A.I. EN MACHINE LEARNING VOOR AUDIOVISUELE ARCHIEVEN

Een onderzoek naar de huidige machine learning toepassingen voor audiovisuele archieven, en de toepasbaarheid op Nederlandstalige bronnen

SCAN ME

CONTACT :
contact@timvanhal.com
www.timvanhal.com

Inhoudstafel

1. Voorwoord

Dankjewel voor de interesse in mijn afstudeerproject voor het behalen van het graduaat informatiebeheer aan de Arteveldehogeschool. Voor het afronden van deze studie worden de studenten gevraagd om een project te verwezenlijken dat een bijdrage levert aan het werkveld. Tijdens mijn stage bij het stadsarchief van Gent heb ik veel bijgeleerd, en was er ook ruimte om mijn eigen interesses in het werkveld verder uit te werken.

De projecten van mijn medestudenten zijn bijzonder uiteenlopend en leunen voornamelijk aan tegen onderwerpen die van toepassing zijn op bibliotheken, aangezien deze richting binnen het graduaat informatiebeheer het populairst is. Mijn keuze voor dit project combineert drie interesses. Een professionele interesse in de archiefwereld; een hobbymatige interesse in de informatica, voornamelijk gericht op nieuwe technologieën; en een interesse in het audiovisuele, voornamelijk film.

In mijn vrije tijd neem ik veel interviews af met mensen die werken in de filmindustrie. Het eigenhandig transcriberen van deze interviews is een zeer tijdrovende taak, en het is moeilijk om tijd vrij te maken voor iets wat weinig oplevert in vergelijking met de investering om het te verwezenlijken. Één uur interview neemt al snel het drievoudige in beslag om te transcriberen. Daar kwam verandering in toen in september 2022 het bedrijf OpenAI de eerste publieke versie van Whisper uitbracht. Dit open source programma kan taalmodellen inzetten die samengesteld werden dmv. machine learning om audio te transcriberen. Het merendeel van de taalmodellen heeft enkel ondersteuning voor de Engelse taal, wat voor mijn persoonlijke doeleinden meer dan voldoende was. Het grootste taalmodel daarentegen heeft ondersteuning voor maar liefst 98 verschillende talen, waaronder ook het Nederlands.

Bij het zoeken van een geschikt project voor mijn stage besloot ik om de Nederlandstalige ondersteuning van Whisper eens op de testbank te leggen, met daarnaast twee commerciële pakketten voor video editing die ook ondersteuning bieden voor automatische transcriptie. Met dit project hoop ik bruikbare resultaten te delen met instellingen in het werkveld die meer uitgebreide metadata voor hun audiovisuele archieven overwegen, en dat het kan dienen als een blik op wat er ondertussen mogelijk is met machine learning (ML) *speech-to-text* toepassingen.

Ik zou graag de medewerkers van het stadsarchief van Gent willen bedanken, met in het bijzonder Peter De Bock, en Matthias Priem van de vzw Meemoo, voor hun tijd en raad in verband met het opzetten van dit project. Peter De Bock leverde de materialen aan, en gaf me goede raad ivm. de omvang van deze test. Bij Meemoo bracht Matthias Priem mij op de hoogte van hun gelijkaardige project, met een véél grotere *scope*, en ook hun testmethode heeft toegelicht. Dit heeft mij geholpen bij de keuze voor de juiste *tools* om dit project uit te voeren.

2. Project

Het doel van dit project is de huidige AI en ML toepassingen voor transcriptie in de Nederlandse taal uit te testen. Er werd gekozen voor programma's die enkel lokaal op de computer de audiovisuele data kunnen verwerken.

De aanzet was OpenAI's Whisper, omdat die niet alleen ondersteuning heeft voor het Engels, maar ook 98 andere talen kan transcriberen. De ondersteuning voor de overige talen is aanzienlijk kleiner, maar volgens de makers loopt het Nederlandse taalmodel voor op de meerderheid van de ondersteunde talen qua resultaten[1].

Het archief van Gent heeft in hun collectie de *'Gentse Filmaktualiteiten'* van de producent Daska. Deze bestaan uit 105 korte films met lokaal nieuws uit de jaren '70 en '80, en werden oorspronkelijk vertoond in de Gentse cinema's. De beeldkwaliteit van deze films is bijzonder goed, en de gesproken taal is vrij zuiver. Daarom leek het me interessant om als steekproef tien Daska films te selecteren, en te laten transcriberen met ML toepassingen.

Tijdens initiële fase van mijn project bracht OpenAI een update uit van het taalmodel met de Nederlandstalige ondersteuning. Volgens het bedrijf zou de ondersteuning voor het Nederlands verbeterd zijn, dus de timing was ideaal om zowel versie 2, als versie 3 van het *Large* taalmodel te vergelijken.

Sinds enkele jaren ondersteunen enkele van de bekendere video editing pakketten ook transcriptie dmv. machine learning. Het lijkt me interessant om eens te kijken hoe deze populaire pakketten presteren, dus daarom heb ik de resultaten van deze twee softwarepakketten meegenomen in mijn test. Het gaat om Premiere Pro van Adobe, en om DaVinci Resolve Studio van het bedrijf Blackmagic Design.

Ondanks dat er veel aanbieders zijn voor automatische transcriptie in de cloud, werd de keuze gemaakt om archiefmaterialen bij het archief te houden, en deze niet te uploaden naar een externe partij. Sommige commerciële partijen, waaronder ook OpenAI, de producent van Whisper, gebruiken aangeleverde data om hun taalmodellen verder te perfectioneren. Daarom werd de keuze gemaakt om enkel voor lokale verwerking te kiezen, zodat de gebruikte data niet in andere toepassingen van deze bedrijven gebruikt kan worden[2].

[1] https://github.com/openai/whisper/discussions/1762

[2] Meemoo maakt wel gebruik van externe aanbieders om hun audiovisuele materialen te verwerken, maar beperken contractueel de bruikbaarheid van deze data, en waar deze data verwerkt zal worden, conform met de GDPR-regels.

3. Werkwijze

OpenAI's Whisper is een open source toepassing die gratis beschikbaar is gesteld als een Python programma. Er zijn ondertussen door de community al verschillende aanpassingen en uitbreidingen gemaakt op Whisper, en via andere ontwikkelaars is het ook beschikbaar gemaakt met een graphical user interface (GUI). Voor deze test maak ik gebruik van het oorspronkelijke programma om een zo zuiver mogelijk resultaat te krijgen.

Ik liet Whisper de tien video's transcriberen met zowel het Large v2-model, als het Large v3-model. Als output koos ik voor iedere video voor *.srt en *.txt. De keuze voor *.srt is voor de hand liggend, aangezien dit de standaard output is van Whisper. Dit soort ondertitelingsbestand is ook bruikbaar met de meest gangbare videospelers voor de courante besturingssystemen. De keuze om deze bestanden ook als *.txt te exporteren komt voort uit de mogelijkheid om ook de transcriptie zonder de tijdscodering [3] te exporteren. Dit maakt de voorbereiding voor de test gemakkelijker, aangezien dit ervoor zorgt dat ik minder tijd moet besteden om de bestanden klaar te maken voor de *benchmarking* software.

Voor de twee commerciële pakketten heb ik gekozen voor een transcriptie zonder aanpassingen aan de standaardinstellingen, met uitzondering van de selectie voor de Nederlandse taal. De geselecteerde output was wederom *.srt en *.txt. In beide pakketten zit er ook ondersteuning om de output op te slaan met of zonder tijdscodering, wat de verwerking weer vergemakkelijkt.

Na de uitvoering van alle transcripties heb ik voor alle tien de video's telkens vier ondertitelingsbestanden, en vier keer de output als tekstbestand zonder tijdscodering. Daarna maak ik een eigen transcriptie van de videobestanden die achteraf nog drie keer is gecontroleerd op correctheid. Deze manueel getranscribeerde bestanden zijn de baseline transcripties waar de output van Whisper, Premiere Pro, en DaVinci Resolve Studio aan zal worden afgetoetst.

Om de prestaties van de softwarepakketten te testen maak ik gebruik van BenchmarkSTT versie 1.1. Deze software werd ontwikkeld door EBU, de European Broadcasting Union; een alliantie die als doel heeft een duurzame toekomst te bieden aan openbare omroepen. De software heeft als doel AI en ML *speech-to-text* toepassingen te benchmarken. De EBU werkt ook aan ondersteuning voor afbeelding- en gezichtsherkenning. Voor dit project beperken we de testen tot *speech-to-text*.

Voor een zo neutraal mogelijk resultaat is het aangeraden om de tekst te normaliseren. Dit wil zeggen dat er geen hoofdletters of leestekens in de *input* mogen staan. Zo kan BenchmarkSTT een beoordeling maken puur op basis van de getranscribeerde input, ongeacht of deze met of zonder hoofdletters geschreven zijn, of andere tekens bevatten zoals bvb.: punten, komma's of aanhalingstekens. Dit geeft een duidelijker beeld van accuraatheid van de modellen.

[3] Deze codering vertelt de videospeler wanneer de ondertiteling in beeld moet komen. Bvb: "00:00:21,640 --> 00:00:27,440"

Voor een zo correct mogelijk resultaat heb ik besloten om koppeltekens in woorden te bewaren, en om beschrijvingen voor geluid, zoals bvb. “MUZIEK”, te verwijderen. Aangezien de gebruikte software verschillende methodes gebruiken om muziek in de bestanden aan te duiden. Zelfs binnen de twee gebruikte Whisper taalmodellen is er een andere hantering om deze weer te geven, en deze zouden de resultaten onterecht negatief beïnvloeden.

De huidige AI en ML toepassingen kunnen soms onderhevig zijn aan hallucinaties. Dit fenomeen komt voornamelijk voor wanneer de toepassing een fout begaat tijdens de verwerking, en hierdoor de output vaak verkeerd weergeeft. Voorbeelden van hallucinaties zijn herhalingen van woorden, of andere interpretaties van bepaalde woorden. Dit is vaak het gevolg van de gebruikte trainingsdata. Een veel voorkomende hallucinatie is het proberen woorden te herkennen in muziek. De *waveforms* waar de modellen op getraind worden kunnen soms overeenkomen met de *waveform* van een bepaald woord, waardoor het model muziek wel eens verkeerd durft te labelen.

In mijn transcripties met Whisper viel het vaak op dat de muziek in de eindgeneriek vaak gelabeld werd als “TV GELDERLAND 2021”. Mogelijk is dit omdat de muziek ook door andere producenten gebruikt werd en op die manier dat label kreeg in de trainingsdata van het taalmodel. Ondanks dat de beschrijvingen van muziek in de films bij het normaliseren wel verwijderd werden, werd de beslissing gemaakt om woorden die toegekend werden aan muziek wel in de transcriptie te laten.

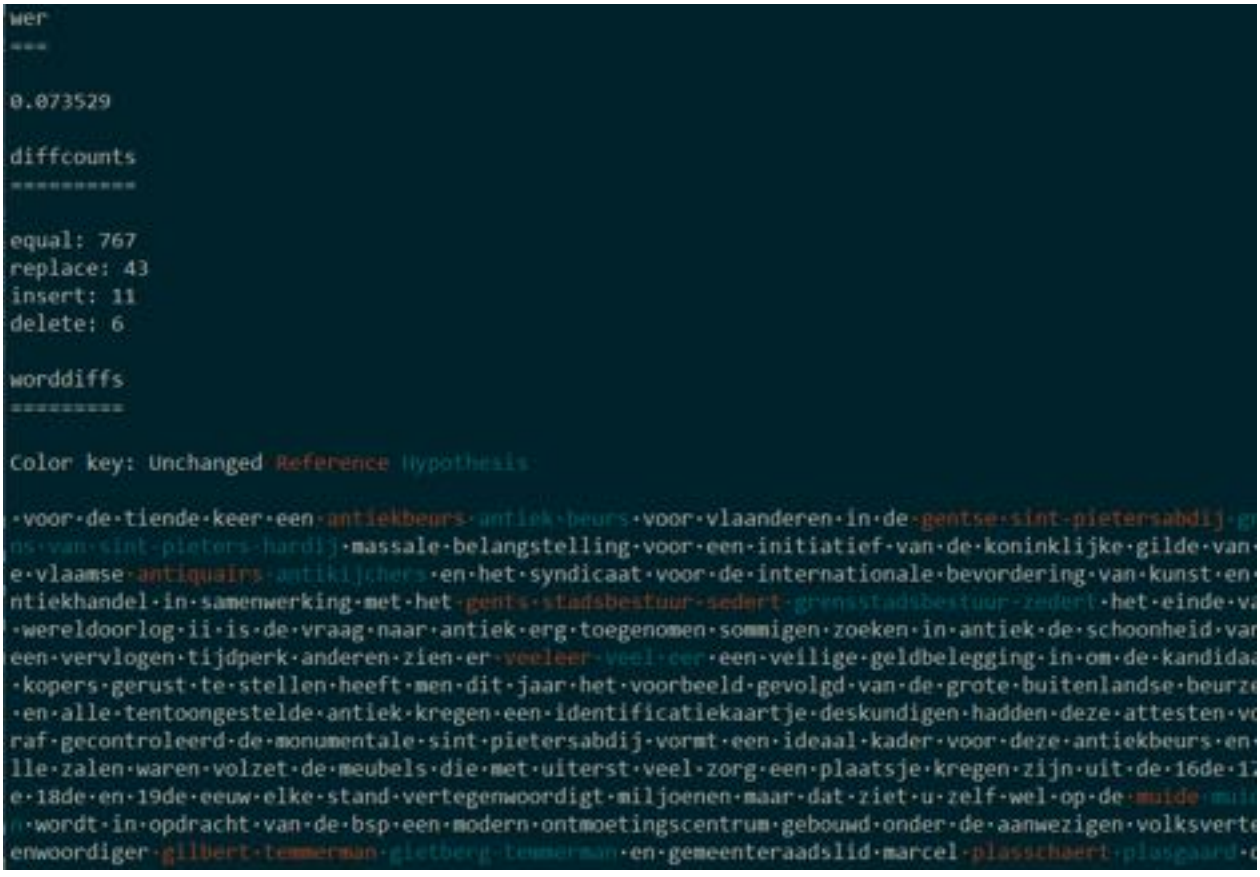
4. Resultaten

Bestand	Software	WER	equal	replace	insert	delete
Daska 01	Whisper Large v2	0.128378	530	55	14	7
	Whisper Large v3	0.060811	561	28	5	3
	Premiere Pro 25.3.1	0.221284	504	84	43	4
	DaVinci Resolve Studio 18.6.4	0.261824	447	102	10	43
Daska 02	Whisper Large v2	0.086799	506	38	1	9
	Whisper Large v3	0.068716	522	28	7	3
	Premiere Pro 25.3.1	0.142857	501	45	27	7
	DaVinci Resolve Studio 18.6.4	0.289331	437	94	44	22
Daska 04	Whisper Large v2	0.131078	836	100	14	10
	Whisper Large v3	0.104651	860	78	13	8
	Premiere Pro 25.3.1	0.206131	815	125	64	6
	DaVinci Resolve Studio 18.6.4	0.276956	718	201	34	27

Daska 05	Whisper Large v2	0.160677	837	90	43	19
	Whisper Large v3	0.071882	884	53	6	9
	Premiere Pro 25.3.1	0.217759	798	142	58	6
	DaVinci Resolve Studio 18.6.4	0.254757	719	193	14	34
Daska 06	Whisper Large v2	0.082278	446	26	11	2
	Whisper Large v3	0.054852	454	18	6	2
	Premiere Pro 25.3.1	0.137131	436	37	27	1
	DaVinci Resolve Studio 18.6.4	0.200422	396	69	17	9
Daska 07	Whisper Large v2	0.156204	597	71	19	17
	Whisper Large v3	0.083212	640	42	12	3
	Premiere Pro 25.3.1	0.261314	570	107	64	8
	DaVinci Resolve Studio 18.6.4	0.313869	520	144	50	21
Daska 09	Whisper Large v2	0.159335	924	86	64	13
	Whisper Large v3	0.069404	959	53	7	11
	Premiere Pro 25.3.1	0.257087	917	102	157	4
	DaVinci Resolve Studio 18.6.4	0.332356	787	207	104	29
Daska 10	Whisper Large v2	0.095827	592	49	7	6
	Whisper Large v3	0.054096	617	27	5	3
	Premiere Pro 25.3.1	0.154560	585	57	38	5
	DaVinci Resolve Studio 18.6.4	0.258114	521	90	41	36
Daska 11	Whisper Large v2	0.126225	730	78	17	8
	Whisper Large v3	0.073529	767	43	11	6
	Premiere Pro 25.3.1	0.174020	724	85	50	7
	DaVinci Resolve Studio 18.6.4	0.254902	632	171	24	13
Daska 12	Whisper Large v2	0.118483	382	38	10	2
	Whisper Large v3	0.075829	395	25	5	2
	Premiere Pro 25.3.1	0.206161	382	40	47	0
	DaVinci Resolve Studio 18.6.4	0.222749	343	71	15	8

Verklarende termen:

- **WER**: *Word Error Rate*. Een calculatie die het aantal fouten deelt door het totale aantal woorden in de tekst
- **Equal**: het aantal woorden dat exact overeenkomt met het bronbestand
- **Replace**: het aantal woorden dat vervangen werd door een woord dat hard lijkt op een woord in het bronbestand. Bijvoorbeeld: “pan” en “pen”
- **Insert**: het aantal woorden dat moet worden toegevoegd om overeen te komen met het bronbestand
- **Delete**: het aantal woorden dat ontbreekt om overeen te komen met het bronbestand.



(de output van BenchmarkSTT voor één van de Whisper Large v3 transcripties)

5. Conclusie

Whisper Large v3 is overduidelijk als beste uit de test gekomen. Alle beste scores voor de *word error rate* (WER) behoren toe aan dit taalmodel. De claim van OpenAI dat het Large v3-model ook beter zou presteren van v2 is bij deze ook duidelijk bewezen. Ondanks dat de prestaties van Large v2 zelden ver liggen van het vernieuwde taalmodel, behaalt dit model over alle tests de tweede plaats.

Op de derde plaats vinden we Premiere Pro van Adobe. Bij een visuele controle van de *.srt-bestanden viel het op dat dit pakket, ondanks slechter te scoren dan de Whisper-modellen, standaard een spell check toepast op de output. Dit zorgde voor meer correcte benamingen van landen, steden en plaatsen; voornamelijk bij het gebruik van hoofdletters op de juiste plaatsen. Helaas worden hoofdletters niet meegenomen in de benchmark, dus heeft Premiere Pro zijn voorsprong op DaVinci Resolve Studio voornamelijk te danken aan de juiste schrijfwijzes.
Een teleurstellend resultaat van Premiere Pro, aangezien Adobe heel sterk inzet op A.I. en machine learning in de recente versies van hun toepassingen.

DaVinci Resolve Studio eindigde bij alle resultaten op de laatste plaats. Bij een visuele controle van de output bleek er bij dit pakket vaker last te zijn van hallucinaties. Deze komen vooral voor als woorden die meermaals herhaald worden in de output, ondanks maar eenmaal of niet voor te komen in het bronbestand. Desalnietemin maakte deze software veel fouten in de transcriptie. Ondanks dat deze software een steeds grote populariteit geniet tegenover andere video editing pakketten, is het duidelijk dat de transcriptie toepassingen niet hoog in de lijst met prioriteiten staat bij Blackmagic Design.

Omdat in vrijwel alle AI en ML toepassingen fouten en hallucinaties kunnen voorkomen, is het risico wel in te calculeren bij het transcriberen. Door middel van *spell checking* en scripting (bvb. door zoeken naar vaak herhaalde woorden) kunnen deze er gemakkelijk uitgefilterd worden.

Ondanks dat deze pakketten inmiddels al enkele jaren bestaan, staat het gebruik van machine learning bij *speech-to-texst* toepassingen nog in de kinderschoenen. Mits controle van de output is deze wel bruikbaar voor archieven. Door het gebruik van scripting bij de controle kunnen er grote batches automatisch gecontroleerd en gecorrigeerd worden, en bestanden die grotere problemen bevatten gemarkeerd worden voor visuele controle. Net zoals bij vrijwel alle digitalisering, is het sowieso aangeraden om ook controles uit te voeren via steekproeven.

Bij de controle van de output van deze test blijkt dat de grootste fouten voornamelijk gemaakt worden bij plaatsnamen (steden, dorpen, straten, wijken) en eigennamen. In theorie is het mogelijk om dit soort problemen af te vangen bij de controle via scripting door de context van de output te combineren met namenlijsten.
In het geval van de Daska films gaat het bijna uitsluitend over Gent, en zouden fouten als “Gentsche”, “Gensche”, “Gens” en “Gend” gemakkelijk gecorrigeerd kunnen worden als deze in

een controlelijst staan die het script kan raadplegen. Dit zou ook toegepast kunnen worden met namen van koningen, ministers, burgemeesters, schepenen, gemeenteraadsleden; en lijsten met steden, wijken en straatnamen. Deze gegevens zouden gekoppeld kunnen worden aan de al bestaande metadata, door bijvoorbeeld het jaartal te controleren, en daar de bijbehorende lijsten te consulteren. Helaas vergt het maken van dit soort controlemiddelen veel tijd en werk, waardoor het voor kleinere instellingen moeilijk haalbaar is om dit te verwezenlijken. Op nationale schaal is dit mogelijk wel haalbaar.

De organisatie Meemoo heeft een vergelijkbare test gedaan voor *speech-to-text* toepassingen, maar op veel grotere schaal. Ze werken momenteel ook aan de toepassing van gezichtsherkenning, en dit zou naast het gebruik van controlelijsten ook een zeer bruikbaar hulpmiddel kunnen zijn om de juiste personen te benoemen in de output.

AI en ML toepassingen staan nog in de kinderschoenen, maar zijn in beperkte mate al toepasbaar binnen archieven. Met het transcriberen van audiovisuele materialen is het met ML nu al mogelijk om de context van het materiaal samen te vatten, en persoonsnamen te onttrekken om te gebruiken in andere metadata velden. Hierdoor kunnen deze bronnen gemakkelijker vindbaar gemaakt worden, en is het mogelijk om audiovisuele collecties beter te ontsluiten voor het publiek.

Voor meer informatie ivm. OpenAI's Whisper:
https://github.com/openai/whisper

Voor meer informatie over BenchmarkSTT:
https://benchmarkstt.readthedocs.io/en/latest/index.html