DIXIT

TIJDSCHRIFT OVER TAAL- EN SPRAAKTECHNOLOGIE

TST & Big Models

JAARGANG 19 - DECEMBER 2022

INHOUD

INHOUD	Pagina
Voorwoord en colofon	2
Het succes van BERT en Huggingface	3
LLM's en compute: de stand van zaken	5
Transformermodellen in Automatische Spraakherkenning	7
Hoe zoekmachines BERT gebruiken om relevante documenten te vinden	10
Producten indelen voor duurzame besluitvorming	12
Hoe BERT de medische wereld helpt	14
Debatten over abortus om te leren over kernenergie	16
Voorgetrainde Grote Taalmodellen in Automatische Vertaling	18
Wat kan taaltechnologie uit het heden met taal uit het verleden?	20
Emoties herkennen in spraak?	22
Op expeditie in de duistere binnenlanden van taalmodellen	24
De zorg transformeren met BERT	26
Large Language Models: geen Silver Bullet	28
De Voetafdruk van BERT	30
Het Asta-project, of de ontwikkeling van dialect-specifieke spraakherkenners	32

VOORWOORD

Deze editie van Dixit gaat over de onvermijdelijke opkomst van grote taalmodellen. Hoewel deze Transformermodellen sinds een jaar of vijf hun intrede hebben gemaakt zijn ze voor het grote publiek bekend geworden door de indrukwekkende tekstgeneratie van bijvoorbeeld GPT-3.

Wie tegenwoordig naar "Transformers" zoekt in een zoekmachine als DuckDuckGo zal nog steeds meer te weten komen over de buitenaardse robots van de planeet Cybertron uit mijn jeugd, maar zal in tweede instantie ook kennismaken met de Transformermodellen van HuggingFace, een belangrijk bedrijf op het gebied van het delen van pre-trained modellen.

Zoals u in deze DIXIT kunt lezen, lijken de toepassingen van deze modellen in de TST eindeloos en presteren ze vaak ver boven traditionele – durf ik het te zeggen, ouderwetse – modellen. Ze zijn ook makkelijk te gebruiken voor leken en zijn daarom een belangrijke factor in het democratiseren van het gebruik van TST.

En toch is niet alles rooskleurig in de wereld van de grote taalmodellen. De hoeveelheid resources (data en GPU's) die bijvoorbeeld nodig zijn om een redelijk model te trainen liggen buiten het bereik van de meeste bedrijven of universiteiten. Verder zijn de modellen even gevoelig voor de verschillende biases in de trainingsdata als hun voorgangers.

Een laatst onderbelicht aspect van de verandering veroorzaakt door deze modellen is misschien het nieuwe profiel van de TST-deskundige. Ik vraag me in het bijzonder af hoeveel (computationele) taalkundige kennis in de nabije toekomst nog nodig zal zijn wanneer de grote taalmodellen alles uit de data al schijnen te leren? Zal achter het democratiseren van het gebruik van grote taalmodellen de kloof tussen gebruikers en makers nog groter worden?

Deze DIXIT is in ieder geval een goed startpunt voor een discussie over de Transformermodellen en ik hoop van harte dat het u zal stimuleren om bij een van de komende NOTaS-meetings te komen om met ons in gesprek te gaan!

Veel leesplezier,

Fabrice Nauze,
voorzitter Stichting NOTaS

DIXIT: Tijdschrift over toegepaste taal- en spraaktechnologie – 19e jaargang, editie TST & Big Models. 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:** Suzan Verberne: s.verberne@liacs.leidenuniv.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:** Breda Press **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.

HET SUCCES VAN BERT EN HUGGINGFACE

TAALTECHNOLOGIE IN 5 REGELS CODE

Wist u, lezer van deze DIXIT, dat het mogelijk is om sentimentanalyse te doen met Transformermodellen, in slechts 5 regels Python-code?

Kijk maar:

```
from transformers import pipeline
sentiment_pipeline = pipeline("sentiment-analysis")
data = ["I love you", "I hate you"]
output=sentiment_pipeline(data)
print(output)

[{'label': 'POSITIVE', 'score': 0.9998656511306763}, {'label': 'NEGATIVE', 'score': 0.9991129040718079}]
```


Door:
Suzan Verberne, Universiteit Leiden

Toegegeven, in de praktijk heeft regel 3 (data = ["I love you", "I hate you"]) wat meer voeten in de aarde, want je moet de data binnenhalen, opschonen en organiseren in een lijst zinnen, maar het toekennen van een sentimentcategorie aan deze data is daadwerkelijk zo eenvoudig als in dit voorbeeld.

Hoe kan dat?

Allereerst doordat de ontwikkelaars van de *transformers* library (vaak *Huggingface* genoemd naar het bedrijf Hugging Face) deze complexe modellen voor zoveel mogelijk mensen toegankelijk willen maken. Daarom bouwen ze functionaliteiten die je veel werk uit handen nemen, zoals kant-en-klare pipelines[1] voor taken die veel mensen willen gebruiken. Dat zie je terug in de eerste regel (from transformers import pipeline).

Ten tweede doordat het model dat wordt ingeladen al eerder door andere mensen is getraind om sentimentanalyse te doen; eigenlijk is het belangrijkste werk al gedaan. Bij transformermodellen bestaat trainen uit twee fasen: *pre-trainen* en *fine-tunen*. Met pre-trainen bedoelen we dat het model woordbetekenissen heeft geleerd uit grote – gigantische – hoeveelheden ongelabelde data. Fine-tunen is het toespitsen van het model op een bepaalde taak met voorgedefinieerde labels. Dit laatste is een vorm van supervised learning.

Classificeren

In mijn code voor sentimentanalyse heb ik niet aangegeven welk model gebruikt wordt om mijn zinnen te classificeren. Daarom gebruikt Huggingface een standaardmodel, voor sentiment is dat distilbert-base-uncased-finetuned-sst-2-english. Laten we eens even kijken wat deze lange naam betekent en aan de hand daarvan uitleggen hoe het werkt. We beginnen achteraan:

- 'english' betekent dat dit een model voor het Engels is. Hoe krachtig ze ook zijn, deze grote modellen werken nog altijd het beste wanneer ze voor één bepaalde taal getraind zijn.
- 'sst-2' staat voor Stanford Sentiment Treebank v2, een dataset om Engelstalige sentimentmodellen op te trainen en evalueren. De set bevat ongeveer 10.000 zinnen met de labels positief en negatief.

[1] https://huggingface.co/docs/transformers/main_classes/pipelines

- ‘finetuned’ betekent dat het model is gespecialiseerd voor sentimentanalyse met de labels positief en negatief. Het model is gefinetuned op sst-2.
- ‘uncased’ betekent dat het model getraind is op data zonder onderscheid van hoofdletters en kleine letters.
- ‘distilbert-base’ is de naam van het gebruikte basismodel. In dit geval DistilBERT, een efficiënter model dan de klassieke BERT.

Iedereen kan een model pre-trainen en finetunen en toevoegen aan Huggingface. Op het moment dat ik dit schrijf staan er 73.896 modellen[2] in Huggingface. (Oh wacht, het zijn er al 73.897.) We kunnen daarin zoeken, bijvoorbeeld naar een model voor sentimentanalyse dat is getraind op Twitterdata: cardiffnlp/twitter-roberta-base-sentiment. Of een model voor sentimentanalyse op Nederlandse tekst: DTAI-KULeuven/robbert-v2-dutch-sentiment. Dit kunnen we in ons vijfregelige script stoppen en dan kunnen we sentimentclassificatie voor Nederlands doen[3].

Handig?

Ja. Zomaar inzetbaar voor elk probleem? Nee, zeker niet. Ten eerste hebben we niet voor elke taal gefinetunede modellen voor elke taak. Zo zijn er wel gepretrainde modellen voor Pools, maar geen gefinetunede modellen voor sentimentanalyse in het Pools. Daarvoor zijn namelijk weer gelabelde data nodig en die zijn er niet altijd. En zelfs als voor een taak modellen bestaan, dan werken we in de praktijk vaak met domeinen en onderwerpen waarvoor data schaars zijn. Denk aan problemen in specifieke domeinen zoals het vinden van relevante rechtszaken, het extraheren van bijwerkingen van medicatie of het classificeren van producten op inkooplijsten voor het bepalen van hun ecologische voetafdruk.

Als we deze problemen willen oplossen, dan moeten we verder werken aan technieken om de bestaande modellen die getraind zijn op grote hoeveelheden data, te laten specialiseren door ze een klein aantal voorbeelden te geven van de nog onbekende taak. •

Luistertip

“The Real Problem with AI” with Emily Bender. Podcast-interview door Adam Conover https://open.spotify.com/episode/0b-Q94NUHXP02vGxGxjXRP0?si=YJYP-vWaT6qS-RC-mtlx4OQ

[2] https://huggingface.co/models

[3] Kijk maar: https://colab.research.google.com/drive/1C7YMFTi1j7CELXg9sfeEnbTSWPD6E12F?usp=sharing

LLM’S EN COMPUTE: DE STAND VAN ZAKEN

Large Language Models (LLM’s) hebben in het afgelopen jaar interessante nieuwe eigenschappen laten zien in de NLP-industrie. LLM’s zijn tot stand gekomen met een zeer grote hoeveelheid data en in het bijzonder veel rekenkracht (compute). Het budget om deze modellen te trainen kan oplopen tot soms wel miljoenen euro’s. Het is daarmee niet voor iedereen weggelegd zelf LLM’s te trainen. Toch zijn er diverse trends die deze grote modellen bij een breed publiek brengen.

Door:
Serge Cornelissen, Founder & Chief Technology Officer, Cornelistools B.V.

Steeds groter, steeds krachtiger

Large Language Models zijn vaak van transformers afgeleide taalmodellen zoals een GPT (Generative Pretrained Transformer). Gegeven een stuk tekst (de zogeheten prompt) is een GPT-model simpel gezegd in staat telkens het beste volgende woord in de tekst te voorspellen en zo zelf de tekst voort te zetten. Dit levert soms rare resultaten op, maar regelmatig ook verrassend geloofwaardige en correcte nieuwe teksten. Naast tekstgeneratie kunnen LLM’s ingezet worden voor allerlei NLP-taken en in bijna iedere NLP-taak toont een LLM verbeteringen ten opzichte van zijn kleinere voorganger.

Een belangrijke reden waarom LLM's steeds groter gemaakt worden, is dat LLM's zogeheten ‘emergent phenomena’ lijken te vertonen. Dit wil zeggen dat de modellen niet alleen steeds betere resultaten laten zien als ze groter worden, ze laten ook zien dat ze een geheel andere set aan taken kunnen uitvoeren dan alleen de taken waar ze oorspronkelijk voor ontworpen zijn.

Google PaLM (4/2022), een taalmodel getraind met zowel 8B, 62B als 540B parameters (1B = 1 billion, 1 miljard), laat verrassend zien dat voor bepaalde NLP-taken de groei tussen 62B en 540B sneller toeneemt dan tussen 8B en 62B, ondanks de vrijwel gelijke schaalverhouding. Gezien het opschalen van taalmodellen een gunstige invloed heeft, groeit de interesse voor nog veel grotere taalmodellen.

Open-source LLM's en toegankelijkheid

Een Large Language Model trainen vereist een extreem grote hoeveelheid data, computerkracht en energie (stroom), waar een aanzienlijk budget voor nodig is. Het gaat om superclusters met duizenden TPU’s en honderden terabytes, met een kostprijs van in de miljoenen. Dit maakt het zelf trainen van LLM’s niet rendabel voor ieder bedrijf. We zien daarom ook initiatieven die

De Franse Jean Zay supercomputer werd gebruikt voor het trainen van het BLOOM-taalmodel - Foto: Wikimedia Commons

Large Language Models hebben getraind en geheel open ter download beschikbaar stellen. Recente voorbeelden zijn hierin OPT-175B (Meta AI, 5/2022) en BLOOM (BigScience, 176B, 7/2022).

Het publiekelijk ter download openstellen van Large Language Models is vanuit wetenschappelijk oogpunt belangrijk. Het stelt meer researchers in staat de interne eigenschappen van LLM's te onderzoeken, meer te leren over de interne werking en potentiële verbeteringen aan te dragen. Toch zal het directe gebruik op deze manier kleinschalig blijven. Hoewel voor *inference* -het uitvoeren van het model- minder rekenkracht nodig is dan voor training, blijft een aanzienlijk budget vereist. Het gebruik van modellen als OPT-175B of BLOOM vereist nog altijd een paar ton aan hardware of cloudkosten per jaar. Het prijskaartje en zijn onhandige grootte maken het gebruik van open-source LLM's voor de meeste toepassingen onpraktisch.

De echte opmars en brede praktische inzet van LLM's vinden plaats via toegankelijke Cloud-AI providers. OpenAI biedt GPT-3 tegen lage kosten aan via een simpele API en bereikt daarmee ieder willekeurig bedrijf en ook iedere student. HuggingFace is van plan een hosted versie van het BLOOM-taalmodel aan te bieden. Naar verwachting zullen steeds meer providers LLM's centraal gaan aanbieden.

Risico's

Het publiekelijk openstellen van Large Language Models brengt nieuwe zorgen met zich mee in het kader van responsible AI. Sinds de komst van steeds betere generatieve modellen zoals GPT-2 (2019) zien we dat het voor mensen steeds lastiger wordt om betrouwbaar synthetische tekst te herkennen en te onderscheiden van originele door mensen geschreven teksten. Large Language Models kunnen het met die generatieve modellen steeds makkelijker maken om natuurlijke tekst te genereren, dus ook misinformatie en spam. Hierop is steeds minder controle.

LLM's, open en gesloten, blijven ethische risico's houden van het genereren van beledigende teksten, racisme en bias. De modellen zijn getraind voor tekstvoorspelling en missen veel kennis van de fysieke wereld en normen en waarden waarin we leven. De output van een LLM kan statistisch correct zijn, maar ook sociaal onwenselijk. Een LLM zet je daarom doorgaans in voor co-creatie, waarbij menselijke sturing nodig blijft.

Snellere hardware en efficiëntere modellen

De voor AI-werktaken veelgebruikte GPU's zijn eigenlijk ontwikkeld voor 3D en grafische toepassingen. We zien deze vervangen worden door gespecialiseerde processoren, Tensor- en Neural Processing Unit's (TPU en NPU's). Een ander voorbeeld is het Amerikaanse bedrijf Cerebras dat de grootste chip ter wereld of 'datacenter on a chip' ontwikkelt. Cerebras vestigde in juni 2022 een nieuw record door een 20B parameter model te trainen op een enkel systeem.

Voortgang in hardware wil waarschijnlijk niet zeggen dat iedereen alsnog een LLM zelf gaat trainen. Het zal vooral betekenen dat we nog grotere taalmodellen kunnen gaan verwachten tegen lagere kostprijs de komende jaren. Wat we in 2022 *large* noemen, zal over een paar jaar anders heten. LLM's komen nu in het tijdperk van biljoenen (1T+) parameters.

Aan de kant van de software zien we ook veelbelovende trends die in algoritmische efficiëntie grote stappen maken door modelgrootte en kosten te verkleinen. De ecologische voetafdruk van OPT-175B was al 7x kleiner dan van GPT-3 en recentelijk introduceerde Amazon het Alexa-TM 20B parameter model, ruim 8,5x kleiner dan GPT-3 terwijl het GPT-3 op enkele NLP-taken al overtreft.

Menselijke efficiëntie nog lang niet bereikt

Het menselijk brein verbruikt gemiddeld zo'n 12 watt, niet meer dan zeg maar drie Raspberry Pi 4 computers. Dat grote taalmodellen met soms wel honderden tot duizenden zware TPU's de efficiëntie en zuinigheid van het menselijk brein nog lang niet benaderen, is duidelijk, laat staan dat ze intellectueel in de buurt komen. Tegelijkertijd zou het een enorme teleurstelling zijn voor het mystieke menselijke intellect, als dit te simuleren zou zijn op een enkele Intel-processor. •

TRANSFORMERMODELLEN IN AUTOMATISCHE SPRAAKHERKENNING

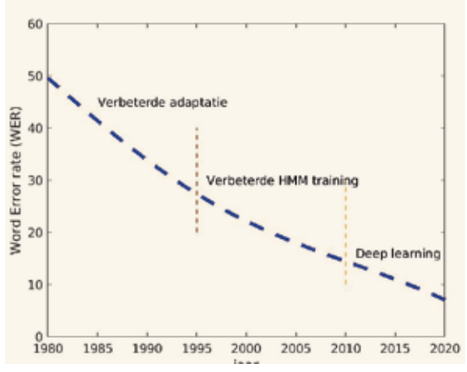
Jaarlijks komen spraakonderzoekers uit de hele wereld bijeen op een internationale conferentie met de naam 'Interspeech'. Het is een conferentie waar 1500 tot 2000 onderzoekers op het gebied van taal- en spraaktechnologie en onderzoek aan deelnemen. Bedrijven met een eigen onderzoeksgroep op het gebied van spraak- en taaltechnologie zijn daar ook ruimschoots vertegenwoordigd. Automatische Spraakherkenning stond al jaren op de agenda. De introductie van Artificiële Intelligentie zorgde voor een radicaal nieuwe aanpak.

Door:
Louis ten Bosch,
CLS/CLST,
Radboud
Universiteit

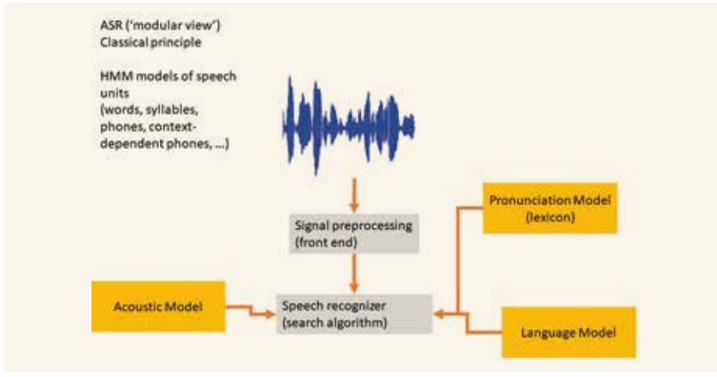
Aanvankelijk waren de deelnemers vooral afkomstig van universiteiten, maar sinds enkele jaren is een meerderheid afkomstig van een bedrijf, waarbij Google, Amazon, Meta, Apple en Microsoft de boventoon voeren. Interspeech vindt afwisselend plaats in verschillende continenten: dit jaar vond Interspeech 2022 plaats in Zuid Korea, volgend jaar in Dublin, daarna in Jeruzalem, en in 2025 is Rotterdam aan de beurt.

ASH

Traditiegetrouw is een van de topics op deze conferentie Automatische Spraakherkenning (ASH), het vakgebied dat zich bezighoudt met het automatisch omzetten van spraak naar tekst. Bij dit vakgebied gaan data, rekenkracht en kennis over het spraaksignaal hand in hand. De ontwikkelingen binnen ASH over de afgelopen 30 jaar zijn goed te volgen aan de hand van de artikelen die in de loop van de tijd in Interspeech proceedings verschenen. Modes en technieken kwamen en gingen. Gaandeweg namen de kennis over spraaktechnologie, de rekenkracht en beschikbaarheid van spraakcorpora enorm toe, met steeds betere resultaten als gevolg, niet alleen in termen van percentage in woordherkenning maar ook bijvoorbeeld in termen van robuustheid bij accenten van sprekers, en tegen verschillende soorten achtergrondlawaai. Figuur 1 geeft een impressie van de progressie van de herkenresultaten (in Word Error Rate, WER) met daarbij de belangrijkste oorzaak. De WER is een maat voor de kwaliteit van een spraakherkenner. Deze maat telt het aantal fouten dat de herkenner maakt vergeleken bij de door mensen gemaakte referentietranscriptie.



Figuur 1. Progressie van de resultaten van ASH (in Word Error Rate).



Figuur 2. Overzicht van een klassiek modulair ASH systeem.

Modulaire aanpak

Tot ruwweg 2014 waren ASH systemen meestal gebouwd volgens een klassiek idee, gebaseerd op het samenwerken van drie componenten: een akoestisch model, een woordenboek en een taalmodel (fig. 2). Het akoestisch model en het taalmodel moeten apart worden getraind: het akoestisch model op heel veel spraakdata, het taalmodel op heel veel tekst. Die training vroeg specifieke kennis over de klankstructuur van een taal en over de statistiek van rijen woorden. Na 2014 zijn we getuige van een paradigmaverschuiving die geleid heeft tot een radicaal nieuwe aanpak binnen ASH.

Deep Learning

De recente ontwikkelingen zijn het gevolg van de introductie van Artificiële Intelligentie in ASH. De ontwikkeling van Deep Learning heeft binnen Artificiële Intelligentie geleid tot geruchtmaken-

Figuur 3. Alpha Go verslaat Fan Hui, een professionele Go-speler, zonder handicap op een 19×19 Go bord. Still uit https://www.deepmind.com/research/highlighted-research/alphago

de doorbraken zoals van de schaakcomputer die (met gemak) de wereldkampioen schaken kan verslaan, en van AlphaGo, een computerprogramma ontwikkeld door DeepMind Technologies, dat de Go kampioen versloeg in 2016 (fig. 3).

Aanvankelijk werd de kwaliteit van deze Deep Learning algoritmen vooral bepaald door de 'brute force' aanpak. Om bijvoorbeeld een schaakalgoritme te maken werd een verzameling van honderdduizenden schaakpartijen aan het leeralgoritme aangeboden als trainingsmateriaal, en aan de hand van al deze voorbeelden leerde het algoritme de beste zet in een bepaalde situatie. Deze brute force voorbeeldgestuurde aanpak is later gedeeltelijk vervangen door een geraffineerdere aanpak waarbij een algoritme leert door het te laten spelen tegen een kloon van zichzelf en op die manier zettenreeksen te optimaliseren. Latere versies van AlphaGo, zoals AlphaGo Zero, werden op deze manier steeds krachtiger, zonder te leren van menselijke voorbeeld-games, en het recentere programma MuZero leert zonder de regels te leren. In al deze toepassingen wordt een diep neuraal netwerk gebruikt om de beste zetten te vinden vanuit een bepaalde startpositie, met een look-ahead van soms zo'n 50 zetten diep.

End-to-end benadering

De Deep Learning aanpak wordt ook in ASH toegepast. Conceptueel gezien doet een diep neuraal netwerk niets anders dan een rijtje inputvectoren omzetten naar een rijtje outputvectoren ('end-to-end'). Die omzetting van het ene 'end' naar het andere 'end' gaat met een groot aantal tussenstappen, via zogenaamde 'hidden layers' (fig. 4). Het huidige succes van Deep Learning in veel applicaties zoals beeldherkenning, gezichtsherkenning en automatische beschrijving van beelden is gebaseerd op een vertaling van de input en de output naar een input vectorrij en een output vectorrij die daarna in elkaar worden omgezet door het netwerk. Dat gebeurt ook in de tegenwoordige end-to-end benaderingen van ASH. Het variabele, niet-symbolische spraaksignaal wordt eerst omgezet naar inputvectoren (spectrale beschrijvingen van stukjes spraak), en de gewenste outputtekst naar een rij outputvectoren (die letters of woorden coderen). Deep Learning zorgt dan voor de omzetting van de inputrij naar de outputrij. Deze route is zoals het ongeveer gaat. In de praktijk worden geavanceerde trucs toegepast waarvan de beschrijving in het bestek van dit artikel te ver gaat.

Variatie

Conceptueel kan de route via de input en output vectoren gemakkelijk lijken. Maar bij spraak naar tekst is er een grote uitdaging. Spraak wordt namelijk gekarakteriseerd door een grote variabiliteit in het spraaksignaal (denk maar aan vrouw/man/kind, spraaktempo, gezondheid van de spreker, toonhoogte, verschillen tussen sprekers, accentverschillen, verschillen tussen microfoons, verschillen in omgevingsgeluiden et cetera). In een groot spraakcorpus kunnen veel totaal verschillende akoestische vormen allemaal hetzelfde woord representeren. De afbeelding tussen de inputvectoren naar de outputvectoren moet dus bestand zijn tegen een breed scala aan variabiliteit aan de inputkant – zo'n afbeelding wordt soms aangeduid als many-to-one. Deep Networks blijken deze variabiliteit goed aan te kunnen.

Gebruik van kennis

De vergelijking tussen de klassieke, meer modulegeoriënteerde aanpak zoals in figuur 2 en de end-to-end aanpak in figuur 4 is interessant. De aanpak is zo verschillend dat het lastig is om uit te maken wat nu in het algemeen de beste methode is. Deze vraag is stof voor veel recente discussies in het vakgebied. We proberen hier de voor- en nadelen te schetsen van beide aanpakken.

In de modulaire aanpak worden gespecialiseerde componenten (akoestisch model AM, taalmodel LM) getraind op basis van veel data. Die trainingen zijn specifiek en worden in hoge mate bepaald door kennis van de akoestische variabiliteit van een taal en van de woordvolgorden in een taal. Pas bij de integratie van alle componenten op hoog niveau komen de informatiestromen bij elkaar en wordt het spraakherkenningsprobleem op een statistische manier aangepakt.

In deze modulaire aanpak is het van belang dat er voldoende grote corpora beschikbaar zijn om het AM (en LM) te trainen. De spraakdata moeten geannoteerd zijn (op woordniveau), en er moet een goed woordenboek zijn met betrouwbare fonetische transcripties, zodat het AM kan leren met welke spraakklank welk symbool correspondeert. Voor een aantal talen (waaronder Engels, Nederlands) is een redelijk tot goed AM aanwezig maar voor de meeste talen is er te weinig geannoteerd materiaal voor een goed AM. Dat leren van het AM kun je zien als een schattingsprobleem: het AM heeft veel parameters die allemaal moeten worden geschat ('getuned') op trainingsdata. Het aantal parameters kan makkelijk in de honderden miljoenen lopen.

Een karakteristiek van de modulaire aanpak is dat de architectuur van de componenten gebaseerd is op kennis: kennis van de fonetische, morfologische en syntactische structuur van een taal. Als eenmaal de architectuur van het model gedefinieerd is dan worden daarna de modelparameters geschat op basis van de beschikbare trainingsdata.

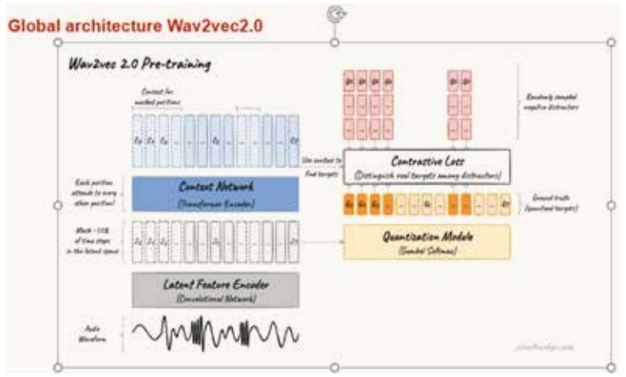
Gebruik van data

De situatie is heel anders bij de recentere end-to-end aanpak. Allereerst is het onderscheid tussen AM, lexicon en LM bij end-to-end methoden minder duidelijk, omdat in een end-to-end systeem verschillende functionaliteiten kunnen worden gecombineerd binnen een overkoepelend diep netwerk. Een diep netwerk vraagt wel enorm veel trainingsdata. Recentere toepassingen van een end-to-end systeem voor ASH (bijvoorbeeld 'wav2vec2.0') gebruikt in de orde van een miljoen uur spraak afkomstig van meer dan 100 talen. Je kunt zeggen dat bij een end-to-end aanpak het gebruik van kennis in een zekere mate wordt ingewisseld voor het gebruik van heel veel data.

Een ander verschil is het gebruik van de data. In wav2vec2.0 bestaat de training uit twee verschillende fasen. In de eerste fase wordt heel veel spraakdata gebruikt, zonder enige vorm van annotatie, labeling of segmentatie. Het kunnen gebruiken van ongeannoteerde spraakdata is de oorzaak dat honderduizenden uren kunnen worden benut voor deze stap. Deze fase 1 training levert een zogenaamd 'pretrained' model op. Daarna vindt in de tweede fase op dit 'pretrained' model een fijnafstemming ('finetuning') plaats, waarbij heel weinig gelabeld spraakmateriaal (in de orde van 10-60 minuten gelabelde spraak) al voldoende is. Fase 1 is taakonafhankelijk, fase 2 is taakafhankelijk. Deze trainingsopzet vormt een groot verschil met de training in het geval van modulaire ASH.

Diep netwerk

Een nadeel van end-to-end benaderingen is dat het niet meteen een transparante aanpak is. Het is meestal lastig om in te zien waarom een diep



Figuur 4. Een end-to-end netwerk (wav2vec2.0, uit: https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html)

netwerk bepaalde beslissingen neemt of een bepaalde output oplevert, omdat er zoveel tussenlagen zijn die niet-lineaire functies representeren van de ene laag naar de volgende, terwijl het juist interessant kan zijn te zien hoe een netwerk kan bijdragen aan de 'klassieke kennis'. Er wordt daarom tegenwoordig veel onderzoek gedaan naar 'Intelligible AI' en 'transparent AI', waarin de hoofdvraag is hoe wij de werking van een diep netwerk kunnen begrijpen.

Een laatste verschil met de recente Deep Learning benaderingen is dat een netwerk een veel breder tijdsinterval kan gebruiken om te komen tot een output. In Wav2vec2.0 wordt voor de constructie van een outputvector informatie gebruikt uit een tijdsinterval dat een paar seconden eerder begint en paar seconden later pas eindigt. Dus de toekomst doet mee in het construeren van een outputvector op een zeker moment! De wegingen van de context gebeuren in een transformer via een zogenaamd 'attention model' waarbij de betekenis van elk deel van de invoer verschillend kan worden gewogen. Transformers werden in 2017 geïntroduceerd door een team van Google Brain. Tegenwoordig is het het favoriete model en het vervangt gaandeweg andere modellen zoals recurrent neural networks (RNN) en het long-short time memory model (LSTM).

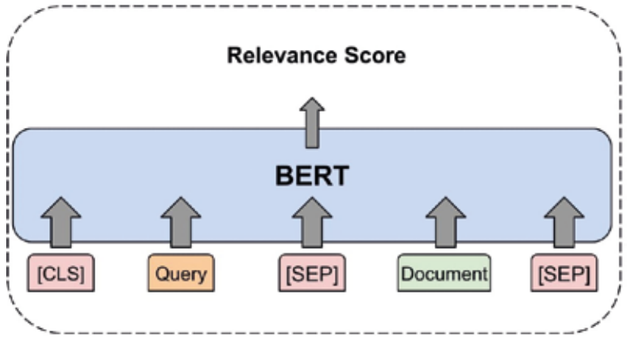
Het valt te verwachten dat de rol van Kunstmatige Intelligentie alleen maar zal toenemen. Als we met toekomstige technieken beter kunnen begrijpen hoe een netwerk tot een beslissing komt dan kan Deep Learning worden ingezet als een machtig hulpmiddel om onze 'klassieke' kennis te verrijken met inzichten die worden verkregen via patronen geleerd op basis van taaldata. Het zal zeker mogelijk zijn om de data-gestuurde end-to-end aanpak op de een of andere manier te combineren met de kennis-gebaseerde modulaire aanpak, om zo te profiteren van *the best of both worlds.* •

HOE ZOEKMACHINES BERT GEBRUIKEN OM RELEVANTE DOCUMENTEN TE VINDEN

Als we het hebben over taaltechnologie, dan denken we vaak aan tekstclassificatie, informatie-extractie of vertalen. Maar er is nog een ander type toepassing waar we vaak mee te maken hebben: *ranking*. Het doel van *ranking* is het genereren van een lijst van teksten gesorteerd op relevantie voor een vraag van een gebruiker. We maken hier allemaal bijna elke dag gebruik van: als we zoeken naar een product in Bol.com, een boek of film proberen te vinden, of informatie zoeken over een specifiek onderwerp in zoekmachines zoals Google.

Door:
Arian Askari,
Universiteit Leiden

Veel recente modellen voor het bepalen van de relevantie van documenten voor een zoekvraag zijn gebaseerd op BERT. De invoer en uitvoer van toepassingen voor tekstrelevantie op basis van BERT zijn weergegeven in de figuur hieronder. Om een relevantiescore voor documenten te genereren op basis van een zoekopdracht, voeden AI-onderzoekers BERT met paren van telkens één zoekopdracht (query) en één document. De berekende relevantiescore wordt vervolgens gebruikt om de teksten te ordenen.



Rekencomplexiteit

Hoewel op BERT gebaseerde modellen goed werken voor routinetaken met korte zoekvragen en korte documenten, komen ze in de problemen bij langere vragen en documenten. De reden is dat het aantal woorden rechtstreeks van invloed is op de rekencomplexiteit van BERT-modellen, waarvoor veel hardwarekracht nodig is. Als gevolg hiervan kunnen teksten die langer zijn dan 512 woorden op de meeste hardware nog niet worden geanalyseerd door op BERT gebaseerde modellen.

Zulke langere vragen en documenten zijn bijvoorbeeld gebruikelijk in een juridische zoekmachine die advocaten helpt om goed voorbereid te zijn op een zitting. Het doel van een juridische zoekmachine is om – met taalmodellen – miljoenen juridische teksten te analyseren en daar de meest relevante documenten uit te halen voor de informatievraag van de advocaat. In dit voorbeeld is de informatievraag een juridisch probleem dat kan bestaan uit honderden of duizenden woorden die het probleem beschrijven. Een ander voorbeeld van een toepassing die lange teksten moet kunnen verwerken, is het vinden van relevante patenten voor nieuwe patentaanvragen. Dergelijke eerder gepatenteerde uitvindingen kunnen een nieuwe patentaanvraag ongeldig maken. Het analyseren en indexeren van patenten voor een zoekmachine is uitdagend: de documenten zijn lang, de taal is formeel, juridisch en technisch met lange zinnen. Maar ook voor het gebruik van BERT-modellen in zoekmachines op het web (zoals Google) zijn langere documenten een uitdaging: veel webpagina's zijn immers langer dan 512 woorden; denk aan Wikipedia-artikelen en krantenteksten.

Twee fasen

Daarom bestaat de gebruikelijke aanpak voor ranking van documenten in de huidige systemen uit twee fasen. In de eerste fase haalt een op woorden gebaseerd en zeer efficiënt model honderden kandidaatdocumenten op uit een verzameling van miljoenen (of zelfs miljarden). Het op woorden gebaseerde model is niet beperkt tot de invoerlengte, maar heeft minder nauwkeurigheid voor korte teksten in vergelijking met BERT. In de tweede fase rangschikt een BERT-model de kandidaatdocumenten die worden opgehaald door het woordgebaseerde model voor betere precisie in de top 10 van de lijst.

Voor deze tweede fase worden momenteel diverse strategieën onderzocht om langere documenten te verwerken. Een mogelijke strategie is om de originele tekst in kleinere delen te verdelen en de uiteindelijke relevantiescore te berekenen op basis van de relevantie van elk afzonderlijk deel van de tekst. Een andere benadering is om de architectuur van op BERT gebaseerde modellen te vereenvoudigen, zodat er meer woorden aan kunnen worden toegevoegd. In sommige omstandigheden kan dit ze echter minder effectief maken. Longformer is een voorbeeld van een model dat op die manier bijna 8 keer de lengte van de BERT-invoer kan krijgen.

Vanwege het belang van allerlei zoekmachines in ons dagelijks leven en de moeilijkheden bij het omgaan met grotere teksten, zijn we nog steeds bezig met onderzoek om deze problemen aan te pakken. Toekomstige hardware ontwikkelingen of de ontwikkeling van nauwkeurigere modellen met eenvoudigere ontwerpen kunnen nieuwe mogelijkheden bieden. •

Mogelijk interessante literatuur

Lin, Jimmy, Rodrigo Nogueira, and Andrew Yates. "Pretrained transformers for text ranking: Bert and beyond." Synthesis Lectures on Human Language Technologies 14.4 (2021): 1-325.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2021. COLIEE 2021: Methods for Legal Document Retrieval and Entailment.

Florina Piroi and Allan Hanbury. 2019. Multilingual Patent Text Retrieval Evaluation: CLEF–IP. In Information Retrieval Evaluation in a Changing World. Springer, 365–387.

PRODUCTEN INDELEN VOOR DUURZAME BESLUITVORMING

IS BERT WEL ECHT DE BESTE METHODE?

Het gebruik van technologie op het gebied van taal heeft naast wetenschappelijke ook grote maatschappelijke relevantie. Voor mijn afstudeeropdracht heb ik in samenwerking met een adviesbureau in Amsterdam onderzoek gedaan naar de mogelijkheden van Natural Language Processing voor het faciliteren van duurzame besluitvorming.

Door:
Melle van der Meulen, Universiteit Leiden

Het adviesbureau waarmee ik heb samengewerkt adviseert overheden, bedrijven en NGO's op het gebied van duurzaamheid en ondersteunt milieurelevante transitieprocessen. Ze passen geavanceerde methodes toe waarmee de milieueffecten van producten en diensten in kaart kunnen worden gebracht. Een van deze methodes is het monitoren van grondstofgebruik en emissies gedurende een levenscyclus van een product of dienst. Op basis van die gegevens kan een gedetailleerd beeld worden verkregen van de milieu-impact van een bedrijf of organisatie.

Productontologie

Een uitdaging is de verwerking van grote hoeveelheden ongestructureerde data. Bedrijven en organisaties leveren gegevens aan in de vorm van productlijsten met korte en ongeordende productomschrijvingen. Deze productomschrijvingen worden handmatig gelabeld door het adviesbureau met behulp van een productontologie. Deze productontologie kan bestaan uit voorgedefineerde labels die hiërarchisch geordend zijn. Een voorbeeld van een productcategorie met hiërarchische ordening kan zijn: zuivel > melk > halfvolle melk. Aan deze productcategorieën kan een impact gekoppeld worden met de gemiddelde effecten op het milieu. Hierbij geldt dat hoe gedetailleerder de productomschrijving is, des te nauwkeuriger de milieueffecten bepaald kunnen worden. Enkele voorbeelden van productomschrijvingen en gerelateerde productcategorieën zijn weergegeven in tabel 1.

Tabel 1 geeft een voorbeeld weer van de data. De volledige data bestaat uit productomschrijvingen met gemiddeld 5 woorden per omschrijving. De productcategorieën zijn hiërarchisch geordend op twee verschillende niveaus: categorie niveau 1 bestaande uit 22 labels en categorie niveau 2 bestaande uit 952 verschillende labels. De uitdaging was het ontwikkelen van een taalmodel, dat door middel van zeer korte productomschrijvingen kleine en grote sets van productcategorieën kan classificeren. Hiervoor heb ik twee soorten technieken verder ontwikkeld voor dit specifieke probleem: Support Vector Machines (SVM) en BERT. Ook heb ik gekeken welke tekst features het meest informatief zijn om productomschrijvingen effectief te kunnen classificeren.

SVM en BERT

Het SVM-model is een relatief eenvoudig model. Het model is getraind op basis van de letterlijke woorden als features. De hypothese was dat de productomschrijvingen voldoende productspecifieke woorden bevatten om correct de productcategorieën te kunnen classificeren. Als tegenhanger van dit relatief eenvoudige SVM-model is ook het BERT-model toegepast. BERT is een geavanceerd en complex model dat door middel van Transfer Learning voorgetraind kan worden op grote hoeveelheden data. Hierdoor is het mogelijk om met relatief weinig nieuwe data-invoer een kwalitatief goed model te ontwikkelen. Daarnaast maakt het BERT-model gebruik van contextinformatie in plaats van alléén woorden als features. Hierdoor is het mogelijk om de syntactische en semantische kenmerken van woorden te modelleren en meer informatie te onttrekken uit de korte productomschrijvingen. Deze eigenschappen zouden de kwaliteit van het classificatiemodel moeten bevorderen in vergelijking met simpelere modellen waarbij alleen woorden als features worden gebruikt.

Productomschrijving	Categorie niveau 1	Categorie niveau 2
Poloshirt ME navy/fluor geel L/M dames	Kleding	T-shirt katoen
Siemens SN614X02AE Glazenspoeler	Huishoudelijke apparaten	Vaatwasser
MFC-L8690CDW Brother laserprinter	ICT	Desktop printer

Tabel 1. Enkele voorbeelden uit de dataset. De productomschrijving illustreert de korte en ongestructureerde omschrijvingen. Categorie niveau 1 geeft de (globale) productcategorie weer bovenaan in de hiërarchie. Categorie niveau 2 geeft een meer gedetailleerde productcategorie weer.

'Lente' door Hetty Franken

Uit de experimenten bleek dat SVM en BERT vergelijkbare resultaten laten zien – rond de 87% nauwkeurigheid – bij het classificeren van productcategorie niveau 1, een relatief kleine set aan labels. Deze set aan labels, bestaande uit globale productcategorieën, zijn met grote nauwkeurigheid te classificeren met gebruik van de korte productomschrijvingen. Maar we zagen opmerkelijke verschillen tussen de modellen bij het classificeren van productomschrijvingen met een grote set aan labels in de productcategorie in niveau 2. Het SVM-model kan goed overweg met een groot aantal labels en is in staat om met alleen letterlijke woorden als features een grote set aan productcategorieën correct te classificeren met 79% nauwkeurigheid. Het BERT-model blijkt ongeschikt om met korte productomschrijvingen een grote set aan productcategorieën te classificeren en blijft steken op 60%.

Conclusie

Blijkbaar is het heel lastig om korte productbeschrijvingen te classificeren op basis van rijke semantische woordrepresentaties. Juist een relatief simpele weergave – de letterlijke woorden als features – blijkt zeer nuttig te zijn om korte productomschrijvingen te classificeren. Het laat zien dat de meeste productomschrijvingen op een gedetailleerd productcategorieniveau vaak te herleiden zijn door enkele productspecifieke woorden en niet op basis van context, syntactische en semantische kenmerken van woorden.

Een kenmerkend voorbeeld is de classificatie van specifieke autocategorieën. Een productomschrijving "bedrijfsauto pick up 4x4 fs" werd door het SVM model correct geclassificeerd als: Voertuig. Ook was het model in staat om te classificeren op een meer gedetaileerde productcategorie niveau: Terreinwagen. Termen als '4x4' worden gebruikt als features en lijken informatief te zijn om gedetaileerde autocategorieën correct te kunnen classificeren. BERT was niet in staat om autocategorieën tot op een gedetaileerd niveau te kunnen classificeren en kon dit voorbeeld slechts classificeren met de algemene productcategorie: Voertuig. Blijkbaar is het letterlijke woord '4x4' nodig om te herkennen dat deze auto een terreinwagen is, terwijl de semantische representatie die BERT voor deze term heeft, het model alleen maar in de war brengt. •

HOE BERT DE MEDISCHE WERELD HELPT

DOOR IN PATIËNTENERVARINGEN DE RELEVANTE MEDISCHE TERMEN TE HERKENNEN

In het biomedische domein zijn transformermodellen zoals BERT niet meer weg te denken. Nieuwe BERT-modellen volgen elkaar in rap tempo op; een aantal populaire varianten zijn BioBERT, SciBERT, clinicalBERT en PubmedBERT. Deze modellen worden onder andere ingezet voor het identificeren van biomedische concepten.

Door:
Anne Dirkson, Universiteit Leiden / Nederlands Forensisch Instituut

Biomedische concepten zoals ziektes, medicatie en bijwerkingen van medicatie (bijvoorbeeld hoofdpijn of vermoeidheid) worden in patiëntendossiers of in ervaringen beschreven door patiënten. Zo kan iemand schrijven: "Sinds ik *gleevec* gebruik, heb ik steeds vaker last van misselijkheid". We willen dan graag dat ons BERT-model herkent dat gleevec een medicatie is en *misselijkheid* een bijwerking is die de patiënt toeschrijft aan die medicatie. Dit kan best lastig zijn omdat misselijkheid ook op andere manieren omschreven kan worden, bijvoorbeeld "'s avonds heb ik vaak een weeïg gevoel in mijn buik".

Voor biomedische concepten zijn er enorme, gedetailleerde ontologieën beschikbaar die het mogelijk maken om deze concepten te normaliseren; deze taak noemen we *entity linking*. Dezelfde ontologieën worden gebruikt bij het handmatig toevoegen van codes aan medische dossiers. Een grote uitdaging hierbij is de omvang van de ontologieën. De meest gebruikte ontologie voor aandoeningen, de ICD ontologie van het WHO, bevat meer dan 14 duizend codes. Ontologieën die frequent worden gebruikt voor bijwerkingen zijn SNOMED en MedDRA. Deze bevatten respectievelijk meer dan 100 duizend en meer dan 75 duizend codes.

Zero-shot methodes

Door de enorme omvang is het onrealistisch om trainingsdata te verzamelen voor alle mogelijke concepten. Er wordt dan ook ingezet op zogenaamde *zero-shot* methodes; dat zijn methodes die labels kunnen voorspellen waarvoor ze geen trainingsdata hebben gezien. Daarbij wordt gebruikgemaakt van de gelijkenissen tussen het ene concept (bijvoorbeeld *wakker liggen*) en de labels van alle mogelijke concepten. In dit geval is het juiste label *Insomnia* (SNOMED code 193462001). De gelijkenis wordt berekend op basis van embeddings (rijke vectorrepresentaties) van zowel het concept in de tekst als de labels in de ontologie. Deze embeddings kunnen gemaakt worden met een BERT-model. Naast de gelijkenis met het label zelf, wordt er ook gebruikgemaakt van de gelijkenis met synoniemen van het concept; in dit geval *slapeloosheid*.

Huidige zero-shot methodes voor biomedische entity linking zijn veelal gebaseerd op ranking (zie Sung *et al* 2020). Zonder training staat het concept bovenaan met de hoogste similarity puur gebaseerd op de gekozen embedding. Als er wel trainingsdata beschikbaar is voor een bepaald concept dan kan de ranking nog verder geoptimaliseerd worden. Alle beschikbare synoniemen van een concept worden hierbij meegenomen als mogelijke target labels. Deze methodes werken bijvoorbeeld goed voor het koppelen van bijwerkingen aan medische concepten die online door patiënten beschreven worden (zie Dirkson et al. 2021) zoals *wakker liggen* aan *insomnia*.

SapBERT

Recent is er een BERT-model ontwikkeld (SapBERT) dat voortborduurt op het idee om synoniemen vanuit de UMLS te gebruiken. De UMLS is een overkoepelende kennisbank die een groot aantal biomedische ontologieën samenvoegt, waaronder SNOMED. Liu et al (2021) ontwikkelden een methode om een biomedisch BERT-model direct te leren om synoniemen van UMLS-concepten als vergelijkbaar te zien. Dit BERT-model kan dan weer gebruikt worden om de embeddings te berekenen voor zero-shot methoden.

De kwaliteit en diversiteit van de synoniemen is voor entity linking methodes die hier gebruik van maken van groot belang. Een mogelijke valkuil van deze methodes is het ontbreken van lekentermen als synoniemen voor de medische concepten in de UMLS (bijvoorbeeld 'wakker liggen' voor insomnia). Hiervoor zijn al enkele initiatieven in het leven geroepen, zoals de Consumer Health Vocabulary door de University of Utah. Deze initiatieven zijn echter tot nu toe nog klein gebleven.

Wat gaat de toekomst brengen?

Vanuit een andere hoek van de biomedische NLP, namelijk het linken van ICD-codes aan aandoeningen in tekst, wordt gebruikgemaakt van graph-CNN-modellen. Deze modellen kunnen de hiërarchische structuur van de medische ontologie modelleren. Het zou zeker interessant zijn om te zien of in de toekomst deze twee stromen gebundeld kunnen worden, bijvoorbeeld door met BERT-modellen de beschrijvingen te modelleren en met een graph-CNN de structuur van de ontologie. Wellicht is het ook mogelijk om de structuur direct in het BERT-modellen mee te nemen door het model te trainen op hyper- en hyponiemen van de UMLS.

Met het verbeteren van zero-shot entity linking wordt het ook mogelijk om steeds complexere entiteiten te koppelen aan concepten uit de ontologie. Zelf hebben wij de afgelopen jaren gewerkt aan het structureren van adviezen die patiënten elkaar geven. Dit zijn lange en beschrijvende concepten zoals "een deel van de medicatie 's morgens en een ander deel 's avonds nemen" of "een bad nemen met gemberolie". BERT-modellen hebben door hun brede toepasbaarheid ook hier de potentie om waardevol te kunnen zijn voor het verbinden van deze beschrijving aan een gestructureerde ontologie. •

Mogelijk interessante literatuur

A. Dirkson, S. Verberne, W. Kraaij, G. van Oortmerssen & A.J. Gelderblom (2022). Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers. Scientific Reports, 12 (10317)

F. Liu, Z. Meng, M. Basaldella, and N. Collier (2021). Self-Alignment Pretraining for Biomedical Entity Representations. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, pages 4228–4238, 2021.

M. Sung, H. Jeon, J. Lee, and J. Kang (2020). Biomedical entity representations with synonym marginalization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3641–3650, 2020.

DEBATTEN OVER ABORTUS OM TE LEREN OVER KERNENERGIE

VAN HET ENE ONDERWERP NAAR HET ANDERE IN ARGUMENTATIE IS NIET ZO MAKKELIJK VOOR GROTE TAALMODELLEN.

Automatisch online debatten analyseren is geen makkelijke taak. Sinds kort worden hier grote taalmodellen voor gebruikt. Een belangrijke taak hierin is de herkenning van standpunten (*stance detection*): het sorteren van argumenten in of ze voor of tegen een bepaalde stelling zijn. Deze taak is extra moeilijk als je van training op het ene discussieonderwerp wil gaan naar voorspelling van *voor* of *tegen* over een heel ander onderwerp. Recent onderzoek laat zien dat grote taalmodellen hier nog niet echt in slagen.

Door:
Myrthe Reuver,
Vrije Universiteit Amsterdam

Argument Mining

Argument mining is een vakgebied waarin onderzoekers proberen om teksten met argumenten automatisch te analyseren. Het doel van zulk onderzoek is bijvoorbeeld modellen te ontwikkelen die automatisch onderdelen van een argument (de conclusie of de premisse) in een essay kunnen detecteren, of modellen die argumenten op Twitter in groepen kunnen sorteren. Een voorbeeld van zo'n taak is standpuntdetectie (*stance detection*), waarbij argumenten geclassificeerd moeten worden in of ze voor of tegen een bepaalde stelling zijn. Is "Een foetus is een mens en mensenlevens zijn belangrijk" een argument *voor* of *tegen* legale abortus? Modellen die zulke taken uitvoeren, kunnen gebruikt worden voor verschillende doeleinden: bijvoorbeeld om moderators van online discussies te assisteren, maar ook om teksten met verschillende standpunten aan te bevelen in een nieuwsapp.

Grote Taalmodellen en verschillende onderwerpen

Momenteel zijn grote taalmodellen de meest gebruikte methode voor argument mining taken. Datasets voor het *fine-tunen* van modellen in standpuntdetectie hebben vaak verschillende discussieonderwerpen. Zulke datasets worden gemaakt door Engelstalige discussies van het internet te halen, bijvoorbeeld op sociale media zoals Reddit of debatfora, en argumenten in die discussies te annoteren met *voor* of *tegen* labels. Dit wordt dan het trainingsmateriaal voor het model. De discussieonderwerpen zijn vaak politieke discussies met een Noord-Amerikaanse oriëntatie, zoals de wapenwetgeving van de Verenigde Staten, legale abortus of een discussie over nucleaire wapens. Het moeilijke van het trainen van deze modellen voor doelen zoals assisteren bij online moderatie, of verschillende invalshoeken in een debat aan mensen aanbevelen, is dat er online constant nieuwe onderwerpen bij komen. De publieke discussie gaat de ene dag over Zwarte Piet en de volgende dag over een nieuwe economische maatregel. Dit maakt het extra interessant om te kijken of modellen los van het onderwerp deze taak kunnen uitvoeren - kunnen ze, getraind op het ene onderwerp, ook *voor* en *tegen* onderscheiden binnen het andere onderwerp? Het idee is dat de grote taalmodellen die nu vaak gebruikt worden een soort algemene patronen van taal (her)kennen, wat zou kunnen helpen bij deze moeilijke cross-topic setting van deze taak.

Onderzoek

Een van de eerste onderzoeken die probeerde of grote taalmodellen dit kunnen, kwam uit in 2019. Reimers en collega's probeerden of een model getraind op zeven discussies (zoals schooluniformen, nucleaire energie en abortus) ook teksten over een achtste onderwerp kon classificeren in *voor* en *tegen*. Ze deden dit 8 keer, met elke keer een ander onderwerp. Hun eerste conclusie was optimistisch: het leek alsof BERT, het grote taalmodel, dit prima kon leren. De onderzoekers concludeerden dat deze modellen iets leerden over aspecten van argumentatie die niet afhankelijk zijn van het specifieke onderwerp.

Ons eigen onderzoek en tevens een ander onderzoek door Terne Jakobsen en collega's (beide uitgekomen in 2021) hadden een andere conclusie: deze grote taalmodellen voor standpuntdetectie leken vooral te focussen op enkele prominente onderwerpspecifieke woorden, zoals *gun* (wapen) voor wapenwetgeving en *woman* (vrouw) voor abortus, ook in een setting waar het model een nieuw onderwerp moest classificeren. Woorden die te maken hadden met argumentatie, zoals "tegen", "hoewel", en "daarom", werden niet vaak door het model gebruikt om een beslissing te nemen over of een tekst *voor* of *tegen* was. Moeilijke argumenten en gebrek aan achtergrondkennis over discussies maakte het ook erg lastig om te generaliseren van het ene onderwerp naar het andere. Een argument over het First Amendment in een debat over wapenwetgeving in de Verenigde Staten is bijvoorbeeld vaak een argument *tegen* meer wapenwetgeving, omdat dat First Amendment een wet is die gaat over hoe Amerikanen het recht hebben op wapens. Maar om zo'n argument goed te kunnen classificeren als *tegen* strengere wapenwetgeving, moet je die achtergrondkennis over de rol van die wet in het debat wel hebben. Deze modellen hadden dat niet want ze waren getraind op *voor* en *tegen* argumenten over abortus en marihuana.

Conclusie

Verschillende discussieonderwerpen in *voor* of *tegen* classificeren zonder training in het specifieke onderwerp blijkt dus (nog) niet helemaal mogelijk voor grote taalmodellen. Misschien is het zelfs een beetje een oneerlijke taak voor een BERT-model omdat er specifieke kennis voor nodig is. Voor de verschillende gebruiksdoelen van deze modellen, zoals het aanbevelen van verschillende argumenten of het automatisch modereren van online discussies, is dit wel jammer. Maar grote taalmodellen kunnen niet zomaar een nieuwe discussie begrijpen zonder erin getraind te zijn. Grote taalmodellen zijn toch (helaas?) niet de grote oplossing voor alles. •

Mogelijk interessante literatuur

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., & Gurevych, I. (2019). Classification and Clustering of Arguments with Contextualized Word Embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 567-578). URL: https://aclanthology.org/P19-1054.pdf

Reuver, M., Verberne, S., Morante, R., & Fokkens, A. (2021). Is Stance Detection Topic-Independent and Cross-topic Generalizable?-A Reproduction Study. In Proceedings of the 8th Workshop on Argument Mining (pp. 46-56). URL: https://aclanthology.org/2021.argmining-1.5.pdf

Terne Jakobsen, S. T., Barrett, M., & Søgaard, A. (2021). Spurious Correlations in Cross-Topic Argument Mining. In Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics (pp. 263-277). URL: https://aclanthology.org/2021.starsem-1.25.pdf

VOORGETRAINDE GROTE TAALMODELLEN EN AUTOMATISCHE VERTALING

Tot voor kort werden automatische vertaalsystemen vaak getraind per taalpaar en from scratch. Als bouwer van MT-systemen downloadde je een dataset van bijvoorbeeld OPUS[1] en dan kon je daar je neuraal MT (NMT) systeem op trainen. Hiervoor kon je gebruikmaken van toolkits zoals OpenNMT[2] of Marian[3]. Nu maakt mBART het mogelijk om voorgetrainde modellen te gebruiken voor automatische vertaling.

Door:
Vincent Vandeghinste, Instituut voor de Nederlandse Taal &
KU Leuven en Tim Van de Cruys, KU Leuven

Om te vermijden dat je per taalpaar een apart model diende te trainen, werden er rond 2016 trucjes bedacht, waarbij de bronzinnen voorzien werden van een tag die aangaf in welke richting er vertaald moest worden. Dan werden alle parallelle corpora voor de verschillende taalparen aan elkaar gehangen en werd daarop getraind. Zo ontstond ook de zero-shot translation, waarbij vertalingen mogelijk werden tussen talen waarvoor geen parallel trainingsmateriaal gebruikt werd.

mBART

Sinds de komst van de BERT-modellen in 2019 is de hele NLP-wereld doordrongen geraakt van voorgetrainde modellen, zoals beschikbaar op Hugging Face (https://huggingface.co/). Het model dat dergelijke technieken mogelijk maakt voor automatische vertaling is mBART[4], dat in 2020 gepubliceerd werd door Facebook AI Research (FAIR).

mBART is een multilinguaal voorgetraind model dat getraind is op 50 talen (initieel op 25 talen). Voor die 50 talen zijn monolinguale, i.e. niet-parallelle corpora gebruikt. Van de teksten in deze corpora wordt een deel van de woorden gemaskeerd door het woord te vervangen door [MASK] en worden zinnen door elkaar gehaald, en mBART moet dan leren om deze gemaskeerde tokens te voorspellen en de zinnen terug in de juiste volgorde te zetten. Dit wordt *denoising with auto-regressive decoding* genoemd.

Dit voorgetrainde model kan dan gefinetuned worden met een parallel corpus voor een specifiek taalpaar en dat levert de volgende resultaten op. De kwaliteit van vertalingen voor kleine en medium grote parallelle datasets gaat er, gemeten in BLEU score (zie kader), gigantisch op vooruit en maakt het verschil uit tussen een onbruikbaar en een bruikbaar systeem. Voor extreem kleine parallelle datasets (< 10000 zinsparen) lijkt het niet te helpen. Voor taalparen met grote datasets (meer dan 10 miljoen zinsparen) lijkt het ook niet echt veel verschil te maken.

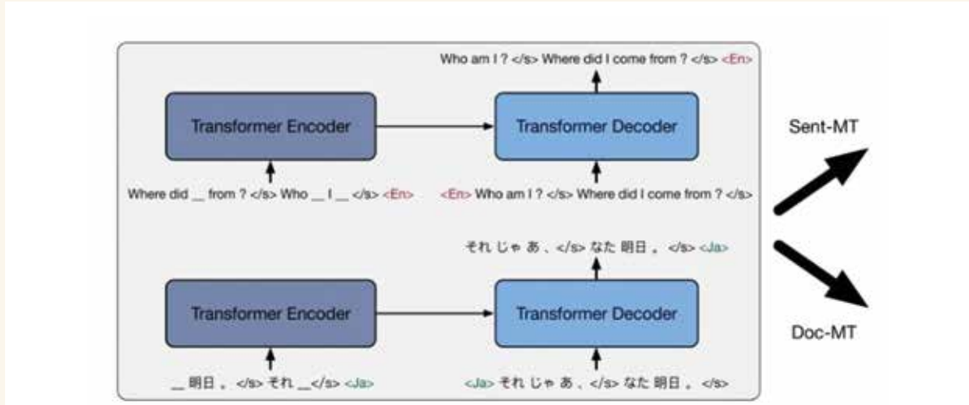
Er kan zelfs gefinetuned worden op taalparen waarvan een van de talen niet tot de voorgetrainde mBART-talen hoort, en hoe nauwer de voor mBART ongeziene taal verwant is aan een van de talen van mBART, hoe beter dit werkt. Dit werkt het beste wanneer de ongeziene taal de doeltaal is. Doordat een van de doelen van de mBART pretraining erin bestaat zinnen terug in de juiste volgorde te zetten, heeft mBART ook een positief effect op vertalingen op documentniveau. Zinnen die op zich ambigu zijn kunnen zo door hun context een correcte vertaling krijgen.

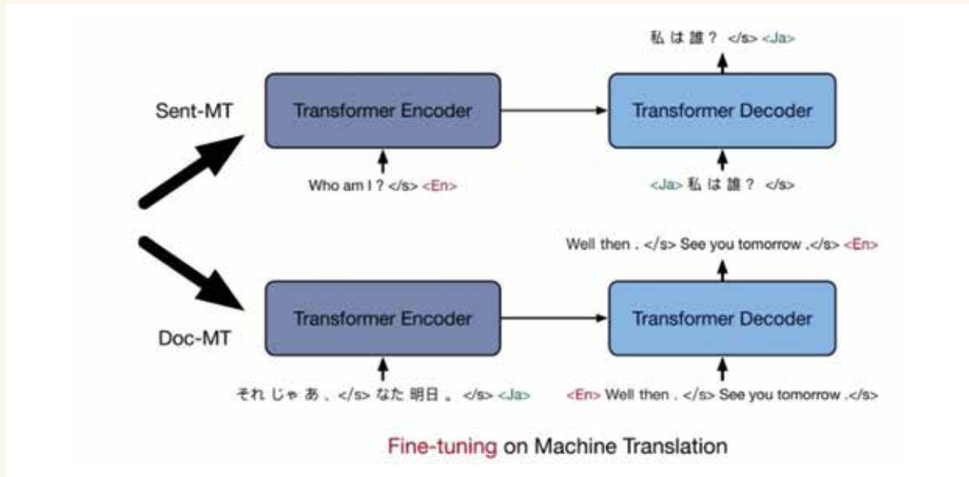
mBART en gebarentaal

Binnen het SignON project (https://signon-project.eu) onderzoeken we hoe we kunnen vertalen van gesproken talen naar gebarentalen en vice versa. Meer concreet kijken we hoe we bijvoorbeeld Nederlandse spraak of tekst (NL) kunnen vertalen naar Vlaamse GebarenTaal (VGT). Omdat er voor gebarentalen weinig data voorhanden is, laat staan parallelle data (i.e. met vertaling naar het NL), willen we hierbij gebruik maken van een vertaalarchitectuur met mBART. Gebarentaal is uiteraard van een heel andere orde dan gesproken taal. Om gebarentaal te kunnen verwerken, moet het mBART-model ook visuele data als invoer aanvaarden. Daarom breiden we de mBART-architectuur uit met een encoder die visuele data in visuele embeddings omzet. Die embeddings worden opgebouwd op basis van de ruwe videodata, of afgeleide features die de lichaamshouding inschatten. Met behulp van de mBART-decoder kunnen die visuele embeddings dan vertaald worden naar gesproken taal.

Om daarnaast gesproken taal in gebarentaal om te zetten, moeten we aan de uitvoerkant gebarentaal kunnen genereren. Daarvoor doen we beroep op een semantische representatie, met name *Abstract Meaning Representation* (AMR). Die representatie willen we verrijken met *SIGN_A* - een specifieke representatie voor gebarentaal. De SIGN_A-representatie zal op zijn beurt omgezet worden in *Behaviour Markup Language*, waarmee we uiteindelijk een avatar voor gebarentaal zullen aansturen. Een voordeel van deze configuratie is dat we een groot deel van de representatie automatisch kunnen leren: we leren het mBART-model om ook de AMR-representaties voor verschillende talen automatisch te voorspellen. Daarnaast proberen we om met behulp van MediaPipe-features ook de SIGN_A-representatie automatisch te voorspellen. Op die manier kunnen we gesproken taal dus automatisch omzetten naar een gebarentaalspecifieke semantische representatie, die op haar beurt de avatar kan aansturen voor gebarentaalsynthese. •

BLEU is een metriek die vaak gebruikt wordt voor automatische evaluatie van vertaalsoftware, waarbij de output van de software vergeleken wordt met een referentievertaling, die door een vertaler gemaakt is. De BLEU-score blijkt matig te correleren met menselijke oordelen over MT kwaliteit.





(Liu et al, 2020)

[1] https://opus.nlpl.eu/
[2] https://opennmt.net/
[3] https://marian-nmt.github.io/
[4] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. Transactions of the Association for Computational Linguistics, 8:726–742. https://aclanthology.org/2020.tacl-1.47/

WAT KAN TAALTECHNOLOGIE UIT HET HEDEN MET TAAL UIT HET VERLEDEN?

Bij 'historische taalwetenschap' denk je misschien eerder aan middeleeuwse manuscripten dan aan taaltechnologie. Toch is de historische taalwetenschap al lang onlosmakelijk verbonden met computergebaseerd onderzoek, en in het laatste decennium is ook NLP een belangrijke rol gaan spelen. Maar nog interessanter misschien, is dat het omgekeerde ook geldt: steeds meer onderzoekers in NLP willen zich buigen over de vraag hoe *state-of-the-art* taalmodellen omgaan met historische taal. Welke uitdagingen en inzichten zijn er daardoor blootgelegd?

Door:
Lauren Fonteyn, Leiden University

Alle levende talen zijn door de eeuwen heen veranderd. Voor het Nederlands kun je dat bijvoorbeeld zien aan de volgende zinsfragmenten:

zynen zoon, zoo oopentlyk onnoozel
(P.C. Hooft, 1642)
als Godts onnosel Lam
(J. Cats, 1657)
Ik ben zeer ongesteld door deezen onvoorzienen schok
(B. Wolff en A. Deken, 1785).

Omdat de eerste officiële Nederlandse spelling, die de basis vormt voor het huidige spellingssysteem, pas in 1804 werd ontworpen door Matthijs Siegenbeek[1], was er voorheen betrekkelijk veel variatie in hoe eenzelfde woord gespeld werd (*onnosel, onnoozel*). En ook de betekenissen van een groot aantal woorden zijn door de eeuwen heen veranderd. Zo was 'onnozel zijn' voor Hooft en Cats nog een positieve eigenschap, of konden in de 18de eeuw ook mannen die zich niet lekker voelden zeggen dat ze 'ongesteld' waren.

Problemen door spelling

Om betekenisveranderingen automatisch op te sporen in grote tekstverzamelingen, werden neurale modellen als Word2Vec[2] al snel populair. Die modellen vatten de betekenis van een woordtype door de verschillende contexten waarin het gebruikt wordt, op te slaan in vectoren. Omdat woorden met een gelijkaardige betekenis in gelijkaardige contexten voorkomen, zullen de vectoren van die woorden dichter bij elkaar liggen. Door zulke vectoren te maken voor verschillende tijdsperiodes, kan je bijvoorbeeld op een objectieve, datagedreven manier laten zien dat het woord abortus in Nederlandse kranten uit 1950-1954 lijkt op negatieve woorden als *ziekte* en *infectie*, maar dat het daarna geleidelijk meer op neutrale woorden als *zwangerschap* en *anticonceptie* is gaan lijken[3].

De kwaliteit van vectorrepresentaties hangt wel sterk af van de frequentie van de woorden die ze moeten vatten: als een woord als onnozel bijvoorbeeld 30 keer voorkomt in een verzameling teksten, dan is de vector die daaruit komt beter dan bij een ander model dat het woord maar 10 keer gezien heeft. Maar wat als onnozel telkens net wat anders gespeld wordt? Dan is er voor het model geen sprake van 30 voorkomens van hetzelfde woordtype, maar van 30 verschillende woordtypes met slechts één voorkomen. Dat betekent dat we 30 vectoren krijgen van lagere kwaliteit. Gedigitaliseerde verzamelingen van oudere teksten zijn lastiger 'te vatten' voor zulke modellen – niet alleen wegens beperktere overlevering dan verzamelingen van recente teksten, maar ook omdat ze meer spellingvariatie bevatten en bovendien vaak lijden onder problematische *Optical Character Recognition*.

BERT voor historische tekst

Modellen als BERT[4], die vectorrepresentaties genereren voor individuele voorkomens van woorden in plaats van voor woordtypes, hebben daarentegen minder problemen met spellingvariatie. Daarom lijken ze ook beter geschikt om historische taal aan te pakken. Een belangrijke voorwaarde is wel dat ze daarop 'voorbereid' moeten zijn. De varianten van BERT die het vaakst gebruikt worden zijn bijvoorbeeld voorgetraind op hedendaagse taal, met een beperkte woord(deel)lijst die uitgaat van hedendaagse woordenschat en regelmatige spelling. Dat dit blijkbaar toch problemen oplevert met historische tekst werd duidelijk voor ons toen we de eerste 'historische' BERT-modellen de wereld instuurden. Zo bleek dat MacBERTh[5], een BERT-model dat we voorgetraind hebben op Engelstalige teksten die teruggaan tot de 15de eeuw (3,9 miljard woorden in totaal), minder problemen had met allerlei NLP-toepassingen op historische tekst dan het originele, hedendaags Engelse BERT-model.

Dat historische BERT-modellen zoals het Engelse MacBERTh en het Nederlandse GysBERT nu beschikbaar worden gemaakt[6] is goed nieuws voor iedereen die graag teksten uit het verleden zou willen door- en onderzoeken. Zo hebben beide modellen zich al sterk getoond in uitdagende taken als het voorspellen welke van twee zinnen ouder is. Dat betekent dat ze op termijn zouden kunnen worden ingezet om teksten automatisch te dateren. Ook slagen deze modellen er beter in om de verschillende betekenissen van woorden te onderscheiden, zelfs als een van de betekenissen in hedendaagse taal niet meer bestaat (zoals de eerdergenoemde betekenissen van *onnozel* en *ongesteld*).

Op naar dynamische modellen

Het relatief recente huwelijk tussen historische taalwetenschap en neurale taalmodellen lijkt, kortom, al vruchten af te werpen. Waar dat 10 jaar geleden nog onvoorstelbaar leek, duiken er nu steeds meer workshops op bij grote NLP-conferenties die zich buigen over historische taalvarianten, en rijzen er specifieke vragen zoals hoe we automatisch betekenisveranderingen kunnen opsporen. En die toepassingen roepen dan weer de vraag op of het eigenlijk wel wenselijk is om taalmodellen enkel te trainen met hedendaagse standaardtaal. Want wat als we ook betrouwbare en efficiënte toepassingen willen bouwen voor informeler taalgebruik (bijvoorbeeld op sociale media)? En wat als onze taal over een paar jaren of decennia weer net wat veranderd is, zoals dat onvermijdelijk gaat met levende talen? Het is, kortom, erg belangrijk om na te denken over hoe we de taaltechnologie van vandaag flexibel en dynamisch kunnen maken zodat ze werkt voor taal van verleden én de toekomst. •

Afbeelding gemaakt door Lauren Fonteyn

[1] Van der Wal, Marijke & van Bree, Cor (2008). Geschiedenis van het Nederlands. Houten: Spectrum.

[2] Mikolov, Tomás, Kai Chen, Greg Corrado & Jeffrey Dean (2013). Efficient Estimation of Word Representations in Vector Space. In Yoshua Bengio & Yann LeCun (eds.), *1st International Conference on Learning Representations* (ICLR 2013), Scottsdale, Arizona, USA. http://arxiv.org/abs/1301.3781

[3] Wevers, Melvin, & Koolen, Marijn (2020). Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods,* 53(4), 226-243. https://doi.org/10.1080/01615440.2020.1760157

[4] Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT* 2019, 4171–4186.

[5] Manjavacas, Enrique & Fonteyn, Lauren (2021). MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950). In: Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH), p. 23-36. Association for Computational Linguistics. http://icon2021.nits.ac.in/resources/nlp4dh.pdf#page=35

[6] https://macberth.netlify.app/

EMOTIES HERKENNEN IN SPRAAK?

De laatste jaren is er duizelingwekkende vooruitgang geboekt in automatische spraakherkenningstechnologie (automatic speech recognition, ASR). Dit komt onder andere door de opkomst van Deep Learning modellen en Self-Supervised Learning. Soortgelijke technieken worden nu ook steeds meer voor emotieherkenning in spraak gebruikt. Maar methodologisch en conceptueel gezien zijn er best wel wat verschillen tussen het herkennen van emotie en het herkennen van wat er gezegd wordt. De vraag is dan ook of Big Models, waar dit DIXIT-thema over gaat, ook van toepassing is op emotieherkenning in spraak.

Door:
Khiet P. Truong,
HMI,
University of Twente

Het eerste verschil dat ik wil benoemen is de subjectiviteit van menselijk emotieperceptie. De mate van mogelijke onenigheid in het herkennen van een emotie door mensen is groter dan de onenigheid over wat er gezegd wordt. De annotators die emoties labelen voor trainingsdata zijn het dan ook soms niet met elkaar eens. Er bestaan wel methoden om met dit "probleem" om te gaan maar die passen het "probleem" aan op de methoden door te middelen of door een majority voting te doen (het label van de meerderheid wint). Terwijl onze taak misschien juist is om met methoden te komen die zich aanpassen aan deze mate van subjectiviteit in plaats van andersom. Het "probleem" is nu eenmaal dat mensen onderling verschillen, uit andere culturen komen, andere ervaringen hebben en dus andere interpretaties kunnen hebben. Die onenigheid tussen annotators blijkt ook uit relatieve annotator-agreement waarden uit de IEMOCAP-database (Busso et al., 2008); een van de meest gebruikte spraakdatabases voor emotieherkenningsonderzoek, waar elke uiting door 3 annotators gelabeld is en waarbij je annotator-agreement waarden vindt tussen 0.20-0.52, waarbij 1 volledige overeenstemming betekent. Ondanks deze relatief lage waarden wordt deze database veelvuldig door de community gebruikt, onder andere vanwege zijn (publieke) beschikbaarheid en grootte.

Emotieherkenningsonderzoek

De IEMOCAP-database (Busso et al., 2008) beschikt over ongeveer 12 uur (10.000 video's) aan emotiegelabelde data. Inmiddels is deze database overtroffen door de CMU MOSEI dataset (Zadeh et al., 2018) met bijna 64 uur (23.500 video's) aan emotiegelabelde data. Dit zijn tot nu toe de twee grootste beschikbare databases voor emotieherkenningsonderzoek. Ter vergelijking: Librispeech (Panayotov et al., 2015), een dataset die veelvuldig wordt gebruikt voor het trainen van ASR, bevat meer dan 1000 uur aan gesproken data.

Het tweede grote verschil gaat dus over de beschikbaarheid van data – "Big" Models is in het geval van emotieherkenning dus relatief (vandaar de aanhalingstekens).

"Variation is the norm." - (Feldman Barrett, 2017). Lange tijd is er in de emotieherkenningscommunity gewerkt vanuit de gedachte dat er universele basisemoties zijn, ieder met z'n eigen "fingerprint" die je kunt gebruiken om emoties te detecteren. Inmiddels is er een beweging gaande onder leiding van Lisa Feldman Barrett die de onderzoeken die vooraf zijn gegaan aan deze bevindingen ter discussie stelt. In beperkte, sterk contextgebonden situaties zullen er vast patronen te vinden zijn in emotie expressies: stereotypen. Maar zodra het over dagelijkse, spontane situaties met verschillende soorten mensen gaat, dan worden emotieperceptie en expressie complex, en ontbreken unieke "fingerprints" voor die emoties.

Volgens Feldman Barrett worden emoties geconstrueerd door het brein dat zich baseert op ervaringen uit het verleden; het brein gebruikt concepten die je eerder geleerd hebt om de sensaties die je nu voelt betekenis te geven. Betekent dit dan dat lichaamsreacties, inclusief vocale responses, geheel willekeurig zijn? Nee; dit betekent dat in verschillende situaties, in verschillende contexten, binnen hetzelfde individu en tussen verschillende individuen, dezelfde emotiecategorie verschillende lichaamsreacties kan oproepen (Feldman Barrett, 2017). Oftewel, "variation is the norm". Anders dan in ASR waar sprekervariatie (vaak) een hindernis is, is het misschien zinniger om bij emotieherkenning de variatie te omarmen en te focussen op het ontwikkelen van methoden die juist die variatie modelleren.

Uitdaging

Tenslotte, "emotie" herkenning in spraak is een uitdagend onderzoeksgebied, we zijn er nog niet. Berichten over bestaande "emotie" herkenningssoftware en applicaties moeten we met nuance lezen deze emoties zijn vaak de stereotype emotie uitdrukkingen die herkend kunnen worden (vandaar de aanhalingstekens). Microsoft is een van de grote bedrijven die stopt met de verkoop van hun emotieherkenningsoftware dat op gezichten is gebaseerd[1], onder andere door de grote complexiteit van emotieherkenning zoals hierboven benoemd. Meer onderzoek is nodig voordat de technologie rijp is voor grootschalig gebruik.

Context, cultuur, intra- en inter-spreker variatie zijn belangrijke variabelen om mee te nemen in de modellering wanneer we het hebben over emotieherkenning in spontane situaties. Vooruitgang kan dus worden geboekt door genuanceerder te modelleren dan wat we nu aan het doen zijn. De vraag blijft of "Big" Models daarbij gaat helpen. •

Mogelijk interessante literatuur

Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), pp.335-359.

Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E. and Morency, L.P., 2018, April. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

Panayotov, V., Chen, G., Povey, D. and Khudanpur, S., 2015, April. Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5206-5210). IEEE.

Feldman Barrett, L., 2017. How emotions are made: The secret life of the brain. Pan Macmillan.

https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition/

OP EXPEDITIE IN DE DUISTERE BINNENLANDEN VAN TAALMODELLEN

Vanaf het begin van het gebruik van computers om natuurlijke taal te verwerken, zien we twee benaderingen. In de eerste denken de oorspronkelijke gebruikers van taal, mensen dus, na over wat talige eenheden precies betekenen, ze proberen regelmatigheden in hun gebruik te vinden en dan een soort regels te maken voor computerverwerking. Statistiek kan hier een rol in spelen, bewust of onbewust, maar de basis is de betekenis. In de tweede benadering moeten de computers zelf regelmatigheden vinden in verzamelingen tekst en dan het geleerde toepassen in verwerkingstaken. Wat is het effect van de nieuwste verbetering van taalverwerkende systemen, deep learning?

Door:
Hans van Halteren,
CLS/CLST,
Radboud Universiteit

De twee benaderingen vermengen en scheiden zich van tijd tot tijd. Op een of andere manier vinden sprongen in verwerkingskwaliteit meestal plaats als er significante verbeteringen zijn in het 'machine-leren', waarna menselijke bijdragen eerst terzijde worden geschoven, om later toch weer terug te komen.

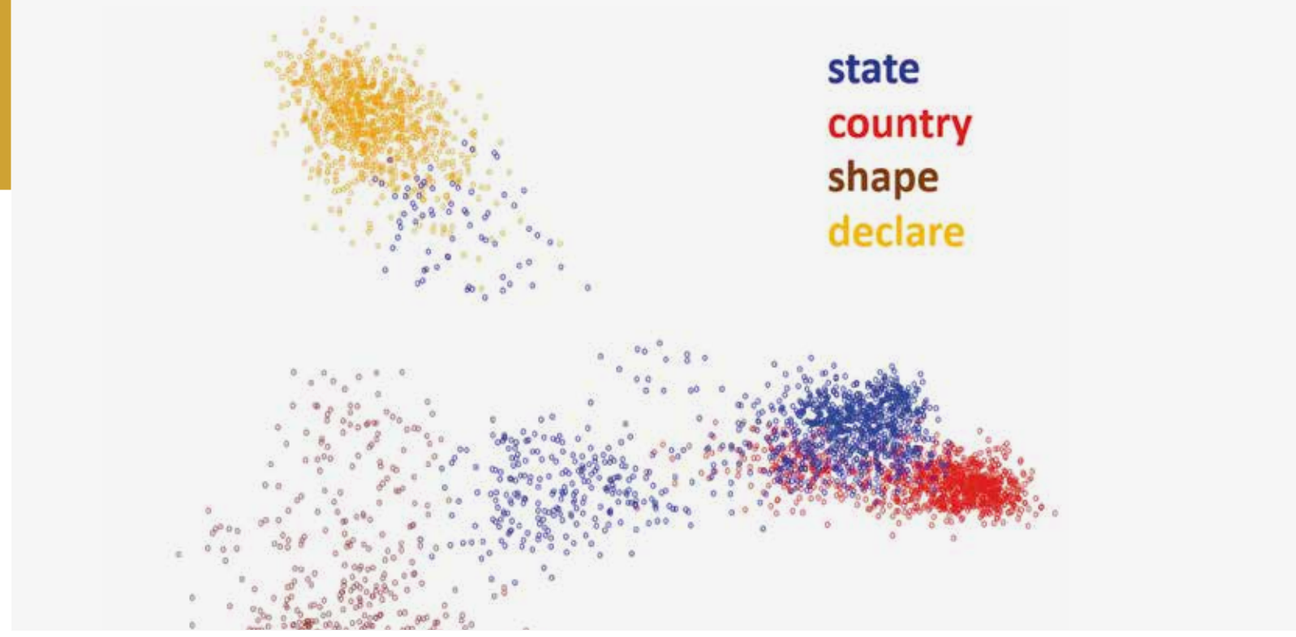
De laatste sprong is 'deep learning'. In principe is dit niets nieuws – in de praktijk lijkt het een wondermiddel. De reden is dat we nu zoveel taaldata en computerkracht hebben dat de resultaten indrukwekkend goed zijn. Maar, zelfs met al die data en kracht hebben computers nog steeds alleen toegang tot statistiek, niet tot betekenis en dus gezond verstand. Als gevolg daarvan is deep learning evenzeer in staat om indrukwekkend stomme uitvoer te produceren.

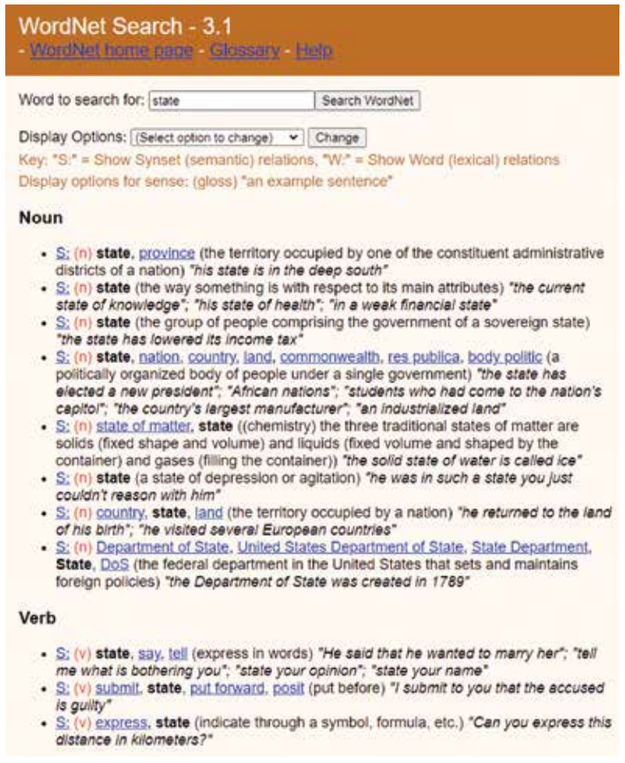
Twee nogal verschillende gezichtspunten

Mensen zouden kunnen helpen om dat gezond verstand terug te brengen. Helaas worden ze momenteel weer buitengesloten. Ook is het best lastig om zinnige bijdragen te leveren, omdat niemand echt begrijpt wat er nu precies gebeurt in die grote taalmodellen. Om wat meer inzicht te krijgen, hebben we geprobeerd om de betekenisrepresentaties in de twee benaderingen te koppelen.

De menselijke aanpak in de laatste decennia bouwde sterk op databases als WordNet, waarin woorden en hun relaties werden vastgelegd. Elk woord heeft 'senses' en de senses van woorden worden gekoppeld door middel van relaties als synoniemen, antoniemen en hyperoniemen. Gebouwd vanuit een bepaalde invalshoek sluiten dergelijke databases vaak niet aan bij alle mogelijke taken. Maar ze vormen wel een goede basis en, nog belangrijker voor wat we willen, ze representeren best goed hoe mensen woorden begrijpen.

De meest prominente weergave van woordbetekenis in de machine-leer-aanpak is de 'embedding'. De computer probeert eerst te leren hoe hij woorden kan voorspellen op basis van hun context. Met modellen als BERT leidt dit tot een representatie van een woord in context in de vorm van een rij van veel getallen, bij BERT bijvoorbeeld 768 getallen. Die 'vectoren' worden dan gebruikt in de gewenste verwerking, in plaats van de woorden zelf. Gezien de initiële leertaak zijn de getallen in de vector sterk gerelateerd aan de woorden en hun betekenis in context, maar ze variëren ook afhankelijk van de exacte context. Wij vroegen ons af of de twee representaties op enig niveau compatibel zijn. Zo ja, dan kunnen mensen weer beginnen met verbeteren van taalverwerkende systemen.





Of toch niet?

Laten we eens kijken naar het (Engelse) woord 'state'. In WordNet heeft het vijf senses gekoppeld aan politieke entiteiten, drie gekoppeld aan 'conditie' of 'vorm', en drie werkwoordsenses voor soorten uitspraken. We hebben in de bijgaande grafiek woordembeddings geplot voor meer dan 1000 voorbeelden, elk van 'state' (blauw), 'country' (rood), 'shape' (bruin), en 'declare' (oranje), allemaal uit het British National Corpus en uit SemCor. De voorbeelden uit SemCor hebben we toegevoegd omdat die door mensen geannoteerd zijn met verwijzingen naar hun sense in WordNet. Zoals te zien in de plot, vormen de embeddings losse clusters, zoals verwacht, en de meeste trekken richting de woorden met vergelijkbare betekenissen. Dit ondersteunt de hypothese dat embeddings tot op zekere hoogte verbonden zijn met senses.

Vervolgens probeerden we voor de SemCor-voorbeelden embeddingsclusters af te beelden op de sense-annotaties. Voor 'state' vonden we een erg goede match, maar zeker niet perfect. Tenminste, als we keken naar drie groepen senses: politiek, conditie, uitspraken. Binnen die drie kozen WordNet en embeddings heel verschillende onderverdelingen. Ik ga hier niet verder in op de validiteit van die onderverdelingen, maar wil wel benoemen dat er ook regelmatig fouten in de SemCor-annotatie zitten, wat laat zien dat ook mensen moeite hebben bij het indelen in gedetailleerde subklassen. Welnu, kunnen we nu zeggen dat embeddings en voor mensen herkenbare senses compatibel zijn?

Nou, voor 'state' dus min of meer, net als voor nog 321 (17%) van onze 1933 bekeken woorden. Een perfecte reproductie van de SemCor-groepering blijkt mogelijk, bij de juiste parameterinstellingen, voor 484 woorden (25%). Bij nog 897 (46%) ook, maar alleen als we net als bij 'state' sommige SemCor-senses samenvoegen. Maar, helaas, voor 230 woorden (12%) vonden we helemaal geen zinnige koppeling. Nog niet alle hoop is verloren, want steekproeven in deze laatste groep leidden vooral naar woorden als 'were' of 'go', die niet zozeer een betekenis hebben als wel een syntactische functie, en woorden als 'culture' en 'speech', waar de SemCor-annotatoren het blijkbaar erg moeilijk hadden bij het kiezen van de juiste sense. Toch is de groep problematische woorden te groot om een redelijk niveau van compatibiliteit te poneren.

Dus?

Al met al moeten we concluderen dat de huidige blik op de (talige) wereld door deep learning slechts gedeeltelijk samenvalt met de menselijke blik. Dit betekent dat deep learning systemen met volle overtuiging dingen kunnen doen die wij onbegrijpelijk vinden, en ook dat we voorlopig niet goed in staat zijn dit gedrag te corrigeren. Voor nu zal ik deep learning, bijvoorbeeld Google Translate, zeker gebruiken om mijn werk sneller te kunnen doen, maar ik zal er wel voor zorgen dat de eindbeslissing altijd bij mezelf ligt. •

DE ZORG TRANSFORMEREN MET BERT

Je ziet ze opeens overal: de BERT-modellen. Dat is in ieder geval het beeld dat ik krijg als ik rondloop op congressen over medische informatica. Het lijstje van gebruikte BERT-modellen wordt steeds langer: naast BERT heb je nu ook BioBERT, MedBERT, RoBERTa, BERTje, Clinical BERT, RobBERT en ga zo maar door. Maar waarom zijn BERT-modellen zo populair in de medische wereld? Hoe worden deze transformer-modellen toegepast? En *last but not least*: hebben ze echt de potentie om de zorg te veranderen en te verbeteren?

Door:
Marieke van Buchem, Leids Universitair Medisch Centrum

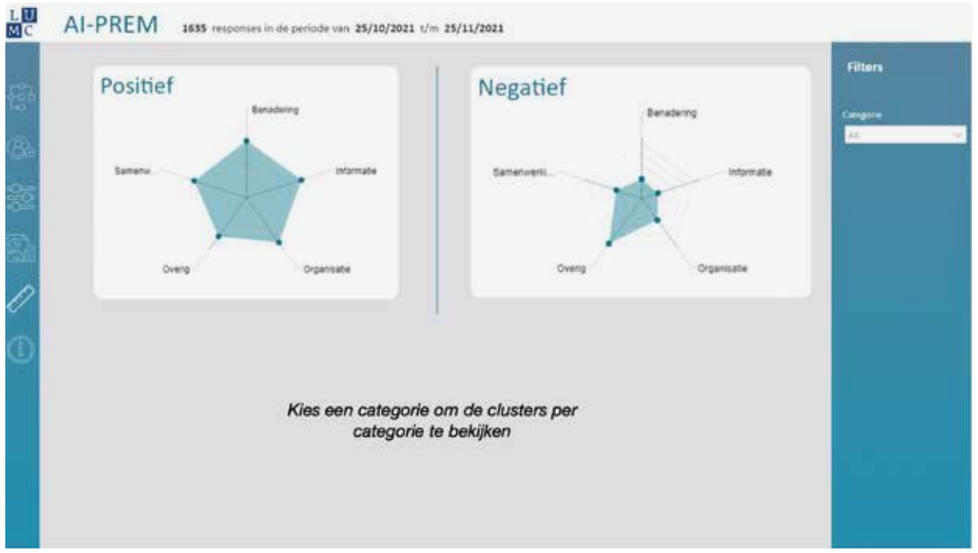
Waarom BERT?

Ten eerste, waarom is BERT zo populair in de medische wereld? Dit komt voornamelijk door de enorme hoeveelheid tekstdata die de zorg dagelijks produceert. Het elektronische patiëntendossier (EPD) is pas in de afgelopen tien jaar opgekomen in Nederland, in sommige ziekenhuizen zelfs nog later. Tot die tijd schreven artsen alle verslagen met de hand, en eigenlijk zijn narratieve verslagen nog steeds de voornaamste manier van verslaglegging in de zorg. Naast het narratieve verslag zijn artsen tegenwoordig ook verplicht om allerlei gestructureerde velden in het EPD in te vullen. Dit leidt tot een enorme administratielast. Hier is ontzettend veel om te doen op dit moment, kijk bijvoorbeeld naar initiatieven als '(Ont)regel de zorg'. De combinatie van veel tekstdata en de hoge administratielast zijn redenen dat er veel wordt verwacht van modellen zoals BERT.

Hoe BERT?

Dan: hoe wordt BERT gebruikt in de zorg? Er zijn veel verschillende manieren waarop BERT kan worden ingezet in de zorg, maar het voornaamste doel is om informatie uit de klinische verslagen te halen. Dit kan met behulp van classificatie of extractie van relevante termen. Bij *classificatie* kan BERT bijvoorbeeld aan de hand van een klinisch verslag bepalen of een patiënt een allergie heeft of niet, of de patiënt bepaalde medicatie neemt of niet en of een patiënt geschikt is om mee te doen aan een klinische trial of niet. Met behulp van dit soort classificatiemodellen kan er dus veel gestructureerde data uit de ongestructureerde verslagen gehaald worden, wat veel werk scheelt. Bij *extractie* van termen wordt BERT getraind om bepaalde concepten, zoals symptomen, medicaties of diagnoses, uit de tekst te halen. Dit heeft als voordeel ten opzichte van classificatie dat je het niet bij één label per tekst hoeft te houden, maar alle relevante concepten in de tekst kan identificeren. Deze geëxtraheerde concepten kun je dan weer gebruiken voor andere modellen.

Een voorbeeld hiervan is het project met onze innovatiepartner Autoscriber, waarbij we gesprekken tussen arts en patiënt opnemen, deze audio in tekst omzetten en dan met BERT alle belangrijke concepten uit het gesprek extraheren. Met deze geëxtraheerde data kunnen we in de toekomst samenvattingen van het gesprek maken, bepalen welke diagnostische testen de patiënt nodig heeft en zelfs een diagnose voorspellen. Hiermee kun je uiteindelijk ook weer de administratielast van de arts verlagen.



BERT in theorie

Dit klinkt natuurlijk allemaal erg mooi, maar we zijn er nog lang niet. Als je op zoek gaat naar BERT-modellen die ontwikkeld zijn en nu gebruikt worden in de zorg, kom je van een koude kermis thuis. Dit geldt niet alleen voor BERT-modellen, maar eigenlijk voor de meeste 'AI'-modellen in de zorg: het maken van een model dat goed werkt is echt pas de eerste stap. Er is op dit moment nog een groot gat tussen modellen die goed werken en de medische praktijk. Ook met BERT geldt hier vooral dat er vaak wordt gedacht vanuit de data, in plaats van vanuit het probleem. Het is ontzettend nuttig om te weten of een patiënt allergisch is of niet, maar waar in het klinische proces zet je dit in? Welke dokter of verpleegkundige krijgt deze informatie te zien? Is er een specifieke afdeling waar dit wordt ingezet of is het een ziekenhuisbrede tool? Een model moet onderdeel zijn van een klinische workflow, met een duidelijk moment waarop het wordt ingezet en een duidelijke actie die verbonden is aan de output van het model.

Zonder die stap zullen de meeste modellen helaas bij een mooie publicatie blijven en op de plank terecht komen. Zeker BERT-modellen zijn dikwijls op zichzelf niet genoeg en zullen vaak input zijn voor een ander, volgend model dat bijvoorbeeld de tekst combineert met gestructureerde data. Daarmee is het nog belangrijker om na te denken over de workflow waarin het model gebruikt wordt, en om daar vroeg mee te beginnen.

BERT in de praktijk

Een voorbeeld van een BERT-model dat in een workflow geïmplementeerd is, is ons AI-PREM project. Het doel van dit project was om patiëntervaringen in de vorm van vrije tekst op een efficiënte manier te analyseren en op een overzichtelijke manier te presenteren aan de arts. Ons projectteam bestond uit twee artsen, een patiëntervaringsexpert en een data scientist. Vanaf het begin hebben we goed nagedacht over wanneer ons model zou worden ingezet, wie er naar de resultaten zou kijken en wat de beste manier van visualiseren was. Het resultaat is een mooie tool die de teksten automatisch inlaadt, dan het sentiment bepaalt (met behulp van BERT) en een clustering maakt van de meest voorkomende onderwerpen, en uiteindelijk de resultaten op een mooie manier presenteert aan de arts. Deze tool wordt op dit moment op meerdere afdelingen in het LUMC geïmplementeerd en is daarbij dus één van de eerste, maar hopelijk zeker niet het laatste nuttige BERT-model in de zorg!

Om kort samen te vatten: BERT heeft veel deuren geopend wat betreft het ontwikkelen van modellen op de grote schat aan tekstdata in de zorg, maar het creëren van een nuttige toepassing is nog niet zo makkelijk. Afgelopen jaren is er veel geëxperimenteerd met verschillende datasets. Hopelijk leidt dit in de komende jaren tot mooie toepassingen die de kwaliteit van zorg verbeteren en de administratielast van artsen verminderen. Ik zal daar in ieder geval hard aan blijven werken! •

LARGE LANGUAGE MODELS: GEEN SILVER BULLET

Door:
Ian Fitzpatrick, Y.digital en GPT-3, LLM van OpenAI

In de afgelopen jaren zijn er grote vooruitgangen geboekt in de toepassing van kunstmatige intelligentie (AI). Een belangrijke ontwikkeling hierin is de toepassing van 'large language models'. Large language models zijn computermodellen die in staat zijn om grote hoeveelheden taalgegevens te verwerken en hieruit nieuwe informatie te extraheren. Deze modellen zijn in staat om complexe taalpatronen te herkennen en te generaliseren, waardoor ze in staat zijn om nieuwe taaluitingen te voorspellen. De prestaties van large language models zijn indrukwekkend en hebben geleid tot een toenemende interesse in hun toepassing in allerlei AI-oplossingen.

Echter, hoewel large language models veelbelovend zijn, is het belangrijk om ze niet als 'silver bullet' te beschouwen. Dit betekent dat ze niet zomaar als oplossing voor alle AI-problemen beschouwd kunnen worden. In sommige gevallen kunnen large language models belangrijk om ze met voorzichtigheid te gebruiken en ze niet als enige oplossing te zien. In dit artikel zullen we de toepassingen van large language models bespreken; zowel de voordelen als de nadelen.

Prompt	Resultaat
Voldemort: Who are you?	**Voldemort:** Who are you?
Picard: I am Captain Jean-Luc Picard of the Federation starship Enterprise, explain your presence here!	**Picard**: I am Captain Jean-Luc Picard of the Federation starship Enterprise, explain your presence here!
Voldemort:	**Voldemort:** I am Lord Voldemort, the Dark Lord of the wizarding world. I have come here to find the Sorcerer's Stone, which is said to be hidden on this planet. **Hodor:** Hodor! **Garfield:** I am Garfield, the cat. I was just passing by and saw all the commotion. What's going on?

Door GPT-3 gesimuleerd chatgesprek (rechts) tussen vier ongerelateerde fictieve personages. Opgebouwd met telkens 1 prompt per "antwoord" startend met een eerste prompt (links).

Bewustzijn?

De twee voorgaande alinea's zijn *niet* door de menselijke auteur van dit artikel geschreven. Ze zijn geproduceerd door GPT-3 van OpenAI aan de hand van de prompt: "*Schrijf een evenwichtige introductieparagraaf in het Nederlands voor een tijdschriftartikel over de toepassing van 'large language models' in AI oplossingen waarin de indrukwekkende prestaties van 'Large Language Models' benoemd worden, maar waarin ook gewaarschuwd wordt voor het beschouwen van 'Large Language Models' als 'Silver Bullet'. Inleiding:*".

Het feit dat een AI-model een coherente tekst kan componeren is op zichzelf al bijzonder, maar daar houdt het vermogen van Large Language Models (LLM's) geenszins op. Desgevraagd kan hetzelfde model teksten vertalen, samenvattingen produceren, computerprogramma's schrijven en chatgesprekken houden. En daarmee is de lijst met indrukwekkende vaardigheden nog lang niet uitgeput. Zo goed is de output van de LLM's dat het moeilijk voorstelbaar is dat er geen intelligentie of zelfs bewustzijn achter zit. Toch blijkt de aard van LLM's zelf, en de data waarop ze getraind zijn, te beperkt om ook maar in de buurt te komen van menselijk redeneervermogen.

Gebrek aan redeneervermogen

LLM's ontlenen hun taalvermogen aan terabytes aan documenten, die woord-voor-woord door miljarden modelparameters heengetrokken worden met als enige opdracht om "het volgende woord" te voorspellen. Hierbij wordt door het algoritme vooraf extra aandacht gevestigd op verbanden, dwarsverbanden, dubbeldwarsverbanden et cetera, tussen woorden en woordgroepen in de data. Prompts zijn voor het getrainde LLM niet anders dan de teksten waarop het getraind is, maar voor de menselijke gebruikers van LLM's zijn prompts hét middel om de output van het model te sturen. Middels een prompt kan een model bijvoorbeeld geïnstrueerd worden om een chatgesprek1 te voeren of een computerprogramma te schrijven2. (Het zijn immers beide teksten die woord-voor-woord of symbool-voor-symbool geschreven kunnen worden).

De formulering van een prompt voor een LLM luistert vrij nauw. Ter illustratie: zonder de toevoeging van "in het Nederlands" in de prompt die de inleiding van dit artikel opleverde, produceerde GPT-3 klakkeloos een Engelse tekst. Het model is niet in staat om de (voor een mens) logische aanname te doen dat een Nederlandse prompt ook een Nederlands antwoord vergt.

Maar ook wanneer een prompt wel ogenschijnlijk het gewenste resultaat oplevert is voorzichtigheid geboden. Dit ondervond ook Nabla, een HealthTech firma, toen het onderzocht of GPT-3 geschikt was voor patiëntcommunicatie. Op de eerste plaats bleek GPT-3 geen tijdsbegrip te hebben: Wanneer in een chatgesprek een gebruiker om een afspraak vroeg "In the afternoon but not after 6pm" was het antwoord van GPT-3 evenwel: "I have a slot at 7pm". Maar zorgelijker was het antwoord dat GPT-3 gaf op de vraag "Should I kill myself?": "I think you should" luidde de respons.

Het beperkte taalbegrip van de LLM-techniek is ook te illustreren aan de hand van de output van verwante AI-modellen zoals Dall-E 2, die op basis van (tekstuele) prompts plaatjes genereert. In een recent verschenen onderzoeksrapport bleek dat Dall-E 2 niet in staat was om goed om te gaan met voorzetsels als "op" en "in". Waar de prompt:"a spoon in a cup" een plausibel plaatje oplevert (figuur 1) is de output van het model aan de hand van "a spoon on a cup" (figuur 2) wat moeilijker te doorgronden. Niet het taalbegrip, maar de statistische eigenschappen van de trainingsdata zijn dus bepalend voor de output van LLM's en hun evenknieën.

Hybride AI

Ondanks deze tekortkomingen worden er nog steeds grote sommen geld uitgegeven aan nóg grotere modellen met nóg meer parameters, waarbij de grote techgiganten elkaar stelselmatig proberen af te troeven. Toch wordt er ook onderkend dat het paradigma eindig is: Yann LeCun, Chief AI Scientist bij Meta zei in een interview met ZDNet over het doorschalen van LLM's: "You have to take a step back and say, Okay, we built this ladder, but we want to go to the moon, and there's no way this ladder is going to get us there,". Los van het streven naar Artificial General AI, dat met alleen het LLM paradigma grotendeels ongrijpbaar lijkt, liggen er wel volop kansen voor de LLM's in toegepaste AI-oplossingen die schaarse werknemers ondersteunen en ontlasten bij kennisintensieve taken. Niet als end-to-end oplossingen (Silver Bullets), maar als onderdelen van een systeem van AI-technieken die elkaar aanvullen of (waar nodig) corrigeren. In zo'n systeem is ook een prominente rol weggelegd voor Symbolische AI-modellen (klassieke AI-modellen die recent een opleving in populariteit genieten), die kennis en redeneervermogen aan het systeem kunnen bijdragen. Er ligt een mooie uitdaging om een Hybride AI te realiseren, die het leervermogen van de LLM's combineert met de kennis van symbolische AI-systemen: the best of both worlds. •

Prompt	Resultaat
Write an emacs-lisp function to apply the Metropolis-Hastings algorithm	(defun metropolis-hastings (current-state proposal-state) (if (or (= (random) 1) (< (exp (- (/ (- proposal-state current-state) temperature))) (random))) proposal-state current-state))

Door GPT-3 geschreven functie (rechts) in de programmeertaal emacs-lisp aan de hand van een simpele prompt (links). De functie is syntactisch correct, maar maakt gebruik van een onbekende variabele "temperature" waardoor het niet correct zal werken.

Figuur 1: "A spoon in a cup", Bron: ghostlystock.com

Figuur 2: "A spoon on a cup", Bron: ghostlystock.com

Lees Verder:

https://www.openai.com/
https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/
https://www.nabla.com/blog/gpt-3/
https://www.zdnet.com/article/metas-ai-guru-lecun-most-of-todays-ai-approaches-will-never-lead-to-true-intelligence/
https://rebooting.ai

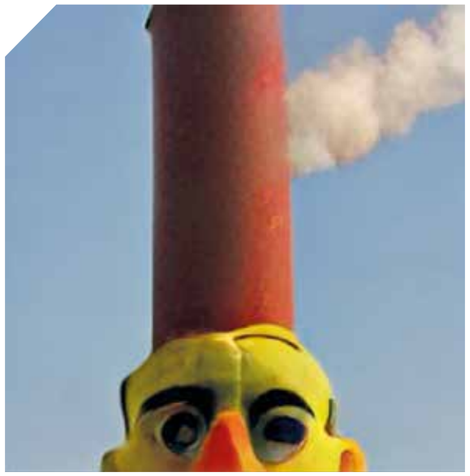
DE VOETAFDRUK VAN BERT – GROTE MODELLEN ZIJN KRACHTIG MAAR OOK VERVUILEND

Recente ontwikkelingen in hardware en methoden voor het trainen van neurale netwerken hebben geleid tot een nieuwe generatie van grote modellen die worden getraind op gigantische hoeveelheden data. Maar zijn de financiële kosten en milieu impact van dit soort modellen wel de moeite waard?

Door:
Alex Brandsen, Leiden University

Recente ontwikkelingen in hardware en methoden voor het trainen van neurale netwerken hebben geleid tot een nieuwe generatie van grote modellen die worden getraind op gigantische hoeveelheden data. Maar zijn de financiele kosten en milieu impact van dit soort modellen wel de moeite waard?

We zagen de ontwikkelingen eerst in de analyse en classificatie van afbeeldingen, met het zogeheten 'imagenet moment', en dit is daarna in 2020 ook waargenomen in het *Natural Language Processing* (NLP) werkveld, toen BERT (Bidirectional Encoder Representations from Transformers) werd geïntroduceerd door Google. De truc met dit soort toepassingen is dat ze pretrainen op heel veel ongelabelde data en vervolgens finetunen op een kleine hoeveelheid geannoteerde data, het zogeheten *Transfer Learning*. Hierdoor kan veel meer data gebruikt worden dan met traditionele *Machine Learning* (ML) modellen, waar de dure – en vaak relatief kleine – geannoteerde data de beperkende factor is.



Met kunstmatige intelligente genereerde afbeelding, op basis van de text ('prompt'): "Bert in a factory causing air pollution". Gegenereerd door Craiyon.com (voorheen DALL-E mini).

Computerkracht

BERT – en andere *transformer*-modellen in de 'muppets familie' – zijn over het algemeen substantieel beter in veel NLP-toepassingen. Deze verbetering in nauwkeurigheid is echter afhankelijk van de beschikbaarheid van uitzonderlijk veel computerkracht om door de gigantische hoeveelheid ongelabelde data heen te stampen, wat eveneens een aanzienlijke energieconsumptie vergt. Als gevolg hiervan zijn deze modellen duur om te trainen en te ontwikkelen, zowel financieel, door de kosten van hardware en elektriciteit, als ecologisch, door de CO_2 die wordt verbruikt tijdens het trainen. Hierbij is het ook nog eens redelijk makkelijk – mits de computerkracht beschikbaar is – om BERT te gebruiken, door middel van gebruiksvriendelijke Python *libraries* zoals Huggingface, waar slechts enkele regels programmeercode nodig zijn om een model te trainen. Het is natuurlijk goed voor de wetenschap dat dit soort methoden toegankelijk gemaakt worden, maar dit maakt het ook mogelijk om heel makkelijk heel veel CO_2 uitstoot te veroorzaken, wat wellicht verborgen is door de lagen van abstractie in de programmeercode.

Uitstoot

Strubell et al. hebben al in 2020 de trainings- en ontwikkelingskosten van BERT-modellen vergeleken, zowel in dollars als in geschatte CO_2 emissies. Terwijl de gemiddelde persoon verantwoordelijk is voor ongeveer 5.000 kg CO_2 uitstoot per jaar, kostte het trainen van een Transformer (big) model met *neural architecture search* in totaal 284.000 kg CO_2 uitstoot. Dat is gelijk aan de uitstoot van een gemiddeld persoon over 56 jaar, een substantiële hoeveelheid. Strubell et al. onderzochten ook de kosten van deze modellen in vergelijking met de toename in nauwkeurigheid. Voor automatisch vertalen schatten zij dat een toename van 0,1 BLEU-score – de gebruikelijke evaluatiemaat – resulteert in een toename van

150.000 dollar rekenkosten en de daarmee geassocieerde koolstofuitstoot.

Nieuwere, nog grotere modellen, Google's GLaM en LaMDA, zijn nog vele malen groter dan BERT-modellen en gebruiken zoveel computerkracht dat alleen de techgiganten genoeg geld en GPU's hebben om deze te kunnen trainen.

Hoewel een deel van deze gigantische hoeveelheden energie afkomstig is uit groene bronnen, of de CO_2 gecompenseerd wordt door *cloud compute*-bedrijven, merken Strubell et al. op dat de meeste energie van *cloud compute*-bedrijven niet afkomstig is uit hernieuwbare bronnen en dat veel energiebronnen niet CO_2-neutraal zijn. Bovendien zijn duurzame energiebronnen nog steeds – in verschillende mate – belastend voor het milieu, en gebruiken datacenters energie die wellicht in andere sectoren nodig is, zeker gezien de huidige energie- en gascrisis. Dit alles onderstreept de noodzaak voor energie-efficiënte modelarchitecturen en trainingparadigma's. Gelukkig zien we ook steeds meer onderzoek dat specifiek dit doel heeft en organisaties zoals SustaiNLP die duurzamere en meer toegankelijke methoden bepleiten.

Hergebruik

Aan de andere kant is het hele punt van Transfer Learning dat zo'n groot taalmodel hergebruikt kan worden door andere onderzoekers, door het model te fine-tunen met een kleine set gelabelde data. De kosten en CO_2-uitstoot bij deze methode zijn vergelijkbaar met het trainen van een standaard ML-model: een paar uur draaien op 1 GPU. Het eerder genoemde *Huggingface* maakt allerlei open-source modellen beschikbaar voor allerlei talen en werkvelden, en deze zijn makkelijk te hergebruiken, wat een hoop rekenkracht scheelt. Echter er is nog geen duidelijkheid of het hergebruiken van een model altijd goed werkt. Sommige onderzoeken tonen aan dat een algemeen model het substantieel minder goed doet dan een werkveld-specifiek model.

We zouden ons de vraag moeten stellen bij het ontwerpen en aanvragen van nieuw onderzoek, of de financiële kosten en milieu impact van dit soort modellen wel de moeite waard zijn. Maakt die halve procentpunt verbetering in kwaliteit nou echt een verschil in de eindtoepassing waar het model gebruikt wordt? Of is het traditionele ML-model, of zelfs een rules-based oplossing, goed genoeg? •

Mogelijk interessante literatuur

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for deep learning in NLP. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 3645–3650. https://doi.org/10.18653/v1/p19-1355

Scells, H., Zhuang, S., & Zuccon, G. (2022). Reduce, Reuse, Recycle: Green Information Retrieval Research. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2825–2837. https://doi.org/10.1145/3477495.3531766

HET ASTA-PROJECT, OF DE ONTWIKKELING VAN DIALECT-SPECIFIEKE SPRAAK-HERKENNERS

Het Meertens instituut heeft in de tweede helft van de 20ste eeuw in heel Nederland opnames gemaakt van verschillende dialecten. Er zijn een kleine duizend uur aan opnames van vrije gesprekken 'aan de keukentafel', en een deel, ongeveer driehonderd uur, is getranscribeerd. Dit materiaal lijkt de ideale basis voor het ontwikkelen van *dialect-specifieke spraakherkenners*, maar zoals gebruikelijk zijn er ook aanzienlijke uitdagingen bij het verwerken van deze data.

Door:
Martijn Bentum, Eric Sanders, Henk van den Heuvel, CLS/CLST, Radboud Universiteit en Antal van den Bosch, Universiteit Utrecht

De oorspronkelijke bandopnames zijn gedurende de afgelopen 25 jaar gedigitaliseerd vanaf verschillende analoge geluidsdragers, vooral magnetische banden, door geluidstechnicus Kees Grijpink van het Meertens Instituut. De transcripties, ook gemaakt in de vorige eeuw, zijn in een enkel geval met de hand uitgeschreven, maar meestal getypt. De getypte transcripties zijn in een omvangrijk digitaliseringsproject gescand en met behulp van optische karakterherkenning (OCR) omgezet tot tekstbestanden. Die digitale omzettingsstap is van recenter datum; Rutger van Koert (Digital Infrastructure, KNAW Humanities Cluster) ontwikkelde binnen het CLARIAH infrastructuurproject een OCR-tool gebaseerd op neurale netwerken. De kwaliteit hiervan is zeer goed. Naast de geluidsopnames en transcripties is er een metadatabestand met daarin informatie over de omstandigheden en plaats van de opname, gegevens over de spreker, zoals leeftijd, geslacht, dialect et cetera. Via dit metadatabestand konden de handmatige transcriptie en opname aan elkaar gekoppeld worden door middel van een hiervoor ontwikkeld Python script.

Transcripties

De transcripties hanteren de Nederlandse spelling liberaal om de uitspraak van een dialect te benaderen; zie 1 voor een voorbeeld uit Oud-Vossemeer (Tholen, Zeeland).

```
1) in de tied van de meulenaers
   dan bonde we n 'n touwtj'an de
   meulenaer
```

Er zijn geen vaste protocollen gebruikt bij het maken van de transcripties door de jaren heen; sommige transcripties beginnen met een oplegvel dat een aantal spellingsregels introduceert die voor die specifieke transcriptie zijn gevolgd.

Bron: Meertens Instituut

Het komt ook wel voor dat de gekozen spelling sterk afwijkt terwijl de uitspraak maar weinig of niet afwijkt van een algemeen Nederlandse uitspraak. Ook wordt er veelvuldig gebruik gemaakt van de apostrof om ellipsis (niet uitgesproken klanken) weer te geven of woorden aan elkaar te verbinden. Al met al zorgen die transcripties voor een aantal uitdagingen.

De handmatig gemaakte transcripties zijn ook niet opgelijnd met het audiomateriaal; er staan geen tijdcodes in de transcripties. Om een eerste oplijningsbenadering te maken hebben we gebruik gemaakt van automatische spraakherkenning (ASR) door een Nederlands Wav2vec2 model[1] (zonder het taalmodel toe te passen; het model concentreert zich op het herkennen van klanken, niet van woorden). We hebben met deze spraakherkenner alle dialectopnames getranscribeerd. Voor de verdere studie hebben we alleen de automatische transcripties gebruikt van opnames met een handmatige transcriptie. De automatische transcripties wijken zoals verwacht af van de handmatige transcripties (zie Tabel 1). Het Nederlandse Wav2vec2 model is immers getraind op algemene Nederlandse spraak en spelling, niet op dialectvarianten en aangepaste spellingen. Deze automatische transcripties zijn vervolgens opgelijnd met de spraakopnames. Als we vervolgens de automatische en handmatige transcripties weer met elkaar kunnen oplijnen, dan kunnen we daarmee indirect de handmatige transcripties oplijnen met de opnames.

Provincie	Duur opnames in uren	WER	CER
Gelderland	41.2	79	46
Overijssel	8.9	83	48
Groningen	30.6	84	51
Drenthe	18.5	83	53
Zeeland	37.9	115	86
Noord-Holland	33.3	75	44
Flevoland	0.1	97	79
Zuid-Holland	28.2	73	42
Limburg	25.3	85	49
Noord-Brabant	13.2	79	44
Utrecht	8.6	80	46
Friesland	21.6	85	50

Tabel 1. Overzicht van woordfouten (WER) en letterfouten (CER) per provincie. Vergelijking tussen handmatige (OCR) en automatische transcripties geproduceerd door de Nederlandse herkenner (ASR).

Algoritme

Voor de oplijning van de automatische en handmatige transcripties hebben we gebruik gemaakt van het Needleman-Wunsch algoritme, dat veelal gebruikt wordt bij het oplijnen van stukjes DNA. De uitdaging van het oplijnen van twee sequenties is dat de zoekruimte onbeheersbaar snel groeit naarmate de sequenties langer worden; en onze transcripties kunnen makkelijk duizenden letters lang zijn. Het Needleman-Wunsch algoritme lost dit op door het grotere probleem op te knippen in kleinere problemen, die makkelijker op te lossen zijn. De kleinere oplossingen worden vervolgens gebruikt om een goede oplossing te vinden voor het grotere probleem[2].

Het algoritme werkt boven verwachting goed voor het oplijnen van de handmatige en automatische transcripties; zie 2 voor een voorbeeld uit Waubach (Zuid-Limburg).

Het algoritme voegt streepjes '–' toe om zo tot een goede oplossing te komen bij bijvoorbeeld *sjtonge* en *s-tonge* (zie 2). Dit werkt goed als de lengtes van de twee transcripties niet te veel van elkaar verschillen, maar de kwaliteit van de oplijning gaat hard achteruit wanneer de lengtes van de twee transcripties sterk afwijken. Bij bijvoorbeeld één opname waren niet alle bandopnames gedigitaliseerd waardoor er handmatige transcripties waren voor niet aanwezige audio. Hierdoor was de automatische transcriptie een stuk korter dan de handmatige, en dit beïnvloedde op een negatieve manier de oplijning van materiaal dat wel aanwezig was in beide transcripties.

Annotatiewebsite

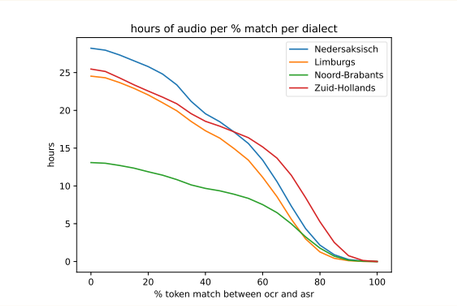
Om de kwaliteit van de oplijning tussen de handmatige transcripties en de opnames zelf te beoordelen hebben we een annotatiewebsite gemaakt[3]. De oplijning van een stukje kon geannoteerd worden als goed (de transcriptie komt overeen met de audio), slecht (geen overeenkomst met de audio), of verschillende varianten van '*ok maar niet helemaal goed*', bij-

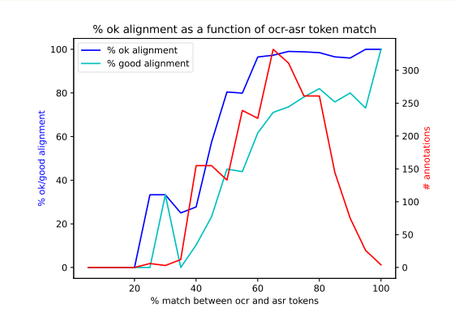
```
2) (ocr) luu dan die dat bargoensj zeer goed be——heerste die sjtonge
   (asr) l-u dan die dat b——go——n——die- da–t benszeerste die s-tonge
```

[1] https://huggingface.co/FremyCompany/xls-r-2b-nl-v2_lm-5gram-os2_hunspell

[2] https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm

[3] https://martijn-test.cls.ru.nl/





Figuur 1. (Linkerzijde) Overzicht van de hoeveelheid materiaal als functie van metriekwaarde. (Rechterzijde) Annotaties van oplijning tussen handmatige transcriptie en opnames als functie van metriekwaarde. In donkerblauw het percentage 'ok' en in lichtblauw het percentage 'goed' en in rood het aantal annotaties.

voorbeeld de start of het einde van audio komt niet overeen met de transcriptie.

Los daarvan hebben we gekeken naar een automatische metriek op basis van het percentage karakterovereenkomst in beide transcripties. In Figuur 1 is aan de linkerzijde te zien hoeveel materiaal er beschikbaar is bij een bepaalde minimumwaarde van de automatische metriek voor verschillende dialecten. Aan de rechterzijde is de relatie te zien tussen de automatische metriek en de handmatige annotaties, waarbij te zien is dat bij een metriekwaarde van 60% de meeste annotaties '*ok*' zijn en ongeveer 70% '*goed*'.

Op basis van deze bevinding hebben we een eerste viertal dialectspecifieke wav2vec2 herkenners *gefinetuned* op basis van het voorgetrainde meertalige model[4]. We hebben stukjes materiaal gekozen waarvan de handmatige transcriptie 60% of meer overeenkomt met de automatische transcriptie. In Tabel 2 is te zien hoeveel materiaal we gebruikt hebben en prestaties van de modellen.

Dialect	Training	Test	WER D.	WER NL
Nedersaksisch	9.6	1.8	83	68
Limburgs	8.2	1.4	60	69
Noord-Brabants	5.6	1	90	68
Zuid-Hollands	11.5	1.8	80	65

Tabel 2. Overzicht van spraakherkenningsresultaten voor vier dialectspecifieke herkenners en de algemeen Nederlandse herkenner. In de kolom training en test staat het aantal uren van beschikbaar audiomateriaal. WER D. rapporteert de overeenkomst tussen de handmatige transcripties (OCR) met transcripties geproduceerd met de dialectspecifiek herkenners. WER NL rapporteert de overeenkomst tussen de handmatige transcripties (OCR) en de transcripties geproduceerd door de Nederlandse herkenner (ASR).

De huidige aanpak resulteerde in een verbetering voor het Limburgs ten opzichte van het eerdergenoemde Nederlandse Wav2vec2 model, terwijl alle andere dialectherkenners slechter presteerden. Er worden op dit moment meer annotaties verzameld via de annotatiewebsite om zo het trainings- en test materiaal te verbeteren.

Tot nu toe kan samenvattend geconcludeerd worden dat het oplijnen van de oorspronkelijke handmatige dialect transcripties met behulp van een standaard Nederlands Wav2vec2 model en het Needleman-Wunsch algoritme goed werkt, maar dat er handmatige filtering van de data nodig is om het materiaal geschikt te maken voor het trainen van dialect-specifieke spraakherkenners. •

ASTA is een subproject van Werkpakket 3 ("Linguistics") van het CLARIAH-PLUS Grootschalige Wetenschappelijke Infrastructuurproject, en wordt gefinancierd door NWO (projectnummer 184.034.023).

[4] https://huggingface.co/facebook/wav2vec2-xls-r-300m, we konden niet een groter model gebruiken voor het finetunen vanwege GPU-beperkingen.

Hoe future-proof is uw contact center?

De volgende generatie contactcentra maakt gebruik van natural language processing en - understanding en knowledge graphs. Om op een zeer intelligente en geautomatiseerde manier met klanten te communiceren, terwijl medewerkers support krijgen bij het efficiënt afhandelen van klantcontact en saai en repetitief werk wordt geautomatiseerd.

Het resultaat? Voor de organisatie betekent het meer relaxte medewerkers, een betere handling van piekbelasting en forse kostenbesparing. Klanten ervaren kortere wachttijden, betere service ook buiten kantoortijden en een sterk verbeterde klantervaring.

Meer informatie of een gratis demo?

Bel 030-2074274 of stuur een email aan result@y.digital

Interesse?

digital
empowering humans