Ricardo Saraiva Grava

Aprimoramento de Agente Conversacional Especializado em Amazônia Azul

São Paulo, SP

2023

Ricardo Saraiva Grava

Aprimoramento de Agente Conversacional Especializado em Amazônia Azul

Trabalho de conclusão de curso apresentado ao Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro.

Universidade de São Paulo – USP

Escola Politécnica

Departamento de Engenharia de Computação e Sistemas Digitais (PCS)

Orientador: Profa. Dra. Anarosa Alves Franco Brandão

São Paulo, SP

2023

Catalogação-na-publicação

Grava, Ricardo
Aprimoramento de Agente Conversacional Especializado em Amazônia Azul / R. Grava -- São Paulo, 2023.
54 p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Question Answering 2.Question Generation 3.Transformers 4.Processamento de Linguagem Natural 5.Aprendizado de máquina I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Dedico este trabalho ao meu irmão Guilherme, meu melhor amigo, escritor favorito e inspiração de uma vida inteira. Aos meus pais pelo incansável apoio por todos estes anos. Aos meus amigos Pedro e Gabriel, pelas incontáveis horas de Rock and Stone que me mantiveram são durante a elaboração deste projeto. Ao meu amigo Steve, por nossa longa amizade, forjada com o fogo da curiosidade e o ferro do conhecimento. Ao meu amigo Eduardo, que me acompanhou nesta jornada a cada passo do caminho. Por fim, dedico a obra a todos os artistas, músicos, escritores, dentre outras mentes criativas que me fizeram quem sou hoje.

Agradecimentos

À professora Anarosa, por sua orientação e estímulo à pesquisa. Este trabalho foi realizado com o apoio do Itaú Unibanco S.A., por meio do Programa de Bolsas Itaú (PBI), vinculado ao Centro de Ciência de Dados da Escola Politécnica da Universidade de São Paulo.

"Nos tempos mais sombrios,
esperança é algo que você dá a si mesmo.
Este é o significado de força interior."
- Iroh

Resumo

Este projeto de pesquisa trata do aprimoramento do DEEPAGÉ, um sistema de *question answering* baseado em *transformers* que será utilizado no núcleo de um agente conversacional especializado em Amazônia Azul, o BLAB, do Centro de Inteligência Artificial (C4AI). Em sua forma original, o DEEPAGÉ apresenta algumas falhas, ocasionando respostas inconsistentes. Este projeto objetivou a minimização de tais falhas a partir de um segundo treinamento, utilizando um *dataset* expandido. Tipicamente, para o treinamento destes respondedores, são necessários pares de pergunta e resposta cuja geração é demorada e dispendiosa. Foram utilizadas perguntas automaticamente geradas para a automatização desta tarefa, eliminando a necessidade de intervenção humana e agilizando o processo. Para este fim, foi criado um outro modelo, o PTT5-QG, capaz de gerar perguntas a partir de textos base em língua portuguesa, para aproveitar a literatura acerca da Amazônia Azul escrita em português. Como resultado, disponibilizou-se uma nova versão do DEEPAGÉ, mais assertivo em suas respostas e que pode ser integrada ao *chatbot* BLAB. O sistema gerador de *datasets* foi avaliado por métricas automáticas, pela quantidade de questões que gera e por seu desempenho na melhoria do DEEPAGÉ. Ambos os modelos criados também foram submetidos a testes manuais por parte de avaliadores humanos, que indicaram bons resultados em diveros critérios estipulados.

Palavras-chave: Chatbot. Agente conversacional. Geração de questões.

Abstract

This research project deals with the enhancement of DEEPAGÉ, a transformer based question answering system which will be used at the core of a conversational agent specialized in the Blue Amazon, BLAB, of the Center for Artificial Intelligence (C4AI). In its original form, DEEPAGÉ displays some flaws, resulting in inconsistent answers. This project aims to minimize such flaws through a second training process, using an expanded dataset. Typically, to train these answerers, question and answer pairs are necessary, which are expensive and time-consuming to generate. Automatically generated questions were used in order to automate this task, eliminating the necessity of human intervention and streamlining the process. For this purpose, an intermediate model was created, capable of generating questions from portuguese texts, to take advantage of the existing literature about the Blue Amazon in portuguese. As a result, a new version of DEEPAGÉ was made available, which is more assertive in its responses and can be integrated with the BLAB chatbot. The dataset generating system was evaluated using automatic metrics, by the amount of questions it produces and its performance in enhancing DEEPAGÉ. Both of the models created were also tested manually by human annotators, which indicated good results in several stipulated criteria.

Keywords: Chatbot. Conversational agent. Question generation.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

QG	*Question Generation*
QA	*Question Answering*
BLAB	*Blue Amazonia Brain*
T5	*Text To Text Transfer Transformer*
PTT5	*Portuguese T5*
SQuAD	*Stanford Question Answering Dataset*
BERT	*Bidirectional Encoder Representations from Transformers*
PAQ	*Probably Asked Questions*
KEML	*Knowledge Enhanced Machine Learning*
AQPL	*All Questions Per Line*
OQPL	*One Question Per Line*
BrWac	*Brazilian Portuguese Web as Corpus*
API	*Application Programming Interface*
NLP	*Natural Language Processing*
JSON	*Javascript Object Notation*
JSONL	*Javascript Object Notation Lines*
IA	Inteligência Artificial
TyDi QA	*Typologically Diverse Question Answering*
Assin	Avaliação de Similaridade Semântica e Inferência textual
HAREM	HAREM é uma Avaliação de Reconhecedores de Entidades Mencionadas
BLEU	*Bilingual Evaluation Understudy*
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*
METEOR	*Metric for Evaluation of Translation with Explicit Ordering*

Lista de símbolos

C: *Context* (contexto)

A: *Answer* (resposta)

Q: *Question* (questão)

μ: Média

σ^2: Variância

γ: Grau de confiança

ϵ: Erro de estimação máximo

Sumário

1 Introdução

Este trabalho de conclusão de curso consiste de um esforço no sentido de aprimorar o DEEPAGÉ, um sistema de *question answering* de domínio aberto. Espera-se que agentes deste tipo sejam capazes de fornecer uma saída factual e correta correspondente a uma pergunta que o usuário faça, em linguagem natural, atuando como sistemas especialistas que podem ser rapidamente consultados para a obtenção de conhecimento. Para o treinamento de um respondedor como este, são necessários exemplos de pergunta e resposta, que normalmente não estão disponíveis para qualquer assunto. De modo a obter os dados com os quais realizar este aprimoramento, fez-se uso de uma ferramenta adjacente, a geração de questões.

A geração automática de questões (Question Generation, QG) é o domínio de tarefas de processamento de linguagem natural que trata da criação de perguntas a partir de um contexto. Apesar de ser menos explorado que sua contraparte, a geração de respostas, modelos geradores de perguntas possuem diversas aplicações de interesse, como a elaboração de sugestões de pesquisa em motores de busca, ou a criação de tarefas para estudantes.

Nesta pesquisa é apresentado o PTT5-QG, um modelo gerador de perguntas em português baseado em *transformers* (VASWANI et al., 2017). O modelo é um *End-To-End Question Generator*, que recebe como entrada um contexto, na forma de um texto corrido, e produz como saída um conjunto de perguntas referentes ao contexto.

Trata-se de uma pesquisa que explora, portanto, o potencial do uso conjunto de modelos de *Question Answering* e *Question Generation* na criação de pares de pergunta e resposta que podem ser utilizados na formação de datasets sintéticos para o treino de novos sistemas inteligentes.

Nas seções a seguir, são discutidos em maior detalhe a motivação, objetivos exatos e justificativa deste projeto.

1.1 Motivação

Este projeto de pesquisa insere-se no contexto do desenvolvimento de DEEPAGÉ (CAçãO et al., 2021), um sistema respondedor de perguntas. Este usa um modelo gerador de questões baseado na arquitetura PTT5 (CARMO et al., 2020), um *transformer*, tecnologia de estado da arte que vem sendo utilizada na criação de modelos de processamento de linguagem natural e de agentes conversadores. Sua criação se deu por meio do treinamento de especialização com questões sobre o meio ambiente brasileiro extraídas do *dataset*

PAQ (LEWIS et al., 2021). O sistema faz uso da combinação do modelo gerador de perguntas com o algoritmo BM-25, que busca documentos fonte em bases de conhecimento, tratando-se portanto de um sistema neuro-simbólico.

O projeto é motivado pela necessidade de realizar melhorias no DEEPAGÉ, alimentando-o com mais dados. Para tanto, foi utilizado como base o trabalho de Alberti et al. (2019), que propõe a utilização de modelos geradores de pergunta e resposta para a criação de corpora sintética para uso em tarefas de processamento natural de linguagem. A presente pesquisa explora uma abordagem semelhante, propondo um sistema análogo, porém mais simples, apoiado na tecnologia moderna de modelos *transformers*.

Para que um sistema desta natureza possa ser criado, é necessário ter um modelo gerador de questões. Até o momento, não havia na literatura registro da criação de modelos *transformers* de geração de questões em língua portuguesa. O trabalho, portanto, envolveu a replicação de um método já existente na literatura em língua inglesa que usa este tipo de arquitetura para o desenvolvimento de agentes geradores de questão.

É, portanto, um trabalho que combina três diferentes interesses: (*i*) o aprimoramento do agente respondedor DEEPAGÉ; (*ii*) a criação do primeiro modelo gerador de questões *transformer* em língua portuguesa e (*iii*) o estudo da viabilidade do uso de dados sintéticos no treinamento de modelos geradores de texto.

1.2 Objetivos

Este trabalho objetiva o aprimoramento do DEEPAGÉ (CAçãO et al., 2021), um sistema respondedor criado a partir do modelo de linguagem PTT5, um *transformer*. O DEEPAGÉ foi especializado no domínio do meio ambiente brasileiro por meio de um processo de *fine-tuning* realizado com pares de pergunta e resposta referentes ao assunto.

O respondedor, em sua forma original, apresenta várias falhas. Dentre elas, a alucinação, a apresentação de erros conceituais e a geração de respostas demasiado curtas. A alucinação é uma característica dos *transformers* (JI et al., 2023), não sendo abordada neste projeto. Já os outros erros podem ser considerados uma consequência da pequena variedade de dados sobre os quais o modelo foi treinado. Sendo assim, este projeto toma como meta principal a obtenção de uma versão melhorada do DEEPAGÉ, retreinado utilizando uma fonte de conhecimento expandida, com mais informações acerca da Amazônia Azul e com uma maior variedade de tipos de resposta.

A criação de um *dataset* de qualidade e tamanho suficientes para uso no treinamento de uma rede neural não é trivial. Deste modo, como objetivo intermediário, o trabalho propõe a implementação de um sistema, também movido por inteligência artificial, capaz de gerar *datasets* de perguntas e respostas que podem ser utilizados como fonte para outros

modelos geradores de texto.

A pesquisa tenciona, de forma conjunta ao aprimoramento do respondedor, produzir um estudo de caso acerca da viabilidade do uso de dados sintéticos no treinamento de um modelo de *question answering*. Tal estudo será pautado em uma série de análises, de ordem quantitativa e qualitativa, que verificam a qualidade dos *datasets* gerados pelo sistema intermediário e também comparam a performance do novo DEEPAGÉ com o antigo.

1.3 Justificativa

Pretende-se com esta pesquisa apoiar o desenvolvimento e implantação do BLAB, um agente conversacional especializado em Amazônia Azul, enriquecido por bases de conhecimento referentes a este assunto. Um agente conversacional é uma aplicação dotada de inteligência artificial, capaz de responder perguntas de alta complexidade em linguagem natural. Este agente deve conseguir conduzir conversas caracteristicamente humanas, fazendo uso de argumentos, raciocínios e explicações para atender da melhor maneira possível aos desejos do usuário. Para tanto, faz uso de modelos de processamento de linguagem natural, busca de resposta e sumarização de textos para aprimorar a qualidade de suas respostas, cumprindo a função de tornar o sistema um especialista no domínio.

Embora aborde especificamente o tema da Amazônia Azul, a construção do BLAB colabora com o estudo de inteligência artificial neuro-simbólica. Esta é uma abordagem de estado da arte em IA, que conecta o aprendizado profundo com o raciocínio simbólico baseado em conhecimento. As técnicas exploradas neste projeto podem futuramente ser generalizadas para sua aplicação na criação de agentes inteligentes especializados nos mais diversos domínios. Ademais, esta pesquisa dialoga com outros estudos atualmente sendo desenvolvidos no KEML, como o de conversão de linguagem natural para SQL.

A escolha da Amazônia Azul como domínio de conhecimento também gera um subproduto prático de grande valia para a comunidade acadêmica e a sociedade: a criação de uma ferramenta poderosa de pesquisa sobre conhecimento da Amazônia Azul. Esta é a região designada território marítimo brasileiro, destacando-se por sua biodiversidade e suas riquezas naturais, sendo portanto de importância econômica, social e política para o país. Todavia, é pouco conhecida pela população em geral. Sendo assim, o BLAB serve tanto para difundir o conhecimento relacionado à Amazônia Azul, quanto para auxiliar em pesquisas acadêmicas e no planejamento de atividades relativas ao oceano.

Por fim, a criação do modelo gerador de questões também é de enorme importância, tendo em vista que ainda não está disponível uma tecnologia aberta equivalente para a língua portuguesa. Este, uma vez publicado, poderá ser utilizado em uma variedade de outras aplicações não relacionadas ao BLAB. Por exemplo, sugestões de perguntas para um usuário em um input de texto.

1.4 Organização do Trabalho

O presente trabalho é organizado em capítulos, cujo conteúdo é descrito em sequência:

- Capítulo 2: revela os conceitos que servem como alicerce ao trabalho, bem como os artigos, tecnologias e dados nos quais se apoiou este projeto.
- Capítulo 3: apresenta a metodologia do projeto, descrevendo detalhadamente as atividades realizadas passo a passo para possibilitar a replicação dos sistemas produzidos. São apresentados também os experimentos e avaliações realizados para verificar a qualidade dos sistemas e comparar o desempenho das diferentes versões do DEEPAGÉ.
- Capítulo 4: detalha os requisitos funcionais e não funcionais da nova versão do DEEPAGÉ e do sistema gerador de *datasets*.
- Capítulo 5: expõe os resultados do desenvolvimento, dos testes, experimentos e uma análise aprofundada dos resultados obtidos.
- Capítulo 6: é uma conclusão, na qual são discutidos os resultados, apresentando os pontos fortes e as deficiências de ambos os sistemas trabalhados. Também são apontados os próximos passos que podem ser explorados em pesquisas futuras.

2 Aspectos Conceituais

Esta pesquisa se apoia em uma série de tecnologias e conceitos já existentes na literatura. Nesta seção, são apresentados artigos contendo os métodos relativos a processamento de linguagem natural seguidos neste trabalho, bem como os modelos e *datasets* utilizados para o desenvolvimento dos sistemas.

2.1 Question Generation usando Transformers

Lopez et al. (2021) exploram pela primeira vez a possibilidade de utilizar Transformers na geração automática de questões, a partir apenas do fine-tuning de um modelo de língua pré-treinado. Até aquele momento, as abordagens típicas para o problema de *Question Generation* envolviam soluções complexas, como modelos Seq2Seq baseados em RNN, que faziam uso de dados adicionais ao contexto (como a presença de uma resposta, no caso de geradores *answer aware*).

Os modelos foram todos treinados utilizando como base o modelo de linguagem GPT-2 (RADFORD et al., 2019) pré-treinado para o inglês. A este, aplicaram-se treinamentos de *fine-tuning* usando o *dataset* SQuAD (RAJPURKAR et al., 2016).

No artigo, são exploradas diversas variações de modelo, no que diz respeito ao tipo de dado de saída. Mais especificamente, é verificada a diferença de performance entre modelos "*All Questions Per Line*"(AQPL), nos quais o texto de saída contém várias perguntas correspondentes ao contexto, separadas por um delimitador, e "*One Question Per Line*"(OQPL), em que cada parágrafo produz uma questão.

Ademais, são testados três diferentes tipos de delimitador, o artificial (token <SEP> entre as questões), natural (palavras "context"e "question") e natural-numérico ("context1", "question1", "question2", etc.).

A conclusão do texto é a de enorme sucesso na tentativa de criar estes modelos. De acordo com as avaliações obtidas, foi possível superar o estado da arte, com complexidade reduzida e sem a necessidade de introdução de metadados (como no caso de geração *answer aware*), que nem sempre estão disponíveis. Houve muito pouca diferença entre a performance dos modelos AQPL e OQPL, bem como entre os diferentes delimitadores. Adicionalmente, os autores criaram uma versão do modelo que faz uso de *answer awareness* e notaram uma perda de desempenho considerável.

Para a pesquisa atual, foi seguido o formato AQPL com delimitador artificial <SEP>, considerando o desejo de obter um grande volume de perguntas por passagem.

2.2 Geração de Corpora sintética usando Question Aswering e Generation

Alberti et al. (2019) introduzem um método novo de geração sintética de corpora de *question answering*, caracterizado pela combinação de modelos de geração de pergunta e de resposta, bem como um extrator de resposta.

O método proposto consiste em, dado um contexto C, extrair uma possível resposta A para uma pergunta. Então, com um modelo de geração de pergunta *answer aware*, usar C e A para gerar uma pergunta Q. Enfim, deve ser usado um modelo de geração de resposta (não extrativo) para a criação de uma segunda resposta. O par de pergunta e resposta somente deve ser considerado válido caso ambas as respostas sejam iguais.

Além disso, ao final da geração, os pares são submetidos a uma filtragem que garante que possuem *roundtrip consistency*, um critério relacionado à otimização da log-verossimilhança das triplas C, A, Q geradas. A validade deste processo de filtragem foi verificada manualmente, demonstrando que, dos pares analisados, apenas 16% dos descartados eram válidos, e 39% dos que não foram eliminados eram.

Para testar a proposta de sistema, cada um dos modelos necessários foi criado usando a arquitetura BERT (DEVLIN et al., 2018). O sistema, uma vez montado, foi empregado na geração de um corpus sintético de milhões de questões. Este foi, então, utilizado para o pré-treino de um modelo de linguagem BERT. O modelo resultante demonstrou performance superior a modelos treinados utilizando os *datasets* NaturalQuestions (KWIATKOWSKI et al., 2019) e SQuAD2 (RAJPURKAR; JIA; LIANG, 2018).

2.3 PTT5

O modelo PTT5 (CARMO et al., 2020) é um modelo de linguagem pré-treinado em língua portuguesa criado por pesquisadores da Unicamp. Foi adotado como base para o projeto desta pesquisa, tendo em vista que a utilização de modelos pré-treinados de língua específica produzem resultados melhores do que o uso de modelos multilíngues (VIRTANEN et al., 2019).

O PTT5 foi obtido por meio do treinamento não supervisionado de um modelo de arquitetura T5 no corpus BrWac (FILHO et al., 2018), uma coleção de páginas da *web* em português. Foram criadas três versões do modelo, small, base e large, que se referem aos *batch sizes* de treinamento 1, 2 e 32, respectivamente.

Para ser testado, o modelo foi avaliado usando outros dois *datasets* a partir dos quais foi realizado um processo de *fine-tuning*. São estes o banco ASSIN 2 (REAL; FONSECA; OLIVEIRA, 2020), que foi feito para as tarefas de similaridade semântica e reconhecimento

de inferência textual, e o HAREM, para reconhecimento de entidade nomeada.

Os testes revelaram que o modelo teve excelente performance, em especial sua versão base. Ainda assim, por pequena diferença, o modelo desempenhou pior que um de seus concorrentes, o BERTimbau Large.

2.4 Dataset SQuAD

Para o treinamento do PTT5-QG, foi utilizada uma versão traduzida do SQuAD (RAJPURKAR et al., 2016). Este é um *dataset* que foi criado por pesquisadores da universidade de Stanford, com o objetivo específico de auxiliar na tarefa de compreensão de leitura por parte de sistemas inteligentes. Conta com mais de cem mil perguntas, elaboradas a partir de um conjunto de mais de quinhentos artigos da Wikipedia.

No *dataset*, cada artigo é dividido em parágrafos. A cada parágrafo, é associada uma lista de perguntas e suas respectivas respostas possíveis, extraídas diretamente do texto. Cada resposta é composta por duas informações: seu texto, e a posição dos tokens de início e fim para uso por parte de modelos extratores.

O *dataset* é dividido em três partes, sendo estes o conjunto de treinamento (correspondente a 80% dos dados), de validação (10%) e de teste (que não é de acesso público, e sim utilizado pelos pesquisadores como forma de verificar a qualidade de modelos treinados no SQuAD).

A criação do dataset foi dada em três etapas, sendo estas (*i*) curadoria de passagens de artigos da wikipedia, (*ii*) coleta de perguntas e respostas sobre estas passagens por *crowdsourcing* e (*iii*) obtenção de respostas adicionais para as perguntas já elaboradas, também por *crowdsourcing*.

Uma etapa de análise e verificação da qualidade do *dataset* também foi realizada, para garantir que a base atendia a critérios relativos à dificuldade das questões. Para determinar a dificuldade das questões do banco, foi treinado um modelo de regressão logística que seleciona uma dentre várias respostas candidatas para cada pergunta. O modelo revelou uma queda de performance quanto maior a complexidade das questões (relativo ao tipo de raciocínio necessário para respondê-las) e quanto maior o grau de divergência sintática entre as questões e suas respostas. Na versão final, obtiveram para o modelo uma F1 *score* de 51%. A nota supera a de outros modelos, porém é consideravelmente inferior à performance humana (86.8%).

A partir dos métodos e tecnologias apresentados nesta seção, foi realizado o desenvolvimento do modelo PTT5-QG e o sistema criador de questionários que o utiliza.

3 Método do trabalho

O presente capítulo discute em detalhe os métodos adotados e atividades realizadas na pesquisa, em ordem de execução. Pretende-se, ao longo do capítulo, descrever o passo a passo dos procedimentos adotados para a obtenção do sistema gerador de *datasets* e da nova versão do DEEPAGÉ. Serve, portanto, como uma documentação de referência técnica, tendo em vista que os processos são altamente generalizáveis para utilização em outros sistemas de inteligência artificial.

A seção 3.1 discute a criação e avaliação do sistema gerador de perguntas e respostas. A seção 3.2 descreve a geração do *dataset* expandido sobre a Amazônia Azul, bem como sua utilização no treino e avaliação do novo respondedor.

3.1 Criação do Sistema QA-QG

Para criar um sistema capaz de elaborar pares de pergunta e resposta a partir de textos base, são necessários ao menos dois componentes centrais, sendo estes um gerador de questões e um gerador de respostas. O caminho escolhido foi o de utilizar redes neurais artificiais para cada um destes componentes, tomando proveito do enorme arcabouço tecnológico já existente para estas tarefas.

Com um breve estudo da literatura recente, foi possível observar a existência de modelos extratores de resposta suficientes para uso no sistema almejado. No entanto, até o momento, não havia modelo gerador de questões específico à língua portuguesa que alcançasse uma performance de estado da arte na tarefa. Decidiu-se, portanto, desenvolver este modelo.

O modelo gerador de perguntas em questão foi obtido por meio do fine-tuning da rede neural PTT5 (CARMO et al., 2020), de arquitetura T5, com o dataset SQuAD (RAJPURKAR et al., 2016) traduzido para a língua portuguesa. O modelo é associado ao respondedor extrativo que, dado um contexto e uma pergunta, elabora uma resposta extraindo uma passagem do contexto. Seguem as descrições das etapas de elaboração e teste do modelo gerador de perguntas e da aplicação geradora de corpora sintética.

3.1.1 Preparação do SQuAD

Como delimitado anteriormente, a pesquisa se iniciou com o desenvolvimento do sistema gerador de *datasets* de *question answering*. Como era necessária primeiro a criação de um modelo gerador de questões para compor este sistema, a atividade inaugural do trabalho foi a escolha e preparação de um *dataset* adequado com o qual treinar a rede

neural do gerador. O *dataset* eleito para tanto foi o SQuAD, pois possui uma enorme (+100.000) quantidade de pares de pergunta e resposta sobre diversos assuntos, criados manualmente e filtrados por um processo de curadoria.

Tendo em vista que o SQuAD é um dataset em língua inglesa, foi necessário escolher especificamente uma versão adequadamente traduzida ao português para seu uso no treinamento do modelo. Assim, escolheu-se uma tradução disponível na plataforma Huggingface[1]. Esta foi obtida por tradução automática, usando a ferramenta Google Translate da Google Translate API. Os pares resultantes foram filtrados manualmente pelos autores da tradução de modo a eliminar anomalias resultantes da tradução automática.

Uma vez obtidos os dados traduzidos, estes foram submetidos a um pré-processamento, utilizando *scripts* escritos em linguagem python, para adequá-los ao formato de entrada e saída esperado pela arquitetura do modelo. O *dataset* foi separado em duas colunas, uma contendo o texto de entrada e a outra o texto alvo para ser usado como rótulos no treinamento supervisionado. A entrada é simplesmente o contexto do parágrafo a partir do qual deseja-se gerar as questões. O texto alvo é também um único bloco de texto, contendo a concatenação de todas as perguntas correspondentes, apartadas por um separador artificial na forma de um token <SEP>. Este é o formato chamado por Lopez et al. (2021) de *All Questions Per Line*. A figura 1 contém um exemplo de uma geração que segue esta forma.

"Marco Úlpio Nerva Trajano (em latim: Marcus Ulpius Traianus; 18 de setembro de 53 – 9 de agosto de 117) foi imperador romano de 98 a 117. Oficialmente declarado optimus princeps ('melhor governante') pelo senado, Trajano é lembrado como um soldado-imperador de sucesso que presidiu uma das maiores expansões militares da história romana e levou o império a atingir sua maior extensão territorial na época da sua morte."

Em que ano Marco Úlpio Nerva Trajano foi imperador?<sep>Quem foi o imperador de 98 a 117?<sep>Qual é o título oficial de Trajano?<sep>Quem presidiu uma das maiores expansões militares da história romana?<sep>Quem foi o soldado-imperador de sucesso?<sep>Quando o império romano atingiu sua maior extensão?

Figura 1 – Exemplo de geração de perguntas no formato AQPL, com separador <sep>

[1] https://huggingface.co/

3.1.2 Treinamento do Modelo PTT5-QG

Para criar um modelo baseado em rede neural, são necessários os dados de treinamento e também uma arquitetura adequada que defina a estrutura da rede. Levando em consideração sua eficácia comprovada para aplicação em sistemas geradores de texto, foi escolhido o T5 (RAFFEL et al., 2020), um *transformer*. Mais especificamente, foi escolhida uma versão do T5 já pré-treinada para a língua portuguesa, o PTT5 (CARMO et al., 2020). O pré-treino é um processo que expõe a rede, em estado inicial, a enormes quantidades de texto não rotulado de modo a prepará-la para tarefas na língua alvo. Basta, assim, realizar um treinamento final de adequação à tarefa específica desejada, procedimento este chamado de *fine-tuning*.

Assim, o treinamento do modelo foi realizado utilizando as frameworks Transformers e Datasets da plataforma Huggingface. Estas permitem o carregamento, uso, treinamento e avaliação de modelos de linguagem e também datasets. O carregamento pode ser realizado a partir de uma plataforma *cloud*, a Huggingface Hub, onde se publicam milhares de modelos de NLP e datasets dos mais variados tipos. Desta plataforma foi obtida a versão do modelo pré-treinado PTT5 utilizada neste trabalho, bem como a versão traduzida do *dataset* SQuAD.

Foi criado um script em linguagem python que recebe como entrada um arquivo de formato JSON. Este especifica o dataset de entrada, o modelo base e os demais parâmetros do algoritmo. É realizado um processo de treinamento e avaliação de uma rede T5 conforme especificado.

O modelo foi treinado utilizando o otimizador AdamW (KINGMA; BA, 2014), com learning rate inicial de $1e^{-04}$, batch size de 8, por um total de 20 épocas, weight decay igual a 0 e acumulação de gradiente por 16 passos.

3.1.3 Avaliação do Modelo PTT5-QG

O modelo, uma vez treinado, depende de uma avaliação de sua performance para garantir que possui capacidade de executar a tarefa para qual foi preparado.

Idealmente, tratando-se de um gerador de texto, seria realizada a avaliação de uma amostra de perguntas por parte de avaliadores humanos, pois estes possuem a perspectiva de um usuário final de uma aplicação que interage em língua natural. Este é um processo demorado e que muitas vezes exige o trabalho de especialistas, sendo assim muitas vezes utilizado em seu lugar o cálculo de métricas automáticas de avaliação.

As métricas mais usuais para sistemas de IA mais simples como classificadores (*accuracy, precision, recall*) apresentam baixa correlação com o julgamento humano, sendo assim necessário utilizar métricas específicas para a tarefa de geração de texto.

Para este trabalho, foram escolhidas as métricas BLEU (PAPINENI et al., 2002), ROUGE-L (LIN, 2004) e METEOR (LAVIE; DENKOWSKI, 2009), por serem métricas de eficácia demonstrada e também muito comuns na literatura, facilitando a comparação desta pesquisa com demais trabalhos da área. A implementação usada destas métricas é a da biblioteca Evaluate, também parte do acervo da Huggingface.

Os resultados destas avaliações foram comparados aos escores obtidos para um outro modelo gerador de questões, multilíngue, o mt5-base-tydi-question-generator, treinado com o *dataset* TyDi QA (CLARK et al., 2020).

Vale notar que as métricas selecionadas foram originalmente propostas para outras tarefas de geração textual (como tradução, no caso do BLEU), e testadas em outras línguas que não o português. Novikova et al. (2017) apontam que sistemas geradores de texto podem se beneficiar imensamente de métricas ajustadas especificamente para sua tarefa. Assim, os sistemas estariam efetivamente sendo treinados para desempenhar sua função adequadamente, e não para performarem bem em testes arbitrários.

3.1.4 Composição do Sistema QG-QA

O sistema completo de geração de corpora é uma aplicação em python criada usando a *framework* Haystack[2]. Esta é uma biblioteca que auxilia na criação de aplicações que façam uso de modelos de NLP, oferecendo uma série de ferramentas para carregamento e pré-processamento de documentos textuais, bem como carregamento e uso dos modelos.

A aplicação, após carregar todos os documentos a serem processados, quebra-os em pedaços utilizando o pré-processador do Haystack. Este separa passagens dos textos em janelas de tamanho configurável, respeitando sempre que possível a separação de sentenças, para evitar que uma frase seja cortada no meio. Excertos acima deste limite máximo são truncados.

Opcionalmente, o usuário pode escolher acionar um tradutor de textos que realiza uma tradução do inglês ao português dos documentos. Para tanto, é utilizado o modelo opus-mt-tc-big-en-pt, do projeto OPUS-MT (TIEDEMANN; THOTTINGAL, 2020).

Em sequência, os fragmentos passam pelo modelo gerador de questões, que elabora uma lista de perguntas. Tanto o fragmento quanto a lista são repassados ao extrator baseado em BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), que extrai uma lista de respostas do contexto.

O *dataset* resultante é enfim salvo em um arquivo JSONL, no qual cada linha representa um fragmento de texto com todas suas perguntas e respostas correspondentes, em formato semelhante ao SQuAD (RAJPURKAR et al., 2016). Para fins de documentação,

[2] https://haystack.deepset.ai/

a aplicação ao final também gera um log contendo o número de perguntas e respostas geradas e filtradas.

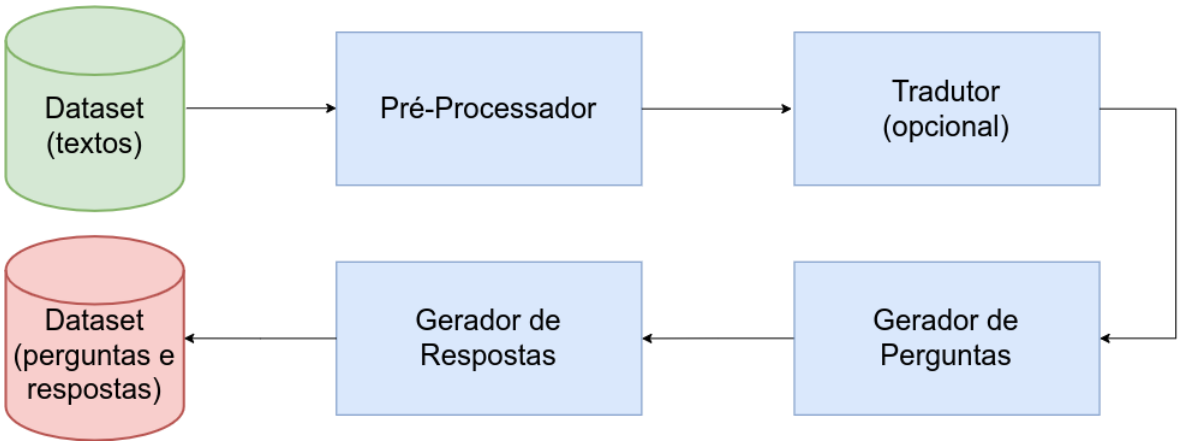

Figura 2 – Diagrama representando os módulos do sistema completo QG-QA

3.1.5 Criação de um Filtro

Para assegurar uma qualidade mínima dos pares gerados pelo sistema em um caso de geração sem supervisão, foi criado também um filtro que pode ser aplicado ao *dataset* gerado para produzir uma iteração mais limpa.

O filtro consiste em calcular e atribuir, para cada resposta, um escore entre zero e um. Respostas cuja nota computada esteja abaixo de um determinado limiar são descartadas, sendo este limite especificado como entrada do filtro. Caso uma pergunta tenha todas suas potenciais respostas descartadas, a pergunta também é eliminada do *dataset*. Por fim, contextos que eventualmente tenham todas suas perguntas rejeitadas são jogados fora.

Evidentemente, o cerne deste componente filtrador é o cálculo da nota de cada resposta. Deseja-se escolher um algoritmo que reflita adequadamente a qualidade da resposta obtida, com a maior segurança possível.

Sendo assim, o candidato ideal é o escore de confiança do modelo extrator de respostas. Este pode ser automaticamente computado pelo Haystack. Este escore opera calculando a soma das probabilidades estimadas pelo modelo para o *token* de início e *token* de fim da resposta. Deste modo, quanto maior a probabilidade estimada para o começo e fim do trecho, maior a confiança do modelo em sua predição. Caso esta confiança seja baixa, é razoável assumir que o modelo teve dificuldade em decidir entre duas ou mais respostas, pondo em dúvida a qualidade do par.

3.1.6 Geração de Corpora de Teste

Com o sistema gerador de *datasets* para *question answering* em mãos, deseja-se comprovar que este atende aos seus requisitos mínimos, especificados com detalhe no

capítulo 4. Tratam-se de requisitos de qualidade e quantidade de pares gerados.

Como teste do sistema completo, foi utilizado o *dataset* TriviaQA(JOSHI et al., 2017) como fonte de documentos para a geração de uma bateria de perguntas e respostas. Este é, assim como o SQuAD, um banco elaborado para a finalidade de uso em sistemas respondedores, contando com noventa e cinco mil pares de pergunta e resposta. Estas respostas não foram geradas por modelos de NLP nem extraídas diretamente de textos, e sim elaboradas organicamente, apoiando-se em evidência disponível na internet.

Não foram utilizadas as perguntas do dataset, todavia, e sim seu banco de evidência, que conta com mais de setenta mil documentos a serem usados para consulta na elaboração das respostas. Estes textos foram usados como uma fonte de dados não rotulados a partir dos quais puderam ser gerados livremente pares de pergunta e resposta. Este banco é dividido nas partições *web* e *wiki*, correspondentes a páginas gerais da *web* e páginas da wikipedia, respectivamente.

Foram selecionados mil documentos da partição wiki para geração de perguntas e respostas pelo sistema. Estes textos foram traduzidos pela aplicação geradora de dataset, usando o modelo tradutor escolhido. Os resultados foram submetidos a filtros com valores 0.7, 0.8 e 0.9., para auxiliar na determinação de um limiar ideal para exclusão de respostas. O número resultante de pares foi comparado ao tamanho do dataset SQuAD para avaliar o potencial gerador desta ferramenta, com o objetivo de determinar sua capacidade de gerar corpora de tamanho suficiente para o uso em treinamento de redes neurais.

3.2 Análise Qualitativa dos Erros

Um importante aspecto da geração automatizada de textos é sua legibilidade, interpretabilidade e utilidade para o usuário final. Portanto, foi conduzido um processo manual de análise qualitativa, pelo autor, dos resultados de ambos os experimentos, de modo a determinar de quais formas o sistema falha em gerar textos úteis. Os critérios adotados para esta análise são melhor detalhados no capítulo 4.

3.3 Aprimoramento do DEEPAGÉ

Esta seção trata do processo de aprimoramento do DEEPAGÉ, bem como seu processo de avaliação.

3.3.1 Geração de Corpus sobre Amazônia Azul

Com a ferramenta de geração de *datasets* em mãos, sua qualidade já aferida, foi realizada a elaboração automática de um corpus sobre a Amazônia Azul. Para tanto, foi

primeiro necessário montar a base de documentos que iriam ser processados pelo sistema. Os textos foram puxados de duas fontes, sendo estas a BLAB-Wiki e a Wikipedia.

A BLAB-wiki é um corpus de atualmente trinta documentos que versam sobre a Amazônia Azul. O material cobre aspectos ambientais, econômicos, legais, dentre outros. Foi realizado um *scraping* completo da plataforma, extraindo todos os documentos em forma de texto corrido para processamento pelo gerador.

Para complementar os resultados fornecidos pela BLAB-wiki, que é um corpus pequeno, foi também realizado uma busca por documentos relativos à Amazônia Azul na Wikipedia. Para tanto, também foi realizado um *scraping*, precedido desta vez por uma busca pelos vinte documentos mais relevantes encontrados utilizando uma lista de palavras chave. As *queries* de busca foram as seguintes:

- Amazônia Azul
- Meio Ambiente do Brasil
- Zona Econômica Exclusiva
- Território Marítimo Brasileiro

Os documentos obtidos foram filtrados manualmente, para assegurar sua relevância ao domínio da Amazônia Azul. Estes documentos também passaram por um processamento que elimina artefatos presentes no texto (como marcadores de citação) e seções irrelevantes, que poderiam contaminar os pares resultantes.

3.3.2 Retreinamento do DEEPAGÉ

Para desenvolver o novo DEEPAGÉ, foram replicadas as mesmas condições de treino do respondedor original, conforme descritas em seu artigo. O processo consiste em tomar o modelo pré-treinado PTT5-Base e submetê-lo a um treinamento de *fine-tuning*, por 30 épocas, com *batch size* de 16, decaimento dos pesos de 0.01 e uma taxa de aprendizado de 2e-5.

Este processo foi realizado duas vezes. Primeiro, com a base de dados antiga, usada para treinar o DEEPAGÉ original. Isto foi realizado com o intuito de produzir uma réplica do antigo DEEPAGÉ, que será chamada de DEEPAGÉ-OLD. Então, foi realizado o treino com o *dataset* expandido com os novos pares de pergunta e resposta, gerando o novo DEEPAGÉ-MERGED.

3.3.3 Avaliação do novo DEEPAGÉ

A etapa final do projeto foi a avaliação das duas versões do DEEPAGÉ, de modo a compará-las. Este processo foi realizado por vias automáticas e manuais de avaliação.

Primeiro, os modelos foram submetidos a avaliação pelo cálculo métricas automatizadas. Assim como no artigo original do DEEPAGÉ, foram usadas as métricas F1-Score, Exact Match e Rouge-L. Adicionalmente, foram também calculadas as métricas BLEU e METEOR, para garantir uma cobertura mais extensa com diferentes métodos de avaliação comumente usados na literatura.

Em sequência, o DEEPAGÉ-MERGED foi submetido a um processo de avaliação manual por parte de anotadores humanos. Este processo foi realizado com o intuito de obter um escore garantidamente relacionado à percepção humana, considerando a distância já observada em pesquisas entre as métricas automáticas e o julgamento por seres humanos (CALLISON-BURCH; OSBORNE; KOEHN, 2006).

Para tanto, foi separado um conjunto de perguntas geradas sobre a Amazônia Azul pelo PTT5-QG, acompanhadas por seu respectivo contexto. Cada uma delas foi então respondida pelo DEEPAGÉ-MERGED, e as respostas foram associadas a cada pergunta, formando triplas contexto, pergunta, resposta.

A cada anotador, foram apresentados um conjunto de triplas geradas. De modo a examinar a qualidade tanto do PTT5-QG quanto do novo DEEPAGÉ, os avaliadores atribuíram notas entre 0 e 10 para cada pergunta e resposta em diferentes quesitos. Os aspectos avaliados estão melhor detalhados no capítulo 4, na forma de requisitos dos sistemas.

Para garantir que fosse avaliado um número de pares suficiente para formar uma visão verdadeiramente representativa da qualidade dos modelos, foi considerado o critério estatístico para determinação de tamanho de uma amostra apresentado por Bussab e Morettin (2010).

Tomando como variável o escore atribuído em um determinado item para avaliar um modelo, temos:

$$n = \frac{\sigma^2 z_\gamma^2}{\epsilon^2}$$

Onde:

- n é o tamanho mínimo da amostra;
- σ^2 é a variância desconhecida da população;
- ϵ^2 é o erro de estimação do escore estipulado;

- γ é o grau de confiança com o qual estimamos n.

Para realizar este cálculo, iremos estimar a variância da população usando uma amostra piloto de 30 pares, considerar um erro de estimação máximo de $\epsilon = 0.75$ com um grau de confiança $\gamma = 0.95$.

4 Especificação de Requisitos

Esta pesquisa trabalha dois sistemas, sendo estes o sistema gerador de *datasets* de *question answering* e o DEEPAGÉ 2. Nas subseções a seguir, são especificados em maior detalhe os requisistos funcionais e não funcionais esperados para cada uma destas aplicações de inteligência artificial.

4.1 Gerador de Perguntas e Respostas

O gerador automático de corpora de QA deve atender a uma série de requisitos, de modo a garantir que produz conteúdo de qualidade e em volume suficiente para ser utilizado no treinamento de um agente respondedor.

A avaliação qualitativa das perguntas e respostas é guiada pelos critérios sugeridos por Sai et. al.(SAI; MOHANKUMAR; KHAPRA, 2022). Estes estabelecem que um gerador deve fornecer textos:

- **Fluentes:** a gramática e escolha de palavras devem ser corretas, incluindo soletração.

Além disso, questões devem ser:

- **Respondíveis:** a pergunta deve ser respondível dado o contexto;
- **Relevantes:** a pergunta deve abordar o contexto sobre a qual é baseada.

Por fim, as repostas devem ser:

- **Corretas:** a resposta deve endereçar a pergunta e respondê-la corretamente de acordo com o contexto.

A quantidade de perguntas e respostas deve ser consideravelmente grande, para que possa ser usada no treinamento do DEEPAGÉ. Para tanto, a quantidade de pares gerados deve ser comparado ao tamanho do *dataset* original do DEEPAGÉ. Neste trabalho, será considerado um sucesso o aumento de ao menos 20% na quantidade de questões.

4.2 BLAB e o DEEPAGÉ

Pretende-se que o BLAB seja um agente conversacional especializado em Amazônia Azul, enriquecido por bases de conhecimento referentes a este assunto. Um agente conversacional é uma aplicação dotada de inteligência artificial, capaz de responder perguntas de

alta complexidade em linguagem natural. Este agente deve conseguir conduzir conversas caracteristicamente humanas, fazendo uso de argumentos, raciocínios e explicações para atender da melhor maneira possível aos desejos do usuário. Para tanto, faz uso de modelos de processamento de língua natural, busca de resposta e sumarização de textos para aprimorar a qualidade de suas respostas imensamente, cumprindo a função de tornar o sistema um especialista no domínio.

O DEEPAGÉ é um destes modelos, sendo seu aprimoramento o objetivo principal deste trabalho. Este deve ser capaz de responder com corretude a qualquer pergunta de um usuário para a qual haja conhecimento dentro de suas bases. Deve ser assertivo e capaz de ser integrado ao BLAB *Chatbot*. Além disso, deve idealmente ser capaz de reconhecer cenários em que não há resposta para uma pergunta, evitando respostas sem embasamento características da halucinação.

5 Resultados

Neste capítulo, são relatados os resultados da pesquisa. Foram produzidas com sucesso duas peças de tecnologia principais, sendo estas o sistema gerador de perguntas e respostas, e a nova versão do DEEPAGÉ. As seções a seguir apresentam os frutos dos testes já descritos para cada um dos sistemas, respectivamente.

5.1 Sistema QG-QA

A presente seção trata dos dados e análises colhidos a partir do sistema completo de geração de *datasets*.

5.1.1 Avaliação do Modelo PTT5-QG

A tabela 1 apresenta os resultados da avaliação dos modelos geradores de questão. Podemos notar que o desempenho do PTT5-QG é consideravelmente superior ao do modelo mT5. A diferença entre os escores é de 29.46 em BLEU(PAPINENI et al., 2002), 10.58 em ROUGE-L(LIN, 2004) e 12.8 em METEOR(LAVIE; DENKOWSKI, 2009).

Há uma forte indicação de que a especialização do modelo em uma língua provoca uma diferença considerável em sua performance. Como a tarefa de geração de corpora agrega o funcionamento de diversos modelos, o desempenho geral sofre um impacto grande com apenas um elo fraco. Pode-se estabelecer o argumento de que é necessário um modelo monolíngue para a tarefa de geração de questões com esta finalidade.

5.1.2 Geração de Corpus de Teste

A tabela 2 apresenta os resultados do experimento com os documentos do banco TriviaQA(JOSHI et al., 2017). Nota-se que o maior número de perguntas foi gerado com uma janela menor. Podemos inferir que o modelo trabalha melhor com passagens menores do que textos longos. Ademais, o número de pares filtrados da janela de 128 foi menor, em

Modelo	BLEU	ROUGE-L	METEOR
PTT5-QG	33.10	30.95	31.80
mt5-base-tydi-question-generator	3.64	20.37	19.00

Tabela 1 – Escores obtidos pelos diferentes modelos.

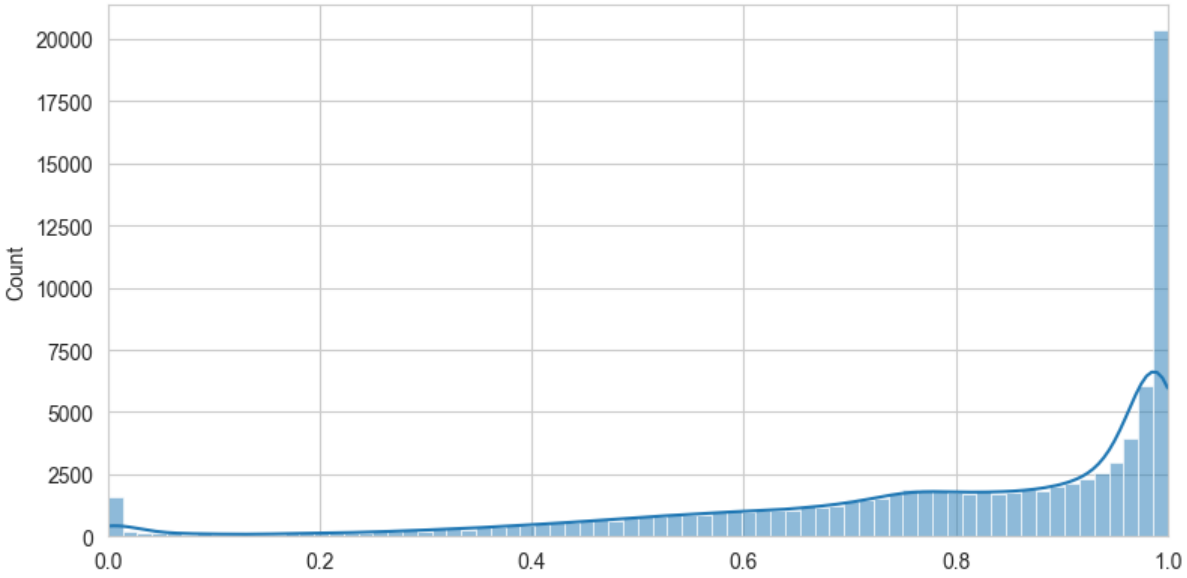

Figura 3 – Distribuição dos pares com base na nota das respostas, para passagens de até 128 palavras.

Limite do Filtro	**Comprimento da Passagem**	**Pares Gerados**	**Pares Filtrados**	**Pares Finais**	Comprimento Médio das Perguntas
0.6	128	**92367**	18063	74304	58.87
0.7	128	**92367**	26272	66095	58.75
0.8	128	**92367**	38473	53894	58.63
0.6	256	58339	23981	34358	58.29
0.7	256	58339	28195	30144	58.17
0.8	256	58339	33898	24441	58.05
0.6	512	54185	38822	15363	57.67
0.7	512	54185	40917	13268	57.67
0.8	512	54185	43643	10542	57.58

Tabela 2 – Quantidade de pares gerados, filtrados e o respectivo tamanho médio de suas perguntas.

termos absolutos e relativos, do que para a janela de 512, sugerindo também que o modelo extrai respostas melhores para as perguntas geradas a partir de passagens mais curtas.

A figura 3 mostra um histograma com a distribuição de perguntas geradas para passagens de janela de 128 palavras. Podemos notar uma concentração de pares próximos a zero. Deste modo, a aplicação de um filtro, ainda que de limite baixo, eliminaria uma parcela considerável do *dataset* da qual o modelo respondedor não tem segurança. Ainda assim, há uma concentração grande de questões entre 0.6 e 1, indicando que há um número satisfatório de pares gerados com confiança.

Com o filtro consideravelmente alto de 0.8 e a janela de 128 palavras, foi possível produzir 53894 pares com o modelo, o que corresponde a cerca de 50.0% dos 107785 pares do SQuAD(RAJPURKAR et al., 2016). Mesmo com a janela de 512 e filtro de 0.8, configuração a partir da qual resultaram apenas 10542 pares, temos a geração de um *dataset* correspondente a 9.78% do tamanho do SQuAD.

É importante considerar também que esta geração utilizou apenas 1000 dos 74021 documentos da partição wikipedia do TriviaQA, correspondendo a apenas 1.35% do *dataset*. Com mais documentos, seria possível a geração de bancos ainda maiores que o próprio SQuAD.

O pequeno declínio de comprimento médio das perguntas conforme o filtro aumenta tende a acusar que perguntas longas costumam produzir respostas de pior qualidade. Esta é uma variação muito pequena, porém.

5.1.3 Análise Qualitativa dos Erros

Os pares gerados são sujeitos a uma série de erros, que foram analisados qualitativamente e separados em categorias. A tabela 3 contém exemplos de cada um dos erros.

- **Erros Gramaticais**: O gerador cria passagens com erros de sintaxe e ortografia.
- **Perguntas Redundantes**: O gerador por vezes cria perguntas que são redundantes, pois já incluem a resposta na própria pergunta.
- **Perguntas Irrelevantes**: O gerador não é capaz de distinguir entre passagens relevantes ao conteúdo do documento, e passagens complementares.
- **Respostas Erradas**: Mesmo com o índice de confiança, o extrator pode escolher respostas gramaticalmente corretas, porém erradas.

Tipo	Contexto	Pergunta	Resposta
Erro Gramatical	-	O que é conhecido por seu governo filantrópico?	-
Redundante	-	Qual o nome da cidade de Londres?	Londres
Irrelevante	Neste capítulo abordaremos equações diferenciais ordinárias (...)	O que é abordado no capítulo?	-
Resposta Errada	Oficialmente declarado optimus princeps ("melhor governante") pelo senado, Trajano é lembrado (...)	Qual é o título oficial de Trajano?	O senado

Tabela 3 – Exemplos de cada tipo de erro do sistema.

5.2 Versão Atualizada do DEEPAGÉ

A corrente seção apresenta os resultados de três etapas da pesquisa. São estes a geração do corpus específico à Amazônia Azul, a avaliação automática do respondedor e a avaliação manual dos modelos criados no trabalho.

5.2.1 Geração de Corpus sobre Amazônia Azul

Utilizando os métodos de obtenção de documentos descritos anteriormente, foram obtidos 30 documentos da BLAB-Wiki e 29 documentos da Wikipedia, configurando um total de 59 documentos.

Destes, foi possível extrair 5445 pares de pergunta e resposta. Submetidos a um filtro de 0.8, o *dataset* resultante foi reduzido a 2849 pares. Isto corresponde a um aumento de 23.3% em relação ao banco original, que possuía 12226 questões. O *dataset* combinado totaliza 15075 questões.

5.2.2 Avaliação Automática do Respondedor

Submetidos aos cálculos das métricas automáticas de avaliação, os modelos obtiveram os escores apresentados na tabela 4. Podemos ver que, em todas as medidas consideradas, houve um aumento significativo de desempenho ao comparar o DEEPAGÉ-OLD com o DEEPAGÉ-MERGED. Isto tende a indicar uma melhoria no desempenho do modelo quando treinado com mais dados, ainda que estes sejam gerados artificialmente.

5.2.3 Avaliação Manual do PTT5-QG e do DEEPAGÉ

Primeiramente, foi obtida uma amostra piloto de 30 pares de pergunta e resposta avaliados. Estes apresentaram as seguintes variâncias, para cada item:

Modelo	F1	EM	BLEU	ROUGE-L	METEOR
DEEPAGE-OLD	23.34	17.39	5.29	23.57	17.67
DEEPAGE-MERGED	32.55 (+39.5%)	26.15 (+50.4%)	22.89 (+32.6%)	32.82 (+39.2%)	24.88 (+40.8%)

Tabela 4 – Quantidade de pares gerados, filtrados e o respectivo tamanho médio de suas perguntas.

	Média	**Variância**
Fluência da Pergunta	9.59	1.79
Respondibilidade da Pergunta	8.45	12.02
Relevância da Pergunta	8.22	10.72
Fluência da Resposta	9.947	0.22
Corretude da Resposta	7.4	16.94

Tabela 5 – Média e variância de cada item avaliado manualmente

- Fluência da pergunta: 1.637
- Respondibilidade da pergunta: 12.328
- Relevância da pergunta: 3.886
- Fluência da Resposta: 0
- Corretude da Resposta: 14.671

Tomamos a maior variância e utilizamos a fórmula determinada para calcular o número mínimo de amostras necessárias:

$$n = \frac{\sigma^2 z_\gamma^2}{\epsilon^2} = \frac{14.671(1.96)^2}{0.75} = 100.196$$

São necessários, portanto, ao menos 101 pares de pergunta e resposta para adequadamente representar a qualidade do sistema. Foram amostrados, ao total, 132 pares. A tabela 5 indica a média e variância de cada item.

Todas as notas são superiores 5, indicando que, em média, os modelos conseguem atingir seus objetivos na maioria dos casos. Em particular, podemos perceber que a fluência dos dois modelos recebe uma nota alta com baixa variância. Não há dúvida, portanto, de que os modelos demonstram alto domínio da língua portuguesa, muito provavelmente em decorrência do extenso pré-treino do PTT5.

Contudo, a corretude da resposta aparenta ser o ponto de maior fraqueza da nova versão do DEEPAGÉ, que tende a flutuar entre respostas totalmente corretas ou totalmente erradas, gerando uma alta variância. Por fim, percebemos que o maior problema do PTT5-QG é o de gerar perguntas relevantes.

6 Considerações Finais

Este capítulo explicita as contribuições tecnológicas e científicas produzidas pelo presente trabalho de conclusão de curso. É providenciado um breve resumo conclusivo da pesquisa e também são ponderados possíveis próximos passos no desenvolvimento de agentes respondedores.

6.1 Conclusões do Projeto de Formatura

Neste trabalho, foi apresentado o PTT5-QG, um gerador de perguntas baseado em uma arquitetura *transformer* para a língua portuguesa, uma língua que tem recebido pouca atenção na pesquisa de processamento de linguagem natural. Este modelo obteve bons resultados em sua avaliação automática, superando modelos multilíngues já existentes.

Unindo o PTT5-QG a um modelo extrator de respostas e um pré-processador, foi criado um sistema gerador de questionários, que serve a muitos propósitos, incluindo mas não limitado à geração de tarefas para estudantes e *datasets* para o treinamento de sistemas de *question answering*. Por meio do experimento com o banco de documentos do TriviaQA, foi demonstrado ser possível utilizar este sistema para criar bancos significativamente grandes de perguntas e respostas de forma automática.

Os *datasets* gerados podem filtrados automaticamente quanto à qualidade das respostas extraídas, eliminando muitos pares ruins, mas não é verificada a qualidade das perguntas geradas. Este filtro verifica a confiança do modelo em sua própria resposta, mas não garante que as respostas sejam corretas, gramaticalmente ou em conteúdo, nem que tenham relevância.

O sistema apresentou seus melhores resultados quando submetido a entradas de passagens de até 128 palavras, tanto em quantidade absoluta de pares gerados quanto em sua qualidade média. As passagens de até 512 palavas resultaram em uma geração de menos pares e de escore geral mais baixo, sofrendo muito com o processo de filtragem.

Subsequentemente, o sistema foi empregado na geração de um *dataset* referente ao assunto da Amazônia Azul, que pôde servir como fonte de treinamento para um agente respondedor, o DEEPAGÉ. Este apresentou melhorias significativas em diversas métricas de avaliação automática para geradores de texto, e uma análise manual por anotadores humanos indicou um julgamento positivo quanto à capacidade do modelo de produzir respostas satisfatórias para questões relacionadas ao seu domínio focal.

Os resultados das avaliações e experimentos dos dois sistemas desenvolvidos demonstram haver grande potencial no uso de dados automaticamente gerados para produção

de modelos de inteligência artificial. O volume de pares gerados é suficientemente grande para ser utilizado no treinamento destes agentes, e foi possível observar uma melhoria objetiva no agente respondedor alimentado pelo corpus gerado. Porém, há, ainda, uma grande preocupação em assegurar a qualidade dos pares gerados, tendo em vista os múltiplos modos de falha aos quais o sistema é sujeito. Sendo o PTT5-QG baseado em uma arquitetura *transformer*, é sujeito ao fenômeno de alucinação, que pode acarretar na produção de dados contaminados.

6.2 Contribuições

A contribuição principal deste trabalho, em um âmbito tecnológico, é o bem sucedido aprimoramento do respondedor DEEPAGÉ. Sob um ponto de vista científico, o trabalho serve também como um estudo de caso que revela ser possível utilizar dados criados artificialmente para expandir o arcabouço de conhecimento fornecido a uma rede neural e, subsequentemente, melhorar seu desempenho.

Podemos destacar também a importante produção intermediária do primeiro modelo de geração de questões em língua portuguesa e baseado em *transformers*. Este pode ser empregado não somente na geração de *datasets*, como neste trabalho, mas também com outros objetivos. Por exemplo, a criação de questionários para estudantes.

Por fim, a produção da análise humana do PTT5-QG e do DEEPAGÉ pode ser utilizada em trabalhos futuros como uma referência com a qual pode ser estudada a eficácia de diferentes métricas de avaliação automática para agentes respondedores.

6.3 Perspectivas de Continuidade

Um possível próximo passo no desenvolvimento do sistema QG-QA é a elaboração de novos métodos de filtragem automática de pares de perguntas e respostas. Preferencialmente, além da medida de confiança, devem ser adotados critérios que possam garantir corretude gramatical, relevância, respondibilidade e corretude das respostas. Uma sugestão é o uso de um modelo de linguagem para calcular a probabilidade de ocorrência de uma frase, de modo a eliminar perguntas extremamente improváveis.

Ademais, como apontado anteriormente, as métricas usadas para avaliação do modelo respondedor podem ser consideradas suficientes para a tarefa abordada neste trabalho, porém possuem falhas já documentadas e foram elaboradas com outra tarefa em mente. Sendo assim, é de interesse que em trabalhos futuros o sistema seja revisitado com testes específicos ao domínio de geração de respostas.

Um passo seguinte, tangencial a este último ponto, é o próprio estudo e desenvolvimento de novas métricas automáticas para avaliação de sistemas de *question answering*,

que levem em conta as particularidades de um sistema gerador de textos específico a essa tarefa.

Referências

ALBERTI, C. et al. Synthetic qa corpora generation with roundtrip consistency. *arXiv preprint arXiv:1906.05416*, 2019. Citado 2 vezes nas páginas 24 e 28.

BUSSAB, W. d. O.; MORETTIN, P. A. Estatística básica. In: *Estatística básica*. [S.l.: s.n.], 2010. p. xvi–540. Citado na página 38.

CALLISON-BURCH, C.; OSBORNE, M.; KOEHN, P. Re-evaluating the role of bleu in machine translation research. In: *11th conference of the european chapter of the association for computational linguistics*. [S.l.: s.n.], 2006. p. 249–256. Citado na página 38.

CARMO, D. et al. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*, 2020. Citado 4 vezes nas páginas 23, 28, 31 e 33.

CAçãO, F. N. et al. *The BLue Amazon Brain (BLAB): A Modular Architecture of Services about the Brazilian Maritime Territory*. 2021. Citado 2 vezes nas páginas 23 e 24.

CLARK, J. H. et al. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 8, p. 454–470, 2020. Citado na página 34.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Citado na página 28.

FILHO, J. A. W. et al. The brwac corpus: a new open resource for brazilian portuguese. In: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. [S.l.: s.n.], 2018. Citado na página 28.

JI, Z. et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, ACM New York, NY, v. 55, n. 12, p. 1–38, 2023. Citado na página 24.

JOSHI, M. et al. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017. Citado 2 vezes nas páginas 36 e 43.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Citado na página 33.

KWIATKOWSKI, T. et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 7, p. 453–466, 2019. Citado na página 28.

LAVIE, A.; DENKOWSKI, M. J. The meteor metric for automatic evaluation of machine translation. *Machine translation*, JSTOR, p. 105–115, 2009. Citado 2 vezes nas páginas 34 e 43.

LEWIS, P. et al. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 9, p. 1098–1115, 2021. Citado na página 24.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: *Text summarization branches out*. [S.l.: s.n.], 2004. p. 74–81. Citado 2 vezes nas páginas 34 e 43.

LOPEZ, L. E. et al. *Transformer-based End-to-End Question Generation*. 2021. Citado 2 vezes nas páginas 27 e 32.

NOVIKOVA, J. et al. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017. Citado na página 34.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2002. p. 311–318. Citado 2 vezes nas páginas 34 e 43.

RADFORD, A. et al. Language models are unsupervised multitask learners. *OpenAI blog*, v. 1, n. 8, p. 9, 2019. Citado na página 27.

RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, v. 21, n. 140, p. 1–67, 2020. Disponível em: <http://jmlr.org/papers/v21/20-074.html>. Citado na página 33.

RAJPURKAR, P.; JIA, R.; LIANG, P. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018. Citado na página 28.

RAJPURKAR, P. et al. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. Citado 5 vezes nas páginas 27, 29, 31, 34 e 45.

REAL, L.; FONSECA, E.; OLIVEIRA, H. G. The assin 2 shared task: a quick overview. In: SPRINGER. *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*. [S.l.], 2020. p. 406–412. Citado na página 28.

SAI, A. B.; MOHANKUMAR, A. K.; KHAPRA, M. M. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, ACM New York, NY, v. 55, n. 2, p. 1–39, 2022. Citado na página 41.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*. [S.l.], 2020. p. 403–417. Citado na página 34.

TIEDEMANN, J.; THOTTINGAL, S. OPUS-MT – building open translation services for the world. In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, 2020. p. 479–480. Disponível em: <https://aclanthology.org/2020.eamt-1.61>. Citado na página 34.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 23.

VIRTANEN, A. et al. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*, 2019. Citado na página 28.