Is de zoekmachine van de toekomst een chatbot?

Verberne, S.

Prof. Dr. Suzan Verberne

Is de zoekmachine van de toekomst een chatbot?

Is the search engine of the future a chatbot?

(English version included)

Universiteit Leiden

Is de zoekmachine van de toekomst een chatbot?

Oratie uitgesproken door

Prof. Dr. Suzan Verberne

bij de aanvaarding van het ambt van hoogleraar

Natural Language Processing

aan de Universiteit Leiden

op maandag 3 juni 2024

Universiteit
Leiden

English text from page 12.

Is de zoekmachine van de toekomst een chatbot?

Mevrouw de rector magnificus, geacht faculteitsbestuur, zeer gewaardeerde toehoorders. Ik wil het in deze lezing hebben over **zoekmachines en chatbots**. Er was een tijd dat dit twee totaal van elkaar gescheiden toepassingen waren, maar tegenwoordig overlappen ze en gaan ze meer op elkaar lijken. Chatbots op basis van grote taalmodellen, zoals ChatGPT, worden door veel mensen gebruikt om antwoorden te krijgen op vragen, en bedrijven achter grote zoekmachines zoals Google en Microsoft zijn bezig om chatbot-technologie in te bouwen in hun interfaces. Is de zoekmachine van de toekomst een chatbot? Vandaag ga ik die vraag beantwoorden. Het korte antwoord is ja. Maar ook het omgekeerde is waar: de chatbot van de toekomst is een zoekmachine.

Deze lezing bestaat uit vier delen: Het eerste deel gaat over grote taalmodellen. Ik begin met uitleg over wat taalmodellen zijn en hoe de technologie achter ChatGPT werkt. Daarna leg ik aan de hand van recent onderzoek uit wat deze grote taalmodellen voor ons kunnen doen, ik vertel wat de problemen en uitdagingen van deze modellen zijn, en in welke richtingen we de oplossingen kunnen zoeken. Het tweede deel gaat over zoekmachines. Mijn groep doet veel onderzoek naar de ontwikkeling van methoden om mensen te helpen specifieke informatie te vinden, zoals antwoorden op medische en juridische vragen. Ik ga uitleggen hoe we dit doen en wat de uitdagingen daarvan zijn. Het derde deel van mijn lezing gaat over hoe zoekmachines grote taalmodellen kunnen helpen en andersom. In het vierde en laatste deel geef ik een doorkijkje naar de toekomst: wat is de huidige focus van ons onderzoek en wat zijn belangrijke richtingen voor de volgende stappen?

Deel 1. Grote taalmodellen

Sinds de release van ChatGPT in het najaar van 2022 heeft iedereen kennis kunnen maken met de kracht van grote taalmodellen. Voor de rest van mijn lezing is het belangrijk om te weten hoe deze modellen werken. ChatGPT is een chatbot gebaseerd op het taalmodel GPT. De G in GPT staat voor *generative*; GPT is een **generatief taalmodel**. Dit is een model dat tekst genereert door na elk woord te bepalen wat het meest waarschijnlijke volgende woord is. Dit soort taalmodellen bestaan al decennia (Van den Bosch, 2012, 2008; Verberne, 2002). Als ik u de zin voorlees

> "De trein kon niet verder want er lag een boom op het…",

dan weet u dat het meest waarschijnlijke volgende woord 'spoor' is. Door grote hoeveelheden taal te horen en te lezen hebben we geleerd welk volgend woord het meest waarschijnlijk is.

Taalmodellen zoals ChatGPT gebruiken de waarschijnlijkheden van woorden om een vloeiende tekst te genereren. Maar er zit meer achter. De P in GPT staat voor *pre-trained* (Radford et al., 2018). Dit verwijst naar een belangrijk kenmerk van deze modellen: ze zijn voorgetraind op heel, heel veel tekst. Door grote hoeveelheden tekst te verwerken leert het model welke woorden vaak samen voorkomen en welke woorden niet. Nu is het zo dat woorden die vaak samen voorkomen qua betekenis meer op elkaar lijken dan woorden die niet vaak samen voorkomen. Dit is een oud principe uit de taalkunde, de **distributional hypothesis** (Harris, 1954). Die hypothese stamt uit 1954 en is nog altijd de basis voor alle moderne taalmodellen. Laat ik nog een voorbeeld geven. Ik ga ervan uit dat u nog nooit gehoord heeft van het woord *gleevec*. Maar als ik de volgende zin geef, dan kunt u waarschijnlijk de betekenis raden:

> "Ik heb gleevec voorgeschreven gekregen, maar ik krijg er een droge huid van."

Kunt u de betekenis raden? Gleevec is een medicijn. Doordat we in ons leven heel veel taal gehoord en gelezen hebben, kunnen we de betekenis van woorden afleiden uit de context. Dit is ook wat taalmodellen doen: door te leren te voorspellen welke

woorden waarschijnlijk zijn op welke plaats in de tekst leert het model welke betekenissen woorden hebben.

Nu even terug naar de afkorting GPT. Ik heb al uitgelegd dat de G staat voor generatief en de P voor pre-trained. De T staat voor Transformer. De transformer is een type computermodel, ontwikkeld in 2017 (Vaswani et al., 2017) dat heel goed is in het verwerken van tekst en het leren van de betekenissen van woorden. Dat komt doordat het model relaties berekent tussen alle woorden in de tekst, ook als de zinnen lang zijn. Een taalmodel dat gemaakt is met een transformer heeft daardoor geleerd dat er een relatie is tussen treinen en sporen, waardoor het, net als mensen, heel goed kan raden wat het volgende woord is in de zin "De trein kon niet verder want er lag een boom op het …".

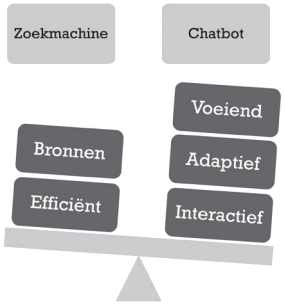
Het voorspellen van het volgende woord is op zichzelf niet een heel handige toepassing, maar nadat een transformermodel is getraind om woorden te voorspellen, kan het ook andere taken die met woordbetekenis te maken hebben leren.

Om een voorgetraind taalmodel een specifieke taak aan te leren, moet het model voorbeelden te zien krijgen: honderden tot duizenden voorbeelden van wat we willen dat het model doet. Dit noemen we **finetunen** (Devlin et al., 2019). Via finetunen kan een model bijvoorbeeld worden aangeleerd om teksten in te delen naar onderwerp of om in een tekst specifieke soorten woorden te vinden, zoals namen van medicijnen. Met mijn studenten en promovendi heb ik op die manier modellen getraind voor allerlei soorten taken. Ook GPT kan gefinetuned worden. Zo maakten we een GPT-model meer empathisch door het te finetunen op empathische dialogen (Chen, 2023). Helaas verloor het model hierdoor kennis van de wereld en kon het vragen zoals "wat is groter, de maan of de zon" niet meer beantwoorden. Maar vriendelijker was het wel.

Tussen 2018 en 2022 werden taalmodellen groter: ze hadden meer parameters en werden getraind op meer data. Daardoor werden ze ook krachtiger en de output vloeiender. Door een groot taalmodel te finetunen op conversaties, kan het dialogen voeren; zo wordt een groot taalmodel een **chatbot**. Toen ChatGPT toegankelijk werd via een online interface, ontdekten de media en de rest van de wereld wat de kracht van grote taalmodellen is.

ChatGPT is weliswaar bereikbaar via een website; over het achterliggende model is niet publiek gemaakt hoe het precies getraind is en op welke data. Voor onderzoek gebruiken we liever niet-commerciële modellen die **open** zijn en die we kunnen downloaden naar onze eigen computers. Via de website HuggingFace zijn transformermodellen beschikbaar voor allerlei talen en taken, waaronder een heel aantal open alternatieven voor ChatGPT.

Deze grote taalmodellen zijn zo krachtig dat finetunen voor veel taken niet eens meer nodig is; ze kunnen tekstuele instructies opvolgen (Brown et al., 2020). Zo ontdekten we dat een groot taalmodel goede samenvattingen van wetenschappelijke artikelen kan maken met duidelijke instructies en slechts twee voorbeelden. Het was zelfs iets beter dan een eerder transformermodel dat was gefinetuned op 5000 voorbeelden (Zakkas et al., 2024). Omdat GPT **vloeiende tekst** kan genereren, is samenvatten een kerntaak waar het best goed in is. Hierdoor lijkt het model intelligent, maar is het dat ook? Ik ben zelf van mening dat wat taalmodellen doen het **simuleren** van menselijke taal is. Een papagaai met wat random variatie, zou je kunnen zeggen (Bender et al., 2021).



Want zoals ik heb uitgelegd, genereren taalmodellen tekst op basis van waarschijnlijkheid. Omdat ze heel veel voorbeelden gezien hebben (miljarden woorden aan teksten op het web), zijn ze heel goed in het produceren van vloeiende taal. Maar vloeiende taal is niet helemaal hetzelfde als menselijke taal. Zo is ChatGPT vaak langdradig, weinig specifiek en gebruikt het vaak literaire, hoogdravende woorden. Samen met een bachelorstudent hebben we laten zien dat het heel eenvoudig is voor een automatische classifier om te onderscheiden welke reacties op nieuwsberichten door ChatGPT geschreven zijn en welke door mensen (Tseng et al., 2023). Maar erger dan dat: dat de tekst vloeiend is, betekent niet dat de inhoud correct is. De term die gebruikt wordt voor het genereren van onjuiste informatie door taalmodellen is **hallucinatie** (Ji et al., 2023). In feite is hallucineren precies wat taalmodellen van nature doen: het woord voor woord genereren van tekst op basis van waarschijnlijkheden.

Het probleem van misinformatie in de uitvoer van generatieve taalmodellen is groter voor **specifieke onderwerpen**, onderwerpen die niet superveel voorkomen op het internet. Als ik ChatGPT vraag om een korte biografie te schrijven van Mark Rutte, dan gaat dit goed, maar als ik een korte biografie vraag van onze rector Hester Bijl, dan staat de tekst vol met misinformatie. ChatGPT verzint een geboortedatum en -plaats, en een studie die niet klopt met de werkelijkheid. Hoe komt dit? Niet omdat de onjuiste informatie online te vinden is, maar omdat de juiste informatie niet vaak genoeg voorkomt en het model dus niet goed de waarschijnlijkheden van de woordcombinaties geleerd heeft. Met andere woorden: ChatGPT genereert plausibele tekst en dat is niet noodzakelijk correcte tekst.

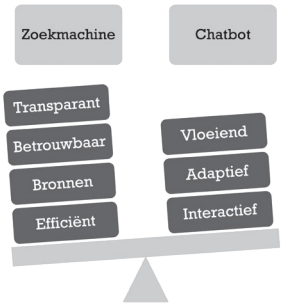
Een ander probleem is dat **bias** die aanwezig is in menselijke taal wordt uitvergroot in grote taalmodellen (Navigli et al., 2023). Bias is van nature aanwezig in mensen: we denken bij de term *hoogleraar* sneller aan een man dan aan een vrouw, en bij de term *dokter* ook. Omdat artsen in teksten vaker man zijn en verpleegkundigen vaker vrouw, zal een taalmodel als het gevraagd wordt een tekst te schrijven over een arts en een verpleegkundige, ook de arts een man laten zijn en de verpleegkundige een vrouw. Is dat erg? Wel als het áltijd zo is in álle teksten. Als mensen ChatGPT gebruiken om suggesties te krijgen voor hun eigen teksten dan verergert dit dus de bias die ze zelf mogelijk al hebben. Dit kan schadelijk zijn, bijvoorbeeld als bedrijven ChatGPT-achtige modellen willen gebruiken voor het analyseren van CV's en vacatures.

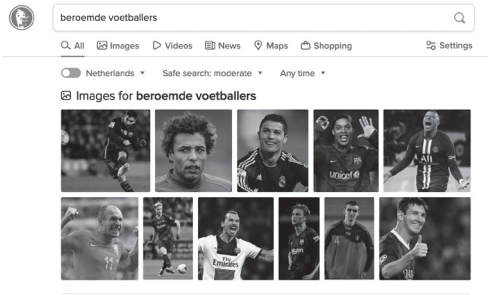
Daarom zijn twee belangrijke pijlers in de toekomstige ontwikkeling van grote taalmodellen: **mensen en bronnen**. Dat brengt me bij:

Deel 2. Zoekmachines

Mensen zijn nodig om de juistheid van informatie te controleren, en hiervoor is de bron van de informatie een noodzakelijke toevoeging. Bronnen maken het mogelijk om in te schatten wat de kwaliteit en de waarde van informatie is. Deze twee aspecten, de mens en de bron, is precies waarom **zoekmachines** zo waardevol zijn: als we een zoekmachine een vraag stellen, zal hij ons een lijst laten zien van op het web gevonden bronnen. De informatie wordt niet 'gehallucineerd' maar opgehaald uit een index. Het is aan de mens om de waarde van de informatie te beoordelen. Met een zoekmachine is er van nature een '*human in the loop*': de mens bepaalt de zoekvraag en selecteert de relevante informatie.

Door het tonen van **bronnen** zorgen zoekmachines voor een vorm van **transparantie**. Snippets, de kleine stukjes van documenten op resultatenpagina's, proberen de gebruiker te tonen waarom het document als relevant voor de zoekvraag beschouwd werd door de termen uit de zoekvraag te markeren. We proberen in ons huidige onderzoek deze vorm van transparantie uit te breiden: kunnen we de gebruiker laten zien hoe de zoekmachine de vraag geïnterpreteerd heeft?





Zoekmachines maken het ook mogelijk om representativiteit en diversiteit expliciet in te bouwen in de resultaten. Zoekvragen zijn vaak kort en ondergespecificeerd. In zoekmachines zoals Google bestaat een zoekvraag gemiddeld uit maar twee woorden. Als iemand de zoekvraag 'leiden' ingeeft in Google, dan kan de zoekmachine niet weten of diegene op zoek was naar informatie over de stad, de universiteit, of het werkwoord 'leiden'; of diegene de website van de gemeente, een Wikipedia-pagina, nieuwsberichten, afbeeldingen, of een kaart wil zien. De oplossing is het diversifiëren van de resultaten (Santos et al., 2015): de zoekmachine toont op de eerste pagina resultaten voor elk van deze opties, in de hoop daarmee zoekers met verschillende intenties achter dezelfde zoekvraag allemaal te helpen.

We kunnen **diversificatie** ook gebruiken om bias tegen te gaan. Als we zoeken naar beroemde voetballers, dan kan het zijn dat een zoekmachine alleen resultaten geeft voor mannelijke voetballers, vanwege het meer voorkomen van mannelijke dan vrouwelijke voetballers op het web. We hebben recentelijk een metriek voorgesteld om te evalueren hoeveel bias er zit in een lijst van resultaten van een zoekmachine (Abolghasemi et al., 2024). Door dergelijke metrieken te gebruiken in de ontwikkeling en optimalisatie van zoekalgoritmes, kan de bias verminderen en kunnen de resultaten diverser worden.

Een andere dimensie van diversiteit is **diversiteit van meningen**. Dit is belangrijk bij het lezen van nieuws. Misschien maakt u wel gebruik een persoonlijke nieuwsfeed. U krijgt dan berichten te zien die aansluiten bij uw interesse en uw eigen perspectief op een onderwerp. Denk aan onderwerpen zoals klimaatacties, boerenprotesten, of vaccinaties. Het zou goed zijn voor het democratische debat als iedereen meerdere perspectieven op een onderwerp te zien kan krijgen. We hebben gewerkt aan methoden om in een krantenartikel te bepalen welke positie in een bepaald debatonderwerp wordt ingenomen (Reuver et al., 2024). We werken samen met politicologen om te bepalen hoe dergelijke methoden een rol zouden kunnen spelen in de algoritmes die de nieuwsfeed samenstellen: in hoeverre willen we mensen stimuleren om andere perspectieven te zien, en in hoeverre is het huidige stimuleren van het 'eigen perspectief' een goed idee (Reuver et al., 2021a)?

Bij het ontwikkelen van dit soort methoden speelt het onderwerp een grote rol. Stel dat je een model hebt getraind dat de verschillende perspectieven op klimaatprotesten kan onderscheiden, dan is het te verwachten dat dit model ook nog wel enigszins de perspectieven op boerenprotesten kan herkennen, maar ditzelfde model krijgt het lastig wanneer het artikelen over vaccinaties te zien krijgt (Reuver et al., 2021b).

Het ontwikkelen van modellen die inzetbaar zijn voor **verschillende domeinen en onderwerpen**, is een van de doelen van ons onderzoek.

We doen dat voor verschillende taken. Een voorbeeld is het identificeren van medicaties en bijwerkingen in ervaringsverhalen van patiënten op sociale media (Dirkson et al., 2022). Een ander voorbeeld is het herkennen van archeologisch relevante termen zoals vindplaatsen en artefacten in archeologische dossiers (Brandsen et al., 2022). In beide gevallen leunen we op het werk van experts om een substantiële hoeveelheid voorbeelden te maken: teksten waarin zij hebben aangegeven waar de medicaties en bijwerkingen, de vindplaatsen en artefacten worden genoemd (Brandsen et al., 2020). Zo maken ze enkele duizenden voorbeelden om een model te finetunen. Dit type model heet **BERT**, waarin de T wederom staat voor *transformer*. BERT-modellen zijn dus transformermodellen, net als GPT, maar in plaats van tekst te genereren, leren ze betekenis te halen uit woorden en teksten.

Een BERT-model dat gefinetuned is op enkele duizenden voorbeelden kunnen we gebruiken om alle relevante termen – bijwerkingen, vindplaatsen of artefacten – in een veel grotere tekstcollectie te ontdekken.

Het wordt uitdagender als de informatie die we uit de tekst halen minder duidelijk gedefinieerd is. Een medicijnnaam zoals *Gleevec* is een duidelijk afgebakend stukje tekst, bestaande uit één of een paar woorden. Een bijwerking is al lastiger te herkennen: iemand kan insomnia omschrijven als slecht slapen, niet in slaap kunnen vallen, of wakker liggen. Nog een stapje verder is het kunnen identificeren wat de *copingstrategieën* zijn die mensen gebruiken om met bijwerkingen om te gaan. Nemen ze een warm bad tegen de spierpijn, drinken ze gemberthee tegen de maagproblemen, nemen ze hun medicatie in kleine doseringen? Dit is relevant om te weten, want artsen willen voorkomen dat mensen strategieën toepassen die hun medicatie verstoren. En onschuldige copingstrategieën kunnen weer van nut zijn voor andere patiënten. Deze copingstrategieën blijken nog lastiger te herkennen in patiëntenervaringen dan bijwerkingen, maar het bleek mogelijk om (weliswaar met een lage precisie) een aantal veel voorkomende copingstrategieën te vinden voor patiënten die gleevec gebruiken en daar bijwerkingen van ondervinden (Dirkson et al., 2023).

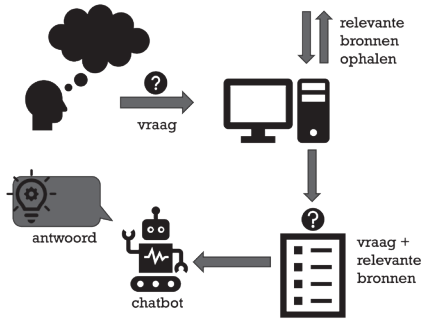
Een vergelijkbare uitdaging is het analyseren van narratieven die mensen gebruiken om te argumenteren waarom ze wel of niet deelnemen aan het bevolkingsonderzoek kanker. De opkomstcijfers zijn relatief laag en uit vragenlijstonderzoek kent het RIVM een deel van de redenen die mensen hebben om niet te gaan. We proberen samen met communicatiewetenschappers te ontdekken welke argumenten in nieuwsmedia en sociale media gegeven worden om wel of niet deel te nemen. Een eerste resultaat is dat nieuwsmedia over het algemeen kankerscreening beschrijven als vervelend en stressvol, maar wel belangrijk. Momenteel vervolgen we dit onderzoek met een analyse van sociale media.

Uitdagingen met specifieke onderwerpen en domeinen spelen ook een rol in *zoekmachines*. Als u informatie zoekt voor uw werk dan doet u dat op een andere manier dan wanneer u even wilt checken hoe laat deze lezing begint: u bekijkt meer resultaten, formuleert meer zoekvragen, en uw collega heeft andere vragen dan u. Hoe specifieker het onderwerp, hoe meer zoekopdrachten uniek zijn (Wiggers et al., 2023). Een zoekmachine zoals *Google Scholar* of *Legal Intelligence* moet daarom andere criteria gebruiken om documenten te sorteren dan reguliere zoekmachines (Wiggers et al., 2022). De **betekenis van zoekvragen** kan verschillend zijn voor zoekers met verschillende soorten expertise.

Zoals ik al eerder zei, komt hallucinatie in grote taalmodellen meer voor bij specifieke dan bij algemene onderwerpen. Als de gebruiker geen expert is op het onderwerp, is dit problematisch omdat diegene niet kan inschatten of de gegenereerde tekst wel correct is. En dat brengt me bij

Deel 3. Hoe zoekmachines grote taalmodellen kunnen helpen, en andersom

Ik had het al over het belang van bronnen. We werken momenteel aan methoden waarbij een zoekmachine 'aan de achterkant' van het taalmodel eerst relevante documenten ophaalt, en het taalmodel daarna op basis van die documenten het antwoord formuleert. Deze techniek heet **Retrieval Augmented Generation (RAG)** (Lewis et al., 2020).



Dit is bruikbaar in veel situaties. Zo kost het veel tijd en energie om een groot taalmodel te trainen en zal een taalmodel dus nooit toegang hebben tot de meest recente informatie – denk aan nieuwsberichten, standpunten van politieke partijen, of de meest recent wetenschappelijke artikelen. Met RAG kan het systeem eerst relevante recente informatie ophalen en het antwoord genereren op basis van die informatie. Een andere setting waar RAG nuttig is, is voor chatbots die vragen beantwoorden over de eigen informatie van een bedrijf of organisatie. Zo ben ik samen met een masterstudent bezig om een chatbot te ontwikkelen die helpdeskmedewerkers van de ICT-dienst van onze universiteit kan helpen om het juiste antwoord te formuleren op binnenkomende vragen, op basis van de grote database van eerder gegeven antwoorden. In het LESSEN project werken we samen met Nederlandse bedrijven aan chatbots die grote taalmodellen gebruiken voor vloeiende interactie met de klant, maar wel betrouwbaar moeten zijn en dus niet zomaar wat verzinnen over producten of diensten (Verberne, 2023). Daarvoor blijft de technologie van zoekmachines essentieel.

Andersom kunnen grote taalmodellen ook zoekmachines van dienst zijn. Daarvoor moet ik eerst uitleggen hoe zoekmachines werken. De kernfunctionaliteit van zoekmachines is om bronnen op te halen die relevant zijn voor de ingegeven zoekvraag. Dit moet heel snel gebeuren, want als de zoeker meer dan een seconde moet wachten op de resultaten dan haakt hij af. Uitgaande van een collectie van miljoenen documenten (of miljarden, als we het over het hele web hebben), moeten we een heel snelle functie hebben voor het vinden van relevante documenten. Traditioneel zijn deze functies gebaseerd op **woorden die voorkomen in zoekvragen en documenten**: de zoekmachine haalt de documenten op uit de index die de woorden van de zoekvraag bevatten (Sparck Jones, 1972). Als ik in Google zoek naar 'leiden' dan verwacht ik websites te krijgen die het woord 'leiden' bevatten. Dit is vaak belangrijk en nuttig, maar er zijn ook veel situaties waarin we niet de letterlijke woorden uit de vraag maar ook documenten met vergelijkbare betekenis willen vinden. Als ik bijvoorbeeld zoek naar 'reparatie fiets' dan wil ik websites van *fietsenmakers* vinden.

Om dit te bereiken zijn BERT-relevantiemodellen ontwikkeld die niet naar het voorkomen van de letterlijke woorden kijken maar naar de betekenis van woorden. In die modellen liggen bijvoorbeeld *reparatie* en *maker* dicht bij elkaar en daardoor kunnen meer relevante documenten gevonden worden. Als we die modellen willen leren om een sortering van documenten in een zoekmachine te maken, dan hebben we voorbeelden nodig. Voorbeelden voor zoekmachines bestaan uit zoekvragen en documenten die relevant zijn voor die zoekvragen. Het getrainde **relevantiemodel** moet vervolgens inzetbaar zijn voor nieuwe zoekvragen die niet voorkomen in de trainingsdata; en om een model die generalisatie naar nieuwe zoekvragen te leren is veel trainingsdata nodig, duizenden voorbeelden (Lin et al., 2022, fig. 19). Zonder al deze trainingsdata is het meestal

niet mogelijk voor een BERT-model om relevantere resultaten te vinden dan een traditioneel model dat kijkt naar het voorkomen van woorden.

Eén van de doelen van mijn onderzoeksgroep is om zoekmachines te ontwikkelen voor specifieke toepassingen. Denk aan juridische (Askari et al., 2023a), archeologische (Brandsen et al., 2019) of medische websites; ICT-helpdesks of webwinkels. Voor deze toepassingen hebben we vaak helemaal niet de beschikking over duizenden voorbeelden van zoekvragen met relevante documenten, en zeker niet voor andere talen dan het Engels.

Wat we daarom doen, is grote taalmodellen gebruiken om **data te genereren** waarmee we zoekmachines kunnen trainen. We doen dat op verschillende manieren: als er voor een taak wel documenten beschikbaar zijn, maar geen zoekvragen, dan gebruiken we een taalmodel om voor elk document mogelijk relevante vragen te genereren die met het document beantwoord kunnen worden. Andersom, als we wel zoekvragen hebben, maar geen relevante documenten, dan laten we een groot taalmodel de relevante antwoorden genereren (Askari et al., 2023b, 2023c). De paren van vragen en antwoorden gebruiken we vervolgens om een relevantiemodel te trainen.

We volgen dezelfde strategie voor het trainen van **chatbots**. Denk bijvoorbeeld aan een groot verzekeringsbedrijf. Zij hebben een chatbot op hun website die klanten kan helpen vragen te beantwoorden. De eerste stap is identificeren wat het onderwerp van de vraag is: gaat het over reisverzekeringen, of over autoverzekeringen? De bot probeert de zoekvraag van de gebruiker in te delen naar onderwerp. Maar niet voor elk van deze onderwerpen heeft het verzekeringsbedrijf voorbeelddata beschikbaar. We hebben een groot taalmodel ingezet om trainingsdata te genereren voor elk mogelijk onderwerp, en met deze data kunnen we een goed classificatiemodel trainen – bijna net zo goed als een model dat is getraind op echte, door mensen gemaakte data.

Misschien fronst u nu en denkt: “ja maar die taalmodellen hadden toch problemen met hallucinatie? Wat als er misinformatie gebruikt wordt om zoekmachines en chatbots te trainen?” Dat klopt, dat is ook wel een zorg en een onderwerp van ons onderzoek. Wat mij betreft is het risico op hallucinatie juist een reden om voor dit indirecte gebruik van generatieve taalmodellen te kiezen: als datagenerator in plaats van het direct gebruik van ChatGPT om vragen te beantwoorden. Een zoekmachine zal immers nog altijd alleen informatie ophalen die beschikbaar is in de index, en dus niet hallucineren. Bovendien kunnen we als we trainingsdata genereren met een taalmodel, kunstmatig een balans aanbrengen in de data die niet in natuurlijke, menselijke data zit, en daarmee de bias in het model verminderen (Abolghasemi et al., 2023). Dat neemt niet weg dat ik wel zorgen heb over het steeds meer voorkomen van teksten op het web die door grote taalmodellen gegenereerd zijn.

Nu komen we bij de vraag:

Deel 4. Is de zoekmachine van de toekomst een chatbot?

Ik heb verteld dat zoekmachines taalmodellen kunnen helpen door bronnen aan te leveren, en dat andersom generatieve taalmodellen zoekmachines kunnen helpen door data te genereren. Waar gaat dit naartoe? **Is de zoekmachine van de toekomst een chatbot?** Ja, ik denk van wel, op verschillende manieren. De zoekmachine van de toekomst zal, ondersteund met de juiste bronnen uit de index, steeds vaker een antwoord genereren in natuurlijke taal.

Daarnaast geloof ik dat generatieve modellen de toekomst hebben voor veel meer soorten taken dan het schrijven van teksten en het geven van antwoorden. Zoals ik eerder heb beschreven, zijn we gewend aan het leren van modellen via duizenden voorbeelden van wat we willen dat het model doet. Dit werkt goed voor problemen waar we veel voorbeelden hebben, maar in de praktijk hebben we die niet altijd. Generatieve grote taalmodellen zijn flexibeler en kunnen leren van veel minder

voorbeelden. We zijn aan het exploreren in hoeverre deze generatieve modellen ook classificatie- en extractietaken kunnen doen, zoals het herkennen van bijwerkingen en copingstrategieën. Voor sommige taken lijkt dit goed te werken, zoals het herkennen welke emoties een stukje tekst bevat (Broekens et al., 2023).

Maar hoe werkt dat dan voor zoekmachines? Het principe van een zoekmachine blijft om op basis van een zoekvraag relevante informatie – documenten – op te halen uit een index. We werken momenteel aan een generatieve manier om dit te doen: als elk document in de index een beschrijvende identificatiecode heeft, kunnen we die dan **genereren** voor een inkomende zoekvraag (Sun et al., 2024)? Deze benadering brengt de potentie van grote taalmodellen (leren uit weinig voorbeelden) samen met de kracht van zoekmachines (het ophalen van bronnen). Ook wil ik onderzoeken of we grote taalmodellen meer geschikt kunnen maken voor informatie zoeken door tijdens het verwerken van grote hoeveelheden tekst in de pre-training meteen ook de bronnen op te slaan waar elk woord uit komt.

Een belangrijk probleem met de huidige generatie grote taalmodellen is dat ze heel groot zijn en veel energie gebruiken. Zo is berekend dat één run in ChatGPT 1000 keer meer energie kost dan één zoekvraag in Google.[1] Ook is heel veel data nodig om een model te pretrainen en dat kan problemen geven rondom auteursrecht en kwaliteit van bronnen. Daarom werken we aan manieren om taalmodellen te trainen op veel minder data. We willen onderzoeken van welke soorten teksten taalmodellen sneller woordbetekenissen kunnen leren, van verhalen bijvoorbeeld of van kindertaal (Van Dijk et al., 2023). Zo hopen we modellen te trainen die **kleiner zijn en daardoor minder energie gebruiken**. Dat heeft als bijkomend voordeel dat de modellen toegankelijker worden en door meer mensen en bedrijven gebruikt kunnen worden. Mensen kunnen dan meer en meer werken met lokaal opgeslagen modellen en ook hun data privé houden.

Zeker voor onderzoek is het belangrijk dat we weten op welke data een model getraind is en met welke instellingen. Voor de betrouwbaarheid van onze resultaten moeten we onderzoek kunnen reproduceren. We willen inzicht krijgen in de bronnen van bias en hallucinatie, en we willen zorgen dat de output van modellen eerlijker en diverser is. Zoals ik al zei werken we daarom met open-source modellen, ook al zijn ze niet voor alle doeleinden even krachtig als de beste commerciële modellen. Het is belangrijk dat voor meer talen goede open-source taalmodellen getraind worden. Er staan tienduizenden generatieve taalmodellen op HuggingFace die Engels kunnen, tegenover slechts enkele honderden modellen voor Nederlands. En dat terwijl Nederlands een taal is waarvoor veel data beschikbaar is, vergeleken met veel andere talen. Dit jaar werken SURF en TNO aan de ontwikkeling van GPT-NL, een groot taalmodel voor het Nederlands dat volledig open en met aandacht voor auteursrecht getraind is. Vanaf volgend jaar kunnen we casussen – zoals rond bias, hallucinatie, en emotie – gaan onderzoeken met GPT-NL.

De zoekmachine van de toekomst zal zeker kenmerken van een chatbot hebben, en gebouwd zijn op grote taalmodellen, maar tegelijkertijd zal de chatbot van de toekomst ondersteund zijn door zoektechnologie. Een belangrijke rol is hierin weggelegd voor de mens: ondanks dat kunstmatige intelligentie steeds meer voor ons kan doen, zullen het stellen van de juiste vraag en het inschatten van de kwaliteit van het antwoord belangrijker zijn dan ooit tevoren. ***The future of AI is human***.

Aan het einde van mijn oratie wil ik graag nog wat zeggen over **impact**. Als wetenschappers wordt ons vaak gevraagd uit te leggen wat de wetenschappelijke en maatschappelijke impact is van ons onderzoek. Ik zou willen beargumenteren dat de grootste impact die we hebben de impact op de mensen om ons heen is, en dan in het bijzonder op de studenten en jonge

1. https://ai.stackexchange.com/questions/38970/how-much-energy-consumption-is-involved-in-chat-gpt-responses-being-generated

onderzoekers die we begeleiden. In de zeven jaar dat ik bij LIACS werk heb ik tientallen scriptiestudenten begeleid en onderwijs gegeven aan vele honderden studenten. Als voorzitter van de examencommissie heb ik met nog eens honderen studentenverzoeken per jaar te maken. In mijn 1-op-1 begeleiding probeer ik altijd oog te hebben voor de individuele talenten en uitdagingen van studenten en promovendi. Het belangrijkste is om inlevingsvermogen te blijven hebben, en vriendelijk en duidelijk te communiceren. Veel studenten die ik begeleid komen van buiten Europa. Na hun master vinden ze vaak werk in Nederland en de economie drijft op hun kennis en ervaring. Ik geloof in het principe van 'pay it forward'; ik hoop dat de studenten die ik begeleid op hun beurt uitgroeien tot waardevolle collega's en mentoren. Het is niet het soort impact waar ons werk op beoordeeld wordt, maar het is wel de impact die we kunnen hebben op de samenleving. Daarom is mijn oproep aan mijn collega's en iedereen die zijn eerste stappen zet in het geven van onderwijs om altijd oog te blijven hebben voor de menselijke kant van ons werk.

Is the search engine of the future a chatbot?

Mevrouw de rector magnificus, geacht faculteitsbestuur, zeer gewaardeerde toehoorders. Dear colleagues.

In this lecture, I want to talk about search engines and chatbots. There was a time when these were two completely separate applications, but nowadays they overlap and are becoming more alike. Chatbots based on large language models, such as ChatGPT, are used by many people to get answers to questions, and companies behind major search engines such as Google Microsoft are incorporating chatbot technology into their interfaces. Is the search engine of the future a chatbot? Today, I will answer that question. The short answer is yes. But the reverse is also true: the chatbot of the future is a search engine.

This lecture consists of four parts: The first part is about large language models. I will start by explaining what language models are and how the technology behind ChatGPT works. Then, based on recent research, I will explain what these large language models can do for us, discuss the problems and challenges they present, and explore potential solutions. The second part is about search engines. My team conducts extensive research on developing methods to help people find specific information, such as answers to medical and legal questions. I will explain our approach and the challenges we face. The third part of my lecture focuses on how search engines can assist large language models and vice versa. In the fourth and final part, I will provide a glimpse into the future: What is the current focus of our research, and what are some important directions for the next steps?

Part 1. Large Language Models

Since the release of ChatGPT in the fall of 2022, everyone has had the opportunity to experience the power of large language models. For the remainder of my lecture, it is important to understand how these models work. ChatGPT is a chatbot based on the GPT language model. The G in GPT stands for *generative*; GPT is a **generative language model**. This means it generates text by determining the most likely next word after each word. Such language models have existed for decades (Van den Bosch, 2008; van den Bosch, 2012; Verberne, 2002). If I read you the sentence

> "The train could not continue because there was a tree on the...",

you know that the most likely next word is 'track'. By processing large amounts of language input, we have learnt which word is most likely to come next.

Language models like ChatGPT use these word probabilities to generate fluent text. But there is more to it. The P in GPT stands for *pre-trained* (Radford et al., 2018). This refers to an important feature of these models: they are pre-trained on a vast amount of text. By processing large amounts of text, the model learns which words often appear together and which do not. It turns out that words that often appear together are more semantically related than words that do not. This is an old principle in linguistics, the **distributional hypothesis** (Harris, 1954). This hypothesis dates back to 1954 and still forms the basis for all modern language models. Let me give you another example. I assume you have never heard of the word *gleevec*. But if I provide the following sentence, you can probably guess the meaning:

> "I have been prescribed gleevec, but it's causing dry skin."

Can you guess the meaning? Gleevec is a medication. Because we have encountered a lot of language in our lives, we can infer the meanings of words from context. This is also what language models do: by learning to predict which words are likely at which positions in the text, the model learns the meanings of words.

Now let's go back to the abbreviation GPT. I have already explained that the G stands for generative and the P for pre-trained. The T stands for Transformer. The transformer is a type of computer model, developed in 2017 (Vaswani et al., 2017), that is very good at processing text and learning the meanings of words. This is because the model calculates relationships between all words in the text, even when the sentences are long. A language model created with a transformer has therefore learned that there is a relationship between trains and tracks. That enabled it, like humans, to very accurately guess the next word in the sentence "The train could not go further because there was a tree on the ...".

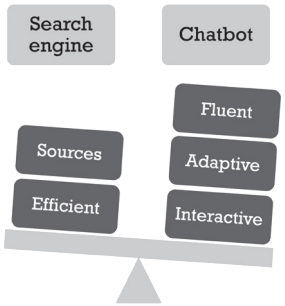
Predicting the next word by itself isn't a very useful application, but after a transformer model has been trained to predict words, it can also learn other tasks related to word meaning.

To teach a pre-trained language model a specific task, it needs to be shown examples: hundreds to thousands of examples of what we want the model to do. This is called **fine-tuning** (Devlin et al., 2019). Through fine-tuning, a model can, for example, be trained to classify texts by subject or to find specific types of words in a text, such as medication names. With my students and PhD candidates, I have trained models for all kinds of tasks in this way. GPT can also be fine-tuned. For example, we made a GPT model more empathetic by fine-tuning it on empathetic dialogues (Chen, 2023). Unfortunately, this caused the model to lose some knowledge of the world and it could no longer answer questions like "which is bigger, the moon or the sun". But it was friendlier indeed.

Between 2018 and 2022, language models became larger: they had more parameters and were trained on more data. As a result, they also became more powerful, and the output became more fluent. By fine-tuning a large language model on conversations, it can engage in dialogues. This way, a large language model becomes a **chatbot**. When ChatGPT became accessible through an online interface, the media and the rest of the world discovered the power of large language models.

While ChatGPT is accessible via a website, the underlying model's training details and data sources are not publicly available. For research purposes, we prefer non-commercial models that are **open** and can be downloaded to our own computers. Through the website HuggingFace, transformer models are available for various languages and tasks, including many open alternatives to ChatGPT.

These large language models are so powerful that fine-tuning is not even necessary for many tasks; they can follow textual instructions (Brown et al., 2020). For example, we discovered that a large language model can produce good summaries of scientific articles with clear instructions and only two examples. It even was slightly better than an earlier transformer model fine-tuned on 5000 examples (Zakkas et al., 2024). Because GPT can generate **text fluently**, summarizing is a core task it excels at. This makes the model seem intelligent, but is it really? Personally, I believe that what language models do is **simulate** human language. You could say it is like a parrot with some random variation (Bender et al., 2021).



As I have explained, language models generate text based on probability. Because they have seen a lot of examples (billions of words of text on the web), they are very good at producing fluent language. But fluent language is not exactly the same as

human language. For instance, ChatGPT is often verbose, not very specific, and often uses literary, pompose words. Together with a bachelor student, we showed that it is very easy for an automatic classifier to distinguish which comments on news articles were written by ChatGPT and which by humans (Tseng et al., 2023). But worse than that: the fact that the text is fluent does not mean that the content is correct. The term used for language models generating incorrect information is **hallucination** (Ji et al., 2023). In fact, hallucinating is exactly what language models do naturally: generating text word by word based on probabilities.

The problem of misinformation in the output of generative language models is bigger for **specific topics**, topics that aren't very widespread on the internet. If I ask ChatGPT to write a short biography of Mark Rutte, it does so well, but if I ask for a short biography of our rector Hester Bijl, the text is full of misinformation. ChatGPT invents a date and place of birth, and an education curriculum that doesn't match reality. How does this happen? Not because the incorrect information is available online, but because the correct information isn't frequent enough, and the model did not have sufficient information to learn the correct probabilities of the word combinations. In other words: ChatGPT generates plausible text, which isn't necessarily correct.

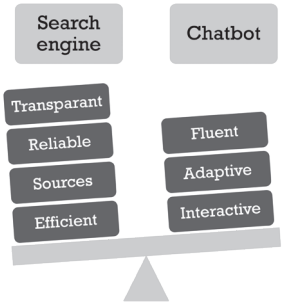
Another problem is that biases from human language are amplified in large language models (Navigli et al., 2023). Bias is naturally present in humans: we more easily associate the term *professor* with a man than with a woman, and the term *doctor* as well. Because doctors are more often male in texts and nurses more often female, a language model, when asked to write a text about a doctor and a nurse, will also make the doctor male and the nurse female. Is that bad? Yes, if it is *always* like that in *all* texts. If people use ChatGPT to get suggestions for their own texts, this magnifies the biases they may already have. This can be harmful, for example, if companies want to use ChatGPT-like models for analysing resumés and job postings.

Therefore, two important pillars in the future development of large language models are: **humans and sources**. That brings me to:

Part 2. Search Engines

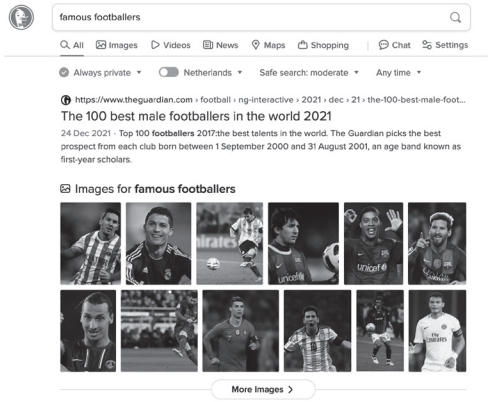
Humans are needed to verify the accuracy of information, and for this purpose, the source of the information is necessary. Sources enable us to assess the quality and the value of information. These two aspects, the human and the source, are precisely why **search engines** are so valuable: when we ask a search engine a question, it will show us a list of sources found on the web. The information is not 'hallucinated' but retrieved from an index. It is up to the human to assess the value of the information. With a search engine there is inherently a '*human in the loop*': the human determines the search query and selects the relevant information.

By showing **sources**, search engines provide a form of **transparency**. Snippets, the small pieces of documents on results pages, try to show the user why the document was considered relevant to the search query by boldfacing query terms. In our current research, we are trying to expand this form of transparency: can we show the user how the search engine interpreted the question?



Search engines also enable us to explicitly incorporate representativeness and diversity into the results. Search queries are often short and underspecified. In search engines like Google, an average query consists of only two words. If someone enters the search query 'leiden' into Google, the search engine cannot know whether the person was looking for information about the city, the university, or the verb 'leiden'; or whether they want to see the municipality's website, a Wikipedia page, news articles, images, or a map. The solution is to diversify the results (Santos et al., 2015); the search engine displays results for each of these options on the first page, hoping to assist searchers with different intentions behind the same search query.

We can also apply **diversification** to combat bias. When searching for famous footballers, a search engine may only provide results for male football players, due to the prevalence of male rather than female footballers on the web. We recently proposed a metric to evaluate how much bias exists in a list of search engine results (Abolghasemi et al., 2024). By using such metrics in the development and optimization of search algorithms, bias can be reduced, and the results can become more diverse.



Another dimension of diversity is **diversity of opinions**. This is important when reading the news. Perhaps you use a personalized news feed. You then see messages that align with your interests and your own perspective on a topic. Think of topics such as climate protests, farmers' protests, or vaccinations. It would be beneficial for the democratic debate if all news readers have access to different perspectives on a topic. We have been working on methods to determine which position is taken on a particular debate topic in a newspaper article (Reuver et al., 2024). We collaborate with political scientists to determine how such methods could play a role in the algorithms that compile the news feed: to what extent do we want to encourage people to see other perspectives, and to what extent is the current encouragement of the 'own perspective' a good idea (Reuver et al., 2021a)?
When developing these types of methods, the topic plays a significant role. Suppose you have trained a model that can distinguish between different perspectives on climate protests, then it is expected that this model can still somewhat recognize the perspectives on farmers' protests, but this same model would struggle when presented with articles about vaccinations (Reuver et al., 2021b).

Developing models that can be deployed across **different domains and topics** is one of the goals of our research.

We do this for various tasks. One example is to identify medications and side effects in patient experiences on online media (Dirkson et al., 2022). Another example is recognizing archaeologically relevant terms such as locations and artifacts in archaeological records (Brandsen et al., 2022). In both cases, we rely on the work of experts to create a substantial number of examples: texts in which they indicated where the medications and side effects, the locations and artifacts are mentioned (Brandsen et al., 2020). This way, they create several thousands of examples to fine-tune a model. This type of model is called **BERT**, where the T stands again for transformer. BERT models are transformer models, just like GPT, but instead of generating text, they learn to derive meaning from words and texts.

A BERT model fine-tuned on several thousands of examples can be used to discover all relevant terms – side effects, sites, or artifacts – in a much larger text collection.

It becomes more challenging when the information we extract from the text is less clearly defined. A drug name like *Gleevec* is a clearly defined piece of text, consisting of one or a few words. A side effect is already more difficult to recognize: someone may describe insomnia as poor sleep, inability to fall asleep, or sleepless nights. One step further is to identify the coping strategies that people use to deal with side effects. Do they take a warm bath for muscle pain, drink ginger tea for stomach problems, or take their medication in small doses? This is relevant to know because doctors want to prevent people from applying strategies that disrupt their medication. And innocent coping strategies can be useful for other patients. These coping strategies are even more difficult to recognize in patient experiences than side effects, but we made it possible (albeit with low precision) to identify some common coping strategies for patients using Gleevec who experience side effects (Dirkson et al., 2023).

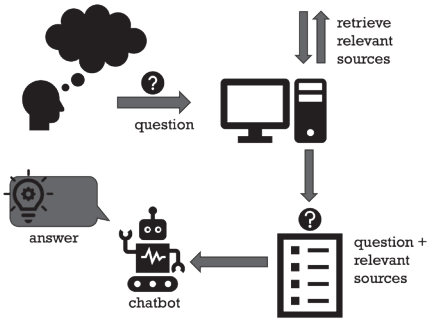
A similar challenge is analyzing the narratives that people use to argue why they do or do not participate in cancer screening. The turnout rates are relatively low, and from questionnaire research, the National Institute for Public Health and the Environment is aware of some of the reasons why people are not participating. We are trying together with communication scientists to discover which arguments are given in news media and social media to participate in screening or not. A first result is that news media in general describe cancer screening as inconvenient and stressful, yet important. Currently, we are continuing this research with an analysis of social media.

Challenges with specific topics and domains also play a role in search engines. When you are searching for information for your work, you do it differently than when you are looking up the time of this lecture: you look at more results, formulate more search queries, and your colleague has different questions than you do. The more specific the topic, the more unique the search queries are (Wiggers et al., 2023). Therefore, a search engine like Google Scholar or Legal Intelligence must use different criteria to sort documents than regular search engines (Wiggers et al., 2022). The **meaning of search queries** can vary for searchers with different types of expertise.

As I mentioned earlier, hallucination in large language models is more common in specific than in general topics. If the user is not an expert in the field, this is problematic because they cannot assess whether the generated text is correct. And that brings me to

Part 3. How search engines can help large language models, and vice versa

I already mentioned the importance of sources. We are currently working on methods where a search engine in the back end of the language model first retrieves relevant documents, after which the language model formulates the answer based on those documents. This technique is called **Retrieval Augmented Generation (RAG)** (Lewis et al., 2020).



This is useful in many situations. It takes a lot of time and energy to train a large language model, so a language model will never have access to the most recent information – think of news articles, political party positions, or the most recent scientific articles. With RAG, the system can first retrieve

relevant recent information and generate the answer based on that information. Another setting where RAG is useful is for chatbots that answer questions about a company or an organization's own information. Together with a master student, I am working on a chatbot that can assist helpdesk staff from our own university's IT service in formulating the correct answer to incoming questions, based on the large database of previously given answers. In the LESSEN project, we are collaborating with Dutch companies on chatbots that use large language models for smooth interaction with the customer, but that at the same time are reliable and not hallucinate information about products or services (Verberne, 2023). For this, the technology of search engines remains essential.

On the other hand, large language models can also be of service to search engines. To explain this, I first need to explain how search engines work. The core functionality of search engines is to retrieve sources that are relevant to the entered search query. This needs to happen very quickly because if the searcher has to wait more than a second for the results, they might abandon their search effort. Considering a collection of millions of documents (or billions, if we are talking about the entire web), we need a very fast function to find relevant documents. Traditionally, these functions are based on **words that occur in search queries and documents**: the search engine retrieves documents from the index that contain the words from the search query (Sparck Jones, 1972). If I search for 'leiden' on Google, I expect to get websites that contain the word 'leiden'. This is often important and useful, but there are also many situations where we want to find not just the literal words from the query but also documents with similar meaning. For example, if I search for 'bike repair', I want to find websites of *bicycle shops*.

To achieve this, BERT relevance models have been developed that do not look at the occurrence of the words themselves but at the meaning of words. In those models, for example, 'bike' and 'bicycle' are close to each other and therefore more relevant documents can be found. If we want to teach those models to make a relevance ranking of documents in a search engine, we need examples. Examples for search engines consist of search queries and documents that are relevant to those search queries. This trained **ranking model** must be applicable to new search queries that are not present in the training data; to teach a model the generalization to different queries, a lot of training data is needed, thousands of examples (Lin et al., 2022, fig. 19). Without all these training data, it is usually not possible for a BERT model to find more relevant results than a traditional model that looks at the occurrence of words.

One of the goals of my research group is to develop search engines for specific applications. Think of legal (Askari et al., 2023a), archaeological (Brandsen et al., 2019), or medical websites, IT helpdesks or webshops. For these applications, we often do not have access to thousands of examples of search queries with relevant documents, especially not for languages other than English.

Therefore, we use large language models to **generate data** with which we can train search engines. We do this in different ways: if there are documents available for a task but no search queries, we use a language model to generate potentially relevant questions for each document that can be answered with the document. Conversely, if we have search queries but no relevant documents, we let a large language model generate the relevant answers (Askari et al., 2023b, 2023c). We then use the pairs of search queries and answers to train a ranking model.

We follow the same strategy for training **chatbots**. Consider, for example, a large insurance company. They have a chatbot on their website that can help customers answer questions. The first step is to identify the topic of the question: is it about travel insurance or car insurance? The bot tries to categorize the user's search query by topic. However, the insurance company does not have example data available for each one

of these topics. We have deployed a large language model to generate training data for every possible topic, and with this data, we can train a good classification model – almost as good as a model trained on real, human-made data.

You might be frowning now, thinking, "But didn't those language models have issues with hallucination? What if misinformation is used to train search engines and chatbots?" That is a valid concern and indeed a topic of our research. In my view, the risk of hallucination is precisely a reason to opt for this indirect use of generative language models: as data generators instead of direct use of ChatGPT to answer questions. After all, a search engine will still only retrieve information available in its index, and not hallucinate. Moreover, when we generate training data with a language model, we can artificially introduce a balance in the training data that is not present in natural, human-generated data, thereby reducing bias in the model (Abolghasemi et al., 2023). However, that doesn't take away my concerns about the increasing prevalence of texts on the web generated by large language models.

Now we come to the question:

Part 4. Is the search engine of the future a chatbot?

I have mentioned how search engines can assist language models by providing sources, and conversely, generative language models can help search engines by generating data. Where is this headed? **Is the search engine of the future a chatbot?** Yes, I believe so, in several respects. The search engine of the future will increasingly generate responses in natural language, supported by the right sources from the index.

Additionally, I believe that generative models have a future in many more types of tasks than writing texts and providing answers. As I described earlier, we are accustomed to training models using thousands of examples of what we want the model to do. This works well for problems where we have many examples, but in practice, we do not always have them. Generative large language models are more flexible and can learn from far fewer and less well-defined examples. We are exploring to what extent these generative models can also perform classification and extraction tasks, such as to identify side effects and coping strategies. For some tasks, this seems to work well, such as recognizing which emotions a piece of text contains (Broekens et al., 2023).

But how does this work for search engines? The task of a search engine remains to retrieve relevant information – documents – from an index based on a search query. We are currently working on a generative way to do this: if each document in the index has a descriptive identification code, can we **generate** those for an incoming search query (Sun et al., 2024)? This approach combines the potential of large language models (learning from few examples) with the power of search engines (retrieving sources). I also want to investigate whether we can make large language models more suitable for information retrieval by storing the sources of each word while processing large amounts of text during pre-training.

A significant issue with the current generation of large language models is their size and energy consumption. It has been calculated that one run in ChatGPT consumes 1000 times more energy than one Google search query.[i] Additionally, a vast amount of data is required to pretrain a model, which can raise concerns regarding copyright and the quality of sources. That's why we are exploring ways to train language models on much less data. We want to investigate from which types of texts language models can learn word meanings more quickly, such as stories or children's language (Van Dijk et al., 2023). By doing so, we hope to train models that are **smaller and therefore consume less energy**. This has the added benefit of making the models more accessible to deploy, allowing them to be used by

i https://ai.stackexchange.com/questions/38970/how-much-energy-consumption-is-involved-in-chat-gpt-responses-being-generated

more people and businesses. People can then increasingly work with locally stored models and keep their data private.

Especially for research, it is important to know on which data a model is trained and with which settings. For the reliability of our results, we must be able to reproduce research. We want to gain insight into the sources of bias and hallucination, and ensure that the output of models is fairer and more diverse. As I mentioned, we therefore work with open-source models, even though they may not be as powerful as the best commercial models for all purposes. It is important that good open-source language models are trained for more languages. There are tens of thousands of generative language models on HuggingFace that can handle English, compared to only a few hundred models for Dutch. And yet Dutch is a language for which much data is available compared to many other languages. This year SURF and TNO are working on the development of GPT-NL, a large language model for Dutch that is fully open and trained with attention to copyright. From next year onwards, we can start investigating cases such as bias, hallucination, and emotion with GPT-NL.

The search engine of the future will certainly have characteristics of a chatbot and will be built on large language models. However, the chatbot of the future will also be supported by search technology. An important role in this is reserved for humans: despite artificial intelligence being able to do more and more for us, asking the right question and assessing the quality of the answer will be more important than ever before. ***The future of AI is human.***

At the end of my inaugural lecture, I would like to say something about **impact**. As scientists, we are often asked to explain the scientific and societal impact of our research. I would like to argue that the greatest impact we have is on the people around us, particularly the students and young researchers we mentor. In the seven years that I have been at LIACS, I have supervised dozens of thesis students and taught hundreds of students in courses. Additionally, as chair of the board of examiners, I deal with hundreds of student requests per year. In my one-to-one supervision, I always aim to recognize the individual talents and challenges of students and PhD candidates. The most important thing is to have empathy and communicate kindly and clearly. Many of the students I mentor come from outside Europe. After completing their master's degrees, they often find a job in the Netherlands, contributing their knowledge and experience to the economy. I believe in the principle of 'paying it forward'; I hope that the students I mentor will in turn become valuable colleagues and mentors. It may not be the type of impact by which our work is judged, but it is the impact we have on society. Therefore, I ask my colleagues and everyone starting their journey in education to always remain mindful of the human aspect of our work.

Dankwoord/Acknowledgements

Aan het einde van mijn oratie wil ik graag allen bedanken die aan de totstandkoming van mijn benoeming hebben bijgedragen. Als eerste wil ik het College van Bestuur en het faculteitsbestuur bedanken voor het in mij gestelde vertrouwen. The LIACS management team I would like to thank for giving me the opportunity to start and grow a research group in 2017, the continuous trust and support, and making this professor position possible.

I would like to thank all my project collaborators and co-supervisors; locally, nationally, and internationally, for providing different viewpoints, asking different questions than me, and challenging my ideas.

My colleagues in the Board of Examiners I would like to thank for their pleasant collaboration, and for continuously making sure that we do not forget the *human* side of our work. Aletta and my intervision group mates, I thank you for everything I have learned about communication and leadership.

My direct co-workers at LIACS I would like to thank for their team spirit and friendship. Academic life at LIACS is never about competition but always in collaboration. Zhaochun, Gijs, Tessa, thank you for joining my group, broadening our scope and way of working, and supporting each other. Frank and Matthijs, thank you for walking the academic road together with me and always having your slack open for me.

All students and colleagues whose work I have mentioned in this lecture I would like to thank for their collaboration, inspiration, and making my work a pleasure. Alex, Anne, Gineke, Arian, Amin, Myrthe, I-Fan, Mert, Yumeng, LanGe, Andreas, Pavlos, Peter, Max, Bram, and Joost: Having discussions with you and co-authoring your papers is the best part of my academic life.

Some colleagues have helped me writing this lecture. Jan, Marco, Mischa, thank you for being so reflective and helpful. Nava, thank you for teaching me to express the importance of kindness.

There are some people who have mentored me in the past and who have taught me to be a good mentor myself. Peter-Arno, Lou, and Nelleke, I would like to thank you for being my PhD supervisors. The voice of Lou is still sometimes in my head: *what is your goal*? It has always remained to be relevant. Other people have mentored me informally over the past 15 years: Antal, Arjen, and Maarten, thank you for all the feedback I have received from you over the years and for still being there for me when I need support and advice. Wessel, thank you for our years-long friendly collaboration, both in research and teaching.

Aske, you have been an incredibly supportive mentor in the past years. You brought me into LIACS and always made me feel that I 100% belonged. You are a key example of kind leadership and I hope that your door will remain open for me.

Ik wil mijn ouders bedanken voor hoe ze mij het leven hebben leren leven, met als belangrijkste waarden zelfstandigheid, je eigen keuzes maken en te doen wat je leuk vindt. “Self supporting” was een gevleugelde uitspraak bij ons thuis; zorg dat je altijd zelf je weg kunt vinden. Ik denk dat dat gelukt is.

Joris en Tijmen, jullie zijn de toekomst. Het is geweldig om te zien hoe jullie de wereld tegemoet treden op jullie eigen manier. Ik wil jullie bedanken voor het altijd zijn van de plek waar geen werk is.

En tot slot: Paul, jou wil ik bedanken voor al je steun, liefde en voor het vormen van een twee-eenheid. Ons huishouden is een logistieke puzzel van duizend stukjes, maar samen krijgen we het altijd voor elkaar.

Ik heb gezegd.

Referenties

Abolghasemi, A., Azzopardi, L., Askari, A., de Rijke, M., Verberne, S., 2024. Measuring Bias in a Ranked List using Term-based Representations, in: European Conference on Information Retrieval. Springer, pp. 3–19.

Abolghasemi, A., Verberne, S., Askari, A., Azzopardi, L., 2023. Retrievability Bias Estimation Using Synthetically Generated Queries, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Presented at the CIKM '23: The 32nd ACM International Conference on Information and Knowledge Management, ACM, Birmingham United Kingdom, pp. 3712–3716. https://doi.org/10.1145/3583780.3615221

Askari, A., Aliannejadi, M., Abolghasemi, A., Kanoulas, E., Verberne, S., 2023a. CLosER: Conversational Legal Longformer with Expertise-Aware Passage Response Ranker for Long Contexts, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Presented at the CIKM '23: The 32nd ACM International Conference on Information and Knowledge Management, ACM, Birmingham United Kingdom, pp. 25–35. https://doi.org/10.1145/3583780.3614812

Askari, A., Aliannejadi, M., Kanoulas, E., Verberne, S., 2023b. A Test Collection of Synthetic Documents for Training Rankers: ChatGPT vs. Human Experts, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Presented at the CIKM '23: The 32nd ACM International Conference on Information and Knowledge Management, ACM, Birmingham United Kingdom, pp. 5311–5315. https://doi.org/10.1145/3583780.3615111

Askari, A., Aliannejadi, M., Meng, C., Kanoulas, E., Verberne, S., 2023c. Expand, Highlight, Generate: RL-driven Document Generation for Passage Reranking, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 10087–10099.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? , in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Presented at the FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, Virtual Event Canada, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Brandsen, A., Lambers, K., Verberne, S., Wansleeben, M., 2019. User requirement solicitation for an information retrieval system applied to Dutch grey literature in the archaeology domain. Journal of Computer Applications in Archaeology 2, 21–30. https://doi.org/10.5334/jcaa.33

Brandsen, A., Verberne, S., Lambers, K., Wansleeben, M., 2022. Can BERT Dig It? Named Entity Recognition for Information Retrieval in the Archaeology Domain. J. Comput. Cult. Herit. 15, 1–18. https://doi.org/10.1145/3497842

Brandsen, A., Verberne, S., Wansleeben, M., Lambers, K., 2020. Creating a dataset for named entity recognition in the archaeology domain, in: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4573–4577.

Broekens, J., Hilpert, B., Verberne, S., Baraka, K., Gebhard, P., Plaat, A., 2023. Fine-grained Affective Processing Capabilities Emerging from Large Language Models, in: 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, pp. 1–8.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Chen, L., 2023. 'I Feel You': Enhancing conversational agents with empathy. Leiden University.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Presented at the NAACL-HLT 2019, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

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

Dirkson, A., Verberne, S., Van Oortmerssen, G., Gelderblom, H., Kraaij, W., 2023. How do others cope? Extracting coping strategies for adverse drug events from social media. Journal of Biomedical Informatics 139, 104228.

Harris, Z.S., 1954. Distributional Structure. *WORD* 10, 146–162. https://doi.org/10.1080/00437956.1954.11659520

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P., 2023. Survey of Hallucination in Natural Language Generation. ACM Comput. Surv. 55, 1–38. https://doi.org/10.1145/3571730

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. Advances in Neural Information Processing Systems 33, 9459–9474.

Lin, J., Nogueira, R., Yates, A., 2022. Pretrained transformers for text ranking: BERT and beyond, 1st ed, Synthesis Lectures on Human Language Technologies. Springer Nature.

Navigli, R., Conia, S., Ross, B., 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality 15, 10:1-10:21. https://doi.org/10.1145/3597307

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training.

Reuver, M., Fokkens, A., Verberne, S., 2021a. No NLP task should be an island: multi-disciplinarity for diversity in news recommender systems, in: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. pp. 45–55.

Reuver, M., Verberne, S., Fokkens, A., 2024. Investigating the Robustness of Modelling Decisions for Few-Shot Cross-Topic Stance Detection: A Preregistered Study, in: Proceedings of the 30th International Conference on Computational Linguistics (LREC-COLING).

Reuver, M., Verberne, S., Morante, R., Fokkens, A., 2021b. Is Stance Detection Topic-Independent and Cross-topic Generalizable? A Reproduction Study, in: Proceedings of the 8th Workshop on Argument Mining. pp. 46–56.

Santos, R.L., Macdonald, C., Ounis, I., 2015. Search result diversification. Foundations and Trends® in Information Retrieval 9, 1–90.

Sparck Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation 28, 11–21.

Sun, W., Yan, L., Chen, Zheng, Wang, S., Zhu, H., Ren, P., Chen, Zhumin, Yin, D., Rijke, M., Ren, Z., 2024. Learning to tokenize for generative retrieval. Advances in Neural Information Processing Systems 36.

Tseng, R., Verberne, S., Van Der Putten, P., 2023. ChatGPT as a Commenter to the News: Can LLMs Generate Human-Like Opinions?, in: Disinformation in Open Online Media, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp. 160–174. https://doi.org/10.1007/978-3-031-47896-3_12

Van den Bosch, A.P.J., 2012. Taal in uitvoering.

Van den Bosch, A.P.J., 2008. Het volgende woord.

Van Dijk, B., van Duijn, M., Verberne, S., Spruit, M., 2023. ChiSCor: A Corpus of Freely-Told Fantasy Stories by Dutch Children for Computational Linguistics and

Cognitive Science, in: Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL). Association for Computational Linguistics, Singapore, pp. 352–363. https://doi.org/10.18653/v1/2023.conll-1.23

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems.

Verberne, S., 2023. Service-chatbots voor het Nederlands: De Onderzoeksagenda van het LESSEN Project. DIXIT 2023 – Tijdschrift over toegepaste taal- en spraaktechnologie 20, 32–33.

Verberne, S., 2002. Context-sensitive spell checking based on word trigram probabilities. Radboud University, Nijmegen.

Wiggers, G., Verberne, S., Van Loon, W., Zwenne, G., 2023. Bibliometric-enhanced legal information retrieval: Combining usage and citations as flavors of impact relevance. Asso for Info Science & Tech 74, 1010–1025. https://doi.org/10.1002/asi.24799

Wiggers, G., Verberne, S., Zwenne, G.-J., Van Loon, W., 2022. Exploration of domain relevance by legal professionals in information retrieval systems. Legal Information Management 22, 49–67.

Zakkas, P., Verberne, S., Zavrel, J., 2024. SumBlogger: Abstractive Summarization of Large Collections of Scientific Articles, in: Advances in Information Retrieval. Springer Nature Switzerland, Cham, pp. 371–386. https://doi.org/10.1007/978-3-031-56027-9_23

Prof. Dr. Suzan Verberne

1998	Eindexamen Gymnasium, Scholengemeenschap De Grundel, Hengelo
2002	Masterdiploma Taal, Spraak en Informatica, Radboud Universiteit, Nijmegen
2003-2004	Linguistic engineer, Polderland Language & Speech Technology
2005-2009	Promovendus, Radboud Universiteit, Nijmegen
2010-2017	Postdoctoraal onderzoeker, Radboud Universiteit, Nijmegen
2017-2020	Universitair Docent (tenure track), Universiteit Leiden
2020-2023	Universitair Hoofddocent, Universiteit Leiden
2023-nu	Hoogleraar Natural Language Processing, Universiteit Leiden

Suzan Verberne is hoogleraar Natural Language Processing bij het Leiden Institute of Advanced Computer Science (LIACS) van de Universiteit Leiden. Ze is in 2010 aan de Radboud Universiteit gepromoveerd op het gebied van slimme zoeksystemen en werkt sindsdien op het raakvlak van taaltechnologie en zoektechnologie. Ze heeft projecten begeleid in een groot aantal toepassingsdomeinen – van sociale media tot rechten en van archeologie tot gezondheidszorg. Haar huidige onderzoek gaat over de relatie tussen zoekmachines en chatbots, met een focus op specifieke domeinen en op toepassingen waar niet veel data en computerkracht beschikbaar is. Ze is actief in Europese consortia en leidt een nationaal project over chatbots voor het Nederlands. Ze geeft twee mastervakken en begeleidt een groot aantal studenten. Daarnaast is ze voorzitter van de examencommissie van LIACS en van de facultaire ethische commissie, en zet ze zich actief in voor internationale conferenties, o.a. op het gebied van diversiteit en inclusie.

Universiteit Leiden