AUTOMATISCH TEKSTEN VERSIMPELEN IS NOG NIET ZO SIMPEL

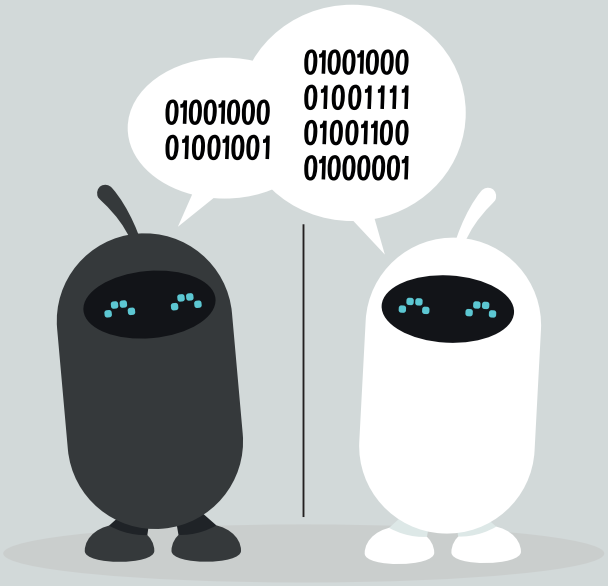
Veel mensen in Nederland hebben moeite met lezen, daarom gebruikte de KB de workshop ICT with Industry om te onderzoeken of AI kan worden ingezet voor het versimpelen van teksten. Inkijkje in het project ARTIST.

In de rubriek 'KB Onderzoekskroniek' beschrijven medewerkers van de afdeling Onderzoek van de Koninklijke Bibliotheek hun resultaten, trends en vondsten.

Elk jaar organiseren de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) en het Lorentz Center in Leiden de workshop ICT with Industry. Tijdens deze een week durende workshop werken studenten, onderzoekers, academici en professionals uit het bedrijfsleven samen aan relevante vraagstukken uit de praktijk. Afgelopen januari vond er weer een ICT with Industry plaats; de KB deed voor de derde keer mee. Vijf enthousiaste deelnemers werkten samen met Marijn Koolen (Huygens ING) en Jie Yang (TU Delft) aan het project ARTificial Intelligence for Simplified Texts (ARTIST), ofwel: kunnen we met behulp van artificiële intelligentie teksten automatisch versimpelen?

Laaggeletterdheid

In Nederland hebben ongeveer 2,5 miljoen mensen tussen de 16 en 65 jaar moeite met lezen. Hierdoor kunnen zij tegen problemen aanlopen in de huidige maatschappij. Een recent voorbeeld hiervan zijn de verstuurde uitnodigingen voor de coronavaccinatie. Voor laaggeletterde mensen kan zo'n brief te moeilijk zijn, waardoor zij de informatie niet, of niet goed, begrijpen.



Op dit moment lopen er verschillende initiatieven waarbij teksten handmatig worden versimpeld. Dit proces duurt echter erg lang, waardoor maar een klein deel van alle gepubliceerde teksten in een eenvoudiger versie beschikbaar komt. Daarom wil de KB graag onderzoeken welke methodes gebruikt zouden kunnen worden om teksten geautomatiseerd te versimpelen en daarmee toegankelijk te maken voor alle inwoners van Nederland.

Veel trainingsdata nodig

Het ARTIST-team ging enthousiast aan de slag met het vraagstuk, maar al snel werd duidelijk dat er geen simpele, snelle oplossing voor dit probleem bestaat. Mensen kunnen namelijk om verschillende redenen moeite hebben met het lezen en begrijpen van teksten. Bijvoorbeeld omdat Nederlands niet hun moedertaal is, of omdat ze laagbegaafd of dyslectisch zijn. Deze groepen hebben ieder baat bij een andere oplossing.

Daarnaast moeten voor een geautomatiseerde oplossing machine learning-modellen worden getraind. Hiervoor heb je een grote hoeveelheid geschikte trainingsdata nodig, en hiervan zijn er helaas maar weinig beschikbaar in het Nederlands. Verder is het vaak niet voldoende om alleen moeilijke woorden te vervangen voor een simpelere variant, omdat de zinsopbouw zelf ook complex kan zijn.

Experimenteren met opties

De deelnemers aan ARTIST kwamen met verschillende oplossingsrichtingen. In hun zoektocht naar mogelijke datasets ontdekten ze de Canon van Nederland (canonvannederland.nl) die bestaat uit teksten met verschillende moeilijkheidsgraden: van basisschool tot voorgezet onderwijs. Ook dachten ze na over verschillende manieren van versimpelen. De meest voor de hand liggende methoden zijn moeilijke woorden vervangen en zinnen verkorten, maar een suggestie was ook om moeilijke teksten te versimpelen door middel van visualisaties.

Verder werden er verschillende al bestaande machine learning-modellen getest. De uitkomsten hiervan zijn alleen nog niet erg betrouwbaar. Dit was echter te verwachten omdat er voor deze specifieke taak niet speciaal een model was getraind. <

Online samenwerken

Normaal gesproken vindt de ICT with Industry-workshop plaats in het Lorentz Center, maar vanwege de coronamaatregelen was hij dit jaar online. Ondanks het gebrek aan fysiek contact ontstond er al snel een teamgevoel en was er al snel gezamenlijk een strategie bedacht voor de rest van de week. Om de deelnemers in de gelegenheid te stellen ook daadwerkelijk te kunnen experimenteren met machine learning, stelde SURF gedurende de week een server beschikbaar waarop kon worden gewerkt.

Mirjam Cuper
Data scientist bij
de Koninklijke Bibliotheek