TAALTECHNOLOGIE ONTRAFELD

pelckmans.be
Brasschaatsteenweg 308, 2920 Kalmthout, België

Omslagontwerp: Armée de Verre Bookdesign
Vormgeving: Crius Group
Illustraties: Studio Sans

D/2024/0055/254
ISBN 978 94 6310 622 1
NUR 616, 620
THEMA CFP, CJAB

pelckmans.be
facebook.com/pelckmans.be
twitter.com/Pelckmans_be
instagram.com/pelckmans.be

TAALTECHNOLOGIE ONTRAFELD

Hoe taal en technologie hand in hand gaan

CYNTHIA VAN HEE & VÉRONIQUE HOSTE (RED.)

m.m.v.

ORPHÉE DE CLERCQ, ELS LEFEVER, LIEVE MACKEN,
ARDA TEZCAN, A. SEZA DOĞRUÖZ & JOKE DAEMS

P E L C K M A N S

INHOUD

VOORWOORD: DE KRACHT VAN TAAL IN HET TIJDPERK VAN TECHNOLOGIE

Wil je meer weten over taaltechnologie, maar weet je niet goed waar te beginnen? Dan heb je het juiste boek in handen! Hoewel je je misschien maar moeilijk iets kunt voorstellen bij taaltechnologie, of je enkel aan chatbots denkt, worden we de dag van vandaag omringd door allerlei apps die gebruikmaken van taaltechnologie. Denk bijvoorbeeld maar aan je smartphone en hoe je die gebruikt. Heb je een vraag? Dan surf je wellicht meteen naar een zoekmachine zoals Google. Wil je je tekst herschrijven in een meer academische stijl? Dan vraag je dat misschien aan ChatGPT. Heb je een bordje gezien met een opschrift in een taal die je niet begrijpt? Dan heb je mogelijk Google Translate of DeepL geraadpleegd. Wil je iemand opbellen achter het stuur van je wagen? Dan kunnen virtuele assistenten zoals Siri en Alexa dat voor je doen. De mogelijkheden zijn talrijk en steeds meer toepassingen worden een deel van ons dagelijks leven. De ontwikkelingen binnen het domein van artificiële intelligentie volgen elkaar razendsnel op en taaltechnologie is daar een onontbeerlijke component van. Taal kan immers beschouwd worden als de sleutel tot echte artificiële intelligentie; zonder een vorm van taalbegrip zullen slimme systemen en robots in hun functioneren namelijk altijd beperkt blijven tot vooraf geprogrammeerde taken. Enige vorm van interactiemogelijkheid opent meteen een breed spectrum van complexere en personaliseerbare taken. Hoe computers nu precies op een menselijke manier kunnen communiceren, lichten we toe in dit boek.

Dit boek wil een leidraad zijn in de snel veranderende digitale maatschappij en introduceert de basisprincipes voor iedereen die zich wil verdiepen in taaltechnologie. Of je doel nu is om de werking van taaltechnologische hulpmiddelen beter te begrijpen, om de tools slimmer te kunnen aanwenden of om ze effectiever toe te kunnen passen binnen communicatie- en marketingstrategieën.

De inhoud bestaat uit drie delen. In het eerste deel reiken we de bouwstenen aan van taaltechnologie, waarmee we toelichten hoe computers taal zowel kunnen begrijpen als produceren, en waarom die taken soms ook moeilijk blijken. In het tweede deel gebruiken we de bouwstenen uit het eerste deel om een aantal taaltechnologische toepassingen in detail te bespreken, zoals chatbots en automatische tekstgeneratie.

Het derde deel bespreekt de implicaties van taaltechnologie voor ons dagelijks leven en onze privacy en snijdt daarmee enkele ethische vraagstukken aan.

Aangezien taaltechnologie niet meer weg te denken is uit onze maatschappij, reiken we in het tweede deel een aantal praktische richtlijnen aan om zelf met taaltechnologie aan de slag te gaan, of je nu actief bent in het onderwijs, als marketeer, taalprofessional of communicatiespecialist, of gewoon interesse hebt in de ontwikkelingen en toepassingen van artificiële intelligentie.

Tijdens het lezen van dit boek zul je verschillende termen tegenkomen die in een **blauw kadertje** staan. De bijbehorende verklaringen vind je terug in de trefwoordenlijst achterin.

We hopen dat dit boek je enthousiasme voor taaltechnologie zal aanwakkeren en dat het je misschien zal inspireren om zelf met taaltechnologie aan de slag te gaan. En als er één boodschap is waarvan we hopen dat ze blijft nazinderen, dan is het dat taal en technologie hand in hand gaan. Vandaag meer dan ooit.

DANKWOORD

Dit boek is het gezamenlijke resultaat van maar liefst acht coauteurs die hun expertise in het domein van de taaltechnologie hebben vertaald naar een boek dat toegankelijk is voor het brede publiek. Bij dat schrijfproces hebben we kunnen rekenen op de hulp van een aantal mensen die we hier uitdrukkelijk willen bedanken.

Speciale dank gaat uit naar onze collega's Colin Swaelens, Joni Kruijsbergen en Amaury Van Parys voor de inhoudelijke en vormelijke revisie van de tekst. Jullie kritische blik en soms ongezouten commentaar hebben de helderheid en coherentie van het boek zonder twijfel verbeterd. Ook van harte dank aan Walter Daelemans voor zijn constructieve inhoudelijke feedback en aan Eva Lievens voor de pointers naar de recente juridische literatuur over artificiële intelligentie.

Dank aan Studio Sans om dit boek te voorzien van sprekende illustraties. Larissa, je creatieve talent heeft dit boek niet alleen een visuele flair gegeven, maar biedt lezers ook een leidraad bij moeilijke concepten. Bedankt!

DEEL 1

WAT IS TAALTECHNOLOGIE?

1. WAT IS TAALTECHNOLOGIE?

Cynthia Van Hee, Véronique Hoste, Orphée De Clercq, Els Lefever, Lieve Macken

Taaltechnologie en bij uitbreiding artificiële intelligentie (AI) zijn alomtegenwoordig in onze maatschappij. Krantenkoppen berichten over nieuwe technologieën, vooral wanneer de grote techbedrijven zoals Alphabet, Meta, OpenAI, IBM, Baidu en Neuralink hun doorbraken delen met de wereld. Denk bijvoorbeeld aan de kwaliteit van Google Translate tegenwoordig, of hoe dankzij de release van ChatGPT (OpenAI) eind 2022 een writer's block misschien voorgoed tot het verleden behoort. Daarnaast snijden opiniemakers ethische kwesties aan en worden jong en oud via sociale media overspoeld met berichten over technologische evoluties en automatisch gegenereerde inhoud.

Artificiële intelligentie wordt in toenemende mate gebruikt om toepassingen te verbeteren die ons dagelijks leven helpen te organiseren: een gps, een slimme thermostaat, een virtuele assistent, streaming- of muziekapps die lijsten opstellen van onze favoriete keuzes, filters die ons gepersonaliseerde reclame- en nieuwsberichten aanbieden enzovoort. Die toepassingen zorgen ervoor dat onze maatschappij wordt gekenmerkt door een overvloed aan digitale data: gegevens die met een duizelingwekkende snelheid het internet verrijken en inhoudelijk zeer gevarieerd zijn. Die gegevens, ook wel *big data* genoemd, bestaan voor een groot deel uit teksten en bevatten een schat aan informatie voor bijvoorbeeld beleidsmakers en bedrijven.

Die laatste zetten dan ook steeds vaker en meer in op (taal)technologie, en de vraag naar expertise in het domein van de natuurlijketaalverwerking stijgt exponentieel. Ook in het onderwijs neemt het belang van digitale geletterdheid toe om leerlingen niet alleen klaar te stomen voor een snel evoluerende arbeidsmarkt, maar ook om hen te helpen bedachtzaam om te springen met nieuwe technologieën en stil te staan bij de impact ervan op maatschappelijke evoluties en onze privacy.

TAALTECHNOLOGIE?

Taaltechnologie is dus een discipline binnen artificiële intelligentie. Wanneer we over taaltechnologie spreken of schrijven, hanteren we vaak de Engelse term *Natural Language Processing* (NLP) of '**natuurlijketaalverwerking**' in het Nederlands. We kunnen natuurlijketaalverwerking omschrijven als technologie aanwenden om taal te analyseren of te genereren. De term wordt doorgaans gebruikt om het onderzoeksdomein te benoemen. Daarnaast wordt de verwante term 'taaltechnologie' gebruikt om te verwijzen naar concrete toepassingen die mensen in staat stellen om te communiceren met computers. De complexe taken die taaltechnologische toepassingen moeten kunnen uitvoeren, kunnen we terugbrengen tot vier belangrijke deelprocessen van natuurlijketaalverwerking:

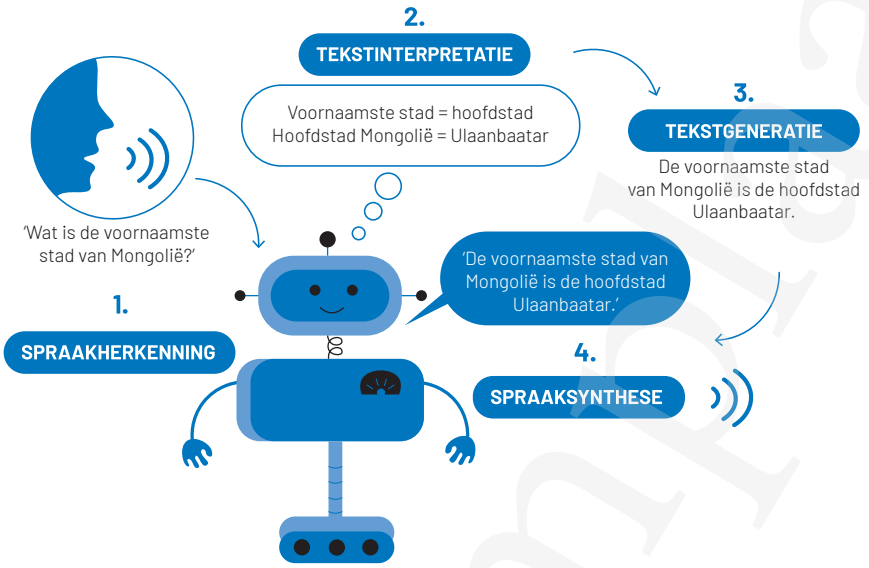
1. **Spraakherkenning**
2. **Tekstinterpretatie**
3. **Tekstgeneratie**
4. **Spraaksynthese**

Als je bijvoorbeeld een vraag stelt aan een virtuele assistent, dan ziet de invulling van die vier processen er als volgt uit:

1. **Spraakherkenning:** een stem herkennen en de klanken vervolgens omzetten in woorden.
2. **Tekstinterpretatie:** een vraag op een correcte manier begrijpen. Met andere woorden: achterhalen wat de intentie van de spreker is.
3. **Tekstgeneratie:** het correcte antwoord op een vraag vinden en dat antwoord vervolgens omzetten in de juiste woorden.
4. **Spraaksynthese:** een woordencombinatie omzetten in klanken.

Die processen vormen de vier belangrijkste onderzoeksdisciplines binnen het domein van de natuurlijketaalverwerking en zijn allesbehalve eenvoudig om te automatiseren. Als mensen communiceren we immers vlot in vreemde talen, leggen we niet altijd meteen onze kaarten op tafel, zijn we bijzonder creatief met taal en beschikken we bovendien over veel wereldkennis, maar daarover later meer.

In dit boek belichten we vooral de interpretatie en de generatie van geschreven taal. Spraakherkenning en spraaksynthese behoren tot het gespecialiseerde domein van de spraaktechnologie en worden niet verder uitgediept in dit boek.



Figuur 1.1. De vier belangrijkste deelprocessen van natuurlijketaalverwerking.

Praten over communicatie tussen mensen en computers leidt al snel tot het concept van kunstmatige of artificiële intelligentie. In Van Dale wordt artificiële intelligentie omschreven als 'het vermogen van een computer om gegevens te verwerken waarbij zoveel mogelijk wordt geprobeerd het menselijk denken na te bootsen'. Als we stilstaan bij het menselijk denken, dan kunnen we stellen dat alle menselijke kennis verweven is met taal (boeken, websites, alledaagse gesprekken enzovoort) en dat het kraken van de taalcode wellicht de ultieme sleutel is om artificiële intelligentie te bereiken. Meer nog, communicatie is cruciaal geweest voor onze historische overleving als mensheid en stelt ons vandaag in staat om verhalen te vertellen, samen te werken en de groepsdynamiek te creëren die essentieel is voor onze vooruitgang.

Om die redenen neemt natuurlijketaalverwerking een belangrijke plaats in binnen het domein van artificiële intelligentie en wordt er nauw samengewerkt tussen natuurlijketaalverwerking en diverse andere AI-domeinen, zoals robotica en beeldherkenning. Toch omvat natuurlijketaalverwerking meer dan alleen het streven naar artificiële intelligentie: het maakt automatische tekstanalyse mogelijk die van belang is voor tal van toepassingen en uiteenlopende domeinen, zoals communicatie, marketing, vastgoed, computerveiligheid en de financiële sector. Ondanks al die mogelijke toepassingen

blijft het modelleren van taal een buitengewoon complexe taak voor computers. Hoe dat komt, leggen we uit in het volgende hoofdstuk.

EEN WOORDJE OVER TERMINOLOGIE

In dit boek komen verschillende termen aan bod om gelijkaardige technologieën te beschrijven. Met de term 'natuurlijketaalverwerking' verwijzen we naar het gelijknamige onderzoeksdomein. Automatische systemen of toepassingen die binnen dat onderzoeksdomein ontwikkeld worden, noemen we 'NLP-systemen' of 'taaltechnologische systemen'. Beide termen vallen onder de noemer 'taaltechnologie'. Als we het hebben over '**machinelearning**' of 'lerende systemen', dan bedoelen we NLP-systemen die ontstaan door het trainen van een algoritme met behulp van een dataset.

In deel 3 van dit boek behandelen we ethische vraagstukken en maatschappelijke implicaties in de ruimere context van artificiële intelligentie. We gebruiken de term 'AI-systemen' om te verwijzen naar applicaties die niet alleen steunen op taaltechnologie, maar ook op andere technologieën, zoals automatische beeldherkenning.

EEN BEKNOPTE GESCHIEDENIS

De term 'artificiële intelligentie' is dan wel behoorlijk recent, vele technologieën en concepten bestaan al meer dan een halve eeuw. Sinds de naoorlogse periode in de jaren 1940 heeft het vakgebied een aantal mijlpalen meegemaakt. De belangrijkste daarvan geven we mee in het volgende overzicht.

WIST JE DAT ...
... het eerste computerprogramma wellicht door een vrouw werd geschreven? De Britse wiskundige Ada Lovelace wordt beschouwd als de eerste programmeur. Ze beschreef in 1843 al hoe een 'analytische machine' ingezet kon worden om getallen te berekenen en wetenschap te bedrijven. Tijdens de Tweede Wereldoorlog waren het bovendien vaak vrouwen die de computers programmeerden (met hendels, knoppen en kabels); veel mannen waren immers naar het front gestuurd.

Stroomversnelling door WO II.

1943

Uitvinding van de Colossus, de eerste elektronische computer, gebouwd om geheime codes van het leger te kraken.

Het domein evolueert razendsnel dankzij krachtiger geworden computers en toenemende interesse in taaltechnologie uit verschillende hoeken.

1938

Ontwikkeling van de eerste (toen nog mechanische) computer.

1954

1ste AI-mijlpaal: regelgebaseerde systemen

Onder leiding van de universiteit van Georgetown en IBM wordt het eerste automatische systeem gebouwd dat zestig Russische zinnen naar het Engels kan vertalen.

2018

5de AI-mijlpaal: generatieve modellen

Krachtige taalmodellen lossen niet alleen classificatietaken op, maar voorspellen woorden en kunnen zo tekst genereren en een gesprek met je voeren.

1990

2de AI-mijlpaal: statistische methodes

Automatische (vertaal)systemen bestaan niet uitsluitend uit regels, maar maken voorspellingen op basis van probabiliteiten.

Stroomversnelling door de uitvinding van het World Wide Web (WWW).

2010

4de AI-mijlpaal: deep learning

Lerende systemen bestaan uit diepe structuren met meerdere lagen en werken autonomer dan tevoren met gigantische taalmodellen als invoer.

1990–2000

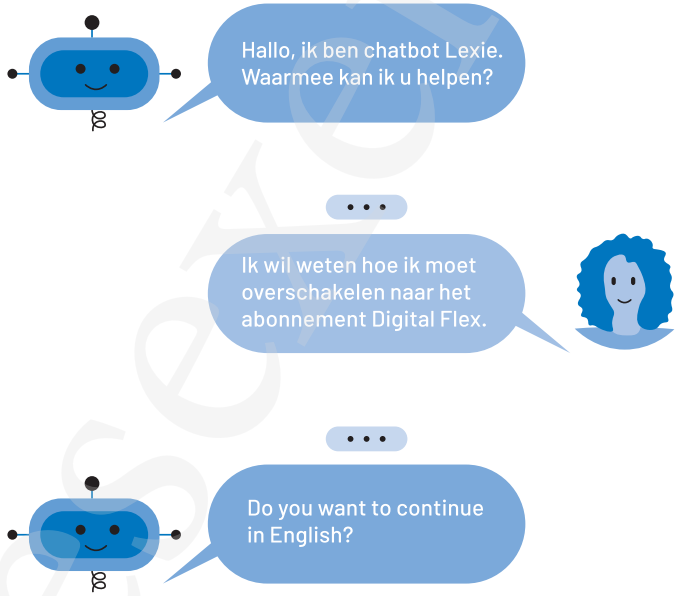
3de AI-mijlpaal: lerende systemen

Automatische systemen leren zelf regels uit data, wat we kennen als machinaal leren of machine-learning. Dankzij het WWW zijn talrijke datasets beschikbaar om systemen voor allerhande taken te trainen.

2. WAAROM IS TAAL ZO MOEILIJK?

Cynthia Van Hee, Véronique Hoste, Orphée De Clercq, Els Lefever, Lieve Macken

Ondanks de enorme technologische vooruitgang in de afgelopen decennia worden we dagelijks geconfronteerd met de beperkingen van intelligente systemen als het aankomt op taal. Denk bijvoorbeeld aan moderatiesoftware die haatboodschappen en pesterijen op sociale media laat passeren, virtuele assistenten of chatbots die maar niet begrijpen wat je zegt, of fouten die generatieve AI-toepassingen maken. Figuur 1.2 illustreert een chatbot die tijdens een Nederlandstalige conversatie plots in het Engels antwoordt omdat de gebruiker een Engelstalige term heeft gebruikt.



Figuur 1.2. Een gesprek tussen een klant en een chatbot waarbij het misloopt.