! De eerste maanden trad er een aantal te verwachten kinderziekten op, maar die werden snel hersteld en op dit moment functioneert het systeem naar behoren. Wat duidelijk werd, is dat als het fout gaat, dit niet zozeer aan de herkenning capaciteit van het systeem ligt, als wel aan het feit dat bellers onverwachte dingen zeggen. Zo hadden we laatst een van onze eigen redacteurs die er met spraakherkenning niet uit kwam omdat hij een klacht voor twee dagen wilde doorgeven. Zijn begrijpelijke, doch *onterechte* reactie was: spraakherkenning werkt niet!

Hiermee komen we gelijk aan de kern van veel problemen met geautomatiseerde systemen: het gebrek aan flexibiliteit. Telefoonapplicaties met spraakherkenning zijn veel flexibeler dan die met IVR, maar voor beiden geldt dat wat er niet in zit, er ook niet uit kan komen. We hebben het systeem ingericht voor incidentele bezorgklachten, waarbij de klant een verzoek tot nabezorging doet. Uiteraard weten we dat er ook andersoortige klachten zijn over de bezorging, bijvoorbeeld dat de krant steeds te laat komt. Het gaat hierbij om enkele gevallen die inderdaad niet via spraakherkenning afgehandeld kunnen worden. Door in ons keuzemenu een extra toelichting te geven, zorgen we er nu voor dat echt alleen de verzoeken voor nabezorging in de spraakherkenner terechtkomen. Zo blijkt maar weer dat het opstellen

Renate van der Toorn en spraakherkenning

Renate van der Toorn was als manager van de *call centers* van de landelijke dagbladen van PCM verantwoordelijk voor de introductie van spraakherkenning bij de afhandeling van bezorgklachten. Hiervóór heeft zij onder andere gewerkt bij Openbaar Vervoer Reisorganisatie (OVR), waar spraakherkenning al lang wordt gebruikt voor het opvragen van informatie over treintijden. Renate van der Toorn mag daarom gezien worden als één van de weinige managers in Nederland met serieuze praktijkervaring met spraakherkenning.

van de dialoogspecificaties erg belangrijk is. Bovendien weet je dat de dialogen nooit perfect zullen zijn. De vraag is steeds: hoeveel flexibiliteit ten koste van welke prijs ga je in het systeem stoppen. Hiermee bedoel ik niet zozeer de prijs in euro's – hoewel dat ook belangrijk is – maar vooral de mate waarin het systeem nog gebruiksvriendelijk en robuust blijft wanneer meer en meer mogelijk wordt. Het heeft niet zo heel veel zin om erg veel tijd te besteden aan vragen die slechts sporadisch voorkomen; voor dat soort zaken blijft de mens toch de ideale, zij het duurdere, oplossing.

Wat waren de verwachtingen van jouw collega's over spraakherkenning bij de krant?
Het merendeel van de medewerkers was sceptisch maar op directieniveau zag men wel de voordelen van spraakherkenning. Kranten hebben het, mede door de opkomst van allerlei dikwijls gratis nieuwsbronnen zoals Internet en *Newswire*, toch al moeilijk en het idee abonnees eventueel af te stoten door een spraakgestuurde klachtenafhandeling stuitte velen tegen de borst. Dat zie ik trouwens duidelijk als een compliment, omdat het aangeeft hoe betrokken onze medewerkers bij het reilen en zeilen van de kranten zijn. Doordat de applicatie goed werkt en aan de verwachtingen voldoet, is de scepsis afgenomen; zeker omdat er altijd de terugvalmogelijkheid naar medewerkers bestaat.

Hebben jullie de abonnees gevraagd wat ze er van vonden?

Ja, maar niet voordat we de dienst introduceerden. Nadat alles in onze ogen naar wens verliep, hebben we wel een redelijk uitgebreid gebruikersonderzoek gedaan waarbij de bellers werden teruggebeld. Hieruit bleek dat 51% geen enkel bezwaar had tegen het gebruik van spraakherkenning, 26% had toch liever een mens en de rest had zich in het geheel niet gerealiseerd dat ze een spraakherkenner aan de lijn hadden gehad. Al met al een prima resultaat. Wat opviel is dat de klanten met name niet blij waren met spraakherkenning als de belofte om na te bezorgen niet werd nagekomen. Klanten associëren in zo'n geval de niet geleverde service direct met het *systeem* dat nu gebruik maakt van



Renate van der Toorn:

"Sommige abonnees realiseren zich niet eens dat ze een spraakherkenner aan de lijn hebben gehad."

spraakherkenning. Maar het probleem staat natuurlijk geheel los van spraakherkenning!

Past spraakherkenning bij het beeld van de kranten?

Ja en nee. Zoals je weet, voeren we verschillende titels met duidelijk verschillende soorten lezers. We zijn begonnen met het uitrollen voor het *AD* en ondertussen

Wat er niet inzit, komt er ook niet uit. Dat geldt voor elk geautomatiseerd systeem

zijn daar alle andere titels van PCM bijgekomen. Het is zonneklaar dat, uitzonderingen daar gelaten, men altijd het liefst een mens aan de telefoon wil krijgen. Helaas is men steeds minder bereid om daar ook voor te betalen waardoor we wel gedwongen worden om delen van de service te automatiseren. Spraakherkenning is dan de meest vriendelijke manier. Ik denk dat met het verstrijken van de tijd men, mits de service goed blijft, wel aan het idee zal wennen: op een gegeven moment weet men gewoon niet beter. Je ziet dat bij veel nieuwe technologie zoals geldautomaten, SMS, IVR, e-mail en Internet. Een deel

U spreekt met de krant

De eerste spraakherkenningapplicatie, waarbij abonnees van het *Algemeen Dagblad* (AD) incidentele klachten over niet bezorgde kranten kunnen melden, is in februari 2004 gerealiseerd. Het spraakherkenningssysteem vraagt de postcode en het huisnummer van de beller. Op basis van deze gegevens wordt het adres van de beller bepaald.

Omdat het systeem voldeed aan de verwachting en er zich geen serieuze problemen hebben voorgedaan, werd afgelopen juni gestart met het toepassen van spraakherkenning bij de overige PCM-titels te weten *NRC Handelsblad*, *De Volkskrant*, *Trouw*, *Rotterdams Dagblad*, *Rijn en Gouw* en *De Dordtenaar*.

Wim Luimes is directeur van
TeleCats

vindt het direct geweldig, een deel vindt het niets maar na verloop van tijd raakt men er aan gewend en weet men eenvoudig niet meer waarom men er in eerste

instantie bezwaar tegen had. Deze gewenning geldt echter alléén wanneer de nieuwe technologie een duidelijke toegevoegde waarde heeft. Nieuwe technologie inzetten alleen omdat het werkt, heeft volgens mij geen enkele zin.

Wat zou je nog meer willen doen met spraakherkenning?

Een van de zaken die ik wel zie zitten, is het verder verhogen van de service door nog meer gebruik te gaan maken van taal- en spraaktechnologie. Wanneer we eenmaal zover zijn dat we de spraak niet alleen kunnen herkennen maar ook, net als e-mails en web-forms, kunnen interpreteren, dan kan het serviceniveau van de krant nog een heel eind omhoog zonder dat het onbetaalbaar is.

Team voor Taal 2005 !

Het congres Team voor Taal is het jaarlijks terugkerende Nederlandstalige ontmoetingsplatform op het gebied van taal- en spraaktechnologie (TST) voor investeerders, managers, bestuurders en wetenschappers. De organisatoren van Team voor Taal streven ernaar om de mensen en organisaties die reeds TST toepassen of van plan zijn deze technologieën te gaan toepassen, bij elkaar te brengen waardoor nieuwe kansen worden gecreëerd.

De succesvolle bijeenkomst van Team voor Taal op 21 januari van dit jaar krijgt een prominent vervolg op donderdag 2 juni 2005. De focus van deze bijeenkomst zal liggen op de toepassing van de combinatie van TST.

De combinatie van taal- en spraaktechnologie

Taal- en spraaktechnologie werden afzonderlijk al een aantal jaren gebruikt, maar het wordt steeds duidelijker dat het gecombineerd gebruiken van TST grote voordelen biedt. Klanten en bedrijven hebben immers vaak contact via meerdere media (telefoon, Internet, e-mail). Toepassingen van deze technologieën kunnen ondersteunend zijn bij de identificatie en segmentatie van de klant, de classificatie van het 'gespreksonderwerp' en bij de juiste verwijzing of afronding van de dialoog. Door het (semi-)automatisch afhandelen van de informatie-uitwisseling via de verschillende media met behulp van taal- en spraaktechnologie is er veel winst te behalen op het gebied van effectiviteit en efficiency.

Tijdens Team voor Taal 2005 kan, onder andere via de introductiemarkt, kennis worden gemaakt met een aantal instellingen en bedrijven die al succesvol TST toepassen. Recente succesvolle toepassingen van TST worden tijdens het congres 's ochtends gepresenteerd, geanalyseerd en bediscussieerd. Tijdens de parallelsessies in de middag zal worden ingegaan op de kansen van TST en op de laatste ontwikkelingen op het gebied van het (commercieel) toepassen van deze technologieën.

Team voor Taal: hét ontmoetingsplatform voor de branche

Team voor Taal wordt hét netwerk-event om kennis te maken met de spelers in de markt en de wetenschap. Reserveer nu alvast 2 juni 2005 in uw agenda. In de volgende uitgave van DIXIT zal het definitieve programma worden gepubliceerd en tegelijkertijd zullen persoonlijke uitnodigingen worden verzonden.

Wilt u ook op de lijst van genodigden komen of meer informatie wensen, stuur dan een e-mail bericht naar m.spitholt@telecats.nl en vermeld daarin uw naam, adresgegevens en motivatie.

Uitgelicht!

Spraakherkenning succes voor Van der Valk

AudioCom heeft een reserveringslijn op basis van spraakherkenning voor Van der Valk ontwikkeld. De afgelopen maanden belden al ruim 4000 mensen naar deze reserveringslijn. Op jaarbasis verwacht men 16.000 bellers.

Klanten die in één van de vestigingen van Van der Valk willen overnachten, vergaderen of dineren, kunnen bellen naar het centrale reserveringsnummer, 0900-123VALK, welke uitkomt op het *voice response*-platform van AudioCom. Bellers worden welkom geheten en kunnen vervolgens via de telefoontoetsen kiezen of ze een kamer, een vergaderzaal of een tafel willen reserveren. Wanneer de beller een keuze heeft gemaakt, wordt gevraagd om de naam van de stad in te spreken waar men wil overnachten, vergaderen of dineren. Vervolgens worden de

dichtstbijzijnde Van der Valk-vestigingen genoemd. Om doorverbonden te worden met de vestiging van zijn keuze kan de beller weer gebruik maken van de telefoontoetsen.

Bij het ontwikkelen van de reserveringslijn is klantvriendelijkheid en gebruiksgemak het uitgangspunt geweest. Hieruit voortvloeiend is gekozen voor de combinatie 'kiezen met de telefoontoetsen' en 'kiezen met behulp van spraakherkenning'. De beller kiest met de telefoontoetsen op het moment dat de keuzemogelijkheden beperkt zijn. Als het aantal keuzemogelijkheden te groot wordt, kiest de beller door middel van spraakherkenning.

Overigens krijgen bellers altijd de mogelijkheid om doorverbonden te worden met het *call center* van Van der Valk.

Taaltechnologie als wapen tegen terrorisme

Human Inference heeft in nauwe samenwerking met ABZ – onderdeel van het Assurantie Data Netwerk – een softwareprogramma ontwikkeld voor het opsporen van geldstromen naar mogelijke terroristische organisaties. Met de oplossing kunnen financiële dienstverleners hun administratie controleren op de aanwezigheid van mogelijke terroristen.

Sinds 9/11 en de bomexplosies in Madrid maart van dit jaar, zijn veiligheidsdiensten volop actief om aanslagen in de kiem te smoren. Eén van de maatregelen die het werk van terroristen moet bemoeilijken, is het opsporen en afsluiten van geldstromen richting terroristische organisaties en het bevriezen van hun tegoeden. In Nederland hebben banken en verzekeraars de onderzoeksplicht om hun administratie door te lichten op de aanwezigheid van mogelijke terroristen. Hiertoe heeft Human Inference samen met ABZ software ontwikkeld, die dit najaar op de markt komt.

De software maakt het mogelijk op volledig geautomatiseerde wijze de

Europese databank te vergelijken met het bestand van personen of entiteiten (bedrijven, organisaties) van de desbetreffende financiële instelling. De software analyseert namen en eventuele aliases van de Europese databank met de aangeleverde persoons- of bedrijvenadministratie. Alle naamselementen en eventuele aanvullende gegevens zoals geboortedatum, geboortejaar of -plaats en land van herkomst, worden met elkaar vergeleken. Financiële dienstverleners zijn verplicht mogelijke treffers, volgens een vastgesteld formaat, door te geven aan de Pensioen- & Verzekeringskamer (PVK). Naast ondersteuning bij deze meldingsplicht, waarborgt de software de integriteit van de administratie.

“Met deze applicatie hopen wij een (kleine) bijdrage te kunnen leveren aan de strijd tegen het terrorisme. De periodieke controle van de databanken kan worden uitbesteed aan ABZ. Verzekeringsmaatschappijen en banken kunnen de software ook zelf implementeren” aldus Gert Dazler, Business Development Director ABZ.

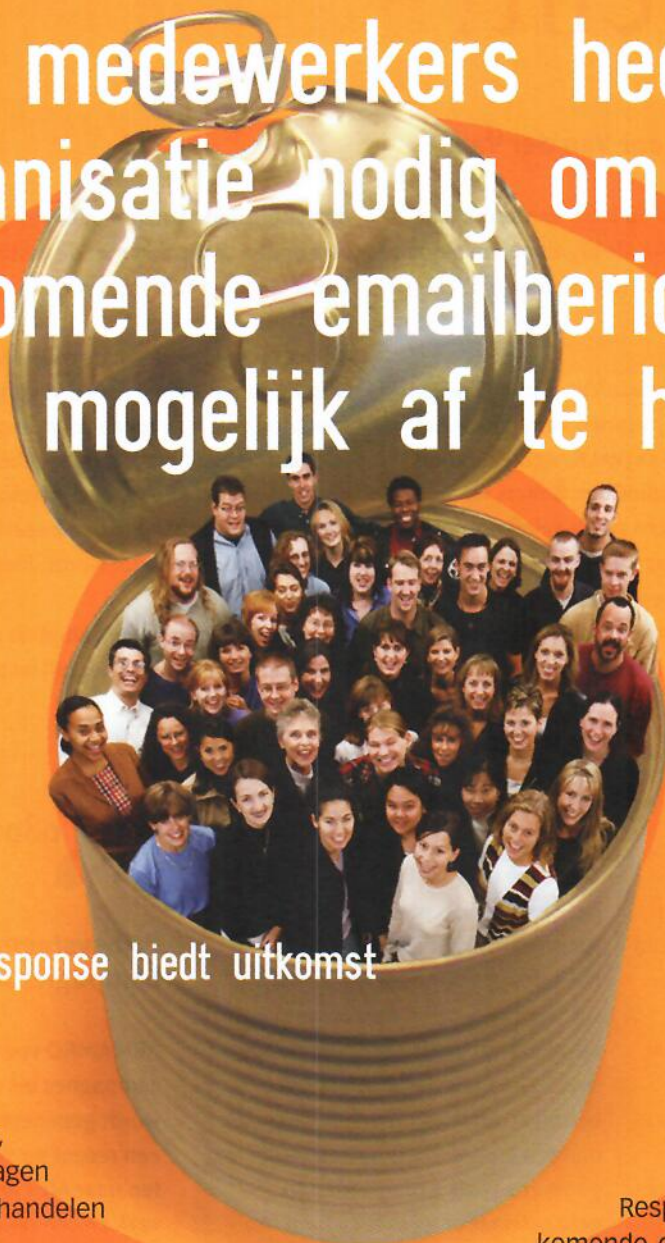


IVR-systeem ondersteunt campagne ABN AMRO

ABN AMRO voert regelmatig marketingcampagnes uit waarbij veel telefoonverkeer wordt gegenereerd. Naar aanleiding van een recent gevoerde campagne bellen klanten naar ABN AMRO om aan te geven hoe vaak ze een rekeningafschrift willen ontvangen. De afhandeling van dit telefoonverkeer wordt verzorgd door een selfservice IVR-systeem van TeleCats, dat in samenwerking met BSC tot stand gekomen is.

Vanaf 1 juni 2004 tot en met medio oktober kunnen ABN AMRO klanten die de ontvangstfrequentie van rekeningafschriften willen aanpassen, bellen met het gratis nummer 0800 - 024 07 11. De telefoongesprekken komen binnen op een IVR-systeem van TeleCats, waar ze volledig automatisch worden afgehandeld. Dit systeem verwerkte alleen al in de eerste maand 400.000 gesprekken. Dagelijks worden callstatistieken aangeleverd aan medewerkers van ABN AMRO, zodat zij het verloop van de actie goed kunnen overzien en kunnen bepalen welke vervolgacties moeten worden gepland.

Hoeveel medewerkers heeft uw organisatie nodig om binnenkomende emailberichten zo efficiënt mogelijk af te handelen?



Smart E-mail Response biedt uitkomst

- Totale dekking en controle over het gehele E-mail proces, van binnenkomende vragen van klanten tot het afhandelen ervan.
- Taalkundige analyse van de binnenkomende berichten, om de berichten automatisch te prioriseren en routeren naar medewerkers met de juiste skills.
- Snelle accurate categorisering en het begrijpen van het bericht door het gebruik van state of the art technieken op het gebied van kunstmatige intelligentie (I.A.) en taaltechnologie.
- Minimaal en geautomatiseerd onderhoud door gebruik van geavanceerde leer algoritmen.

Smart E-mail Response van AskNow Solutions biedt uitkomst om dit proces professioneel te beheersen. Smart E-mail Response is in staat om binnenkomende elektronische berichten auto-

matisch te classificeren, te routeren, te prioriseren en waar mogelijk automatisch te beantwoorden. AskNow's Smart E-mail Response is gestoeld op een zeer robuuste state-of-the-art-techniek uit de Taaltechnologie en Kunstmatige Intelligentie. De combinatie hiervan maken dat Smart E-mail Response slechts éénmalig 'getraind' hoeft te worden voor de uit te voeren taken waarna verder onderhoud tot een minimum beperkt blijft.

AskNow is een Nederlandse organisatie die een totaaloplossing biedt op het gebied van klantcontact management en in het bijzonder Selfservice, E-mail management, Agent Support en CRM+.

Kostenbesparend, servicegericht en gemakkelijk



SOLUTIONS www.asknow.nl

Tool helpt medewerkers schrijven in één bedrijfsstijl

Organisatie, let op uw woorden!

Organisaties zijn zich steeds bewuster van hun imago. Ze doen er alles aan om dat imago positief 'uit te stralen'. Een nieuw terrein dat daarbij aandacht krijgt, is de geschreven taal van de organisatie. Taaltechnologie helpt een handje. Ruim baan voor de 'elektronische eindredacteur'.

Dat merken geld waard zijn, is geen nieuws. Het merk Coca Cola staat bijvoorbeeld voor vele miljarden op de balans van de gelijknamige *company*. En dat merken staan of vallen bij hun imago, daar kan iedereen zich wel wat bij voorstellen. KLM is betrouwbaar, Rabobank staat midden in de samenleving en Philips *makes things better*. Aan zulke merkimago's wordt zorgvuldig gebouwd, door middel van (marketing)communicatie. Een geheel nieuw aandachtsterrein daarbij vormen schriftelijke uitingen als correspondentie, offertes, documentatie, nota's en plannen van aanpak. Zulke uitingen worden door individuele medewerkers geschreven en staan veelal nauwelijks of niet onder 'toezicht' van de afdeling marketing, communicatie of PR, die het merk 'beheert'. En daarin schuilt een afbreukrisico. Het gevaar is dat de luchtvaartmaatschappij in een externe nota helemaal niet meer zo betrouwbaar overkomt, dat de correspondentie van die bank ineens heel afstandelijk is en dat de elektronicagigant in zijn productdocumentatie het er helemaal niet beter op maakt.

GEVOEL

Neem nou die Zweedse autofabrikant: fraaie modellen, *glossy* reclame-uitingen. Van website tot foldermateriaal: alles is grondig doordacht en tot in de puntjes verzorgd. Hoewel? De correspondentie

van een lokale dealer geeft een ander beeld. Bijvoorbeeld het eenregelige briefje waarmee een groene kaart wordt opgestuurd: *'Mijne dames, heren, bijgaand treft u de bovengenoemde groene kaart. Hoogachtend, de dealer.'* Einde brief. Dat is niet alleen onduidelijk (wat moet er met de kaart gebeuren?), de toon past niet bepaald bij het servicegerichte gevoel dat dit merk wil oproepen. Nee, dan de eerdergenoemde bank. Die begreep dat de gewenste merkbeleving niet gediend is met onvriendelijke en onduidelijke correspondentie (*'Wij verzoeken u per omgaande zorg te dragen voor ondertekening van de betreffende bescheiden.'*). Daarom liet de bank alle standaardcorrespondentie die via lokale banken naar cliënten wordt verstuurd herschrijven (*'Wilt u de documenten ondertekenen?'*). En daarom kregen alle medewerkers met klantencontacten ondersteuning in het leren corresponderen conform de huisstijl.

| ROBBERT JAN SABEL |

INVLOED

Geschreven taal is een van de krachtigste uitingvormen van bedrijfscommunicatie. Taal speelt een zeer belangrijke rol in het ordenen van ervaringen. De manier waarop we onze ervaringen verwoorden heeft alles te maken met de wijze waarop we informatie opnemen en verwerken. Dat hangt soms af van kleine woordjes. De manager die zijn collega's wil aanmoedi-

Hoera! Een bijzonder moment...



...al 5.000.000 vragen beantwoord!

En daarmee zijn we marktleider in de financiële sector. Wilt u ook weten wat uw klanten van u willen weten? Wilt u uw website gebruiken om uw klantenservice verder te verbeteren? Of om uw klanten te behouden en meer aan hen te verkopen? En daarbij toch de kosten van het klantcontact snel omlaag brengen?

Q-go garandeert u drie resultaten: een ROI van 6 maanden, een volledig inzicht in de wensen van uw klanten en een zeer korte implementatietijd.

De Online Marketing & Self Service applicaties van Q-go zijn gebaseerd op een uniek natuurlijk taalplatform. Uw klant kan intuïtief en in eigen woorden zijn vraag stellen op uw website. Q-go zorgt ervoor dat

iedere vraag wordt begrepen en uw klant automatisch razendsnel het beste antwoord krijgt.

Tal van grote, maar ook kleinere organisaties binnen en buiten Nederland zijn u voorgegaan. Wilt u ook binnen korte tijd uw website gebruiken om kosten te besparen, service te verbeteren en meer te verkopen?

Meld u aan voor één van de workshops via www.q-go.com of bij Astrid Smedema astrid.smedema@q-go.com. Bellen kan uiteraard ook: 020 531 3800 en vragen naar Susanne.



Online Marketing & Self Service

gen met 'we moeten er tegenaan', bouwt onbewust een blokkade in. 'Moeten' is een negatief woord, dat energie ontnemt.

Waarom niet 'we gaan er tegenaan'? En wat te denken van al die overbodige woorden die ondernemingen gebruiken? Alsof er al niet genoeg woorden op ons afkomen. ~~Veel zinnen kunnen over het algemeen aanzienlijk korter worden opgeschreven.~~ Hoe *beknopter* geformuleerd, hoe groter de *impact*. Minder is meer. Dat geldt ook voor jargon. Zeer geliefd bij vakspecialisten die er graag status aan ontlelen. Toch toont onderzoek aan dat ook onder vakgenoten klare taal de voorkeur verdient. Met andere woorden: in 'gewone mensentaal' komt de boodschap beter over en neemt het aanzien van de zender juist toe. Dat is niet alleen van belang bij commerciële merken, maar ook bij overheidsorganisaties die steeds bewuster met hun imago omgaan. Of het nu gaat om een ministerie, provincie, waterschap of gemeente, de kerntaak is het ontwikkelen van beleid en het uitdragen en naleven daarvan. Dat zijn allemaal taken waarin het geschreven woord 'basismateriaal' is. Dan luistert het zeer nauw hoe die woorden worden gebruikt.

KRACHT VAN TAAL

Taaltechnologie kan helpen bij het benutten van de kracht van taal. In samenwerking met Polderland ontwikkelde Sabel Communicatie een tool die iedereen in een organisatie ondersteunt bij het schrijven volgens één afgesproken stijl: de 'teksthuisstijl' van het bedrijf. Deze teksthuisstijl is een vertaling van de communicatiedoelstellingen, het imago en de gewenste merkbeleving van het bedrijf in de tekstkenmerken woordgebruik (geen *geachte* maar *beste*), stijl (niet '*De provincie zorgt voor de uitvoering van het project*' maar '*De provincie voert het project uit*') en spelling (geen *Rabo bank* maar *Rabobank*). De tool analyseert tekstdocumenten (in Word-, txt- of html-formaat) op de genoemde punten en beoordeelt automatisch of het woordgebruik en de toon van de tekst passen bij de stijl van de organisatie. Daarnaast helpt de tool om toegankelijk, communicatief Nederlands te schrijven, bijvoorbeeld door omslachtige (passieve) constructies en omhaal van woorden te voorkomen. Evenals het

gebruik van jargon, holle kreten en archaïsche formuleringen.

'ELEKTRONISCHE EINDREDACTEUR'

Technisch gezien bestaat de tool uit twee delen: een statistisch gedeelte en het zogenoemde administrator's-gedeelte. In dat laatste deel staan alle afspraken over de teksthuisstijl van het betreffende bedrijf. Eén persoon – meestal de beheerder van de tool – kan deze gegevens aanpassen, bijvoorbeeld wanneer het bedrijf een begrip wil toevoegen aan de lijst met 'verboden' jargon. Het statistische gedeelte van de tool analyseert de tekst op het voorkomen van bepaalde woorden of woordcombinaties. Wanneer de tool bijvoorbeeld aangeeft dat in een tekst van zeshonderd woorden twintig keer *worden* in combinatie met een voltooid deelwoord voorkomt, dan kan dat erop wijzen dat de auteur zinnen te vaak in de lijdende vorm zet. En als een organisatie jong en dyna-

'Moeten' is een negatief woord, dat energie ontnemt. Waarom niet 'we gaan er tegenaan'?

misch wil overkomen, zal de tool een schrijver erop wijzen als deze woorden als *thans* en *weldra* gebruikt. Voor zulk taalgebruik draagt de tool alternatieven aan. Daarmee is het een 'elektronische eindredacteur', die op een vriendelijke wijze iedere gebruiker van dienst is. Van secretaresse tot directeur en van beleidsmedewerker tot administrator. Iedereen die zich namens de organisatie schriftelijk uit, kan de tool benutten. Op dit moment wordt met de tool proefgedraaid bij organisaties waar taal een sleutelrol speelt. De ervaringen zijn positief en de verwachtingen hooggespannen. Wie weet gloort voor deze nieuwe toepassing van taaltechnologie een gouden toekomst.

Robbert Jan Sabel is directeur van Sabel Communicatie in Utrecht

Uit de boeken



L E E S W I J Z E R

De praktijk is altijd een goede leermeester, maar zonder gedegen achtergrondkennis – trends, ontwikkelingen, toegepast onderzoek, noem maar op – kan diezelfde praktijk al snel een zeer beperkte leermeester worden. Het praktische blikveld heeft al gauw de neiging zich te vernauwen tot de waan van de dag. Vandaar dat Antal van den Bosch – verbonden aan de Sectie Taal en Informatica van de Letterenfakulteit in Tilburg – telkens een aantal relevante boeken of artikelen kort zal bespreken. Ditmaal een recensie van het artikel 'Finding predominant senses in untagged text' van Diane McCarthy, Rob Koeling, Julie Weeds, en John Carroll, dat tijdens de jaarlijkse conferentie van de Association for Computational Linguistics werd bekroond.

Nog niet zo lang geleden hadden vaktijdschriften, vakboeken en conferentiebundels dezelfde graad van toegankelijkheid: ze stonden allemaal in kasten van universiteitsbibliotheken. Tijdschriften en conferentiebundels mochten vaak niet eens de deur uit. De geurmix van warme inkt en zweet in het kopieerhokje van de bibliotheek waar ik het als student van moest hebben staat me nog helder voor de geest. Er moest fysiek geleden worden voor de trofee, een pakje kopieën van artikelen waarvan ik mocht hopen dat ze het antwoord boden op mijn vragen van dat moment, die voortkwamen uit weer andere artikelen. Een soort heel traag internet met papieren hyperlinks en trage zoekmachines op menskracht.

Vijftien jaar later onttrekken alleen boeken zich aan elektronische toegang. Boeken moet je nog steeds lenen of kopen, en zijn doorgaans niet *on-line*. Artikelen uit tijdschriften en conferentiebundels zijn dat wel, hoewel vaak door de burgerlijke ongehoorzaamheid van de auteurs die hun eigen werk op hun eigen webpagina beschikbaar stellen naast de gedrukte versie. Artikelen zijn in het dagelijks werk de voornaamste bron van informatie, vooral van recente onderzoeksresultaten, want die doen er even over om in boeken terecht te komen.

Vandaar in deze rubriek de bespreking van een recent artikel. Ik leun op de mening van een groep competente collega's door de keuze te laten vallen op het artikel 'Finding predominant senses in untagged text' van Diane McCarthy, Rob Koeling, Julie Weeds, en John Carroll van de University of Sussex (Brighton). Dit artikel werd verkozen tot de beste van ACL-2004, de jaarlijkse conferentie van de Association for Computational Linguistics, die

in juli in Barcelona plaatsvond. Wat maakt dit paper tot het beste werk van dat moment, gegeven de ruime keuze die de programmacommissie had?

Taaltechnologische halfproducten

Het onderwerp van het paper is 'word sense disambiguation' (WSD): het bepalen van de betekenis van een woord in context. Dit is, net als *part-of-speech tagging*, zo'n typisch taaltechnologisch halfproduct dat nuttig zou kunnen zijn in grotere systemen – *information retrieval*, automatisch vertalen. Om 'blik' te vertalen naar het Engels is het nodig om te bepalen of het om een 'glance' gaat of een 'can'. WSD is met name bekend geworden door de SENSEVAL-competities, waarin teams strijden om de beste scores op verschillende WSD-taken in verschillende talen. De eerste auteur van het paper, Diane McCarthy, is een bekende deelnemster aan SENSEVAL. Een flink aantal gepresenteerde papers in ACL-2004 gaat over WSD, en tijdens de ACL werd de derde SENSEVAL-workshop gehouden. Op grond van discussies tijdens die laatste workshop zou je de enigszins negatieve conclusie kunnen trekken dat WSD in een lastig parket is gekomen, omdat langzamerhand is gebleken hoe ongelofelijk moeilijk de taak is en hoe moeilijk het nut van WSD in grotere taken is aan te tonen.

Betekenis zonder het woord

Een van de tekenen dat WSD zo moeilijk is, is dat de beste systemen weinig beter scoren dan een simpel systeem dat van alle woorden de meest voorkomende betekenis gokt. De meest voorkomende betekenis van een woord is eenvoudig te meten als er data beschikbaar is waarin betekenissen zijn geannoteerd door mensenhanden. Maar dit soort data is uiterst schaars. Het artikel stelt een methode voor om zonder

vooraf beschikbare tellingen en annotaties, op grond van ruwe ongeannoteerde teksten, van willekeurige woorden hun meest voorkomende betekenis vast te stellen in die teksten, zodat in ieder geval die 'baseline' WSD-score gehaald kan worden. Het maakt daarbij gebruik van een bestaande techniek (van Dekang Lin) om automatisch een thesaurus af te leiden uit ruwe tekst. Gegeven een woord kan deze techniek een groep 'naaste buur'-woorden vinden die veel met het woord te maken hebben (synoniemen, maar ook in mindere mate gerelateerde woorden). Deze naaste buren worden opgezocht in WordNet, een alom bekend en veel gebruikt betekeniswoordenboek, en al hun mogelijke betekenissen worden verzameld. Met een gewogen methode (die gebruik maakt van het WordNet:: Similarity Package) worden alle betekenissen op een rij gezet, en de hoogst scorende betekenis wordt door het systeem teruggegeven als de waarschijnlijke meest voorkomende betekenis van het woord in kwestie. Let wel – het woord zelf is niet opgezocht.

Tie: neckwear of toch draw

De onderzoekers laten zien dat hun methode beter werkt dan een aantal andere methoden die ook geen gebruik te maken van geannoteerde WSD-data ('unsupervised' methoden) – het systeem haalt het overigens niet bij 'supervised' systemen die op grond van voorbeelden met *machine-learning* en statistische methoden de juiste betekenis gokken. Al met al is het systeem

een redelijke deelnemer aan de 'unsupervised' competitie binnen SENSEVAL en is het een fraai voorbeeld van een combinatie van bestaande technieken. Wat het – volgens mij – boven het niveau van andere papers uittilt, is dat de onderzoekers ook buiten de gebaande SENSEVAL-paden treden en laten zien dat hun aanpak precies zo werkt zoals je het hoopt, en dat bij verschillende teksten. Het is een bekend feit dat een woord binnen een bepaald domein (wetsteksten, jaarverslagen, notulen, kinderboeken, sportberichten enzovoort) een hele andere verdeling aan betekenissen kan hebben, en een hele andere meest voorkomende betekenis, dan in een 'gemiddelde' tekst. Het systeem komt inderdaad met verschillende meest voorkomende betekenissen als het domein verandert van 'gemiddeld' naar 'sport' of 'financieel'. Het woord 'tie' is 'neckwear' in gemiddelde tekst, maar 'draw' in teksten over sport. Het woord 'head' is een 'body part' in gemiddelde tekst, maar een 'leader' in teksten over financiële onderwerpen. Een 'goal' is een 'end' in algemene teksten, maar een 'score' in sportteksten.

De toepassingstaal is Engels, maar het zou voor de Nederlandse context goed mogelijk moeten zijn om de aanpak te kopiëren, als je gebruik maakt van het Nederlandse deel van EuroWordNet, een meertalige Europese versie van WordNet. Een leuk afstudeerproject lijkt me.

NOTaSnieuws

NOTaS: de zomermaanden voorbij

Ondanks de vakantieperiode, is het voor NOTaS een drukke tijd geweest. De lancering van het BATAVO-programma op 15 september speelde hierbij de hoofdrol, maar ook op het gebied van de TST-centrale hebben er allerlei ontwikkelingen plaats gevonden.

1. STEVIN van start!

Eindelijk is het dan zover: op 15 september is STEVIN gestart. STEVIN staat voor: Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands. STEVIN is een meerjarig onderzoeks- en stimuleringsprogramma voor Nederlandstalige taal- en spraaktechnologie dat gezamenlijk door de Vlaamse en Nederlandse overheid wordt gefinancierd.

De bedoeling van STEVIN is het stimuleren van de taal- en spraaktechnologische sector in Vlaanderen en

Nederland, zodat de innovatiecapaciteit van deze sector kan worden vergroot, en tegelijkertijd ook de positie van het Nederlands in de moderne informatie- en communicatiewereld kan worden gevrijwaard en verder versterkt. Het meerjarenprogramma heeft een financiële omvang van 11,4 miljoen euro en heeft een grensoverschrijdend karakter. De financiële middelen worden onder verantwoordelijkheid van de Nederlandse Taalunie uitgezet. De eerste oproep met een omvang van maximaal twee miljoen euro zal eind september 2004 worden gepubliceerd en is gericht op kortlopende specifiek gerichte (onderzoeks)projecten. Begin 2005 zal de tweede oproep met een omvang van maximaal 4,3 miljoen euro worden gepubliceerd. Tijdens de bijeenkomst op 15 september werden workshops gehouden waarin de deelnemers van bedrijven en kennisinstelling nieuwe projectideeën ont-

wikkelden die in de tender aan het eind van het jaar als project ingediend worden door Vlaamse en Nederlandse kennisinstellingen.

Meer informatie over STEVIN en aanvullende informatie over het brokerage event op 15 september is te vinden op: <http://taalunieversum.org/taal/technologie/stevin>

2. Het indienen van subsidieaanvragen: do's en dont's

In het kader van STEVIN kunnen partijen hun projectaanvragen indienen. Tijdens de NOTaS-deelnemersvergadering van 9 juli gaf Professor Lou Boves (vanuit CLST) een nuttig en interessant inzicht in de belangrijkste aspecten van het indienen van een aanvraag.

Zo noemde hij onder andere de volgende punten om in gedachten te houden bij het opstellen van een aanvraag:

- Geen enkele financier geeft 100% van de projectkosten.
- Onderzoeks- en ontwikkelprojecten vereisen een substantiële investering.
- Een aanvraag is alleen zinvol als de staf voldoende tijd heeft om mee te werken.

Voor ondersteuning bij het schrijven van voorstellen kan men terecht bij de volgende partijen:

- Senter – voor het uitvoeren van een projecttoets en advies.
- Subsidiologen.
- Collega's en relaties met ervaring.

3. Werkgroep Flankerend Beleid

Op 10 augustus kwam de werkgroep Flankerend Beleid bijeen voor een eerste vergadering. Deze werkgroep is onderdeel van het STEVIN-programma en is verantwoordelijk voor het voorbereiden van voorstellen voor activiteiten die ontwikkelingen op het gebied van taal- en spraaktechnologie kunnen bevorderen. Hierbij zijn de speerpunten:

- Kennisoverdracht – naar partijen buiten het programma;
- Netwerkvorming – contacten leggen met en tussen anderen en samenwerking stimuleren met anderen;
- Verankering – de toekomst van taal- en spraaktechnologie verzekeren;
- Vraagstimulering – door middel van demonstrators, gebruikersvoorlichting, markteducatie en de overheid als een 'launching customer'.

Debbie Kenyon-Jackson, voorzitter van NOTaS, is tevens voorzitter van de werkgroep Flankerend Beleid.

Het definitieve plan voor Flankerend Beleid met een tijdsplanning wordt op dit moment uitgewerkt. Ook was de werkgroep verantwoordelijk voor de brokerage event op 15 september. Deze wordt tijdens de volgende vergadering op 6 oktober geëvalueerd. Daarna volgt ook meer

informatie over de plannen en ideeën voor vraagstimulering en de andere speerpunten.

4. TST-centrale: stand van zaken

De TST-centrale is een INL-project, waarbij betrokkenen zorgdragen voor de verzameling van digitale materialen voor het Nederlands die centraal worden beheerd, onderhouden en beschikbaar worden gesteld. De TST-centrale is met andere woorden hét Nederlands-Vlaams loket waar men terecht kan voor onder meer geluidsopnamen, gedigitaliseerde teksten, corpora voorzien van allerlei taalkundige annotaties, computationele mono- en multilinguale lexica en thesauri, woordsoort-taggers, lemmatiseerders, parsers, enzovoort.

De TST-centrale speelt een belangrijke rol in het STEVIN-meerjarenprogramma. Van het totale budget komt 3% toe aan de TST-centrale. Ieder project moet voor het wordt ingediend met de TST-centrale overleggen over IPR, beheer, onderhoud en beschikbaar stellen van materiaal.

Tijdens de zomer hebben er diverse ontwikkelingen plaats gevonden, waaronder:

- Het aanstellen van acht nieuwe personeelsleden. De teamleden werken in Leiden, Antwerpen, en Nijmegen. Vanaf januari 2005 wordt bijkomend personeel aangeesteld, in totaal 1 FTE.
- Het operationeel maken van de website www.tst.inl.nl, sinds begin augustus. Ook is er een informatiebrief over de werking van de TST-centrale.
- Het instellen van twee van de te verwachten vijf adviescommissies, namelijk de ALVV-commissie (01/04/2004) en de CGN-commissie (01/07/2004).

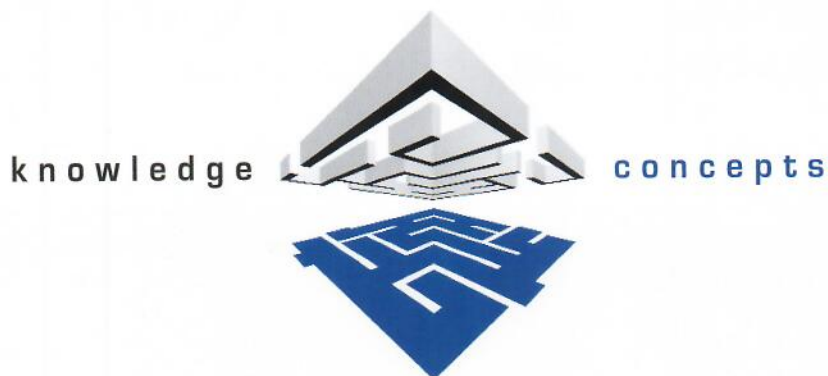
NOTaS

Deelnemers van Stichting NOTaS zijn:

AskNow Solutions - CLST - Comsys - EliTech - Human Inference -
Institute for Logic Language and Computation (ILCC) - Knowledge
Concepts - ORCAvoice - PAT Document Systemen - Polderland
Language and Speech Technology - Q-go - TeleCats - TNO TPD -
Universiteit van Tilburg Faculteit Letteren - Universiteit Twente
Faculteit Informatica - Van Dale Lexicografie - Vocalibur Language
& Speechtechnology

Meer informatie: www.stichtingnotas.nl

Wij werken aan de derde generatie Intranet



Willem van Beuningenlaan 3, 5261 NG Vught.
Tel: +31(0)73 6840199 | Fax: +31(0)73 6840328
E-mail: info@knowledge-concepts.com | Web: www.knowledge-concepts.com

Spraakherkenning

Selfservice Telefonie

Interactive voice response



Veel organisaties besparen enorme kosten door het inzetten van selfservice telefoniesystemen. Selfservice telefoniesystemen handelen telefoongesprekken geheel of gedeeltelijk af m.b.v. IVR en spraakherkenning.

TeleCats is de partij om zaken mee te doen als we praten over selfservice met behulp van IVR en spraakherkenning. Organisaties die dat dagelijks ondervinden zijn o.a.: De Telegraaf, Eneco Energie, OHRA, SNT, KWF, het Ministerie van VROM, De Belastingdienst en NS reizigers.

Wilt u weten hoeveel kosten u kunt besparen door de selfservice telefoniesystemen van TeleCats in te zetten? Maak dan een afspraak met ons.



Internet: www.TeleCats.nl info@TeleCats.nl Telefoon: 053 4889900