Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Geração Automática de Questões no formato de Exames com base em Aprendizado profundo

Pablo Arruda Araujo

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Dr. Daniel Guerreiro e Silva

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Geração Automática de Questões no formato de Exames com base em Aprendizado profundo

Pablo Arruda Araujo

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Dr. Daniel Guerreiro e Silva (Orientador)
ENE/UnB

Prof. Dr. Janaína Angelina Teixeira
Universidade de Brasília

Prof. Dr. Alexandre Ricardo Soares Romariz
Universidade de Brasília

do Curso de Engenharia da Computação

Brasília, 15 de dezembro de 2023

Dedicatória

Primeiramente, dedico este trabalho à minha família, especialmente ao meu pai, Silvio, à minha mãe, Mirian, ao meu irmão, Pedro, à minha tia, Lidiane, e à minha avó, Lidia. Aos meus pais, serei eternamente grato por todo o amor e suporte que sempre me proporcionaram, permitindo toda a estrutura para que eu e meu irmão pudéssemos nos concentrar nos estudos e no desenvolvimento pessoal.

Dedico este trabalho também aos meus colegas do Cosseno e aos grandes amigos que me acompanharam na jornada da Engenharia da Computação. Em especial, agradeço a Alexandre, Arthur e Nicholas, membros da empresa onde trabalho, o Cosseno, e que contribuíram imensamente para a minha evolução durante a graduação. Agradeço também a Laura e ao Pedro, grandes amigos que me acompanham na vida pessoal desde o ensino fundamental, tornando todo o processo mais leve.

Ao Deus que rege tudo isso, agradeço!

Agradecimentos

Agradeço ao meu orientador, Prof. Dr. Daniel Guerreiro e Silva, pela oportunidade de pesquisa e a contribuição para tornar esse projeto possível. Fica também o meu agradecimento a Universidade de Brasília (UnB) e especificamente ao Grupo de Processamento Digital de Sinais (GPDS) que permitiu a utilização da máquina utilizada no projeto.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Este trabalho aborda a geração automatizada de questões (GAQ) como uma área de pesquisa em destaque no cenário educacional e de Processamento de linguagem natural (PLN). É apresentado um modelo inovador que utiliza técnicas de Aprendizado profundo para criar questões de maneira eficiente e personalizada, visando a melhora da experiência de aprendizagem dos alunos, em especial no cenário educacional brasileiro. Nesse sentido, o objetivo desse trabalho é realizar a GAQ na língua portuguesa. Foi explorada a importância da prática ativa na retenção de informações, destacando como a GAQ pode reduzir o tempo e esforço dedicados pelos educadores à criação manual de perguntas. Além disso, foi discutida a relevância da GAQ em contextos educacionais brasileiros, evidenciando seu potencial para superar desafios específicos do ensino no país. Por fim, foi enfatizada a contribuição significativa da GAQ na personalização do processo de aprendizado, promovendo uma abordagem mais eficaz e adaptada às necessidades individuais dos alunos.

Palavras-chave: Geração Automatizada de Questões, Processamento de Linguagem Natural, Deep Learning, Educação Personalizada, Desafios Educacionais Brasileiros.

Abstract

This paper discusses automated question generation (AQM) as a prominent research area in the educational and NLP scene. An innovative model is presented that uses Deep Learning techniques to create questions in an efficient and personalized way, aiming to improve the learning experience of students, especially in the Brazilian educational scenario. In this sense, the aim of this work is to carry out the GAQ in the Portuguese language. The importance of active practice in retaining information was explored, highlighting how the GAQ can reduce the time and effort dedicated by educators to manually creating questions. In addition, the relevance of GAQ in Brazilian educational contexts was discussed, highlighting its potential to overcome specific teaching challenges in the country. Finally, the significant contribution of the GAQ in personalizing the learning process was emphasized, promoting a more effective approach adapted to student's individual needs.

Keywords: Automated Question Generation, Natural Language Processing, Deep Learning, Personalized Education, Brazilian Educational Challenges.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

B-1 BLEU-1.

B-2 BLEU-2.

B-3 BLEU-3.

B-4 BLEU-4.

GAQ Geração automatizada de questões.

GQ Geração de questões.

IA Inteligência artificial.

LLMs Modelos grandes de linguagem.

LSTM Long short-term memory.

MTR METEOR.

PLN Processamento de linguagem natural.

RNP Redes Neurais Profundas.

SQuAD Stanford Question Answering Dataset.

T5 Text-to-text-transfer-transformer.

TL Transfer Learning.

Capítulo 1

Introdução

No atual cenário educacional, destaca-se a área de Processamento de linguagem natural (PLN), que engloba um conjunto de técnicas e metodologias para aprimorar a interação entre computadores e linguagem humana. No contexto específico da PLN, emerge a Geração automatizada de questões (GAQ) como uma área de pesquisa proeminente. A GAQ refere-se à capacidade de desenvolver sistemas automáticos capazes de criar perguntas de forma contextualizada, com base em conteúdos específicos.

Este capítulo introduz o contexto e a relevância do modelo proposto para a tarefa de GAQ. Apesar de sua relevância, essa tarefa ainda é desafiadora mesmo com os recursos considerado estado da arte. Ademais, quando analisado o cenário brasileiro, poucas publicações são encontradas nesse tema.

Esse tipo de tecnologia desempenha um papel transformacional no ensino e aprendizagem, tornando a educação mais dinâmica, personalizada e alinhada com os objetivos educacionais. A Geração automatizada de questões não só otimiza o tempo dos professores, como também oferece uma abordagem de ensino adaptativa para o estudante.

Além disso, no contexto do PLN, a GAQ desempenha um papel crucial no avanço das técnicas de compreensão de linguagem natural, processamento semântico e geração de texto. A criação de modelos eficazes para gerar questões requer uma compreensão profunda das variações linguísticas, estruturas sintáticas e semânticas, representando, assim, um desafio significativo no campo da pesquisa em PLN.

Diante disso, a introdução de um modelo de GAQ, principalmente na língua portuguesa, se torna essencial não apenas para a comunidade acadêmica, mas também para profissionais da educação, pesquisadores e desenvolvedores interessados em explorar o potencial da tecnologia na educação. Este capítulo busca explorar o contexto amplo em que a GAQ está inserida, destacando sua relevância tanto para aprimorar a prática educacional quanto para impulsionar o avanço científico no campo do PLN.

1.1 Motivação

A razão para o desenvolvimento e implementação de sistemas de Geração automatizada de questões (GAQ) é profundamente enraizada na importância pedagógica e na busca por aprimorar os métodos de avaliação dos estudantes no processo educacional. A questão desempenha um papel vital no contexto educacional, indo além da avaliação do conhecimento adquirido para reforçar o engajamento e o pensamento crítico dos alunos durante o ensino efetivo (Prince (2004)). Essa abordagem interativa não apenas avalia o entendimento dos alunos, mas também proporciona uma oportunidade valiosa para desenvolver habilidades analíticas e reflexivas.

A automatização do processo de geração de questão tem uma série de benefícios potenciais, e no contexto educacional isso se expande além das vantagens diretas. A redução nos custos, tanto em termos financeiros quando do esforço humano, relacionado a criação manual de perguntas é uma das vantagens significativas oferecidas pela GAQ (Prince (2004)). Isso libera o tempo dos educadores e permite que os mesmos deem mais atenção para outras atividades, aprimorando a qualidade do ensino.

No contexto do ensino brasileiro, marcado por desafio diversos e um grande déficit educacional, a introdução de tecnologias inovadoras, como a GAQ, torna-se ainda mais impactante. Notícias como a falta de verba na educação e a piora no desempenho escolar são constantes [1]. Em frente a isso, a possibilidade de gerar questões automáticas em português, se mostra uma forte motivação científica e pouco explorada até então.

Outra forte motivação desse projeto é a contribuição para uma plataforma de ensino brasileira, o Cosseno [2]. Essa plataforma, cujo *slogan* é "A plataforma oficial do estudante", traz uma série de recursos inovadores visando otimizar o tempo de estudo dos estudantes brasileiros durante a sua jornada dos vestibulares. No Brasil, a grande maioria das universidades públicas exigem exames para o seu ingresso, estes que em sua maioria são muito concorridos, principalmente em cursos disputados como a medicina. É nesse contexto que a plataforma atua, trazendo inúmeras ferramentas que facilitam os estudos.

Desde 2020, o Cosseno vem sendo desenvolvido com a disponibilização de diversas ferramentas como banco de questões, simulados e materiais. Atualmente a plataforma conta com dezenas de milhares de estudantes de todos os estados brasileiras, e com a penetração em diversas cidades e municípios. A proposta da GAQ tem uma grande relação com o propósito da plataforma que visa à autoaprendizagem e ensino priorizando a prática.

[1] https://www.correiobraziliense.com.br/brasil/2023/12/6665774-aluno-brasileiro-esta-entre-os-piores-em-matematica.html

[2] https://cosseno.com

1.2 Objetivos

O principal objetivo desta pesquisa, é construir um modelo de GAQ na língua portuguesa, mostrando a sua capacidade e desempenho. A partir dessa viabilidade, serão explorados os limites desse modelo para entender os reais benefícios pedagógicos da solução.

Serão investigadas as principais abordagens e técnicas utilizadas na Geração automatizada de questões, para entender essa tarefa em sua completude. Diante disso, é construído um modelo baseado em *transformers* que visa a alcançar resultados eficientes nesse contexto. O modelo T5 é especificamente escolhido para essa atividade, para explorar a sua flexibilidade.

Depois da construção do modelo, a utilização de métricas como *BLEU*, *METEOR* e *ROUGE* são usadas para verificar a real eficácia da solução. A geração das questões tem por objetivo estabelecer métricas comparáveis aos estudos existentes. A primeira parte dessa pesquisa é a reprodução de resultados, para então expandir o modelo para otimizar as métricas em questão.

Diante do impacto educacional da solução, é esperado explorar possíveis aplicações desse modelo diante dos resultados gerados. A partir disso, ficará a reflexão sobre a adaptabilidade e escalabilidade da solução. Além disso, sendo o Cosseno [3] um bom canal de alcance, será discutida a possibilidade de produtos envolvendo a GAQ na plataforma.

Por fim, como essa tarefa é recente no campo do PLN, serão propostas melhorias e recomendações para futuras pesquisas na área de Geração automatizada de questões. Diante disso, ficará evidente a contribuição pedagógica da solução, bem como o seu papel na área da tecnologia.

1.3 Contexto técnico

Além de toda motivação e objetivos expostos ao longo desse capítulo, é importante visualizar a ideia do modelo computacional em sua completude. Como antes dito, será utilizado como base um modelo considerado estado da arte em diversas tarefas de GAQ, as arquiteturas baseadas em *transformers* (Vaswani et al. (2017)).

A arquitetura escolhida para esse modelo foi o T5 em sua versão em português do Brasil, o PTT5 (CARMO et al., 2020). Esse modelo é altamente flexível, característica que será mostrada ao longo desta pesquisa. Diante disso, foi proposto a *pipeline* desse modelo na figura 1.1.

A ideia geral do modelo, é receber um contexto base, a partir dele selecionar as palavras-chave, que servirão de resposta, e por fim realizar a geração da questão efetiva-

[3]Empresa na área de educação da qual faço parte: <https://cosseno.com>

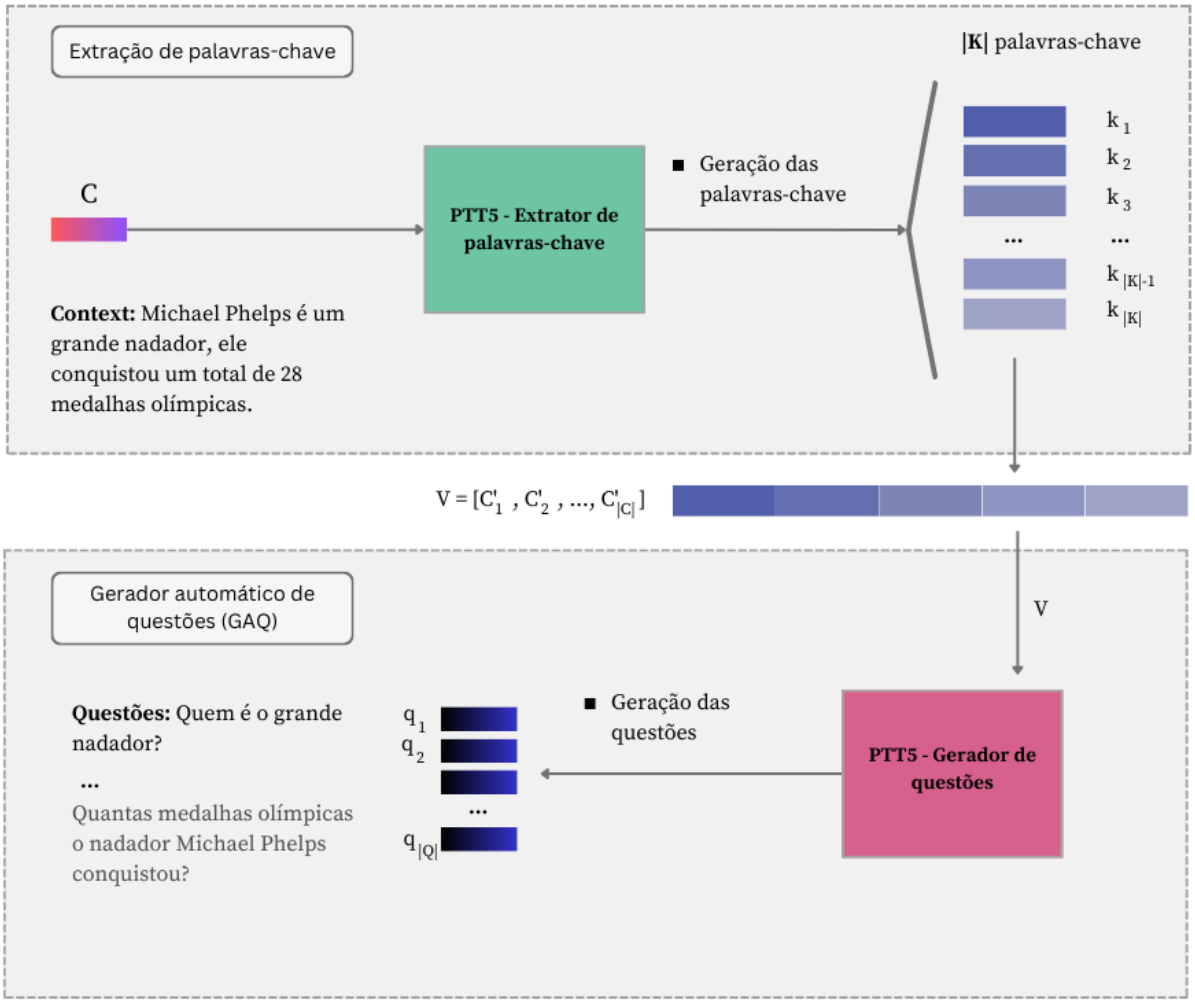


Figura 1.1: *Pipeline* do modelo de GAQ

mente. É possível observar que existem dois principais módulos na arquitetura proposta, o módulo de extração da palavra-chave e o módulo que gera a questão. No primeiro módulo, um modelo é especializado para retirar as palavras mais apropriadas do texto. A partir dessas palavras, o outro modelo responsável por gerar as questões recebe esses termos destaques e produz várias possibilidades de perguntas.

Em ambos os módulos, métricas de avaliação serão utilizadas para avaliar individualmente o desempenho. Além disso, como se trata de uma tarefa de Processamento de linguagem natural, uma análise qualitativa dos resultados será efetuada. Por fim, visando explorar os limites do modelo, será analisada qualitativamente a possibilidade da geração de questões de múltipla escolha, que se assemelham às questões de exame e possibilitam inúmeras aplicações reais da solução de GAQ.

Capítulo 2

Fundamentação teórica

Este capítulo sintetiza os conceitos essenciais que fundamentam este projeto, cujo propósito é a geração automatizada de questões na língua portuguesa. São abordados tópicos iniciais que fundamentam a escolha do tema, por exemplo, o papel pedagógico de solucionar questões e os formatos existentes dessas perguntas. Além disso, são explorados tópicos que embasam a arquitetura da solução, esses que em sua maioria tem relação com a área de Aprendizado profundo.

No que se refere ao papel pedagógico da GAQ, será apresentada a importância educacional dessa atividade, tanto para os estudantes quanto para os educadores. Além disso, será exposto a literatura que ressalta as principais vantagens do aprendizado através de questões.

A tarefa de geração de perguntas e respostas a partir de um determinado contexto já é conhecida, e ademais, já existem soluções clássicas que utilizam principalmente o *Machine Learning* para solucionar esse problema (Kurdi et al. (2020)). Com o forte avanço na área de Aprendizado profundo nas últimas décadas, modelos como o "Long short-term memory"e arquiteturas próprias foram as responsáveis por um avanço nessa área. E indo além, na última década os modelos baseados em *transformers* se consolidaram como o estado da arte para a solução desse tipo de tarefa. Especificamente a atividade de gerar uma questão a partir de um contexto tem se mostrado um desafio, por isso as especificidades desse tópico serão abordadas ao longo deste capítulo.

2.1 O aprendizado mediante perguntas

No contexto desse trabalho, é importante ressaltar a importância da prática no campo da educação, especificamente através do uso de questões. De acordo com Karpicke e Blunt (2011), a eficácia da prática ativa na retenção de informações é mais efetiva do que métodos passivos de aprendizagem.

Em consonância, Thalheimer (2003), oferece os benefícios educacionais dessa abordagem. De acordo com Thalheimer, Will, os benefícios são:

1. Oferecer a oportunidade de praticar a recuperação de informação da memória;
2. Fornecer aos alunos *feedback* sobre as suas concepções erradas;
3. Concentrar a atenção dos alunos no material de aprendizagem importante;
4. Reforçar a aprendizagem através da repetição de conceitos fundamentais;
5. Motivar os alunos para se envolverem em atividades de aprendizagem (por exemplo, a em atividades de aprendizagem (por exemplo, leitura e debate).

Ao incorporar essas perspectivas, sublinho a ideia de que a prática envolvendo questões não apenas enriquece a retenção de conhecimento, mas também contribui para um método ativo e envolvente de aprendizado. Esta abordagem se revela como uma estratégia eficaz para a assimilação e aplicação efetiva do conteúdo em questão.

Os benefícios exibidos por Thalheimer (2003), mostram diversas vantagens educacionais, principalmente voltadas para o estudante. Durante esse processo de prática, o professor também é beneficiado. Um exemplo disso é o tópico três, que tem como consequência a facilitação de produção de materiais mais específicos pelo lado do professor, a partir da identificação dos temas de déficit do estudante.

Apesar de todo o benefício do uso de questões no ensino, o seu processo produtivo exige um extensivo esforço humano e a necessidade de especialistas para sua criação (Zou et al. (2022)). Além disso, com o constante crescimento de recursos educacionais na internet e a popularidade de cursos onlines, a criação manual de questões se torna uma tarefa de grande complexidade.

Diante disso, a Geração automatizada de questões (GAQ) surge como uma solução para facilitar a tarefa dos educadores de criar perguntas, como, por exemplo, na cadeia produtiva de um exame. Além disso, no contexto atual, com a existência de estruturas como os MOOC's [1], a GAQ se consolida como uma ferramenta viável para conseguir personalizar a experiência dos estudantes.

No que tange aos vestibulares e exames, é vital compreender como essas avaliações seguem padrões específicos em termos de tipos de perguntas, estrutura e níveis de dificuldade. Como exemplo, o Exame Nacional do Ensino médio (ENEM) acontece há mais de

[1]Os MOOC's (Massive Online Open Courses) são cursos online abertos que estão disponíveis para qualquer pessoa com acesso à internet e não exigem requisitos mínimos para quem pretende realizá-los.

20 anos e possui uma estrutura específica de 180 questões divididas em 4 áreas e com questões de múltipla escolha com cinco alternativas [2]. A automação na geração de questões pode se alinhar a esses requisitos, proporcionando eficiência e consistência.

Além disso, a GAQ não apenas atende às demandas específicas dos vestibulares, mas também se adapta às tendências educacionais modernas. Isso inclui a crescente popularidade de cursos online e plataformas de aprendizagem digital, onde a automação na geração de questões se mostra como uma ferramenta relevante e eficaz.

Ademais, a partir do propósito de contribuir para a plataforma do Cosseno [3], e tendo em vista as suas necessidades educacionais práticas, podemos ressaltar a GAQ como uma solução para a necessidade de personalização dos estudos. Diante de uma estrutura com dezenas de milhares de alunos, especificamente a tarefa de geração de questões específicas para o estudante se torna um desafio, sendo a produção automática das questões uma potencial alternativa.

2.2 Geração automatizada de questões

Para entender a Geração automatizada de questões (GAQ) em sua completude, é preciso compreender o principal objeto de estudo dessa atividade, a questão. Para isso, será utilizada uma revisão bibliográfica sistemática (Kurdi et al. (2020)) que possui as categorias relevantes de questão no escopo da GAQ.

2.2.1 Formatos de questão

No contexto de uma questão, ela pode ser classificada quanto ao propósito, domínio, fonte de dados, método de geração, tipo de questão e formato da resposta (Kurdi et al. (2020)). Cada um desses tipos tem um importante significado e possuem exemplos específicos que são relevantes a atividade de GAQ, explicitados na Tabela 2.1.

Em particular, na dimensão relativa ao formato de resposta, é importante ressaltar as questões de múltipla escolha. Essa categoria de questão se baseia em criar uma pergunta com múltiplas alternativas de resposta, em que só uma tem a informação completamente correta. Portanto, a diferença desse tipo de questão, é que além da resposta correta, é necessário a criação de distratores, alternativas semanticamente relacionadas, porém que não configuram uma alternativa correta.

As questões por extenso e as questões múltipla escolha serão avaliadas em detalhes ao longo deste trabalho. Além disso, o propósito das tarefas de GAQ será discutido ao longo

[2] <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>
[3] <https://cosseno.com>

Tabela 2.1: Dimensões existentes na GAQ (Kurdi et al. (2020))

Dimensão	**Significado**	**Exemplos**
Propósito	A motivação para o uso da questão.	Provas, aquisição de conhecimento, validação e geral.
Domínio	O assunto que permeia a questão.	Domínio específico e genérico.
Fonte de conhecimento	Tipo do dado para a geração da questão.	Texto, ontologia e outros.
Método de geração	Método para gerar a questão.	Sintático, semântico e baseado em modelo.
Tipo de questão	Tipos de questões.	Questões factuais, completar espaço em branco e problemas aplicados à matemática.
Formato de resposta	Formatação da resposta para a questão.	Resposta por extenso, múltipla escolha e verdadeiro ou falso.

do texto com uma abordagem de domínio mais genérico, em que não será especializado um tipo específico de conteúdo.

Independentemente das opções disponíveis nas diversas dimensões da Geração automatizada de questões (GAQ), é fundamental reconhecer que o processo avaliativo permanece uma constante. De forma mais abrangente, na pedagogia, esses métodos avaliativos podem ser divididos em três principais frentes: avaliação formativa, diagnóstica ou somativa.

No âmbito da avaliação formativa, o foco recai sobre o acompanhamento contínuo do progresso do estudante durante o processo de aprendizagem. Nesse contexto, as questões geradas automaticamente desempenham um papel crucial ao fornecerem feedback imediato, permitindo ajustes e melhorias ao longo do percurso educacional.

A avaliação diagnóstica, por sua vez, concentra-se na identificação de habilidades, conhecimentos e lacunas específicas do aluno. As questões geradas nesse contexto têm o propósito de diagnosticar pontos fortes e fracos, orientando intervenções pedagógicas personalizadas para promover um desenvolvimento mais eficaz.

No contexto da avaliação somativa, o enfoque volta-se para a mensuração do aprendizado alcançado ao final de um determinado período. As questões geradas automaticamente desempenham um papel crucial na elaboração de avaliações abrangentes, fornecendo uma visão consolidada das conquistas acadêmicas do estudante.

Dessa maneira, na seção de metodologia será discutido os tipos de questões escolhidas para esse trabalho tendo em vista o seu papel pedagógico e o seu impacto na realização deste projeto de engenharia.

2.2.2 O desafio da geração de questões

De acordo com Raina e Gales (2022), a tarefa de Geração de questões tem alta entropia dentre as atividades de sequência para sequência já que existem muitas possibilidades de perguntas para uma determinada passagem de entrada. Ademais, comparado a clássica tarefa de geração de respostas a partir de uma pergunta, a criação de uma pergunta a partir de um contexto é um desafio substancialmente maior, tendo em vista que a entropia de distribuições posteriores no processo de geração de uma sequência é grande, deixando poucas referências anteriores para as novas sequências (RAINA; GALES, 2022).

Além disto, a escolha das métricas é diretamente influenciada pela entropia da atividade em questão. Por exemplo, quando se trata de geração de questões do tipo *answer-aware* [4], formato que será detalhado no capítulo de Metodologia, métodos como BLEU-4, METEOR e ROUGE podem ser utilizados com uma boa eficácia, o que é inválido para questões geradas a partir de um contexto sem a resposta dada previamente (RAINA; GALES, 2022). A escolha das métricas no contexto deste trabalho serão justificadas na Seção 2.4.

2.3 Aprendizado profundo no Processamento de linguagem natural

As Redes Neurais Profundas (RNP), enquadradas no contexto de Inteligência artificial (IA), e mais especificamente no Aprendizado profundo, constituem uma categoria avançada de modelos computacionais que se inspiram no funcionamento do cérebro humano para realizar tarefas complexas. Compostas por múltiplas camadas de unidades computacionais interconectadas, as RNP exibem uma capacidade única de aprendizado e representação de padrões complexos.

No contexto da Geração automatizada de questões, onde o uso de IA é prevalente, as RNP desempenham um papel fundamental. Sua arquitetura complexa permite a extração de características específicas e a compreensão de contextos complexos, tornando-as ferramentas poderosas para lidar com o desafio da criação automática de perguntas.

Dentro do espectro de aplicações das RNP, o Processamento de linguagem natural (PLN) destaca-se como um importante campo de estudo. É a área responsável por interpretar e gerar linguagem natural, aquela entendível pelo ser humano. No contexto da GAQ, o PLN torna possível a geração de questões, utilizando linguagem natural, a partir da interpretação de um texto base.

[4] Diz respeito ao tipo de questão criada a partir de um contexto e de uma resposta pré-concebida.

Na última década, essa área tem tido um enorme crescimento, principalmente com o estado da arte no campo da Inteligência artificial generativa [5] com o surgimento dos modelos baseados em *transformers* (VASWANI et al., 2017). E essa área tende a continuar nesse ritmo de crescimento, principalmente pelos incentivos de mercado [6].

2.3.1 Transformers

Os modelos baseados em *transformers* têm revolucionado o campo de Aprendizado profundo, especialmente em tarefas relacionadas ao processamento de sequências, como na GAQ. A arquitetura baseada em *transformers*, introduzida por Vaswani et al. (2017), é conhecida por sua capacidade de lidar eficientemente com dependências de longo alcance em dados sequenciais. Esse fator foi essencial para superar o desempenho de arquiteturas anteriores como o Long short-term memory na tarefa de GAQ.

Arquitetura baseada em *transformers*

A arquitetura dos *transformers* é um modelo de rede neural intitulado sequência para sequência em que a função principal é converter uma sequência em outra sequência. A tarefa de tradução de línguas é um bom exemplo de atividade realizada por esse modelo, tendo bons resultados nesta arquitetura (VASWANI et al., 2017).

Em suma, como pode ser visto na Figura 2.1, a estrutura do *transformer* é composta basicamente de um *encoder*, representado na esquerda, e um *decoder*, representado na direita. O papel do *encoder* é transformar a sequência original em uma representação intermediária, geralmente expressa por uma estrutura de dados com espaço n-dimensional. O *decoder* por sua vez, recebe a representação intermediária e a converte na sequência final.

Nas estruturas tradicionais de sequência para sequência, a estrutura de *encoder* e *decoder* é normalmente utilizada. Um exemplo comum é colocar nessas duas estruturas o modelo de Long short-term memory (LSTM) para realizar a tarefa de tradução de sequências. Durante a execução desses modelos, um importante processo ocorre: a **atenção**. A atenção, de forma resumida, é um mecanismo que analisa a sequência de entrada e decide a cada iteração quais outras partes da sequência também são relevantes.

No artigo de (VASWANI et al., 2017) é proposta uma nova arquitetura para lidar eficientemente com a tradução de sequências, o *transformer*. Assim como nos modelos

[5] A inteligência artificial generativa é uma inteligência artificial capaz de gerar texto, imagens ou outros meios de comunicação, utilizando modelos generativos. Os modelos de IA generativa aprendem os padrões e a estrutura dos seus dados de treino de entrada e, em seguida, geram novos dados com características semelhantes.

[6] <https://finance.yahoo.com/news/natural-language-processing-nlp-market-060100921.html4>

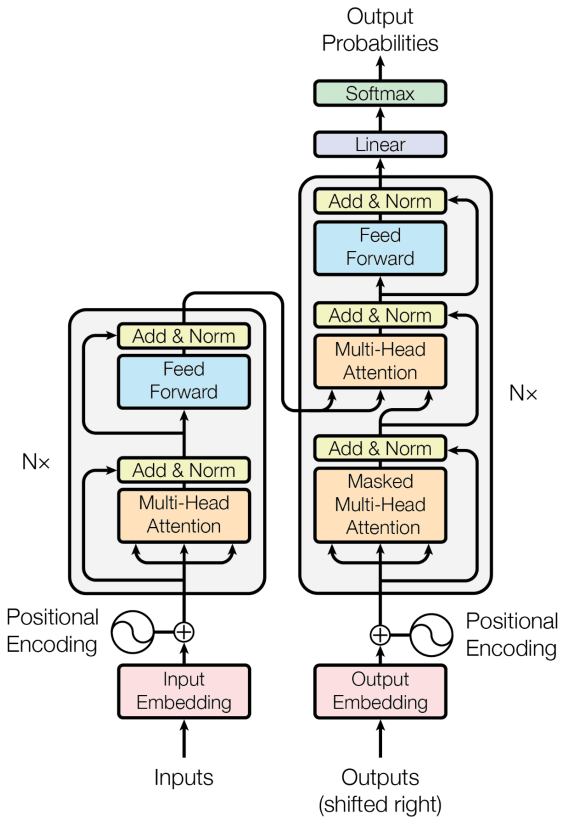


Figura 2.1: Arquitetura do *transformer* (VASWANI et al., 2017)

utilizando LSTM, o *transformer* possui a estrutura de *encoder/decoder* bem como o mecanismo de atenção. Porém, assim como o título do artigo deixa evidente, a nova proposta é priorizar o mecanismo de atenção, de forma que não são utilizadas estruturas de redes neurais recorrentes [7], como LSTM por exemplo.

Nessa nova proposta é esclarecido como essa arquitetura de *transformer* pode ser superior a Long short-term memory para algumas atividades como tradução. Uma grande vantagem dessa nova construção é a sua capacidade de paralelizar as tarefas durante o treinamento, permitindo uma execução muito mais rápida.

O princípio básico de atenção pode ser descrito por:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{2.1}$$

em que:

- **Q:** Matriz de consulta. Q representa uma palavra específica na sequência para a qual queremos calcular a atenção.

[7] Uma rede neural recorrente (RNN) é um tipo de rede neural artificial que usa dados sequenciais ou de séries temporais. Esses algoritmos de Aprendizado profundo são comumente usados para problemas ordinais ou temporais, como tradução de idiomas.

- **K:** Matriz de chaves. K são todas as chaves, ou seja, as representações vetoriais de todas as palavras na sequência.
- **V:** Matriz de valores. V são os valores associados às palavras na sequência. Para módulos de atenção simples, V é a representação vetorial da mesma sequência que Q. No entanto, para módulos de atenção mais complexos que consideram diferentes sequências, como do codificador e decodificador, V pode ser diferente da sequência representada por Q.
- d_k: dimensão das chaves. O resultado do produto de pontos QK^T é escalonado por $\sqrt{d_k}$ para estabilizar os gradientes durante o treinamento.

Podemos resumir a proposta da fórmula como:

1. **Produto de Pontos (Dot Product):** QK^T - Isso representa a similaridade entre a consulta e cada chave.
2. **Escalonamento:** Divida o resultado pelo fator de escala $\sqrt{d_k}$. Este passo ajuda a estabilizar os gradientes durante o treinamento.
3. **Função Softmax:** Aplica a função softmax ao resultado escalonado. Isso converte os valores em pesos de atenção normalizados (valores entre 0 e 1, somando 1).
4. **Ponderação dos Valores:** Multiplica os valores (V) pela matriz softmax resultante. Isso pondera os valores conforme a importância atribuída pela atenção.

A estrutura que realiza esse mecanismo de atenção é o *Multi-head Attention*, presente tanto no *encoder* quanto no *decoder*. A outra estrutura vital dessa arquitetura é a camada *Feed Forward*, presente em vários pontos do processo, é essencial para processar as informações da sequência e melhorar a generalização do modelo.

2.3.2 Large Language Models

O surgimento dos Modelos grandes de linguagem (LLMs) marcou um importante capítulo na evolução do Processamento de linguagem natural. Um grande marco veio com o modelo BERT (Bidirectional Encoder Representations from Transformers), apresentado por Devlin et al. (2018). Essa arquitetura revolucionou o processamento de sequências com a introdução do conceito de pré-treinamento bidirecional, permitindo que o modelo analisasse um contexto de uma palavra considerando tanto as palavras à esquerda quanto a sua direita.

Outro modelo destaque foi o Text-to-text-transfer-transformer (T5), proposto por Raffel et al. (2019) . O T5 inovou ao trabalhar com conversões diretas de texto para texto, proporcionando uma maior flexibilidade nos resultados de geração e análise de sequências.

Em conjunto com a capacidade computacional dos modelos, a quantidade de recursos necessários para criar e executar esses modelos também aumenta. Para treinar essas arquiteturas, são necessários um enorme volume de dados, bem como uma grande capacidade computacional para executá-los. Recentemente, a quantidade de parâmetros nos modelos vem quase atingindo a marca dos trilhões como pode ser visto na Figura 2.2.

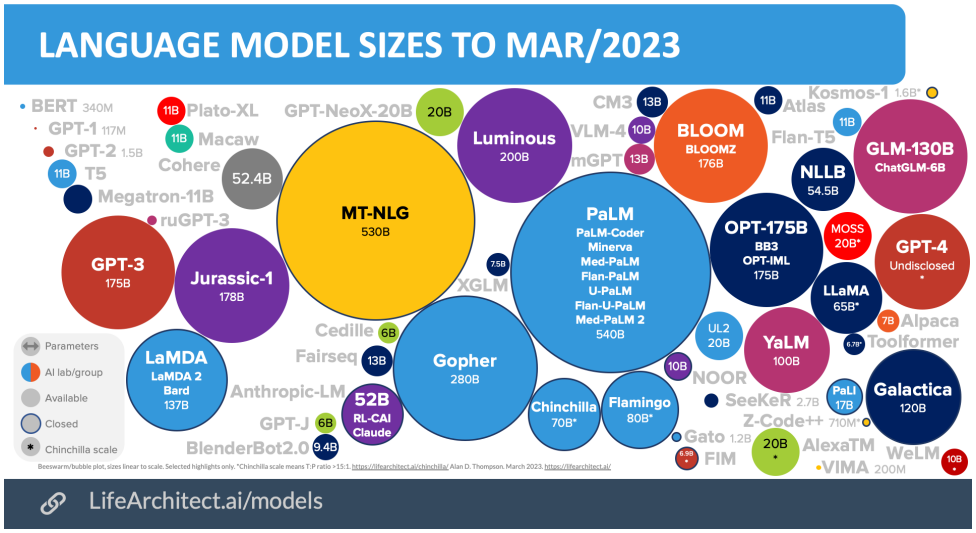


Figura 2.2: Evolução dos parâmetros das LLMs[8]

Ademais, os resultados dos LLMs se mostram excelentes na área de PLN. Essa eficiência, somada à capacidade de aprender representações abstratas e contextuais, torna-os valiosos para a Geração automatizada de questões. A habilidade desses modelos em compreender e gerar texto de alta qualidade é particularmente relevante para a criação automática de perguntas que demandam compreensão semântica e sintática do contexto analisado.

No contexto de LLMs, o Transfer Learning (TL) representa uma importante abordagem para especializar modelos em determinadas tarefas. Nesse caso, os modelos que são pré-treinados em grandes conjuntos de dados, são ajustados em uma nova base de dados. No caso da GAQ, isso permite que o modelo adquira conhecimentos específicos da prática de gerar uma pergunta, permitindo sua aplicação no contexto educacional.

Entrando em mais detalhes sobre o processo, o TL é possível nesse cenário através da prática de *fine-tuning*. Esse processo se baseia em treinar novamente o modelo com dados específicos da sua aplicação. Nesse projeto o *fine-tuning* foi realizado em uma base de

[8]<https://lifearchitect.ai/models/>

dados envolvendo perguntas e respostas em cima de vários contextos, deixando o modelo específico para o seu propósito na educação.

2.3.3 Text-to-text-transfer-transformer (T5)

O T5, ou text-to-text-transfer-transformer, representa um marco significativo no desenvolvimento de Modelos grandes de linguagem (LLMs). Desenvolvido pelo *Google Research*(RAFFEL et al., 2019), o T5 adota uma abordagem inovadora, tratando todas as tarefas de Processamento de linguagem natural como problemas orientados de texto para texto.

A arquitetura do T5 permite a geração de texto a partir de uma variedade de entradas, tornando-o altamente flexível para várias tarefas, incluindo a Geração automatizada de questões. Sua capacidade de entender contextos complexos e gerar saídas de alta qualidade faz do T5 uma escolha promissora para a GAQ.

Em termos de estrutura, o T5 é baseado na arquitetura *transformer*. A arquitetura desse modelo utiliza elementos muito parecidos com o BERT (DEVLIN et al., 2018). O seu treinamento é com o *C4 dataset* [9], uma poderosa base de dados de aproximadamente 750GB originada de uma extensa coleção de dados da web, obtida pelo projeto Common Crawl [10], contendo informações variadas de domínios e idiomas.

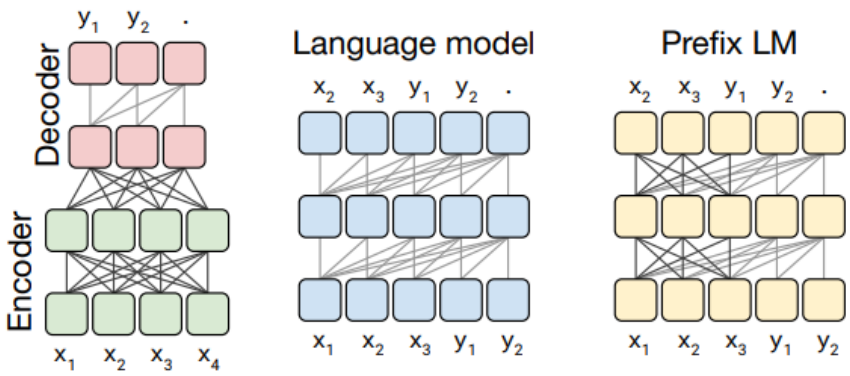


Figura 2.3: Arquitetura do T5 Transformer (RAFFEL et al., 2019)

A Figura 2.3 ressalta a versatilidade do T5 no que tange a sua construção. As variantes de arquitetura ilustradas no esquema demonstram diferentes formas de aplicar o *transformer*. Na arquitetura padrão encoder/decoder, o *fully-visible masking* permite que todos os elementos da sequência de entrada sejam visíveis durante a atenção no encoder e na atenção encoder-decoder, possibilitando uma interação completa entre os elementos. Por outro lado, o *causal masking* restringe a visibilidade no decoder, assegurando que

[9] <https://www.tensorflow.org/datasets/catalog/c4>
[10] <https://commoncrawl.org/>

cada elemento só possa atentar para elementos anteriores na sequência, sendo útil em contextos onde a ordem temporal é crítica. No modelo de linguagem, uma única pilha de camadas *transformer* é alimentada com a concatenação da entrada e do alvo, usando uma máscara causal para preservar a ordem temporal. Além disso, a inclusão de um prefixo no modelo de linguagem permite o *fully-visible masking* sobre a entrada, possibilitando considerar todos os elementos da sequência durante a atenção. Essas adaptações sublinham a flexibilidade do *transformer* no T5, permitindo uma abordagem unificada para diversas tarefas de processamento de linguagem natural.

2.4 Métodos de Avaliação

Nesta seção, serão descritos os métodos de avaliação utilizados para medir a qualidade e desempenho do modelo de Geração automatizada de questões. A primeira etapa de avaliação necessária é na tarefa de extração de palavras-chave. Para isso, serão utilizadas métricas clássicas como *precision*, *recall* e *F1 score*. No começo dessa seção será discutido o motivo de escolha e detalhe sobre essa avaliação.

Na tarefa de GAQ, serão utilizadas métricas comuns na avaliação automática de geração de linguagem natural. Dessa forma foram selecionados os métodos: METEOR, BLEU-4 e ROUGE-L. A validade desses métodos para o problema foi sublinhada na subseção 2.2.2, que ressalta que esses métodos são válidos na construção de modelos de GAQ do tipo *answer-aware* (RAINA; GALES, 2022), que coincide com o objeto de estudo desta pesquisa.

2.4.1 Precision, Recall e F1 score

No contexto da extração de palavras-chave, métricas de avaliação desempenham um papel crucial na mensuração da eficiência de modelo propostos. *Precision*, *recall* e *F1 score* representam indicadores clássicos para avaliar esses modelos, amplamente utilizado na literatura (Firoozeh et al. (2020)).

Para o melhor entendimento, seguem as definições dessas fórmulas:

$$Precision = \frac{\text{VP}}{\text{VP + FP}} \tag{2.2}$$

$$Recall = \frac{\text{VP}}{\text{VP + FN}} \tag{2.3}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.4}$$

- **Verdadeiro positivo (VP)** - representa verdadeiros positivos (palavras-chave corretamente identificadas).
- **Falso positivo (FP)** - representa falsos positivos (palavras-chave identificadas erroneamente).
- **Falso negativo (FN)** - representa falsos negativos (palavras-chave não identificadas erroneamente).

Essas métricas foram escolhidas devido à sua aplicabilidade comum na avaliação de tarefas de extração de palavras-chave (FIROOZEH et al., 2020).

Métricas mais profundas, como *Mean Reciprocal Rank* (MRR) (FIROOZEH et al., 2020), não foram utilizadas, pois a atividade de extração de palavras-chave não é o foco central deste estudo. Foi realizada uma avaliação eficaz sem aumentar desnecessariamente a complexidade da análise.

2.4.2 BLEU-4

O BLEU-4 (Bilingual Evaluation Understudy) (Papineni et al. (2002)) é uma métrica bastante utilizada na avaliação de tradução automática. Esta métrica avalia a qualidade de uma tradução candidata em relação a uma ou mais traduções de referência, levando em consideração a correspondência de unigramas, bigramas, trigramas e quadrigramas. O BLEU-4 (B-4) é calculado considerando a precisão das *n-grams*, onde n varia de 1 a 4, e uma **Brevity Penalty**.

***N-gram* Precision (P)**

A precisão de *n-gram* (P) é calculada para cada valor de n (1 a 4). Representa a razão entre o número de *n-grams* correspondentes na tradução candidata e o número total de *n-grams* na tradução candidata.

$$P_n = \frac{\text{matches}_n}{\text{candidate n-grams}} \tag{2.5}$$

Penalidade de brevidade

A penalidade de brevidade penaliza traduções candidatas que são mais curtas do que as referências. A ideia é evitar que a métrica favoreça traduções muito curtas.

$$\text{Brevity Penalty} = \begin{cases} 1, & \text{se } c > r \\ e^{(1-\frac{r}{c})}, & \text{se } c \leq r \end{cases} \tag{2.6}$$

Onde c é o número de palavras na tradução candidata e r é o número de palavras na referência mais próxima.

Pontuação Final do BLEU-4

A pontuação final do BLEU-4 é calculada multiplicando a média geométrica das precisões *n-gram* pela penalidade de brevidade.

$$\text{BLEU-4} = \text{Brevity Penalty} \cdot \left(\prod_{n=1}^{4} P_n \right)^{\frac{1}{4}} \tag{2.7}$$

O BLEU-4 oferece uma avaliação abrangente da qualidade da tradução, considerando a correspondência de diferentes tamanhos de n-gramas e aplicando uma penalidade para promover traduções mais longas e informativas.

Além dessa métrica, também serão utilizados o BLEU-1 (B-1), BLEU-2 (B-2) e BLEU-3 (B-3) para uma avaliação mais qualitativa de todas as granularidade de *n-gram*.

2.4.3 METEOR

O METEOR (Métrica para Avaliação de Tradução com Ordem Explícita) (Banerjee e Lavie (2005)) é uma métrica de avaliação de tradução automática que leva em consideração a correspondência de unigramas, bigramas e chunks entre a tradução candidata e a tradução de referência. A pontuação METEOR (MTR) é calculada considerando a *precision*, *recall* e uma penalidade de alinhamento.

Essa métrica foi criada buscando corrigir alguns pontos de falha do BLEU-4. Alguns estudos comparativos reforçam a maior correlação desta métrica com a avaliação humana em comparação com o BLEU (Liu et al. (2017)).

Unigram Precision (P) e Recall (R)

A precisão *unigram* (P) é a razão entre o número de unigramas correspondentes e o número total de unigramas na tradução que é candidata. O recall *unigram* (R) é a razão entre o número de unigramas correspondentes e o número total de unigramas na tradução de referência.

É importante ressaltar

$$P = \frac{\text{matches}}{\text{candidate unigrams}} \tag{2.8}$$

$$R = \frac{\text{matches}}{\text{reference unigrams}} \tag{2.9}$$

F-measure

A medida F é uma média harmônica ponderada da precisão e recall *unigram*.

$$F_{\text{mean}} = \frac{10PR}{9P + R} \tag{2.10}$$

Penalidade de alinhamento

A penalidade de alinhamento leva em consideração a ordem dos unigramas e penaliza por grandes descontinuidades.

$$\text{Penalty} = 0.5 \left(\frac{c}{u_m} \right)^3 \tag{2.11}$$

Onde c é o número de chunks e u_m é o número de unigramas mapeados.

Pontuação Final do MTR

A pontuação final do MTR é calculada multiplicando a medida F pela penalidade de alinhamento.

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty}) \tag{2.12}$$

Cada uma dessas componentes contribui para a avaliação global da qualidade da tradução candidata em comparação com a tradução de referência, levando em consideração não apenas as correspondências *unigram*, mas também a ordem e continuidade dos termos.

2.4.4 ROUGE-L

O ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) (Lin (2004)) é uma métrica frequentemente empregada para avaliar a qualidade de sistemas de geração automática, em especial para tarefas relacionadas a sumarização. Essa métrica avalia o quão similar um texto gerado automaticamente está em comparação com um texto de referência, com base na sequência mais longa de palavras em comum, também conhecida como "Longest Common Subsequence"(LCS).

Longest Common Subsequence (LCS)

O LCS representa a maior sequência de palavras comuns entre o texto gerado e o texto de referência. A precisão do ROUGE-L é então calculada como a razão entre o comprimento do LCS e o número total de palavras no texto de referência.

$$\text{ROUGE-L Precision} = \frac{\text{Comprimento do LCS}}{\text{Total de palavras no texto de referência}} \tag{2.13}$$

Recall

O recall do ROUGE-L é a razão entre o comprimento do LCS e o número total de palavras no texto gerado.

$$\text{ROUGE-L Recall} = \frac{\text{Comprimento do LCS}}{\text{Total de palavras no texto gerado}} \tag{2.14}$$

F-measure

A medida F do ROUGE-L é a média harmônica ponderada da precisão e recall.

$$\text{ROUGE-L F-measure} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{2.15}$$

Onde β é tipicamente definido como $1,2$.

O ROUGE-L oferece uma avaliação considerando a similaridade na sequência de palavras mais longa entre o texto gerado e o texto de referência. Essa métrica é valiosa para avaliar a qualidade do texto gerado em termos de conteúdo e cobertura.

Dessa maneira, utilizando-se dessas métricas, é possível obter uma boa avaliação automática da tarefa de GAQ.

Capítulo 3

Metodologia

Nesse capítulo, será abordada a base de dados e a metodologia de treinamento para a Geração automatizada de questões (GAQ). Serão analisadas duas fontes principais de dados: o Common Crawl's Continuously Crawl Corpus (C4) que é a base de dados do modelo T5 e o Stanford Question Answering Dataset (SQuAD) usado para a tarefa específica de GAQ.

O C4, uma extensa coleção de dados da web, forneceu o treinamento necessário para o T5. No SQuAD, o foco foi o *fine-tuning*, ajustando a estrutura dos dados para a entrada do modelo. Foi evidenciado também o método de destaque das respostas em conjunto com o formato *answer-aware*, que ressalta as respostas nos dados de treinamento para otimizar a geração de questões.

Quanto à arquitetura do modelo, o T5 foi utilizado devido à sua flexibilidade e eficiência. A versão *T5-small* foi escolhida para contornar limitações de hardware.

Na etapa de geração de sequências, foi explorado o uso do T5 com ênfase na extração de respostas. Na avaliação da extração de palavras-chave, as métricas de *precision*, *recall* e *F1 score* foram utilizadas. Para gerar as sequências foi adotado o método de *beam search* [1] para obter as opções de perguntas.

Além disso, foi desenvolvida questões de múltipla escolha, incorporando a geração de distratores. Foi apresentado o pré-processamento de respostas, edições para criar variação semântica e a aplicação da biblioteca *sense2vec* [2] para enriquecer as opções de resposta.

Para avaliação, métricas como *BLEU*, *ROUGE-L* e *METEOR* foram utilizadas. Foi avaliado o desempenho do modelo em inglês, português e português traduzido, proporcionando *insights* sobre sua eficácia em diferentes idiomas. Essa metodologia inicial orienta os resultados apresentados nas próximas seções.

[1] Em informática, a *beam search* é um algoritmo de pesquisa heurística que explora um grafo expandindo o nó mais promissor num conjunto limitado. A *beam search* é uma otimização da *best-first search* que reduz os seus requisitos de memória.

[2] <https://spacy.io/universe/project/sense2vec>

3.1 Base de dados

Na esfera do *Machine learning*, o tratamento de dados por si só representa uma grande parcela da construção de uma solução (Géron (2017)). No contexto desse projeto, as bases de dados assumem uma grande relevância também, e, portanto, esta seção visa entrar em detalhes a respeito disso.

As duas principais fontes de dados desse projeto podem ser resumidas pelos domínios:

1. **Modelos grandes de linguagem (LLMs)** - Os LLMs por si só dependem de um grande volume de dados de treinamento.

2. **Geração automatizada de questões (GAQ)** - A tarefa de geração de questões, na abordagem do uso de *fine-tuning*, necessita de dados para se especializar nesta tarefa específica.

Portanto, em um primeiro momento serão discutidos os detalhes de cada um desses domínios na seção a seguir. Ademais, serão explorados o processamento de dados necessário nesses processos e a separação de dados utilizada no processo de treinamento do modelo de GAQ.

3.1.1 Fonte de dados

É importante conhecer as fontes de dados que permeiam o modelo utilizado neste projeto. O modelo escolhido é o Text-to-text-transfer-transformer (T5), detalhado na seção 2.3.3 da fundamentação teórica. Como este modelo faz parte dos LLMs, o seu treinamento é feito em uma base massiva de dados. No caso do T5, essa base de dados é o C4 [3].

A base de dados C4, *Common Crawl's Continuously Crawl Corpus*, é uma coleção extensa de dados públicos coletados da web. Essa base representa uma grande parte do conteúdo disponível na internet atualmente e recebe constantes atualizações. A ordem de grandeza dessa base é na ordem de bilhões de páginas. Um aspecto importante dessa base é a inclusão de metadados, como data de coleta e origem, que proporciona a transparência das informações. Essa base é frequentemente usada no campo de Processamento de linguagem natural, principalmente por sua variedade de informações e riqueza de domínios. Originalmente, o C4 foi criado com dados de abril de 2019, e conta com aproximadamente 750GB de dados utilizados para o seu treinamento (Raffel et al. (2019)).

Para que o T5 execute a Geração automatizada de questões, foi realizado o *fine-tuning* desse modelo utilizando-se a base de dados do SQuAD [4]. Essa base de dados reúne perguntas e respostas colocadas por contribuidores em páginas da *Wikipédia* [5]. O

[3] <https://www.tensorflow.org/datasets/catalog/c4>
[4] <https://rajpurkar.github.io/SQuAD-explorer/>
[5] <https://pt.wikipedia.org/>

artigo da *Wikipédia* é utilizado como contexto e cada uma das perguntas a respeito desse artigo possui uma resposta que representa um segmento de texto ou *span* que faz parte do próprio contexto. Ao todo essa base de dados conta com *98.169* linhas de dados [6].

Tabela 3.1: Dados do conjunto de teste do SQuAD

ID	*56d6edd00d65d21400198250*
Título	Super_Bowl_50
Contexto	In early 2012, NFL Commissioner Roger Goodell stated that the league planned to make the 50th Super Bowl "spectacular"and that it would be "an important game for us as a league".
Pergunta	Who is the commissioner of the NFL?
Respostas	{ "text": ["Roger Goodell", "Roger Goodell", "Goodell"], "answer_start": [32, 32, 38] }

É possível observar que cada entrada dessa base de dados, exibida na tabela 3.1, conta com os campos:

- **Id** - identificação única do campo;
- **Título** - título do contexto;
- **Contexto** - texto utilizado como referência;
- **Pergunta** - pergunta sobre o contexto;
- **Respostas** - possíveis respostas para a pergunta com suas respectivas localizações no texto original.

Algumas modificações nessa base foram realizada e serão especificadas na seção 3.1.2, que trata a respeito do processamento desses dados. Além disso, como este projeto tem por objetivo gerar questões também na língua portuguesa, foi utilizado a versão do SQuAD em português do Brasil [7]

3.1.2 Processamento dos dados

A arquitetura do T5 não foi modificada nesse projeto, portanto, a sua base de treinamento, o C4, não foi alterada, sendo integralmente utilizada. A necessidade de modificação vem no processo de *fine-tuning*, tendo em vista que o formato padrão dos dados do SQuAD teve de ser modificado para a entrada do modelo de GAQ.

[6] <https://huggingface.co/datasets/squad>
[7] <https://huggingface.co/datasets/ArthurBaia/squad_v1_pt_br>

Existem diversos formatos para treinar um modelo para realizar a tarefa de GAQ. Alguns desses formatos são especificados na Seção 3.2.1, além disso, é determinado que este trabalho utilizará a abordagem de treinamento a partir dos elementos de contexto, pergunta e resposta, chamado também de *answer-aware*. Seguindo este padrão e para otimizar a geração de questões, foi utilizada a estrutura explicada pela Equação 3.1.

Segundo Chan e Fan (2019), podem ser geradas questões automáticas de melhor qualidade ao se utilizar um separador especial de parágrafos no trecho do contexto em que a resposta está presente. Dados que se tem um contexto $C = [c_1, c_2, ..., c_{|C|}]$, e uma frase de resposta $A = [a_1, a_2, ..., a_{|A|}]$, pode-se reescrever C como C' da seguinte maneira:

$$C' = [c_1, c_2, ..., [HL], a_1, ..., a_{|A|}, [HL], ..., c_{|C|}] \tag{3.1}$$

Para tal função, o SQuAD foi pré-processado para conter esse destaque nas respostas. Em resumo, todo contexto dessa base de dados foi separado em sentenças, e para cada pergunta de um determinado contexto, o período com a resposta foi envolvido por um *token* especial de destaque, representado na fórmula 3.1 por *[HL]*.

Ademais, como o SQuAD é uma base de dados amplamente utilizada [8], a mesma já possui uma divisão clássica entre treinamento e teste. Esses conjuntos são representados pela proporção de 10% para teste e 90% para treinamento aproximadamente. Essa mesma proporção foi adotada neste projeto, de forma que o conjunto de treinamento foi utilizado para o ajuste fino e o teste para a avaliação e estabelecimento de métricas.

3.2 Arquitetura do modelo

O modelo-base escolhido para esse projeto foi o T5 - explicado em detalhes na seção 2.3.3. A principal motivação dessa escolha foi a sua flexibilidade e eficiência. O T5, na sua essência, foi criado visando explorar os limites dos *transformers* na geração de sequências, para tal foram utilizados diferentes variações da arquitetura de *transformers*, empregando assim uma alta flexibilidade ao modelo. Além disso, essa estrutura se mostra inovadora ao tratar tanto a entrada quando saída no formato textual. A sua eficiência foi demonstrada por Nguyen et al. (2022), mostrando a possibilidade do seu uso e os resultados satisfatórios na tarefa de GAQ.

Além da escolha do T5, foi estabelecido o uso da versão *T5-small*. Essa versão do modelo conta com 60.5M de parâmetros e foi escolhida pelas limitações de *hardware*, contornando problemas como a possível falta de memória na GPU. A título de comparação,

[8] Centenas de milhares de downloads, dezembro, 2023 - <https://huggingface.co/datasets/squad>

o modelo-base possui 223M de parâmetros, e existem outras versões podendo chegar até 11B de parâmetros [9].

Para o uso do português do Brasil, foi utilizado a versão do T5 com o fine-tuning para suportar esse idioma específico, o PTT5 (CARMO et al., 2020). Ademais, foi escolhida sua versão *unicamp-dl/ptt5-small-portuguese-vocab*, que também conta com aproximadamente 60M de parâmetros.

3.2.1 Ajuste fino

Na arquitetura proposta neste projeto, a geração de questões foi realizada no formato *answer-aware*, isso caracteriza que a entrada do modelo, além de receber o contexto e a pergunta, também deve receber a resposta esperada. Essa escolha tem um grande impacto na solução, tendo em vista que antes de executar a solução de geração de questões efetivamente, é preciso extrair qual a reposta esperada para um determinado contexto. Esse processo de extração de respostas foi feito utilizando outro modelo T5 com um fine-tuning para especialização nessa tarefa.

Portanto, existem dois principais domínios a serem considerados no processo de fine-tuning:

1. **Atividade do modelo** - extração de resposta ou geração de questões;
2. **Idioma** - Português do Brasil ou inglês.

Diante dos domínios existentes, serão realizados quatro vezes o procedimento de *fine-tuning*, combinando a atividade do modelo com o idioma desejado. Dessa forma, os detalhes dessa execução serão comparados no capítulo 4 de resultados.

No modelo de extração de respostas, a entrada foi construída utilizando um termo especial *"extract answer"* seguido do contexto que se deseja extrair as palavras-chave, as possíveis respostas. A saída do modelo será um conjunto de termos presentes no texto que teoricamente são os mais importantes semanticamente.

Todas as tarefas de fine-tuning foram feitas diante das mesmas configurações de parâmetros, definidos com base na literatura. A seguir é possível visualizar a descrição destes:

- **Otimizador:**
  - AdamW
  - Learning Rate: 3×10^{-4}

[9] <https://huggingface.co/docs/transformers/model_doc/t5>

  - Epsilon (eps): 1×10^{-8}

- **Hiperparâmetros do Treinamento:**
  - Batch Size: 4
  - Número de Trabalhadores: 32

Além disso, o treinamento do modelo foi possível devido à disponibilização de hardware feito pelo Grupo de Processamento Digital de Sinais (GPDS) [10] da Universidade de Brasília (UnB). A máquina utilizada para o *fine-tuning* conta com uma placa de vídeo de 24 GB de RAM do modelo *Nvidia Quadro P6000*. Além disso, a máquina possui dois processadores de 8 cores (e 16 threads) do modelo *Xeon*.

3.3 Geração de sequências

Um dos principais objetivos desse projeto é gerar questões na língua portuguesa com base em um contexto fornecido. Inicialmente, foi conduzida uma etapa de reprodução de resultados, fundamentada em trabalhos prévios que empregaram o modelo T5 na tarefa de Geração automatizada de questões (GAQ) [11]. Nessa abordagem inicial, o modelo T5 foi especializado por meio do processo de *fine-tuning* com a base de dados SQuAD, utilizando um contexto e uma resposta (manual) como entrada.

Após essa etapa de reprodução, foram propostas algumas otimizações do processo. Primeiramente, a etapa de extração de respostas foi levada em consideração, já que esta é fundamental para a seleção de bons trechos de um contexto para ser utilizado depois na geração de questões.

A tarefa de extração de palavras-chave, de forma isolada, representa uma grande área de estudo. É possível observar na revisão de Firoozeh et al. (2020) que é uma tarefa desafiadora, mas que vem atingindo bons avanços no atual estado da arte. No contexto desse projeto, a preocupação principal é que essa extração seja feita para retirar ao menos os principais conceitos de um dado contexto.

Diante disso, e tendo em vista a necessidade de avaliar esse modelo de extração de palavras-chave, foi-se utilizada as métricas de *precision*, *recall* e *f1 score* para computar a eficiência da solução. Essas três métricas, assim como explorado no referencial teórico, trabalham bem em conjunto para metrificar resultados.

No contexto da GAQ, as classes dessa geração serão classificadas de acordo com os tópicos a seguir:

[10] <http://www.gpds.ene.unb.br/>

[11] <https://colab.research.google.com/drive/1-QZr80BN597BAtsVV3n8nrvh30HSjqN6?usp=sharing>

- **Verdadeiro positivo (VP)** - O modelo extraiu uma palavra que está na referência.
- **Falso positivo (FP)** - O modelo extraiu uma palavra que não está na referência.
- **Falso negativo (FN)** - O modelo não extraiu uma palavra que está na referência

O conjunto de referência, nessa tarefa de extração de respostas, muitas vezes é construído pelo trabalho humano (Firoozeh et al. (2020)), com o objetivo de ter uma métrica supervisionada. Porém, no contexto desse trabalho, tendo a base de dados do SQuAD, é possível utilizar as respostas de cada contexto como o conjunto de referência. Consequentemente, as palavras extraídas pelo modelo serão o conjunto a ser avaliado.

Na etapa de gerar a questão em si, além do ajuste do processo de ajuste fino, foi utilizado o método evidenciado pela Equação 3.1, em que trechos contendo a resposta são sublinhados no processo de treinamento. Na Figura 3.1 é possível uma melhor visualização desse destaque feito usando o termo especial *[HL]*.

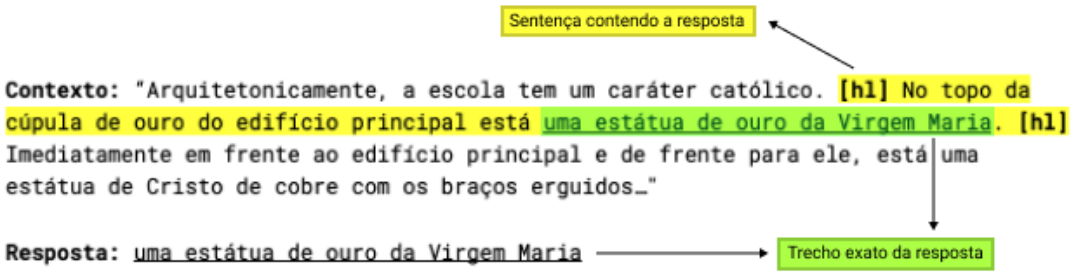


Figura 3.1: Método de otimização da GAQ sublinhando a resposta

No modelo final, a solução segue um método para criar questões automaticamente. Inicialmente, cada resposta identificada é combinada com o contexto da sentença original, resultando em várias instâncias de texto que incorporam o contexto e a resposta associada. Esses textos são então codificados e processados pelo modelo utilizando o método de geração implementado.

O modelo utiliza uma técnica de *beam search*, com *num_beams* definido como 4, para explorar múltiplas opções de perguntas durante a geração das sequências. Além disso, o comprimento máximo de decodificação é limitado a 150 tokens para garantir a concisão das questões geradas. O processo de decodificação resulta em saídas, posteriormente refinadas, removendo tokens especiais e espaços, e extraindo assim as perguntas geradas.

Para que no Capítulo 4 uma análise mais completa da GAQ seja feita, esse modelo final proposto será executado em inglês, português e português traduzido. Cada uma dessas categorias seguirá a seguinte estrutura:

1. Inglês - Será utilizado o modelo T5 com fine-tuning no SQuAD original.

2. Português - Será utilizado o modelo PTT5 com fine-tuning no SQuAD em versão português.

3. Português traduzido - Será utilizado o método do modelo em inglês em conjunto com um processo de tradução usando o *Google tradutor* [12]

Com os modelos construídos, serão feitas análises especificadas na Seção 3.4. Porém, visando reforçar o papel pedagógico desse trabalho, foi também explorado o formato de questão de múltipla escolha. Esse tipo de questão consiste basicamente em colocar opções distratoras nas possibilidades de resposta de forma que a pergunta fica no formato de uma prova. A intenção disso é reforçar o aprendizado através da GAQ como especificado no Capítulo 2 de fundamentação teórica.

A diversidade e relevância dos distratores são fundamentais para testar efetivamente o conhecimento dos respondentes. Nesse contexto, e tendo em vista que a geração automática de questões de múltipla escolha é uma atividade complexa, foi adotada uma solução inicial a ser descrita a seguir.

A estratégia de geração de distratores começa com a pré-processamento da resposta correta. Removem-se pontuações, converte-se para minúsculas e os caracteres especiais são removidos. Em seguida, são aplicadas edições na resposta, envolvendo deleções, transposições, substituições e inserções, considerando diversas possibilidades semânticas e ortográficas.

Para enriquecer ainda mais as opções, incorporou-se a biblioteca *sense2vec* [13], que utiliza *embeddings* semânticos para identificar o sentido mais próximo da palavra. A partir desse sentido, busca-se palavras semanticamente similares que possam servir como distratores. Essa etapa visa criar variação semântica nas opções de resposta, tornando o processo de escolha mais desafiador.

Além disso, para evitar distratores muito próximos à resposta correta, um filtro foi implementado baseado na distância de Levenshtein. Esse mecanismo garante que os distratores não sejam tão ortograficamente similares à resposta correta, mantendo o desafio da escolha da reposta.

O resultado é um conjunto de dois distratores - além da resposta - que não apenas cumprem o papel de alternativas de resposta, mas também enriquecem a experiência de avaliação, proporcionando desafios cognitivos variados e mantendo a qualidade e relevância necessárias para uma avaliação eficaz do conhecimento do respondente.

[12] <https://translate.google.com/>
[13] <https://spacy.io/universe/project/sense2vec>

3.4 Métricas do modelo

Assim como explicado na seção 2.4, serão utilizadas três principais métricas de análise para o modelo: *BLEU*, *ROUGE-L* e *METEOR*. Essas métricas são complementares e são bons indicativos da eficiência do modelo na tarefa de GAQ. Na literatura, os trabalhos de Geração automatizada de questões são feitos na língua inglesa em sua maioria (Kurdi et al. (2020)). Diante disso, uma grande proposta desse trabalho é viabilizar essa tarefa utilizando a língua portuguesa. Dessa forma, cada uma das métricas mencionadas serão combinadas com os modelos em: inglês, português e português traduzido.

Todas as métricas mencionadas, foram analisadas com base no conjunto de teste do modelo. Portanto, para todos os modelos gerados, foram construídas sequências de referência com os dados do conjunto de teste e estas foram colocadas em comparação com as questões padrões da base de dados SQuAD. O resultado para cada modelo também é comparado, a fim de verificar se o modelo em português confere o resultado desejado.

Ademais, para comparar qualitativamente o resultado do modelo com a base de dados, será utilizado um gráfico comparativo em relação às palavras mais frequentes nesses conjuntos. Isso contribui para analisar a similaridade entre estes, além de avaliar se o modelo se enquadra nos tipos de perguntas mais comuns da literatura (Raina e Gales (2022))

Além disso, questões de múltipla escolha foram geradas. Para este propósito não foram criadas métricas, mas um método de geração foi utilizado para a avaliação geral do resultado em que foram selecionadas três áreas do conhecimento: química, geografia e história. Cada área foi associada a um texto de referência criado por um software de geração de texto ChatGPT [14]. Para cada uma dessas sentenças foi utilizado o modelo construído em conjunto com a dinâmica de distratores. Esses resultados foram coletados e serão evidenciados no capítulo seguinte.

[14] <https://chat.openai.com/>

Capítulo 4

Resultados

Nesta seção, serão avaliados os resultados obtidos na tarefa de Geração automatizada de questões (GAQ). Primeiro a extração de respostas será avaliada, métricas como *precision*, *recall* e *F1 score* serão utilizadas para verificar a eficiência do modelo. Na sequência, será feito uma análise das questões abertas, estas que geraram as principais métricas de análise. Ademais, será feito um estudo exploratório das questões de múltipla escolha geradas. Em ambos os cenários, será feita uma comparação do modelo construído em inglês e em português, além de que serão exibidos as avaliações pedagógicas em cima dos resultados, tendo em vista que esse projeto tem um propósito educacional. Durante as discussões, será feita uma comparação com outros trabalhos para exibir um pouco da contribuição científica desse projeto. Por fim, serão discutidas possibilidades de aplicações práticas do modelo construído e o seu impacto em situações reais.

4.1 Análise da extração respostas

A tarefa de Geração automatizada de questões proposta nesta pesquisa segue o modelo *answer-aware*, ou seja, é necessário o conhecimento da resposta para a geração efetiva de uma questão. Sendo assim, nesta seção, será avaliado a eficácia do modelo construído para retirar palavras-chave, que serão possíveis respostas, de um dado contexto.

Primeiramente, para compreender o resultado do modelo, será feita uma análise qualitativa do exemplo a seguir:

Contexto:
Michael Phelps, o lendário nadador, gravou o seu nome na história com um recorde sem paralelo de 23 medalhas de ouro olímpicas, demonstrando o seu extraordinário talento e dedicação ao desporto da natação.

Keywords extraídas: *Michael phelps* e *23*

É possível observar que para o texto retratando as conquistas do nadador Michael Phelps, o modelo extraiu dois termos destaques: *Michael Phelps* e *23*. Essas palavras configuram de fato pontos relevantes no texto, tendo em vista que o principal objeto de estudo é o nome do nadador e o número *23* que representa numericamente a sua grande conquista histórica.

A fim de validar a eficácia da extração desses termos destaques, o conjunto de teste do SQuAD foi utilizado para verificar a proximidade entre os termos gerados pelo modelo em comparação com as repostas que existem na base de dados. As métricas de *Precision*, *Recall* e *F1 score* foram utilizadas para avaliar a eficácia da extração de palavras-chave.

Tabela 4.1: Comparação entre modelos de extração de palavras-chave

Modelo	**Precision**	**Recall**	**F1 Score**
Inglês	0.3111	0.31284	**0.3120**
Português	0.1824	0.3145	**0.2309**

Na tabela 4.1, é possível observar as métricas dos modelos de extração de respostas. Como esperado, o modelo em inglês se sobressaiu em relação ao português, porém por uma margem menor que 30% na métrica de F1 - que será utilizada como destaque nesta análise - indicando um resultado próximo entre ambos os modelos.

A tarefa de extração de palavras-chave de um documento é comum na literatura e é amplamente estudada (Firoozeh et al. (2020)). Essa tarefa tem por objetivo criar um modelo que seja capaz de extrair o máximo de palavras-chave precisas de um determinado documento. No contexto desse trabalho, o objetivo principal é conseguir palavras-chave relevantes para conseguir atestar a possibilidade do modelo de GAQ. Dito isso, o score de 0.3120 no modelo inglês e de 0.2309 no modelo português na *F1 score* se mostrou compatível com a revisão feita por Firoozeh et al. (2020). Nessa revisão, a maior métrica encontrada na *F1 score* foi de 0.56, sendo que o estado da arte tem o score de médio por volta de $0.2 - 0.3$, mesma faixa encontrada neste modelo.

A seguir, no bloco de texto é possível visualizar um dos exemplos que foram avaliados no conjunto de teste do SQuAD.

Contexto:
O princípio de inclusões e componentes afirma que, com rochas sedimentares, se inclusões (ou **clastos**) são encontradas em uma formação, as inclusões devem ser mais antigas que a formação que as contém. Por exemplo, em rochas sedimentares, é comum o **cascalho** de uma formação mais antiga ser rasgado e incluído em uma camada mais nova. Uma situação semelhante com rochas ígneas ocorre quando são encontrados **xenólitos**. Esses corpos estranhos são recolhidos como **fluxos de magma ou lava** e são incorporados posteriormente para resfriar a matriz. Como resultado, os **xenólitos** são mais antigos que a rocha que os contém.

Nesse contexto, conforme a base de dados, existem cinco possíveis palavras-chave, da qual o modelo foi capaz de reconhecer quatro destas (Verdadeiros positivos (VP)) e deixou de reconhecer apenas uma (Falso negativo(FN)), segundo a tabela 4.2.

Tabela 4.2: Resultado da extração de respostas

Palavra-chave	Referência	Extraído	Resultado
clastos	Sim	Sim	VP
xenólitos	Sim	Sim	VP
cascalho	Sim	Sim	VP
fluxos de magma ou lava	Sim	Sim	VP
O princípio de inclusões e componentes	Sim	Não	FN

Diante desse exemplo, foi possível obter uma *F1 score* de 0.89 para um processo de extração de palavras-chave. Sendo assim, dado que o modelo de extração de respostas se mostrou capaz na sua tarefa, nas seções a seguir será trabalhado a tarefa de GAQ efetivamente.

4.2 Análise das questões abertas

Nesta etapa, será abordada a avaliação das questões abertas geradas pelo modelo de Geração automatizada de questões (GAQ). O intuito central é verificar a capacidade do modelo de criar questões frente a um contexto.

Tabela 4.3: Avaliação dos modelos de GAQ

Modelo	**B-1**	**B-2**	**B-3**	**B-4**	**MTR**	**ROUGE-L**
Inglês	0.37055	0.26369	0.19960	0.15085	0.35646	0.44682
Português	0.31214	0.22385	0.17021	0.12905	0.30026	0.38871
Português traduzido	0.30868	0.21985	0.16590	0.12468	0.30032	0.38874

As métricas de desempenho, apresentadas na Tabela 4.3, refletem a qualidade das respostas geradas pelo modelo. É importante notar que a métrica *BLEU* foi avaliada para 1-gram até 4-gram. A métrica BLEU-4 representa a avaliação acumulativa de todas as granularidades, portanto as outras métricas *BLEU* foram utilizadas para avaliar a coesão geral de cada unigrama. Além disso, na sequência são apresentadas as métricas METEOR (MTR) e *ROUGE-L*.

Assim como exposto na fundamentação teórica, a métrica METEOR possui uma maior correlação com a avaliação humana nas tarefas de traduções, sendo o parâmetro de maior importância a ser avaliado. É possível observar que a métrica METEOR teve um resultado de 0.36 para o modelo em inglês e 0.30 para o modelo em português, demonstrando um bom resultado quando comparado à literatura (Kurdi et al. (2020)).

Tabela 4.4: Exemplos de GAQ com a métrica METEOR

Contexto	**Referência**	**Gerado**	**MTR**
... [HL]Newton foi limitado pela defesa de Denver, que o demitiu sete vezes e o forçou a três turnovers, incluindo um fumble que eles recuperaram para um touchdown. [HL] ...	Quantas rotações Cam Newton teve?	Quantos turnovers o Broncos forçou a fazer?	**0.0**
[HL]Em muitas partes dos Estados Unidos, após a decisão de 1954 no processo judicial Brown v. Conselho de Educação de Topeka, que exigia que as escolas dos Estados Unidos dessegregassem "com toda velocidade deliberada", as famílias locais organizaram uma onda de "academias cristãs"privadas.[HL]...	Que processo judicial desagregou as escolas nos Estados Unidos?	Que processo judicial foi feito em 1954?	**0.3346**
Em economia, ... [HL] Paul Samuelson, o primeiro americano a ganhar o Prêmio Nobel de Ciências Econômicas, [HL] ...	Quem foi o primeiro americano a ganhar o Prêmio Nobel de Ciências Econômicas?	Quem foi o primeiro americano a ganhar o Prêmio Nobel de Ciências Econômicas?	**0.9998**

Na tabela 4.4, é possível observar os exemplos diversos de variação na métrica METEOR. Foi coletado o pior e melhor resultado no SQuAD em cima dessa métrica, bem como um exemplo próximo à média obtida em toda a base. Na primeira linha da tabela, é possível observar que no exemplo que o METEOR é zero, a pergunta gerada usa termos completamente diferentes da referência, justificando a métrica nula. No exemplo representativo da média do METEOR, na segunda linha, é possível observar que a pergunta, apesar de diferente, é passível da mesma resposta. E no último exemplo, a referência e o modelo tem a mesma pergunta para o contexto exibido. Diante disso, a avaliação geral é que o modelo conseguiu desempenhar bem a tarefa de GAQ.

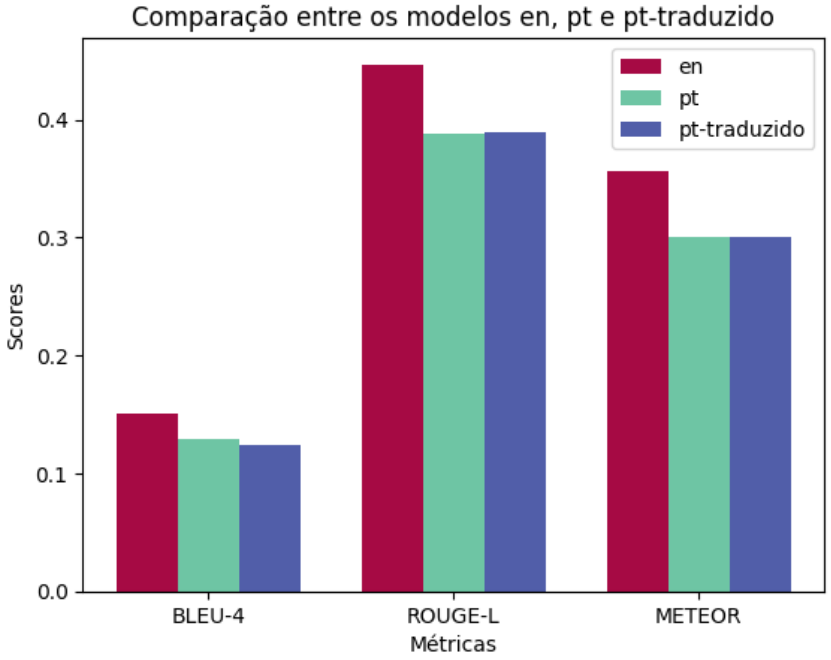


Figura 4.1: Comparação entre os diferentes modelos de GAQ

Para relacionar melhor o desempenho de cada modelo, a Figura 4.1 coloca em comparação os três diferentes domínios linguísticos. É possível visualizar que os modelos em português traduzido e português obtiveram uma maior proximidade, sendo que o modelo inglês se destacou em todas as métricas. É possível observar que a métrica METEOR do português representa 84.2% do valor obtido pelo modelo em inglês, sendo um percentual satisfatório. O modelo do português traduzido obteve resultados muito próximos do modelo em português, com variações percentuais inferiores a 5% em todas as métricas.

Na literatura, é comum que se tenham avaliações de caráter humano para a tarefa de GAQ (Kurdi et al. (2020)). Diante disso, a seguir serão expostos alguns exemplos das gerações de questões abertas. Primeiro, avalia-se o modelo português:

Contexto:
Michael Phelps, o lendário nadador, gravou o seu nome na história com um recorde sem paralelo de 23 medalhas de ouro olímpicas, demonstrando o seu extraordinário talento e dedicação ao desporto da natação.

Questão: Quem é o lendário nadador?
Resposta: Michael Phelps

É possível observar que o modelo em português foi capaz de gerar uma pergunta semântica e sintaticamente coerente. Apesar disto, a pergunta fica um pouco ampla e muito dependente do contexto. Ao exercitar o mesmo exemplo no modelo em inglês, obtém-se o resultado:

Contexto:
Michael Phelps, the legendary swimmer, has etched his name in history with an unparalleled record of 23 Olympic gold medals, showcasing his extraordinary talent and dedication to the sport of swimming.

Questão: Who has etched his name in history with an unparalleled record of gold medals?

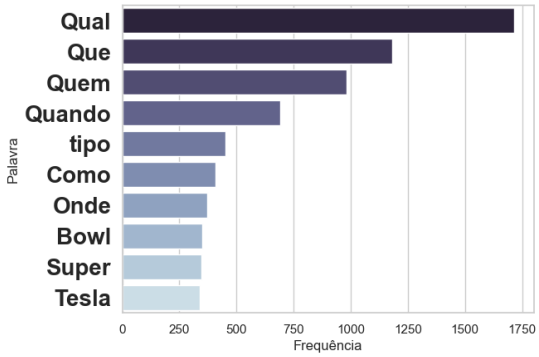
Resposta: Michael Phelps

Nesse caso, é notável que o modelo em inglês atribuiu um maior contexto a pergunta, deixando-a menos dependente do contexto e evidenciando métricas melhores que o modelo em português. De toda maneira, ambos conseguem construir perguntas corretas.

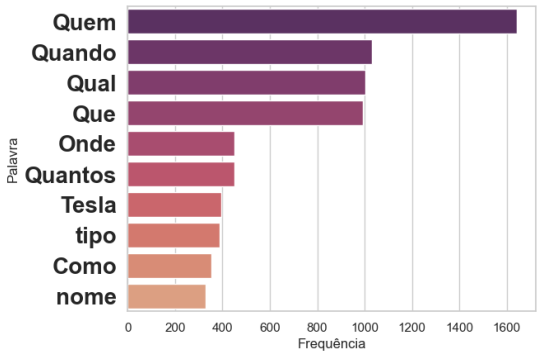
Além da análise de métricas e a avaliação, é possível observar a correlação entre palavras mais frequentes nas perguntas do conjunto de teste em comparação com a saída do modelo construído no mesmo conjunto de dados. A Figura 4.2 compara essa frequência.

É possível analisar que *nove* termos dentre os *dez* mais frequentes gerados pelo modelo, também estão presentes nas perguntas da base de dados. Isso contribui para o resultado positivo do modelo. Além disso, de acordo com Raina e Gales (2022) "Para um conjunto de questões, elas podem ser categorizadas nos tipos: O quê, quem, quando, onde, por que, como, qual, sim/não [...]". Esse discurso concorda com os resultados obtidos tendo em vista a presença de termos como "Quem", "Onde", "Quando"dentre outros.

Ao selecionar questões que possuem um enunciado mais contextualizado, é possível observar que a GAQ assume o seu papel de contribuição para um processo avaliativo. Diante disso, as vantagens propostas por (THALHEIMER, 2003) podem ser aplicadas de



(a) Palavras mais frequentes no conjunto de teste do SQuAD



(b) Palavras mais frequentes geradas pelo modelo no conjunto de teste do SQuAD

Figura 4.2: Frequência de palavras no conjunto do SQuAD

uma maneira muito mais prática com a geração automática de uma pergunta, diminuindo o tempo de redação de uma questão pelo educador e ampliando as opções de estudo do aluno.

Por fim, assim como esperado, o modelo em português é capaz de gerar perguntas automáticas a partir de um contexto. Diante dessa realidade, e a fim de explorar mais a fundo a sua capacidade, no capítulo seguinte será trabalhado a possibilidade da geração de questões de múltipla escolha. É possível avaliar outros exemplos de geração no apêndice A.

4.3 Análise de questões de múltipla escolha

Nesta Seção, serão abordados os resultados obtidos na geração de questões de múltipla escolha. Como retratado no Capítulo 3 de metodologia, o objetivo dessa tarefa é a geração de dois distratores que sirvam de alternativas para responder uma dada questão. É importante que esses termos sejam próximos semanticamente da resposta, porém não pode representar uma resposta alternativa, tendo em vista que a pergunta deve conter apenas uma alternativa correta.

Ao utilizar o mesmo contexto dos exemplos da seção anterior, é possível observar os seguintes resultados:

Contexto:
Michael Phelps, o lendário nadador, gravou o seu nome na história com um recorde sem paralelo de 23 medalhas de ouro olímpicas, demonstrando o seu extraordinário talento e dedicação ao desporto da natação.

Questão: Quem é o lendário nadador?
Opções:
A. Usain Bolt
B. Michael phelps *
C. Phil Heath

Além da pergunta e da resposta já obtida anteriormente, são criados dois distratores complementares. Observe que no exemplo acima, a reposta está marcada com *'*'*. Em uma avaliação inicial, é possível perceber que as alternativas distratoras têm uma concordância semântica. No exemplo em questão, a resposta para o contexto é *Michael Phelps*, que foi um grande nadador norte-americano com o maior número de medalhas olímpicas. Os distratores colocam como opção *Usain Bolt*, que também é um grande atleta olímpico, porém no atletismo, e o *Phil Heath*, outro grande atleta, porém na área do fisiculturismo. Isso mostra que é possível gerar distratores semanticamente corretos, porém que não conflitam com a resposta certa.

Para uma melhor avaliação, três grandes áreas do conhecimento foram escolhidas a fim de realizar uma análise qualitativa da geração das múltiplas escolhas, são elas: química, geografia e história. Para cada uma dessas áreas, textos foram gerados a fim de representar um dado contexto para o modelo de GAQ. Dessa vez, foi escolhido o modelo em inglês para expandir a análise geral do modelo. A seguir tem-se a visualização das questões obtidas pelo modelo para cada uma dessas frentes na tabela 4.5

É possível observar que nas três frentes trabalhadas, o modelo foi capaz de gerar perguntas com relevância sintática e semântica. Além disso, o modelo foi capaz de incorporar os distratores, gerando assim questões de múltipla escolha corretas.

Na questão de química, foi perguntado a respeito da fórmula da água. É possível observar que a resposta correta é H_2O, e como distratores, foram colocadas as opções *Nitrogênio* e *Hidrogênio*, que não são respostas corretas por não contemplarem a fórmula da água por completo.

Na questão de Geografia, a partir de um texto descritivo sobre o Brasil, foi gerado uma pergunta com opções que tratam sobre possibilidades de países. A resposta correta é Brasil, porém opções distratoras foram criadas, é o caso da opção Gana e da África do Sul.

Tabela 4.5: Resultados das questões de múltipla escolha

Contexto	Referência	Gerado
Química	Water, a fundamental substance for sustaining life, is denoted by the chemical formula H2O, comprising two hydrogen atoms and one oxygen atom.	What chemical formula denotes water? **A)** H20 * **B)** Nitrogen **C)** Oxygen
Geografia	Brazil, the largest country in South America, boasts the Amazon Rainforest, the world's largest tropical rainforest, and is renowned for its vibrant culture, diverse ecosystems, and iconic landmarks like the Christ the Redeemer statue in Rio de Janeiro.	What country has the largest rainforest in South America? **A)** South Africa **B)** Ghana **C)** Brazil *
História	The French Revolution, a pivotal event in history, unfolded in the late 18th century, transforming the socio-political landscape of France. Marked by fervent calls for liberty, equality, and fraternity, it led to the overthrow of the monarchy and the rise of radical political ideologies.	What did the French Revolution call for liberty, equality, and what other element of the French Revolution? **A)** Greek Life **B)** Sorority **C)** Fraternity *

Na questão de história, o texto embasa a respeito da revolução francesa. O modelo gerou uma clássica pergunta: além da liberdade e igualdade, qual o outro principal lema da revolução francesa? E a resposta para essa questão, também encontrada no contexto, é Fraternidade.

Sendo assim, foi qualitativamente revelado a possibilidade de se trabalhar com as questões de múltipla escolha no modelo em questão. Esse formato de pergunta, que se assemelha ao modelo de exames como os vestibulares, é utilizado em sua maioria para a avaliação do conhecimento adquirido. Uma visualização mais completa desses resultados podem ser vista no apêndice A. Diante disso, as questões de múltipla escolha tem um forte apelo pedagógico que será explorado na seção a seguir.

4.4 Possíveis aplicações da GAQ

Os resultados revelados nas seções anteriores, validam a hipótese que é possível a Geração automatizada de questões com uma boa qualidade por modelos baseados em *transformers*. Diante disso, cabe a discussão de quais as aplicações possíveis para essa geração de questões, na prática, no contexto educacional.

A tabela 4.6 revela quais são as aplicações mais comuns dentro da literatura de GAQ segundo Kurdi et al. (2020). Esses dados serão utilizados como referência para justificarem as possíveis aplicações das GAQ.

O maior número de estudos na área de GAQ tem por objetivo auxiliar na avaliação de estudantes, contando com 40 estudos dentre os 86 avaliados, representando quase 50% do total. Tendo isso em vista, uma aplicação clara desses modelos é no auxílio da criação de exames.

No Brasil, o Exame Nacional do Ensino Médio (ENEM) [1] acontece desde 1998, e representa uma das maiores portas de entrada para estudantes nas universidades públicas. Diante disso, e tendo em vista que o exame do ENEM se pauta em questões de múltipla escolha, uma aplicação possível seria a utilização do modelo de GAQ para construção de questões desse sistema avaliativo.

Tabela 4.6: Relação entre estudos de GAQ e o seu propósito

Propósito	**Número de estudos**
Avaliação	40
Ensino geral	10
Autoaprendizagem	9
Apoio à aprendizagem	9
Tutoria assistida por computador	7
Fornecimento de questões práticas	8
Fornecimento de questões para *MOOC's* [2]	2
Aprendizagem ativa	1

Os outros propósitos existente corroboram principalmente com o suporte ao aluno, seja em tutorias assistidas ou em outras formas de apoio, e o estímulo da autoaprendizagem, seja por aplicações diretas ou exemplos como os processos de autoavaliação durante a execução de *MOOC's*.

A plataforma Cosseno [3], apresentada nas seções anteriores, tem como principal objetivo estimular a autoaprendizagem através de uma plataforma digital acessível. Tendo

[1] <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico>

[2] Os MOOCs (Massive Online Open Courses) são cursos online abertos que estão disponíveis para qualquer pessoa com acesso à internet e não exigem requisitos mínimos para quem pretende realizá-los.

[3] <https://cosseno.com/>

isso em vista, e sabendo que o site conta com ferramentas avaliativas para os estudantes, esse modelo de GAQ tem um grande valor para a plataforma.

De forma direta, uma aplicação plausível do modelo no Cosseno é a confecção de processos avaliativos no formato de *quiz*, reforçando conceitos básicos a respeito de uma determinada área do conhecimento. Sendo assim, em um visão de produto, essas questões poderiam ser diariamente resolvidas pelo aluno a depender das suas áreas de déficit. A Tabela 4.5 ilustra bem essa possibilidade, já que no tema de história, por exemplo, ao retratar sobre um texto da revolução francesa, o modelo construiu uma pergunta a respeito dos três pilares desse acontecimento histórico. Esse mesmo conceito é explorado em vários exames de vestibulares existentes[4].

Outra ideia de produto, explorando a questão do apoio contínuo ao aluno, é a criação de um assistente de perguntas e respostas durante a leitura de materiais. Dessa maneira, durante esse estudo, o aluno poderia consultar uma ferramenta para criar questões sobre aquele tema, reforçando palavras-chave essenciais e reforçando o ensino ativo do aluno, que configura um dos métodos mais eficientes de estudo (Karpicke e Blunt (2011)).

Por fim, é possível observar as diversas aplicações da tarefa de GAQ e a sua relevância educacional. Foi observado possibilidades de aplicações nacionalmente, como no caso do ENEM, bem como a aplicação em soluções online como o Cosseno.

[4]Exemplo de questão explorando os pilares da revolução francesa: <https://cosseno.com/q/f7rozw138>

Capítulo 5

Conclusão e sugestões

Durante esse trabalho, foi discutido a estrutura necessária para realizar a tarefa de Geração automatizada de questões (GAQ). Um foco especial foi dado para a abordagem na língua portuguesa, até então pouco explorada nesse tipo de atividade. O desenvolvimento de um modelo eficaz de GAQ representa um avanço significativo no campo de Processamento de linguagem natural (PLN), especialmente no cenário educacional. Esse trabalho reforçou todo esse ganho pedagógico, especialmente aplicado ao cenário brasileiro.

No primeiro módulo da solução, em que é necessário extrair palavras-chave de um contexto para a construção de respostas, foi observado que o modelo em inglês e em português atingiu métricas compatíveis com a literatura. Além disso, foi observado que o modelo em inglês se sobressai nessa tarefa, que já era esperado pelo fato do modelo T5, utilizado na solução, ser majoritariamente treinado na língua inglesa.

No módulo seguinte, na abordagem da geração de questões efetivamente, comprovou-se a efetividade do modelo de GAQ em português. Para atingir isso, além do processo de reprodução de resultados, otimizações como o destaque de respostas nos contextos foram essenciais para o sucesso da solução. Foi avaliado que o modelo inglês teve métricas superiores, mas apesar disto, o modelo em português teve um resultado próximo.

Após a extração das respostas e da geração das questões, foi explorada a geração de questões de múltipla escolha. Nessa análise, em sua maioria qualitativa, foi percebido que existe a possibilidade de questões desse tipo, que se assemelham em sua maioria com questões de exames, tais como os vestibulares.

Por fim, diante de uma solução fundamentalmente educacional, foram discutidas quais seriam as possíveis aplicações práticas do modelo na realidade. Em cima disso, foram avaliados os casos comuns de uso na literatura, tal como o uso para realização de sistemas avaliativos, bem como outras funções mais específicas da realidade brasileira. Um exemplo prático foi a possível contribuição do sistema de questões múltipla escolha para produzir questões do ENEM.

Ademais, foram discutidas ideias de produto utilizando a GAQ na plataforma Cosseno. Nesse contexto, as soluções buscam resolver áreas como a realização de testes, o estímulo da autoaprendizagem e o estudo de uma maneira ativa e dinâmica.

A partir de todo esse desfecho, o trabalho atingiu o resultado esperado e comprovou a hipótese de que é possível realizar a tarefa de Geração automatizada de questões na língua portuguesa tendo um impacto na educação e no contexto do Processamento de linguagem natural.

5.1 Possibilidades de estudos futuros

O estudo da tarefa de GAQ é relativamente recente, e atingiu seu estado da arte principalmente com arquiteturas mais robustas como as baseadas em *transformers* (VASWANI et al., 2017). Além disso, por se tratar de uma área com diversas aplicações, são inúmeras as possibilidades de estudos futuros para melhorar e incrementar esse estudo. Diante disso, nessa seção serão apresentados os principais tópicos observados como possibilidade de estudo futuro.

Nesse trabalho, o *T5* foi escolhido pela sua flexibilidade e eficiência, comprovada durante o estudo. Apesar disso, existem outras grandes arquiteturas capazes de resolver esta tarefa, até mesmo outros exemplos de modelo baseada em *transformers*, como no caso do *BERT*. Logo, uma possibilidade é a realização de um estudo comparativo entre modelos na tarefa de GAQ. Para isso, é importante o estudo mais aprofundado de métricas e otimizações para comparar os modelos.

Ademais, existem outros formatos de geração além do *answer-aware*, trabalhado nesse estudo. Um exemplo disso é o *pipeline end-to-end (E2E)* , em que um único modelo é responsável por extrair palavras-chave e gerar as questões. Sendo assim, em outras oportunidades de estudo, é interessante realizar outras abordagens e até mesmo compará-las.

Dependendo da aplicação do GAQ, pode ser que o modelo necessite de um conhecimento mais específico. Diante disso, é possível melhorar os resultados do modelo nessa realidade utilizando conhecimento externo (Jia et al. (2021)), se consolidando como outra possibilidade de estudo.

E por fim, o estudo das questões de múltipla escolha é uma atividade de maior complexidade em comparação com questões de resposta aberta (RAINA; GALES, 2022). É preciso a melhoria das métricas de avaliação com métodos mais robustos como *diversidade* e *complexidade* (RAINA; GALES, 2022) para ser capaz de uma boa avaliação. Portanto, esse tema pode ser facilmente melhor discutido em trabalhos futuros.

(LIN, 2004) (PAPINENI et al., 2002) (BANERJEE; LAVIE, 2005)

Referências Bibliográficas

BANERJEE, S.; LAVIE, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: GOLDSTEIN, J. et al. (Ed.). *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. p. 65–72. Disponível em: <https://aclanthology.org/W05-0909>. 17, 41

CARMO, D. et al. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*, 2020. 3, 24

CHAN, Y.-H.; FAN, Y.-C. A Recurrent BERT-based Model for Question Generation. In: FISCH, A. et al. (Ed.). *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 154–162. Disponível em: <https://aclanthology.org/D19-5821>. 23

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. Disponível em: <http://arxiv.org/abs/1810.04805>. 12, 14

FIROOZEH, N. et al. Keyword extraction: Issues and methods. *Natural Language Engineering*, v. 26, n. 3, p. 259–291, maio 2020. ISSN 1351-3249, 1469-8110. Publisher: Cambridge University Press. Disponível em: <https://www.cambridge.org/core/journals/natural-language-engineering/article/keyword-extraction-issues-and-methods/84BFD5221E2CA86326E5430D03299711>. 15, 16, 25, 26, 30

GÉRON, A. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2017. ISBN 9781491962299. Disponível em: <https://books.google.com.br/books?id=I6qkDAEACAAJ>. 21

JIA, X. et al. *Enhancing Question Generation with Commonsense Knowledge*. arXiv, 2021. ArXiv:2106.10454 [cs]. Disponível em: <http://arxiv.org/abs/2106.10454>. 41

KARPICKE, J. D.; BLUNT, J. R. Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science*, v. 331, n. 6018, p. 772–775, fev. 2011. Publisher: American Association for the Advancement of Science. Disponível em: <https://www.science.org/doi/10.1126/science.1199327>. 5, 39

KURDI, G. et al. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in*

Education, v. 30, n. 1, p. 121–204, mar. 2020. ISSN 1560-4306. Disponível em: <https://doi.org/10.1007/s40593-019-00186-y>. x, 5, 7, 8, 28, 32, 33, 38

LIN, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: <https://aclanthology.org/W04-1013>. 18, 41

LIU, C.-W. et al. *How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation*. arXiv, 2017. ArXiv:1603.08023 [cs]. Disponível em: <http://arxiv.org/abs/1603.08023>. 17

NGUYEN, H. A. et al. Towards Generalized Methods for Automatic Question Generation in Educational Domains. In: HILLIGER, I. et al. (Ed.). *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*. Cham: Springer International Publishing, 2022. v. 13450, p. 272–284. ISBN 978-3-031-16289-3 978-3-031-16290-9. Series Title: Lecture Notes in Computer Science. Disponível em: <https://link.springer.com/10.1007/978-3-031-16290-9_20>. 23

PAPINENI, K. et al. Bleu: a Method for Automatic Evaluation of Machine Translation. In: ISABELLE, P.; CHARNIAK, E.; LIN, D. (Ed.). *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. p. 311–318. Disponível em: <https://aclanthology.org/P02-1040>. 16, 41

PRINCE, M. Does Active Learning Work? A Review of the Research. *Journal of Engineering Education*, v. 93, n. 3, p. 223–231, jul. 2004. ISSN 1069-4730, 2168-9830. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/j.2168-9830.2004.tb00809.x>. 2

RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. Disponível em: <http://arxiv.org/abs/1910.10683>. ix, 13, 14, 21

RAINA, V.; GALES, M. *Multiple-Choice Question Generation: Towards an Automated Assessment Framework*. arXiv, 2022. ArXiv:2209.11830 [cs]. Disponível em: <http://arxiv.org/abs/2209.11830>. 9, 15, 28, 34, 41

THALHEIMER, W. The Learning Benefits of Questions. 2003. 6, 34

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. ix, 3, 10, 11, 41

ZOU, B. et al. Automatic True/False Question Generation for Educational Purpose. In: KOCHMAR, E. et al. (Ed.). *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Seattle, Washington: Association for Computational Linguistics, 2022. p. 61–70. Disponível em: <https://aclanthology.org/2022.bea-1.10>. 6

Apêndice A

Exemplos detalhados de GAQ

Com o objetivo de deixar mais transparente os resultados da Geração automatizada de questões, serão expostos vários exemplos de geração a seguir.

A.1 Questões abertas

Para o tipo de questões abertas, serão visualizadas as gerações para o modelo inglês, português e inglês traduzido. Para exibir cada categoria, foi selecionado um exemplo do conjunto de teste do SQuAD, presente na tabela A.1

Tabela A.1: Exemplo do conjunto de teste do SQuAD

ID SQuAD	*56d6edd00d65d21400198250*
Título SQuAD	Super_Bowl_50
Contexto SQuAD (en)	In early 2012, NFL Commissioner Roger Goodell stated that the league planned to make the 50th Super Bowl "spectacular"and that it would be "an important game for us as a league".
Contexto SQuAD (pt)	No início de 2012, o comissário da NFL Roger Goodell afirmou que a liga planejava tornar o 50º Super Bowl "espetacular"e que seria "um jogo importante para nós como liga".
Pergunta SQuAD (en)	Who is the commissioner of the NFL?
Pergunta SQuAD (pt)	Quem foi o comissário da NFL em 2012?
Respostas SQuAD	"text": ["Roger Goodell", "Roger Goodell", "Goodell"], "answer_start": [32, 32, 38] }
Pergunta gerada (en)	Who was NFL commissioner in 2012?
Pergunta gerada (pt)	Quem foi o comissário da NFL em 2012?
Pergunta gerada (pt - traduzido)	Quem é o comissário da NFL no início de 2012?

Além das gerações, complementando a abordagem feita no capítulo de resultados utilizando a tabela 4.4, serão exibidos também os resultados em inglês na tabela A.2:

Tabela A.2: Exemplos diversos de GAQ com a métrica METEOR

Contexto	**Referência**	**Gerado**	**MTR**
... [HL] The American Football Conference (AFC) champion **Denver Broncos** defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. [HL] ...	Which NFL team won Super Bowl 50?	Who defeated the Carolina Panthers 24-10?	0.0000 (menor)
[HL] Within the genitourinary and gastrointestinal tracts, commensal flora serve as biological barriers by competing with pathogenic bacteria for food and space and, in some cases, by changing the conditions in their environment, such as **pH or available iron** [HL] ...	Commensal flora can change what specific conditions of their environment in the gastrointestinal tract?	What conditions do commensal flora change in their environment?	0.3553 (média)
... [HL] During the construction of a building, **the municipal building inspector** inspects the building periodically to ensure that the construction adheres to the approved plans and the local building code. [HL] ...	Who inspects the building periodically to ensure that the construction adheres to the approved plans and the local building code?	Who inspects the building periodically to ensure that the construction adheres to the approved plans and the local building code?	0.9999 (maior)

A.2 Questões de múltipla escolha

Durante a discussão qualitativa das questões de múltipla escolha, foi priorizado o uso da língua inglesa. Para demonstrar a validade desse tipo de questão na língua portuguesa, segue um resultado de exemplo:

Contexto:
Quando a FCC impôs suas regras de fin-syn em 1970, a ABC criou proativamente duas empresas: a Worldvision Enterprises como distribuidora de distribuição e a ABC Circle Films como produtora. No entanto, entre a publicação e a implementação desses regulamentos, a separação do catálogo da rede foi realizada em 1973. Os direitos de transmissão de produções anteriores a 1973 foram transferidos para a Worldvision, que se tornou independente no mesmo ano. A empresa foi vendida várias vezes desde que a Paramount Television a adquiriu em 1999, e mais recentemente foi absorvida pela CBS Television Distribution, uma unidade da CBS Corporation. No entanto, a Worldvision vendeu partes de seu catálogo, incluindo as bibliotecas Ruby-Spears e Hanna-Barbera, para o Turner Broadcasting System em 1990. Com a compra da ABC pela Disney em 1996, a ABC Circle Films foi absorvida pela Touchstone Television, uma subsidiária da Disney que, por sua vez, foi renomeada ABC Studios em 2007.

Questão 1: Quem comprou a ABC em 1996? **Resposta 1:** Disney
A. Pixar
B. Universal
* C. Disney

Questão 2: Quando a FCC impôs suas regras de fin-syn? **Resposta 2:** 1970
A. 1968
B. 1965
* C. 1970

Questão 3: Quando a ABC se separou do catálogo da rede? **Resposta 3:** 1973
A. 1968
* B. 1973
C. 1966

Indo além, o mesmo processo foi repetido para o modelo em inglês, a fim de validar o potencial de gerar múltiplas perguntas para o mesmo contexto:

Contexto:
"The Islamic State", formerly known as the "Islamic State of Iraq and the Levant"and before that as the "Islamic State of Iraq", (and called the acronym Daesh by its many detractors), is a Wahhabi/Salafi jihadist extremist militant group which is led by and mainly composed of Sunni Arabs from Iraq and Syria. In 2014, the group proclaimed itself a caliphate, with religious, political and military authority over all Muslims worldwide. As of March 2015[update], it had control over territory occupied by ten million people in Iraq and Syria, and has nominal control over small areas of Libya, Nigeria and Afghanistan. (While a self-described state, it lacks international recognition.) The group also operates or has affiliates in other parts of the world, including North Africa and South Asia.

Questão 1: What religious affiliation did the Islamic State declare itself?

Resposta 1: Caliphate
A. Islamic State
B. Holy War
C. Caliphate *

Questão 2: Who is the main members of the Islamic State?

Resposta 2: Sunni Arabs
A. Kurds
B. Sunnis
C. Sunni Arabs *

Questão 3: What does the Islamic State lack?

Resposta 3: International recognition
A. International recognition *
B. Regional Autonomy
C. Independence