DIXIT

TIJDSCHRIFT OVER TAAL- EN SPRAAKTECHNOLOGIE

TST van betekenis

JAARGANG 17 - DECEMBER 2020

INHOUD

VOORWOORD

Het spreekt voor zich dat 2020 voor ons allemaal een bijzonder jaar is geweest. Door de uitbraak van de COVID-19-pandemie is ons leven ingrijpend veranderd. We zijn massaal thuis gaan werken, het onderwijs is flink ontregeld en vindt nog steeds voor een groot deel op afstand plaats, en de gezondheidszorg is onder enorme druk komen te staan.

Ook voor NOTaS zijn de consequenties voelbaar geweest. Vergaderingen en evenementen werden uitgesteld, naar later verplaatst, om dan wel of niet online aangeboden te worden. Een bekende ontwikkeling voor de meesten van ons. Het zat niet altijd mee en het blijft vaak nog puzzelen, namelijk met het vraagstuk: hoe houd je een stichting springlevend zonder persoonlijk contact tussen de leden?

Laten we hopen dat deze DIXIT een gedeeltelijk antwoord op deze vraag biedt. Het is en blijft een belangrijke verbindingsfactor - door de inspanning van de redactie, de inhoudelijke contributies van de TST-gemeenschap en door jullie, de lezers.

De gezondheidszorg was - met vooruitziende blik - het thema van de vorige DIXIT. Terugkijkend kunnen we ons afvragen hoeveel van deze TST-toepassingen de zorg hebben geholpen (zoals dialoogsystemen met patiënten) en welke nieuwe applicaties de toekomst zal brengen.

Denk bijvoorbeeld aan het bron- en contactonderzoek, een prachtig speelveld om alle kanten van TST toe te passen, van spraakherkenning tot entiteitenresolutie, en van zoek- en dialoogsystemen tot Linked Data. Het zoeken naar verbanden en betekenis in grote hoeveelheden verhalen van patiënten als bewijs voor het nut van TST.

Deze DIXIT heeft als thema "TST van betekenis", een zeer relevant thema in onzekere tijden.

Veel leesplezier,

Fabrice Nauze, voorzitter Stichting NOTaS

DIXIT: Tijdschrift over toegepaste taal- en spraaktechnologie – 17e jaargang, editie TST van betekenis. DIXIT is een uitgave van Stichting NOTaS, Toernooiveld 100, 6525 EC Nijmegen. Tel. 024-3512108– E-mail info@notas.nl – www.notas.nl **Redactieadres:** Stichting NOTaS, Toernooiveld 300, 6525 EC Nijmegen **Redactie:** Arjan van Hessen: a.j.vanhessen@utwente.nl; Henk van den Heuvel: h.vandenheuvel@let.ru.nl; Martha Hofman: MAHofman@hetnet.nl; Marieke den Os (secretariaat NOTaS): info@notas.nl. **Hoofdredactie themanummer:** Piek Vossen: p.t.j.m.vossen@vossen.vu.nl **Advertenties:** Stichting NOTaS – info@notas.nl, 024-3512108 Abonnementen: Voor een gratis abonnement kunt u zich wenden tot een van de NOTaS-deelnemers **Druk:** Leonard Marketing & Communicatie **Verantwoording:** DIXIT is een uitgave van Stichting NOTaS. Overname van de artikelen is alleen toegestaan met bronvermelding en na toestemming van Stichting NOTaS. Stichting NOTaS en de bij deze uitgave betrokken redactie en medewerkers aanvaarden geen aansprakelijkheid voor mogelijke gevolgen die zouden kunnen voortvloeien uit het gebruik van de in deze uitgave opgenomen informatie.

WHAT'S IN A WORD?

Woorden *hebben* geen betekenis van zichzelf en zinnen of verhalen opgebouwd uit woorden evenmin. Woorden *krijgen* betekenis in een samenleving, een context, voor een spreker of een hoorder: betekenis in het *gebruik* zoals de filosoof Wittgenstein het stelde en niet a priori.

Door:
Piek Vossen, hoofd-redacteur van deze DIXIT, Vrije Universiteit, Amsterdam

Er valt veel te analyseren als we kijken naar dat gebruik. Zo gebruiken we een relatief klein aantal woorden voor veel verschillende dingen en zaken. Woorden en taal zijn vreselijk ambigu en vaag. De meest voorkomende woorden hebben ook de meeste betekenissen als we mogen geloven wat er in een woordenboek staat. Denk aan woorden zoals "band" [12], "stuk", "zin" [12], "spelen" [12], "slaan" [9]. Een zin als "Het heeft geen zin om te spelen met een band die stuk is" is dan een puzzel van 12x12x12x9=15,552 betekeniscombinaties en dan tellen we de mogelijke relaties tussen die woorden niet eens mee.

Aan de andere kant is taal ook rijk en gevarieerd. Volgens datzelfde woordenboek hebben wij meer dan 8.000 woorden (niet allemaal even vriendelijk) in het Nederlands die naar een of andere persoon kunnen verwijzen maar ook meer dan 5.000 woorden voor bewegingen en nog eens 5.000 voor geluiden. Stel je voor, je ziet een persoon die een beweging maakt en je hoort een geluid. Welk woord moet je dan kiezen?

Als er twee verschijnselen zijn die onze natuurlijke taal kenmerken dan zijn dat ambiguïteit en variatie. In deze tijd van Big Data zijn neurale netwerken in staat om uit miljoenen teksten patronen te leren. Zogenaamde woord embeddings reduceren de betekenis van woorden tot een paar honderd dimensies (100 tot 300). Woorden die dicht bij elkaar komen te staan kunnen gezien worden als variaties voor dezelfde concepten. Dergelijke modellen doen het goed in taaltechnologische toepassingen omdat ze het probleem van de variatie gedeeltelijk oplossen. Bovendien krijg je deze modellen gratis en voor niets als een afgeleide van ons taalgebruik. Wat ze nog niet oplossen is het probleem van ambiguïteit. Daarvoor worden nu nog meer geavanceerde modellen gemaakt die woorden modelleren in de zinnen waarin ze voorkomen.

Toch blijft het een spel van woorden. Taal refereert echter ook naar de buitenwereld: beelden, geluiden, situaties, emoties en sociale interactie om ons heen. Er liggen nog veel uitdagingen voor ons om taaltechnologie naar een menselijk niveau te tillen. Toch is dat de weg die voor ons ligt om intelligente systemen te kunnen maken die communicatie niet alleen duiden maar gebruiken om ons van dienst te zijn.

Betekenis in taal is een soort heilige graal. Is die eenmaal gevonden, dan zijn de toepassingen legio. In deze DIXIT komen veel aspecten van taal en betekenis naar voren, en in het bijzonder wat taal- en spraaktechnologie kan bijdragen om betekenis te ontsluiten. Het vormt een interessant palet dat laat zien in welke gaten en hoeken men die graal zoekt. •

De betekenissen van "band" in Wikipedia

HET REFERENTIEBESTAND NEDERLANDS: EEN DIGITALE SCHATKAMER VOOR HET NEDERLANDS

Door:
Hennie van der Vliet en Willy Martin, Vrije Universiteit Amsterdam

Voor de automatische verwerking van taal is een zekere hoeveelheid kennis van die taal nodig, of die verwerking nu gebeurt op basis van machineleren, neurale netwerken of een 'ouderwets' regelgebaseerd systeem. Linguïstische kennis op het gebied van zowel klank, woordvorming, syntaxis, betekenis als gebruik, kan relevant zijn, afhankelijk van het soort taalverwerking en het doel ervan. In deze bijdrage stellen we het Referentiebestand Nederlands (RBN) voor, een machineleesbaar lexicon van het Nederlands dat rijk is aan zulke lexicale informatie.

We zullen vooral aandacht besteden aan één van de troeven van het RBN, de uitvoerige manier waarop de combinatoriek van de zelfstandige naamwoorden en de werkwoorden in het lexicon is beschreven. Dat is, zoals gezegd, slechts één van de troeven van het RBN. Wie meer wil weten van de rijkdom aan informatie in dit lexicon kan terecht op de website van het Instituut van de Nederlandse Taal, waar de uitvoerige documentatie van het woordenboek te vinden is[1]. Wie zelf eens in het RBN wil rondkijken kan de demoversie van Cornetto[2] raadplegen, waarin het RBN integraal is opgenomen.

Bijzondere informatiebron

Het RBN was een initiatief van de Commissie Lexicografische Vertaalvoorzieningen (CLVV) en functioneerde succesvol als referentiebestand bij de totstandkoming van vertaalwoordenboeken met het Nederlands als bron- of doeltaal[3]. Daarnaast is het RBN ontworpen voor gebruik in de taaltechnologie. Met het oog op die doelstellingen is bij de woorden in het RBN heel veel informatie opgenomen, die elders vaak niet of niet in die mate te vinden is. Met name vanwege dat laatste is het RBN nog steeds heel relevant voor taaltechnologische toepassingen.

Lexical Units

Het RBN bevat ongeveer 30.000 zelfstandige naamwoorden, 6.300 werkwoorden en 6.300 bijvoeglijke naamwoorden. Daarnaast zijn er nog zo'n 1.450 functiewoorden en 2.450 geografische namen. Elke lexicale ingang in het RBN is een *lexical unit* (LU), een eenheid van vorm en betekenis. Voor de vorm *koe* bijvoorbeeld, zijn drie betekenissen, en dus ook drie aparte LU's opgenomen: het zoogdier op de boerderij, het vrouwtjesdier van bepaalde zoogdieren (giraf, buffel) en de betekenis domoor. In elk lemma is steeds de volledige lexicale informatie van de LU opgenomen.

Informatie voor nomina

Hieronder volgt een kort en onvolledig overzicht van de informatie in het RBN voor nomina:

- Algemene informatie over het lemma:
  - de woordsoort, de vorm (is het een afkorting, acroniem), het geslacht (zowel pronominaal als grammaticaal), het lidwoord
- Woordvorm (vervoeging, verbuiging):
  - de meervoudsvormen
- Een korte definitie
- Semantiek
  - onder meer de telbaarheid en een semantisch type
- Syntaxis
  - onder meer de complementatie
- Combinatoriek
  - grammaticale en lexicale collocaties en andere vaste combinaties
- Pragmatiek
  - onder meer een specificatie van het domein en het register

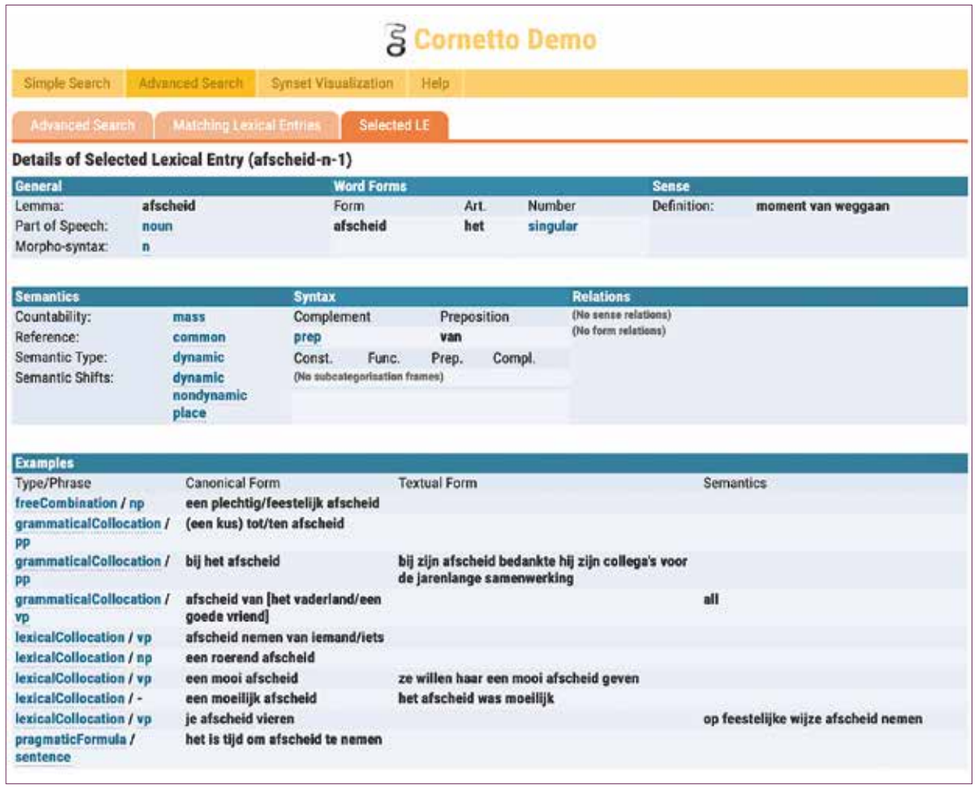
We hebben niet de ruimte om volledig te zijn en dat is hier ook niet nodig. In publicaties[4] en in de al genoemde documentatie is een tamelijk compleet overzicht te vinden van de beschrijving van nomina, adjectiva en verba. We gaan hier wel wat dieper in op collocaties, min of meer vaste combinaties van woorden.

[1] https://taalmaterialen.ivdnt.org/download/tstc-referentiebestand-nederlands/

[2] http://cornetto.clarin.inl.nl/index.html

[3] Martin, W. en J. Ploeger (1999). *Tweetalige Woordenboeken voor het Nederlands: het beleid van de Commissie Lexicografische VertaalVoorzieningen* (CLVV). In: Neerlandica Extra Muros, 37 (3), 23-33.

[4] Vliet, H.D. van der (2007). The Referentie Bestand Nederlands as a multi-purpose lexical database. In: International Journal of Lexicography, 20(3): 221-238.



Een voorbeeld van een RBN-ingang zoals weergegeven in de demoversie van het Cornetto lexicon (http://cornetto.clarin.inl.nl/index.html)

Soorten collocaties

Zowel voor de zelfstandige naamwoorden als voor de werkwoorden zijn collocaties en andere woordcombinaties opgenomen. Voor begrip en productie van taal zijn die onmisbaar. Collocaties zijn combinaties van woorden die vaak voorkomen en die een voorkeurspositie innemen in taalgebruik. Een paar voorbeelden maken veel duidelijk:

- Een zwerm bijen, een school vissen, een baar goud.
- De radio speelt, de motor loopt, het licht brandt.
- Hard werken, streng bewaken, zwaar belasten.

Het ene woord (*bij, radio, werken*) roept het andere (*zwerm, spelen, hard*) op. In het RBN worden dit lexicale collocaties genoemd en ze worden semantisch ingedeeld in een kleine vijftig categorieën. Daarnaast bevat het RBN ook een grote hoeveelheid grammaticale collocaties, doorgaans combinaties met voorzetsels zoals *kijken naar*, *de hoop op* en *een aanklacht tegen*. Ten slotte zijn in het RBN natuurlijk ook ondoorzichtige lexicale combinaties opgenomen, zoals *van de hak op de tak (springen)* en pragmatische uitdrukkingen zoals *wie weet* en *toe nou*.

Het RBN bevat een rijkdom aan collocaties en andere combinaties die in de meeste andere lexicale bronnen niet te vinden is[5]. Collocaties zijn taalspecifiek, hoogfrequent, onvoorspelbaar, niet altijd transparant en ze zijn aan specifieke woorden gekoppeld. Ze leveren veel semantische informatie en zijn onmisbaar bij de productie van taal. En dit is nog maar één voorbeeld van lexicale informatie in het RBN die elders niet of veel minder uitgebreid te vinden is. Een ander voorbeeld is de zeer uitvoerige beschrijving van de syntactische werkwoordcomplementatie. Wat dat betreft, is dit lexicon een voorloper van het Open Dutch FrameNet. Maar daarover leest u elders in dit nummer. •

[5] Het Algemeen Nederlands Woordenboek (ANW, http://anw.ivdnt.org/about) en Combinatiewoordenboek van Piet de Klein (https://combinatiewoordenboek.nl) zijn ook goede bronnen van collocaties, zij het dat het Combinatiewoordenboek zich beperkt tot combinaties van zelfstandig naamwoorden met werkwoorden.

POLYSEMIE EN BETEKENISRELATIES

Door:
Dirk Geeraerts, Onderzoeksgroep QLVL (Quantitative Lexicology and Variational Linguistics), KU Leuven

Rekening houden met betekenis in taaltechnologische toepassingen veronderstelt dat we een goed idee hebben van wat zoal schuilgaat achter het begrip 'betekenis'. Daar kan de taalkunde toe bijdragen: de semantiek is weliswaar niet altijd de centrale subdiscipline van het taalkundig onderzoek, maar dat onderzoek heeft wel een arsenaal aan begrippen en methodes ontwikkeld voor het beschrijven en classificeren van betekenisverschijnselen. Een deel van dat arsenaal heeft betrekking op de manier waarop woordbetekenissen in zinnen gecombineerd worden. Dat is de 'compositionele semantiek'. Het andere deel – lexicale semantiek – heeft betrekking op die woordbetekenissen zelf, en daarover vertellen we hier iets meer.

Lexicale semantiek

Een voorbeeld kan helpen om de veelheid aan verschijnselen te illustreren die in de lexicale semantiek een rol spelen. Een alledaags woord als stof komt voor met alvast de volgende toepassingen (de omschrijving wordt telkens gevolgd door enkele voorbeelden):

1. weefsel: *wollen, katoenen stoffen*
2. materie, substantie van een bepaald type: *giftige stoffen, vaste stof, grijze stof*
3. onderwerp waarover men spreekt, schrijft, nadenkt: *stof voor een roman, een college, een handboek; stof tot onenigheid*
4. massa zeer kleine droge deeltjes van verschillende oorsprong, door de lucht meegevoerd: *een wolk stof, stof afnemen*
5. massa zeer kleine deeltjes als toestand van een specifieke substantie: *iets tot stof vermalen, tot stof verpulveren.*

Oorsprong van woorden

Tussen deze betekenissen bestaan verbanden van verschillende aard. Zo hangen 1, 2 en 3 onderling samen, en 4 en 5 eveneens. De historische oorsprong van die twee groepen is verschillend. De 'weefsel'-betekenis gaat etymologisch terug op het Oudfranse woord *estoffe*, dat zelf z'n oorsprong vindt in een verre voorloper van ons huidige woord *stoppen*. De 'kleine deeltjes'-lezing aan de andere kant heeft historisch een relatie met het woord *stuiven*. De technische term voor betekenis(groep)en die zo ver uit elkaar liggen is **homonymie**. We gaan er dan van uit dat 1-3 en 4-5 betekenisgroepen zijn die bij een ander woord (zeg maar, $stof_1$ en $stof_2$) horen. Ieder van die woorden afzonderlijk kent zelf ook nog verschillende betekenissen, en die meerduidigheid van woorden noemen we **polysemie**. Het herkennen van polysemie is voor taaltechnologische toepassingen natuurlijk niet zonder belang: het maakt echt wel wat uit of *stofjes* in een tekst gaat over textieltypes dan wel partikeltjes die door de lucht zweven.

Betekenisverbanden

Polysemie zit niet willekeurig in elkaar; de onderscheiden betekenissen die we bij een woord aantreffen, vertonen onderlinge verbanden. Die verbanden verklaren ook waarom we een woord in meer dan een betekenis kunnen gebruiken. Een begrijpelijk associatief verband tussen een bestaande betekenis en een nieuwe, maakt aannemelijk hoe die nieuwe heeft kunnen ontstaan. Enkele van de vaakst optredende betekenisverbanden kunnen we nu ook weer illustreren met het stof-voorbeeld.

- Tussen 1 'weefsel' en 3 'onderwerp waarover men spreekt, schrijft, nadenkt' bestaat een figuurlijk verband. Zoals textiel het materiaal is waarmee we kledingstukken (of overgordijnen of tafelkleden et cetera) maken, zo zijn onderwerpen, topics, thema's de substantie waaruit teksten, uiteenzettingen, verhalen en dergelijke opgetrokken worden. Bij zo'n betekenisrelatie op basis van een beeld spreken we van een **metafoor**.
- Een metaforische betekenisrelatie is gebaseerd op gelijkenis: we zien een analogie tussen de rol van textielweefsels ten opzichte van kleding en de rol van onderwerpen ten opzichte van teksten. Maar de gelijkenis is niet-letterlijk: het materiaal van de roman, het college of het handboek is alleen in figuurlijke zin 'materie'. Heel vaak zijn betekenisrelaties echter ook gebaseerd op letterlijke **gelijkenis**. Dat is bijvoorbeeld het geval tussen 4 en 5. De vuiltjes die spontaan aan komen dwarrelen en het gruis dat we produceren als we iets afbreken lijken heel sterk op elkaar. Tot op zekere hoogte verschillen ze slechts in de manier waarop ze tot stand zijn gekomen. Er is dus een gelijkenisverband, maar we zullen niet willen zeggen dat de ene betekenis een figuurlijke, metaforische betekenis is van de andere.
- Tussen 1 en 2 bestaat een verband dat we **generalisatie** noemen. We blijven in het domein van de concrete materie, maar de basisbetekenis 'materiaal van textiele aard' wordt verruimd tot 'materiaal in het algemeen'. In een hiërarchische ordening van begrippen is 'textiel' een subcategorie van 'type materie'. De betekenis van stof verschuift in die hiërarchie dus van een ondergeschikt naar een overkoepelend begrip. (De omgekeerde verschuiving, van ruimer naar specifieker bestaat ook. Dan spreken we van *specialisatie*.)
- **Metonymie** is een laatste belangrijk type van betekenisverschuiving. We kunnen het illustreren met de uitdrukking *in het stof bijten*. Binnen de uitdrukking sluit *stof* aan bij betekenis 5, maar wat is de relatie tussen het op de grond neervallen (de initiële betekenis van de uitdrukking als geheel) en de afgeleide betekenis 'verliezen, overtroffen worden, verslagen worden, het onderspit delven'? De 'verliezen'-interpretatie is een veralgemening van een situatie waarin iemand in een letterlijk, fysiek gevecht overwonnen wordt, maar tussen 'op de grond neervallen' en 'een (fysiek) gevecht verliezen' bestaat nog een ander soort van verband: het op de grond neervallen is oorzakelijk verbonden met de nederlaag. Of men kan ook zeggen dat het een onderdeel is van de mislukking. Betekenissen die aan elkaar gelieerd zijn door zulke causale of deel/geheel- verbanden (of nog andere, zoals nabijheidsrelaties) noemen we metonymisch.

Positie van woorden

Hoe kun je nu taaltechnologisch greep krijgen op de veelheid van lexicaal-semantische verschijnselen? Corpus gebaseerde modellen voor de representatie van woordbetekenis gaan uit van de gedachte (terug te voeren tot J.R. Firth) dat woorden die in vergelijkbare omgevingen voorkomen, vergelijkbare betekenissen hebben. In functie van de woorden waarmee ze samen optreden worden woorden dan mathematisch voorgesteld door de positie die ze innemen in een vectorruimte met een groot aantal dimensies. Als we daarbij ook de verschillende polyseme toepassingen van een woord willen kunnen onderscheiden, dan moeten we in staat zijn om te herkennen dat, bijvoorbeeld, zinnetjes met *wollen stof* en *katoenen stof* dichter bij elkaar liggen dan zinnetjes met *wollen stof* en *giftige stof*.

Taalrijkdom

Daarbij moeten we ter afsluiting op een belangrijke moeilijkheid wijzen. Het beeld van betekenisverschillen dat in het bovenstaande overzicht geschetst wordt, is behoorlijk simplistisch. Het suggereert dat het gebruik van een woord als *stof* altijd makkelijk in te delen is in betekenissen als geïllustreerd door 1-5. Maar een onbetwistbare manier om tot zo'n ordening te komen is er niet, en het semantische bereik van woorden laat dan ook meestal verschillende alternatieve verdelingen toe. Dat kan onder andere daaruit blijken dat verschillende woordenboeken met een verschillende classificatie werken. Bij de Grote Van Dale wordt 3 bijvoorbeeld in twee gesplitst, door 'grond, reden, aanleiding' (als in *stof tot onenigheid*) apart te plaatsen van 'datgene wat de inhoud uitmaakt van een geschrift, rede of les'. In het Verschueren-woordenboek worden 4 en 5 onder één definitie gevat, terwijl de Grote Koenen 4 juist uitsplitst tussen 'in de lucht voorkomende kleine deeltjes van verschillende aard' en 'een neergeslagen laag daarvan'. Woordbetekenissen laten alternatieve indelingen toe, en die fundamentele flexibiliteit betekent ook voor de taaltechnologie een bijzondere uitdaging. •

NEURALE TAALMODELLEN EN TRANSFER LEARNING

In 2017 schreef ik in DIXIT een artikel getiteld 'De Deep Learning revolutie – en wat nu?'. Ik eindigde dat artikel met een blik op de toekomst:
Op dit vlak moeten deep-learning-methodieken de komende jaren nog grote stappen gaan zetten, met generieke methoden voor specifieke problemen en met verklarende modellen die ons inzicht verschaffen in de uitdagingen en verbeteringen.
Nu, drie jaar later, is het zo ver: we hebben grote neurale taalmodellen die inzetbaar zijn voor specifieke taken waarvoor we niet veel data beschikbaar hebben.

Door:
Suzan Verberne, Universiteit Leiden

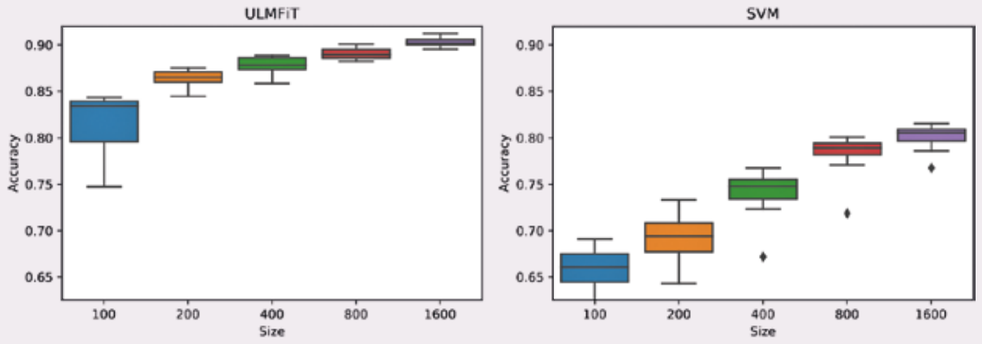
Beeldclassificatie

Om uit te leggen hoe die methoden werken en wat we ermee kunnen, is het handig om eerst iets te vertellen over modellen voor de classificatie van afbeeldingen. Die modellen worden getraind op *ImageNet*. ImageNet is een reusachtige database van afbeeldingen, handmatig ingedeeld in duizenden categorieën. Er zitten bijvoorbeeld 1474 afbeeldingen van zebra's in ImageNet. Met deze gigantische database kun je een neuraal netwerk trainen om zebrafoto's te kunnen vinden in een nieuwe fotocollectie. Maar het is ook mogelijk om van zo'n getraind neuraal netwerk de laatste laag te verwijderen – de laag die een afbeelding het label 'zebra' geeft – en het netwerk opnieuw te trainen op een veel kleinere dataset, met nieuwe categorieën die ImageNet niet heeft. Dit kunnen heel specifieke categorieën zijn. Denk bijvoorbeeld aan een gemeente die uit foto's van de openbare ruimte graag wil kunnen afleiden rondom welke afvalbakken afval is gedumpt. Omdat het neurale model een enorme hoeveelheid informatie heeft opgeslagen uit de ImageNet-afbeeldingen, kan het model met weinig voorbeelden van afvalbakken de nieuwe classificatietaak succesvol leren. Dit proces heet *transfer learning*.

Van ImageNet naar taalmodellen

In juli 2018 schreef Sebastian Ruder: "NLP's ImageNet moment has arrived"[1]. Het was al langer mogelijk om neurale taalmodellen te trainen op grote hoeveelheden tekst. Die taalmodellen, zoals *word2vec*, leren welke woorden voorkomen in elkaars context en dus qua betekenis en syntactische functie op elkaar lijken. Voor het trainen van zo'n taalmodel is geen gelabelde data nodig; alleen een grote collectie lopende tekst – de volledige Wikipedia bijvoorbeeld. Taalmodellen zoals word2vec zijn ontzettend populair en kunnen interessante inzichten opleveren, maar ze hebben ook beperkingen. Zo zijn ze niet direct te gebruiken als classificatiemodel na het 'afknippen' van de laatste laag zoals bij de op ImageNet getrainde neurale modellen.

In 2018 vond een revolutie plaats in NLP: de ontwikkeling van taalmodellen die dit wél kunnen.



ULMFiT en SVM vergeleken voor verschillende datasetgroottes

[1] https://ruder.io/nlp-imagenet/

Ruder publiceerde ULMFiT: *Universal Language Model Fine-tuning*, een effectieve *transfer learning* methode die ingezet kan worden voor elke tekstclassificatietaak. In 2019 vergeleken wij ULMFiT voor het classificeren van Nederlandstalige boekrecensies met een traditioneel model, Support Vector Machines (SVM), en keken hoe goed ze allebei om konden gaan met weinig trainingsdata[2]. Uit de resultaten (zie pagina 8) bleek duidelijk dat ULMFiT in het voordeel is, dankzij het taalmodel dat is voorgetraind op 92 miljoen woorden Wikipediatekst.

BERT

Maar de ontwikkelingen stonden niet stil en labs bleven zoeken naar nog efficiëntere manieren om nog betere taalmodellen te trainen. Onderzoekers van Google presenteerden eind 2018 een universeel taalmodel op basis van de *Transformer*-architectuur: BERT[3]. De Transformer is een model dat een hele zin in één keer kan verwerken en tijdens het verwerken opslaat welke stukjes van de zin het belangrijkst zijn om te onthouden. Dit is iets wat bijvoorbeeld van pas komt bij automatisch vertalen, omdat het vaak nodig is om het begin van de zin nog te weten om het einde goed te produceren.

BERT leert representaties van woorden door in het corpus steeds willekeurige woorden te verbergen en dan te proberen die woorden te voorspellen. Na het lezen van de hele Wikipedia weet BERT behoorlijk goed welke woorden hij kan verwachten in welke context. En dat niet alleen, het getrainde netwerk kan elke classificatie- of extractietaak leren uitvoeren. Denk aan het vinden van positieve en negatieve berichten op sociale media, het ontdekken van relaties tussen medicaties en bijwerkingen, of het anonimiseren van teksten door persoons- en plaatsnamen te identificeren.

BERT is ongekend populair, mede dankzij de direct inzetbare implementaties[4] en een groot aantal voorgetrainde BERT-modellen die klaarstaan om gefinetuned te worden.

Er bestaan ook Nederlandse BERT-modellen, genaamd BERTje, RobBERT en BERT-NL. Hoe groter het corpus is dat gebruikt is voor het voortrainen, hoe hoger de nauwkeurigheid op de uiteindelijke taak[5]. De classificatie van boekrecensies, die ULMFiT met 93.8% nauwkeurigheid kon maken, kan RobBERT met 95.1% volbrengen. En de Transformer-architectuur maakt het bovendien mogelijk om te zien welke woorden in een recensie het meest bijdroegen aan de positieve of negatieve toon van de tekst.

Zijn al onze problemen opgelost?

Nee, nog niet. Er zijn nog veel open vragen over BERT-modellen. Wat leren ze precies over syntax en semantiek? Hoe kunnen verschillende BERT-en elkaar aanvullen? Ook lijken BERT-modellen kwetsbaar voor slordigheden in de data, of classificatietaken met heel veel categorieën (bijvoorbeeld medische diagnoses). Hoe kunnen we die kwetsbaarheid verminderen?

En dan is er nog een andere uitdaging: de computerkracht die nodig is om diepe modellen zoals BERT te trainen. Het voortrainen van een nieuw BERT-model op een krachtige GPU-computer duurt dagen. Het finetunen kan in enkele uren op een GPU-processor, maar een SVM-model kun je op je eigen laptop in hooguit een half uur trainen. Voor grootschalig gebruik van de nieuwste generatie taalmodellen is een lichtgewicht trainingsproces nodig, dat de modellen inzetbaar maakt voor mensen en organisaties die minder computerkracht tot hun beschikking hebben. Hopelijk kan ik over drie jaar in DIXIT schrijven dat dit doel bereikt is. •

[2] Benjamin van der Burgh and Suzan Verberne (2019). The merits of Universal Language Model Fine-tuning for Small Datasets -- a case with Dutch book reviews. https://arxiv.org/abs/1910.00896
[3] Jacob Devlin et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805
[4] https://huggingface.co/
[5] Pieter Delobelle et al. (2020). RobBERT: a dutch RoBERTa-based language model. https://arxiv.org/abs/2001.06286

CONTEXT, COMPLEXITEIT EN CONTEMPORANITEIT

Waarom computers nog steeds moeite hebben met lezen en hoe Linked Data hierbij gaat helpen

Het domein van de taaltechnologie heeft de afgelopen jaren enorme vorderingen gemaakt, maar toch lukt het de computer vaak nog niet om teksten helemaal te doorgronden. Dit komt vooral doordat veel informatie impliciet blijft. Als we taaltechnologie willen inzetten om teksten gemakkelijker doorzoekbaar, analyseerbaar en vergelijkbaar te maken, dan zullen we ervoor moeten zorgen dat die technologie zich meer bewust wordt van de context van een tekst zodat deze minder fouten zal maken.

Door:
Marieke van Erp - DHLab, KNAW Humanities Cluster

Outbreak Management Team of Outright Monetary Transactions

Ons brein kan razendsnel koppelingen leggen tussen nieuwe informatie en kennis die we al hadden. Dit zorgt ervoor dat berichtgeving niet steeds de basisconcepten rondom een bepaald onderwerp hoeft te herhalen. Zo is het voor de meeste lezers van een krant niet nodig om bij een nieuw artikel rondom COVID-19 steeds weer uit te leggen dat 'OMT' staat voor 'Outbreak Management Team' omdat zij de context inmiddels wel kennen. Deze betekenis van 'OMT' is in Nederland vrij recent en er zijn dus niet veel encyclopedieën die bij 'OMT' een definitie geven als: 'specialisten en experts met verschillende achtergronden en kennis over een bepaalde ziekte die (op basis van actuele informatie, hun vakkennis en beschikbare wetenschappelijke literatuur) bespreken, hoe de uitbraak bestreden kan worden.' Maar in zo'n encyclopedie vind je bijvoorbeeld wel een verwijzing naar het 'Outright Monetary Transactions programma van de Europese Centrale Bank', dat dezelfde afkorting heeft.

Tekstverrijking en waar het spaak loopt

Als we taaltechnologie willen inzetten om teksten gemakkelijker doorzoekbaar, analyseerbaar en vergelijkbaar te maken, dan is het vaak zinvol om elementen in teksten te koppelen aan bronnen die deze elementen beschrijven. Zo kan een koppeling gelegd worden tussen een verwijzing van OMT in een tekst naar de Wikipedia pagina, die uitlegt dat dit staat voor ofwel het Outbreak Management Team in Nederland, ofwel het Outright Monetary Transactions programma van de Europese Centrale Bank, afhankelijk van de context. Veel van dit soort informatie is beschikbaar en bruikbaar als 'Linked Data', een methode van vormgeven van data die begonnen is in het vakgebied van het Semantic Web.

Er zijn machine-leesbare varianten van Wikipedia, zoals DBpedia en Wikidata, die door veel taalverrijkingstools gebruikt worden om zulke koppelingen te leggen. Dit is enorm waardevol, omdat het ervoor kan zorgen dat je het niveau van zoeken op trefwoord kunt ontstijgen en kunt gaan zoeken op informatie over categorieën. Als je dan bijvoorbeeld 10.000 krantenberichten hebt verrijkt met koppelingen naar politici, dan kun je ook vergelijkingen gaan maken tussen hoe vaak politici van de ene partij worden genoemd versus politici van een andere partij.

Echter, de informatie in Wikipedia (en dus DBpedia en Wikidata) is niet compleet en ook niet altijd up-to-date. Wikipedia bevat bijvoorbeeld veel informatie over hedendaagse politici en media-persoonlijkheden, maar als je nu juist geïnteresseerd bent in 19e eeuwse vrouwelijke wetenschappers, dan vang je vaak bot. En het kan gebeuren dat morgen in de krant een nieuw concept wordt genoemd, dat nog niet in Wikipedia staat. Hierdoor is het niet mogelijk om een koppeling te leggen naar meer achtergrondinformatie.

Daarnaast is de informatie in dit soort bronnen op zo'n manier vormgegeven dat deze niet altijd flexibel kan schakelen tussen contexten. Veranderingen zijn in DBpedia en Wikidata zeer minimaal gepresenteerd. Er wordt bijvoorbeeld wel beschreven dat Arnold Schwarzenegger tot de categorieën bodybuilders, acteurs en politici behoort, maar wanneer hij precies wat deed is maar moeilijk of zelfs niet te achterhalen.

Betere Linked Data en Betere Taaltechnologie

Het precies beschrijven van wat een concept of entiteit precies is, blijkt ontzettend moeilijk te zijn. Sinds de tijd van Aristoteles wordt al geprobeerd om 'de wereld te volledig beschrijven'. Natuurlijk blijft iedere dataset die we maken incompleet, evenals een representatie van een stukje van de wereld, maar er wordt wel hard gewerkt aan het verbeteren van Linked-Data-bronnen. In de

context van het Europees gefinancierde NewsReader project is bijvoorbeeld begonnen met het creëren van Linked Data-databases die niet uitgaan van een concept, maar van een gebeurtenis, waardoor verandering centraal staat en niet de statische karakteristieken van een concept.

Door middel van 'fuzzy' kennismodellering en het gebruik van Word embeddings slaan onderzoekers van de UvA en het KNAW Humanities Cluster een brug tussen Linked Data en hoe concepten in verschillende facetten in teksten beschreven worden. Hierbij kun je denken aan het uitsplitsen van teksten die het hebben over Nederland als geografische locatie ('In Nederland wonen 17 miljoen mensen'), Nederland als voetbalteam ('Nederland won het EK van 1988') of Nederland als politieke entiteit ('Nederland akkoord met strengere controles EU'). Deze verschillende dimensies van concepten worden met elkaar verbonden in een Linked Data bron. Vervolgens kunnen tekstverrijkingstools koppelingen maken naar het hoofdconcept of de relevante dimensie. Dit is nog 'work in progress' maar op deze manier komen we steeds dichter bij databronnen die niet alleen maar snapshots presenteren van hoe de wereld er nu uitziet. Ze tonen ook betekenisveranderingen en de verschillende dimensies van een concept die ervoor zorgen dat taalverrijkingstools een tekst aan de juiste context kunnen koppelen. •

NAMED ENTITIES: BELANGRIJKE BOUWSTENEN IN DE OPBOUW VAN AUTOMATISCH TEKSTBEGRIP

Named Entity Recognition (NER) of automatische naamherkenning is de NLP-taak waarbij de computer eigennamen in teksten identificeert en vervolgens toekent aan vooraf gedefinieerde categorieën. Wat kunnen we verstaan onder 'namen'? Uiteraard zal een standaard NER-systeem namen van personen, organisaties en locaties detecteren, maar named entities kunnen ook vele andere vormen aannemen. Zo worden in het biomedische domein proteïnen, genen en chemische substanties gezien als named entities en in een banktoepassing kunnen ook data en bedragen beschouwd worden als named entities. In dit artikel blikken we terug op het ontstaan van automatische naamherkenning, haar vele toepassingen en de uitdagingen.

Door:
Veronique Hoste,
Vakgroep Vertalen, tolken en communicatie,
Universiteit Gent

Korte geschiedenis

NER is ontstaan als een subtaak binnen de zesde Message Understanding Conferentie (MUC-6) in 1995. Deze conferenties werden in het leven geroepen en gefinancierd vanuit het Amerikaanse defensie-agentschap DARPA. Terwijl de focus in het begin lag op de ontwikkeling van NLP-systemen voor militaire toepassingen, verschoof die later naar meer 'civiele' thema's. Opmerkelijk aan die MUC-conferenties was, dat het niet alleen bijeenkomsten waren waarin onderzoekers hun bevindingen rapporteerden, maar dat ook ter plekke evaluaties werden uitgevoerd. Hierbij kregen deelnemende onderzoeksteams data waarop ze een bepaalde methodologie moesten uittesten. Dat was de geboorte van de zogenaamde 'shared tasks', die anno 2020 met competities zoals SemEval en vele anderen, populairder zijn dan ooit en een katalysator zijn van vooruitgang in verschillende NLP-taken.

NER ingeburgerd

25 jaar na MUC-6 is NER nog steeds een van de hoekstenen van vele NLP-toepassingen. Vraag-antwoord-systemen zoals IBM Watson of automatische zoeksystemen draaien vaak rond named entities, al dan niet gekoppeld aan externe kennisbanken (wikification). Corporate branding toepassingen die de opinies en emoties van klanten, investeerders of zelfs het brede publiek monitoren over specifieke entiteiten, zoals bedrijven, politieke personen of partijen, producten of diensten, zijn intussen breed ingeburgerd. En stel dat u een nieuwsgierige burger bent die goed geïnformeerd wenst te zijn over de actualiteit van de dag. Als u uw krant digitaal leest, zal u geconfronteerd worden met een aanbevelingsalgoritme dat bepaalde artikelen voorstelt op basis van uw leesgedrag. Named entities zijn cruciale componenten in nieuwsgebeurtenissen, zoals mooi geïllustreerd in de zin *Von der Leyen en Michel overleggen met Britse premier Johnson over Brexit*. Het ligt het voor de hand dat ze een steeds belangrijkere rol gaan spelen in meer geavanceerde nieuwsaanbevelingssystemen. NER wordt idealiter ook gekoppeld aan toepassingen zoals coreferentieresolutie en automatische standpuntanalyse zodat je als burger een breed inzicht krijgt in verschillende standpunten over actuele nieuwsitems.

Lerende systemen

Op methodologisch vlak zien we dat NER ongeveer dezelfde evolutie heeft doorlopen als vele andere taaltechnologische toepassingen. Terwijl men in het begin vooral werkte met eenvoudige lexicongebaseerde aanpakken waarbij lijsten van persoonsnamen en plaatsnamen (zogenaamde gazetteers) gebruikt werden om teksten te scannen op de aanwezigheid van die namen, worden nu toch ook al meer dan 10 jaar lerende systemen gebruikt. De standaard aanpak hierbij is een gesuperviseerde methodologie, waarbij een lerend systeem wordt getraind op basis van een corpus dat manueel is gelabeld met named entities. Uit die teksten worden dan kenmerken of 'features' afgeleid (zoals bepaalde vormkenmerken van woorden of informatie over de omringende woorden) die op een gestructureerde manier aan het systeem gevoed worden. Uit die features en de manueel gelabelde klasse kan een predictief model worden afgeleid dat dan kan worden toegepast op ongeziene data. Terwijl deze feature-engineering-aanpak jarenlang de state-of-the art uitmaakte, wordt ze nu steeds meer verdrongen door neurale aanpakken, die ook voor de taak van NER goede resultaten boeken.

NER succesvol

Nu de automatische naamherkenning haar zilveren jubileum heeft gevierd, zouden we dan niet mogen verwachten dat dit probleemloos functioneert? Kan ik met andere woorden ergens een named entity recognizer vinden die met een hoge accuraatheid namen van organisaties, locaties en personen kan detecteren in teksten? Natuurlijk! Als je namen wilt herkennen in krantenmateriaal, geschreven in een standaardtaal, en dan nog bij voorkeur het Engels, dan zal die zoektocht eenvoudig zijn. Ook voor het Nederlands mogen we toch wel van geluk spreken; een aantal jaren geleden hebben Vlaamse en Nederlandse TST-onderzoekers een mooi corpus geassembleerd, SoNaR, waarin 1 miljoen woorden werd voorzien van manuele codering van de named entities. Dit heeft geleid tot de bouw van verschillende automatische naamherkenners voor het Nederlands. Neem zeker een kijkje op de site van het CLARIN voor beschikbare tools (https://www.clarin.eu/resource-families/tools-named-entity-recognition).

Rothko in het Witte Huis

Het verhaal wordt iets minder rooskleurig als je kijkt naar talen waarvoor minder geannoteerde data of gazetteers beschikbaar zijn. Bovendien brengt elke shift van domein of genre weer nieuwe uitdagingen met zich mee. Daarbovenop wil je voor sommige applicaties, zoals vraag-antwoordtoepassingen of automatische ontologieconstructie, soms ook een fijnmazige naamherkenning tot je beschikking hebben. En ook in NER krijgen we te maken met de creativiteit van de taalgebruiker: soms is de letterlijke betekenis van de named entity niet dezelfde als de bedoelde betekenis. Dergelijk metonymisch gebruik van namen is schering en inslag. Denk maar aan namen

Rothko in het Witte Huis

van landen die continu worden gebruikt om nationale sportteams te benoemen, bijvoorbeeld: *Spanje staat met 1-0 voor*.

Of neem het volgende voorbeeld: *Het Witte Huis heeft een aantal hedendaagse kunstwerken gekocht, waaronder een Rothko. Het schilderij werd voor maar liefst 1,7 miljoen dollar verkocht*. Een traditioneel NER systeem zou het 'Witte Huis' classificeren als een locatie en 'Rothko' als een persoon, wat geen steek houdt. In deze context fungeert de locatie namelijk metonymisch als de president van de V.S. of eventueel als het personeel tewerkgesteld in het Witte Huis. En het is voor ons als taalgebruiker ook evident dat het genoemde bedrag niet werd neergeteld om de mens Rothko ergens aan een muur in het Witte Huis te hangen. Voor een NLP-systeem ligt hier de uitdaging om te begrijpen dat 'Rothko' en 'het schilderij' co-refereren: ze verwijzen beide naar dezelfde buitentalige realiteit. En dat brengt ons bij de grote uitdaging die vele NLP-onderzoekers kopbrekens bezorgt: hoe modelleer je wereldkennis? •

FRAMING, BETEKENIS EN PRAGMATIEK

Ligt deze boot aan land of op de grond? Frames maken duidelijk hoe taal onze blik op de werkelijkheid 'kleurt' (afb. bewerkt van: https://www.pikist.com/free-photo-vgiof/nl, CC)

Frames zijn een belangrijk hulpmiddel om de betekenis van taal te beschrijven. Een team van onderzoekers aan de Vrije Universiteit Amsterdam en de Rijksuniversiteit Groningen werkt momenteel in het kader van het NWO-project 'Framing situations in the Dutch language' aan een database over talige frames in het Nederlands, en aan taaltechnologie om automatisch frames in teksten te kunnen herkennen.

Door:
Gosse Minnema, Rijksuniversiteit Groningen,
Levi Remijnse, Vrije Universiteit, Amsterdam

Wat zijn frames?

'Framing' met taal is een fenomeen dat we allemaal kennen uit bijvoorbeeld de media of uit geschiedenisboeken. Een "verzetsdaad" of een "terroristische aanslag" kan zomaar dezelfde gebeurtenis zijn, maar op verschillende manieren geframed worden. Kort gezegd is een frame een beschrijving van alle conceptuele informatie die wordt opgeroepen bij het lezen van een woord of zin. Welke deelnemers horen er (stereotypisch) bij de gebeurtenis? Bij een verzetsdaad verwachten we informatie over degenen tegen wie het verzet gericht was ("een verzetsdaad tegen de nazi's"). Bij terrorisme verwachten we eerder informatie over de slachtoffers ("twintig mensen kwamen om bij de aanslag"). Ook wordt er informatie opgeroepen over gerelateerde woorden en frames: bij een verzetsdaad denken we aan oorlog en bezetting, bij een aanslag eerder aan criminaliteit en rampen.

Frames hoeven lang niet altijd politiek getint te zijn en kunnen door allerlei woorden en uitdrukkingen worden opgeroepen. "Land" en "grond" betekenen bijvoorbeeld ongeveer hetzelfde maar worden gebruikt in verschillende frames. Schipbreukelingen kunnen aan land gaan en een bal kan op de grond vallen, maar andersom ("de bal viel op het land") klinkt gek. Net als bij verzet en terrorisme hebben we dus te maken met verschillende verwachtingen over 'deelnemers' en gerelateerde woorden: "land" verwachten we in de context van "zee", "grond" verwachten we in de context van "lucht".

Deze opvatting van frames is algemeen geaccepteerd, maar sommige taalwetenschappers gaan een stap verder: de Amerikaanse taalkundige Charles J. Fillmore (1929-2014) ontwikkelde vanaf de jaren '80 de '*frame semantics*'-theorie, die stelt dat frames de belangrijkste 'bouwstenen' zijn van de betekenis van menselijke taal. Deze theorie is later ook in de computationele taalkunde opgepikt en gebruikt als basis voor lexicale databases en software voor het begrijpen van taal.

FrameNet en automatisch parsen

De belangrijkste computationele toepassing van *frame semantics* is FrameNet, een Engelstalige database van 1226 frames en meer dan 13.000 woordbetekenissen die beschreven worden aan de hand van deze frames. Ieder frame bevat een lijst van woordbetekenissen, semantische rollen en relaties met andere frames. Bijvoorbeeld: het frame KOPEN bevat de woorden "kopen", "aanschaffen" en "klant", heeft als deelnemers "Koper", "Verkoper", "Geld" en "Goederen", en is gerelateerd aan andere frames zoals VERKOPEN (met woorden als "verkopen" of "veiling").

FrameNet omvat niet alleen een database maar ook een corpus van teksten die geannoteerd zijn met informatie over de opgeroepen frames en hun deelnemers, bijvoorbeeld "$[\text{Harry}]_{\text{Koper}}$ $\text{kocht}_{\text{KOPEN}}$ $[\text{een boek}]_{\text{GOEDEREN}}$". Dit corpus kan gebruikt worden voor het trainen van statische modellen die automatisch zinnen kunnen analyseren aan de hand van frames. Zulke software kan nuttig zijn voor het vinden van relevante informatie in een tekst (*information extraction*) of het automatisch genereren van parafrases.

FrameNet en de daarop gebaseerde frame-analyse software zijn oorspronkelijk ontwikkeld aan de Berkeley-universiteit in Californië voor het Engels. Het project is ook vertaald naar een aantal andere talen zoals het Braziliaans-Portugees en het Zweeds, maar bestaat nog niet of nauwelijks voor het Nederlands. Dit is jammer omdat FrameNet, net als bijvoorbeeld WordNet en PropBank, een belangrijke semantische hulpbron is die gebruikt wordt in de taaltechnologie. Het NWO-project 'Framing Situations in the Dutch Language' heeft als doel om dit gat op te vullen.

Indirecte frames

In ons project maken we niet alleen een Nederlands FrameNet, maar onderzoeken we ook variatie in framing van bekende gebeurtenissen. We willen weten hoe teksten met verschillende perspectieven andere frames gebruiken om dezelfde gebeurtenis te beschrijven. Om dit te doen willen we FrameNet-annotaties koppelen aan een databank met gedocumenteerde gebeurtenissen. We willen bijvoorbeeld een tekst over een aanslag kunnen koppelen aan gestructureerde informatie over die aanslag. Op grond van deze informatie doen we aannames over relevante FrameNet-frames die in de tekst opgeroepen moeten worden om de inhoud van de tekst te begrijpen, zoals VERMOORDEN of WAPEN in het geval van aanslagen. Uit ons onderzoek blijkt vervolgens dat deze frames vaak niet op een directe manier worden opgeroepen, maar wel pragmatisch geïmpliceerd worden met behulp van gedeelde achtergrondkennis. Zo wordt in een tekst over de nasleep van de tramaanslag in Utrecht slechts één keer naar de aanslag zelf verwezen, namelijk met "het slachtoffer van een laffe daad". Dit zinsdeel bevat geen woorden die direct VERMOORDEN of WAPEN oproepen, maar deze frames wel impliceren op basis van gedeelde achtergrondkennis over de aanslag. We werken momenteel aan een nieuwe annotatiemethode die kan omgaan met dit soort 'indirecte frames'.

Taaltechnologie

Een ander onderdeel van het project is het ontwikkelen van automatische frame-parsers die aansluiten bij de twee doelen van het project: enerzijds willen we bestaande technieken uitbreiden naar het Nederlands, anderzijds willen we automatische parsers laten werken met indirecte framing. Deze technieken kunnen in een later stadium ook worden gebruikt om sneller documenten in ons corpus te kunnen annoteren. Voor het ontwikkelen van parsers voor het Nederlands is het een probleem dat er momenteel nog maar een beperkte hoeveelheid data voorhanden is; om dit te ondervangen willen we gebruikmaken van bestaande annotatie voor het Engels, en die met verschillende *transfer learning* technieken bruikbaar maken voor het trainen van parsers voor het Nederlands. •

Meer lezen

- Dutch FrameNet www.dutchframenet.nl
- Remijnse, L., & Minnema, G. (2020, May). Towards Reference-Aware FrameNet Annotation. In Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet (pp. 13-22).
- Vossen, P., Ilievski, F., Postma, M., Fokkens, A., Minnema, G., & Remijnse, L. (2020, May). Large-scale Cross-lingual Language Resources for Referencing and Framing. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 3162-3171).

[1] Bron: https://getfunded.nl/campagne/4321-kleine-bijdrage-in-kosten-afscheid-lieve-roos?

MEER DAN AUTOMATISCHE SPRAAK-HERKENNING: BETEKENIS IS MEER DAN ALLEEN WOORDEN

Stel je deze beurtwisseling voor. A vraagt: “Heb je het druk?” B zegt: “Nee hoor, helemaal niet.” Uit dit korte gesprek uitgedrukt in tekst zou je kunnen opmaken dat B het niet druk heeft. Maar, als je B had kunnen horen, dan hoorde je: “Neeee hoor, helemaaaal niet.” En dan had je waarschijnlijk bedacht, oh, B bedoelt het sarcastisch, B is juist wel druk, ook al zegt B van niet. Dus de manier waarop je iets zegt, speelt een belangrijke rol in de interpretatie van wat er gezegd wordt. Dit is iets wat automatische spraakherkenners nog niet goed kunnen: het interpreteren van die verborgen laag van betekenis aan de hand van stemanalyse.

Door:
Khiet Truong,
Human Media Interaction,
Universiteit Twente

Paralinguïstiek

De paralinguïstiek bestudeert allerlei aspecten in gesproken communicatie die meer interpretatie geven aan de lexicale betekenis van wat er gezegd wordt. Denk daarbij aan toonhoogte en intonatie, luidheid, spreeksnelheid, stemkwaliteit, pauzes, lachen, zuchten, interrupties et cetera. Deze paralinguïstische kenmerken kunnen bewust maar ook onbewust worden gecommuniceerd door de spreker. Je kunt bijvoorbeeld vaak een inschatting maken van iemands leeftijd en geslacht op basis van de stem – dit is informatie die onbewust wordt doorgegeven in de stem. Men maakt vaak ook bewust gebruik van paralinguïstische kenmerken in de stem om bijvoorbeeld emoties of sarcasme te uiten – dit zijn manieren om de lexicale betekenis te nuanceren, of te veranderen. Op dit moment maken automatische spraakherkenners weinig tot geen gebruik van deze paralinguïstische informatie om tot een transcriptie te komen; vocalisaties zoals lachen en zuchten behoren tot de groep van niet-lexicale vocalisaties en worden dus (ten onrechte) genegeerd.

Ten onrechte, omdat die paralinguïstische informatie heel belangrijk is in de interpretatie, niet alleen in mens-mens interactie, maar ook in mens-machine interactie.

Ik zal een aantal voorbeelden geven van toepassingen waarbij het automatisch herkennen van die paralinguïstische informatie zo belangrijk is.

Spraakherkenning

Het beeld dat iedereen met een automatische spraakherkenner in z’n broekzak loopt, was zo’n 20-30 jaar geleden nog science fiction. We komen steeds vaker gesproken dialoog systemen tegen (ook wel bekend als conversational agents). De vraag naar conversational agents die jou goed begrijpen, groeit; naast het goed verstaan van wat de gebruiker zegt, groeit ook de behoefte bij gebruikers en wetenschappers om technologie te ontwikkelen die ook de intentie en socio-emotionele aspecten van de gebruiker begrijpt zodat de (conversational) agent een passende reactie kan geven. Dit is met name van belang in de context van mens-agent interactie bij ouderen.

Uit eerder onderzoek is gebleken dat hulpbehoevende ouderen het makkelijkst interacteren met technologie via spraak. Taken van conversational agents voor ouderen kunnen dan zijn: hulp bieden aan het organiseren van en structuur geven aan een dag, het gevoel van isolatie en eenzaamheid verminderen, en het stimuleren van cognitieve, sociale, en fysieke activiteiten. Bij deze taken is het van belang dat de agent empathie toont in de dialoog; dit impliceert dat de agent socio-emotionele aspecten kan herkennen bij de gebruiker, zoals de gemoedstoestand en ‘engagement’ (mate van interesse / betrokkenheid in het gesprek).

Emotieherkenning

Er wordt volop onderzoek gedaan naar het automatisch herkennen van emoties in spraak. Ook hier heeft deep learning zijn intrede gedaan – de accuraatheid staat of valt met de hoeveelheid data.

Maar, emoties zijn complexer dan dat. Hoewel er pogingen zijn gedaan om een set van akoestische correlaten te vinden die emoties kunnen meten, is deze set is nog niet zo eenduidig zoals facial action units (die worden gebruikt om gezichtsexpressies te beschrijven) dat wel zijn. Bovendien zijn relaties die gevonden worden tussen deze akoestische correlaten en bepaalde socio-emotionele aspecten, vaak

Gezichtsexpressie. Beeld ANP Uit: Het Parool 18 maart 2020

niet reproduceerbaar in andere studies. Over het algemeen is men het er wel over eens dat je er met spraak alleen niet komt bij emotieherkenning; als er multimodale informatie beschikbaar is (lexicale informatie, gezichtsexpressies), dan zou het goed zijn om die informatie ook mee te nemen – uit veel onderzoeken blijkt dat de accuraatheid hiermee verbetert.

Ook moeten we goed beseffen dat context een belangrijke rol speelt: de situatie waarin een emotie wordt geuit is van invloed op de interpretatie. Wanneer een foto wordt getoond met een hordeloper die over een horde aan het springen is, classificeren mensen de gezichtsexpressie van deze hordeloper als ‘vastberaden’. Als er wordt ingezoomd en alleen het hoofd zichtbaar is, wordt dezelfde expressie door 91% van alle ondervraagden geclassificeerd als ‘boosheid’.

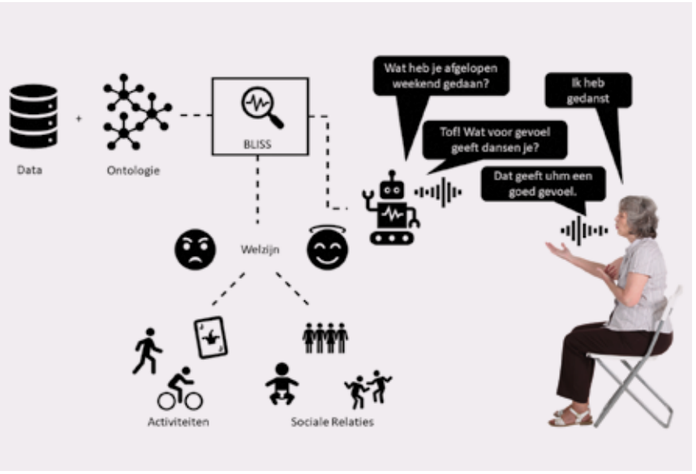
Naast het herkennen van socio-emotionele aspecten in spraak, groeit ook de interesse in het herkennen van fysieke en mentale gesteldheid in spraak. Zo wordt er onderzoek gedaan naar het herkennen van ziektebeelden zoals dementie, depressie, stress, autisme en psychische stoornissen (bijvoorbeeld bipolaire stoornis) met behulp van paralinguïstische informatie.

Sociale intelligentie

Bij Human Media Interaction aan de Universiteit Twente doen we onderzoek naar gesproken interacties tussen mensen en mens-machine en is ons doel om automatisch interpretatie te geven aan paralinguïstische kenmerken die geuit worden door de sprekers. Daarbij focussen we niet alleen op de socio-emotionele, fysieke en mentale gesteldheid, maar kijken we ook naar hoe paralinguïstische informatie wordt gebruikt om structuur te geven aan dialogen (denk bijvoorbeeld aan automatische generatie van de timing van backchannels of end-of-turn voorspelling). Dit doen we met als uiteindelijk doel om agents sociaal intelligenter te maken en om gesproken archieven te ontsluiten en doorzoekbaar te maken op een rijkere manier. Een sociale agent die emoties of sarcasme kan herkennen staat dus hoog op onze agenda, en zoals je net hebt kunnen lezen: <sarcasm=TRUE>dit is een fluitje van een cent</sarcasm>. •

EEN SPREKENDE INTERACTIEVE ASSISTENT DIE ZORGT VOOR MEER AUTONOMIE IN JE EIGEN WELBEVINDEN

In het project BLISS (Behaviour-based Language-Interactive Speaking Systems) werken onderzoekers van de Radboud Universiteit samen met collega's van de Universiteit Twente, het bedrijf Games for Health en het bedrijf ReadSpeaker. Dit onderzoeksproject heeft als doel het ontwikkelen van een spraakgestuurd interactief, Nederlandssprekend systeem dat mensen helpt zelf hun gevoel van welzijn en geluk te verhogen. Zie de BLISS website: https://bliss.ruhosting.nl/.



Door:
Catia Cucchiarini, Iris Hendrickx, Helmer Strik en Louis ten Bosch, CLST, Radboud Universiteit, Nijmegen;
Mariët Theune en Jelte van Waterschoot, HMI Universiteit Twente;
Rob Tieben, Games for Health

Virtuele sprekende assistenten

Veel diensten maken tegenwoordig gebruik van chatbots: apps die vragen stellen en antwoorden verzamelen. De meeste van deze diensten zijn gebaseerd op getypte conversatie, die meestal verloopt volgens een min of meer vastliggend stramien. Interactie via tekst heeft echter zijn beperkingen. Het recente succes van spraakgestuurde virtuele assistenten zoals Google Assistant, Amazon Alexa en Siri maakt duidelijk dat gesproken interactie groot potentieel heeft. De belangrijkste reden hiervoor is de relatieve laagdrempeligheid en toegankelijkheid van gesproken ten opzichte van geschreven interactie.

Aan de andere kant blijkt uit recente berichtgeving ook dat er nog veel nadelen kleven aan het gebruik van bovengenoemde spraakassistenten. Ten eerste blijken de herkenprestaties van spraakherkenners voor Nederlandse spraak tegen te vallen. Ten tweede klinkt de spraaksynthese in het Nederlands nog vaak erg kunstmatig in conversaties. Ten derde, en dat is waarschijnlijk het grootste nadeel, is het onduidelijk of privacy en beveiliging van de spraakopnames die via commerciële spraakassistenten worden verzameld, gegarandeerd kan worden. Ten slotte zijn veel systemen weliswaar geschikt om eenmalig een gesprek met een cliënt te voeren, maar veel minder geschikt voor meerdere gesprekken met een cliënt over een langere periode waarbij belangrijke informatie wordt opgeslagen en meegenomen in een nieuw gesprek.

Aanpak en domein

Om de bovengenoemde vier aspecten verder te onderzoeken en de technologie een stap verder te brengen, is in het kader van het programma NWO DATA2PERSON het project BLISS (Behaviour-based Language-Interactive Speaking Systems) gefinancierd. In dit project werken onderzoekers van het Centre for Speech and Language Technology (CSLT) van de Radboud Universiteit samen met collega's van Human Media Interaction (HMI) van de Universiteit Twente, het bedrijf Games for Health (GfH) en het bedrijf ReadSpeaker. Dit onderzoeksproject heeft als doel het ontwikkelen van een spraakgestuurd interactief, Nederlandssprekend systeem dat mensen helpt zelf hun gevoel van welzijn en geluk te verhogen. Daartoe gaat het project uit van een brede definitie van gezondheid en welzijn als indicatoren voor geluk.

Big Data

BLISS maakt gebruik van verschillende typen data. Het beschikbaar komen van grote hoeveelheden persoonlijke data (Big Data) en de enorme verbeteringen die recentelijk zijn geboekt in taal- en spraaktechnologie bieden nieuwe kansen voor autonomie van cliënten in de zorg. Het idee in BLISS is om persoonlijke data van cliënten (teksten, interviews, dialogen) te doorzoeken, om daaruit informatie te halen over hun gezondheid en welzijn en om te weten te komen wat hen gelukkig maakt. Om dit mogelijk te maken wordt technologie ontwikkeld die relevante informatie uit Nederlandse teksten en audio-opnames kan halen. Vervolgens kan deze informatie gebruikt worden om computerprogramma's te ontwerpen die met mensen laagdrempelig, in gesproken Nederlands kunnen communiceren om hen te ondersteunen in hun dagelijks handelen. Deze informatie kan gebruikt worden om het systeem te personaliseren voor individuele cliënten zodat het rekening houdt met persoonlijke behoeftes en achtergrond.

Welzijn van ouderen

Specifiek werken we aan een casus om het welzijn te verbeteren van oudere volwassenen in een zorginstelling. Ons doel hier is om systemen en hulpmiddelen te ontwikkelen waarmee verpleegkundigen nog beter persoonlijke zorg kunnen bieden aan senioren. Ons sprekende systeem kan persoonlijke, prettige gesprekken aangaan met een oudere cliënt, om er samen met de cliënt achter te komen wat deze persoon echt gelukkig maakt. Deze gesprekken kunnen bijvoorbeeld gaan over passies, interesses en sociaal participeren. Deze kennis kan vervolgens door verpleegkundigen gebruikt worden om de cliënt beter te leren kennen, contacten te leggen tussen cliënten met dezelfde interesses of om activiteiten te organiseren.

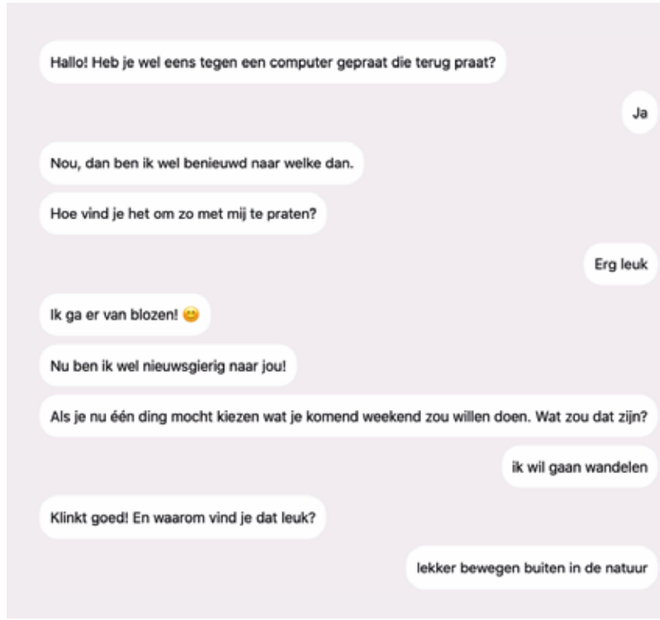
De nieuwste versie van het systeem wordt op dit moment geëvalueerd en doorontwikkeld met cliënten en professionals van een zorginstelling. Met hen ontwikkelen we samen een systeem dat makkelijker geaccepteerd en geïntegreerd kan worden op de werkvloer.

Testresultaten

Een eerste versie van het sprekende systeem is getest tijdens de Games for Health Europe conferentie en het DRONGO-talenfestival in Nijmegen in het najaar van 2019. Tijdens deze bijeenkomsten werd deelnemers gevraagd om met het BLISS systeem te praten over hun activiteiten en motivaties (Figuur 1). In totaal werden er 55 conversaties opgenomen en geanalyseerd. Dit gaf ons een eerste inzicht in welke activiteiten mensen leuk vinden om te doen en hun reden ervoor. Zo bleek dat de meerderheid van de genoemde motivaties betrekking had op de kwaliteit van leven en dagelijks functioneren van deelnemers.

Uitdagingen

Het BLISS-project heeft inmiddels een aantal struikelblokken ervaren. Zo bleek de zorginstelling die tijdens de projectaanvraag had ingestemd om mee te werken, bij de start van het project niet langer beschikbaar te zijn voor onze onderzoeken. Vervolgens werd het door de coronacrisis een heel stuk moeilijker om in contact te komen met ouderen. Zo werd het lastiger om te testen of het prototype van ons sprekende systeem goed werkt en of het ontwikkelde systeem mensen daadwerkelijk kan ondersteunen om inzicht in hun geluk en welzijn te krijgen. Afgelopen zomer hebben we online en op afstand getest; hieruit bleek dat ons systeem prima met ouderen kan praten via videobellen.



Een gesprek tussen BLISS en een deelnemer op een conferentie.

Bij het uitvoeren van studies met gevoelige data moet men ook altijd rekening houden met het duidelijk verwoorden van de doelen van de studies om ethische goedkeuring te krijgen. Vanwege het type data dat we verzamelen (spraak) en de omslag van offline naar online studies hebben we lang moeten wachten op ethische goedkeuring van de studies.

Effect van corona

Door de coronacrisis is het aantal sociale contacten dat mensen hebben aanzienlijk kleiner geworden en dat geldt vooral voor ouderen. Wij zien daardoor goede mogelijkheden voor onze toepassing, niet alleen als metgezel om mee te praten, maar ook als een methode om meer inzicht te krijgen in de effecten van de coronamaatregelen op iemands welzijn en dagelijks leven. We zijn momenteel bezig om een studie op te zetten voor het najaar van 2020 waarin het sprekende systeem zal praten met ouderen over hun sociale contacten en activiteiten voor en tijdens de coronacrisis. •

WAT IS ER GEBEURD IN DE WERELD?

De betekenis van het nieuws

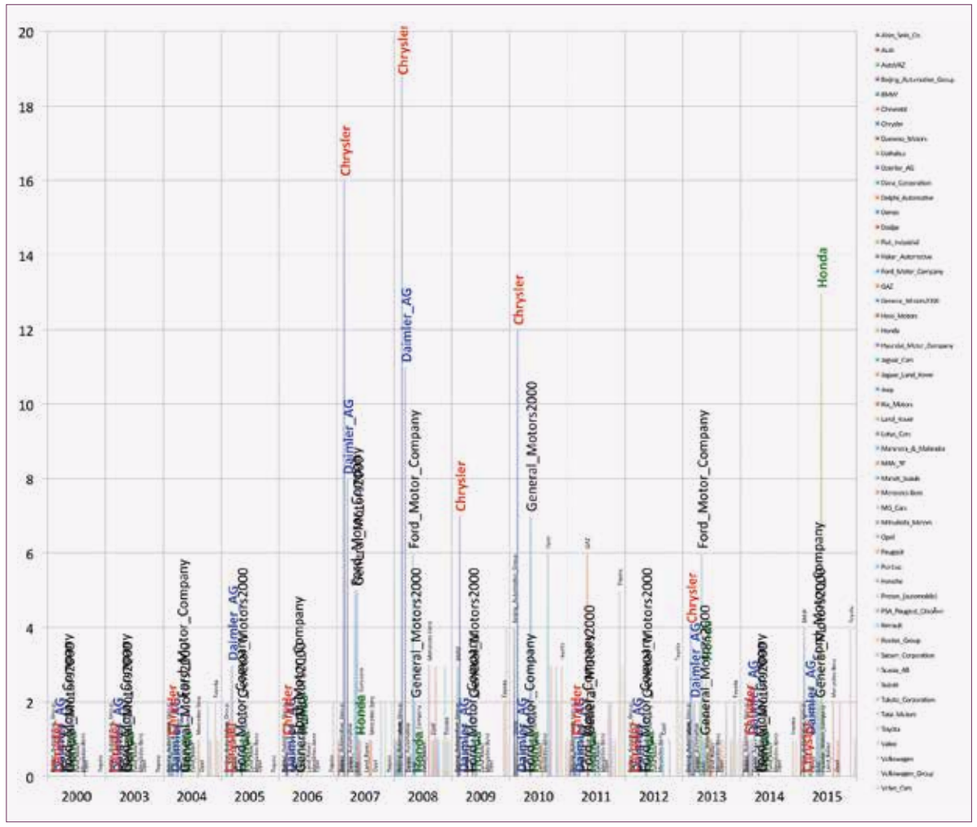
Vroeger las je het nieuws en wist je wat er speelde in de wereld. Tegenwoordig worden er meer dan een miljoen nieuwsartikelen gepubliceerd per werkdag in meer dan dertigduizend bronnen en in verschillende talen. Wie gaat dat allemaal lezen, wie kan dat allemaal bevatten om er zijn voordeel mee te doen?

Door:
Piek Vossen,
Vrije
Universiteit
Amsterdam

Het NewsReader project ontwikkelde een leesmachine die per dag meer dan honderdduizend nieuwsartikelen kan lezen in vier talen. Deze leesmachine bepaalt het ***wat***, ***wie***, ***waar*** en ***wanneer*** van ieder artikel en verzamelt het resultaat in een zogenaamde Knowledge Graph (zie figuur-1). Knowledge Graphs registreren die gebeurtenissen volgens een betekenismodel zodat je weet **wie** de betrokken personen en organisaties zijn, om **wat** voor soort gebeurtenissen het gaat en **waar** op de wereldbol het heeft plaatsgevonden.

Automobielindustrie

Deze kennis cumuleert dag na dag en vormt een geschiedenis van de wereld. De leesmachine wordt daarom ook wel een 'geschiedenisrecorder' genoemd. NewsReader heeft bijvoorbeeld meer dan twee miljoen artikelen gelezen over de automobielindustrie voor de periode 2000 tot 2015 en daar 1.2 miljard feiten uit gedistilleerd. Met een druk op de knop haal je uit de Knowledge Graph alle juridische gebeurtenissen waarbij een manager van een automobielbedrijf is betrokken.



Alle juridische gebeurtenissen in 2.3 miljoen nieuwsartikelen waarbij een manager van een autobedrijf betrokken is tussen 2000 en 2015.

Nieuwsartikelen verhalen over dezelfde wereld en deels over dezelfde gebeurtenissen. De grootste uitdaging van het project is om van ieder artikel te bepalen of ze het over dezelfde gebeurtenis hebben. De ene overname of acquisitie is de andere niet. Heeft een bank een keer een investering gedaan van een miljoen euro of ging het om twee of drie verschillende investeringen? Was het de ene keer een miljoen en de andere keer anderhalf miljoen of ging het om verschillende bedrijven?

NewsReader leest het nieuws in het Engels, Nederlands, Spaans en Italiaans maar interpreteert die artikelen in hetzelfde model. Er wordt dus ook over taalgrenzen heen bepaald of men het over dezelfde zaken heeft en wie wat beweert.

Taaltechnologie

NewsReader analyseert taal om tot een interpretatie te komen in de vorm van een kennisgraaf met eigenschappen van gebeurtenissen. Een breed scala aan taaltechnologie wordt ingezet die uiteindelijk voor iedere zin bepaalt welke woorden we gebeurtenissen noemen, wat entiteitsnamen zijn en wat de identiteit van die entiteitsnamen is. Ook wordt bepaald welke rollen entiteiten vervullen, hoe plaatsaanduidingen en tijdsuitdrukkingen te interpreteren en ten slotte wie de bron is. Dus is "Ford" een president, een automerk, het bedrijf zelf of de oprichter van het bedrijf; vervult "Ford" de rol van koper of verkoper en is "Ford" de auteur of is het een aangehaalde bron en wat vindt die bron ervan?

GRaSP

Het model om die interpretaties van het nieuws te kunnen vastleggen heet GRaSP, wat staat voor Grounded Representation and Source Perspective framework. In GRaSP wordt de berichtgeving in het nieuws opgeslagen als een bewering van een bron. Zo kan van iedere gebeurtenis worden nagegaan welke bron wanneer wat heeft beweerd, wie elkaar tegenspreken, bevestigen of aanvullen. Op die manier kan inzicht worden gekregen in het perspectief op de wereldgebeurtenissen vanuit de verschillende bronnen. •

VAN HERKENNEN NAAR BEGRIJPEN

Niet wat zeg je, maar wat bedoel je…

Door:
Nadine Glas,
Michel Boedeltje en
Arjan van Hessen,
Telecats

Abstract

Twee decennia geleden hielden we bij presentaties van spraakherkenningsresultaten onze klanten voor dat eenvoudige vragen het best met toetsen afgehandeld konden worden. Iets complexere vragen met veel mogelijke antwoorden konden dan met spraakherkenning gedaan worden en voor de echt lastige en emotionele zaken had je altijd een mens nodig. Medewerkers in het Klantcontactcentrum hoefden niet te vrezen voor hun baan! Maar dat was toen en in de jaren erna veranderden stap voor stap zowel het niveau als het type van de dialogen waarbij Automatische Spraakherkenning (ASR) toch succesvol gebruikt kon worden.

En dat zie je terug in de dialogen zoals de meeste bedrijven die tegenwoordig voeren. Natuurlijk, de mens speelt daarin nog steeds een dominante rol, maar we zien een verschuiving naar het tijdstip waarop de mens in de dialoog gaat bijdragen.

Waar 10 jaar geleden het bijna altijd of mensen of machines waren die de dialoog voerden, zien we tegenwoordig een duidelijke toename van hybride dialogen. De computer start het gesprek en vraagt een aantal zaken. Afhankelijk van het antwoord wordt het gesprek, inclusief de gegeven antwoorden, naar een specifieke persoon of afdeling gestuurd. Maar over het algemeen levert de huidige ASR erg goede resultaten en mogen we verwachten dat, wanneer de volgende grens succesvol genomen wordt, de hoeveelheid dialogen die al-dan-niet geheel of op z'n minst gedeeltelijk door de software gevoerd zullen worden, sterk zal blijven stijgen. Hierdoor schuift het moment waarop het door mensen wordt overgenomen snel op en wordt er steeds meer door de computer gedaan.

Hybride

De vraag rijst dan: gaan we op korte termijn de mensen er tussenuit halen? Gaan we naar een situatie toe waarbij de software het geheel zonder menselijke tussenkomst kan en, als het kan, willen we dat dan ook?

Het antwoord hierop is niet eenduidig te geven. Er zullen altijd diensten blijven bestaan waarbij mensen het gesprek (gedeeltelijk) zullen voeren. Enerzijds zal dit gebeuren wanneer het call volume zo klein is dat het niet de moeite waard is om het te automatiseren, anderzijds zullen bedrijven zich willen onderscheiden door aan te geven dat je bij hen altijd direct met een mens kunt spreken. Ook is het goed voorspelbaar dat in dialogen met een hoog emotioneel karakter, de computer geen of slechts een ondersteunende rol krijgt toebedeeld.

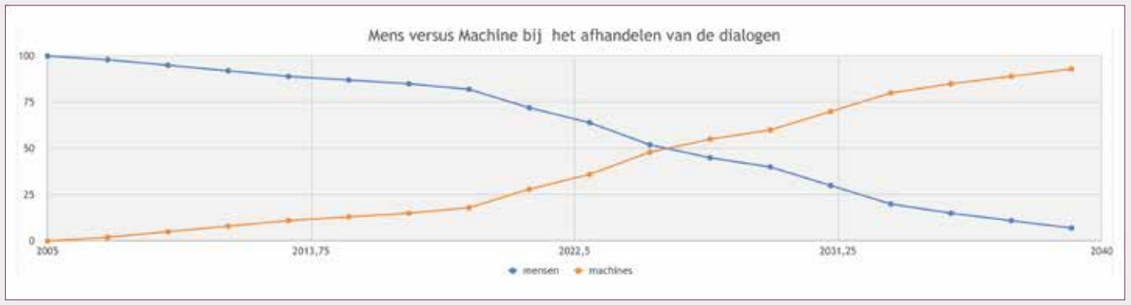
Bovendien zal er (als het goed is) altijd de mogelijkheid moeten blijven bestaan dat mensen in fout lopende gesprekken kunnen ingrijpen. Denk bijvoorbeeld aan gesprekken waarin mensen onduidelijk spreken, geen correct antwoord geven op gestelde vragen, rare woorden gebruiken (bijvoorbeeld veel jargon), of wanneer er erg veel achtergrondlawaai is. In de ideale situatie kan een medewerker dan of echt zijn intrede doen (*sorry, het gaat verkeerd, ik neem het gesprek nu even over*) of kan de ASR vervangen worden door menselijke spraakherkenning zonder dat dit direct in de dialoog duidelijk gemaakt wordt.

Grens

Deels zal de toename van de door software geleverde dialoog worden veroorzaakt door een betere verwerking van het geluid (spraak, achtergrondgeluid, ruis), betere taalmodellering en betere dialoogvoering.

Maar de belangrijkste verbeteringen zullen tot stand komen door het toenemende gebruik van DNN's: Deep Neural Networks die a) de verschillende parameters die uit de spraak gedestilleerd kunnen worden, steeds beter weten om te zetten in geschreven woorden en daarnaast b) de stap kunnen maken naar wat de vraag-achter-de-vraag precies is.

We zien immers dat we bij het verwerken van de vragen zoals we die nu herkennen, tegen een grens aanlopen. Zelfs met 100% correcte herkenning is het voor de huidige generatie software soms lastig om 'chocola te maken' van een ingesproken vraag of opmerking. Ook voor mensen is dit soms onmogelijk maar die zullen dan reageren met een vraag als "maar wat wilt u nu precies" en natuurlijk kunnen we dit soort vragen nu al door de computer laten stellen. Maar we moeten daar erg mee oppassen want je kunt dit soort vragen eigenlijk alleen maar stellen als je heel zeker weet dat je de spraak wel goed hebt herkend maar er desondanks niet uitkomt.



Ruwe schatting van het verloop van de verhouding mens-machine in de dialogen van de afgelopen 15 en komende 20 jaar. Wat we denken dat er zal gaan gebeuren, is dat de komst van goede AI + het gebruik van non-verbale features (alles dat NIET tot een woord leidt) er voor zal kunnen zorgen dat het aandeel van door de machines geleide dialogen, sterk zal toenemen.

Hoe doen mensen dat?

In de meeste gevallen begrijpen mensen slecht gesproken spraak veel beter dan de huidige ASR-engines doordat zij de spraak kunnen interpreteren: ze kunnen betekenis geven aan de uitgesproken reeks woorden en filteren als het ware de niet ter zake doende woorden eruit. En dat heeft niets met de herkenning van spraak te maken maar alles met het **begrijpen** ervan. Mensen horen **en** verwerken de spraak en kunnen vervolgens daar een bericht uit destilleren. De onvolkomenheden in de spraak worden vervolgens gewoon genegeerd.

Bij dit verwerken van de spraak gebruiken wij mensen echter meer dan alleen de herkende tekst. Wij horen de pauzes, de aarzelingen, de verbeteringen, andere akoestische zaken en passen die direct toe op het beeld dat wij van het gesprek vormen. We updaten als het ware constant en onthouden de gestelde vraag, maar minder de manier waarop die vraag gesteld werd.

Tekstprocessing

De volgende stap is dus het 'begrijpen' van de gesproken spraak, maar de vraag is "hoe doe je dat precies?" Om dit met software te kunnen nadoen, moeten we dus niet alleen de gesproken woorden proberen te herkennen, maar ook goed luisteren **hoe** die woorden werden uitgesproken, welke woorden vervolgens verbeterd werden en hoe andere spraakuitingen gerealiseerd werden.

Intent bepaling

Van alle binnengekomen berichten ga je als eerste een 'intent' bepalen: je vraagt aan menselijke beoordelaars die normaal gesproken ook met dit soort berichten werken, waar het bericht precies over gaat. Als je dat voor een paar honderd berichten hebt gedaan, dan geef je de data (de tekstuele vorm van het spraakbericht, de non-verbale parameters en de door mensen gegeven intent) aan een Machine Learning-algoritme (ML) dat vervolgens aan de slag gaat.
En dat kan op allerlei manieren. Zo kan er naar de woorden en cijfers gekeken worden, maar ook naar de pauzes, het ontbreken van bepaalde woorden, verdubbelingen, gebruik van tegenwoordige tijd of juist de verleden tijd. Maar, doordat ook de non-verbale features meegenomen worden, kan het systeem ook daar gebruik van maken: dus niet alleen **wat** maar ook **hoe** iets gezegd wordt.

De combinatie van deze twee manieren om naar het antwoord te kijken, maakt het dikwijls mogelijk om een redelijk betrouwbare voorspelling te doen van de intent van de gesproken boodschap.

Voorbeelden

Om een en ander duidelijk te maken, zullen we een aantal echte voorbeelden geven uit een Telecats-project waar wij de automatische intent-bepaling doen. Het gaat hier om een project waarin we enerzijds **vreemde** vragen krijgen en anderzijds **lastige** vragen.

Vreemde vragen

Dit zijn **moeilijke vragen** uit het oogpunt van de **mens**. Vragen waarvan de context geheel bekend moet zijn om ze te kunnen beantwoorden. Een voorbeeld van een vreemde vraag is: *Hallo, ik heb te weinig streepjes* (ik heb een probleem met mijn mobiele dekking).

Met de opmerking over streepjes wordt gerefereerd aan het symbool voor de internetsterkte op een mobiel apparaat. In het ideale geval

heb je drie streepjes maar dikwijls zijn het er minder. Als jij als helpdeskmedewerker niet precies weet wat er met die streepjes bedoeld wordt, dan kun je eigenlijk geen goed antwoord geven want wij mensen koppelen "streepjes" niet automatisch aan internetsterkte. Voor een algoritme is het echter geen punt dit correct te classificeren. De vragen over streepjes worden geclassificeerd als 'problemen met internet' en iedere vraag die over streepjes gaat kan dan het label 'problemen met internet' krijgen.

Iets anders wat we redelijk veel tegenkomen is dat er een 'te' ruim antwoord wordt gegeven op de gestelde vraag. Veel klanten gaven een antwoord als "ik wil iets weten over mijn laptop". Maar wat wilde iemand weten? Iets over zijn reparatie, een bestelde laptop of nog iets anders? Doorvragen is in die gevallen absoluut noodzakelijk. Bijvoorbeeld met: "en wat wilt u over uw laptop weten?"

Lastige vragen

Dit zijn **moeilijke vragen** uit het oogpunt van de **machine**. Vragen waarin de 'tone-of-voice', de pauzes en andere spraakeigenschappen nadrukkelijk een rol spelen in het duidelijk maken van de vraag en waar een puur schriftelijke versie dikwijls niet voldoet. Denk hierbij aan vragen waar we eigenlijk niet veel meer mee kunnen dan nog een keer vragen of gokken wat er bedoeld wordt.

- *U moet mij helpen met uh bepaalde gegevens snel mogelijk en uh wat kan ik zelf niet doen dat moet u doen dus vandaar dus een medewerker.*
- *Als ik op 4G ze kan uh opent app ze niet.*
- *Dag ik dan van mijn mobiele wifi is er mee opgehouden ik heb gebeld en hij zou uh van de op afstand worden uh opgeladen uh mij wij zijn nu een uur verder.*

Extra lastig zijn de zinnen waarin meerdere onderwerpen worden benoemd, en wij moeten bepalen welk onderdeel in dit geval de meeste prioriteit heeft.

- *Ik belde om op te zeggen omdat het elke keer niet werkt maar kreeg als oplossing een nieuw kastje toegestuurd maar nu wil ik weten wanneer mijn kastje komt.*
- *Een storing uh waarschijnlijk uh omdat uh zodat het op dat wil ik kan geen pakket ik kan geen pakket opzeggen daarvoor al daardoor.*

Soms lijkt een vraag heel makkelijk te beantwoorden, ook met selfservice, maar dan blijkt er meer achter te zitten. Neem bijvoorbeeld de veel gestelde volgende vraag:

- *Wat is mijn abonneenummer?*

Dat nummer zou de selfservice eenvoudigweg uit een systeem kunnen halen en oplezen, maar na analyse bleek dat mensen hun abonneenummer opvroegen omdat zij via een formulier hun lidmaatschap wilden opzeggen. Door deze gesprekken te routeren naar een medewerker (en dus eigenlijk een andere betekenis aan deze zin te koppelen) werd gezorgd voor behoud van veel klanten.

Andere vragen

Het zal duidelijk zijn dat je van een enkel voorbeeld nooit kunt zeggen of het makkelijk of juist erg moeilijk is om die vraag te classificeren. Het classificeren gebeurt namelijk als onderdeel van alle binnenkomende vragen. Machine Learning gebruikt hier de grote hoeveelheden data voor om dit zo goed mogelijk te leren, maar mensen kunnen dit veel beter omdat ze meestal de echte vraag kunnen herkennen. En dat is natuurlijk wat wij ook graag zouden willen doen met onze ASR!

Conclusie

Bij Telecats zijn we al jaren bezig om de verhouding mens-machine in de door ons gemaakte dialogen richting machine (= software) te sturen. We doen dat door in eerste instantie zoveel mogelijk gebruik te maken van betere ASR-technieken (zowel aanpassing van het akoestisch model als het taalmodel). Dat gaat goed maar we merken dat we toch tegen een grens aanlopen die vaak te maken heeft met wat er nu precies bedoeld werd met de gestelde uiting.

OpenSmile

Om deze grens te passeren, moeten we niet alleen maar naar de herkende tekst kijken maar ook gebruikmaken van de manier **waarop** die uiting geformuleerd werd. We zijn nu hard aan het experimenteren met softwarepakketten als OpenSmile waarbij we de verschillende akoestische features die in het spraaksignaal zitten, proberen te gebruiken voor het verbeteren van de intentdetectie. Daarmee begeven we ons nog meer op weg naar het begrijpen van wat er bedoeld (en helaas niet altijd gezegd) werd. We gebruiken zaken als amplitude- en toonhoogte-verloop, pauzes, aarzelingen, herhalingen en andere, typisch akoestische zaken die nu nog verdwijnen wanneer we de boodschap alleen weergeven als tekst. •

De Furhat-robot

GESPROKEN KIND-ROBOT INTERACTIE

Maar dan wel op een verantwoorde manier

Zoeken naar informatie doen we allang niet meer door alleen maar zoektermen in te tikken op een zoekscherm. Informatiesystemen weten steeds meer van ons. Hierdoor kunnen ze al voorsorteren op onze informatiebehoefte. Het zoeken is 'browsen' geworden; dat is zoeken door informatie die gepersonaliseerd wordt aangeboden of aanbevolen. We hoeven ook niet meer de zoektermen in te tikken; de informatiesystemen maken het ons makkelijk door een luisterend oor te bieden. Het project *Kind in Gesprek met Media* gebruikt het hierboven geschetste scenario als startpunt om onderzoek te doen naar het gebruik van een robot om *kinderen* te helpen zoeken.

Informatiebehoefte

Met behulp van onze stem kunnen we ons een weg banen door een keur aan informatiebronnen die online beschikbaar zijn - van het weerbericht tot hoe je een ei kookt en van een moppentrommel tot Wikipedia. Linksom of rechtsom gaat het bij het zoeken naar informatie vooral om het scherp krijgen van onze eigen informatiebehoefte, om die behoefte vervolgens te koppelen aan de beschikbare informatie. Een slim zoeksysteem probeert zoveel mogelijk informatie te achterhalen die kan helpen om te bepalen wat we precies nodig hebben. Stel je nu voor dat dit slimme zoeksysteem een robot is, een 'embodied agent', die een scherm laat zien en vragen kan stellen. Dat biedt interessante mogelijkheden om onze informatiebehoefte nauwkeurig af te vangen. Zo'n robot kan jou voor zichzelf gaan personaliseren: wie ben jij? Wat kan ik over jou te weten komen op internet? Ben je een man of vrouw? Hoe oud zou je ongeveer zijn? Wat vraag je nu precies? Je kijkt teleurgesteld, ben je op zoek naar iets anders? Waar ben je precies? Helpt dat om beter in te schatten waar jij naar op zoek bent? Enzovoorts.

Interactie

Het project Kind in Gesprek met Media onderzoekt hoe een robot kinderen het beste kan helpen zoeken. Een belangrijk aandachtspunt daarbij is de natuurlijke interactie tussen kind en robot met behulp van spraak. De robot herkent wat het kind zegt, maar kan ook signalen detecteren in de stem van het kind die iets kunnen zeggen over hoe het de interactie ervaart. Want we weten nog weinig over wat de beste manier is om kinderen en robots met elkaar te laten communiceren, met name hoe kinderen dit zelf ervaren.

Door:
Roeland Ordelman en Khiet Truong, HMI Universiteit Twente

De Furhat-robot

Perceptie

We kijken naar het vormgeven van gesproken interactie van kinderen met een robot en dus ook naar de perceptie van de interactie door de kinderen. Die zijn snel geneigd om robots als gelijken te zien en te vertrouwen. Maar is dat wel zo goed? Wat gebeurt er met hun spontane vertrouwen als we het gezicht veranderen, gebruikmakend van de mogelijkheden die de Furhat robot (zie afbeelding) hiervoor biedt? Wat gebeurt er als de robot een uitgesproken onbetrouwbaar uiterlijk heeft? Kunnen we het vertrouwen dat een kind in de robot stelt, aflezen uit de spraak van of conversatie met een kind?

Een belangrijk uitgangspunt bij dit onderzoek is dat het gebruik van de technologie niet alleen informerend is op een manier die kinderen snappen en leuk vinden, maar dus ook dat de technologie op een verantwoorde manier wordt toegepast. En misschien zelfs wel bewustwording bij kinderen stimuleert over het gebruik van dit soort 'artificiële intelligentie'.

Dit onderzoek kunnen we doen dankzij financiering van het Responsible AI programma van het SIDN fonds. We werken in het project samen met het Nederlands Instituut voor Beeld en Geluid waarbij we ons de situatie voorstellen dat kinderen in het Museum in Hilversum met de hulp van een robot willen zoeken in het rijke Beeld en Geluid archief. Ook werken we samen met het bedrijf WizeNoze dat in deze technologie is geinteresseerd vanuit het perspectief om leerlingen op school betrouwbare toegang tot informatie te bieden. •

Project Informatie:
https://www.utwente.nl/en/eemcs/hmi/hmi-projects-folder/hmi-project-child-robot-media-interaction/Human Media Interaction, Universiteit Twente

Promotoren:
Vanessa Evers en Theo Huibers

Onderzoekers:
Ella Velner, Thomas Beelen, Khiet Truong en Roeland Ordelman (PI)

Fondsen:
https://www.sidnfonds.nl/responsible-ai, CLICKNL, Nederlands Instituut voor Beeld en Geluid

DIAMANT, EEN DIACHROON SEMANTISCH LEXICON VAN DE NEDERLANDSE TAAL

De digitalisering van historisch tekstmateriaal heeft een enorme vlucht genomen. Bibliotheken en archieven spannen zich in om kranten, tijdschriften, boeken en handschriften online beschikbaar en doorzoekbaar te maken voor onderzoek en voor een groot publiek. Een van de uitdagingen hierbij is niet alleen de spelling van het historisch Nederlands, maar ook het feit dat de Nederlandse woordenschat in de loop der eeuwen behoorlijk is veranderd. Sommige woorden hadden in het verleden een andere betekenis en er zijn begrippen die met andere woorden werden uitgedrukt. Met het computationeel lexicon DiaMaNT (Diachroon seMantisch lexicon van de Nederlandse Taal) wordt inzichtelijk gemaakt hoe een bepaald concept in de geschiedenis van het Nederlands verwoord werd. Het lexicon vormt onderdeel van de infrastructuur voor historisch Nederlands, ontwikkeld bij het Instituut voor de Nederlandse Taal (INT) en is beschikbaar als Linked Open Data[1].

Door:
Katrien Depuydt en Jesse de Does;
Instituut voor de Nederlandse Taal

De historische woordenboeken van het INT bevatten een schat aan informatie voor iedereen die geïnteresseerd is in historisch Nederlands en historische documenten. Je kunt er de herkomst en de betekenis van woorden in opzoeken en door de voorbeeldzinnen die de betekenissen illustreren, krijg je ook een beeld van de spelling die toen werd gebruikt. Om de data bruikbaar te maken voor computationele toepassingen wordt er sinds 2005 gewerkt aan een lexicografische infrastructuur voor historisch Nederlands. De kern van de infrastructuur zijn de vier historische wetenschappelijke woordenboeken voor het Nederlands, die gezamenlijk een periode bestrijken van ca. 500 tot en met ca. 1976: het *Oudnederlands Woordenboek* (ONW), het *Vroegmiddelnederlands Woordenboek* (VMNW), het *Middelnederlandsch Woordenboek* (MNW) en het Woordenboek der Nederlandsche Taal (WNT).

Historische woordenboeken online

Het woordenboekportaal (gtb.ivnt.org) was de eerste module van de infrastructuur. De woordenboeken zijn in eenzelfde XML-structuur omgezet en beter doorzoekbaar gemaakt door aan ieder woordenboeklemma ook een modern Nederlands lemma toe te kennen.

GiGaNT

De volgende stap was de ontwikkeling van het diachrone morfosyntactische lexicon GiGaNT (Groot Geïntegreerd lexicon van de Nederlandse Taal), een computationeel lexicon dat gestructureerde informatie over de woordenschat geeft. GiGaNT geeft informatie over woorden, hun verbuiging en spellingvariatie in de Nederlandse taal vanaf de 6e eeuw tot heden. Bouwmateriaal voor de historische component van GiGaNT zijn de lemmata en de voorbeeldzinnen (attestaties) van de historische woordenboeken, momenteel MNW en WNT. De voorbeeldzinnen zijn citaten uit diverse bronnen en voorzien van bronvermelding en datering. GiGaNT blijft gelinkt aan de oorspronkelijke woordenboeken via persistent identifiers (unieke en onveranderlijke identificatiecodes) van de woordenboekartikelen en de attestaties. Omdat de attestaties gedateerd zijn, is het duidelijk welke spelling in welke periode van toepassing is. Aan GiGaNT wordt nog steeds gewerkt. Het wordt onder andere ingezet door de KB in Delpher (www.delpher.nl) om alternatieve spellingvarianten bij een zoekterm aan te bieden.

DiaMaNT

In 2015 werd gestart met de derde module, DiaMaNT (Diachroon seMantisch lexicon van de Nederlandse Taal). Daar waar het woordenboekenportaal een oplossing biedt om de betekenis van een woord op te zoeken en GiGaNT informatie geeft over mogelijke spellingen en verbuigingen van woorden, biedt dit computationeel lexicon de gebruiker middelen om in historische teksten te zoeken naar een concept waarvoor hij of zij alleen een modern Nederlands woord kent. Het lexicon wil ook historisch taalkundigen beter ondersteunen om diachrone semantische variatie op een systematische manier te bestuderen.

In het lexicon worden de *diachrone onomasiologie* (de verandering in de lexicale expressie van concepten) en de *diachrone semasiologie* (de verandering in de betekenis van woorden) geïncorporeerd op een manier die geschikt is voor gebruik door mensen en computers. Het ono-

[1] https://www.w3.org/DesignIssues/LinkedData.html

masiologische deel van het lexicon ondersteunt het zoeken in tekstmateriaal door verschillende termen voor een concept aan te bieden (*slager → houwer, beenhakker, vleeshouwer; boer → landman, dorper*, ...). De diachrone semasiologische component stelt de gebruiker in staat rekening te houden met semantische veranderingen zoals de verschuiving van de betekenis van appel van vrucht in het algemeen naar de meer specifieke moderne betekenis.

Hoe komt het lexicon tot stand

Het lexicon voegt een semantische laag toe aan het woordvormenlexicon GiGaNT op basis van de definities uit de woordenboekartikelen. Ten behoeve van de onomasiologische laag werden alternatieve termen voor een concept geëxtraheerd door een patroongebaseerde analyse van de definities. Deze extracties (met een accuratesse van rond de 80%) zijn vervolgens in een lexicografische bewerkingsomgeving gecorrigeerd. Het heeft geresulteerd in een lexicon met meer dan 300.000 (bijna) synoniemrelaties uit het MNW en het WNT.

Voor de realisatie van de diachronie zijn de gedateerde attestaties van de woordenschat van groot belang. Terwijl in andere benaderingen van diachronie met periodelabels voor een woordbetekenis gewerkt wordt, laten we hier de periodisering bottom-up uit de data ontstaan. Net als voor GiGaNT wordt de relatie met de woordenboekdata via de persistent identifiers behouden. De semasiologische component wordt momenteel verder vormgegeven door de betekenisrelaties binnen de woordenboeken te formaliseren.

Linked Open Data

Voor gebruik in computationele toepassingen, en om de data makkelijk koppelbaar te maken aan andere (lexicale) datasets, is ervoor gekozen de data te publiceren als Linked Open Data, gebaseerd op het Ontolex-Lemon[2] model dat zich de afgelopen jaren snel tot een de facto standaard ontwikkeld heeft. Daarbij is het nodig gebleken dit model uit te breiden met een attestatie-component, die inmiddels verder ontwikkeld wordt door een internationale werkgemeenschap[3].

Toepassingen

Met de publieksapplicatie op diamant.ivdnt.org kan de onomasiologische component worden geraadpleegd. Gebruikers kunnen kiezen voor een globaal chronologisch overzicht, een gedetailleerd overzicht inclusief de attestaties en een grafische weergave van het semantisch netwerk rond een concept.

Het lexicon is reeds gebruikt in verschillende onderzoekspilots[4] in het CLARIAH-project. De lexiconinhoud is opgeleverd binnen de CLARIAH infrastructuur en gepubliceerd als de dataset concepten (https://anansi.clariah.nl/).

Verdere ontwikkeling

Op de agenda staat onder andere een verdergaande koppeling van concepten en verbalisaties door het aanbrengen van een synset-structuur en verdere invulling van de semasiologische component. Hierbij zal gekeken worden of bestaande technieken om woordbetekenissen te koppelen kunnen helpen. •

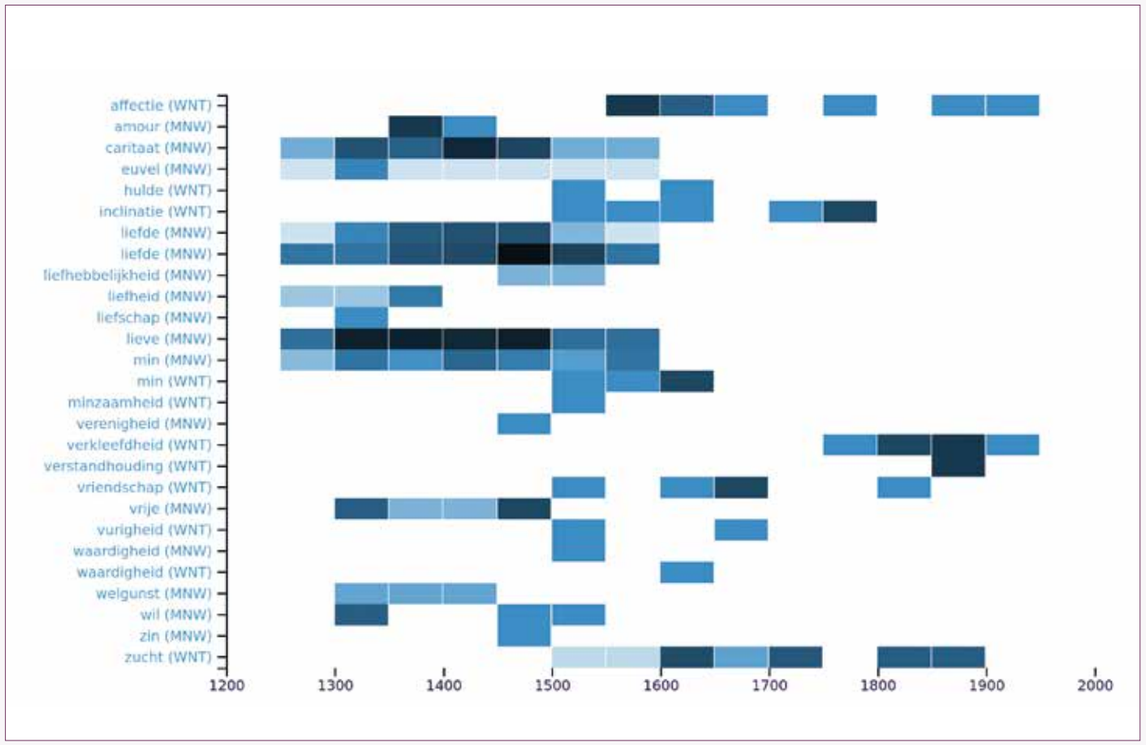
---

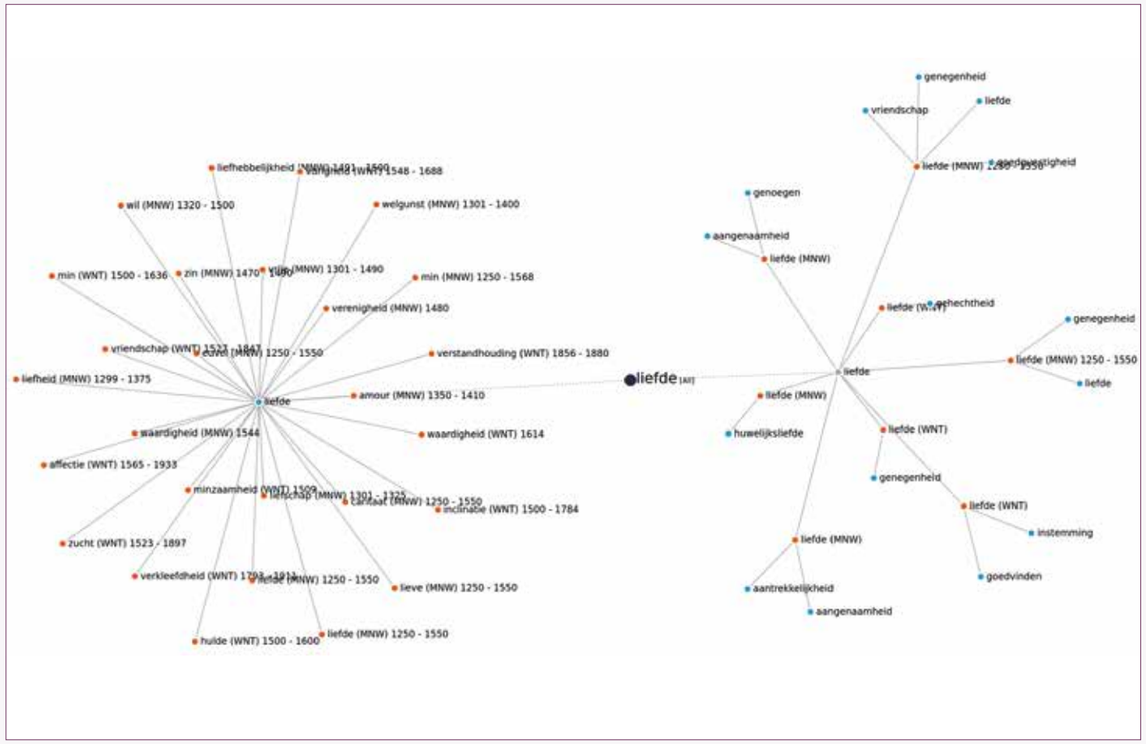
[2] https://www.w3.org/2016/05/ontolex/

[3] K. Depuydt en J. de Does (2018), "The Diachronic Semantic Lexicon of Dutch as Linked Open Data." In: I. Kernerman and S. Krek, Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets". [Miyazaki], 2018, pp. 23-28.

[4] Karin Hofmeester, Ashkan Ashkpour, Katrien Depuydt en Jesse de Does, "Diamonds in Borneo: Commodities as Concepts in Context." In: DATeCH2019: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage. Association for Computing Machinery, New York 2019, p. 45-50. https://doi.org/10.1145/3322905.3322924 ; Marieke van Erp, Jesse de Does, Katrien Depuydt , Rob Lenders and Thomas van Goethem (2018) . "Slicing and Dicing a Newspaper Corpus for Historical Ecology Research." In: Proceedings of European Knowledge Acquisition Workshop – EKAW (2018).

Lemmas for which liefde is a synonym / temporal overview





amour

Dictionary: MNW

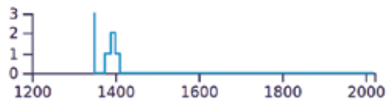
Back to top

1350 - 1410

Meaning: Liefde, inzonderheid mingenot.

Citations: 1350, 1350, 1390-1410, 1350, 1376-1400

Citations per year:



Go to dictionary entry

caritaat

Dictionary: MNW

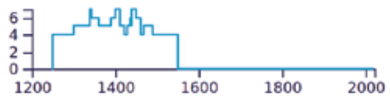
Back to top

1250 - 1550

Meaning: Liefde, de hoogste liefde, reine liefde, christelijke liefde, liefde tot den naaste, in tegenstelling met de aardsche of zinnelijke liefde. Zoowel de liefde van God en Christus tot den mensch, als de liefde van den mensch tot God en tot zijne medemenschen om den wil van God. Zie eene uitvoerige beschrijving der caritate, D. Doct. I, 614;687. Duc. "agapè Christianorum."

Citations: 1250-1550, 1430-1450, 1470-1490, 1301-1400, 1339-1341, 1340-1360, 1440-1460, 1390-1420, 1401-1410, 1250-1550, 1250-1550, 1438, 1401-1410, 1250-1550, 1440-1460

Citations per year:



Go to dictionary entry

*Termen voor het concept **liefde** en verschuivingen in betekenis*

NOTaS

STICHTING NOTAS

Secretariaat Stichting NOTaS
Toernooiveld 300,
6525 EC Nijmegen
T: 024-351 21 08
W: www.notas.nl
E: info@notas.nl

Stichting NOTaS behartigt de belangen van bedrijven en kennisinstellingen in de sector Taal- en Spraaktechnologie (TST). Dit doet zij onder meer door het organiseren van bijeenkomsten, lobbyactiviteiten en de uitgave van het tijdschrift DIXIT.

DEELNEMERS STICHTING NOTAS

CLST, Radboud Universiteit
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 en de inzet van TST ten behoeve van het onderwijs en de zorg. Tevens rekenen wij ontwikkeling en curatie van data en tools tot onze expertise.

Data Archiving and Networked Services (DANS)
Anna van Saksenlaan 51,
2593 HW Den Haag
T: 070-349 44 50
W: www.dans.knaw.nl
E: info@dans.knaw.nl

DANS is het Nederlands instituut voor permanente toegang tot onderzoeksgegevens. Onze activiteiten zijn gecentreerd rond drie kerndiensten: data-archivering, datahergebruik en training en consultancy inzake FAIR datamanagement. Hiertoe bieden we het gecertificeerde langetermijn-archief EASY aan en DataverseNL voor het opslaan en delen van data tijdens het onderzoek, evenals het NARCIS-portal voor Nederlandse onderzoeksinformatie. DANS is een instituut van KNAW en NWO.

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

Stichting Dedicon gelooft in een wereld waarin iedereen moet kunnen lezen en leren wat hij wil in een vorm die hem past. Daarom ontwikkelen en produceren wij aantrekkelijke alternatieve leesvormen. Inmiddels maken tienduizenden mensen met veel plezier gebruik van onze diensten. Binnen de diensten van Dedicon speelt de moderne taal- en spraaktechnologie een belangrijke rol.

Fluency
Prins Hendrikkade 159a,
1011 TB Amsterdam
T: 06-14502731
W: www.fluency.nl
E: info@fluency.nl

Fluency maakt tekst-naar-spraaksoftware voor het Nederlands en het Fries. De belangrijkste producten zijn Fluency TTS en Spika. Spika is een handig hulpmiddel bij dyslexie en andere leesproblemen, dat direct vanaf een usb-stick werkt. Fluency TTS is een uitgebreider product, dat via SAPI ook gebruikt kan worden in combinatie met screen readers voor blinden, en andere toepassingen. Beide producten bieden de gebruiker een ruime keuze aan stemmen.

Instituut voor de Nederlandse Taal
Rapenburg 61,
2311 GJ Leiden
Postbus 9515,
2300 RA Leiden
T: 071 – 514 16 48
W: www.ivdnt.org
E: secretariaat@ivdnt.org

Het Instituut voor de Nederlandse Taal is dé plek voor iedereen die iets wil weten over het Nederlands door de eeuwen heen. Het is een breed toegankelijk wetenschappelijk instituut dat alle aspecten van de Nederlandse taal bestudeert, waaronder de woordenschat, grammatica en taalvariatie. Het instituut verzamelt de nieuwste Nederlandse woorden, actualiseert belangrijke naslagwerken zoals de Algemene Nederlandse Spraakkunst en maakt vaktaal toegankelijk via terminologielijsten. Daarnaast fungeert het instituut als kennis- en distributiecentrum voor digitale Nederlandstalige tekstverzamelingen, woordenlijsten, wetenschappelijke woordenboeken, spraakcorpora en taal- en spraaktechnologische software.

Knowledge Concepts
Roond 22,
5281 SH Boxtel
T: 06-22969561
W: www.knowledge-concepts.co
E: sales@knowledge-concepts.com

Knowledge Concepts realiseert oplossingen ten behoeve van informatieverwerking en data-analyse. Onze oplossingen omvatten het verzamelen, samenvoegen, verrijken, classificeren, categoriseren en ontsluiten ten behoeve van intelligent (her) gebruik van informatie en kennis. Door innovatief gebruik van taaltechnologie te combineren met zoekmachines en document management systemen realiseren wij oplossingen voor zowel numerieke, tekstuele als engineering gerelateerde informatie. Wij doen dit voor zowel bedrijfsbrede data als Big Data oplossingen en voor zowel bedrijven als overheden.

LIACS, Universiteit Leiden
Niels Bohrweg 1,
2333 CA Leiden
T: 071-527 57 72
W: www.liacs.leidenuniv.nl
E: secr@ liacs.leidenuniv.nl

Het Leiden Institute of Advanced Computer Science (LIACS) is het instituut voor onderzoek en onderwijs op het gebied van informatica en kunstmatige intelligentie aan de Universiteit Leiden. Het LIACS werkt actief samen met overheden en bedrijven. Achter de oplossingen voor maatschappelijk relevante problemen zitten theoretische ontdekkingen, met statistiek als belangrijke basis. Taaltechnologie heeft een groter wordend aandeel in het onderzoek en onderwijs van het LIACS.

Martha Hofman
Skoallestrjitte 16,
9178 GR Wânswert
T: 0519 333468,
06 19 858338
W: www.marthahofman.nl
E: mahofman@hetnet.nl

Taalkundige werkzaamheden voor taal- en spraaktechnologie, voor het Nederlands en het Fries. Lexicons, datasets en part-of-speech tagging voor tekst- en sentimentanalyse en spraaktechnologie.

Nederlands Instituut voor Beeld en Geluid
Media Parkboulevard 1,
1217 WE Hilversum
T: 035- 677 5555
W: www.beeldengeluid.nl
E: klantcontactcentrum@beeldengeluid.nl

Het Nederlands Instituut voor Beeld en Geluid is een van de grootste audiovisuele archieven van Europa en herbergt ruim 1 miljoen uur aan radio, televisie, film en muziek. Beeld en Geluid initieert ook onderzoek dat audiovisueel erfgoed open en doorzoekbaar maakt. We volgen relevante innovaties in audiovisueel archiveren, nemen deel in onderzoeksprojecten en experimenteren met nieuwe technologie. Daarvoor wordt bijvoorbeeld gebruik gemaakt van spraaktechnologie, zoals spraakherkenning en sprekerherkenning.

Oracle Cloud Service, Oracle Nederland
Hertogswetering 163,
3543 AS Utrecht
T: 030 6699 000
W: www.oracle.com/nl
E: fabrice.nauze@oracle.com

The Oracle Cloud Service Digital Assistant enables the most efficient way to engage with self-service consumers online, shaping next-generation customer experiences. The Digital Assistant provides a deeper understanding of customer intent and helps guide customers to high-value interactions to help improve customer satisfaction, increase customer loyalty, and drive higher conversion.

ReadSpeaker
Dolderseweg 2A,
3712 BP Huis ter Heide
T: 030 692 4490
W: www.readspeaker.com
E: nederland@readspeaker.com

ReadSpeaker is een wereldwijde speler op het gebied van tekst-naar-spraaktechnologie (TTS). Als enige provider levert ReadSpeaker het totale palet van stemmen, software en (maatwerk)oplossingen. ReadSpeaker is gevestigd in 15 landen en heeft wereldwijd meer dan 10.000 klanten in 65 landen en levert zowel SaaS- als licentieproducten. Voor verschillende kanalen, platformen en apparaten biedt ReadSpeaker bedrijven en organisaties een stem voor on- en offline content, embedded-, server- en desktopoplossingen, spraakproductie en meer. Met meer dan 20 jaar ervaring is het team van ReadSpeaker leidend in tekst-naar-spraak.

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 levert innovatieve klantcontactoplossingen met spraaktechnologie en kunstmatige intelligentie. Wij assisteren de klant door met spraakherkenning slim te routeren naar de meest geschikte medewerker. Voor en tijdens het gesprek tonen we de medewerker relevante informatie en antwoordsuggesties. We signaleren en rapporteren o.b.v. gesprekskarakteristieken en woordkeuze en leveren optimaal inzicht in klantcontact.

Tilburg University – Cognitive Science & Artificial Intelligence
Warandelaan 2,
5037 AB Tilburg
Postbus 90153,
5000 LE Tilburg
T: 013-466 81 18
W: www.uvt.nl
E: A.H.Verschoor@tilburguniversity.edu

In onderwijs en onderzoek van het Departement Cognitive Science & Artificial Intelligence (CS&AI) ligt het accent op aspecten van interactieve technologieën, zoals robotics en avatars, serious gaming, virtual & mixed reality, en taal- en data technologiën. Lopende onderzoeksprojecten betreffen computationele semantiek, dialoogmodellen, en computationele modellen van verbale en non-verbale menselijke communicatie. Het departement heeft nauwe contacten met het Jheronimus Academy of Data Science (www.jads.nl) en Mind Labs (www.mind-labs.eu). Het departement verzorgt onder andere de bloeiende bachelor- en masteropleidingen Cognitive Science & Artificial Intelligence, en is verantwoordelijk voor de interfacultaire Master Data Science & Society.

Universiteit Twente - HMI
Drienerlolaan 5,
7522 NB Enschede
Postbus 217,
7500 AE Enschede
T: 053-489 37 40
W: www.utwente.nl/en/eemcs/hmi/
E: HMI-SECR-L@lists.utwente.nl

De Human Media Interaction (HMI) groep van de Universiteit Twente (UT) is een onderzoeksgroep binnen de afdeling Informatica. HMI doet onderzoek op het gebied van mens-machine interactie naar het ontwerpen, ontwikkelen, en evalueren van (sociaal) interactieve systemen (o.a. gesproken dialoogsystemen voor virtuele agents en sociale robots, interactive storytelling, multimedia retrieval) die de gebruiker ondersteunen op een plezierige en efficiënte manier. Het automatisch analyseren en interpreteren van menselijk verbaal en non-verbaal gedrag op basis van taal en spraak, fysiologische maten en lichaamstaal speelt daarbij een belangrijk rol. De generatie van verbaal en non-verbaal gedrag van robots en virtual agents behoort tevens tot het onderzoek. Daarnaast verzorgt de HMI groep veel onderwijs voor de masteropleiding Interaction Technology (met cursussen zoals Natural Language Processing, Speech Processing, Conversational Agents, en Affective Computing) en de bacheloropleiding Creative Technology.

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.

SAMENWERKING

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

Als je niet weet wat telt, hoe maak je dan het verschil? De mooiste dingen die het verschil maken, liggen niet altijd voor de hand. Deykerhoff Accountants helpt ondernemers in het MKB en zelfstandig professionals hier persoonlijk bij.

Leonard Marketing & Communicatie
W: www.leonard.nl
E: hello@leonard.nl

Leonard Marketing & Communicatie, Online Marketing Communicatie specialist, zorgt voor de vormgeving en het drukwerk van DIXIT.

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 biedt office management en secretariaat en is gespecialiseerd in de ondersteuning van verenigingen, stichtingen, besturen en commissies.

Telecats by Webhelp

Spraaktechnologie Vanzelfsprekend in klantcontact