Vergelijking en evaluatie van modellen voor het gamificeren van het Info Support Guidance Framework met een automatische quiz generator.

Optionele ondertitel.

Manu Vleurick.

Scriptie voorgedragen tot het bekomen van de graad van
Professionele bachelor in de toegepaste informatica

Promotor: Mevr. L. De Mol
Co-promotor: Dhr. Y. Van Damme
Academiejaar: 2022–2023
Eerste examenperiode

Departement IT en Digitale Innovatie .

HO
GENT

Woord vooraf

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Samenvatting

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor.

Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Inhoudsopgave

Lijst van figuren

1

Inleiding

De inleiding moet de lezer net genoeg informatie verschaffen om het onderwerp te begrijpen en in te zien waarom de onderzoeksvraag de moeite waard is om te onderzoeken. In de inleiding ga je literatuurverwijzingen beperken, zodat de tekst vlot leesbaar blijft. Je kan de inleiding verder onderverdelen in secties als dit de tekst verduidelijkt. Zaken die aan bod kunnen komen in de inleiding (**Pollefliet2011**):

- context, achtergrond
- afbakenen van het onderwerp
- verantwoording van het onderwerp, methodologie
- probleemstelling
- onderzoeksdoelstelling
- onderzoeksvraag
- ...

1.1. Probleemstelling

Uit je probleemstelling moet duidelijk zijn dat je onderzoek een meerwaarde heeft voor een concrete doelgroep. De doelgroep moet goed gedefinieerd en afgelijnd zijn. Doelgroepen als “bedrijven,” “KMO’s”, systeembeheerders, enz. zijn nog te vaag. Als je een lijstje kan maken van de personen/organisaties die een meerwaarde zullen vinden in deze bachelorproef (dit is eigenlijk je steekproefkader), dan is dat een

indicatie dat de doelgroep goed gedefinieerd is. Dit kan een enkel bedrijf zijn of zelfs één persoon (je co-promotor/opdrachtgever).

1.2. Onderzoeksvraag

Wees zo concreet mogelijk bij het formuleren van je onderzoeksvraag. Een onderzoeksvraag is trouwens iets waar nog niemand op dit moment een antwoord heeft (voor zover je kan nagaan). Het opzoeken van bestaande informatie (bv. “welke tools bestaan er voor deze toepassing?”) is dus geen onderzoeksvraag. Je kan de onderzoeksvraag verder specifiëren in deelvragen. Bv. als je onderzoek gaat over performantiemetingen, dan

Onderzoeksvragen:

- Welke nlp modellen kunnen toegepast worden om automatisch vragen te genereren op basis van IT-gerelateerde kennis in het Guidance Framework van Info Support?
  - Welke factoren zijn van invloed op de prestaties van deze modellen?
- Welke evaluatiemethoden zijn het meest geschikt voor het evalueren van de getrainde NLP modellen?

1.3. Onderzoeksdoelstelling

Wat is het beoogde resultaat van je bachelorproef? Wat zijn de criteria voor succes? Beschrijf die zo concreet mogelijk. Gaat het bv. om een proof-of-concept, een prototype, een verslag met aanbevelingen, een vergelijkende studie, enz.

1.4. Opzet van deze bachelorproef

De rest van deze bachelorproef is als volgt opgebouwd:

In Hoofdstuk 2 wordt een overzicht gegeven van de stand van zaken binnen het onderzoeksdomein, op basis van een literatuurstudie.

In Hoofdstuk 3 wordt de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen.

In Hoofdstuk 4, tenslotte, wordt de conclusie gegeven en een antwoord geformuleerd op de onderzoeksvragen. Daarbij wordt ook een aanzet gegeven voor toe-

komstig onderzoek binnen dit domein.

2

Stand van zaken

In deze literatuurstudie wordt er onderzocht hoe automatische question generation het best kan bijdragen aan het efficiënt en effectief creëren van quizvragen. Eerst wordt besproken wat de voordelen van automatische question generation zijn voor het opfrissen van kennis.

Daarna wordt er gekeken naar de Natural Language Processing(NLP) technieken achter het automatisch markeren van belangrijke zinsdelen, een belangrijke stap in het proces van automatische question generation. Er wordt onderzocht hoe NLP-technieken, zoals NER en POS-tagging, kunnen worden gebruikt om belangrijke informatie in zinnen te identificeren en te markeren.

Vervolgens wordt er gericht op de modellen die speciaal zijn ontworpen voor het automatisch genereren van quizvragen. Er wordt gekeken naar verschillende transformer-gebaseerde modellen, waaronder T5, BART en GPT-2. Deze modellen zijn pre-trained op grote datasets om zo een breed scala aan taalkundige taken te kunnen uitvoeren. Er wordt onderzocht hoe deze modellen kunnen worden gefinetuned op taakgerichte datasets om zo nauwkeuriger quizvragen te genereren.

Om de kwaliteit van gegenereerde quizvragen te beoordelen, worden er verschillende evaluatiemethoden onderzocht, waaronder BLEU, ROUGE en METEOR. Daarnaast wordt ook de evaluatie van quizvragen door mensen besproken en hoe deze kan worden gebruikt om de prestaties van automatische question generation modellen verder te verbeteren. Bovendien wordt er een question answering(QA) model besproken, een model dat specifiek is ontwikkeld om vragen te beantwoorden op basis van de gegeven context. Dit model wordt gebruikt als evaluatietool om de kwaliteit van gegenereerde vragen te controleren.

Door deze verschillende aspecten van automatische question generation te onderzoeken, hopen we inzicht te krijgen in de mogelijkheden en beperkingen van deze technologie en hoe deze kan worden gebruikt om quizvragen efficiënt en effectief te genereren.

2.1. Question Generation

Question Generation(QG), ook wel bekend als Automatic Question Generation(AQG), is een tak van Natural Language Processing(NLP) die zich richt op het automatisch genereren van vragen uit een gegeven input zoals een stuk tekst, een foto of een kennisbank. Het doel is om vragen te produceren die relevant zijn voor het onderwerp en de gegeven input en die het begrip van de lezer kunnen bevorderen. Question Generation heeft verschillende toepassingen, waaronder onderwijs, quizzen en chatbots. Het kan ook gebruikt worden voor question answering taken waarbij het de trainingsdata van question answering modellen kan uitbreiden met gegenereerde vragen. Voor chatbots kan het ook gebruikt worden om verder te gaan met een conversatie of er een te starten.

In de beginjaren van de ontwikkeling van automatische question generation werd voornamelijk gebruik gemaakt van vooraf gedefinieerde, op regels en sjablonen gebaseerde methoden om de automatische generatie van vragen uit te voeren. Dit was zeer tijdrovend en moest aangepast worden per specifiek domein. Recentelijk wordt er meer gebruikt gemaakt van neurale netwerken (Zhang e.a., 2021). Het automatisch genereren van vragen kan tijd besparen als er geen handmatige vraagcreatie nodig is.

Een studie heeft aangetoond dat een AQG-systeem kan helpen bij het verbeteren van het begrijpend lezen van gebruikers in een andere taal. Interessant is dat het positieve leerresultaat behouden bleef, zelfs als de gebruikers wisten dat de vraag door het AQG-systeem was gegenereerd (Steuer e.a., 2022).

2.1.1. Automatisch markeren van belangrijke zinsdelen

QG-modellen genereren automatisch vragen uit gegeven tekst en hebben daarvoor relevante zinsdelen nodig. Om deze reden is het belangrijk om automatisch belangrijke zinsdelen te kunnen markeren. Dit kan bijvoorbeeld gedaan worden door Named Entity Recognition (NER) en Part-of-Speech (POS) tagging. Deze technieken worden vaak gebruikt in combinatie met elkaar om de kwaliteit van de gemarkeerde zinsdelen te verbeteren. In de rest van deze sectie zullen we deze technieken in meer detail bespreken en hun toepassingen in het automatisch markeren van belangrijke zinsdelen uitleggen.

Named Entity Recognition (NER) is een NLP-techniek die benoemde entiteiten in tekst identificeert en classificeert. Dit kan gebruikt worden om vragen te genereren over belangrijke personen, locaties, organisaties, enzovoort. Het NER-proces omvat twee stappen:

- Detecteren van de benoemde entiteiten
- Classificeren in verschillende categorieën

Er zijn verschillende algoritmes beschikbaar voor NER, waarvan Maximum Entropy Markov Model (MEMM) en Common Random Fields (CRF) de meest gebruikte zijn. MEMM is een probabilistisch sequentieel model dat vaak wordt gebruikt in NLP voor taken zoals NER. Het gebruikt maxent-classificatie om de kans te berekenen dat een bepaalde observatie (woord of token) in een bepaalde staat (een NER-label) wordt getagd, afhankelijk van de huidige en vorige observaties en de huidige en vorige toestanden. CRF is een ander probabilistisch sequentieel model dat ook vaak wordt gebruikt voor NERtaken. Het verschil met MEMM is dat CRF de hele sequentie van observaties tegelijkertijd bekijkt en gebruik maakt van de joint probability distribution van alle observaties en toestanden om de meest waarschijnlijke uitvoer te berekenen. Hierdoor kunnen CRF-modellen beter omgaan met afhankelijkheden tussen toestanden en mogelijke onvolledige of onjuiste waarnemingen in de invoer (Analytics, 2022). Om MEMM en CRF te gebruiken, moeten ze worden getraind op een dataset met gelabelde entiteiten zoals de Annotated Corpus for Named Entity Recognition (Walia, 2017).

POS-tagging is een proces waarbij elk woord in een zin wordt gelabeld met zijn syntactische rol, zoals werkwoord, zelfstandig naamwoord, bijvoeglijk naamwoord, enzovoort. Dit kan handig zijn voor het identificeren van de grammaticale structuur van een zin, wat op zijn beurt weer kan helpen bij het begrijpen van de betekenis ervan. POS tagging gebruikt Markov chains om te voorspellen welke rol een woord krijgt op basis van de rol van het vorige woord. Markov chains zijn wiskundige modellen die de waarschijnlijkheid voorspellen van een bepaalde gebeurtenis op basis van de voorgaande gebeurtenis. In het geval van POS tagging, wordt de voorgaande gebeurtenis gevormd door de vorige woorden in de zin en de waarschijnlijkheid van de huidige gebeurtenis (het taggen van het huidige woord) wordt bepaald door de waarschijnlijkheid van de voorgaande gebeurtenissen. Dit principe vormt de basis van hoe een Markov-model werkt. Een Hidden Markov Model (HMM) is een probabilistisch model dat wordt gebruikt in machine learning en dat kan worden toegepast op POS-tagging. Het model bestaat uit twee soorten toestanden: verborgen toestanden en waarnemingstoestanden. Verborgen toestanden zijn toestanden die niet direct waarneembaar zijn, zoals de grammaticale structuur van een zin. Waarnemingstoestanden daarentegen zijn de toestanden die

wel waarneembaar zijn, zoals de individuele woorden in een zin. Een HMM werkt door een reeks van waarnemingstoestanden te gebruiken om de verborgen toestanden te modelleren. Het model begint met een initiële verborgen toestand, waarbij elke toestand een bepaalde kans heeft om de volgende waarnemingstoestand te genereren. Na elke waarnemingstoestand wordt het model geüpdatet om de kans te berekenen dat de huidige verborgen toestand overgaat naar de volgende verborgen toestand. Deze kansen worden geschat op basis van de overgangskansen tussen de verborgen toestanden in de trainingsgegevens. Op deze manier leert het HMM om de verborgen toestanden te voorspellen op basis van de waarnemingstoestanden. Voor POS-tagging kan het HMM bijvoorbeeld leren welk woord meestal een bepaalde verborgen toestand (bijvoorbeeld een werkwoord) begeleidt, en hoe vaak deze verborgen toestand verandert in een andere toestand. In het kort is een HMM een model dat de relatie tussen verborgen toestanden en waarnemingstoestanden leert op basis van statistische informatie. Door gebruik te maken van de statistieken van waarnemingstoestanden, kan het model de meest waarschijnlijke verborgen toestanden voorspellen (Pykes, 2020).

In tegenstelling tot NER (Named Entity Recognition), dat zich richt op het identificeren van benoemde entiteiten in een tekst, richt POS tagging zich op de identificatie van de rol van woorden in een zin. Hoewel deze technieken verschillende doelen hebben, kunnen ze elkaar wel aanvullen en gebruikt worden in combinatie om meer geavanceerde NLP-taken uit te voeren.

2.2. Modellen voor het automatisch genereren van quizvragen

Er zijn over het algemeen twee typen modellen voor question generation: rule-based modellen en neural network-based modellen. Rule-based modellen maken gebruik van vooraf gedefinieerde sjablonen en regels om vragen te genereren, terwijl neural network-based modellen machine learning technieken gebruiken om te leren van trainingsdata en vragen te genereren. De neural network-based modellen kan je nog verder onderscheiden op basis van architectuur.

Encoder-decoder of seq2seq architecturen bestaan uit twee hoofdcomponenten: een encoder en een decoder. De context wordt gecodeerd door de encoder en vervolgens door de decoder gedecodeerd om de vragen te genereren.

Een andere neural network-based architectuur is een Generative Adversarial Network(GAN). GANs bestaan uit ook uit twee hoofdcomponenten: een generator en een discriminator. De generator genereert nieuwe vragen, terwijl de discriminator beoordeelt of de vragen gegenereert zijn of niet. Het trainen van een GAN-model

kan uitdagend zijn omdat de generator en discriminator elkaars prestaties beïnvloeden en het model kan vastlopen in een lokaal optimum.

Deep reinforcement learning is een andere benadering voor het genereren van vragen. In tegenstelling tot rule-based en neural network-based modellen, leert een deep reinforcement learning-model door middel van feedback van een omgeving. Het model probeert de beste actie te selecteren om een beloning te maximaliseren. In het geval van question generation kan de omgeving de inputtekst zijn en de acties kunnen het genereren van een vraag zijn. Het model krijgt feedback op basis van hoe goed de gegenereerde vraag overeenkomt met de gewenste vraag en past zijn interne parameters aan om in de toekomst betere vragen te genereren. Deep reinforcement learning-modellen hebben het voordeel dat ze in staat zijn om complexe patronen in de inputtekst te begrijpen en daaruit af te leiden welke vraag gesteld kan worden. Echter, de training van deze modellen vereist veel data en computationele kracht, wat kan leiden tot hoge kosten.

Joint question answering-question generation (QG-QA) benaderingen zijn ontworpen om zowel vragen te genereren als antwoorden te vinden op basis van een gegeven context. In deze aanpak worden QG- en QA-taken gecombineerd in één model, wat leidt tot een betere samenhang tussen de gegenereerde vragen en de antwoorden. Een van de belangrijkste voordelen van deze benadering is dat het de noodzaak voor extra informatiebronnen elimineert, omdat de antwoorden worden gevonden in de gegeven context. Bovendien kan de gegenereerde vraag worden gebruikt als een hulpmiddel om de relevantie en nauwkeurigheid van het gevonden antwoord te verifiëren. Een van de uitdagingen van de QG-QA-aanpak is het vinden van een geschikte balans tussen de kwaliteit van de gegenereerde vragen en de accuraatheid van de gevonden antwoorden. Het model moet in staat zijn om vragen te genereren die zowel relevant als begrijpelijk zijn, en tegelijkertijd antwoorden te vinden die correct en volledig zijn.

Transformers is een populaire keuze voor question generation taken vanwege hun vermogen om coherente en contextueel relevante vragen te genereren. Het belangrijkste voordeel van het gebruik van transformers is hun self-attention mechanisme, waardoor het model zich kan concentreren op de meest relevante delen van de invoertekst. Deze modellen hebben een encoder-decoder architectuur die specifiek ontworpen is om rekening te houden met de volgorde van woorden in een tekst. Het zijn state-of-the-art modellen in verschillende NLP-taken, zoals machinevertaling, samenvatting en question generation. Binnen de Transformers zijn er verschillende modellen die veelbelovend zijn voor question generation. T5 is een transformer-model dat speciaal is ontworpen voor taakgerichte NLP-taken, waaronder question generation. BART is een ander transformer-model dat gebruik maakt van een denoising autoencoder-architectuur en kan worden gebruikt voor verschil-

lende taalgeneratietaken, waaronder question generation. GPT-2 is een transformer-model dat bekend staat om zijn indrukwekkende prestaties in taalmodellering en het genereren van natuurlijke taal. Door deze transformers te gebruiken voor question generation, kan het model leren om vragen te genereren die grammaticaal correct zijn, semantisch relevant zijn en passen bij de context van de tekst (Mulla & Gharpure, 2023).

2.2.1. Transformers

In 2017 werd de ”Attention Is All You Need” paper gepubliceerd, die de Transformer-architectuur introduceerde. Het belangrijkste voordeel van de Transformer-architectuur is dat deze geen recurrente of convolutionele lagen nodig heeft om lange-termijnafhankelijkheden in een tekst te modelleren. Dit is mogelijk door middel van het self-attention mechanisme, waarmee de relatie tussen elk woord in de zin wordt berekend. In plaats van afhankelijkheden tussen woorden in een zin stapsgewijs te modelleren, berekent het self-attention mechanisme gecontextualiseerde representaties van elk invoertoken door aandacht te besteden aan alle andere tokens in de invoersequentie. Dit maakt de training van het model efficiënter en maakt het ook mogelijk om langere zinnen te modelleren dan bij traditionele sequentiële modellen.

De Transformer-architectuur is ontworpen om sequentiële invoer, zoals tekst, te verwerken door de invoersequentie om te zetten in een set continue representaties, die vervolgens kunnen worden gedecodeerd naar de uitvoersequentie. De self-attention mechanismen stellen de Transformer in staat om de informatie uit de hele invoersequentie te integreren in elke stap van de decodering, waardoor het model in staat is om betere voorspellingen te doen over het volgende woord of de volgende zin in de tekst.

Een andere belangrijke innovatie in de Transformer-architectuur is de introductie van een nieuwe trainingsdoelstelling genaamd ”masked language modeling”. Tijdens het trainingsproces worden willekeurig enkele invoertokens gemaskeerd, waarna het model wordt getraind om deze tokens te voorspellen op basis van de omringende context. Dit helpt het model om betere gecontextualiseerde representaties te leren voor elk woord in de tekst, aangezien het niet alleen afhankelijk is van de directe context van het woord, maar ook van de bredere context van de zin of het document.

Tot slot tonen de auteurs van de paper aan dat de Transformer-architectuur efficiënt kan worden getraind op grote datasets met behulp van parallelle computing, wat niet mogelijk was met eerdere recurrente of convolutionele architecturen. Dit heeft geleid tot een grote vooruitgang in de prestaties van automatische vertalingssystemen en andere taalgerelateerde taken (Vaswani e.a., 2017).

2.2.2. T5

Het T5 model is een pre-trained transformer-based nlp model. T5 maakt gebruik van transfer learning. Bij transfer learning wordt een model eerst getraind op een grote, algemene dataset om zo kennis en vaardigheden te ontwikkelen die relevant zijn voor een breed scala aan taken. Het model leert dan om bepaalde patronen en structuren in de ongelabelde data te herkennen. Vervolgens wordt het model gefinetuned op een kleinere, specifieke dataset die relevant is voor de beoogde taak. Door het model zo te trainen op een kleinere dataset, kan het model betere prestaties leveren op deze specifieke taak dan wanneer het vanaf nul zou worden getraind. T5 werd pre-trained op het c4 dataset die bestaat uit pagina's die opgehaald zijn via webscraping die vervolgens gefilterd werd op duplicaten, kwetsende inhoud en onvolledige zinnen. T5 krijgt de voorkeur als één van de NLP-technieken voor dit onderzoek omdat het gf gebruik maakt van niet gelabelde tekst. gf werkt met een relatief kleine dataset waardoor het een geschikte nlp-techniek kan zijn voor dit onderzoek. Elke taak die uitgevoerd wordt door het T5 model wordt omgezet in een tekst-naar-tekst formaat. Dit betekent dat de input en output als tekst worden weergegeven. Er worden prefixes toegevoegd aan de input om te specificeren welke soort taak het T5 model moet uitvoeren (Raffel e.a., 2019). Wegens het tekst-naar-tekst formaat kan het T5 model tijd besparen tijdens het ontwikkelingsproces (Rajapakse, 2020). Door een T5 model te trainen op data in de vorm van vraag-antwoord paren, kan het T5 model gebruikt worden als een QG-model.

2.2.3. BART

BART staat voor "Bidirectional and Auto-Regressive Transformer". Het model is ontwikkeld door Facebook AI Research en wordt beschouwd als een hybride van BERT en GPT, twee andere bekende transformer-modellen. BART maakt gebruik van zowel een encoder als een decoder, waarbij de encoder bidirectioneel is en de decoder autoregressief is. Dit betekent dat de encoder informatie kan verwerken vanuit beide richtingen van de input, terwijl de decoder één voor één tokens genereert in de output (Mohammed, 2022).

BART is getraind met een methode genaamd "denoising autoencoding", waarbij de inputdata wordt vervormd door ruis toe te voegen. Het BART-model wordt vervolgens getraind om de oorspronkelijke data te reconstrueren. Dit helpt het model om robuuster te worden en beter te presteren bij het genereren van tekst en het begrijpen van natuurlijke taal.

BART heeft bewezen een effectief model te zijn voor verschillende toepassingen, waaronder samenvattingen van teksten, vertalingen tussen talen en question generation. In het onderzoek waar BART voor het eerst werd geïntroduceerd, be-

haalde BART state-of-the-art prestaties op verschillende taaltaken, waaronder samenvattingen van teksten.

Kortom, BART is een veelzijdig transformer-model dat kan worden toegepast op verschillende taken op het gebied van NLP. De combinatie van een bidirectionele encoder en een autoregressieve decoder, samen met de denoising autoencoding training, maakt het model bijzonder effectief bij het genereren van tekst en het begrijpen van natuurlijke taal (Lewis e.a., 2019).

2.2.4. GPT-2

GPT-2, wat staat voor "Generative Pre-trained Transformer 2", is een taalmodel dat in 2019 werd geïntroduceerd door OpenAI. Het model is gebaseerd op het transformer-architectuur en heeft 1,5 miljard parameters, wat het op dat moment het grootste taalmodel ter wereld maakte. GPT-2 is ontworpen om een breed scala aan taaltaken aan te pakken, waaronder natuurlijke taal generatie, vraag-antwoord en machinevertaling.

Net als zijn voorganger, GPT-1, werd GPT-2 vooraf getraind op een enorme hoeveelheid ongelabelde tekst. Dit pre-trainingsproces, dat bekend staat als "unsupervised learning", stelt het model in staat om taalstructuren en -patronen te leren zonder specifieke labels of aanwijzingen.

GPT-2 is ontworpen als een "multitask" -model, wat betekent dat het in staat is om verschillende taaltaken aan te pakken zonder dat er een specifieke finetuning nodig is. Dit wordt bereikt door het gebruik van een zogenaamde "prompt-based" aanpak, waarbij het model wordt gepresenteerd met een prompt of een beginzin en vervolgens wordt gevraagd om de rest van de tekst te genereren. Het model past zich automatisch aan aan de aard van de taak op basis van de informatie die wordt gepresenteerd in de prompt.

Een opvallende functie van GPT-2 is het vermogen om zeer coherente en realistische tekst te genereren. Dit is te danken aan het gebruik van een techniek genaamd "n-grams", waarbij het model de waarschijnlijkheid van elk woord berekent op basis van de vorige n-woorden in de zin. Dit helpt het model om de context van de zin beter te begrijpen en meer coherente en natuurlijke tekst te genereren.

GPT-2 is getest op verschillende taaltaken en heeft indrukwekkende resultaten laten zien. Zo heeft het model bijvoorbeeld een relatief hoge score behaald op de bekende GLUE-benchmark, een taalmodelprestatietest die gebruikmaakt van een verscheidenheid aan taaltaken, waaronder vraag-antwoord, samenvatting en sentimentanalyse. Bovendien heeft GPT-2 ook laten zien dat het in staat is om zeer overtuigende nepnieuwsartikelen te genereren, wat heeft geleid tot enige bezorgdheid

over het gebruik van dergelijke modellen in de verkeerde handen.

Natuurlijke taalgeneratie is een van de belangrijkste toepassingen van GPT-2. Door gebruik te maken van de transformer-architectuur en pre-training op grote hoeveelheden ongelabelde tekstdata, kan GPT-2 op indrukwekkende wijze natuurlijk klinkende tekst genereren. Het model is in staat om coherent te schrijven en complexe structuren en relaties in de tekst te begrijpen en te reproduceren. Bovendien heeft GPT-2 aangetoond dat het een breed scala aan taaltaken aankan, zoals vertalingen, samenvattingen, vraaggeneratie en nog veel meer (Radford e.a., 2019).

2.3. Evaluatie van gegenereerde quizvragen

Een belangrijk aspect bij het genereren van quizvragen is de evaluatie van de kwaliteit van de gegenereerde vragen. Het is van cruciaal belang om de kwaliteit van de gegenereerde quizvragen te beoordelen om ervoor te zorgen dat de vragen geschikt zijn voor het beoogde publiek en het beoogde doel. Bovendien kan evaluatie helpen bij het verbeteren van de kwaliteit van het model en het aanpassen van de parameters.

Er zijn verschillende methoden voor het evalueren van gegenereerde quizvragen, waaronder automatische evaluatiemethoden, menselijke evaluatie en evaluatie door een Question Answering (QA) model. Automatische evaluatiemethoden zijn gebaseerd op het vergelijken van de gegenereerde quizvragen met referentievragen, en kunnen worden uitgevoerd met behulp van verschillende metrieken zoals BLEU, ROUGE en METEOR. Hoewel deze methoden snel en efficiënt zijn, kunnen ze beperkingen hebben in het evalueren van de semantische kwaliteit van de gegenereerde vragen.

Menselijke evaluatie is een andere methode die veel wordt gebruikt om de kwaliteit van gegenereerde quizvragen te beoordelen. Menselijke evaluatie maakt gebruik van beoordelaars die de kwaliteit van de gegenereerde quizvragen beoordelen op basis van verschillende criteria, zoals grammatica, relevantie en moeilijkheidsgraad. Menselijke evaluatie kan echter tijdrovend en duur zijn, vooral als grote aantallen quizvragen moeten worden geëvalueerd.

Evaluatie door een QA-model is een andere methode die wordt gebruikt om de kwaliteit van gegenereerde quizvragen te beoordelen. In dit geval wordt het QA-model gebruikt om de gegenereerde vragen te beantwoorden en te evalueren op basis van de nauwkeurigheid en relevantie van de gegenereerde antwoorden. Deze methode kan snel en efficiënt zijn en kan helpen bij het evalueren van de semantische kwaliteit van de gegenereerde quizvragen. Echter, deze methode heeft beperkingen, aangezien de prestaties van het QA-model van invloed kunnen zijn

op de evaluatie van de gegenereerde quizvragen.

Het evalueren van gegenereerde quizvragen is van cruciaal belang om de kwaliteit van de gegenereerde vragen te beoordelen en de prestaties van de gebruikte modellen te verbeteren. Er zijn verschillende methoden voor het evalueren van gegenereerde quizvragen, waaronder automatische evaluatie via metrics zoals BLEU, ROUGE, en METEOR, evenals menselijke evaluatie en evaluatie door Question Answering-modellen. Hoewel deze methoden nuttig kunnen zijn, blijft het moeilijk om de kwaliteit van gegenereerde quizvragen nauwkeurig te beoordelen, gezien de subjectiviteit van wat een ”goede” vraag is. Bovendien kunnen menselijke evaluatoren bevooroordeeld zijn en kunnen Question Answering-modellen de prestaties van de gegenereerde vragen mogelijk niet nauwkeurig weergeven. Desondanks blijft het belangrijk om evaluatiemethoden te blijven ontwikkelen en verfijnen om de kwaliteit van gegenereerde quizvragen te verbeteren en zo de potentie van question generation te maximaliseren.

2.3.1. BLEU

Een van de meest gebruikte evaluatiemethoden voor automatisch gegenereerde quizvragen is BLEU (Bilingual Evaluation Understudy). BLEU is oorspronkelijk ontwikkeld om de kwaliteit van machine vertalingen te meten, maar wordt tegenwoordig ook veelvuldig gebruikt voor het evalueren van de kwaliteit van gegenereerde tekst in het algemeen, waaronder gegenereerde quizvragen (Papineni e.a., 2002a).

Om de BLEU-score te berekenen, moeten we eerst de gegenereerde vragen en de referentie-vragen in n-gram sequenties opsplitsen. Een n-gram is een reeks van n opeenvolgende woorden in een tekst. Meestal worden 1-gram (ook wel unigram genoemd), 2-gram (bigram), 3-gram (trigram), enzovoort gebruikt. Vervolgens worden de n-gram sequenties van de gegenereerde vragen vergeleken met die van de referentie-vragen en wordt het aantal overeenkomende n-grams geteld.

Het resultaat van deze telling is de BLEU-score. Hoe hoger de BLEU-score, hoe hoger de overeenkomst tussen de gegenereerde vragen en de referentie-vragen. Een perfecte overeenkomst resulteert in een BLEU-score van 1, terwijl totaal geen overeenkomst een BLEU-score van 0 oplevert.

Stel dat we de volgende gegenereerde vraag hebben: ”Wat is de hoofdstad van Nederland?” en dat onze referentie-vraag is: ”Wat is de naam van de hoofdstad van Nederland?”. Om de BLEU-score te berekenen, zullen we eerst beide vragen opsplitsen in n-gram sequenties. Laten we 2-grams gebruiken als voorbeeld:

Gegenereerde vraag: [”Wat is”, ”is de”, ”de hoofdstad”, ”hoofdstad van”, ”van Nederland”]

Referentie-vraag: [”Wat is”, ”is de”, ”de naam”, ”naam van”, ”van de”, ”de hoofdstad”, ”hoofdstad van”, ”van Nederland”]

Nu tellen we het aantal overeenkomende 2-grams in de gegenereerde vraag en de referentie-vraag:

Overeenkomende 2-grams: [”Wat is”, ”is de”, ”de hoofdstad”, ”hoofdstad van”, ”van Nederland”]

De precisie van deze gegenereerde vraag is 100% omdat elk 2-gram in de gegenereerde vraag voorkomt in de referentie-vraag. Om de BLEU-score te berekenen, moeten we deze precisie-waarde normaliseren door de lengte van de gegenereerde vraag te gebruiken. Ook wordt de BLEU-score berekent voor verschillende individuele n-grammen en hiervan het geometrisch gemiddelde genomen. Meestal wordt er 1 tot en met 4-grammen genomen omdat dit het best aansluit met menselijke evaluatie:

$$\textit{BLEU}\text{-score} = \text{BP} \cdot \text{Geometrisch Gemiddelde van individuele n-grammen} \tag{2.1}$$

BP is de ’brevity penalty’ om rekening te houden met het feit dat kortere gegenereerde vragen meer kans hebben om hoge precisie-waarden te behalen. Als de lengte van de gegenereerde vraag korter is dan de lengte van de kortste referentie-vraag, wordt BP = 1 - (lengte van kortste referentie-vraag/lengte van gegenereerde vraag). Als de lengte van de gegenereerde vraag langer of even lang is dan de lengte van de kortste referentie-vraag, is BP gelijk aan 1. Bij het geometrisch gemiddelde worden de scores van de individuele n-grammen met elkaar vermenigvuldigd, waarna het resultaat tot de macht van 1/n wordt genomen.

Laten we het voorbeeld van de hoofdstad van Nederland nemen. Onze referentie-vraag heeft meer woorden dan onze gegenereerde vraag, dus BP is minder dan 1. Onze precisie-waarde is ook niet 100%, dus we hebben:

score 1-gram = 1

score 2-gram = 1

score 3-gram = 0.75

score 4-gram = 0.33

BP = $\exp(1\text{-}9/6) = 0.61$

$$\textit{BLEU}\text{-score} = 0.61 \cdot (1 \cdot 1 \cdot 0.75 \cdot 0.33)^{\frac{1}{4}} = 0.43 \tag{2.2}$$

Een BLEU-score tussen 0.6 en 0.7 wordt over het algemeen als optimaal beschouwd. Als de BLEU-score hoger is, bestaat het risico dat het question generation-model overfit en alleen de referentie-vragen overneemt (Doshi, 2021).

Hoewel BLEU een populaire evaluatiemethode is, heeft het ook enkele beperkingen. Zo houdt BLEU geen rekening met de semantische overeenkomst tussen de gegenereerde vragen en de referentie-vragen, wat kan leiden tot lage BLEU-scores voor gegenereerde vragen die wel degelijk van hoge kwaliteit zijn. Daarnaast is BLEU niet altijd geschikt voor het evalueren van vragen die inhoudelijk verschillen van de referentie-vragen. In deze gevallen kan het beter zijn om gebruik te maken van andere evaluatiemethoden, zoals ROUGE of METEOR, die rekening houden met de semantische overeenkomst tussen de gegenereerde vragen en de referentie-vragen.

In de praktijk wordt BLEU vaak gecombineerd met andere evaluatiemethoden en menselijke evaluatie om een vollediger beeld te krijgen van de kwaliteit van de gegenereerde quizvragen. Door het combineren van verschillende evaluatiemethoden kan een meer nauwkeurige en betrouwbare evaluatie van de gegenereerde vragen worden verkregen (Chiusano, 2022b).

2.3.2. ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is een familie van automatische evaluatiemethoden die zijn ontworpen om de kwaliteit van samenvattingen en andere generatieve tekst te beoordelen. Het doel van ROUGE is om de relevantie en dekking van een gegenereerde tekst ten opzichte van een referentietekst te meten (Lin, 2004).

ROUGE meet de overeenkomst tussen de gegenereerde en de referentie-tekst door gebruik te maken van verschillende n-gram-modellen. Net als bij BLEU.

We nemen als voorbeeld: Referentie-vraag: ”Ligt de vogel onder de kast?”

Gegenereerde vraag: ”Is de vogel onder de kast?”

Bij ROUGE-1 worden er unigrammen gebruikt om de ROUGE-score te bepalen. Het maakt gebruik van precisie, recall en F1-score om de ROUGE-score te bepalen. De precisie wordt bepaald door het aantal unigrammen die in de gegenereerde vraag en de referentie-vraag voorkomen te delen door het aantal unigrammen in de gegenereerde vraag. De recall wordt bepaald door het aantal unigrammen die in de gegenereerde vraag en de referentie-vraag voorkomen te delen door het aantal unigrammen in de referentie-vraag. In het gegeven voorbeeld zou de precisie, recall en F1-score het volgende zijn:

ROUGE-1 precisie = 5/6

ROUGE-1 recall = 5/6

ROUGE-1 F1-score = 2 * (precisie*recall) / (precisie + recall)

Als we de waarden gebruiken van het voorbeeld is de F1-score:

ROUGE-1 F1-score = 2 * (0.83*0.83)/(0.83+0.83) = 0.83

De zelfde manier van werken kan gebruikt worden voor ROUGE-2.

ROUGE-L is een ROUGE variant die de langste gemeenschappelijke subreeks tussen de gegenereerde tekst en de referentie-tekst meet. In tegenstelling tot ROUGE-1 en ROUGE-2, is ROUGE-L gevoeliger voor lange woordreeksen. ROUGE-L werkt op dezelfde manier als ROUGE-N, maar in plaats van het gebruik van n-gram, maakt het gebruik van de langste gemeenschappelijke subreeks.

ROUGE-S is een andere ROUGE-maatstaf die de semantische overeenkomst tussen de gegenereerde en referentie-tekst meet, in plaats van een letterlijke overeenkomst tussen woorden. Het kan opeenvolgende woorden die zijn gescheiden door één of meer woorden, terugvinden in de referentie-tekst. Bijvoorbeeld, in de zinnen ”De blauwe broek ligt op tafel” en ”De broek hangt op de stoel”, zou de 2-gram ”de broek” alleen als een gemeenschappelijk 2-gram worden aangemerkt in ROUGE-S.

ROUGE is een evaluatiemethode die positief correleert met menselijke evaluatie en die goedkoop en taalonafhankelijk is. Echter, ROUGE heeft ook enkele beperkingen waar rekening mee moet worden gehouden. ROUGE houdt bijvoorbeeld geen rekening met verschillende woorden die dezelfde betekenis hebben, omdat het syntactische overeenkomsten meet in plaats van semantiek. Dit kan leiden tot vertekende resultaten in evaluaties van teksten waarbij de inhoud en betekenis belangrijker zijn dan de exacte woordkeuze. Het is daarom belangrijk om zorgvuldig af te wegen wanneer ROUGE te gebruiken en welke andere evaluatiemethoden moeten worden gebruikt om een zo volledig mogelijk beeld te krijgen.

ROUGE legt de nadruk op recall, terwijl BLEU zich meer richt op precisie. Hoewel deze evaluatiemethoden verschillen, vullen ze elkaar aan in de resultaten die ze opleveren.

ROUGE kan worden gebruikt om verschillende generatieve teksten te evalueren, waaronder samenvattingen, machine vertalingen en chatbot-antwoorden. Net als bij BLEU is ROUGE echter niet perfect en kan het resulteren in onnauwkeurige evaluaties als gevolg van de complexiteit van natuurlijke taal. Het is daarom belangrijk om de resultaten van ROUGE te combineren met menselijke beoordelingen om een nauwkeurig beeld te krijgen van de kwaliteit van de gegenereerde tekst (Chiusano, 2022a).

2.3.3. METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is een automatische evaluatiemethode voor machine vertalingen die in 2005 werd voorgesteld. METEOR is ontworpen om de kwaliteit van machine vertalingen te meten door de overeenkomst tussen de machine vertaling en de referentie vertaling te berekenen. Het belangrijkste verschil tussen METEOR en andere evaluatiemethoden zoals BLEU en ROUGE is dat METEOR rekening houdt met de volgorde van woorden (Banerjee & Lavie, 2005a).

De reden waarom we de METEOR-metric nodig hebben, is dat de populaire BLEU-score, die wordt gebruikt voor machine vertaling, verschillende tekortkomingen heeft. Ten eerste is de BLEU-score meer gericht op precisie dan op recall. Ten tweede houdt de BLEU-score geen rekening met semantische overeenkomst. De BLEU-score zoekt naar exacte woord overeenkomsten, maar houdt geen rekening met het feit dat bijvoorbeeld 'auto' en 'voertuig' vergelijkbaar zijn. Ten derde is de BLEU-score niet goed in het opvangen van goede matches die clusters van woorden hebben die nauw overeenkomen met clusters van woorden in de referentie. Zoals bijvoorbeeld 'Wat is de omtrek van de aarde?' en 'Geef de omtrek van de aarde aan.' hebben weinig overlappende woorden maar de bedoeling van de zinnen is hetzelfde.

De METEOR-metric wordt berekend door eerst een uitlijning te berekenen tussen de gegenereerde tekst en de referentie, door woord-voor-woord overeenkomsten te zoeken of door hulpmiddelen voor vergelijkbaarheid te gebruiken, zoals woordembedding en woordenboeken. Een "chunk" in een uitlijning is een aangrenzende reeks woorden die overeenkomen met een aangrenzende reeks woorden in de referentie. METEOR houdt rekening met zowel de precisie als de recall bij het evalueren van de overeenkomst en berekent een F-score. Ten slotte wordt er een straf toegepast op basis van het aantal chunks in de kandidaattekst die overeenkomen met chunks in de referentie. De uiteindelijke METEOR-score combineert de F-score met de straf op de chunks om de totale score te berekenen (MLNerds, 2021).

2.3.4. Evaluatie door mensen

Het evalueren van de kwaliteit van gegenereerde vragen met behulp van technieken alleen is vaak niet effectief genoeg gebleken. Daarom is het van groot belang om menselijke evaluatietechnieken toe te passen om een betrouwbare beoordeling te krijgen. Het paper "Human evaluation of automatically generated text: Current trends and best practice guidelines" biedt richtlijnen voor menselijke evaluatietechnieken die gebruikt kunnen worden om de kwaliteit van gegenereerde vragen

te beoordelen.

Deze technieken kunnen variëren van het beoordelen van de grammaticale en semantische correctheid van de gegenereerde vraag, tot het beoordelen van de relevantie en bruikbaarheid ervan in een bepaalde context. Daarnaast kunnen menselijke beoordelaars vragen beoordelen op parameters als vloeiendheid, natuurlijkheid, moeilijkheidsgraad en originaliteit. Dit kan bijvoorbeeld gedaan worden door middel van het beoordelen van vragen op een schaal van 1 tot 5 op deze verschillende parameters.

Het is belangrijk om een gestandaardiseerde procedure te volgen bij het uitvoeren van menselijke evaluaties, om zo betrouwbare resultaten te verkrijgen. Dit omvat onder andere het selecteren van een representatieve steekproef van evaluatoren, het opstellen van duidelijke beoordelingscriteria en het trainen van de evaluatoren om consistente resultaten te krijgen.

Over het algemeen kan gesteld worden dat het gebruik van menselijke evaluatietechnieken een waardevolle aanvulling is op technische evaluatiemethoden om de kwaliteit van gegenereerde vragen te beoordelen. Het is belangrijk om te beseffen dat geen enkele evaluatiemethode perfect is, maar door verschillende methoden te combineren, kan een beter beeld verkregen worden van de kwaliteit van de gegenereerde vragen (van der Lee e.a., 2021).

2.3.5. Evaluatie door een question answering model

Een andere manier om de kwaliteit van gegenereerde vragen te evalueren, is door gebruik te maken van een question answering (QA) model. In plaats van te kijken naar de kwaliteit van de gegenereerde vragen op zichzelf, kan een QA-model worden gebruikt om een antwoord te genereren op de gegenereerde vraag. Als het gegenereerde antwoord overeenkomt met het juiste antwoord, kan de gegenereerde vraag als kwalitatief worden beschouwd.

Het gebruik van een QA-model voor evaluatie van gegenereerde vragen heeft enkele voordelen. Het zorgt voor een meer objectieve evaluatie van de kwaliteit van de gegenereerde vragen en stelt onderzoekers in staat om de prestaties van verschillende QG-modellen op een gestandaardiseerde manier te vergelijken. Bovendien biedt het gebruik van een QA-model de mogelijkheid om de relevantie van de gegenereerde vragen te evalueren door te kijken naar de nauwkeurigheid van de gegenereerde antwoorden.

Het is belangrijk op te merken dat het gebruik van een QA-model voor evaluatie van gegenereerde vragen niet zonder uitdagingen is. Zo kunnen QA-modellen soms moeite hebben met het genereren van juiste antwoorden op moeilijke vragen en

is het niet altijd mogelijk om de relevantie van een gegenereerde vraag alleen op basis van het antwoord te bepalen.

Ondanks deze uitdagingen kan het gebruik van een QA-model voor evaluatie van gegenereerde vragen een waardevolle toevoeging zijn aan de evaluatiemethoden voor QG-modellen.

3

Methodologie

Etiam pede massa, dapibus vitae, rhoncus in, placerat posuere, odio. Vestibulum luctus commodo lacus. Morbi lacus dui, tempor sed, euismod eget, condimentum at, tortor. Phasellus aliquet odio ac lacus tempor faucibus. Praesent sed sem. Praesent iaculis. Cras rhoncus tellus sed justo ullamcorper sagittis. Donec quis orci. Sed ut tortor quis tellus euismod tincidunt. Suspendisse congue nisl eu elit. Aliquam tortor diam, tempus id, tristique eget, sodales vel, nulla. Praesent tellus mi, condimentum sed, viverra at, consectetuer quis, lectus. In auctor vehicula orci. Sed pede sapien, euismod in, suscipit in, pharetra placerat, metus. Vivamus commodo dui non odio. Donec et felis.

Etiam suscipit aliquam arcu. Aliquam sit amet est ac purus bibendum congue. Sed in eros. Morbi non orci. Pellentesque mattis lacinia elit. Fusce molestie velit in ligula. Nullam et orci vitae nibh vulputate auctor. Aliquam eget purus. Nulla auctor wisi sed ipsum. Morbi porttitor tellus ac enim. Fusce ornare. Proin ipsum enim, tincidunt in, ornare venenatis, molestie a, augue. Donec vel pede in lacus sagittis porta. Sed hendrerit ipsum quis nisl. Suspendisse quis massa ac nibh pretium cursus. Sed sodales. Nam eu neque quis pede dignissim ornare. Maecenas eu purus ac urna tincidunt congue.

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

Maecenas non massa. Vestibulum pharetra nulla at lorem. Duis quis quam id lacus dapibus interdum. Nulla lorem. Donec ut ante quis dolor bibendum condimentum. Etiam egestas tortor vitae lacus. Praesent cursus. Mauris bibendum pede at elit. Morbi et felis a lectus interdum facilisis. Sed suscipit gravida turpis. Nulla at lectus. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Praesent nonummy luctus nibh. Proin turpis nunc, congue eu, egestas ut, fringilla at, tellus. In hac habitasse platea dictumst.

Vivamus eu tellus sed tellus consequat suscipit. Nam orci orci, malesuada id, gravida nec, ultricies vitae, erat. Donec risus turpis, luctus sit amet, interdum quis, porta sed, ipsum. Suspendisse condimentum, tortor at egestas posuere, neque metus tempor orci, et tincidunt urna nunc a purus. Sed facilisis blandit tellus. Nunc risus sem, suscipit nec, eleifend quis, cursus quis, libero. Curabitur et dolor. Sed vitae sem. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Maecenas ante. Duis ullamcorper enim. Donec tristique enim eu leo. Nullam molestie elit eu dolor. Nullam bibendum, turpis vitae tristique gravida, quam sapien tempor lectus, quis pretium tellus purus ac quam. Nulla facilisi.

4

Conclusie

Curabitur nunc magna, posuere eget, venenatis eu, vehicula ac, velit. Aenean ornare, massa a accumsan pulvinar, quam lorem laoreet purus, eu sodales magna risus molestie lorem. Nunc erat velit, hendrerit quis, malesuada ut, aliquam vitae, wisi. Sed posuere. Suspendisse ipsum arcu, scelerisque nec, aliquam eu, molestie tincidunt, justo. Phasellus iaculis. Sed posuere lorem non ipsum. Pellentesque dapibus. Suspendisse quam libero, laoreet a, tincidunt eget, consequat at, est. Nullam ut lectus non enim consequat facilisis. Mauris leo. Quisque pede ligula, auctor vel, pellentesque vel, posuere id, turpis. Cras ipsum sem, cursus et, facilisis ut, tempus euismod, quam. Suspendisse tristique dolor eu orci. Mauris mattis. Aenean semper. Vivamus tortor magna, facilisis id, varius mattis, hendrerit in, justo. Integer purus.

Vivamus adipiscing. Curabitur imperdiet tempus turpis. Vivamus sapien dolor, congue venenatis, euismod eget, porta rhoncus, magna. Proin condimentum pretium enim. Fusce fringilla, libero et venenatis facilisis, eros enim cursus arcu, vitae facilisis odio augue vitae orci. Aliquam varius nibh ut odio. Sed condimentum condimentum nunc. Pellentesque eget massa. Pellentesque quis mauris. Donec ut ligula ac pede pulvinar lobortis. Pellentesque euismod. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent elit. Ut laoreet ornare est. Phasellus gravida vulputate nulla. Donec sit amet arcu ut sem tempor malesuada. Praesent hendrerit augue in urna. Proin enim ante, ornare vel, consequat ut, blandit in, justo. Donec felis elit, dignissim sed, sagittis ut, ullamcorper a, nulla. Aenean pharetra vulputate odio.

Quisque enim. Proin velit neque, tristique eu, eleifend eget, vestibulum nec, lacus. Vivamus odio. Duis odio urna, vehicula in, elementum aliquam, aliquet laoreet, tel-

lus. Sed velit. Sed vel mi ac elit aliquet interdum. Etiam sapien neque, convallis et, aliquet vel, auctor non, arcu. Aliquam suscipit aliquam lectus. Proin tincidunt magna sed wisi. Integer blandit lacus ut lorem. Sed luctus justo sed enim.

Morbi malesuada hendrerit dui. Nunc mauris leo, dapibus sit amet, vestibulum et, commodo id, est. Pellentesque purus. Pellentesque tristique, nunc ac pulvinar adipiscing, justo eros consequat lectus, sit amet posuere lectus neque vel augue. Cras consectetuer libero ac eros. Ut eget massa. Fusce sit amet enim eleifend sem dictum auctor. In eget risus luctus wisi convallis pulvinar. Vivamus sapien risus, tempor in, viverra in, aliquet pellentesque, eros. Aliquam euismod libero a sem.

Nunc velit augue, scelerisque dignissim, lobortis et, aliquam in, risus. In eu eros. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Curabitur vulputate elit viverra augue. Mauris fringilla, tortor sit amet malesuada mollis, sapien mi dapibus odio, ac imperdiet ligula enim eget nisl. Quisque vitae pede a pede aliquet suscipit. Phasellus tellus pede, viverra vestibulum, gravida id, laoreet in, justo. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer commodo luctus lectus. Mauris justo. Duis varius eros. Sed quam. Cras lacus eros, rutrum eget, varius quis, convallis iaculis, velit. Mauris imperdiet, metus at tristique venenatis, purus neque pellentesque mauris, a ultrices elit lacus nec tortor. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent malesuada. Nam lacus lectus, auctor sit amet, malesuada vel, elementum eget, metus. Duis neque pede, facilisis eget, egestas elementum, nonummy id, neque.

A

Onderzoeksvoorstel

Het onderwerp van deze bachelorproef is gebaseerd op een onderzoeksvoorstel dat vooraf werd beoordeeld door de promotor. Dat voorstel is opgenomen in deze bijlage.

A.1. Introductie

Quizzen zijn een handig middel om kennis op te frissen en te testen. Wanneer de hoeveelheid informatie echter blijft groeien, wordt het handmatig creëren van quizzen onhaalbaar. Een dergelijk probleem wordt ervaren bij InfoS met hun gf. Het gf dient als een verzamelpunt voor kennis die InfoS verkrijgt tijdens het werken aan meerdere projecten in de IT-consultancy wereld. InfoS wil deze kennis gamificeren om zo de kennis van hun medewerkers te testen en op te frissen. Nieuwe medewerkers krijgen de mogelijkheid tot het raadplegen van bestaande kennis in het gf om zo de leercurve op een laagdrempelige manier te verbeteren.

Een mogelijke oplossing voor het probleem die InfoS ervaart is het creëren van een Proof of Concept (PoC) voor een quiz generator die aan de hand van nlp technieken verschillende soorten vragen kan genereren over de informatie aanwezig in het gf. De quiz generator moet in staat zijn om vragen te genereren die de essentie van de onderwerpen in het gf toetsen. In deze thesis wordt onderzocht hoe verschillende soorten vragen kunnen worden gegenereerd en hoe deze vragen kunnen worden beoordeeld op relevantie. Er zal samengewerkt worden met een andere student, die zich zal focussen op het verzamelen en verwerken van de data die nodig is voor de quizgenerator.

Tijdens dit onderzoeksproces zullen we antwoorden formuleren op de volgende vragen:

- Welke nlp modellen kunnen toegepast worden om automatisch vragen te genereren op basis van IT-gerelateerde kennis in het Guidance Framework van Info Support?
- Hoe kunnen we de gegenereerde vragen het best evalueren?

A.2. State-of-the-art

A.2.1. Question generation

Het genereren van vragen met behulp van nlp-technieken is een uitdaging op het gebied van kunstmatige intelligentie. De nlp-technieken die gebruikt worden om vragen te genereren verschillen afhankelijk van het type vraag.

NLU (Natural Language Understanding) en NLG (Natural Language Generation) zijn twee belangrijke onderdelen van nlp. NLU houdt zich bezig met het begrijpen van natuurlijke taal, terwijl NLG zich richt op het automatisch genereren van taal (Kavlakoglu, 2020). Een toepassing van NLU en NLG is Question Generation (QG). QG is een technologie die automatisch vragen kan genereren op basis van een gegeven tekst. Hieruit volgt dat QG gebruikt kan worden bij het efficiënt creëren van vragen voor educatieve en trainingsdoelen zoals het maken van quizzen en testen.

Text-to-Text Transfer Transformer (T5)

Het T5 model is een pre-trained transformer-based nlp model. T5 maakt gebruik van transfer learning. Dit is een proces waarbij het nlp model eerst pre-training doet op een niet gelabelde dataset. T5 krijgt de voorkeur als één van de NLP-technieken voor dit onderzoek omdat het gf gebruik maakt van niet gelabelde tekst. T5 werd pre-trained op het c4 dataset die bestaat uit pagina's die opgehaald zijn via web-scraping die vervolgens gefilterd werd op duplicaten, kwetsende inhoud en onvolledige zinnen. Een gevolg hiervan is dat er hier kan getraind worden op een kleinere niet gelabelde dataset om een specifieke taak te kunnen uitvoeren. gf werkt met een relatief kleine dataset waardoor het een geschikte nlp-techniek kan zijn voor dit onderzoek. Elke taak die uitgevoerd wordt door het T5 model wordt omgezet in een tekst-naar-tekst formaat. Dit betekent dat de input en output als tekst worden weergegeven. Er worden prefixes toegevoegd aan de input om te specificeren welke soort taak het T5 model moet uitvoeren (Raffel e.a., 2019). Wegens het tekst-naar-tekst formaat kan het T5 model tijd besparen tijdens het ontwikkelingsproces (Rajapakse, 2020). Door een T5 model te trainen op data in de vorm van

vraag-antwoord paren, kan het T5 model gebruikt worden als een QG-model.

Bidirectional and Auto-Regressive Transformer (BART)

Het BART model wordt beschouwd als autoregressief en bidirectioneel. Hier worden de bidirectionele BERT en de autoregressive GPT gecombineerd (Mohammed, 2022). Bij de pre-training wordt noise gebruikt om de trainingsdata aan te passen op een willekeurige manier. Vervolgens moet het BART model de trainingsdata reconstrueren. Het BART model is geschikt voor text generation en comprehesion tasks zoals QG (Lewis e.a., 2019).

Meerkeuzevraag

Voor een meerkeuzevraag te genereren kan je 2 verschillende T5 modellen trainen. Het eerste T5 model wordt getraind op een dataset van vraag-antwoord paren. Dit model accepteert een context en een antwoord om daarmee een meerkeuzevraag te genereren. Het tweede T5 model wordt gebruikt om distractors te genereren. Distractors zijn de opties bij een meerkeuzevraag die foutief zijn. Het wordt getraind op een dataset van meerkeuzevragen en per vraag verschillende opties. Dit model accepteert de context, meerkeuzevraag en juiste antwoord als input. Als output geeft het een paar distractors (Vachev e.a., 2022).

Waar/onwaar-vraag

Bij een waar/onwaar-vraag vraag wordt een stelling gegeven die met Waar of Onwaar moet worden beantwoord. Vragen die waar zijn kunnen via abstractieve of extractive summarization opgehaald worden. Wordnet kan ook in dit geval gebruikt worden om vragen die onwaar zijn te genereren. Het genereren van een onwaarvraag kan op verschillende manieren. Hieronder worden een paar gegeven:

- Veranderen van benoemde entiteit (bv. Caesar naar Nero)
- Het toevoegen of verwijderen van een negatie (bv. Het verandert ..., het verandert niet ...)
- Veranderen van bijvoeglijk naamwoord (bv. grootste naar kleinste)
- Veranderen van werkwoord (bv. aantrekken naar wegduwen)

Invulvraag

Een invulvraag bestaat uit een zin waarbij een woord of een stuk van de zin wordt weggelaten om vervolgens het juiste woord te kunnen vullen. Één van de mogelijkheden om dit proces automatisch te laten verlopen is door de Python Keyword

Extraction Library te gebruiken. Het proces begint met het afleiden van sleutelwoorden uit een tekst. Deze library heeft de keuzes om supervised of unsupervised modellen te gebruiken. Er wordt gebruik gemaakt van een unsupervised model omdat er geen gelabelde data beschikbaar is in dit geval (AIEngineering, 2021). De sleutelwoorden kunnen vervolgens weggelaten worden en gebruikt worden als invulvragen.

Combinatievraag

Bij een combinatievraag krijg je een aantal woorden en definities waarbij je ze juist met elkaar moet verbinden. Voor het genereren van combinatievragen wordt bijna hetzelfde proces doorlopen als bij een invulvraag. Er wordt opnieuw gebruik gemaakt van het Python Keyword Extraction Library om de sleutelwoorden in een gegeven tekst te bepalen. Een probleem hierbij is dat sommige sleutelwoorden verschillende betekenissen kunnen hebben. Bijvoorbeeld een muis kan een computermuis zijn of een dier. De juiste keuze kan worden gemaakt aan de hand van het BERT WSD model waarbij het woord in kwestie wordt geanalyseerd aan de hand van de meegegeven context (AIEngineering, 2021).

A.2.2. Gamification

Badges, punten en rapporten kunnen als beloning dienen voor het bereiken van bepaalde doelen of het voltooien van specifieke taken door gebruikers bij het gebruiken vaan een educatieve spel. Het gebruik van dergelijke gamificatie technieken kan leiden tot betere prestaties van studenten (Smiderle e.a., 2020). Het kan interessant zijn om soortgelijke gamificatie technieken toe te passen in een automatische quiz generator.

Duolingo is een educatief spel dat gebruikers in staat brengt om een nieuwe taal te leren of een bestaande taalkennis uit te breiden of testen. Duolingo maakt gebruik van een gamificatie techniek namelijk ranking (Smiderle e.a., 2020), zoals bijvoorbeeld een leaderboard. Een leaderbord heeft als doelstelling gebruikers te motiveren op een competitieve wijze om een taal te leren. Een gelijkaardige systeem kan toegepast worden in een automatische quiz generator. Door de totaalscore te berekenen op basis van het aantal correcte antwoorden op de quizvragen per gebruiker kan een leaderboard worden opgesteld.

Een voorbeeld van een succesvol uitgevoerde software dat gebruik maakt van machine learning en gamification technieken is de chatbot Mya. Mya is een AI gedreven chatbot die als doelstelling heeft bedrijven te helpen met de werving van nieuwe talent voor hun organisatie. De chatbot gebruikt verschillende gamificatie technieken om de gebruikers betrokken en geïnteresseerd te houden tijdens

het sollicitatie-proces. Dergelijke technieken zijn onder andere het belonen van de gebruikers door punten uit te delen bij het voltooien van bepaalde taken zoals het indienen van CV 's of het plannen van sollicitatiegesprekken (Jamal, 2022). Mya zoals andere chatbots maakt gebruik van een nlp-techniek namelijk Named Entity Recognition (NER). NER wordt gebruikt in machine learning om specifieke entiteiten zoals namen, locaties, functietitels uit bijvoorbeeld gebruikersberichten te identificeren (Jurafsky & Martin, 2009) om de inhoud van een bericht te begrijpen om achteraf een nauwkeuriger antwoord te genereren voor een gebruiker.

A.2.3. NLP-evaluatie scores voor een AI model

Om de kwaliteit van een machine-vertaling te kunnen bepalen hebben we een objectief meetsysteem nodig. De kwaliteit van een vertaling kan aan de hand van de BLEU-, ROUGE- of METEOR-score gemeten worden. De scores worden gebruikt bij QG technieken.

BLEU score berekenen is een rekentechniek die gebruikt wordt bij nlp om te beoordelen hoe betrouwbaar een automatisch vertaal systeem presteert. Het is een criterium voor het meten van de nauwkeurigheid van een bepaalde machine-vertaling ten opzichte van een manuele vertaling van een tekst door een persoon. De BLEU score is gebaseerd op het aantal overeenkomende n-grammen (reeksen van n woorden) tussen de machine-vertaling en de referentie tekst. Hoe hoger de BLEU score hoe accurater de machine-vertaling van een referentie tekst. De BLEU score is een cijfer tussen 0 en 1 (Papineni e.a., 2002b).

METEOR (Metric for Evaluation of Translation with Explicit Ordering) metriek werd ontwikkeld om de tekortkomingen van de BLEU score aan te pakken. Zoals de BLEU score is de METEOR methode is een maatstaf voor het evalueren van de kwaliteit van een machine-vertaling. METEOR beoordeelt vertalingen door ze te vergelijken met één of meer referentie vertalingen, waarbij gebruik wordt gemaakt van expliciete woord-tot-woord overeenkomsten. Om dit te kunnen bereiken wordt er een afstemming tussen de machine-vertaling en de referentie vertaling gecreëerd door middel van een proces in meerdere fasen. Het doel van de uitlijning is om unigram van elke string zodanig in kaart te brengen dat elk unigram overeenkomt met nul of één unigram in de andere string en geen enkele in dezelfde string. Ten slotte wordt de beste score gerapporteerd voor de gegeven vertaling wanneer deze wordt vergeleken met elke referentie op een onafhankelijke manier. Het doel van de methode is om een nauwkeurigere evaluatie te geven van de kwaliteit van door een machine gegenereerde vertalingen ten opzichte tot andere methodologieën zoals bijvoorbeeld de BLEU score, die niet expliciet rekening houdt met de woordvolg-

orde en semantische verschillen tussen de door de machine gegenereerde vertaling en de referentie vertaling (Banerjee & Lavie, 2005b).

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is een maatstaf die gebruikt wordt voor het evalueren van de kwaliteit van samenvattende generatiemodellen (summary generation models). Om de overlap tussen door de machine gegenereerde samenvattingen en referentie-samenvattingen te meten, worden de precisie-, recall- en F1-scores berekend voor verschillende N-gram groottes (Papineni e.a., 2002b).

A.3. Methodologie

A.3.1. Modellen en evaluatie scores definiëren

In deze stap worden de modellen gedefinieerd die gebruikt zullen worden om vragen te genereren op basis van de beschikbare IT-gerelateerde kennis in het Guidance Framework van InfoS. Mogelijke modellen zijn bijvoorbeeld seq2seq-modellen, transformer-modellen en neurale netwerken. Het is belangrijk om verschillende modellen te overwegen en hun eigenschappen te vergelijken om het meest geschikte model te kiezen.

Daarnaast worden de evaluatie scores gedefinieerd om de gegenereerde vragen te beoordelen op kwaliteit. Voorbeelden van evaluatie scores zijn BLEU, ROUGE en METEOR, die gebruikt worden om te beoordelen hoe goed de gegenereerde vragen overeenkomen met referentievragen.

A.3.2. Modellen ontwikkelen en trainen

In deze fase zal er samengewerkt worden met een andere student die verantwoordelijk is voor het verzamelen en verwerken van de datasets die nodig zijn voor de quizgenerator. Het doel is om verschillende NLP-modellen te ontwikkelen en te trainen op de verzamelde datasets, om zo vragen te genereren over IT-gerelateerde kennis in het Guidance Framework van InfoS. Het uiteindelijke doel van deze stap is om modellen te creëren die in staat zijn om diverse soorten vragen te genereren over de informatie in het Guidance Framework.

A.3.3. Evaluatie modellen

Bij deze stap zal de prestaties van de ontwikkelde modellen beoordeeld worden en vergeleken worden aan de hand van de eerder gedefinieerde evaluatiescores. Er zal ook gekeken worden naar de interpretatie van de resultaten om te bepalen of er

aanpassingen nodig zijn in de modellen. Er zal in deze fase ook aandacht besteed worden aan de robuustheid van de modellen, oftewel hoe goed de modellen presteren op nieuwe en onbekende vragen die van de testset komen. De resultaten van deze fase zullen ons helpen bij het selecteren van het best presterende model om te implementeren in de Proof of Concept.

A.3.4. Analyse resultaten

In deze fase worden de resultaten van de evaluatie van de modellen geanalyseerd en geïnterpreteerd. Er wordt gekeken naar hoe goed de verschillende modellen presteren op de verschillende soorten vragen en contexten. Daarnaast wordt er gekeken naar de verschillende evaluatiemethoden en -scores en wordt er bepaald welke methoden het meest geschikt zijn voor de analyse van de resultaten. De resultaten worden geïnterpreteerd en de sterke en zwakke punten van de verschillende modellen worden geïdentificeerd. Op basis hiervan worden conclusies getrokken en aanbevelingen gedaan voor de ontwikkeling van de quiz generator.

A.4. Verwacht resultaat, conclusie

Uit dit onderzoek moet blijken welke modellen het meest geschikt zijn voor een quiz generator voor het InfoS gf. Door middel van evaluatiecriteria zoals BLEU, ROUGE en METEOR scores en feedback van medewerkers van InfoS, kan er een duidelijke ranking gemaakt worden van de gegenereerde vragen.

Door middel van de ontwikkelde quiz generator kunnen de medewerkers van InfoS op een laagdrempelige manier hun kennis testen en verbeteren. Dit zorgt voor een meerwaarde voor InfoS omdat medewerkers met up-to-date kennis beter in staat zijn om klanten te helpen en projecten uit te voeren.

Bibliografie

AIEngineering. (2021, maart 6). *Question Generation Using Natural Language Processing*. QuestGen. Verkregen 12 maart 2023, van https://www.youtube.com/watch?v=hoCi_bJHyb8&ab_channel=AIEngineering

Analytics, E. (2022). Uncover Hidden Insights: Advanced Named Entity Recognition. https://www.expressanalytics.com/blog/what-is-named-entity-recognition-ner-benefits-use-cases-algorithms/

Banerjee, S., & Lavie, A. (2005a). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. https://aclanthology.org/W05-0909

Banerjee, S., & Lavie, A. (2005b). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. https://aclanthology.org/W05-0909

Chiusano, F. (2022a). Two minutes NLP — Learn the ROUGE metric by examples. https://medium.com/nlplanet/two-minutes-nlp-learn-the-rouge-metric-by-examples-f179cc285499

Chiusano, F. (2022b). Two minutes NLP — Learn the BLEU metric by examples. https://medium.com/nlplanet/two-minutes-nlp-learn-the-bleu-metric-by-examples-df015ca73a86

Doshi, K. (2021). Foundations of NLP Explained — Bleu Score and WER Metrics. https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b

Jamal, A. (2022). Mya Systems – "Using conversational AI to solve talent acquisition challenges". https://d3.harvard.edu/platform-digit/submission/mya-systems/

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall. http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y

Kavlakoglu, E. (2020). NLP vs. NLU vs. NLG: the differences between three natural language processing concepts. Verkregen 14 februari 2023, van https://www.

ibm.com/blogs/watson/2020/11/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts/

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. https://doi.org/10.48550/ARXIV.1910.13461

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013

MLNerds. (2021). METEOR metric for machine translation. https://machinelearninginterview.com/topics/machine-learning/meteor-for-machine-translation/

Mohammed, A. (2022). What is BART model in transformers? Verkregen 13 februari 2023, van https://www.projectpro.io/recipes/what-is-bart-model-transformers

Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, *12*(1), 1–32. https://doi.org/10.1007/s13748-023-00295-9

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002a). Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002b). Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Pykes, K. (2020). Part Of Speech Tagging for Beginners. https://towardsdatascience.com/part-of-speech-tagging-for-beginners-3a0754b2ebba

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. https://doi.org/10.48550/ARXIV.1910.10683

Rajapakse, T. (2020). Asking the Right Questions: Training a T5 Transformer Model on a New task. Verkregen 12 februari 2023, van https://towardsdatascience.com/asking-the-right-questions-training-a-t5-transformer-model-on-a-new-task-691ebba2d72c

Smiderle, R., Rigo, S. J., Marques, L. B., de Miranda Coelho, J. A. P., & Jaques, P. A. (2020). The impact of gamification on students' learning, engagement and

behavior based on their personality traits. *Smart Learning Environments*, *7*(1). https://doi.org/10.1186/s40561-019-0098-x

Steuer, T., Filighera, A., Tregel, T., & Miede, A. (2022). Educational Automatic Question Generation Improves Reading Comprehension in Non-native Speakers: A Learner-Centric Case Study. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.900304

Vachev, K., Hardalov, M., Karadzhov, G., Georgiev, G., Koychev, I., & Nakov, P. (2022). Leaf: Multiple-Choice Question Generation. https://doi.org/10.48550/ARXIV.2201.09012

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech Language*, *67*, 101151. https://doi.org/https://doi.org/10.1016/j.csl.2020.101151

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. https://arxiv.org/abs/1706.03762

Walia, A. (2017). Annotated Corpus for Named Entity Recognition. https://www.kaggle.com/datasets/abhinavwalia95/entity-annotated-corpus

Zhang, R., Guo, J., Chen, L., Fan, Y., & Cheng, X. (2021). A Review on Question Generation from Natural Language Text. *ACM Trans. Inf. Syst.*, *40*(1). https://doi.org/10.1145/3468889