DIXIT

TIJDSCHRIFT OVER TAAL- EN SPRAAKTECHNOLOGIE

BIG DATA en TST

JAARGANG 9, DECEMBER 2012

INHOUD

Voorwoord

Taal- en spraaktechnologie relevanter dan ooit

Voor u ligt een zeer relevante DIXIT: relevant voor uw werk, de maatschappij en ook voor u persoonlijk. Is het mogelijk om dreigtweets op te sporen? Dit nummer laat de nieuwste ontwikkelingen op dit voor de maatschappij belangrijke gebied zien.

En hoe zit het met het bewaken van wat (jonge) mensen allemaal op de social media zetten? Deze editie van DIXIT zit vol met de laatste ontwikkelingen en concrete voorbeelden uit de praktijk. Kunt u weten wanneer er iets vervelends over uw merk of bedrijf op internet wordt geplaatst? Leonoor van de Beek geeft waardevol inzicht in de (on)zin van sentiment mining. Dit is tevens de laatste DIXIT voor Leonoor als redactieteamlid. We zullen haar kennis en inspirerende kijk op taal- en spraaktechnologie missen en wensen haar veel succes in haar nieuwe functie bij bol.com.

CSI The Hague en de Tweede Kamer

Na het enorme internationale succes van CSI The Hague, is het voor TST in Nederland bemoedigend te ervaren hoe taal- en spraaktechnologie een belangrijke rol kan spelen bij forensisch onderzoek. Wauter Bosma van het Nederlands Forensisch Instituut geeft een kijkje in de fascinerende keuken van zijn werk. De deelnemers van NOTaS hebben recentelijk ook het geluk gehad dat zij te gast mochten zijn bij het NFI. Het bezoek aan NFI was de tweede in een reeks deelnemersbijeenkomsten nieuwe stijl. Hierbij zoeken we interessante, indrukwekkende locaties met innovatieve toepassingen op het gebied van TST. Presentaties en demonstraties worden, waar mogelijk, gecombineerd met concrete vragen voor oplossingen voor actuele problemen waaraan TST een bijdrage zou kunnen leveren. In de vorm van een interactieve sessie en een netwerk-lunch worden medewerkers van de gastlocatie in contact gebracht met onderzoekers en ontwikkelaars op de voorhoede van TST in Nederland. De eerste van deze sessies in de imposante omgeving van de Tweede Kamer werd druk bezocht door zowel NOTaS-deelnemers als vertegenwoordigers van o.a. de Dienst Informatie Voorziening van de Tweede Kamer.

Vooraankondiging Taal in Bedrijf 2013

Omdat TST relevantie heeft voor alle aspecten van ons dagelijks leven, hebben wij ook stappen ondernomen om ervoor te zorgen dat NOTaS ook relevant blijft voor haar deelnemers en stakeholders. De nieuwe bijeenkomsten zijn een voorbeeld hiervan. Verder, na het succes van Taal in Bedrijf in 2011, zijn we druk bezig met de voorbereidingen voor de volgende editie waarbij u weer verzekerd zult zijn van een inspirerende dag vol met de nieuwste praktijkvoorbeelden die ook van belang zijn voor u, uw organisatie en uw omgeving.

Namens NOTaS wens ik u veel relevant leesplezier!
Debbie Kenyon-Jackson
Voorzitter

COLOFON

DIXIT: Tijdschrift over toegepaste taal- en spraaktechnologie – 9e jaargang, editie Big Data & TST. DIXIT is een uitgave van Stichting NOTaS, Postbus 31070, 6503 CB Nijmegen. Tel. 024-352 88 88 – E-mail info@notas.nl – www.notas.nl **Redactieadres:** Stichting NOTaS, Postbus 31070, 6503 CB Nijmegen **Redactie:** Arjan van Hessen: hessen@cs.utwente.nl, Henk van den Heuvel: h.vandenheuvel@let.ru.nl, Remco van Veenendaal: rvanveenendaal@taalunie.org, Leonoor van der Beek: vdbeek@gmail.com, Anne Wijnen: info@notas.nl **Advertenties:** Stichting NOTaS – Anne Wijnen: info@notas.nl, 024-352 88 88 **Abonnementen:** Voor een gratis abonnement dient u zich te wenden tot een van de NOTaS-deelnemers **Druk:** Leonard bv

Analyse van ongestructureerde data wordt belangrijk!

Dit nummer van DIXIT gaat in op de wijze waarop we met grote hoeveelheden ongestructureerde data kunnen omgaan. De voorzitter geeft in haar voorwoord al een mooie voorzet. De proliferatie van data en databestanden leidt tot een steeds dringender behoefte aan gepersonaliseerde zoekmogelijkheden en gepersonaliseerde samenvattingen, die beslissers op alle fronten voorzien van snelle, betrouwbare en relevante kennis en informatie. Het lezen van teksten en het bepalen van acties zal in een aantal gevallen in eerste instantie worden overgelaten aan slimme systemen die de lezer/beslisser ontslaan van de taak om zelf alle relevante bronnen te bestuderen.

Kees Groeneveld
Partner
Content2Context BV
Xperts-4U BV

Werpen we een blik terug in de tijd dan zien we dat de laatste paar jaar een enorme omwenteling plaatsgevonden heeft en nog steeds plaatsvindt wat betreft de wijze waarop mensen onderling met elkaar communiceren. Denk ik aan de wijze waarop er in mijn gezin wordt gecommuniceerd, dan zie ik mijn kinderen de gehele avond sms'en, e-mailen, foto's uitwisselen en Twitteren met vriendjes met hun mobiele, van internet voorziene, telefoon of iPad. En dat ook nog eens terwijl ze televisie kijken…

Ook zakelijk zie je dat communicatie met e-mail niet weer weg te denken is en er steeds meer gebruik gemaakt wordt van verschillende social media zoals Facebook, Twitter en LinkedIn. Kortom: er worden steeds meer gegevens met elkaar gedeeld. Gegevens die interessant zijn voor bedrijven en instellingen. En soms interessant voor kwaadwillenden of voor hen die het (potentiële) kwaad juist willen bestrijden. Het lastige echter van e-mail en de diverse sociale media is dat de gegevens op een ongestructureerde wijze in vrije tekstvelden vermeld staan. Hierdoor is de inhoud lastig te analyseren. De toename van ongestructureerde data neemt door het gebruik van e-mail en sociale media echter explosief toe. Gartner geeft aan dat de hoeveelheid ongestructureerde data de komende vijf jaren zelfs zal stijgen naar 650% van wat er nu reeds geproduceerd wordt.

Tekstanalyse

De nieuwe vormen van communicatie waarbij gebruik wordt gemaakt van (vrije) tekst (velden) vragen nieuwe vormen van het analyseren van de enorme hoeveelheden aan gegevens. Taaltechnologie gebaseerd op natuurlijke taal biedt hiervoor een oplossing. Een voorbeeld is het geautomatiseerd kunnen analyseren van tekstberichten. Hierdoor kan je als bedrijf of instelling zien hoe er over jouw merk of jouw product of dienst gedacht wordt en kan je op basis van deze gegevens gaan voorspellen in welke richting deze meningen zich verder ontwikkelen. Ook kunnen op basis van tekstanalyse (be)dreigingen ontdekt worden (lees maar eens de artikelen over de digitale speurhond of over de tweetanalyse door de politie Drenthe elders in deze DIXIT).

Een ander voorbeeld is het gebruik van taaltechnologie bij het analyseren van e-mail. Door gebruik te maken van tools die geautomatiseerd kunnen onderscheiden welke e-mail klachten bevat of voorstellen voor productinnovatie, kan de e-mail daarna automatisch doorgezet worden naar de juiste afdeling, projectgroep of behandelaar binnen de organisatie. Hierdoor kan sneller ingespeeld worden op klantklachten of –wensen!

Geautomatiseerd samenvatten

Daarnaast zie je een overload aan informatie ontstaan. Tijd ontbreekt om alle informatie volledig te lezen. Samenvatten is een mogelijkheid. Handmatig alle informatie samenvatten, is echter een te dure bewerking. Geautomatiseerd samenvatten op basis van taaltechnologie is dan een optie. Een ontwikkeling die steeds meer ingezet wordt op redacties van uitgeverijen om uit nieuwspagina's van over de hele wereld de relevante tekst in samengevatte vorm aangeleverd te krijgen. Deze toepassing zien we ook binnen organisaties waar veel met dossiers wordt gewerkt. Samenvattingen worden steeds belangrijker om in korte tijd in staat te zijn om grote hoeveelheden informatie op een kernachtige wijze tot je te kunnen nemen. Daarbij is belangrijk dat de gebruiker de vrijheid heeft om de samenvatting naar eigen inzicht langer of korter gepresenteerd te krijgen en zelf kan bepalen welke onderwerpen voor hem in een samenvatting relevant zijn.

Relevantie in context

Dit soort tools op basis van taaltechnologie heeft zich lange tijd in een experimentele

sfeer bevonden. We zien echter daarin de laatste drie jaar een radicale ommekeer. Taaltechnologie wordt steeds vaker als middel ingezet om op basis van zinsanalyse, onderwerpbepaling, contextanalyse, etc. relevante gegevens effectief uit de berg ongestructureerde data te halen en te analyseren. En dat is niet altijd eenvoudig! Tekst is grillig en je moet goed kunnen analyseren wat er feitelijk door de gebruiker bedoeld wordt. De context is daarbij zeer belangrijk. Wanneer in een bericht over een garagebedrijf een zin staat als *"De service bij garage ABC is vet"* dan heeft het woord 'vet' een andere lading dan het woord 'vet' in de zin *"Net een pot vet besteld"*. Wanneer je een analyse maakt over de berichten die complimenten bevatten, dan moet de eerste maar niet het tweede bericht worden meegenomen. Kortom: woorden krijgen relevantie in de context van andere woorden en het is belangrijk dat de samenhang daartussen wordt meegenomen in de analyse.

Geautomatiseerd metadateren

Taaltechnologie wordt ook steeds vaker ingezet in i-postkamers, waar gedigitaliseerde post of e-mail nu nog vaak door postkamermedewerkers fysiek gelezen en handmatig gemetadateerd moet worden. Het gebruik van goed opgezette en goeddoordachte metadata (data over data) is cruciaal om gegevens later weer snel terug te kunnen vinden. Wat betreft de metadata zijn verschillende standaarden voorhanden. Vaak wordt een combinatie van de volgende velden gehanteerd: Documentsoort, Datum, Selectielijst, Identificatienummer, Dossiernummer, Auteur, Bewaar- of vernietigingscategorie, Bewaartermijn, Vernietigingsjaar, Onderwerp toekenning, Taalcode, etc. In de markt zijn nu op taaltechnologie gebaseerde tools voorhanden die geautomatiseerd metadata kunnen toekennen. Het voordeel van het geautomatiseerd metadateren is dat een veel uniformere wijze van het toekennen van metadata ontstaat. Nu gebeurt het frequent dat door de handmatige toekenning van metadata fouten worden gemaakt of door verschillende interpretaties van soortgelijke documenten andere metadata worden toegekend. Door geautomatiseerd metadateren kan echter veel tijd bespaard worden en kunnen fouten worden voorkomen. Ook kunnen bijvoorbeeld klachten sneller afgehandeld worden, doordat ze door de automatische inhoudsherkenning direct naar de juiste afdeling of behandelaar kunnen worden gestuurd.

Taal evolueert

Taal evolueert. Dat betekent ook dat taaltechnologische oplossingen blijvend getraind moeten worden. Trainen op nieuw vakjargon en trainen op nieuwe woorden of wijziging van de betekenis van woorden. Taaltechnologie gebaseerd op natuurlijke taal heeft de toekomst, omdat je op basis hiervan op een geautomatiseerde wijze met ongestructureerde data kunt omgaan. Het is het niveau van experimenteren ontstegen. Met deze tools kunnen bedrijven en instellingen uit enorme hoeveelheden ongestructureerde data op eenvoudige wijze informatie halen en analyseren waardoor zij sneller kunnen acteren in hun markt of daarbuiten.

In de hierna volgende artikelen in DIXIT wordt onder andere ingegaan op de mogelijkheden die taaltechnologie biedt voor toezicht op sociale netwerken, de zin en onzin van sentiment analyse, het leren van regelmatigheden in grote hoeveelheden natuurlijke taaldata en een artikel met een verrassende insteek over het digitaal uitsterven van veel Europese talen. Kortom, deze uitgave van DIXIT bevat een aantal zeer interessante artikelen waarmee we steeds meer en beter grip krijgen op het verschijnsel BIG Data en de analysemogelijkheden daarvan. Veel leesplezier.

Taalgeneratie: van (big) data naar tekst

Van de grote hoeveelheden informatie die tegenwoordig geproduceerd worden, is een groot deel niet beschikbaar in een door mensen leesbare vorm. Om dit soort data toegankelijk te maken kan gebruik gemaakt worden van taalgeneratie: taaltechnologie die zich richt op het omzetten van niet-talige informatie naar een welgevormde tekst in natuurlijke taal.

Mariët Theune
HMI, Universiteit Twente

Traditionele toepassingen van taalgeneratie zijn onder meer het genereren van weerberichten op basis van meteorologische data en het genereren van sportverslagen op basis van wedstrijdgegevens. Een van de eerste in de praktijk gebruikte taalgeneratiesystemen was 'FOG' dat vanaf begin jaren negentig werd ingezet door de Canadese meteorologische dienst voor het genereren van weerberichten in het Frans en Engels. Een voorbeeld van het genereren van sportverslagen, ook uit de jaren negentig, is het Nederlandse systeem 'GoalGetter', dat korte (gesproken) verslagen genereerde van voetbalwedstrijden uit de toenmalige PTT Telecompetitie. Dit gebeurde op basis van gestructureerde data over de wedstrijd, afkomstig van Teletekst. Naast de Teletekstdata maakte dit systeem ook gebruik van een database met achtergrondgegevens over de teams om de geproduceerde verslagen verder 'aan te kleden'.

Getalsmatige gegevens

De Teletekstgegevens die werden gebruikt door GoalGetter zijn ook goed leesbaar in hun oorspronkelijke vorm. Een groot deel van de tegenwoordig beschikbare data bestaat echter uit voor mensen onleesbare getalsmatige gegevens. Zulke gegevens kunnen worden gepresenteerd door middel van grafieken en tabellen, maar die hebben als nadeel dat ze niet geschikt zijn voor mensen met een visuele beperking, en ook voor anderen vaak moeilijk te doorgronden zijn.

Presentatie van de gegevens door middel van tekst is een laagdrempelig alternatief, met als bijkomend voordeel dat teksten met behulp van tekst-naar-spraaktechnologie ook in gesproken vorm beschikbaar kunnen worden gemaakt.

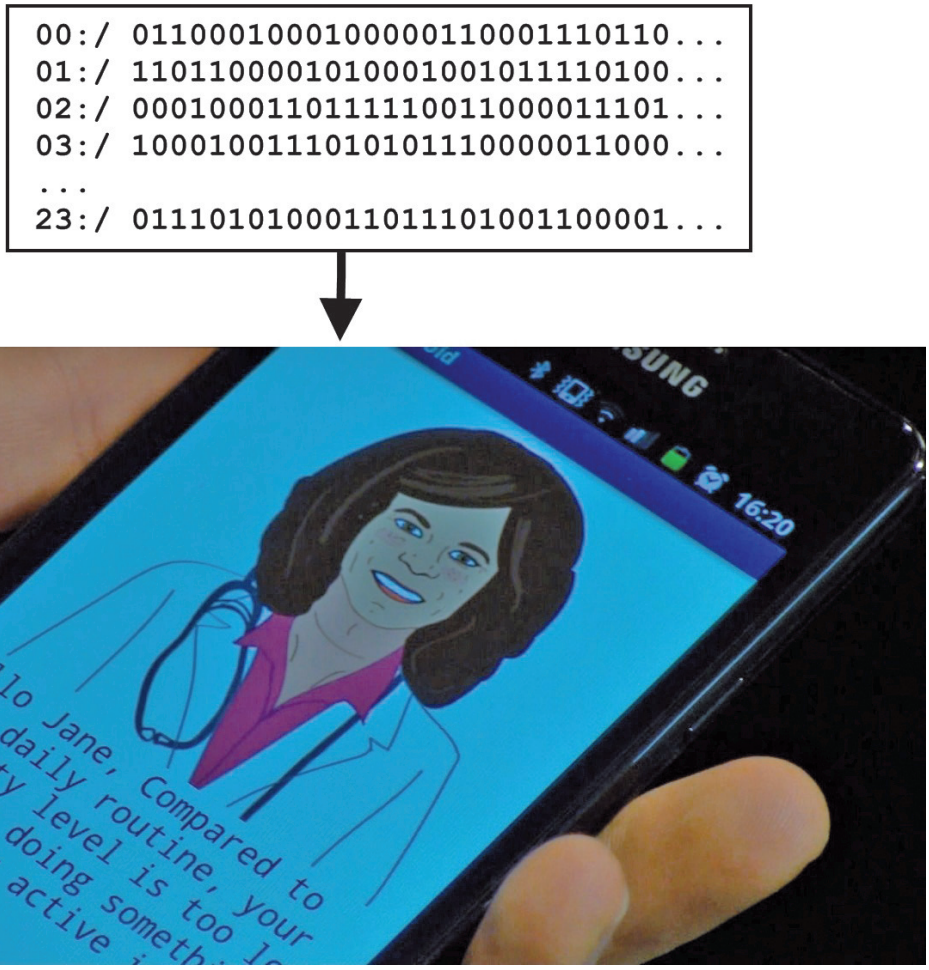
Een zakelijke toepassing van het naar tekst omzetten van getalsmatige data is het automatisch genereren van rapportages op basis van statistieken over bedrijfsprocessen, aandelenkoersen, enzovoort, die door bedrijven op grote schaal worden bijgehouden. In het medische domein wordt onderzoek gedaan naar het genereren van patiëntrapportages op basis van klinische gegevens. Een voorbeeld is het BabyTalk systeem van de Universiteit van Aberdeen, dat verslagen genereert over de toestand van babies op de neonatale intensive care. De invoer van dit systeem bestaat uit grote hoeveelheden gegevens over onder meer hartslag en bloeddruk, die automatisch door medische apparatuur gemeten worden. In het kader van het EU-project Smarcos wordt aan de Universiteit Twente gewerkt aan het geven van automatische feedback aan diabetespatiënten. De bewegingen van de patiënten worden gedurende de gehele dag geregistreerd door een op het lichaam gedragen sensor. Deze data worden vertaald naar korte teksten over hun bewegingspatroon, die op gezette tijden door een virtueel karakter aan de gebruikers worden gepresenteerd, bijvoorbeeld via hun mobiele telefoon.

Go Ahead Eagles ging op bezoek bij Fortuna Sittard en speelde gelijk. Het duel eindigde in twee-twee. Vijfenveertighonderd toeschouwers kwamen naar "de Baandert".

De ploeg uit Sittard nam na zeventien minuten de leiding door een treffer van Hamming. Een minuut later bracht Schenning van Go Ahead Eagles de teams op gelijke hoogte. Na chtenveertig minuten liet de aanvaller Hamming zijn tweede doelpunt aantekenen. In de vijfenzestigste minuut bepaalde de Go Ahead Eagles speler Decheiver de eindstand op twee-twee.

De wedstrijd werd gefloten door scheidsrechter Uilenberg. Hij deelde geen rode kaarten uit. Marbus van Go Ahead Eagles liep tegen een gele kaart aan.

GoalGetter: van wedstrijdgegevens naar voetbalverslag.



Smarcos: persoonlijke feedback op basis van sensordata.

Wat en hoe

Wat komt er zoal kijken bij het omzetten van data naar tekst? Het proces bestaat grofweg uit twee fasen. Eerst wordt bepaald wat er gezegd moet worden (inhoud) en vervolgens hoe dit gezegd moet worden (vorm). Wanneer de invoer van het taalgeneratiesysteem bestaat uit beperkte datastructuren is de eerste fase nagenoeg triviaal; zo dienen in het geval van GoalGetter simpelweg alle beschikbare gegevens in de tekst vermeld te worden. Bij grote aantallen data is het maken van een selectie echter essentieel. Deze moeten eerst door speciale software geanalyseerd worden om belangrijke patronen of gebeurtenissen te herkennen die in de tekst genoemd moeten worden.

De daadwerkelijke taalgeneratie vindt plaats in de tweede fase, die een aantal gerelateerde deeltaken omvat. Ten eerste moet de structuur van de tekst bepaald worden: in welke volgorde wordt de informatie verteld en hoe wordt deze verdeeld over de tekst? Vaak worden hier vaste tekstschema's voor gebruikt. Bij GoalGetter wordt bijvoorbeeld altijd begonnen met algemene gegevens over de wedstrijd, en geëindigd met informatie over de scheidsrechter en de uitgedeelde kaarten. Schema's zijn echter weinig flexibel. Een meer geavanceerde methode is om de tekststructuur afhankelijk te maken van de inhoud. Een medische rapportage kan bijvoorbeeld beginnen met de belangrijkste informatie, die per patiënt zal verschillen.

Een volgende taak is het verdelen van informatie over zinnen. Om samenhangende teksten te creëeren, kunnen gerelateerde gegevens bij elkaar worden genomen en uitgedrukt in één samengestelde zin, bijvoorbeeld een opsomming van alle gele kaarten in een voetbalverslag, of van alle informatie over de bloeddruk in een medische rapportage. Een andere manier om de samenhang van de gegenereerde tekst te bevorderen is het gebruik van persoonlijke voornaamwoorden om te verwijzen naar eerder genoemde personen of zaken.

De laatste stap voor het produceren van een tekst is het vormen van grammaticale zinnen. Hiervoor zijn in veel talen algemeen toepasbare zinsgeneratiesystemen beschikbaar. Zo is voor het Nederlands aan de Rijksuniversiteit Groningen een zinsgeneratiesysteem

ontwikkeld dat gebruik maakt van de Alpino-grammatica van het Nederlands. Optioneel kan de gegenereerde tekst nog naar spraak worden omgezet, en/of worden gepresenteerd door een virtueel karakter, zoals bij Smarcos.

Na jarenlang onderzoek leveren de meeste van bovengenoemde taalgeneratietaken weinig problemen meer op. Het genereren van goed leesbare, samenhangende en grammaticale teksten is inmiddels goed mogelijk. Er zijn echter nog diverse punten waarop de kwaliteit van de gegenereerde teksten verder verbeterd kan worden.

Uitdagingen

Een eerste uitdaging is het op maat maken van teksten voor het publiek waarvoor deze bedoeld zijn. Automatisch gegenereerde teksten zijn meestal neutraal geformuleerd, maar het persoonlijke advies van een virtuele 'lifestyle coach' vereist een andere stijl dan een rapportage van de aandelenkoersen van een bedrijf. Bovendien is het wenselijk om dezelfde gegevens op verschillende manieren te kunnen presenteren aan verschillende doelgroepen. Een verpleegkundige stelt immers andere eisen aan een medisch rapport over een zieke baby dan een bezorgde ouder. Tot nu toe is het onderzoek naar het automatisch aanpassen van gegenereerde teksten meer gericht geweest op de inhoud dan op de vorm. Enkele recente uitzonderingen zijn onderzoek naar het genereren van teksten met een variabele moeilijkheidsgraad, en het automatisch aanpassen van het beleefdheidsniveau van een dialoogsysteem aan dat van de gebruiker.

Een andere uitdaging is het genereren van narratieve teksten, die geen simpele aaneenrijging van feiten zijn maar die causale en temporele samenhang vertonen. Dit omvat onder andere het kiezen van de juiste voegwoorden die de relatie tussen verschillende tekstelementen uitdrukken. Een ander narratief aspect is het omgaan met verschillende verhaallijnen. Als de invoer voor het taalgeneratiesysteem gegevens bevat over meerdere 'hoofdrolspelers' (bijvoorbeeld verschillende personen of bedrijven) dan moet het systeem keuzes maken op wie de focus gelegd wordt, en wanneer. Het is duidelijk dat vaste tekstschema's hiervoor niet volstaan; meer complexe narratieve technieken zijn nodig.

Onderzoek naar dergelijke technieken is tot nu toe grotendeels beperkt gebleven tot het gebied van computationele creativiteit, dat wil zeggen het automatisch genereren van fictieve teksten. Een voorbeeld hiervan is de Virtuele Verhalenverteller die aan de Universiteit Twente is ontwikkeld. Omdat bij dit soort systemen de inhoud van de teksten door het systeem zelf is bedacht, is het relatief eenvoudig om inhoudelijke relaties te bepalen en uit te drukken, en om verschillende verhaallijnen te selecteren. Voor het genereren van teksten op basis van 'echte' data is dit echter nog een grotendeels onontgonnen gebied.

Social Media in SoNaR

Tot voor kort waren er weinig tot geen corpora van sociale media zoals chats, tweets en sms. Er was dus nauwelijks corpusonderzoek mogelijk naar deze nieuwe vormen van communicatie. Het SoNaR corpus opende nieuwe onderzoekswegen naar sociale media, doordat sociale media werden verzameld als onderdeel van het corpus. De auteurs waren verantwoordelijk voor de collectie van chats, tweets en sms.

Eric Sanders en Maaske Treurniet
CLST, Radboud Universiteit Nijmegen

Het SoNaR Nederlandstalig referentiecorpus is ontwikkeld als onderdeel van STEVIN. Eén van de doelen van het STEVIN programma was het realiseren van een adequate digitale taalinfrastructuur voor het Nederlands. Het ontwerpen van een referentiecorpus werd beschouwd als één van de vereisten voor het ontwikkelen van andere bronnen, tools en applicaties. Het corpus is van groot belang voor verder onderzoek in natural language processing. Toepassingen en onderzoek op het gebied van informatie-extractie, question-answering, documentclassificatie en automatisch samenvatten, die gebaseerd zijn op corpus gebaseerde technieken, kunnen profiteren van de grootschalige analyse van het corpus.

SoNaR bevat 500 miljoen woorden, uit Nederland en Vlaanderen. Het corpus is opgebouwd uit teksten in het hedendaagse Nederlands (vanaf 1954), verdeeld over uiteenlopende domeinen en genres. Bij het verzamelen van teksten is ook aandacht uitgegaan naar teksten waar gebruikers mee in aanraking komen via nieuwe, digitale media.

Nieuw taalgebruik

Met de komst van internet en mobiele communicatieapparatuur zijn er nieuwe manieren van communicatie ontstaan en daarmee ook nieuwe manieren van taalgebruik. De eerste methode die ontstond om met elkaar te communiceren via computernetwerken is e-mail. Deze vorm van communicatie heeft nog een grote gelijkenis met een conventionele brief. Snel daarna kwam echter het chatten, dat een compleet nieuwe manier van communiceren is. Chat is geschreven tekst, waarbij de deelnemers elkaar normaliter niet zien, maar die wel verloopt met de snelheid en beurtwisselingen die voorheen alleen bij spreken normaal waren. Men zou chats kunnen beschrijven als een soort getypte telefoongesprekken.
Sms is ontstaan toen de mobiele telefoon gemeengoed werd. Sms'en zijn losse berichten, met een maximale lengte van 160 karakters. Omdat invoer via een telefoon een lastig proces is, is het een relatief moeizame manier van communicatie.

Twitter is pas recent, sinds 2006, in gebruik, maar heeft sindsdien een duizelingwekkende vlucht genomen. Op moment van schrijven worden er dagelijks wereldwijd 300 miljoen tweets verstuurd, maar dit aantal verandert met de dag. Tweets hebben een grote gelijkenis met sms'en. Tweets bevatten maximaal 140 karakters en zijn normaliter losse berichten. Het grootste verschil met sms betreft het aantal ontvangers. Daar waar een sms meestal bij één persoon belandt, zijn tweets doorgaans openbaar en te lezen voor een soms grote schare volgers.

Deze nieuwe vormen van communicatie hebben elk geleid tot nieuwe specifieke vormen van taalgebruik. Het taalgebruik in chats is informeler dan eerder in geschreven taal gewoon was. Daarbij zijn chats zeer belangrijk geweest voor de ontwikkeling en het gebruik van emoticons (de zogenaamde 'smileys'). Vanwege de snelheid van chatconversaties en de beperkte lengte van sms- en twitterberichten is er een compact taalgebruik ontstaan met afkortingen en het weglaten van woorden en letters. In het Nederlands taalgebied is nog weinig onderzoek gedaan naar het taalgebruik in sociale media, onder

andere vanwege de afwezigheid van geschikte corpora.

Daar is nu verandering in gekomen. Als onderdeel van het SoNaR corpus, zijn chats, tweets en sms verzameld. Een collectie van chats en tweets aanleggen is, in tegenstelling tot sms, relatief eenvoudig. De meerwaarde van het corpus ligt echter in het verkrijgen van de rechten op gebruik en verspreiding, het verkrijgen van de persoonsgegevens van schrijvers en de anonimisering van de teksten. Hieronder staat kort beschreven hoe we de data, metadata en gebruikerstoestemming hebben verkregen en hoe de anonimisering al dan niet werd uitgevoerd.

Chats

Het Nederlandse gedeelte van de chats bestaat uit vier delen. Voor drie delen was een speciale chat-server ingericht om doelgroepen deel te laten nemen aan chat-sessies. In twee gevallen betrof dit middelbareschoolleerlingen en in het derde geval collega's van de onderzoeksgroep Taal&Spraak aan de Radboud Universiteit in Nijmegen. Het vierde deel bestaat uit msn-conversaties van vrienden en bekenden, die waren opgeroepen hun chats te verzamelen. Eén corpus met chatdata van middelbareschoolleerlingen was al verzameld voor het SoNaR-project was begonnen. Dit betreft het ChatIG corpus, in 2005 en 2006 verzameld door Wilbert Spooren en Tessa van Charldorp van de Vrije Universiteit in Amsterdam.
Alle deelnemers, of hun ouders, van de Nederlandse chatdata hebben toestemming gegeven voor distributie van de data en hebben hun metadata (leeftijd, geslacht, woonplaats) gegeven. De (nick)namen van de chatters zijn onherkenbaar veranderd. Verdere anonimisering bleek te tijdrovend om handmatig uit te voeren en te complex om te automatiseren.

Voor het Vlaams is toestemming bemachtigd om data van de grote chatservice chat.be te gebruiken. Er is collectieve toestemming gegeven door de eigenaar van de chatsite. Gedurende een aantal maanden in 2011 is data van het chatkanaal verzameld. Voor deze data is geen individuele toestemming en er is ook geen metadata van de gebruikers beschikbaar.

Al met al bevinden zich 2.194.592 chatregels in SONaR.

Tweets

Voor Twitter bestaat een API waarmee tweets verzameld kunnen worden. Er is enige onduidelijkheid over toestemming voor hergebruik, maar wij hebben de gebruiksvoorwaarden zo geïnterpreteerd dat de tweets in SoNaR opgenomen mogen worden, mits de tekst en de gebruikersnaam onveranderd blijven. Daarmee zijn zowel het toestemmings- als het anonimiseringsprobleem opgelost.

Er zijn twee manieren gebruikt om de metadata te verzamelen. Bij de eerste methode is een tweet verstuurd met daarin de oproep aan twitteraars hun persoonsgegevens te sturen vanwege de aanleg van het SoNaR-corpus. Deze tweet werd geretweet en zo ontstond een sneeuwbaleffect dat uiteindelijk resulteerde in berichten op de hoofdpagina van de Radboud Universiteit (ru.nl) en de bekendste nieuwswebsite van Nederland (nu.nl), waarin mensen werd gevraagd

hun gegevens te mailen. Daarnaast hebben we van bekende en semi-bekende Nederlanders en Vlamingen geslacht, leeftijd en woon- of geboorteplaats opgezocht op hun persoonlijke homepage of op Wikipedia.

In SONaR zijn 1.532.251 tweets opgenomen.

Sms

Het verzamelen van sms-berichten vraagt een specifieke aanpak vanuit het oogpunt van privacy en anonimisering. Er is samengewerkt met de National University of Singapore, School of Computing, die al eerder een verzameling sms'en had aangelegd. Er werd een website opgezet, waar instructies te vinden waren over drie verschillende manieren waarop sms'en bijgedragen konden worden. Allereerst was er de mogelijkheid om sms'en handmatig te kopiëren in een online formulier. Daarnaast konden eigenaars van een Android smartphone een zogenaamde App downloaden, waarmee de sms'en automatisch geëxtraheerd werden uit de telefoon en vervolgens via Gmail naar

SoNaR verstuurd werden. Tenslotte bevatte de website instructies om de sms'en via de PC te downloaden vanuit een iPhone of Nokiatoestel, waarna dit bestand via de SoNaR dropbox ingestuurd kon worden. Om toestemming te verkrijgen van de auteurs van iedere sms, werden de voorwaarden gemeld bij het uploaden van de sms, respectievelijk in de te verzenden mail of door een melding bij de dropbox. Anonimisering van de berichten werd grotendeels overgelaten aan de eigenaars van de sms'en. De originele telefoonnummers werden vervangen door een unieke code en bij sms die via de App werden ingezonden, werden daarnaast automatisch bankrekeningnummers en IP-adressen onherkenbaar gemaakt.

De website werd gepromoot via communicatieafdelingen van de verschillende partneruniversiteiten van SoNaR, via flyers, persberichten, sociale media en onder de eigen kennissenkring van medewerkers.

Al met al bevat SoNaR 42.358 sms-berichten.

De digitale speurhond

Het gebruik van internet en de sociale media heeft de afgelopen jaren een enorme vlucht genomen. Het brede publiek heeft vandaag de dag vrijwel ongelimiteerd toegang tot allerlei informatie en berichtgeving. Nog niet eerder was het zo eenvoudig de actualiteit te volgen, daarop te reageren of in te spelen. Ook bij het onderhouden van sociale contacten zijn de moderne media niet meer weg te denken: we houden elkaar doorlopend op de hoogte van wat we aan het doen zijn, hoe we over dingen denken, we maken afspraken, etc. Je zou bijna geneigd zijn te denken dat er aan deze ontwikkeling alleen maar positieve kanten zitten. Toch zijn er ook nadelen. Voor wie er op uit is, bieden de nieuwe media een enorm platform waarop je ongecensureerd je gang kan gaan, en ook nog eens je identiteit kunt verhullen.

Nelleke Oostdijk
CLST, Radboud Universiteit Nijmegen

Internetsurveillance

De politie monitort al enige tijd het internet op dreigingen gericht tegen personen (bijv. kabinetsleden), objecten (bijv. Schiphol), diensten (bijv. openbaar vervoer) en evenementen (bijv. Nationale Herdenking). De internetsurveillant houdt zich bezig met het identificeren van berichten die vanwege de ernst en de waarschijnlijkheid van de (be)dreiging nadere aandacht van de politie behoeven. Eén van de problemen daarbij is het enorme volume aan data. Nederlanders zijn immers massaal actief op o.a. Facebook, LinkedIn, YouTube en Twitter. Er is dan ook behoefte aan een instrumentarium dat de politie kan helpen internet(be)dreigingen te vinden, te verwerken en te valideren.

Dreigtweets

Dreigtweets zijn een typisch voorbeeld van het soort berichten dat de politie graag en liefst zo snel mogelijk op het spoor zou willen komen. Het verzenden van een dreigtweet staat namelijk gelijk aan iemand offline bedreigen en is strafbaar volgens Artikel 285 van het Nederlandse Wetboek van Strafrecht. Hoewel nog onbekend is om welke aantallen het precies gaat, is al wel duidelijk dat het in de honderden tweets per dag loopt. Zo waren er bijvoorbeeld in de aanloop van de onlangs gehouden Tweede Kamerverkiezingen dagelijks enkele tientallen dreigtweets alleen al aan het adres van Geert Wilders.

Dreigtweetdetectie

In opdracht van het Ministerie van Justitie en Veiligheid werd onlangs door onderzoekers van het CLST een (pilot) project uitgevoerd dat tot doel had een methode te ontwikkelen die ingezet zou kunnen worden om dreigtweets te onderscheiden van niet-dreigtweets. Daarnaast zou de methode ondersteuning moeten bieden bij het inschatten van de ernst en waarschijnlijkheid van de dreiging. Het onderzoek richtte zich specifiek op de Nederlandstalige tweets (incl. Nederlandse straattaal en dialect) die verzonden worden in het Nederlandse domein.

Het onderzoek kent een reeks van uitdagingen. Allereerst gaat het om uiterst korte berichten met een maximale lengte van slechts 140 tekens. Bovendien wijkt het taalgebruik in veel gevallen nogal af van wat we gewend zijn vanuit het standaard Nederlands. Zo zien we een soms nogal eigenzinnig gebruik van de orthografie, een enorme variatie in spelling, typische woordkeuzen en formuleringen.
Een extra complicatie is dat in de huidige opzet van elke tweet afzonderlijk, dat wil zeggen zonder enigerlei context, bepaald moet kunnen worden of het een potentiële dreiging betreft of niet.

Hybride benadering

De gebruikte methode combineert machine learning met een linguïstisch gemotiveerde, regelgebaseerde benadering. De twee benaderingen hebben ieder zo hun sterkten en zwakten. In het geval van machine learning blijft de menselijke inspanning tot een minimum beperkt en kan de computer gebruik maken van 'kennis' opgedaan uit de data,

@...: die wilders moet kapot joh kogel door ze kop #verrijking daarom #PVVop1
@...: moet nu een keer klaar zijn met Geert Wilders. / kill hem
Die geert geeft echt een kogel door ze kop nodig.
Wollah als ik oooit in mn leven Wilders tegen kom djoek ik hem gelijk jood
@... wollah jij gaat dood door een moslim die jou nie meer trek anus likker
Wedden als Geert Wilders zo door gaat hij vroeg of laat word vermoord.

Figuur 1. Voorbeelden van dreigingen aan het adres van Geert Wilders

af bek bom doden **dood** echt **ga** gaan **gaat** kanker kapot keer kil **kill** kk kkr kogel kom komt kop **maak** maken man **mes** moeder moord morgen plan school slaan **steek** steken stoer vallen **vermoord** vermoorden we zie

Figuur 2: Woordenwolk van de meest voorkomende woorden in dreigtweets

kennis die wij als mens niet expliciet voorhanden hebben. Daar staat tegenover dat de computer, om met enig succes automatisch te kunnen leren, grote hoeveelheden trainingsdata nodig heeft en die zijn niet altijd voorhanden. De regel gebaseerde benadering is juist niet afhankelijk van trainingsmateriaal. Het opstellen van de regels gebeurt door de linguïst die daarbij put uit zijn/haar kennis en is daardoor nogal arbeidsintensief. Doordat de regels in meer generieke termen patronen beschrijven, worden tevens voorkomens beschreven die nog niet werden geobserveerd.

De machineleermethode is in dit geval een aangepaste vorm van Linguistic Profiling, een methode die eerder door Hans van Halteren werd ontwikkeld en die met succes werd toegepast o.a. voor auteursherkenning [1]. Vooralsnog blijft de performance van de machine learning component achter bij wat op basis van eerdere ervaringen verwacht mag worden. Dit is te verklaren doordat er nog maar weinig trainingsmateriaal beschikbaar is dat voor de huidige taak geschikt is.

Hoewel de regelcomponent verdere uitbreiding behoeft, blijkt hij toch al zeer effectief. De grootste winst t.o.v. machine learning ligt erin dat de mens ook zonder concreet voorbeeldmateriaal goed in staat is om zeer veel mogelijke verwoordingen van dreiging te omschrijven. Het probleem bij deze aanpak is eerder dat er nog een relatief groot aantal tweets ten onrechte als dreiging wordt aangemerkt. Zo passeerden er bijvoorbeeld in verkiezingstijd nogal wat tweets waarin

melding gemaakt werd van een aanval op een politicus. In de meeste gevallen was er geen sprake van een echte dreiging, maar ging het om metaforisch taalgebruik. Verdere aanscherping van de regelcomponent is dan ook nodig om dergelijke gevallen in de toekomst van detectie uit te sluiten.

Een en ander is geïmplementeerd in een softwaremodule, informeel aangeduid als 'de dreigingsdetector'.

Gebruikerstest

Hoewel er bij de ontwikkelaars al tijdens het pilot project verschillende ideeën rezen over hoe de dreigingsdetector verder uitgebouwd en verbeterd zou kunnen worden, is ervoor gekozen eerst een gebruikerstest te doen op basis van het resultaat van de pilot, alvorens verder te gaan met de ontwikkeling. Een belangrijke overweging hierbij is dat de introductie van een instrument zoals de dreigingsdetector een enorme impact heeft op de huidige werkwijze van de professionals. Geïntegreerd in groter systeem waarin tweets aan de module worden aangeboden heeft een team van de KLPD op Prinsjesdag de module uitgetest. De evaluatie is hiervan is nog niet afgerond, maar de resultaten zijn bemoedigend.

Doorontwikkeling

Behalve dat de gebruikerstest de effectiviteit van de module laat zien en de nodige feedback ontlokt aan de gebruikers ervan, levert de test ook waardevolle inzichten, onder andere v.w.b. de volumes waar we in de praktijk mee te maken hebben. Daarnaast komt er een belangrijke aanvulling op het beschikbare trainingsmateriaal. Voor de verdere doorontwikkeling van de module zijn de plannen momenteel in voorbereiding. Een uitbreiding naar andere media zoals Facebook en LinkedIn ligt daarbij voor de hand, terwijl ook het betrekken van de context van de berichten en auteursprofielen kan bijdragen aan de duiding van potentiële dreigtweets.

[1)] *http://acl.ldc.upenn.edu/acl2004/main/pdf/183_pdf_2-col.pdf*

Monitoring-tool voor de politie Assen

Masterstudent Jimmy Meijer van de Rijksuniversiteit Groningen ontwikkelt voor zijn afstudeeronderzoek een monitoring-tool voor de politie Assen.

"Ik ben in contact gekomen met de politie van Assen nadat mijn fiets in mei 2012 ontvreemd was uit de schuur. De fiets was teruggevonden en stond op het politiebureau, maar twee maanden later had ik mijn fiets nog steeds niet terug. Ik deed via Twitter mijn beklag over de werkwijze van de politie Assen en aan de hand van de beschrijving bij mijn Twitter-profiel ben ik toen uitgenodigd voor een gesprek over social media op het bureau.

De insteek voor de politie was het 24-7 beschikbaar zijn via social media, waarbij mij direct opviel dat Twitter bij de politie Assen voornamelijk gebruikt wordt als uitgaand kanaal. Na enkele brainstormsessies is besloten om Twitter juist als inkomend kanaal te gebruiken door een monitoring-tool te ontwikkelen.

De eerste opzet voor mijn onderzoek was om te kijken naar relevante tweets voor de politie Assen. Daarvoor moeten de tweets dus binnen de regio Assen vallen. Maar hoe dit te bewerkstelligen? Deze vraag bleek iets te specifiek voor mijn scriptie, dat een meer algemeen karakter moet hebben. De onderzoeksvraag is nu dan ook: 'Hoe kunnen we tweets herkennen die moeten leiden tot een actie?'. Zo'n actie kan ontstaan uit een vraag, een klacht of een opmerking. Mijn scriptie is erop gericht een tool te ontwikkelen dat deze tweets automatisch herkent."

Toezicht op sociale netwerken met behulp van taaltechnologie

Eind 2011 publiceerde het Europese 'Kids Online' project [1] een studie over internetgebruik door kinderen in de leeftijd van 9 en 16 in 25 Europese landen. De statistische informatie stemt tot nadenken: gemiddeld zijn kinderen bijna anderhalf uur per dag online, 1 op 2 ook in hun slaapkamer. Een derde had ooit mensen toegevoegd als vriend die ze niet kenden. Tot 15% had al eens persoonlijke informatie gedeeld met vreemden (ook foto's of video). Jongere kinderen bleken de vaardigheden niet te hebben om op een doeltreffende manier hun gebruikersprofielen of privacy-instellingen te beheren of ongewenste contacten te blokkeren. Ouders spelen te weinig de controlerende rol die ze zouden moeten aannemen. Dat alles maakt kinderen en jongeren een gemakkelijk slachtoffer voor ernstige misdrijven zoals cyberpesten en 'grooming' door pedofielen, en laat serieuze signalen van depressie (bijvoorbeeld zelfmoordaankondigingen) en ongewenste inhoud, zoals racistisch, beledigend of pornografisch materiaal gericht aan kinderen, ongedetecteerd.

Walter Daelemans
CliPS, Universiteit Antwerpen

Uiteindelijk is het de verantwoordelijkheid van de beheerders van de sociale netwerken (SN) om de veiligheid van hun site te controleren, en de meeste hebben uitgebreide (meestal manuele) procedures voor het monitoren van interacties en inhoud van hun site. Helaas is het tijdig opsporen van ongewenst gedrag en materiaal in sociale netwerken zo goed als onmogelijk door het reusachtige volume dat moet gecontroleerd worden en het feit dat de incidentie van ongewenste gebeurtenissen (gelukkig) relatief klein is ten opzichte van dit volume. Het monitoren van een SN is dan ook een goed voorbeeld van een 'big data' probleem waar taaltechnologie een oplossing kan bieden, bijvoorbeeld met tekstcategorisatie.

Zelflerende tekstcategorisatietechnieken behoren tot de best ontwikkelde onderdelen van de taaltechnologie. Voor sommige problemen zoals spam filtering en e-mail routering worden routinematig al uitstekende resultaten bereikt. Voor andere problemen, variërend van auteursherkenning tot detectie van valse productreviews wordt snel vooruitgang geboekt in onderzoek en ontwikkeling. Ook de detectie van ongewenste of verontrustende inhoud in sociale netwerken kan met deze technieken geautomatiseerd worden.

In het Daphne project van de Universiteit Antwerpen ontwikkelen we tekstcategorisatietechnieken om pedofielen die actief zijn op SN te ontmaskeren [2]. Het systeem bestaat uit twee delen: profielcontrole en detectie van 'grooming' (het geheel van acties en strategieën waarmee pedofielen proberen iets gedaan te krijgen van kinderen, bijvoorbeeld een afspraak maken of de webcam aanzetten). De *profielcontrole* gebeurt door de chattaal van gebruikers automatisch te analyseren op leeftijd-, sekse- en locatiekenmerken. Als het uit het taalgebruik afgeleide profiel niet klopt met het opgegeven profiel kan dat gemeld worden aan de moderator van het SN. Bijvoorbeeld, een volwassen man die zich voordoet in zijn profiel als een tienermeisje. Vooral leeftijd kan in chattaal met een zeer hoge accuraatheid gedetecteerd worden, zeker wanneer het probleem wordt beperkt tot differentiatie tussen wettelijk relevante categorieën (bijv. minderjarig en meerderjarig). In onze resultaten halen we tot 90% accuraatheid bij het bepalen van leeftijd van de auteur uit de tekstkenmerken van de chat. Chattaal is overigens een moeilijk tekstgenre. Het zijn meestal korte berichten, en er wordt gebruik gemaakt van specifieke afkortingen en acroniemen (LOL, w8, wrm, hjg, …) en in Vlaanderen ook van dialectisch taalgebruik. In

samenwerking met de LT3 onderzoeksgroep in Gent wordt gewerkt aan normalisatietechnieken voor chattaal, zodat taaltechnologische hulpmiddelen zoals woordsoortanalyse, zinsanalyse en dergelijke ook voor dit tekstgenre gebruikt kunnen worden. Maar voorlopig kan in elk geval leeftijd dus al met hoge accuraatheid bepaald worden door tekstcategorisatie op basis van oppervlakkige analyse (tokenisering). Een volgende stap is de ontwikkeling van methodes om pedofielen die zich ook in hun taalgebruik voordoen als kinderen (wat eigenlijk weinig gebeurt), te detecteren. De tweede module in het Daphne systeem is de detectie van grooming. Dit komt neer op de identificatie van chatuitingen die wijzen op een pedofiel in actie. Dit gebeurt met een tekstcategorisatiesysteem dat uitgaat van terminologie-analyse (op basis van wat bekend is over hoe pedofielen te werk gaan). In een recente 'gedeelde taak' (voor het Engels), haalden we goede resultaten met onze methode, zowel voor de identificatie van pedofielen (zesde plaats op zestien met een f-score van ongeveer 70%, het beste systeem haalde 87%) als voor de identificatie van verdachte uitingen (eerste plaats op zestien met een f-score van 30%) [3]. Zo'n gedeelde taak, waarbij verschillende onderzoeksgroepen hun systeem testen op dezelfde data, is een uitstekende gelegenheid om ontwikkelde technieken te benchmarken en snel kennis te vergaren over wat er wel of niet werkt.

Vanaf januari 2013 gaat, gecoördineerd door de Universiteit Antwerpen, in Vlaanderen het project AMICA (Automatic Monitoring for Cyberspace Applications) van start waarbij taaltechnologie in combinatie met beeldtechnologie zal worden ontwikkeld voor de detectie van meer veeleisende problemen in SN, zoals de detectie van cyberpesten, van grensoverschrijdend seksueel gedrag, en van indicaties van psychische problemen zoals aankondigingen van zelfdoding, beelden van zelfverminking en dergelijke. In samenwerking met bedrijven als VRT ketnet (het kinderkanaal van de Vlaamse Radio en Televisie dat ook sociale media onderhoudt) en Netlog (een SN) zullen de ontwikkelde technieken meteen uitgetest worden in een realistische omgeving. Vanzelfsprekend is een belangrijk onderdeel van dit onderzoek het efficiënt genoeg maken van de ontwikkelde technologie zodat ze in reële tijd toegepast kan worden op reusachtige volumes tekst- en beelddata.

Tot besluit: het snel controleren van wat er gebeurt in sociale media is een 'big data' probleem bij uitstek. Voor het tijdig ingrijpen bij ongewenste fenomenen als pedofilie en cyberpesten en bij noodsituaties zoals aankondigingen van zelfdoding is het belangrijk om het werk van de moderatoren zo effectief mogelijk te verlichten. Taaltechnologie in combinatie met data mining kan daar een belangrijke rol bij spelen. Door met filters met een hoge recall in reële tijd een SN te analyseren wordt het mogelijk voor de moderator om op tijd te interveniëren. Een hoge recall betekent dat de filter geoptimaliseerd is om geen gevallen te missen, ook al gaat dat ten koste van een aantal keer 'vals alarm'. Het belangrijkste is dat de zoekruimte van de moderator inkrimpt van tienduizenden tot enkele tientallen interacties. Op die manier kan de taaltechnologie bijdragen tot een veiliger internet.

1) *www.eukidsonline.net*
2) *www.clips.ua.ac.be/projects/daphne*
3) *PAN 2012 Lab 'Uncovering Plagiarism, Authorship, and Social Software Misuse', Rome, 2012. www.webis.de*

Automatische extractie van gebeurtenissen in strafdossiers

In de zomer van 2008 verscheen het bericht dat Saban B. is veroordeeld voor het leiden van een gewelddadige mensenhandelbende [1]. Naast mensenhandel werd hij beschuldigd van mishandeling, poging tot moord, afpersing en witwassen. Om de zaak tegen hem en zijn organisatie rond te krijgen verzamelde het Openbaar Ministerie meer dan twaalfhonderd dossiers van ongeveer twintig verdachten [2]. Het omvangrijke dossier doet beetje bij beetje het hele verhaal uit de doeken, over Saban B., zijn criminele organisatie en zijn slachtoffers. Niet als een chronologisch verslag, met een duidelijke structuur, maar als een boek dat uit elkaar gevallen is en waarvan we slechts enkele pagina's hebben teruggevonden. Een team van rechercheurs heeft maandenlang gewerkt om de pagina's op volgorde te leggen, de verhalen te reconstrueren en met behulp daarvan de strafbare feiten aan te tonen.

Wauter Bosma
Kennis- en expertisecentrum voor intelligente data-analyse, NFI

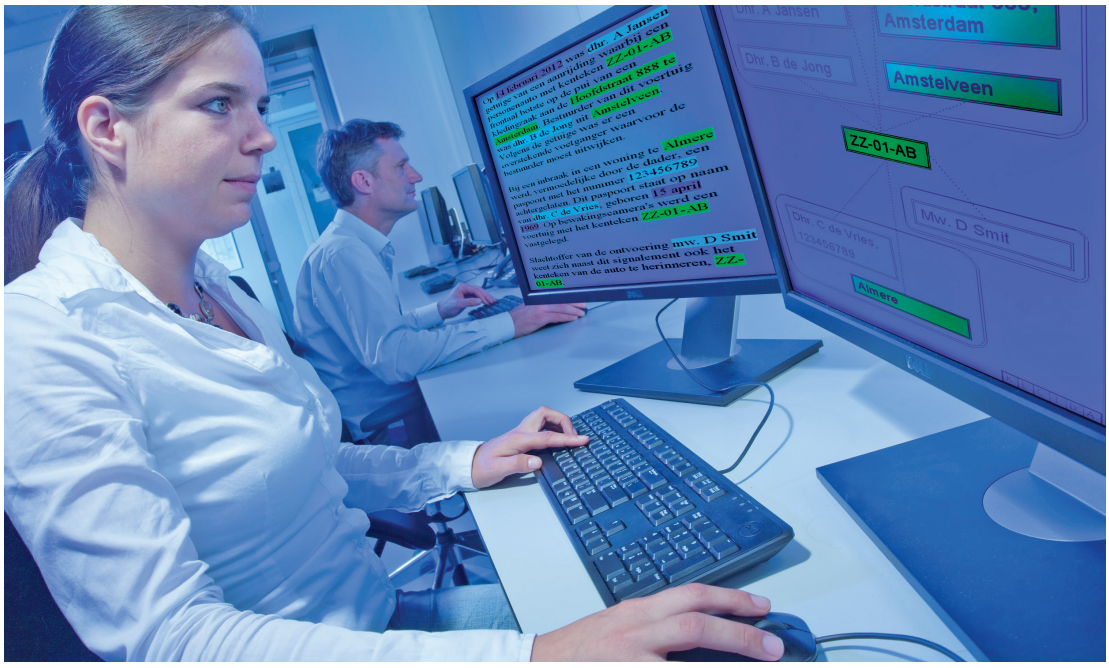
Kecida

Het Kennis- en expertisecentrum voor intelligente data-analyse (Kecida), ondergebracht bij het Nederlands Forensisch Instituut (NFI), biedt opsporingsdiensten ondersteuning bij het analyseren en organiseren van grote hoeveelheden gegevens. In een zaak zoals die van Saban B. wil de rechter uiteindelijk weten wat er precies gebeurd is en hoe dat heeft bijgedragen aan de ten laste gelegde strafbare feiten. Het is de taak van de analist is om dit uit te zoeken. Daarbij moet ook aan het licht komen welke verdachte welke rol heeft vervuld binnen hun samenwerkingsverband. Veel van die informatie is afkomstig van tekst. In persoonlijke communicatie, zoals in e-mail en SMS, maar ook in registratiesystemen waarin over incidenten wordt gerapporteerd, is tekst nog steeds het belangrijkste medium. De informatie die de analist zoekt, over verhoudingen tussen personen en de opeenvolging van gebeurtenissen, is met standaard zoekmachines echter moeilijk te vinden.

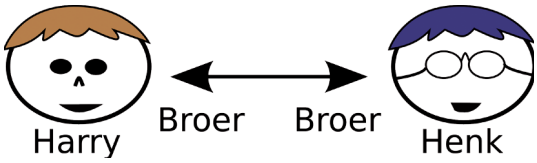
Daarom ontwikkelt Kecida technologie om automatisch relaties en gebeurtenissen uit de tekst te extraheren. Een relatie kan bijvoorbeeld een familieverhouding zijn, en een gebeurtenis een ontmoeting tussen personen, een financiële transactie, of een gewelds-incident. Vervolgens maken we de relaties en gebeurtenissen doorzoekbaar en verwerken we ze in visualisaties die toegesneden zijn op de vraagstelling. Statistieken van financiële transacties over een langere periode maken bijvoorbeeld geldstromen zichtbaar. De individuele transacties blijven nodig om het algemene beeld te onderbouwen.

Gebeurtenissen

Aan de basis van extractie van relaties en gebeurtenissen staan entiteiten. Entiteiten



verwijzen naar iets of iemand in de fysieke wereld, zoals een persoon of een locatie, maar het kan ook een tijdstip of een telefoonnummer zijn. Een entiteit die genoemd wordt in tekst heeft op zichzelf nog beperkte waarde, en krijgt pas echt betekenis in relatie tot andere entiteiten. Bijvoorbeeld, als vermeld wordt dat Harry de broer van Henk is, komen we iets te weten over die twee in relatie tot elkaar. Deze relatie is schematisch weergegeven in Figuur 1. We extraheren relaties automatisch nadat we eerst de afzonderlijke entiteiten herkend hebben. Bij de relatie van broers zijn alleen twee personen betrokken, maar bij andersoortige relaties kunnen andere entiteiten betrokken zijn. Ook kunnen ze andere rollen vervullen.



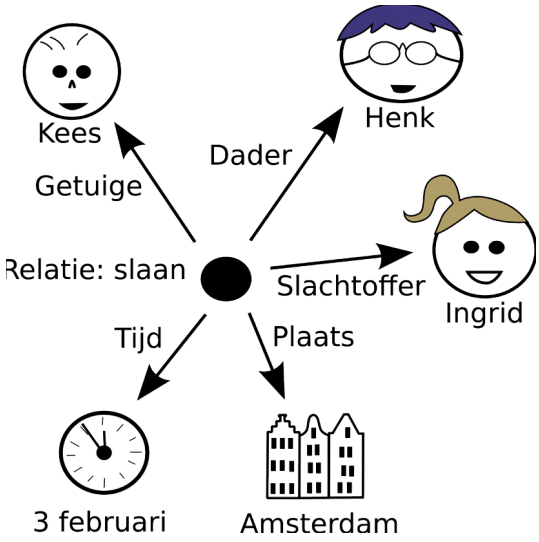
Figuur 1 Weergave van een relatie tussen broers

Ook een gebeurtenis zoals een ontmoeting of een geweldsincident is in feite een relatie tussen entiteiten, maar daarbij spelen naast personen ook plaats en tijd een rol. Stel, Kees is getuige van een geweldsincident en verklaart tegenover de politie dat hij op 3 februari in Amsterdam heeft gezien dat Henk Ingrid heeft geslagen. Henk en Ingrid spelen daarbij de rol van respectievelijk dader en slachtoffer, en daarnaast is er een plaats en een tijd. De derde persoon, Kees, speelt de rol van getuige. De gegevens van deze gebeurtenis zijn verzameld in Figuur 2. In tegenstelling tot Harry en Henk in de relatie van broers spelen de betrokken personen in het geweldsincident dus geen gelijkwaardige rol.

Meta-informatie

De meeste informatie die we gebruiken om relaties te extraheren komt uit de lopende tekst. Soms gebruiken we meta-informatie, zoals bij een getuigenverklaring waarin de getuige in de eerste persoon spreekt. Ook de plaats en tijd waar het brondocument is geschreven kan van belang zijn. Soms wordt het tijdstip van de gebeurtenis bijvoorbeeld niet genoemd of is het relatief, zoals de datum van 3 februari. Het jaartal wordt daar niet expliciet genoemd, maar moet vermoedelijk afgeleid worden van de context.

Maar zelfs als we dit alles meenemen kunnen we niet altijd alle elementen van de relatie invullen, simpelweg omdat niet alle informatie beschikbaar is. Dat we niet alle informatie



Figuur 2 Weergave van een gebeurtenis: slaan

kunnen vinden maakt de gebeurtenis echter niet per se irrelevant. Daarom nemen we ook onvolledige gebeurtenissen mee, waarbij een of meerdere velden niet konden worden ingevuld.

Verborgen informatie

Door dossiers te combineren kunnen we ook verbanden tussen zaken vinden die anders onopgemerkt waren gebleven. De verklaring van Kees van het geweldsincident zal terecht komen in een proces-verbaal, maar op het moment zelf is deze gebeurtenis mogelijk bijzaak, en het feit wordt verborgen in een bijzin te midden van een berg aan informatie. Nu verschijnen signalen dat Harry betrokken is bij mensenhandel, en we weten dat Henk en Harry broers zijn. Deze combinatie van feiten kan aanleiding geven tot een nader onderzoek. Automatische extractie van relaties en gebeurtenissen helpt dergelijke verbanden aan het licht te brengen.

Conclusie

Extractie van entiteiten, relaties en gebeurtenissen helpt analisten gericht te zoeken naar relevante informatie. Entiteitextractie wordt al verwerkt in standaard zoektechnologie. Bij wijze van proef zet Kecida nu ook extractie van relaties en gebeurtenissen in als hulpmiddel voor analisten. De technologie is feilbaar en de nauwkeurigheid hangt sterk af van het soort tekst. Niettemin is onze indruk dat het per saldo een positieve bijdrage levert, omdat het helpt verbanden zichtbaar te maken die anders mogelijk onopgemerkt waren gebleven.

[1] *NRC Handelsblad, 11-7-2008*
[2] *Openbaar Ministerie, www.om.nl/actueel-0/terugkijken/@150256/jan_%27lachende%27/*

BIG DATA behapbaar gemaakt met spraakherkenning

Stel je bent journalist en hebt een drie uur durend interview met iemand gehad. Je artikel (deadline!!) moet over een halfuur klaar zijn en je wilt nog snel die ene quote in je artikel verwerken, maar je weet niet meer precies wanneer het gezegd werd in het interview… Dit voorbeeld laat goed zien dat zoeken in audio niet gemakkelijk is. Als je geen aanknopingspunten hebt kun je het niet zomaar à la Google doorzoeken. Dankzij spraaktechnologie is dit echter tegenwoordig WEL mogelijk! Dit artikel legt uit hoe dit werkt en waar deze techniek toegepast kan worden.

Hans Jongebloed
Dutchear

Een interview van drie uur is natuurlijk nog geen BIG data. Wat nu als je in een call center alle gesprekken met klanten wilt terugzoeken die specifiek gingen over een mislukte aflevering van een bestelling. Stel dat er totaal 200 uur (bedenk hoe lang het een groep mensen kost om dit af te luisteren!) aan gesprekken opgenomen zijn, hoe vind je dan net die 10 – 20 gesprekken die je nodig hebt?

De technieken die in dit artikel worden genoemd zijn zeer waardevol in dit soort gevallen: (heel) veel audio-data en geen tijd/mankracht om alles te gaan afluisteren/uittypen, op zoek naar die ene speld in de hooiberg….

Spraakherkenning methodes voor BIG data

Er is een aantal methodes om een grote hoeveelheid opgenomen spraak te kunnen doorzoeken. We onderscheiden de *'fonetische zoekmethode'* en de 'groot vocabulaire spraakherkenning'. Met de fonetische zoekmethode wordt de spraak herkend met een speciale spraakherkenner die op elk moment in de tijd aangeeft welke klanken (fonemen) er waarschijnlijk zijn uitgesproken. Bijvoorbeeld, er wordt herkend (tussen [] de mogelijke klanken op hetzelfde moment):

[h/w] [A/a:] [l] [o:] [h/w] [A/O] [n/m] [s/f/z]

Korte simpele fonetisch introductie: hoofdletters zijn korte klanken (A uit tak), kleine letters met dubbele punt de lange klanken (a: uit kaas). De medeklinkers worden gewoon geschreven.

Deze fonetische informatie wordt in een database gestopt, zodat er met zoekwoorden kan worden gezocht. Voor elk zoekwoord moet dan wel de fonetische transcriptie bekend zijn. In bovenstaand voorbeeld zijn de volgende woorden terug te vinden:

Hans /hAns/
hallo /hAlo:/
won /wOn/
wal /wAl/
Waal /waal/

Het grote voordeel van deze methode is dat je op elk mogelijk woord kunt gaan zoeken, zonder dat je opnieuw alle spraakopnames hoeft te verwerken met een spraakherkenner. Stel: je wilt het (fictieve) nieuwe zoekwoord 'Lowans' zoeken, dan zou je dit in het voorbeeld kunnen vinden. Het nadeel van de methode is dat het moeilijk te bepalen is welk woord echt gezegd is. Bovenstaande woorden kunnen niet allemaal goed zijn; waarschijnlijk is gezegd 'Hallo Hans' (aangenomen dat de herkende klanken de hele zin vormen).

De andere populaire methode (groot vocabulaire spraakherkenning) lijkt meer op de standaard spraakherkenning in allerlei toepassingen (zoals dicteren). De spraakherkenner is dan voorzien van een groot woordenboek van een bepaalde taal (inclusief de fonetische uitspraak). Tevens 'weet' de spraakherkenner iets over goede zinsopbouw en woordvolgordes door toepassing van een zogenaamd taalmodel. Een taalmodel is een statistisch model waarin veelvoorkomende woordvolgorden een hoge kans krijgen en niet bestaande woordvolgordes een kans van bijna nul. Deze kansen zijn berekend op een grote verzameling teksten die over hetzelfde onderwerp (met hetzelfde woordgebruik) gaan als de audio-opnames. Zodra de audio-opnames zijn herkend door de spraakherkenning (resultaat is tekst) kan op de 'normale' Google manier gezocht worden in het herkenresultaat. Het nadeel van de groot vocabulaire spraakherkenning methode is dat – afhankelijk van de kwaliteit van de audio en het taalmodel – de kans op fouten in de herkenning behoorlijk groot is, waardoor woorden gemist worden

bij het zoeken. Het voordeel van deze zoekmethode is dat de spraakherkenning veel meer rekening houdt met de context van de gezochte woorden en deze daardoor nauwkeuriger te herkennen zijn. Het is daarmee ook mogelijk om zogenaamde 'snippets' te tonen; dat wil zeggen dat je naast het gevonden zoekwoord ook de context laat zien. Als voorbeeld (gezocht is op 'tenlastelegging'):

*… advocaat reageerde op de **tenlastelegging** vanuit het openbaar ministerie …*

Het is uiteraard ook mogelijk om bovenstaande methodes te combineren en de beste elementen van beide te gebruiken. In de volgende paragraaf worden een aantal toepassingen genoemd die gebaseerd zijn op deze technologieën.

BIG data cases met spraakherkenning

Hieronder – ter inspiratie – een variatie aan toepassingen om met spraakherkenning in BIG data te zoeken.

Bij de *Rechtbank van Amsterdam* kunnen de griffiers op basis van hun (getypte/handgeschreven) aantekeningen van een zitting nog eens terug luisteren welke uitspraken ter zitting werden gedaan. Hiermee is het mogelijk om een veel betere kwaliteit van de verslagen te realiseren. Ook andere soorten *vergaderingen* (bedrijven, gemeentes) worden door Dutchear verwerkt.

In *call centers* wil men graag weten of de medewerkers zich aan de regels (script) houden en waarover de klanten van een bedrijf vaak bellen. Met de genoemde methodes is het mogelijk om op specifieke woorden en woordcombinaties te zoeken en ook een classificatie van gesprekken te maken. Het aantonen of medewerkers zich aan de regels houden, is ook voor de overheidsautoriteiten met de spraakherkenningsmethodes toegepast.

Zeer grote verzamelaars van BIG data zijn ongetwijfeld de *politie* en de *inlichtingendiensten*, aangezien Nederland bekend staat als tapland nummer 1 (bron: de Volkskrant 23 mei 2012). In samenwerking met Fox-IT en NFI heeft Dutchear aangetoond dat spraakherkenning hierin een meerwaarde heeft. In de vorige editie van DIXIT is deze case van Dutchear verder toegelicht. Naast telefoontaps kan hier ook gedacht worden aan *afluisteropnames* en *verhoren*.

Naast telefonie zorgt ook de media voor steeds meer data: *radio, televisie, YouTube*. Een leuke toepassing is het laten zoeken op voor jou persoonlijk relevante termen die, nadat ze 's avonds genoemd zijn, de volgende ochtend in e-mail verschijnen met links naar de relevante radio/tv fragmenten.

Conclusie

De genoemde cases in de inleiding - zowel op kleine schaal (interviews), als op grote schaal (call center opnames) – kunnen met spraakherkenning toegankelijk en doorzoekbaar worden gemaakt. Hierdoor is het niet nodig om door mensen alles uit te laten luisteren (wat vaak onbetaalbaar is..).

“De meerwaarde zit in het slim combineren van databronnen”

Harrie Vollaard is innovatiemanager bij de Rabobank. Zijn zes man sterke team heeft als doelstelling om proactief relevante ontwikkelingen in de maatschappij en de technologie te identificeren en daarop in te springen. Sinds vorig jaar houdt het team zich intensief bezig met Big Data en de eerste pilotprojecten zijn inmiddels een feit.

Leonoor van der Beek
bol.com

Geïnterviewde:
Harrie Vollaard
Innovatie-manager Rabobank

Plofkraken

“Ik was heel verbaasd te horen dat vrijwel geen van de bezoekers van het VINT congres over Big Data zelf al echt iets doet op dit gebied. Mijn team heeft zich vorig jaar verdiept in het onderwerp en het papierwerk gedaan, dit jaar doen we onze eerste proefprojecten, en volgend jaar willen we het dan ook echt gaan inzetten. We zijn begonnen met werken met interne Rabobank-data. Zo hebben we op basis van het volume van transacties geanalyseerd welke betaalautomaten het meest effectief zijn. En dan gaat het niet alleen om aantal transacties, maar ook om aantal verschillende gebruikers: we vonden ook een betaalautomaat die bijna voor één persoon in stand gehouden werd - zo vaak maakte die persoon er gebruik van.

Daarna zijn we ook met externe en minder gestructureerde databronnen gaan werken. Zo hebben we plofkraken geanalyseerd op hun ligging (afstand van snelweg), de urbanisatiegraad en de weersomstandigheden. We weten nu dat de meeste plofkraken voorkomen in rurale gebieden dichtbij de snelweg en op dagen dat het niet glad is maar wel mistig.

In de toekomst willen we bijvoorbeeld kennisprofielen bouwen van onze medewerkers. Niet op basis van een profielpagina die ze zelf invullen – want we weten inmiddels wel dat mensen zo’n pagina heel slecht bijhouden – maar op basis van hun output. En die data is voornamelijk ongestructureerd.”

Big Data vs. BI

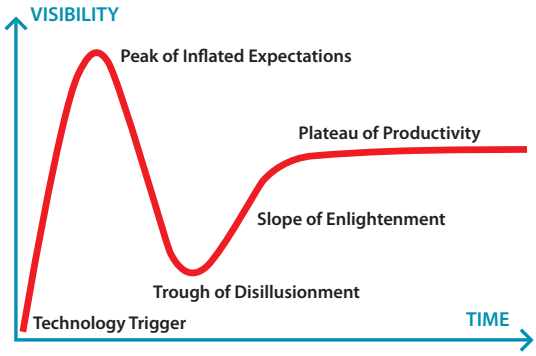
“De hoeveelheid data die wij tot onze beschikking hebben neemt snel toe, maar is natuurlijk nog relatief klein in vergelijking met de webgiganten zoals Google of Yahoo. Wat wel opvalt, is dat die hoeveelheid jarenlang min of meer lineair gestegen is, maar sinds een jaar of 2, 3 ineens exponentieel stijgt. Zeker wanneer je ook nog verschillende typen data met elkaar wilt gaan combineren, red je dat niet meer in een gewone Business Intelligence (BI) omgeving. In een BI systeem ontwerp je een database op basis van vooraf gedefinieerde ideeën over de vragen die je eraan wilt kunnen stellen. Maar als je een heel nieuw idee hebt, of een nieuwe databron, dan kun je daar niet direct iets mee. Dan moet je eerst weer een nieuwe database of een wijziging in een bestaande database aanvragen, dat moet dan gebouwd worden, en dan ben je een hoop tijd kwijt voordat je überhaupt kan beginnen te kijken of je idee zinnig was.

We hebben daarom een apart Hadoop-cluster opgezet, waar heel gemakkelijk nieuwe databronnen ingehangen kunnen worden en die efficiënt bevraagd kan worden. Een vraag die vroeger misschien tien uur duurde om beantwoord te worden, duurt nu nog 3 minuten. Daar wil je nog wel op wachten, om eventueel je vraag wat aan te passen en opnieuw te stellen. Dan wordt ineens exploratief onderzoek mogelijk. Maar het meest interessante is toch wel het combineren van verschillende databronnen. Daar zit het meest waardevolle in voor onze organisatie.”

Sentiment mining

“Social media genereren veel data waar snel op gereageerd moet worden, dus in die zin valt sentiment mining onder Big Data. Bij de Rabobank is er echter al enige tijd een webcare-team operationeel. Dat bekijkt het sentiment op bijvoorbeeld Twitter en Facebook vanuit imago-oogpunt. Sentiment mining helpt ons problemen vroeg te signaleren en biedt de mogelijkheid om, bijvoorbeeld in het geval van een storing in internetbankie-

ren, direct te laten weten dat we ermee bezig zijn. Het is wel interessant hoe zich dat gaat ontwikkelen als we meer willen gaan doen met die data. Je kan dan je niche-oplossing oprekken en bijvoorbeeld social-media data importeren in je CRM-systeem. Maar ik verwacht dat wanneer Big Data een integraal onderdeel van de de organisatie geworden is, deze de niche-oplossingen zal opslokken, zodat alle data in één geïntegreerde omgeving beschikbaar is."



Privacy en Security

"Big Data vereist extra inspanning op het gebied van security. In de verschillende applicaties is security bijvoorbeeld netjes geregeld door rolscheiding: verschillende gebruikerstypen hebben verschillende rechten. Maar wanneer je data uit al die applicaties gaat combineren in je Big Data-cluster, dan kan dat niet meer. Voor mijn team zijn dan ook speciale voorzorgsmaatregelen getroffen. We zitten in een afgescheiden omgeving, alle teamleden zijn gescreend, en we werken met tijdelijke datasets. Wanneer een van onze pilots in productie genomen wordt, moet daar opnieuw uitgebreid naar gekeken worden.

Daarnaast is er het privacy-vraagstuk. Dat is meer een emotioneel probleem dan een juridisch probleem. Wettelijk is er niets op tegen om publieke tweets van klanten te minen en in het CRM-systeem op te slaan, maar klanten vinden het niet OK als ze bellen en er door de agent gerefereerd wordt aan informatie uit die tweet. Het maakt bovendien nogal uit waarvoor de data gebruikt wordt. Als het voor fraudedetectie is, heeft men er meer begrip voor dan wanneer het voor commerciële aanbiedingen gebruikt wordt. Jeffrey Rayport heeft een richtlijn opgesteld die ik helemaal onderschrijf. Hij stelt dat je gebruikers direct duidelijk moet maken welke data verzameld wordt, dat gebruikers moeten kunnen kiezen voor bepaalde niveaus van privacy, en er voor de klant ook waarde moet zitten in het verstrekken van gegevens, zoals bijvoorbeeld bij de herkenning van fraude. Daarnaast moet privacy in de gehele productcyclus, dus vanaf het eerste design, een onderwerp van aandacht zijn."

Toekomst

''Als je naar Gartner's hype cycle kijkt, dan zit Big Data nog in de eerste fase, die van de belofte [1]. De desillusie moet nog komen. Maar daarna, in een jaar of vijf, zal Big Data een integraal onderdeel vormen van iedere organisatie. In de projecten zullen steeds meer verschillende soorten data gecombineerd worden en veel daarvan zullen ongestructureerde data zijn. Een project dat ik zelf graag zou doen is het automatisch herkennen van *social engineering*, het ontfutselen van data voor fraudedoeleinden. Daar ligt zeker een toepassing van taal- en spraaktechnologie."

[1] *http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp*

‘Best’ is lang niet altijd goed

Als paddenstoelen schieten bedrijven uit de grond die automatisch analyseren of er positief of juist negatief over een bedrijf, een product, een politieke beweging of actuele kwestie gesproken wordt. Marketeers publiceren op basis van deze data persberichten en rapporten met klinkende koppen als ‘Mazda het meest positief beoordeelde automerk’ of ‘Telfort meest negatieve merk’. Maar hoe betrouwbaar zijn die gegevens eigenlijk? Over de zin en onzin van sentimentsanalyse.

Leonoor van der Beek
bol.com

Impact van social media

Individuen, politici en bedrijven hebben altijd graag willen weten hoe anderen over hen denken. Maar de afgelopen jaren is er ineens heel veel data beschikbaar gekomen waarin gezocht kan worden naar wat men over jouw nieuwe kapsel, verkiezingsprogramma of product zegt. Data uit online enquetes bijvoorbeeld. Het is voor een groot bedrijf niet moeilijk om bezoekers van een bepaalde pagina te vragen een paar korte vragen te beantwoorden, en zo heel snel enkele tienduizenden reacties te verzamelen. Maar ook data van sociale media. Op ieder moment kan door eenieder op blogs, Facebook of Twitter over je geschreven worden. De hoeveelheid data is enorm en het belang om de data snel te herkennen als positief of negatief is groter dan ooit tevoren: een negatief verhaal op Twitter tot enorme imagoschade leiden. Bekend is het voorbeeld van McDonalds, dat had gehoopt met de hashtag ‘#McDStories’ een trending topic te maken van positieve ervaringen, maar in plaats daarvan een stroom aan negatieve publiciteit ontketende [1]. Positieve voorbeelden zijn er ook: de populaire blogger Peter Shankman gaf een enorme boost aan het imago van Morton’s Steakhouse door uitgebreid te publiceren over de geweldige ervaring die hij met het restaurant had [2]. Automatische sentimentsanalyse vaart wel bij deze combinatie van enorme hoeveelheden data en de noodzaak deze snel te analyseren.

Goedkoop

Sentimentsanalyse is echter niet eenvoudig. Niet alleen zijn er de beruchte, complexe gevallen, zoals de ironie in ‘Dus nu krijg ik iedere keer dat ik in wil inloggen een nieuw paswoord. Geweldig!’. Ook rechttoe-rechtaan zinnen zijn niet altijd gemakkelijk te herkennen als positief of negatief. Woorden lijken misschien duidelijke indicators voor positief of negatief sentiment, maar zijn dat vaak niet. Is ‘goedkoop’ een indicator voor positief sentiment? Misschien wel als het over benzine gaat, maar niet als het over een grap (of over je dochter) gaat. En ‘duidelijker’? Misschien wel in ‘het nieuwe schema is een stuk duidelijker’, maar niet in ‘dat had best wat duidelijker gemogen’. En wie denkt dat ‘best’ misschien wel het meest positieve woord is dat er bestaat, die moet het laatste voorbeeld maar eens herlezen.

Als het al lukt om van een woord te bepalen of het positief of negatief is, dan zijn er nog allerlei constructies die deze waarde veranderen. Versterkende woorden (heel, erg, bijzonder), verzwakkende woorden (beetje, ietwat, tikkeltje) en ontkenningen (niet, geen, nooit) kunnen vaak nog min of meer in regels gevat worden – al is dat met name voor negatie alles behalve eenvoudig. Maar wat te denken een van voorbeeld als ‘dat laat nog wel wat ruimte voor verbetering’? Hoe dit geautomatiseerd te herkennen als een negatieve opmerking, ondanks het positieve woord ‘verbetering’?

Taalspecifiek

Taaltechnologie ondervangt deze moeilijkheden tot op zekere hoogte, al ligt de precisie van sentimentanalysesystemen zelden boven de 80%, en wordt soms zelfs de 70% niet gehaald – cijfers die voor andere taken volstrekt onaanvaardbaar zouden zijn. Belangrijk is het dat resultaten in de ene taal (vaak het Engels) niets hoeft te zeggen over de performance in een andere taal, zoals het Nederlands. Want veel van de uitdagingen op het gebied van sentimentsanalyse zijn taalspecifiek. Waar het Engels bijvoorbeeld heel handig de bijwoorden van bijvoeglijk naamwoorden onderscheidt door het achtervoegsel –ly (‘awful’ vs. ‘awfully’), doet het Nederlands dat niet. Zonder Part-of-Speech-tagger is het dan ook lastig om het negatieve ‘Het is verschrikkelijk wat er met dat leuke meisje gebeurd is’ te onderscheiden van het positieve ‘Ik vind dat meisje verschrikkelijk leuk’.
Sowieso heeft het Nederlands veel meerduidige woorden. Vergelijk ‘Wat een stuk!’ met ‘Mijn auto is stuk’, en ‘De lucht is al een stuk lichter geworden’. Dit wordt nog versterkt door het voorkomen van scheidbare prefixen. ‘Aan’ kan van het werkwoord ‘aanraden’ gescheiden worden in bijvoorbeeld ‘Ik raad

het je van harte aan'. Maar hoe dit positieve 'raden' te onderscheiden van het neutrale 'Kun jij raden aan wie zij de envelop gaat geven'? Hiervoor is specifieke kennis van het Nederlands nodig.

Kortom: het automatische bepalen van sentiment is een heel moeilijke taak, waarbij slechts matige resultaten worden behaald, en waarbij resultaten uit de ene taal geen enkele garantie geven voor precisie in een andere taal. Maar zelfs als het mogelijk is om correct te bepalen of een zin overwegend positief is, of overwegend negatief, dan nog is de vraag in hoeverre dat iets zegt over het imago van een bepaald merk.

Interpretatie

Automatische sentimentsanalyse kan plaatsvinden op het niveau van de phrase, de zin of het document. Een document is vaak een combinatie van positieve en/of negatieve uitlatingen over verschillende onderwerpen, wat analyse lastig maakt. Analyse op phraseniveau – in theorie het nauwkeurigst – breekt de zin op in stukken die los van elkaar geclassificeerd worden, wat de leesbaarheid en interpreteerbaarheid niet ten goede komt. Er wordt dan ook vaak gekozen voor classificatie op zinsniveau. Wanneer een marketeer nu wil weten hoe positief er over een merk gesproken wordt, dan worden alle zinnen opgehaald waarin de productnaam voorkomt, waarna geteld wordt hoe vaak die positief, negatief of neutraal waren. Maar deze resultaten hoeven niets te zeggen over het sentiment rondom een merk.

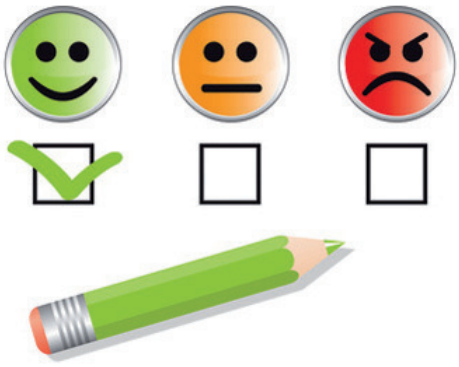
Allereerst is soms moeilijk te garanderen dat de zoekterm ook echt verwijst naar het merk. Wanneer supermarkt Deen tweets gaat analyseren, zullen zij ook tweets tegenkomen echt niet over de supermarkt gaan:

@_MichelH nopp was hier wel een lekkere alleen was een deen daar heb ik niks aan ;p

Daarnaast garandeert het voorkomen van een merk in een zin nog niet dat de polaire (positieve of negatieve) uitdrukking ook over dat merk gaat. Om bij onze supermarkt te blijven:

Net f die starbucks ding gehaald bij deen. Heb er één woord voor: GATVERDAMME.

Tenslotte kan het nog zijn dat de polaire uitdrukking wel over jouw merk gaat, maar toch eigenlijk weinig zegt over de waardering van jouw product. Zo scoorde Carglass in een recent onderzoek bijzonder slecht, maar bij nader onderzoek bleek een zeer groot gedeelte



van de negatieve opmerkingen te gaan over de als irritant ervaren televisiespot [3]. Zullen mensen om die reden niet naar Carglass gaan wanneer zij een sterretje in hun voorruit hebben? Ik betwijfel het.

In de bovenstaande drie gevallen is de sentimentsanalysesoftware niets te verwijten. Elk van de drie uitlatingen was negatief en bevatte de gewenste zoekterm. Geen van de drie voorbeelden moet echter geïnterpreteerd worden als onvrede met het aangeboden product of service.

Wanneer sentimentsanalyse zulke matige resultaten behaalt, en zelfs bij correcte analyse de resultaten niet per se direct bruikbaar zijn voor het monitoren van sentiment rondom een merk, heeft het dan helemaal geen zin om sentimentsanalyse toe te passen? Toch wel. Sentimentsanalyse kan de hoeveelheid potentieel relevante data enorm verkleinen tot een hoeveelheid die handmatig de verwerken is, en alvast grove trends laten zien. Bovendien kunnen met Machine Learning-technieken nog wel iets betere resultaten behaald worden. Performance is erg afhankelijk van de toepassing, maar op de Internet Movie Database (IMDb) worden accuratessecijfers tot 90% gerapporteerd (dit gaat dan wel ten koste van de controle: de mogelijkheid om met handgeschreven regels opvallende en storende classificatiefouten te ondervangen) [4]. Maar ook met die betere cijfers zal altijd iemand met gezond verstand naar de data moeten kijken om de geïdentificeerde trends te verifiëren en te duiden.

1) *www.forbes.com/sites/kashmirhill/2012/01/24/mcdstories-when-a-hashtag-becomes-a-bashtag*
2) *www.huffingtonpost.com/2011/08/18/peter-shankman-mortons-steak-tweet_n_930744.html*
3) *www.slideshare.net/Greenberrynl/top-100merken-whitepaper-greenberry*
4) *www.ijarcsse.com/docs/papers/June2012/Volume_2_issue_6/V2I600263.pdf*

Bouwstenen voor het Analyseren van BIG DATA

Big Data is mainstream geworden. De New York Times analyseert hoe de term Big Data de sprong heeft gemaakt van technologie kringen naar het bedrijfsleven, de media en het grote publiek [1]. Volgens de krant is 2012 het jaar waarin Big Data echt groot wordt. De term mag dan veel gehypet zijn en zelfs opduiken in lijstjes van meest vervelende management buzz-words, de groeiende hoeveelheid beschikbare gestructureerde en ongestructureerde data is een feit en de potentiële waarde van deze data wordt steeds breder erkend. Maar waar zit die waarde precies in?

Thijs Westerveld
Teezir Search Solutions

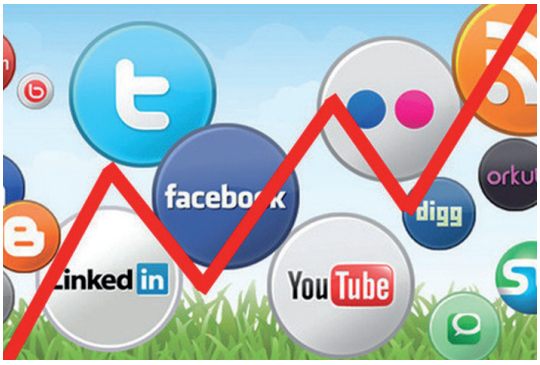
Combineren en aggregeren

De term Big Data refereert slechts aan de omvang van datasets. Het kan gaan om grote databases met biochemische of astronomische data, om RFID informatie, om e-commerce transacties of om video archieven. Maar een groot deel van de Big Data discussies gaat over gebruikersinteractie op social media en andere online platformen. Informatie gedeeld op sociale netwerken, blogs en forums, en interactiepatronen tussen gebruikers op die platformen vormen een rijke bron van informatie. Zeker als je die informatie weet te combineren. Combinaties binnen een platform kunnen al leiden tot interessante inzichten, maar het wordt pas echt interessant als je informatie uit verschillende bronnen weet te combineren.

Een analyse van alle tweets rondom project-X Haren laat bijvoorbeeld zien op welk moment de sfeer omsloeg. Aan het begin van de avond wordt er vooral getweet over 'treinen' en 'de straat van Merthe', later op de avond gaat het vaker over 'mobiele eenheid', 'charges' en 'relschoppers' [2]. Wanneer we deze informatie combineren met de profiel-informatie van deze twitteraars kunnen we bijvoorbeeld het twittergedrag van verschillende demografische groepen onderscheiden. Twitteren jongens over andere onderwerpen dan meisjes? Zijn er verschillen tussen de tweets van de feestgangers in Haren en die van thuisblijvers? Of tussen die van twitteraars die de film gezien hebben en de rest? Hoe meer we weten van de personen achter de tweets, hoe rijker de informatie.

Bouwstenen

Teezir biedt haar klanten de bouwstenen om de hierboven beschreven analyses uit te voeren. Via de Analystics dashboards en API hebben klanten toegang tot ruim 200 miljoen Nederlandstalige documenten verzameld van social media, blogs, forums, en nieuwssites. Dagelijks komen daar bijna een half miljoen nieuwe documenten bij. De dashboards en API bieden klanten standaard de mogelijkheid te filteren en aggregeren op sentiment, bron, publicatietijdstip, auteur of onderwerp en geven bovendien de mogelijkheid in te zoomen tot op het niveau van individuele berichten. Veel klanten hebben genoeg aan die basisfunctionaliteit, maar de echte Big Data pioniers combineren deze informatie met hun eigen data op een slimme manier. Ik noem een paar voorbeelden.



Voorbeelden

Het webcare-team van een telecomaanbieder vraagt klanten die via Twitter of Facebook klagen of een probleem rapporteren om hun klantgegevens. Op deze manier kunnen ze de problemen van klanten in veel gevallen snel oplossen. Bovendien zijn hiermee de social media accounts gekoppeld aan de klanten-database. Monitoren van deze social media accounts levert meer informatie over de klanten. Dit maakt het mogelijk om een klant aanbiedingen op maat te doen bij eventuele contractverlenging (een gratis iPhone voor de Apple fan, maar liever een Android telefoon voor de Apple hater). Ook ziet het webcare-team in het vervolg de context van een klantbericht en is daarmee beter in staat in te spelen op de specifieke klantsituatie. Een uitgebreidere analyse van alle social media berichten van alle klanten helpt het bedrijf een vinger aan de pols te houden bij wat er leeft onder haar klanten.

Een marktonderzoeksbureau integreert traditioneel panelonderzoek en online analyse. Door haar panelleden om hun social media accounts te vragen zijn ze in staat twitteraars en Facebook gebruikers te groeperen op basis van profielinformatie verkregen uit traditionele vragenlijsten. In die vragenlijsten wordt het panel bijvoorbeeld expliciet gevraagd hoe vaak ze vliegen en wat ze van KLM vinden. Door de online verzamelde social media berichten rondom KLM te groeperen op basis van deze auteursinformatie is het mogelijk om onderscheid te maken tussen de tweets van twitteraars binnen en buiten het panel en bovendien om de tweets van panelleden onder te verdelen in tweets van fans of frequent flyers en overige tweets. Dit geeft allerlei inzichten die zonder de combinatie van online en traditioneel onderzoek niet beschikbaar waren. Twitteren frequent flyers vaker over KLM? Zijn de zelfverklaarde fans ook online positiever over de luchtvaartmaatschappij?

Toekomst

Big Data is meer dan een hype. De term zal worden ingehaald door het volgende buzzword, maar de informatiestroom zal blijven groeien en de markt vraagt om steeds meer en steeds uitgebreidere analyses en om steeds snellere resultaten.

Onze klanten vragen meer en meer om een brede dataset. Sinds kort combineert Teezir de analyses van online en offline bronnen.

Op deze manier kan bijvoorbeeld de discussie rondom project-X Haren verder geanalyseerd worden. Hoe was de interactie tussen online en offline bronnen? Na welke krantenberichten zien we de grootste pieken in de communicatie op Twitter en Facebook? Naast offline content zijn ook andere bronnen interessant, zoals radio en televisie feeds of afgeschermde klant-specifieke datastromen (e-mail, inkomende calls of interne forums).

Daarnaast horen we steeds vaker de roep om meer metadata rondom de auteur achter een bericht. Wie is die auteur? Welke groepen spreken over mijn merk? Of waar spreken mijn klanten over? Een bedrijf als Teezir zal steeds meer data en metadata verzamelen en de benodigde bouwstenen voor de analyses blijven doorontwikkelen.

[1] *www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html?ref=technology&_r=0*

[2] *http://storify.com/setuputrecht/harenhackathon-live-blog*

BIG DATA en TST

Deze DIXIT wordt u aangeboden door de deelnemers van NOTaS:

- CLST, Radboud Universiteit Nijmegen
- Data Archiving and Networked Services
- Dedicon
- Dutchear
- Gridline
- Knowledge Concepts
- Lexima
- Linguistic Systems BV
- Nederlands Forensisch Instituut
- OSTT, Sint Maartenskliniek
- Readspeaker
- RightNow Development, Oracle
- TeleCats
- TST-Centrale
- Universiteit van Tilburg, CIW/TiCC
- Universiteit Twente, HMI
- Universiteit Utrecht, Utrecht Institute of Linguistics OTS

"Het is moeilijk de impact van BIG DATA te overschatten"

Michiel Boreel is CTO van ICT-bedrijf Sogeti, dat met haar trendlab VINT opvallende bewegingen in de wereld van de ICT op de voet volgt. In juni 2012 publiceerde het bedrijf het onderzoeksrapport 'Creating clarity with Big Data' over de kansen en uitdagingen van werken met grote hoeveelheden diverse data [1].

Leonoor van der Beek
bol.com

Geïnterviewde:
Michiel Boreel
Sogeti

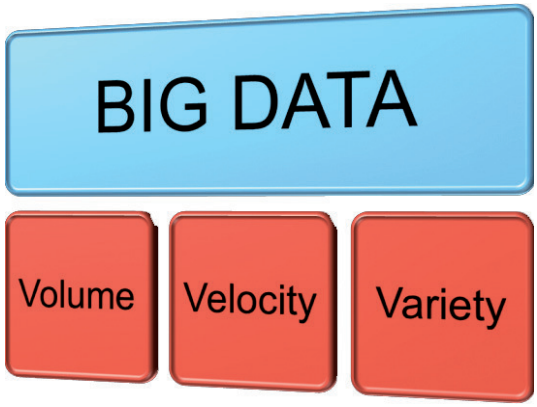
Volume, variety, velocity

"De hoeveelheid data die beschikbaar is voor analyse groeit exponentieel. Op dit moment gaat de aandacht erg uit naar de grote hoeveelheid data uit sociale media, maar in de toekomst gaan de volumes nog veel harder groeien wanneer meer devices uitgerust worden met sensoren, die vrijwel continu informatie zullen registreren. Maar Big Data gaat niet alleen om volume. De drie kernbegrippen zijn *volume*, *variety* en *velocity*. Er komt niet alleen méér data beschikbaar, ook steeds meer verschillende soorten data, en er zijn steeds meer mogelijkheden die aan elkaar te koppelen. Voor een gemiddeld bedrijf is de meerderheid van die data ongestructureerd en afkomstig van externe bronnen. Vroeger werd die data opgeslagen in een database om later eens een rapport uit te genereren, maar nu moet die data real-time geanalyseerd worden, zodat er direct op gereageerd kan worden.

Het is niet zo dat Big Data bestaat uit één nieuwe technologie of uitvinding. Het is een verzameling van ontwikkelingen en technologieën die op dit moment samenkomen en momentum krijgen. Tezamen bewerkstelligen zij een fundamentele verandering. Het gevoel is hetzelfde als rond 1994, aan de vooravond van de doorbraak van het internet. Niemand wist precies wat er zou gaan gebeuren, maar het was wel duidelijk dat er iets heel wezenlijk zou gaan veranderen. Datzelfde gevoel, diezelfde urgentie, die voelen we nu ook. Maar ook nu is nog niet helemaal duidelijk welke toepassingen er zullen komen. Wel is duidelijk dat de huidige praktijk van met grote hoeveelheden data omgaan, namelijk door middel van samples, niet voor iedere onderzoeksvraag geschikt is. Voor sommige vragen moet je echt de hele populatie bekijken. Een voorbeeld is detectie van fraude, waarbij de signalen elk voor zich zo klein en infrequent zijn, dat de fraude met sampling niet opgemerkt zou worden."

Uitdagingen

"Vrijwel iedere organisatie is op een of andere manier bezig met Big Data. Iedereen realiseert zich dat data en de technologie om inzicht uit data te halen cruciaal zijn in de concurrentiestrijd. Maar de meeste organisaties zijn eigenlijk nog helemaal niet klaar voor Big Data. Ze hebben de middelen vaak nog niet om kennisextractie toe te passen. Maar vooral: de stijl van managen op intuïtie is zo ingesleten in organisaties, dat zelfs wanneer de mogelijkheden er zijn om beslissingen te nemen op basis van feiten, op basis van informatie, die kans niet altijd wordt aangegrepen en er nog steeds besloten wordt op basis van *gut feeling*. Nog interessanter wordt het wanneer Big Data zover is dat op basis van Machine Learning kennis wordt verworven die niet te begrijpen is. Gaan managers daar dan hun beslissingen op baseren? Maar we staan nog pas aan het begin van de Big Data beweging – ik verwacht dat het nog wel tien tot vijftien jaar duurt voordat deze beweging voorbij is.



Er zijn een aantal grote hobbels te nemen voordat Big Data een integraal onderdeel is van iedere organisatie. Allereerst de technologie en computing power die nodig is om die enorme hoeveelheden data te verwerken. Er zullen nieuwe manieren gevonden moeten worden om de inzichten die hieruit gewonnen worden te presenteren: een nieuwe vorm van datavisualisatie. Daarnaast vormt de menskracht die Big Data moet gaan realiseren een behoorlijke uitdaging: er zal een grote vraag komen naar data scientists, mensen die kennis hebben van databronnen en hun kenmerken, die statistisch goed

onderlegd zijn, en die de uitkomsten van machine learning algoritmen kunnen terugvertalen naar inzichten en business benefits. De verandering in managementstijl hadden we al genoemd als uitdaging. En dan is er nog privacy. Consumenten ervaren Big Data al snel als Big Brother – maar aan de andere kant geven ze zonder blikken of blozen allerlei persoonlijke informatie vrij. De regelgeving is nog niet compleet op dit vlak. Ik verwacht dat er een manier komt waarop gebruikers zelf kunnen managen welke informatie voor wie beschikbaar en toegankelijk is en voor welke doeleinden."

Financiële impact

"Het is moeilijk de impact van Big Data te overschatten. De voorspellingen uit het rapport – bijvoorbeeld 300 miljard dollar winst in de Amerikaanse gezondheidszorg, of 250 miljard euro voor Europese overheden – komen van McKinsey. Maar wat ze ook zeggen: het zal te weinig blijken. Ik trek opnieuw de vergelijking met de opkomst van het internet. Toen werd voorspeld dat er wel een miljard mee verdiend kon worden, en sommige heel moedige mensen schatten het potentieel in op tien miljard. We zitten nu in eenzelfde situatie. De precieze winst zal per industrie verschillen, maar enorme efficiencyslagen kunnen gemaakt worden, en met concurrentie op basis van inzicht uit Big Data kan heel veel verdiend worden."

Michiel Boreel

Rol TST

"Spraakherkenning is feitelijk een vorm van patroonherkenning, en dat is dé methode van analyse van Big Data. Dat matcht dus direct. Spraakherkenning en tekstanalyse zijn belangrijk in de analyse van ongestructureerde of semigestructureerde data. Maar taal- en spraaktechnologie kunnen ook een rol spelen in de interface tussen Big Data en de gebruiker. We zullen steeds meer naar natuurlijke interfaces toegaan. Taal en spraak zouden daar een belangrijke rol kunnen gaan spelen.

We staan aan de vooravond van een grote doorbraak, waarvan niemand nog precies kan zien waar hij naar toe zal leiden. We zullen verbaasd worden door de snelheid waarmee groter inzicht tot nieuwe producten kan leiden. Maar niemand heeft het ultieme antwoord op dit moment. Ik zou graag horen wat anderen hierover denken, en nodig iedereen uit zijn ideeën te delen op de onze blog over Big Data op www.sogeti.com/vint/bigdata/questions."

[1] *http://blog.vint.sogeti.com/wp-content/uploads/2012/06/Creating-clarity-with-Big-Data-VINT-Research-Report.pdf*

Leren van regelmatigheden in grote hoeveelheden natuurlijke taaldata

Een grammatica van een taal is een heel nuttig hulpmiddel voor taalkundigen, taalbeheersers en taaltechnologen. Maar het ontwikkelen van een grammatica is een behoorlijke klus. Hiervoor zijn experts nodig die kennis hebben van de taal die beschreven moet worden, van het taalkundig formalisme waarin de taal beschreven wordt, en van de tools die het mogelijk maken de grammatica te schrijven en te evalueren. Het is dan ook interessant om te onderzoeken of deze grammatica's helemaal of gedeeltelijk automatisch gegenereerd zouden kunnen worden op basis van heel grote hoeveelheden natuurlijke taaldata.

Menno van Zaanen
Universiteit van Tilburg

Structuur

Structuur is een belangrijk onderdeel van taal: het bepaalt bijvoorbeeld in welke volgorde we letters of morfemen tot woorden kunnen vormen of hoe we woorden kunnen samenvoegen tot frasen of zinnen. Grammatica's worden gebruikt voor het beschrijven van deze structuren en kunnen gebruikt worden om de structuur van woorden en zinnen door een automatische analyse te achterhalen. Hoe een analyse er precies uit ziet is onder meer afhankelijk van de achterliggende taalkundige theorie en het formalisme op basis waarvan de grammatica is geschreven. De resulterende beschrijving van de structuur van een sequentie kan voor veel taken gebruikt worden. De structuur van woorden bijvoorbeeld is soms nodig om te bepalen hoe een woord uitgesproken moet worden en de structuur van een zin kan helpen om de betekenis ervan te achterhalen.

Inductie

Grammatica-inductie (GI), ook wel grammatica- of grammaticale inferentie genoemd, is een alternatief voor het lastige en tijdrovende handmatige ontwikkelen van grammatica's door experts. GI probeert het ontwikkelen van een grammatica op een andere manier aan te pakken: het onderzoekt de (on)mogelijkheid van het efficiënt automatisch leren van talen en de daarbij behorende grammatica's. Het onderzoeksgebied van GI kan verdeeld worden in twee delen: formele en empirische GI.

Formele GI onderzoekt de leerbaarheid van klassen van talen (zoals contextvrije talen) op een mathematische manier. De resultaten zijn wiskundige bewijzen die laten zien of formele talen uit een specifieke klasse wel of niet efficiënt (in de betekenis van tijd complexiteit) leerbaar zijn.

Het uitgangspunt bij empirische GI is dat natuurlijke talen efficiënt leerbaar zijn (want mensen kunnen dat). Het onderzoek richt zich dan ook op het ontwikkelen van implementeerbare systemen die structuur (delen van een grammatica) zoeken in ongeannoteerde voorbeeldsequenties. De geleerde structuur zou overeen moeten komen met de structuur die taalkundigen aan de sequenties toe zouden kennen.

Geannoteerde data

Bij het ontwikkelen van empirische GI systemen willen we natuurlijk weten hoe goed ze zijn. Er zijn verschillende manieren om de evaluatie uit te voeren, maar de meest gebruikte aanpak is om de uitvoer van een systeem te vergelijken met een gouden standaard, een door taalkundigen geannoteerde dataset. Op basis hiervan kunnen maten van precisie en compleetheid berekend kunnen worden. Helaas zijn deze geannoteerde datasets beperkt in omvang en beschikbaarheid, omdat het

annoteren van taalkundige data tijdrovend is en gedaan moet worden door experts op het gebied van de gebruikte taalkundige theorie en formalisme.

De huidige generatie empirische GI-systemen (ABL (Alignment-Based Learning), ADIOS, CCM-DMV, EMILE, en U-DOP) zijn relatief 'duur' in rekenkracht. Omdat daarnaast de evaluatie plaats vindt op geannoteerde datasets van beperkte omvang, concentreren de meeste systemen zich op het leren van structuur vanuit een beperkte hoeveelheid data. Deze systemen proberen dan ook zo snel mogelijk structuur te leren. Hierbij wordt mogelijk structuur geïntroduceerd waarvan het systeem nog onzeker is. Immers, om structuur uit kleine hoeveelheden data te kunnen leren, zullen generalisaties op basis van kleine hoeveelheden bewijsmateriaal gemaakt moeten worden.
Het leren uit weinig data heeft nadelen. Natuurlijke taal volgt namelijk voor een groot gedeelte de wet van Zipf: er zijn maar een paar woorden of constructies die heel vaak voorkomen, terwijl er heel veel zijn die maar zelden voorkomen. Dit wordt ook wel de 'long tail' genoemd. Deze eigenschap heeft directe invloed op de werking van GI-systemen: er is slechts weinig data nodig om informatie te verzamelen over woorden die vaak voorkomen, maar er is heel veel data nodig om ook eigenschappen van woorden die zelden voorkomen te kunnen leren.

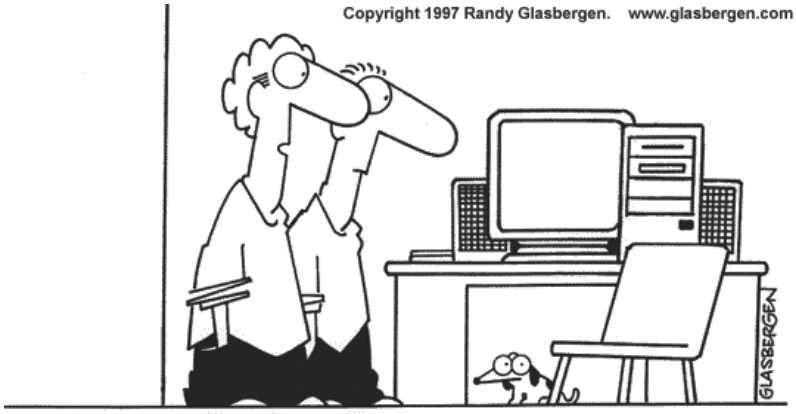
Ongeannoteerde data

Er is een simpele oplossing voor de problemen met het gebruik van beperkte hoeveelheden data: gebruik meer data. Empirische GI-systemen kunnen dan structuur leren op basis van grote hoeveelheden ongeannoteerde data, waarna met de geleerde kennis een kleine hoeveelheid data geannoteerd wordt voor evaluatie. Hierdoor is het mogelijk de systemen te evalueren zonder dat daar grote hoeveelheden handmatig geannoteerde data voor nodig zijn.

De vraag is nu of deze grote hoeveelheden data wel beschikbaar zijn. Op dit moment komen op verschillende plaatsen grote hoeveelheden teksten elektronisch beschikbaar. Denk hierbij aan het scannen en OCRen van bestaande papieren documenten, zoals kranten of boeken, maar ook sociale media genereren veel data. Op Twitter, een microblog-applicatie, worden per uur miljoenen tweets (korte tekstberichten) verstuurd. Deze zijn voor een groot deel te downloaden.
Het leren van structuur in grote hoeveelheden data levert wel een aantal praktische problemen op: de data moet ergens opgeslagen worden. Hiervoor is veel opslagruimte nodig, die bovendien efficiënt toegankelijk moet zijn en die (vanwege repliceerbaarheid van experimenten) gearchiveerd moet kunnen worden. Om de data te kunnen kopiëren of verplaatsen zijn ook snelle verbindingen tussen computers nodig.

Gedistribueerde systemen

Empirische GI-systemen bouwen intern een model van de taal op. Bij grotere hoeveelheden data zal het model complexer worden. Dit vraagt om veel grotere intern geheugens, maar ook om snellere computers. Een mogelijke richting om dit probleem op te lossen is om gedistribueerde systemen te gebruiken. Het probleem van het leren van structuur wordt dan in kleine problemen opgesplitst, dat elk afzonderlijk, op verschillende computers, aangepakt kan worden.



"This computer is very fast! It became outdated faster than any computer I've ever owned."

Het ontwikkelen van gedistribueerde systemen vereist kennis van de technische (on)mogelijkheden van de hardware, maar ook van gespecialiseerde programmeertalen of technieken om bestaande systemen gedistribueerd te maken. Niet alle onderzoekers zullen deze gespecialiseerde kennis hebben. In de (nabije) toekomst zal hier dus nog heel wat werk verzet moeten worden.

Het nut van de enorme meertalige tekstcollecties van de EU om taaltechnologische oplossingen voor alle EU-talen te bouwen

Volgens een studie door gerenommeerde Europese experts in taaltechnologie (TT) worden 21 van de 30 daarin bestudeerde talen (± 70%) bedreigd met digitale uitroeiing omdat de digitale ondersteuning van deze talen niet of nauwelijks aanwezig is [1]. Dit oordeel is gebaseerd op onderzoek in vier gebieden: automatische vertaling, spraakinteractie, tekstanalyse en de beschikbaarheid van tekstcorpora. Tekst-corpora zijn noodzakelijke ingrediënten voor de ontwikkeling van de drie genoemde meer complexe taal- en spraaktechnologieën. Zulke waardevolle corpora zijn echter dun gezaaid, zelfs voor de meerderheid van de 23 officiële Europese talen. De EU maakt en bezit grote hoeveelheden meertalige corpora die gebruikt kunnen worden voor de ontwikkeling van taal-gerelateerde toepassingen. De EU is dus in de positie om het TST-veld een flinke steun in de rug te geven en dat doet ze dan ook! Wat stellen de EU-instituties dan ter beschikking? En hoe kan zelfs een eenvoudig verzameling tekstbestanden gebruikt worden om taal- en spraaksoftware te ontwikkelen? Dit zijn een aantal van de vragen die we in deze bijdrage zullen trachten te beantwoorden. Laten we bij het begin beginnen.

Ralf Steinberger
European Commission Joint Research Centre - Ispra
Italië

Hebben we daadwerkelijk behoefte aan TST-tools voor alle Europese talen?

Is het werkelijk noodzakelijk dat we tools hebben voor Nederlands, Portugees, Litouws en Sloveens? Kunnen we niet beter allemaal goed Engels leren zodat het probleem is opgelost?

Dit is min-of-meer de situatie in de VS, een meertalig land met meerdere nationaliteiten waar men besloten heeft allemaal het Engels als nationale taal te gebruiken. Iedere buitenlandse tekst wordt gewoon vertaald in het Engels. Maar, zouden we dat als Europeanen wel willen? Waarom Engels en niet Nederlands, Duits of Frans? Volgens de Eurobarometer [2], spreekt slechts 38% van de Europeanen het Engels voldoende goed als tweede taal om te kunnen converseren en is 58% in staat om in een willekeurig andere taal te kunnen converseren. Je kunt dus stellen dat we nog lang niet het ideaal van één taal bereikt hebben. Bovendien komen in een internationale setting niet-moedertaalsprekers dikwijls als minder ontwikkeld en dommer over dan mensen die hun eigen taal spreken.

Een ander argument tegen de eenzijdige focus op het Engels is iets dat we gezien hebben tijdens het jarenlang bestuderen van multinationale mediamonitoring, namelijk dat de verstrekte informatie in het nieuws in de verschillende talen complementair [3] is. Alleen wereldomspannende grote gebeurtenissen worden in de verschillende talen besproken maar de meeste plaatselijke gebeurtenissen worden nooit vertaald en halen niet de internationale pers.

De getoonde kaart van de Europese Media Monitor [4], toont de plaatsen die genoemd werden in een momentopname van het live nieuws. Elk van de bijna 50 nieuwstalen heeft een eigen kleur gekregen hetgeen duidelijk maakt dat gebeurtenissen in bepaalde gebieden alleen in sommige talen verslagen worden en in andere talen niet. Wanneer we alleen het nieuws in het Engels zouden monitoren, dan zouden we de meeste gebeurtenissen en de meeste details domweg missen.

De EU staat op meertaligheid

Ongeveer tien jaar geleden onderkende Richtlijn 2003/98/EC [5] van het Europese Parlement en van de Raad voor het hergebruik van publieke informatie, dat meertaligheid een van Europa's fundamentele principes is die de culturele en talige diversiteit garandeert. De wetgevers merkten vervolgens op dat vertalingen en taal-overstijgende informatietechnologie een potentiële bijdrage kan leveren aan het transparanter, gelijker, verantwoordelijker en democratischer maken van de EU omdat het de burgers toegang geeft tot beleids- en wetgevende voorstellen in alle Europese talen.
Echt, zou het niet mooi en interessant zijn wanneer we weten wat de (geplande) wetgeving is in de ons omringende landen zegt over genetisch gemodificeerde organismen, over het

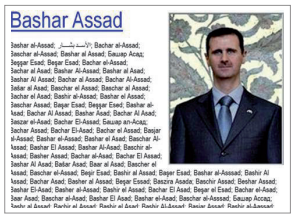
dragen van een boerka in het publieke domein, en over subsidies voor alternatieve energie? De Richtlijn stelt verder dat taal-overschrijdende toegang een positief effect kan hebben op het verwijderen van hindernissen voor concurrentie in de interne markt van de EU. Om al deze redenen plaveide de wetgever alweer negen jaar geleden de weg voor een onbelemmerde toegang voor R&D tot de enorme Europese collectie meertalige teksten.

Hoe kan een eenvoudige collectie documenten helpen TST-tools te ontwikkelen?

We hebben dus TST-applicaties in vele talen nodig. Om ze te ontwikkelen hebben we basisresources zoals corpora en woordenboeken nodig en we hebben behoefte aan softwarecomponenten zoals tools voor morfologische analyse, POS-labelers, parsers, enz. enz. The EU beschikt over een groot aantal parallelle corpora, documenten en hun handmatig geproduceerde vertalingen. Parallelle data is bijzonder nuttig omdat het de training mogelijk maakt van statistische vertaalcomputers (niet alleen voor Engels, Duits of Frans maar ook voor minder gebruikte talen). Het kan bovendien gebruikt worden voor het automatisch genereren van woordenboeken. Het staat projectie van annotaties over talen toe zodat het goedkoper wordt om TST-programma's te maken en te testen. De hierboven genoemde EU-richtlijn van 2003 erkent het nut van EU-data voor het ontwikkelen van TST-hulpmiddelen en het effent de weg voor de vrije en wijdverspreide distributie ervan. In 2006 heeft het eigen *Joint Research Centre* (JRC) van de EC in Italië een groot aantal parallelle corpora gemaakt en beschikbaar gesteld hetgeen een significante bijdrage leverde voor het - en wel voor de eerste keer - ontwikkelen van een automatisch vertaalsysteem voor 462 taal-paren waarvan ook de minder gebruikte taalparen zoals Portugees-Litouws en Fins-Sloveens deel uitmaakten [6]. Sindsdien hebben verscheidene veelal meertalige EU corpora het licht gezien [7].

Heeft de EU meer dan ruw tekstmateriaal?

Ja, dat hebben ze! EU organisaties hebben het zeer grote meertalige inter-institutionele terminologie-gegevensbestand van IATE beschikbaar gemaakt in een voor computers leesbare vorm [8]. Daarnaast zijn er verscheidene meertalige thesaurussen en classificatieschema's (inclusief EuroVoc [9] en enkele ondersteunende gereedschappen en informatie voor vertalers [10]). In 2011 werd het JRC-Namen [11] corpus gelanceerd, een corpus dat bestaat uit automatisch gegenereerde meertalige namenlijsten (zie de verschillende spellingsvarianten van de naam Bashar Assad in het kader).



Ook de bijbehorende software werd beschikbaar gesteld die gebruikt kan worden voor het verbeteren van automatische vertalingen van namen. JRC-Namen helpt ook bij het vinden van gelijke namen die verschillend gespeld worden (inclusief de verschillende schrijfwijze in de verschillende schriftsystemen) in dataverzamelingen zoals pers- en fotoarchieven. Bovendien helpt het bij het trainen en testen van zogeheten Named Entity herkenningssoftware in verschillende talen. In 2012, werd de tekstcategorisatie tool JRC EuroVoc Indexer (JEX) [12] gelanceerd. Deze software, die getraind is op 22 talen, heeft ten doel de snelheid en consistentie van het werk van bibliothecarissen te verbeteren (zie het EuroVoc screenshot, die de Engelse descriptoren geeft van een Hongaarse tekst). Als softwarecomponent, kan JEX bijdragen aan het vinden van verwante teksten in verschillende talen en van gevallen van plagiaat over taalgrenzen heen.

Deze EU-resources lossen niet alle problemen op, maar ze brengen ons wel dichter bij het uiteindelijke doel: namelijk dat mensen en machines gemakkelijk en in verschillende talen met elkaar kunnen communiceren! Het grootste voordeel van deze EU-corpora is dat ze een bijna even grote hoeveelheid data bevatten voor de minder gebruikte talen.

1) *www.meta-net.eu/%20whitepapers*
2) *http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf*
3) *www.springerlink.com/content/86656518k7116r2u/*
4) *http://emm.newsbrief.eu/geo?format=html&type=cluster&language=all*
5) *http://eur-lex.europa.eu/LexUriServ/LexUriServ.o?uri=CELEX:32003L0098:EN:NOT*
6) *www.mt-archive.info/MTS-2009-Koehn-1.pdf and www.euromatrixplus.net/*
7) *http://langtech.jrc.ec.europa.eu/JRC_Resources.html*
8) *http://iate.europa.eu/*
9) *http://eurovoc.europa.eu/*
10) *http://ec.europa.eu/dgs/translation/publications/*
11) *http://langtech.jrc.ec.europa.eu/JRC-Names.html*
12) *http://langtech.jrc.ec.europa.eu/Eurovoc.html*

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

META-Net oktober 2011

baar te maken voor kleinere talen. Alleen zo kan Europa haar talen, een essentieel onderdeel van ons culturele erfgoed, ook in het digitale tijdperk toekomstbestendig maken.

META-NET Witboekserie

De META-NET Witboekserie 'Talen in de Europese Informatiemaatschappij' rapporteert over de toestand van 30 Europese talen met betrekking tot taaltechnologie en legt uit wat de meest urgente risico's en mogelijkheden zijn. Het is voor het eerst dat er per taal een dergelijk veelomvattend overzicht verschijnt. Ieder taalwitboek is geschreven in de taal waarover het rapporteert en bevat ook een volledige Engelse vertaling.
De witboeken zijn geschreven voor de volgende Europese talen: Baskisch, Bulgaars, Catalaans, Deens, Duits, Engels, Ests, Fins, Frans, Galicisch, Grieks, Hongaars, Iers, IJslands, Italiaans, Kroatisch, Lets, Litouws, Maltees, Nederlands, Noors (bokmål en nynorsk), Pools, Portugees, Roemeens, Servisch, Sloveens, Slowaaks, Spaans, Tsjechisch, en Zweeds.

De witboeken zijn zowel digitaal als in drukvorm beschikbaar.

Over META-NET en META
META-NET, een excellentienetwerk van 60 onderzoekscentra uit 34 landen, beoogt de technologische fundamenten te ontwikkelen van een meertalige Europese informatiemaatschappij. META-NET wordt mede gefinancierd door de Europese Commissie. META-NET is ook de initiatiefnemer van META, de Meertalige Europese Technologische Alliantie. Meer dan 600 organisaties uit 55 landen, waaronder onderzoekscentra, universiteiten, midden- en kleinbedrijven en verschillende grote ondernemingen, hebben zich al aangesloten bij deze open technologische alliantie.

Digitale versies van de META-NET taalwitboeken: www.meta-net.eu/whitepapers

NOTaS

Stichting NOTaS
Secretariaat:
Postbus 31070, 6503 CB Nijmegen
T: 024-352 88 88
W: www.notas.nl
E: info@notas.nl

NOTaS behartigt de belangen van bedrijven en kennisinstellingen die actief zijn op het terrein van Taal- en Spraaktechnologie (TST). Dit doet zij o.a. door middel van bijeenkomsten, lobbyactiviteiten en het tijdschrift DIXIT.

Deelnemers Stichting NOTaS

CLST Radboud Universiteit Nijmegen
Erasmusplein 1,
6525 HT Nijmegen
Postbus 9103, 6500 HD Nijmegen
T: 024-361 16 86
W: www.ru.nl/clst
E: clst@let.ru.nl

Het Centre for Language and Speech Technology (CLST) onderzoekt en adviseert op het gebied van taal- en spraaktechnologie (TST). Belangrijke speerpunten in het onderzoek zijn audio- en tekstontsluiting, de inzet van TST ten behoeve van het onderwijs en de zorg, en het gebruik van TST ten behoeve van forensische toepassingen. Tevens rekenen wij ontwikkeling en curatie van data en tools tot onze expertise.

Data Archiving and Networked Services (DANS)
Anna van Saksenlaan 10,
2593 HT Den Haag
Postbus 93067, 2509 AB Den Haag
T: 070-344 64 84
W: www.dans.knaw.nl
E: info@dans.knaw.nl

DANS bevordert duurzame toegang tot digitale onderzoeksgegevens. Hiertoe stimuleert DANS dat wetenschappelijke onderzoekers gegevens duurzaam archiveren en hergebruiken. Tevens biedt DANS met Narcis.nl toegang tot duizenden wetenschappelijke datasets, e-publicaties en andere onderzoeksinformatie in Nederland. DANS is een instituut van KNAW en NWO en deelnemer aan (inter)nationale projecten en netwerken.

Dedicon
Traverse 175,
5361 TD Grave
Postbus 24, 5360 AA Grave
T: 0486-486 486
W: www.dedicon.nl
E: info@dedicon.nl

Dedicon is in Nederland dé organisatie die lectuur en informatie toegankelijk maakt. Zo kan iedereen lezen wat hij wil in de vorm die hem past. Binnen de diensten van Dedicon speelt de moderne taal- en spraaktechnologie een belangrijke rol.

Dutchear
Olof Palmestraat 16-18,
2616 LR Delft
T: 015-219 20 90
W: www.dutchear.nl
E: info@dutchear.nl

Dutchear past spraaktechnologie toe voor het automatisch doorzoeken van grote audio/video verzamelingen. Typische voorbeeldtoepassingen die we dagelijks doen zijn het zoeken in rechtzaken, vergaderingen en radio/televisieuitzendingen. Daarnaast levert en optimaliseert Dutchear spraaksynthese (TTS) en sprekerherkenning/-verificatie.

GridLine B.V.
Keizersgracht 520, 1017 EK Amsterdam
T: 020-616 20 50
W: www.gridline.nl
E: info@gridline.nl

GridLine is een succesvol ITbedrijf dat gespecialiseerd is in taaltechnologie voor het Nederlands. Een aantal van onze producten en diensten: Klinkende Taal - plugin die ambtenaren helpt begrijpelijke brieven te schrijven; FAST; SharePoint; Lucene; NedLex - informatieplatform voor advocaten; de TaalServer - voor alle taalopgaven in uw bedrijf; opinion mining, information retrieval, kennismanagement. GridLine, voor professionals die het Nederlands gebruiken.

Kecida, NFI
Laan van Ypenburg 6,
2497 GB Den Haag
T: 070-888 64 00
W: www.forensischinstituut.nl
E: m.israel@nfi.minvenj.nl

Als onderdeel van het Nederlands Forensisch Instituut (NFI) ondersteunt het Kennis- en expertisecentrum voor intelligente dataanalyse (Kecida) overheidsinstellingen uit de openbare orde- en veiligheidssector (OOV) bij de informatiegestuurde uitvoering van opsporing-, handhaving en toezichthoudende taken. Diverse state-of-the-art technologieën voor datamining, tekstmining en sociale netwerkanalyse worden ingezet om nieuwe inzichten te geven.

Knowledge Concepts
De Handboog 9,
5283 WR Boxtel
T: 0411-610 802
W: www.knowledge-concepts.com
E: sales@knowledge-concepts.com

Gedreven door passie en ervaring in dit vakgebied ontwikkeld en realiseert Knowledge Concepts vooruitstrevende systemen ten behoeve van informatieverwerking. Onze oplossingen onderscheiden zich in aspecten als: verzamelen, samenvoegen, uitlijnen, opslaan, ontsluiten en verspreiden van informatie en kennis. Dit realiseren wij met name door innovatief gebruik van taaltechnologie in combinatie met o.a. zoekmachines, document management systemen en portals.

Lexima
Kastanjelaan 6,
3833 AN Leusden
T: 033-432 44 52
W: www.lexima.nl
E: info@lexima.nl

Lexima is dé specialist in ondersteunende technologie voor dyslexie en taalproblematiek. Onze activiteiten zijn erop gericht kinderen, jongeren en volwassenen met een leerstoornis een eerlijker kans op succes te bieden: thuis, op school of in het werk. Dochterbedrijf Lexima Reinecker Vision is leverancier van de betere hulpmiddelenzorg bij slechter zien.

Linguistic Systems BV
Bijleveldsingel 58,
6524 AE Nijmegen
T: 024 -322 63 02
W: www.euroglot.nl
E: info@euroglot.nl

Linguistic Systems BV, opgericht in 1985, is een van de eerste taaltechnologiebedrijven in Nederland. Het bedrijf ontwikkelde het product Euroglot, een meertalig vertaalhulpmiddel, dat door grote multinationals in binnen- en buitenland dagelijks wordt gebruikt. Daarnaast ontwikkelde het bedrijf een parser-generator voor de constructie van morfologie- en grammatica-modules in zes Europese talen. Momenteel werkt het bedrijf aan de voltooiing van Euroglot Automatic, een systeem om hele teksten te vertalen.

OSTT/Sint Maartenskliniek
Hengstdal 3,
6574 NA Ubbergen
T: 024-365 97 18
W: www.ostt.eu /www.maartenskliniek.nl
E: ostt@maartenskliniek.nl

Het "Ontwikkelcentrum voor Spraak- en Taaltechnologie ten behoeve van spraak- en Taalpathologie en revalidatie algemeen" is erkend als expertisecentrum door het Ministerie van het VWS aan de Sint Maartenskliniek. In het kader van het OSTT werkt de Sint Maartenskliniek samen met de afdeling Taalwetenschap van de Radboud Universiteit en met het UMC St. Radboud in Nijmegen.

ReadSpeaker
Dolderseweg 2A,
3712 BP HUIS TER HEIDE
T: 030 - 692 44 90
W: www.readspeaker.com
E: nederland@readspeaker.com

ReadSpeaker is een internationale organisatie opgericht in Zweden, met inmiddels kantoren in Nederland, Duitsland, Engeland, Frankrijk, Spanje, de VS en Zweden. ReadSpeaker is marktleider op het gebied van het omzetten van tekst op het internet naar spraak via TTS, en richt zich met haar diensten op toegankelijkheid en gemak. Wij leveren innovatieve producten voor het voorlezen van websites (de eerste ooit), het toepassen van spraak voor andere web-based applicaties en RSS feeds en applicaties voor het gebruik op mobiele apparaten.

RightNow Intent Guide Development, Oracle Nederland
Eekholt 40,
1112 XH Diemen
T: 020-398 38 69
W: www.oracle.com/nl
E: fabrice.nauze@oracle.com

Understanding true customer intent through Oracle RightNow Intent Guide with Natural Language Technology will enable you to match customer goals with strategic corporate targets. You will make it easier for customers to engage during key moments of the customer journey in commercial and support-oriented transactions to create an optimal return per online customer interaction. By identifying and recognizing the intent of customer engagements these become more relevant, efficient and insightful, supporting your business goals in terms of transactions, cost-reduction and online engagement.

Telecats
Colosseum 42,
7521 PT Enschede
Postbus 92,
7500 AB Enschede
T: 053-488 99 00
W: www.telecats.nl
E: info@telecats.nl

Telecats ontwikkelt producten en diensten op basis van taal- en spraaktechnologie en VoIP. Met oplos¬singen op basis van IVR en (open vraag) spraakherkenning worden bij telefoongesprekken bellers geïdentificeerd, gesprekken geclassificeerd en (deels) geautomatiseerd. Met spraakanalyse kunnen inhoud en toon van live gesprekken en audioarchieven worden ontsloten. De taal- en spraaktechnologieën zijn als webservices beschikbaar voor derden die deze willen integreren in hun eigen propositie.

TST-Centrale, p/a Nederlandse Taalunie
Lange Voorhout 19,
2514 EB Den Haag
Postbus 10595,
2501 HN Den Haag
T: 070 3469548
W: www.tst-centrale.org
E: servicedesk@tst-centrale.org

Wanneer u op zoek bent naar (informatie over) digitale taalkundige bronnen dan bent u bij ons aan het juiste adres. Of u voor een kennisinstelling werkt of voor het bedrijfsleven, of u geïnteresseerd bent in taal of in spraak, of u een taalkundige bent of een spraaktechnoloog, wij zijn u graag van dienst.

Universiteit van Tilburg – CIW/TiCC
Warandelaan 2,
5037 AB Tilburg
Postbus 90153, 5000 LE Tilburg
T: 013-466 91 11
W: www.uvt.nl
E: uvt@uvt.nl

In onderwijs en onderzoek van het Departement Communicatie- en Informatiewetenschappen (CIW) ligt het accent op aspecten van menselijke communicatie, zowel talige als niet-talige, inclusief mens-machine interactie, kennisrepresentatie, text mining, machine learning, kunstmatige intelligentie, computationele semantiek, semantische annotatie, en dialoogmodellen en -systemen. Het department omvat sinds September 2008 het Tilburg Center for Cognition and Communication (TiCC). Het departement verzorgt onder andere de bloeiende bachelor- en masteropleidingen Bedrijfscommunicatie en Digitale Media, de masteropleiding Human Aspects of Information Technology, en samen met de Radboud Universiteit Nijmegen de research masteropleiding Language and Communication.

Universiteit Twente - HMI
Drienerlolaan 5,
7522 NB
Postbus 217,
7500 AE Enschede
T: 053-489 37 40
W: hmi.ewi.utwente.nl
E: hmi_secr@ewi.utwente.nl

De HMI-groep van de Universiteit Twente (UT) is een onderzoeksgroep binnen de afdeling Informatica. HMI doet onderzoek op het gebied van mens-machine interactie, taal- en spraaktechnologie en zoektechnologie voor multimedia collecties (tekst, spraak, video). In de buurt van de campus van de Universiteit Twente bevinden zich enkele spin-off bedrijven die op hetzelfde terrein actief zijn.

Universiteit Utrecht – Utrecht Institute of Linguistics OTS
Trans 10,
3512 JK Utrecht
T: 030-253 60 06
E: uil-ots@uu.nl
W: www.hum.uu.nl/uilots

Het UiL OTS is een onderzoeksinstituut binnen de Faculteit Geesteswetenschappen van de Universiteit Utrecht en stelt zich ten doel onderzoek te doen naar menselijke taal. De volgende drie dimensies bepalen het onderzoek: (i) de architectuur van het menselijke taalsysteem, (ii) de cognitieve systemen die ten grondslag liggen aan de verwerving en de verwerking van taal, en (iii) de wijze waarop taal wordt gebruikt in communicatie tussen mensen en tussen mens en machine.

Sponsoren/Samenwerking

CumLingua Taal & Communicatie
Full-service taalbureau
Bronkhorstweg 48,
5363 TZ Velp (N.B.)
T: 0486 – 471 554
W: www.cumlingua.com
E: info@cumlingua.com

- Copywriting, redactie- en vertaaldiensten
- Taalles op maat en in-company training
- Proefschriftvertaling en correctie

Deykerhoff/Accountants
Rompertsebaan 70,
5231 GT, 's-Hertogenbosch
Postbus 2325,
5202 CH, 's-Hertogenbosch
T: 073-623 24 50
W: www.deykerhoff.nl
E: denbosch@deykerhoff.nl

Deykerhoff/Accountants is een accountantskantoor dat zich in de regio's 's-Hertogenbosch/Waalwijk en Nijmegen vooral richt op de dienstverlening aan ondernemingen en ondernemers in het MKB en het (medisch) vrije beroep. Onze proactieve instelling is gericht op een grote betrokkenheid en een op maat gesneden persoonlijk contact met de klanten.

Leonard Communicatie bv
W: www.leonard.nl
E: info@leonard.nl

Maakt communicatie zichtbaar

Malta & de Keyzer
Toernooiveld 300,
6525 EC Nijmegen
T: 024-351 21 08
W: www.malta-online.nl
E: info@malta-online.nl

Malta & de Keyzer, specialisten in officemanagement en secretariaat.

Nederlandse Taalunie
Lange Voorhout 19,
2514 EB Den Haag
Postbus 10595, 2501 HN Den Haag
T: 070-346 95 48
W: www.taalunie.org
E: info@taalunie.org

De Nederlandse Taalunie is een beleidsorganisatie waarin Nederland, België en Suriname samenwerken op het gebied van de Nederlandse taal, onderwijs en letteren.

VANZELFSPREKEND

INTERACTIVE VOICE RESPONSE

OPEN SOURCE TELEFONIE

SPRAAKTECHNOLOGIE

KLANTCONTACTCENTER

SPRAAKHERKENNING

VOIP GATEWAYS

TELECATS

TAALTECHNOLOGIE

EMOTIEDETECTIE

SPRAAKANALYSE

MUZIEKDETECTIE

AUDIOMINING

RAPPORTAGES

TURNDETECTIE

SELFSERVICE

OPEN VRAAG

NASPREKEN

BEGRIJPEN

OPLIJNEN

VERSTAAN

TELIOS

ACD

KTO