PAULO HENRIQUE DOS SANTOS E SILVA
VICTOR BENITO RAFAEL ALVES PEREIRA

CHATBOT SOBRE A AMAZÔNIA AZUL BASEADO EM MODELO DE RESPOSTAS A PERGUNTAS

São Paulo
2022

PAULO HENRIQUE DOS SANTOS E SILVA
VICTOR BENITO RAFAEL ALVES PEREIRA

CHATBOT SOBRE A AMAZÔNIA AZUL BASEADO EM MODELO DE RESPOSTAS A PERGUNTAS

Trabalho apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro Mecatrônico.

Orientador:
Fabio Gagliardi Cozman

São Paulo
2022

Este trabalho é dedicado a Maria Francisca dos Santos, Alcebino Viera dos Santos, Herceli Inocencio da Silva, Maria Helena Francisca dos Santos e Silva e Brunna Caroline dos Santos e Silva, pedras fundamentais na minha formação.

Paulo Henrique

Dedico este trabalho a minha mãe e pai por todo o incentivo, apoio e suporte que sempre me deram, a minha irmã, que se tornou grande companheira de vida, e cujas conversas tornaram muitos trânsitos toleráveis, e aos gatos, que me trazem alegria mesmo nos breves momentos que esqueço da existência desse sentimento.

Victor Benito

AGRADECIMENTOS

Agradecemos ao professor Fabio Cozman por todo o suporte, orientação e preocupação durante a execução do trabalho.

A todos os membros do projeto BLAB, que apoiaram o trabalho, nos receberam de braços abertos, e forneceram sugestões e feedbacks valiosos.

Aos grandes amigos do Centro Moraes Rêgo, que acompanharam toda nossa jornada, sendo peças fundamentais do início à fase final da graduação na Escola Politécnica da USP. Esperamos que a amizade que construímos seja como os banquinhos de pedra: resistente, dura, quando necessário, mas sempre lar de boas risadas, momentos e histórias ao fim do dia.

À Cristina Rodrigues por todo apoio ao autor Paulo Henrique durante a execução do trabalho.

"O céu noturno é interessante. Há desenhos. Mesmo sem tentar, podemos imaginá-los. No céu setentrional, por exemplo, há um desenho ou constelação, que parece um ursinho. Alguns povos chamam-na A Grande Ursa. Outros vêem imagens bem diferentes. Estes desenhos não são, naturalmente, reais no céu noturno, nós é que os colocamos lá. Fomos caçadores e vemos caçadores e cães, ursos e jovens mulheres, objetos do nosso interesse. [...] Se as constelações tivessem recebido nomes no século XX, suponho que veríamos bicicletas e refrigeradores no céu, "estrelas" de rock-and-roll e talvez até nuvens em cogumelo - um novo grupo de esperanças e receios humanos entre as estrelas."

-- Carl Sagan

RESUMO

É de amplo conhecimento que a região costeira do Brasil é de suma importância para o país, dada sua larga extensão, valor econômico e riquezas ambientais. Tal região foi denominada Amazônia Azul. Assim, com a finalidade de disseminar o conhecimento a respeito da região bem como contribuir para a pesquisa científica no âmbito de *Natural Language Processing* em português, desenvolveu-se um agente conversacional capaz de responder perguntas relacionadas ao tema da Amazônia Azul. Foi utilizada a infraestrutura do *Watson Assistant* da IBM, populando uma base de perguntas e respostas relacionadas ao tema a fim de obter um sistema de *chatbot* capaz de endereçar grande parte dos assuntos potencialmente questionados por usuários interessados em aprender sobre o tema. Usuários diversos testaram o *chatbot* desenvolvido e avaliaram qualitativamente suas experiências. Os resultados qualitativos demonstraram suficientes graus de satisfação e os resultados quantitativos explicitaram elevados níveis de performance em métricas de avaliação de sistemas de resposta a perguntas.

Palavras-Chave – Amazônia Azul, *chatbot*, agentes conversacionais, *Watson Assistant*

ABSTRACT

It is widely known that the coastland of Brazil is of high importance for the country, given its large extension, economic value and environmental wealth. This region was called the Blue Amazon. Thus, with the goal of disseminating knowledge about the region as well as contributing to scientific research within the scope of Natural Language Processing in Portuguese, a conversational agent capable of answering questions related to the theme of the Blue Amazon was developed. IBM's Watson Assistant infrastructure was used, populating a base of questions and answers related to the topic in order to obtain a chatbot system capable of addressing most of the subjects potentially asked by users interested in learning about the theme. Different users tested the developed chatbot and qualitatively evaluated their experiences. The qualitative results showed sufficient degrees of satisfaction and the quantitative results also showed high levels of performance in metrics for evaluating question answering systems.

Keywords – Blue Amazon, chatbot, conversational agents, Watson Assistant

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

É clara a importância da consciência social sobre recursos naturais e sua relevância para o sustento de nossa e outras espécies. Com informação a respeito das riquezas naturais e do impacto humano na devastação das mesmas, a sociedade é capaz de tomar ações focadas em preservar o meio, dado que, para que se possa agir na solução de um problema, primeiro é necessário que sua existência e impacto sejam devidamente difundidos.

Assim, especialmente em um país como o Brasil, que apresenta reservas ambientais chaves para a manutenção dos principais ecossistemas do planeta, nota-se a clara necessidade de desenvolvimentos no sentido de difundir informações relevantes, solucionar dúvidas diversas e promover consciência social a respeito da preservação do meio ambiente. Além disso, vale ressaltar o impacto considerável e cada vez mais visível das mudanças climáticas ocasionadas pela devastação desse meio.

No cenário científico do problema a ser abordado pelo trabalho, atualmente, a maioria dos recursos para pesquisa sobre *Natural Language Processing* (NLP) existentes focam na língua inglesa, e dependem de uma corpora cuidadosamente produzida [18], de modo que existe ainda considerável carência de conteúdos NLP na língua portuguesa. Dado que o desenvolvimento de soluções utilizando essa ciência depende diretamente do avanço científico em línguas específicas, a escassez de recursos vem sendo uma barreira para o desenvolvimento de soluções em português. Isto é, o fato de soluções voltadas à língua inglesa não serem suficientemente reaproveitáveis para outros idiomas torna a pesquisa científica relacionada a NLP para o idioma português defasada em comparação com o que já foi desenvolvido para idiomas mais utilizados em uma escala global.

Com tal escassez, nota-se consequentemente a carência de desenvolvimento científico em NLP afetando também no contexto de sistemas de interação automatizadas de máquina e humano, através da tecnologia de *chatbots*. Uma das competências que podem ser atribuídas aos *chatbots* seria a capacidade de responder de maneira automatizada perguntas de usuários. Porém tal habilidade depende do desenvolvimento de NLP voltada para a língua portuguesa, para que a solução possa ser empregada no contexto brasileiro.

1.1 Motivação

1.1.1 Interesse Social

Com as evidentes mudanças climáticas e exploração desenfreada que o planeta vem sofrendo no decorrer do século, pautas ambientais têm sido de extrema relevância nos tempos atuais. Com isto, o mundo tem voltado esforços para o desenvolvimento de soluções efetivas a fim de mitigar os efeitos adversos no ambiente devido a ações humanas.

No Brasil, tal pauta se torna ainda mais relevante, considerando nossa imensa biodiversidade e recursos naturais, sendo grande foco dos olhares dado seu grande potencial de destruição mas também de ser foco de preservação mundial. Tratando a respeito do foco de preservação brasileiro, o país possui uma imensa região costeira, repleta de recursos naturais, os quais possuem grande necessidade de preservação. Tal região é denominada como Amazônia Azul.

Segundo o site Portal Amazônia [19], a Amazônia Azul abrange uma área de 4,5 milhões de quilômetros, com uma extensa biodiversidade de vida marinha, e contando com espécies marítimas só existentes na região, como corais, esponjas e peixes. Além disso, a página do ecycle [16] cita exemplos de recursos energéticos capazes de desacelerar o aquecimento global na região, como a geração de energia elétrica a partir de processos marinhos dinâmicos como ondas, correntes e marés.

Estas e outras informações evidenciam a relevância da Amazônia Azul no cenário ambiental global. Essas informações, apesar de robustas tanto em qualidade quanto em quantidade, não são amplamente difundidas no país, assim, o objetivo do projeto tem como principal motivação difundir, facilitar e democratizar o acesso da sociedade brasileira às informações a respeito da Amazônia Azul, grande patrimônio ambiental do país.

1.1.2 Mérito Científico

O desenvolvimento de modelos e aplicações de QA e NLP é relevante para a resolução do problema proposto, dado que podem ser elaborados de forma a serem capazes de atingir a população geral, agregando com aplicações pouco exploradas das técnicas mencionadas. Além disso, a execução desse projeto significa também a ampliação do conteúdo sobre sistemas QA, NLP e *chatbot* focados na língua portuguesa na literatura.

2 OBJETIVOS

Dado o problema em questão, tem-se como o objetivo geral do trabalho o desenvolvimento de um agente conversacional em português, o qual coerentemente é capaz de responder a perguntas de usuários a respeito da Amazônia Azul, de forma fluida e simples para o usuário, proporcionando, também, aprendizados sobre o tema. O público-alvo do trabalho consiste em pessoas com suficiente capacidade de leitura, escrita e interpretação de textos, interessadas em aprender mais sobre conceitos gerais básicos envolvendo a Amazônia Azul.

Modelos de *Machine Learning* e inteligência artificial (IA) no geral são responsáveis por fazer computadores performarem sofisticadamente tarefas sem qualquer intervenção de seres humanos, através de aprendizado [2]. Mais detalhadamente, em nosso caso de uso, ao final do trabalho, o agente deve ser capaz de, a partir de uma base de dados a respeito da Amazônia Azul, extrair uma resposta coerente para a pergunta de entrada. Caso não haja a resposta à pergunta dentro da base, o modelo deverá retornar uma resposta padrão ao usuário, informando que não possui informações suficientes para responder a indagação.

Por fim, o sucesso deste Trabalho de Conclusão de Curso significa uma singela contribuição ao projeto de pesquisa em *Knowledge-Enhanced Machine Learning for Reasoning on Ocean Data* conduzido pelo *Center for Artificial Intelligence* (C4AI) do InovaUSP. Tendo como um dos objetivos do projeto de pesquisa "desenvolver uma estrutura para agentes conversacionais que possam responder a consultas de alto nível ao longo do tempo em um domínio particular, incluindo questões, argumentos, causas, explicações, inferências e planos sobre tarefas específicas" [7].

3 CONCEITOS TEÓRICOS

Esta seção é dedicada à definição e aprofundamento a respeito de conceitos importantes para a compreensão do projeto.

3.1 *Chatbot*

Chatbots são agentes conversacionais usados como interfaces de linguagem natural para o fornecimento de dados e/ou serviços, que oferecem respostas instantâneas ao usuário [3, 17].

3.2 *Natural Language Processing*

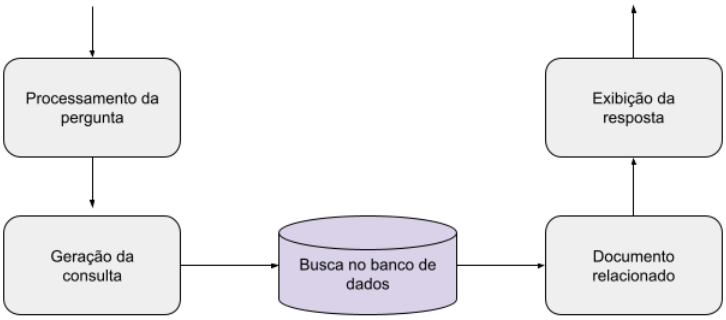
[13] define *Natural Language Processing* (NLP) como um conjunto de técnicas computacionais dedicadas a analisar e representar textos os quais ocorrem naturalmente em um ou mais níveis de análise linguística.

A partir disto, nota-se a usabilidade da utilização de técnicas de NLP neste trabalho, para a compreensão da pergunta do usuário bem como a obtenção da resposta adequada.

3.3 Sistemas de respostas a perguntas (*QA Systems*)

Em [11], sistemas de resposta a perguntas (Sistemas QA) é basicamente detalhado como sistemas os quais usuários fazem perguntas, o sistemas retira a resposta mais correta e apropriada àquela pergunta e a retorna ao usuário. A arquitetura geral de sistemas QA é denotada na figura 1:

Figura 1: Arquitetura básica de sistemas QA

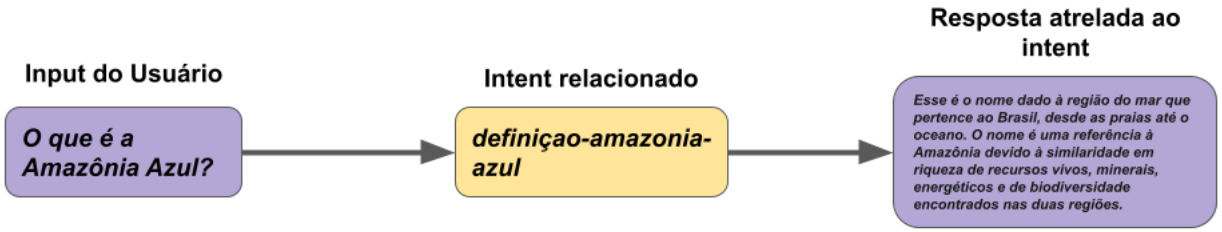


Fonte: Adaptada de [11]

3.4 *Intent*

Em sua documentação, a IBM define *intent* como propósitos ou objetivos que são expressados em um *input* de um usuário, como por exemplo, responder uma pergunta [9]. Reconhecendo o *intent* expressado pelo *input* do usuário, o *Watson Assistant* consegue direcionar o melhor fluxo de diálogo para responder ao usuário.

Figura 2: Exemplo de caso de uso de um *intent*



Fonte: Os autores

3.5 Habilidade de Conversação

Segundo a IBM, as habilidades de conversação do *Watson Assistant* são responsáveis por entender e endereçar perguntas ou pedidos feitos pelos usuários através de NLP e *machine learning*. Na interface do *Watson Assistant* os desenvolvedores fornecem as informações sobre determinado assunto ou tarefa e como o usuário solicita tal informação. A partir disso, o *Wastson Assistant* é capaz de construir dinamicamente um modelo de aprendizado de máquinas customizado para entender as solicitações iguais ou parecidas com as definidas pelos desenvolvedores.

3.6 Habilidade de Diálogo

A IBM fornece as habilidades de diálogos, que são um conjunto de editores, a fim de definir dados de treinamento e conversa. A conversa no assistente é representada como uma árvore de diálogo.

3.7 Métricas de avaliação de sistemas QA

3.7.1 *Exact-Match Score*

No contexto de sistemas QA, a métrica de avaliação de modelos *Exact-Match Score* corresponde a: para cada par pergunta-resposta, caso a resposta prevista pelo modelo seja exatamente igual à resposta correta, temos EM = 1, caso contrário EM = 0. Com isso, tal métrica é útil para a avaliação de modelos onde não há uma variação e abrangência de respostas corretas possíveis. A tabela 1 a seguir evidencia exemplos de possíveis resultados para essa métrica:

Tabela 1: Exemplos de *EM-Score* para diferentes perguntas e respostas

Pergunta	**resposta Correta**	**Previsão de Resposta do Modelo**	**EM Score**
Qual é a cor do céu?	Azul	Azul	1
Em qual continente se localiza a Alemanha?	Europa	África	0
Onde se localiza a cidade Recife?	Pernambuco	Pernambuco	1
Onde se localiza a cidade Salvador?	Bahia	Brasil	0

Fonte: Os autores

Note que para a pergunta "*Onde se localiza a cidade Salvador?*" o *Exact-Match Score* da predição é 0, porém "*Brasil*" é também uma resposta correta para a pergunta em questão. Sendo assim, deve-se ter muita cauleta ao utilizar-se do *EM-Score* já que o mesmo pode retornar uma avaliação ruim a modelos com bons desempenhos ao responder perguntas. Com isso, nota-se a importância de utilizar-se de mais de uma métrica de avaliação em sistemas QA.

3.7.2 *F1-Score*

O *F1-Score* é apropriado e muito utilizado quando os resultados de precisão e *score* são igualmente importantes para a boa performance de modelos, Mais especificamente no contexto de modelos QA, o *F1-Score* é calculado comparando as palavras individuais na previsão do modelo as palavras na resposta verdadeira. O número de palavras compartilhadas entre a previsão e a resposta correta serve como base para o cáculo do *F1-Score*.

$$F_1 = \frac{2 * precisão * recall}{precisão + recall} \tag{3.1}$$

Conceitualmente, o *F1-Score* é uma média harmônica entre a precisão e o recall, como

mostra a equação (3.1). Para modelos QA, a precisão será a razão entre o número de palavra compartilhadas entre predição e resposta real e o número de palavras na previsão. Já o recall é a razão entre o número de palavra compartilhadas entre predição e resposta reale o número de palavras da resposta real.

3.7.3 *Rouge-L Score*

A métrica *Rouge-L Score* é também um bom indicador para medir a qualidade de um modelo de NLP. Para sistemas QA, tal *score* mede a mais longa subsequência de palavras comum entre a resposta prevista pelo modelo QA e a resposta correta da pergunta em questão. Assim, o *score* conta a mais longa sequência de *tokens* compartilhados entre predição e resposta.

Tabela 2: Exemplos de *Rouge-L Score* para diferentes perguntas e respostas

Pergunta	**resposta Correta**	**Previsão de Resposta do Modelo**	**Rouge-L Score**
Qual é a cor do céu?	Azul	A cor do céu é azul	1
Qual é a maior árvore do mundo?	Sequoia Gigante	A maior árvore do mundo é uma sequoia gigante	2
Em qual Estado se localiza a cidade Recife?	Estado do Pernambuco	Recife é localizado no Pernambuco	1
Qual é a moeda da Inglaterra?	Libra Esterlina	Libra	1

Fonte: Os autores

O racional por trás é de que quanto mais longa é uma sequência compartilhada entre resposta e predição, mais eficiente se torna o modelo. Além disso, é possível, a partir do racional do *Rouge-L Score*, calcular o recall e a precisão do modelo com as equações (3.2) e (3.3) a seguir:

$$Recall = \frac{Rouge\text{-}L\ Score}{\text{número de tokens da resposta correta}} \tag{3.2}$$

$$Precisão = \frac{Rouge\text{-}L\ Score}{\text{número de tokens da previsão}} \tag{3.3}$$

Assim, como visto na seção 3.7.2 com o recall (3.2) e a precisão (3.3), é possível calcular o *F1-Score*. Por fim, na tabela 3 é exibida a precisão e recall dos exemplos da tabela 2:

Tabela 3: Exemplos de recall e precisão a partir do *Rouge-L Score* para diferentes perguntas e respostas

Pergunta	**resposta Correta**	**Previsão de Resposta do Modelo**	**Rouge-L Score**	**Recall**	**Precisão**
Qual é a cor do céu?	Azul	A cor do céu é azul	1	100%	17%
Qual é a maior árvore do mundo?	Sequoia Gigante	A maior árvore do mundo é uma sequoia gigante	2	100%	22%
Em qual Estado se localiza a cidade Recife?	Estado do Pernambuco	Recife é localizado no Pernambuco	1	33%	20%
Qual é a moeda da Inglaterra?	Libra Esterlina	Libra	1	50%	100%

Fonte: Os autores

4 REVISÃO DO ESTADO DA ARTE

Esta seção apresenta uma revisão da literatura a respeito de sistemas que utilizam NLP com a finalidade de responder perguntas de usuários de maneira automatizada, assim como fatores e conceitos relevantes que circundam o tema.

4.1 Sistema de respostas à perguntas em português

Dada a disponibilidade de consideráveis bases textuais sobre temas ecológicos, a utilização de sistemas QA e dos avanços em NLP têm importante espaço para aumentar a consciência social, entendimento geral e até mesmo promover o desbanque de informações falsas sobre o meio ambiente.

Nesse sentido, [4] apresenta o primeiro sistema QA na literatura focado em problemas ambientais do Brasil baseado na língua portuguesa. Foram montados dois sistemas *Retriever-Reader*: um apenas com o módulo de *Reader* contendo um módulo de linguagem, de modo que o modelo responde à questões apenas acessando informações guardadas em seus próprios parâmetros (*Reader-only*), baseado em T5 Pré-treinado (PTT5), enquanto o outro consiste em um *Reader* e um *Retriever* baseado no algoritmo BM25 com acesso à uma base de documentos, que, dada uma busca, tinham suas relevâncias estimadas pelo algoritmo. Além disso, também foi testada a influência de *fine-tuning* com pares QA de treino em cada sistema.

Inicialmente, como base de dados, foram gerados *datasets* baseados em documentos textuais e um outro composto por pares QA. Os primeiros se resumem a 17k artigos associados ao tema "Meio ambiente do Brasil" filtrados da Wikipédia em português através de um *script* recursivo, assim como 29k notícias relacionadas ao tema retiradas dos três principais jornais do Brasil, e, unindo os artigos às notícias formou-se um novo *dataset* com 46k elementos. Já os pares QA, utilizados para aplicar *fine-tuning* nos modelos, foram obtidos por meio da filtragem e posterior tradução com API do Google da base PAQ [12], que originalmente apresenta 65M de pares QA, mas após filtragem dos temas

relevantes, formaram-se 14k pares, de qualidade posteriormente averiguada por amostras.

Assim, foram realizados experimentos com os dois sistemas (*Reader-only* e *Retriever-Reader*) e a presença ou não de *fine-tuning*. Como resultados, notou-se que a presença de *fine-tuning* foi responsável por aumentos relevantes na métrica de *F1-score*, além disso, os sistemas suportados apenas pelo *dataset* composto pelos artigos da Wikipédia obtiveram resultados melhores do que os com acesso também aos dados de notícias, contrariando as expectativas. Os experimentos com o sistema *Reader-only* obtiveram os menores resultados de *F1-score* em suas categorias, e o melhor modelo apresentou um *F1-score* de 36,2.

Finalmente, possíveis melhorias em futuros trabalhos baseados em [4], como este, seriam a integração com módulos como *chatbots* sociais, que poderiam tornar o sistema mais apelativo para usuários. Como o *dataset* de treino utilizado apresentava respostas curtas e sucintas, outra melhoria poderia ser a construção de um sistema que fornece respostas mais elaboradas e completas, filtrando as absurdas.

Em [5], foi realizado o desenvolvimento de um *chatbot* com capacidade de responder perguntas sobre a Amazônia Azul, baseado no *dataset* Question Answering V2.1 do MS MARCO, informações de artigos e outras fontes relevantes de informações ambientais. Após validação do *chatbot* por usuários, notou-se que o instrumento é, de fato, relevante no sentido de aumentar a consciência social e difundir conhecimentos acerca da Amazônia Azul. Como melhorias relevantes para futuros trabalhos, destaca-se popular a base de dados com informações e opções de expansão de respostas.

4.2 Formulação de sistemas de respostas a partir da geração de pares perguntas-respostas

Para atingir requisitos similares ao do projeto vemos diversas abordagem tomadas. Entre elas, [12] aborda o desenvolvimento das nomeadas *PAQs* (*Probably-Asked-Questions*). Para defini-las, primeiramente devemos apresentar algumas definições. Primeiramente, *Open-Domain Question Answering*, que em [22] é definido como uma técnica para encontrar uma resposta a determinada pergunta a partir de um massivo conjunto de documentos. Agora, podemos abordar outra técnica chamada *Closed Book Question Answering models*, os quais são capazes de diretamente mapear perguntas e respostas a partir de pares *Question-Answer*, sem a necessidade de um corpo de texto por trás. Tal método vem se mostrado principalmente capaz de memorizar pares QA dos dados de treino, sendo

menos efetivo na resposta de perguntas que não coincidem com os dados de treino. Dado isto, *Probably-Asked-Questions* em [12] é apresentado como uma base de conhecimentos semi-estruturada de 65 milhões de pares QA, de onde modelos podem memorizar ou até aprender a retirar respostas da base.

Assim, PAQ é um conjunto de pares perguntas-respostas gerados automaticamente a partir da Wikipedia que pode ser usado como base de dados para a formulação de sistemas de respostas a perguntas. Os pares perguntas-respostas são gerados em quatro passos. Primeiramente, o sistema utiliza-se de um modelo de seleção para encontrar passagens que são prováveis de serem perguntadas a seres humanos, sendo passíveis de gerar uma pergunta. No segundo passo, a partir da passagem selecionada, o componente de extração de resposta identifica possíveis trechos que são prováveis de constituir uma resposta a uma pergunta. Para tal, utiliza-se dois métodos: o já existente *Named Entity Recognizer* (NER) ou treinando um modelo extrator de respostas BERT [6]. NER somente extrai entidades nomeadas da passagem. Já o modelo BERT extrai a resposta modelando o inicio e o final da resposta independentemente, concatenando os dois trechos. Após isto, é gerada a provável pergunta a partir da passagem e da resposta.

Por fim, é passado um modelo de filtro o qual melhora a qualidade das perguntas geradas, assegurando suas consistências. O modelo verifica se a resposta obtida é uma provável resposta válida à pergunta. Ao aplicar este processo inúmeras vezes, foi assim alimentado o banco de dados do PAQ [12].

Assim, seria possível construir sistemas automatizados de respostas a perguntas de usuários com o projeto citado anteriormente. Para isso, basicamente seria necessário treinar o sistema a navegar e encontrar dentre os pares perguntas-respostas, a pergunta que mais se assemelha à pergunta do usuário e devolver como saída o par resposta daquela pergunta. Desse modo, seria possível se aproximar do cumprimento dos requisitos de projeto, porém, com a possibilidade de uma acurácia não satisfatória, já que tal modelo seria apenas eficiente em responder perguntas existentes no banco de dados. Visto que o contexto do projeto deste trabalho é demasiadamente específico (Amazônia Azul), somente a utilização do PAQ como banco de dados pode não ser uma alternativa.

Contudo, a riqueza de [12] para este trabalho também está em sua metodologia. Tal estudo se mostrou eficientemente capaz de, a partir de uma base extensa de conhecimentos que é a Wikipedia, extrair 65M de pares QA a respeito de diversos temas com níveis satisfatórios de coerência. Assim, o desenvolvimento do PAQ pode ser útil neste projeto ao aplicarmos sua metodologia de criação de pares perguntas-respostas a partir de nossa

base de conhecimentos específica da Amazônia Azul.

Datasets como o referido PAQ, ou mesmo RACE, SWAG e SQuAD (v 1.1 ou 2.0) são comumente usados como métricas para avaliação da capacidade de modelos QA responderem corretamente a perguntas [23]. A partir desses bancos, os modelos podem ser submetidos a testes QA padronizados que medem quão corretas são as respostas geradas.

O SQuAD consiste em um *dataset* de compreensão de leitura composto por pares pergunta/resposta gerados por *crowdsourcing* em uma série de artigos da Wikipédia. A partir de 2018, com sua versão 2.0, passou a conter também questões sem respostas com estruturas parecidas com as das respondíveis [20], de modo que é testada também a capacidade do modelo NLP de reconhecer se a reposta para a pergunta está disponível.

Já o RACE possui maior proporção de questões que requerem considerável raciocínio para compreensão de leitura em comparação com os demais, e as questões do SWAG avaliam a inferência e raciocínio fundamentados em senso comum, a partir de 113k questões múltipla escolha sobre situações fundamentadas. Assim, é possível se fazer uma análise do modelo a partir dos *scores* obtidos pelas métricas comuns [23].

Mesmo que existam alguns multilíngues, a maioria dos *datasets* e documentos para QA disponíveis atualmente são exclusivamente em inglês, de modo que o bilíngue no qual português é uma das línguas pioneiro é o *Pirá*, que também contém questões irrespondíveis para estudos com *answer triggering* (AT) [18]. Seus resultados incluem a geração manual de 2261 pares QA sobre a costa brasileira e o oceano divididas em três versões, fornecimento de dados de treino para sistemas QA bilíngues, e um método replicável para geração de *datasets*.

4.3 Modelos *Retriever-Reader*: Extração de informações de textos

Outra possível solução do problema gira em torno de retirar passagens de textos que potencialmente podem responder a perguntas dos usuários [1]. Conceitualmente, o modelo de NLP poderia receber uma pergunta de entrada do usuário, e, a partir de uma base de conhecimentos textual, procurar, extrair uma passagem com boa probabilidade de responder coerentemente à pergunta e formula a resposta a ser retornada. Isso se diferencia modelos de criação ou consulta a pares perguntas-respostas, já que nestes sistemas, o sistema procura nos pares a pergunta que melhor se assemelha à pergunta inserida pelo usuário. Assim, nos modelos chamado de *Retriever-Reader*, o sistema busca diretamente

pela resposta, e não por perguntas semelhantes.

O artigo [22] aborda a estruturação de um modelo para *chatbot* baseado neste conceito de *Retriever-Reader*. Mais especificamente, [22] se propõe a desenvolver um chatbot que atual como um agente virtual para *troubleshooting*, fornecendo a informação correta a técnicos baseado em conteúdos de documentações de sistemas.

A adoção de processos *Retriever-Reader* podem ser divididas em três subtarefas. A primeira consiste em buscar documentos relevantes na base, os quais podem potencialmente conter a resposta à pergunta. Para esta tarefa [22] cita que pesquisadores tendem a usar técnicas de NLP mais tradicionais como *Text-Frequency Inverse Document Frequency* (TF-IDF) e *Elastic-Search*.

O segundo passo consiste em, dentro dos documentos relevantes obtidos, extrair potenciais candidatos a respostas. Esta tarefa envolve a aplicação de modelos de linguagem. Tais modelos pré-treinados com bases de textos massivas têm se mostrados extremamente eficientes na execução de diversas tarefas de NLP, especialmente aqueles afinados para propósitos específicos. No contexto em questão, é visto que muitos modelos de QA com bons resultados utilizam-se de modelos de linguagem pre-treinados.

Por fim, [22] explicita o terceiro passo como e re-ranqueamento dos candidatos a respostas extraídos com o objetivo de identificar a resposta correta. Para esta tarefa, trabalhos existentes na literatura costuma utilizar-se de redes neurais, que possuem como entradas a pergunta a ser respondida e o contexto em torno da candidata a resposta.

Novas técnicas de *Open-Domain Question Answering* propõem alternativas que incorporam as três tarefas citadas anteriormente em uma só solução [22]. No estudo, é abordado o *Haystack Framework*. O *Haystack Framework* consiste em uma estrutura *open-source* que possibilita o usuário criar modelos QA. A vantagem do *Haystack* está na disponibilização de inúmeras APIs que podem ser customizadas para a criação do modelo mais apropriado a diferentes casos de uso. *Haystack* é constituído por quatro componentes. O primeiro é denominado *DocumentStore*, o qual tem a função de armazenar documentos textuais e seus metadados. O segundo documento é o chamado *Retriever*, que consiste em um algoritmo simples e rápido com a função de identificar a passagem a ser candidata de conter a potencial resposta à pergunta a partir de uma grande coleção de documentos. Tais algoritmos do *Haystack* incluem o TF-IDF, BM25, *Elastic-Search*, entre outros. O terceiro componente do *Haystack* é o chamado *reader*, que tem como objetivo receber múltiplas passagens de texto e retornar as top-k respostas e suas pontuações de confiança. O *reader* o composto por diversos modelos pre-treinados capazes de fazer uma

pesquisa completa no documento selecionado para encontrar a melhor resposta. Por fim, o último componente do *Haystack* é conhecido como *finder*, que é uma API fornecida pela estrutura do *Haystack* que combina o *retriever* e o *reader* em uma sequência a fim de promover um interface de QA *user friendly*.

Com isto, sistemas *retriever-reader* se mostram como uma boa alternativa para o cumprimento do objetivo do trabalho. O *Haystack Framework* poderia ser usado como estrutura para o modelo de QA do *chatbot* sobre a Amazônia Azul, porém deve-se analisar como o mesmo performaria mo caso de uso deste estudo, já que a base de conhecimento textual a respeito da Amazônia Azul utilizada seria limitada. Assim, existe a possibilidade a abordagem de formulação de pares pergunta-resposta possuir uma melhor performance neste caso de uso.

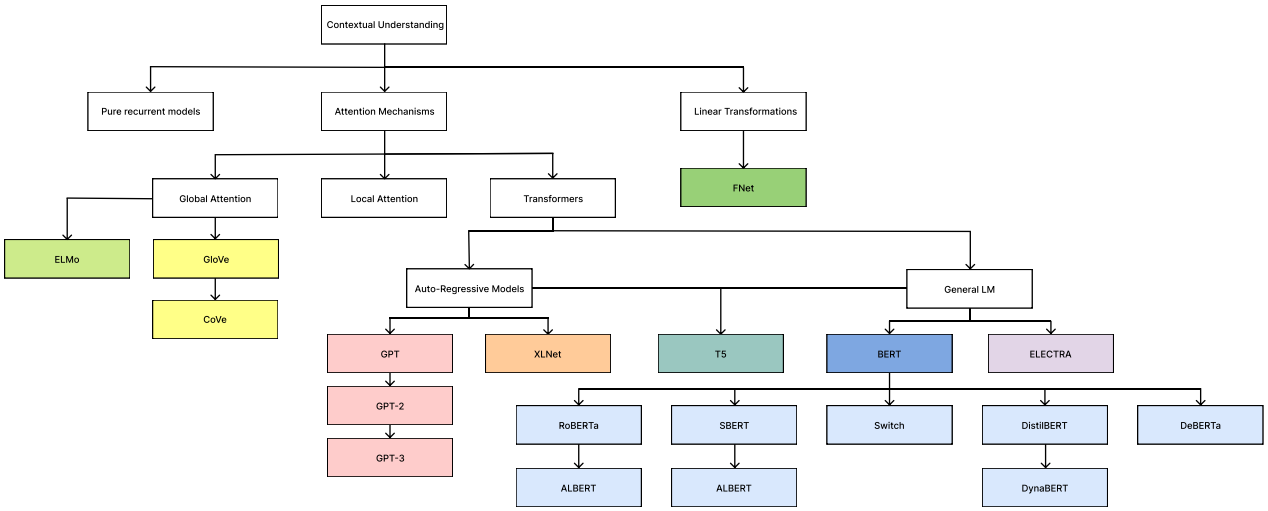
4.4 Mecanismos de Atenção

Usualmente, usa-se modelos *Encoder-Decoder* para o processamento de linguagem natural. *Encoder-Decoders* são redes neurais de multi-camadas às quais as primeiras camadas são respectivas à parcela do *encoder* recebe parcelas textuais como os dados de entradas e às codificam diminuindo a complexidade em sua saída. Já o *decoder* recebe a saída do *encoder* e o processa para sua saída.

Os modelos *Encoder-Decoder* têm se mostrado muito promissores. Porém, em [23], vemos que possuem pontos fracos ao lidar com maior número de parâmetros, como maiores *datasets* ou tempo de treinamento. Esse déficit se da pelo fato de no processo de codificação, uma parte da informação ser perdida, já que o *decoder* recebe apenas a saída da última camada do *encoder*. Assim, nesse processo, importantes informações de contexto são deixadas para trás, afetando a performance do modelo. Para lidar com esse problema, e para que o *decoder* possa receber mais informações, além da saída do *encoder*, foi desenvolvido um mecanismo que possibilite o *decoder* receber mais do que a saída, mecanismo denominado "Modelo de Atenção".

Mecanismos de Atenção são um complemento dos modelos *Encoder-Decoder*. Eles dão suporte para tais modelos no modo em que possibilitam a passagem de informações de mais de um nó, além da saída do *encoder*. Com isso, é visto uma clara melhora de performance dos modelos, se mostrando fortes ferramentas usadas em sistemas de resposta a pergunta

Figura 3: Taxonomia de diversos modelos de NLP



Fonte: Adaptada de [23].

A figura 3 mostra uma taxonomia dos diversos modelos de NLP baseados em compreensão de contextos, evidenciando principalmente os que se utilizam de Mecanismos de Atenção.

4.5 *Chatbots*

Significativa parcela das empresas e marcas líderes no mercado, assim como agências governamentais, utilizam conceitos de IA para criar agentes inteligentes com capacidade de interagir com clientes e automatizar processos sem a necessidade de intervenção humana, os chamados *chatbots*. Estima-se que, de todas as interações online, um terço envolve algum tipo de *chatbot* [21], estatística que tende a aumentar com o avanço de tecnologias que possibilitam a expansão do uso e acessibilidade desse tipo de ferramenta.

Apesar do crescente emprego e diversidade nas aplicações, os *chatbots* têm tido reações mistas dos consumidores, devido principalmente a fatores como respostas erradas ou irrelevantes, perguntas não entendidas, e capacidades que não correspondem às expectativas gerais dos usuários [10]. Assim, levanta-se o questionamento: Como fazer um *chatbot* útil e suficientemente aceito pelo público-alvo?

Diversos estudos foram realizados com a finalidade de avançar em direção a resposta dessa pergunta. [21] parte do princípio de que as pessoas são mais receptíveis a outras com o mesmo tipo de personalidade, estratégia historicamente empregada para influen-

ciar o comportamento de consumidores, mas cuja consistência não havia sido devidamente estudada na interação humano-computador (IHC). Para realizar esse estudo, foram examinadas mais de 57,000 interações com *chatbot*, obtendo-se como resultado que, além de ser possível manipular *chatbots* para assumir uma personalidade através do uso da linguagem, combinar a personalidade do consumidor com a do *chatbot* congruente ocasionou maior média de engajamento. Considerado a personalidade como um aspecto humanizador do agente robótico, essas conclusões parecem estar em acordo com [3], que constata a predisposição de usuários para fazerem mais esforços para reparar equívocos ou mal-entendidos quando em contato com *chatbots* percebidos como humanos do que os vistos como automáticos.

Em uma outra abordagem, focada no método de implementação dos *chatbots*, [10] propõe um *framework* baseado na expertise compartilhada por 15 experts nesse campo que já haviam se envolvido em processos de implementações de *chatbots*, assim como revisões da literatura. A partir disso, desenvolveram um *framework* que compreende 101 questões (disponíveis em inglês no link `https://bit.ly/Implementation_Framework`) para o levantamento de requisitos, desenvolvimento e implementação de um *chatbot* focado na experiência do usuário e compreendendo estruturação fundamentada em pessoas, atividade, contexto e tecnologia (PACT). Os resultados da pesquisa fornecem um guia compreensível de como pode acontecer uma implementação bem sucedida de *chatbots*, assim como auxiliar no levantamento de problemas relevantes durante o processo.

Chatbots podem servir diversos propósitos, desde pedir comidas e caronas até registrar reclamações e fornecer conselhos médicos, situações nas quais são aplicações preferíveis dada sua conveniência e imediatez [3]. Nota-se que a aplicação e implementação de *chatbots* tem setores nos quais parecem fazer mais sentido, com maior potencial de aceitação dos consumidores para o nível tecnológico atual.

No caso da aplicação de *chatbots* na educação, [17] aborda uma revisão sistemática de estudos prévios sobre o uso de *chatbots* na educação a partir de 53 artigos retirados de fontes digitais reconhecidas. Foram elaboradas 5 questões de pesquisa como forma de guia para o estudo, que seguiu um protocolo de orientações para revisões sistemáticas em engenharia de software. Os resultados apontaram que a tecnologia de *chatbots* apresenta não somente potencial de melhorar o processo de ensinar e aprender, mas também todos os outros aspectos da educação, com benefícios como integração de conteúdos, motivação, envolvimento, possibilidade de usuários múltiplos e simultâneos, assim como assistência imediata e aprendizado personalizado. Contudo, foram também relatadas limitações de implementação e uso de *chatbots* em educação relacionadas a fatores éticos, comporta-

mento do usuário, supervisão e problemas de manutenção.

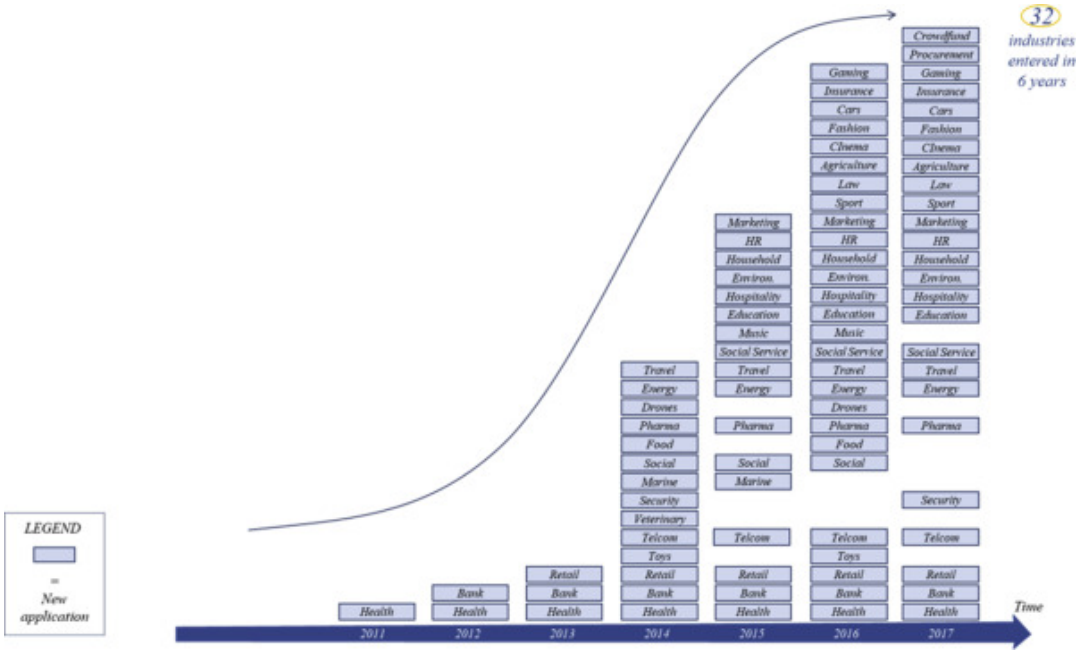
4.6 IBM Watson

O Watson foi revelado em 2005 e lançado em 2010 pela IBM, tendo diversas extensões significativas sido adicionadas na tecnologia desde então, mas sem perder sua funcionalidade como uma máquina de perguntas e respostas [15], que, de forma geral, deve compreender todas as características de uma pergunta para então retornar a resposta mais provável de uma lista de respostas para a mesma, para isso utilizando aplicações avançadas de NLP, aquisição de informações, representação de conhecimento, entre outros.

Fazendo uso de processadores massivamente paralelos (MPP, na sigla em inglês) e capaz de trilhões de operações em segundos, o Watson é "um sistema de gestão altamente sofisticado para conhecimentos preexistentes que pode, se propriamente configurado, prestar auxílio em decisões" [15]. A tecnologia ficou famosa em 2011, quando foi utilizada em um show de perguntas e respostas de temas variados chamado *"Jeopardy!"* para desafiar 2 campeões do programa, posteriormente derrotados pela máquina, que por sua vez não foi conectada à internet, mas tinha acesso à 200 milhões de páginas de conteúdo, demonstrando capacidade de rápida análise estatística e utilização de NLP.

A tecnologia apresentou potencial crescente de aplicações e qualidade de serviço, se tornando referência para aplicações variadas de QA e NLP:

Figura 4: Novas aplicações do Watson ao longo do tempo



Fonte: Retirado de [14]

A figura 4 ilustra a evolução dos campos de aplicações do Watson, que se iniciaram com Saúde e Bancos, mas rapidamente, a partir de 2013, passaram a integrar recursos como reconhecimento rápido de imagens, detecção de som e mineração de textos [14]. De forma geral, a tecnologia se mostrou capaz de processar e analisar grandes quantidades de informação, aplicar conceitos de IA importantes para sistemas QA, como NLP e análises estatísticas, além de fornecer uma interface gráfica e *framework* na web para implementações.

5 METODOLOGIA

Esta seção inicialmente aborda os requisitos definidos para o trabalho, e então, nos tópicos seguintes, detalhamentos do processo para a realização do mesmo.

5.1 Requisitos

Dado que a proposta do trabalho envolve o desenvolvimento de um *chatbot*, este deve atingir um *score* satisfatório em métricas de avaliação de sistemas QA. Além disso, em relação ao contato com usuários, espera-se que o *chatbot* seja capaz de responder perguntas variadas acerca do domínio definido (Amazônia Azul) em uma interface fluida e guiada pela experiência do usuário, que deve, após utilização do agente conversacional, aprimorar seus conhecimentos sobre o tema.

Dessa forma, os requisitos do *chatbot* desenvolvido neste trabalho podem ser resumidos a:

- Atingir *score* satisfatório em métricas de avaliação de sistemas QA (como *EM-Score*, abordado na seção 3.7);
- Ser capaz de responder perguntas variadas relacionadas à Amazônia Azul;
- Promover conversas fluidas e focadas na experiência de usuário;
- Possibilitar aprendizado dos usuários.

Para avaliação do cumprimento dos requisitos listados, serão analisados indicadores obtidos por respostas de usuários em um formulário com as seguintes peguntas:

"Em uma escala de 0 a 5, ...

- qual seria a probabilidade de você recomendar o *chatbot* para uma pessoa que deseja aprender mais sobre a Amazônia Azul?"

- com que frequência o *chatbot* compreendeu sua pergunta?"
- quão fluida foi sua conversa com o *chatbot*?"
- quão confiáveis você considera que foram as respostas?"
- quanto você sente que aprendeu com a experiência?"

Assim como uma pergunta aberta e opcional para futuras possibilidades: "O que você sentiu falta no chatbot?". Com esses indicadores, será avaliado o desempenho do *chatbot*, tendo como foco principal do trabalho a experiência e aprendizado dos usuários ao utilizá-lo.

Finalmente, para cálculo de métricas de avaliação de sistemas QA, serão analisadas as interações entre usuários e *chatbot*, e, para considerações demográficas, também será coletada a idade dos participantes.

5.2 Metodologia detalhada

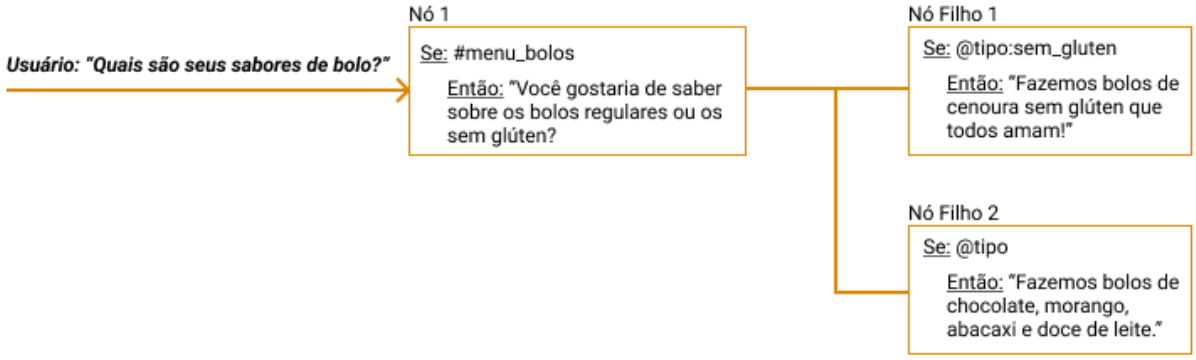
Nesta seção será apresentado e detalhado o processo de desenvolvimento do trabalho.

5.2.1 Inicialização de sistema para configuração do Watson

O *IBM Watsson Assistant*, uma divisão do Watson focada na construção de agentes conversacionais virtuais com utilização de IA (*chatbots*), fornece uma interface web (disponível em `https://www.ibm.com/br-pt/products/watson-assistant`) para utilização do *framework*, assim como diversas possibilidades de integrações. O uso dessa tecnologia pode ser explicado como um diálogo entre usuário e *chatbot*, no qual são ligadas intenções (*intents*, como definido na seção 3.4), contendo o que os usuários dizem querer, à respostas correspondentes do *chatbot*.

Para que esse processo ocorra da maneira esperada, o Watson é composto por nós de diálogos, estruturas que são ativadas quando se identifica determinada informação (que pode ser um *intent* específico, um tipo ou valor de entidade ou valor em uma variável de contexto) no *input* do usuário. Com a condição de ativação do nó satisfeita, este pode tomar diversas ações, como retornar uma resposta textual, uma imagem, fornecer opções, salvar um contexto, entre outros. Após sua ativação, o nó pode guiar a conversa para seus filhos, como estratégia para responder perguntas mais complexas.

Figura 5: Exemplo de estrutura de nó

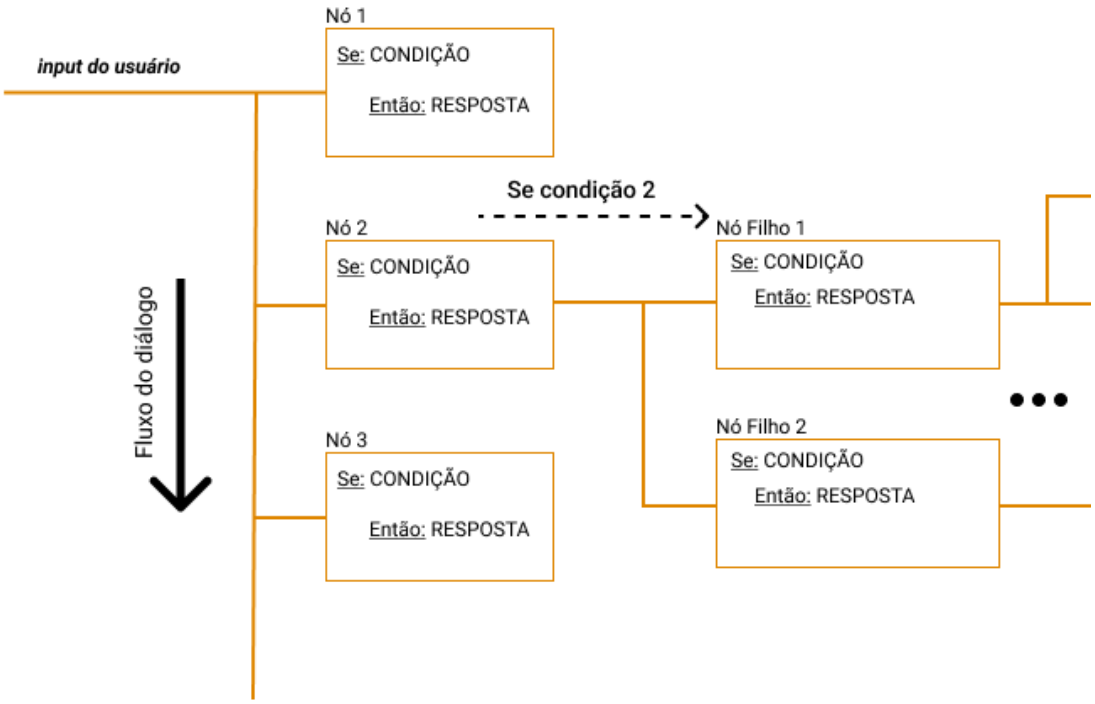


Fonte: Adaptado de [8]

No exemplo abordado pela figura 5, o usuário faz um pedido que requer mais informações para que seja respondido de forma correta. Para a ativação do "Nó 1", a condição era a detecção do *intent* "menu_bolos", enquanto a de ativação dos nós filhos dependia da entidade "tipo", sendo todas as respostas dos nós, nesse caso, no formato de respostas textuais, mas poderiam adicionar ou regular variáveis, salvar contextos ou mesmo pular para outros nós, para eventuais manipulações do fluxo conversacional.

O assistente virtual inicia o fluxo do diálogo a partir do primeiro nó até o último, em uma estrutura de árvore, de cima para baixo. Quando um nó é ativado, o Watson pode começar uma nova busca nos filhos do nó, aguardar um novo *input* do usuário para realizar a busca nos filhos, ou pular para outro nó, sendo o arranjo compreendido pelo nó inicial e todos os filhos chamados de "ramos" do fluxo em árvore, dado que o diálogo é direcionado de filhos em filhos até o fim do ramo, para então recomeçar o processo a partir do primeiro nó da árvore, permitindo que o diálogo seja tão complexo quanto o necessário enquanto mantém possibilidade de flexibilidade nas respostas, conforme ilustrado na figura 6.

Figura 6: Exemplo de fluxo de diálogo em árvore



Fonte: Os autores

A IBM oferece a possibilidade de construção dos elementos do *chatbot*, como seus diálogos, *intents*, nós, contexto e entidades de forma manual em sua interface web, contudo, para a extensão da aplicação desejada nesse trabalho, esse processo tomaria significativo tempo e estaria sujeito a erros, sendo uma alternativa a utilização da API do Watson (cuja documentação se encontra em `https://cloud.ibm.com/apidocs/assistant/assistant-v2`). Para a recepção de configurações dos referidos elementos do *chatbot* geradas de forma automática, o *Watson Assistant* permite o *upload* de dados estruturados em formato JSON (JavaScript Object Notation), que são interpretados pelo sistema e usados para montar o modelo do *chatbot* com o *framework* do Watson.

A partir de revisões na literatura, encontrou-se um programa que permite, entre outros, a geração de *intents* de forma mais ágil e estruturada que o processo puramente manual [5]. Seu código gera um arquivo JSON capaz de programar o Watson com uma estrutura básica de diálogo, dividida em 6 conjuntos de nós: "Diálogo", "Sem contexto", "Intenção", "Respostas" e "Não entendi", além de 186 *intents* obtidos majoritariamente a partir da filtragem do *dataset* Question Answering V2.1 do MS MACRO.

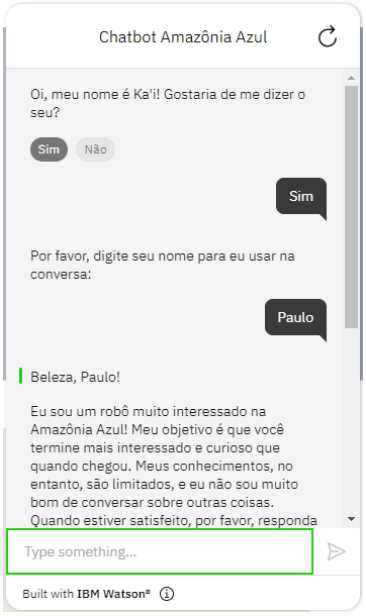
O programa em questão foi usado como base para a geração de configurações do Watson, e os desenvolvimentos realizados estão contidos em `https://github.com/VictBenito/chatbot-Victor-Paulo`.

5.2.2 Adequação do sistema

No processo de adequação do sistema, partiu-se do sistema *as-is* e foi trabalhada uma adaptação focando na melhor experiência do usuário. No sistema inicial, o *chatbot* expressou alta fricção durante as primeiras interações dos usuários, o que prejudica a fluidez da conversa durante a utilização.

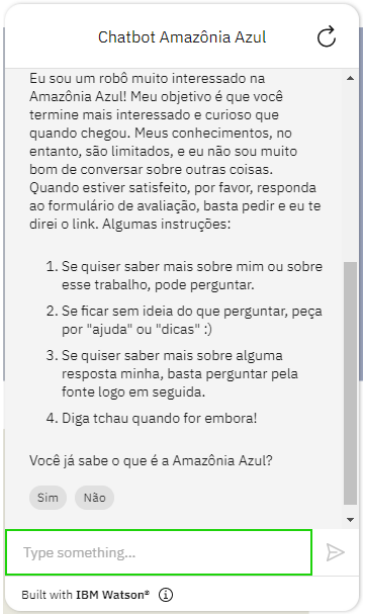
Na inicialização, o sistema primeiramente se apresentava, perguntava o nome do usuário e retornava uma resposta com um denso texto detalhando o objetivo do *chatbot* e aí sim sugerindo uma primeira pergunta, demonstrando considerável latência até a interação do usuário.

Figura 7: Sistema *as-is*



Fonte: Os Autores

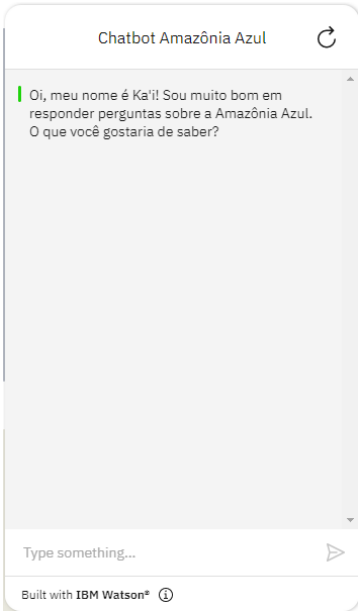
Figura 8: Sistema *as-is* - continuação



Fonte: Os Autores

Pensando nisso, foi alterado o primeiro nó de diálogo para que o sistema seja inicializado de maneira mais simples, podendo assim o usuário já ser capaz de fazer uma pergunta que lhe interesse diretamente, a fim de aprimorar a fluidez da conversa.

Figura 9: Sistema com inicialização adaptada

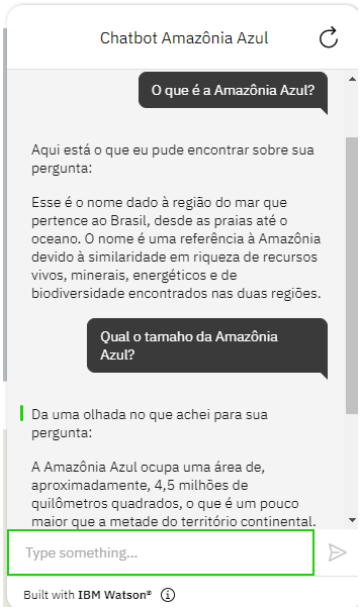


Fonte: Os Autores

Em seguida, foi atualizado o nó de contexto para que sempre que o *chatbot* retornasse qualquer resposta, uma frase inicial fosse exposta na tela para que a sensação de fluidez

e de conversa fosse aprimorada, com um tom mais amigável e familiar ao usuário.

Figura 10: Sistema com frases preliminares às respostas



Fonte: Os Autores

Por fim, notou-se que o sistema *as-is* somente retornava opções de dicas do que perguntar a partir da terceira pergunta a qual o sistema não pôde responder. Para diminuir a frustração do usuário por não obter respostas às suas perguntas, bem como possibilitar maior direcionamento, o sistema foi adaptado para que disponibilizasse a opção de dicas já na primeira vez em que o *chatbot* não fosse capaz de retornar uma resposta ao usuário, para que a fluidez da conversa fosse menos prejudicada.

5.2.3 Geração de *intents* e desenvolvimento do protótipo

O programa desenvolvido em [5], utilizado para adição de *intents* no *chatbot*, divide as perguntas em quatro partes:

- Contexto ou rótulo: tema sobre o qual a pergunta faz parte (ex: ”Amazônia azul”, ”oceano”, ”litoral”);
- Modificador: responsável pela definição do tipo de questão e atribuição de elocuções parciais (ex: ”causa”, ”efeito”, ”detalhar”);
- Substantivo e Recipiente: utilizados para fornecer mais informações de modo a possibilitar definições de perguntas com complexidades maiores (ex: ”ambiente”, ”economia”, ”produção”).

A partir da união dessas partes, forma-se um *intent*, de modo que cada pergunta corresponde a uma junção única das mesmas, enquanto um *intent* pode ser identificado pelo *Watson* por mais de uma pergunta (as referidas elocuções).

Para a adição de novos conjuntos perguntas-respostas no *chatbot* com esse método, foi utilizada uma planilha, na qual cada linha forma um *intent*, e nas colunas estão a pergunta base, sua resposta, fonte, contexto, modificador, substantivo, recipiente e elocuções, permitindo certa agilidade e replicabilidade no processo de inserção, que consiste no preenchimento manual de cada coluna para cada intent na planilha, seguido pela execução do código citado na seção 5.2.1 para conversão dos novos *intents* em formato aceito pelo *Watson Asssistant*.

Como definido na seção 3.4, os *intents* e suas respostas são a base de conhecimentos utilizada no modelo do Watson para identificação e resposta das perguntas feitas pelos usuários. Dessa forma, a quantidade de informação efetivamente coberta pelo *chatbot*, nesse caso, é um reflexo do número de *intents* e respostas associadas presentes na estrutura do Watson.

Para a geração de *intents*, considerou-se que, no caso de uso em questão, um usuário arbitrário entraria em contato com o *chatbot* em duas situações principais: quando em busca de alguma informação ou assunto específico, ou movido pela curiosidade e vontade de aprendizado de assuntos diversos relacionados ao tema. Ambos os casos são beneficiados por uma lógica de expansão da quantidade de *intents* e respostas. Para isso, o procedimento adotado como forma de aumento da cobertura de respostas foi o de, inicialmente, adicionar por volta de 100 *intents* a partir de uma base QA chamada "Questions&Answers_BLAB", que continha 120 pares pergunta-resposta sobre temas como atividade petrolífera, desastres ambientais no ambiente costeiro e marinho, erosão e sedimentação costeiras, esportes marítimos, entre outros, disponibilizada em `https://github.com/VictBenito/chatbot-Victor-Paulo/blob/main/Questions%26Answers_BLAB%20(1).xlsx`.

Então, com uma base de *intents* e respostas mais consolidada, o processo para expansão consistiu em, a partir de respostas, gerar novas perguntas e respostas. Por exemplo, uma das perguntas é "O que é pesca extrativa?", cuja resposta é "A pesca extrativa é aquela que normalmente se associa à palavra "pesca", ou seja, consiste na retirada de recursos pesqueiros do ambiente natural, seja no mar ou no continente (em lagos, rios e outros corpos hídricos). Em oposição a esse conceito existe a aquicultura, que também integra a pesca no geral.", essa resposta pode levar o usuário a questionar frações da

mesma, gerando novas perguntas como "O que são recursos pesqueiros?" ou mesmo "O que é aquicultura?", e, a partir dessas, foram analisadas possibilidades de extensões relacionadas à localidades ("Onde mais se pratica aquicultura?"), maiores figuras relacionadas ("Quais os maiores exploradores de recursos pesqueiros?"), números ("Quanto as aquiculturas produzem em média?"), entre outros, com esse processo feito para diversas perguntas.

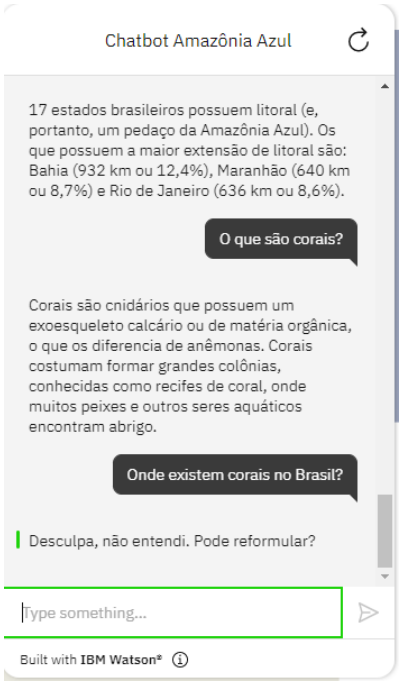
Assim, seguindo esse procedimento de forma exaustiva, chegou-se a um total de 386 *intents* e respostas.

5.2.4 Testes preliminares e validações com usuários

Após uma primeira adequação da estrutura do *chatbot*, voltada à fluidez da experiência e a geração de *intents*, populando a base do sistema, a fim de endereçar as potenciais dúvidas e perguntas dos usuários, foi exposto o sistema a um grupo restrito de usuários a fim de coletar primeiras impressões e *feedbacks* de melhoria.

O processo de uso do *chatbot* foi observado pelos desenvolvedores para compreender o que satisfazia as necessidades dos usuários e o que os causava qualquer tipo de frustração. Notou-se que, para usuários interessados no aprendizado, algumas perguntas diversas não respondidas não geravam demasiada frustração, dado que os usuários possuíam ciência de que o sistema não seria capaz de responder todas as suas dúvidas. Maior frustração foi observada quando o usuário, após receber uma resposta, perguntava informações adicionais sobre a resposta recebida e o *chatbot* não as sabia. Um exemplo claro deste comportamento é explicitado na figura 11.

Figura 11: Exemplo de falha de testes preliminares



Fonte: Os Autores

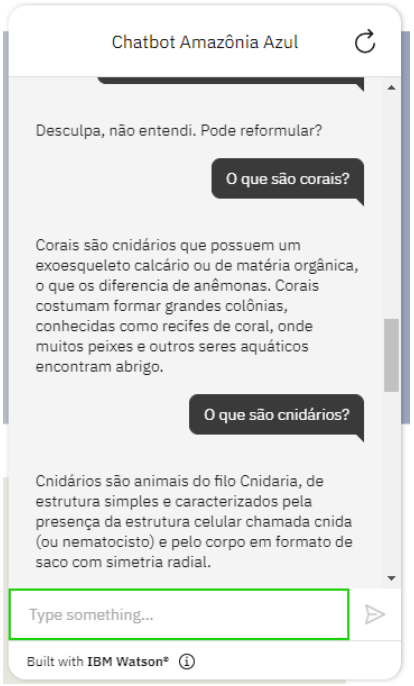
Além disso, foram observadas utilizações de linguagem informal não suficientemente cobertas pelos modificadores e elocuções já estabelecidos, por exemplo, a pergunta base "Quem deu o nome à Amazônia Azul?" foi feita de diversas formas: "De onde vem amazônia azul?", "Por que o nome amazônia azul?" e "Por que Amazônia azul?", impossibilitanto o reconhecimento por parte do *chatbot*.

Estes *insights* foram utilizados como insumos para as adaptações finais do sistema prévias à disponibilização para os usuários finais.

5.2.5 Adaptações finais

Com os insumos coletados nos testes preliminares, notou-se a importância de o *chatbot* possuir maior robustez ao responder perguntas relacionadas às respostas já existentes em sua base de dados. Assim, foi adotado um método para a geração de mais *intents* destes assuntos. Com isso, foram analisadas as respostas existentes na base e a partir de informações contidas nelas gerou-se *intents* relacionados a estas informações, como explicado na seção 5.2.3, a fim de endereçar casos como o da figura 12.

Figura 12: Exemplo de sucesso em perguntas de temas relacionados



Fonte: Os Autores

Além disso, para que o sistema seja ainda mais robusto, foi feito também um trabalho de adição de elocuções dos *intents*. As elocuções são as diferentes estruturas de perguntas as quais possuiriam a mesma resposta.

5.2.6 Testes finais com usuários e coleta de resultados

Após a finalização do desenvolvimento do *chatbot*, foi disponibilizado o acesso ao sistema para os usuários. Criou-se um formulário de avaliação com cinco perguntas qualitativas bem como o link de acesso ao sistema para que usuários pudessem interagir e terem suas impressões à respeito de sua funcionalidade. Após o uso, as pessoas respondiam o fomulário qualitativo para avaliar suas respectivas experiências.

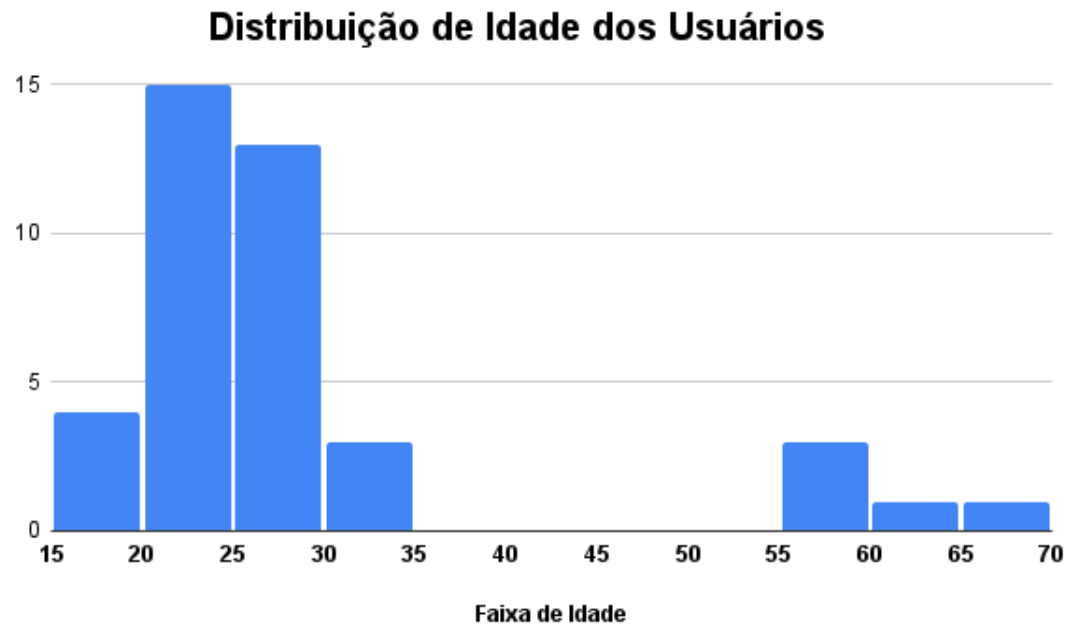
Além disso, é possível ter acesso aos *requests* feitos pelos usuários bem como às respostas fornecidas pelo *chatbot*. Assim, com a coleta das avaliações qualitativas foi possível calcular os resultados de performance explicitados em 6 e também discutir e interpretá-los na seção 7.

6 RESULTADOS

Nesta seção, serão expostos os resultados coletados através de formulário a partir da utilização do projeto por usuários diversos, a fim de analisar a impressão dos usuários finais sobre o funcionamento do *chatbot*.

O *chatbot* foi utilizado e avaliado por 40 pessoas, com uma média de idade de 29 anos, mediana de 25, moda de 24, e distribuídas entre 18 e 65 anos.

Figura 13: Distribuição de idade dos usuários



Fonte: Os autores

Após sua interação, todos os 40 usuários responderam um formulário avaliando qualitativamente suas respectivas experiências durante a interação com o *chatbot*. Além disso, foram armazenadas 76 conversas e 365 *requests*. Dentre estes *requests*, foram consideradas válidas 267 perguntas feitas pelos usuários, desconsiderando perguntas completamente fora do tema Amazônia Azul bem como *requests* sem sentido como, por exemplo, “tama-

nho“ .

A partir disso, serão avaliadas as métricas quantitativas a partir da performance do *chatbot* ao fornecer um retorno às 267 perguntas, bem como as avaliações qualitativas a partir das respostas ao formulário.

6.1 Resultados Quantitativos

Avaliando as 267 perguntas válidas feitas ao *chatbot*, 148 foram respondidas corretamente, 96 não foram respondidas e 23 foram respondidas de maneira incorreta.

Tabela 4: Casos válidos de resposta do *chatbot* para os usuários

Categoria	**Número de saídas do chatbot**
Perguntas respondidas corretamente	148
Perguntas não respondidas	96
Perguntas respondidas incorretamente	23
Total	**267**

Fonte: Os autores

A seguir será detalhado o que estes resultados significam em métricas de avaliação para sistemas de resposta a perguntas.

6.1.1 *Exact-Match Score*

Como definido na seção 3, o *Exact-Match Score* retorna 1 quando a resposta é igual à resposta correta e 0 quando não retorna a resposta correta.

No caso de uso desse trabalho, foi avaliado como EM = 1 quando o *chatbot* retornou uma resposta correta à pergunta e EM = 0 quando o *chatbot* não responde ou responde incorretamente a pergunta.

Assim, foi calculado um *Exact-Match Score* de:

$$EMScore = 0.554 \tag{6.1}$$

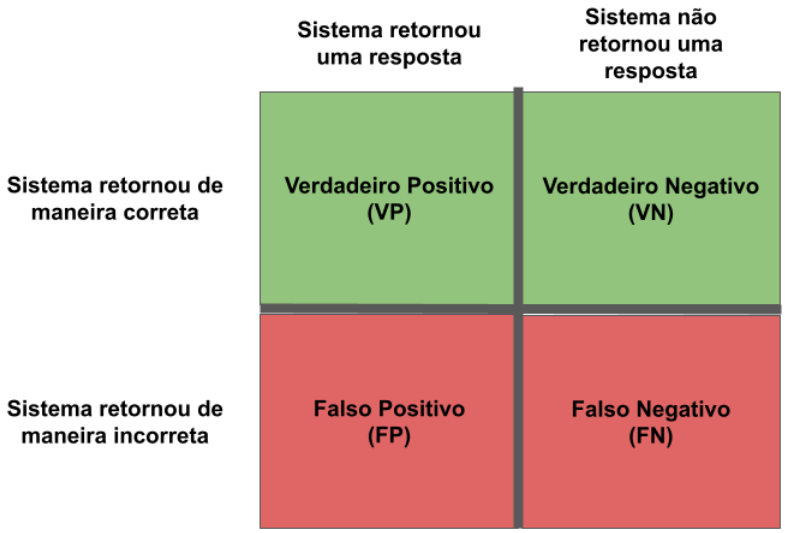
Esse resultado significa que, dentre todas as perguntas feitas ao *chatbot*, aproximadamente 55,4% foram respondidas corretamente.

6.1.2 Precisão, Recall e F1-Score

Foi calculado também a precisão, *recall* e *F1-Score*. Como definido na equação 3.1, o *F1-Score* é baseado puramente na precisão e no *recall* do sistema, porém, é preciso definir esses conceitos para o caso de uso deste trabalho.

Tanto a precisão quanto o *recall* são baseados em verdadeiros e falsos positivos e verdadeiros e falsos negativos. Assim, será considerado um verdadeiro positivo (VP) quando o *chatbot* foi capaz de responder à pergunta e responder de maneira correta. Já um falso positivo (FP) foi considerado quando o sistema retorna uma resposta ao usuário, porém uma resposta incorreta. Um verdadeiro negativo (VN) é interpretado como o *chatbot* retornando que não soube responder a pergunta quando de fato não existe nenhum *intent* capaz de endereçar a questão. Por fim, um falso negativo (FN) significa que o sistema retornou que não sabia responder a pergunta quando na verdade existia um *intent* capaz de endereçá-la.

Figura 14: Matriz de interpretação de resultados



Fonte: Os autores

Os números de ocorrências de cada segmento foram obtidos a partir de análises dos registros de conversas entre o *chatbot* e os usuários, e são exibidos na tabela 5.

Tabela 5: Ocorrência dos resultados

Resultado	Número de ocorrências
Verdadeiro-positivo (VP)	148
Verdadeiro-negativo (VN)	72
Falso-positivo (FP)	23
Falso-negativo (FN)	24

Fonte: Os autores

A precisão é definida por:

$$Precisão = \frac{VP}{VP + FP} \tag{6.2}$$

Já o *recall*, é calculado como:

$$Recall = \frac{VP}{VP + FN} \tag{6.3}$$

A partir das equações 6.2, 6.3 e 3.1, obteve-se os seguintes resultados para o *chatbot* desenvolvido:

$$Precisão = 0.8655 \tag{6.4}$$

$$Recall = 0.8605 \tag{6.5}$$

$$F_1 = 0.8630 \tag{6.6}$$

6.2 Resultados Qualitativos

Como avaliação qualitativa, o questionário possuía 5 perguntas de avaliação de experiência, as quais eram possíveis de serem avaliadas com notas de 0 a 5. Os resultados das 40 respostas obtidas foram:

Tabela 6: Resultados qualitativos

Pergunta de avaliação	Média	Desvio Padrão
Qual seria a probabilidade de você recomendar o chatbot para uma pessoa que deseja aprender mais sobre a Amazônia Azul?	4.15	1.12
Com que frequência o chatbot compreendeu sua pergunta?	3.48	1.22
Quão fluida foi sua conversa com o chatbot?	3.65	1.41
Quão confiáveis você considera que foram as respostas?	4.45	1.06
Quanto você sente que aprendeu com a experiência?	4.05	1.18

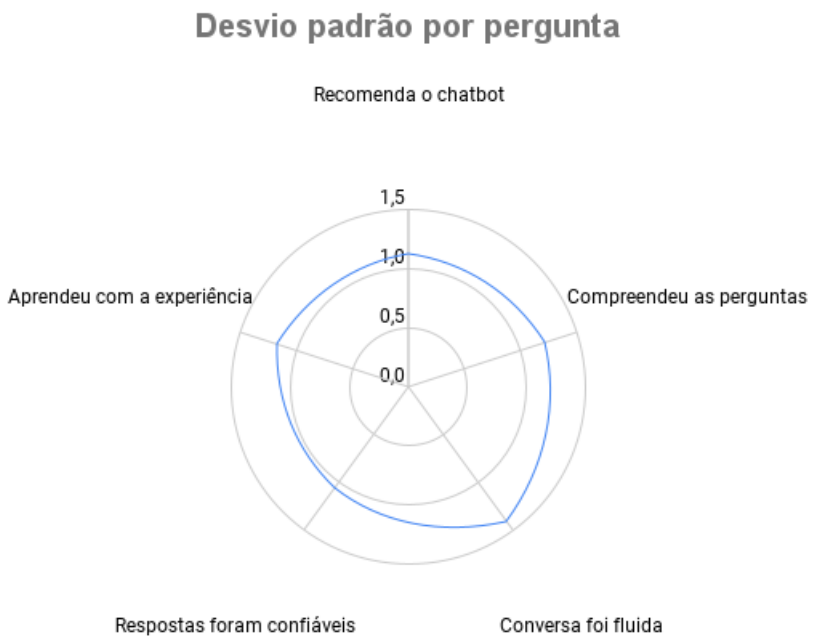
Fonte: Os autores

Assim, temos também explicitados os resultados pelos seguintes gráficos:

Figura 15: Média dos resultados qualitativos

Fonte: Os autores

Figura 16: Desvio padrão dos resultados qualitativos



Fonte: Os autores

Por fim, além das perguntas de avaliação o formulário possuía uma pergunta aberta de "O que você sentiu falta no chatbot?". 21 pessoas preencheram este campo.

Nove pessoas fizeram comentários à respeito da falta de informação com relação à pergunta que foi feita, como por exemplo, "Faltaram respostas sobre o tema de legislação (se existem leis que protegem a biodiversidade da amazônia azul, por exemplo)" ou "faltaram resposta para todas as perguntas".

Sete pessoas comentaram sobre a estruturação do *chatbot* e sobre como ela podia ser melhor para suas respectivas experiências, como por exemplo "Uma introdução para que leve o usuário a saber o que perguntar, como um texto que te leve a uma trilha do conhecimento. " ou "mais dicas relacionadas as dúvidas que nao foram encontradas".

Por fim, duas pessoas comentaram sobre o design do *chatbot* como ausência de uma foto ou imagem representando o sistema, outras duas pessoas comentaram sobre a falta de compreensão de suas perguntas e um usuário comentou que não sentiu falta de nada durante sua interação.

7 DISCUSSÕES

Nesta seção, os resultados descritos na seção 6 serão analisados e detalhados a fim de avaliar a performance do *chatbot*, como ela satisfaz ou não os requisitos listados na seção 5, bem como pontos de atenção a serem desenvolvidos e aprimorados para um melhor funcionamento do sistema.

7.1 Avaliação e discussão dos Resultados Quantitativos

A partir dos valores exibidos na seção 6.1, obteve-se os seguintes resultados em métricas de avaliação de sistemas de *question-answering* e *machine learning*:

Tabela 7: Resumo dos resultados quantitativos

Métrica	**Resultado**
Exact-Match Score	0.554
Precisão	0.8655
Recall	0.8605
F1-Score	0.8630

Fonte: Os autores

Analisando primeiramente o *EM-score*, vemos que, das perguntas válidas, o *chatbot* foi capaz de responder corretamente cerca de 55.4% das perguntas. O *chatbot* desenvolvido em [5], também sobre a Amazônia Azul, obteve um resultado de 44.5% na mesma métrica. Com isso, vemos que o *chatbot* desenvolvido neste trabalho foi capaz de atingir um resultado satisfatório nessa métrica, atingido o requisito definido em 5.1.

Este resultado é reflexo do sistema não ter sido capaz de retornar uma resposta para 96 das 267 perguntas válidas (35.9%). Além disso, 23 das 267 perguntas (8.6%) foram respondidas erroneamente.

Os 35.9% de perguntas não respondidas se deram por duas maneiras. Primeiramente, 72 das 96 perguntas não respondidas não possuíam respostas para elas no sistema do *chatbot*. Isto evidencia que um trabalho de adicionar ainda mais *intents* na base poderia melhorar a performance do sistema na métrica de *EM-score*. Porém, vale ressaltar que é importante seguir um método nesse processo de população de *intents*. Existe uma vasta diversidade de possíveis perguntas a respeito da Amazônia Azul, sendo assim, é importante seguir algum método o qual seja possível direcionar o processo de adição de *intents* para perguntar que usuários potencialmente podem fazer. Um exemplo disso, seria iterar liberando o sistema para usuários, coletando perguntas reais e endereçando tudo que for perguntado e não tiver resposta no sistema.

Além das perguntas as quais o sistema não é capaz de endereçar, 24 das 96 perguntas não respondidas existiam na base de *intents*, porém o *chatbot* não foi capaz de identificar que o mesmo era capaz de responder. Esta interpretação se conecta ao *recall* de 86.05%, que justamente explicita a ocorrência de falso negativos. Apesar de acima de 86%, o *recall* teria potencial de ser ainda maior, dado que o sistema do *Watson Assistant* é considerado um dos melhores sistemas disponíveis no mercado.

Como estratégia para otimização do *recall* é possível utilizar-se do recurso de elocuções disponibilizado no *Watson*. As elocuções são basicamente adicionar variações de estrutura para uma mesma pergunta, como por exemplo, perguntados pelos usuários, "Quantos quilômetros quadrados possui Amazônia Azul? " ou "Qual a área da Amazônia Azul¿', que seriam elocuções para a pergunta "Qual o tamanho da Amazônia Azul? ". Assim, adicionando diferentes elocuções para *intents* já existentes, potencialmente diminuiriam o número de falsos negativos.

Analisando a precisão do *chatbot* foi visto um resultado de 86.55%. Assim, nos casos em que o sistema retornou uma resposta para o usuário, 86.55% deles, a resposta retornada foi correta. Assim, apesar de o sistema não retornar uma resposta em 35.9% dos casos, quando uma resposta é retornada, em sua grande maioria ela é retornada corretamente, dada a alta precisão. Ainda assim, em 13.45% dos casos, a resposta retornada foi incorreta ou não relacionada com a pergunta feita, o que é prejudicial para a experiência do usuário.

Os casos de falso positivo, os quais afetam a precisão, ocorrem quando o *Watson Assistant* não possui o *intent* daquela pergunta ou quando o *intent* existe porém o sistema entende que outra resposta é a que endereça a pergunta. Na definição dos diálogos, é definido uma confiança mínima chamada de "*intent.confidence*" para que o *Watson* ative determinado *intent*. No caso do sistema deste trabalho, foi definido um *intent.confidence*

maior do que 0.8 (80%) para todos os *intents*, assim, uma forma de melhorar a precisão seria definir um valor mais alto para tal parâmetro. Vale ressaltar que, o aumento da precisão alterando o *intent.confidence* irá potencialmente impactar negativamente na métrica de *recall*, dado que quanto maior o *intent.confidence*, maior a quantidade de falsos negativos.

Por fim, o F1-score é responsável por medir o equilíbrio do *tradeoff* entre precisão e *recall*, citado anteriormente. Para o sistema, vemos que o F1-score foi de 0.8630, que pode ser considerado satisfatório. Ainda assim, esta métrica pode ser melhorada ao aprimorar o *recall* pelo uso das elocuções.

7.2 Avaliação e discussão dos Resultados Qualitativos

Na tabela 6 vemos um resumo das avaliações qualitativas feitas pelos usuários. Sobre a primeira pergunta, na qual questiona-se a recomendação ou não do *chatbot*, foi obtida uma média de 4.15 pontos. O resultado para indicador foi positivo, visto que, em sua maioria, os usuários se demonstraram promotores do sistema, se declarando propensos a indicar o *chatbot* para pessoas interessadas.

Na pergunta relacionada a frequência em que o *chatbot* compreendeu as perguntas feitas, a nota média foi de 3.48 pontos, sendo um valor satisfatório porém com resultado qualitativo de pior média entre os indicadores. Isto também foi evidenciado pelo *EM-score* discutido na subseção anterior. Lá também foram levantadas ações para elevarem a performance de tal métrica, as quais certamente elevariam a nota média desta pergunta.

A pergunta sobre o grau de fluidez da conversa obteve uma nota média de 3.65 pontos, também satisfatório, já que ainda permanece acima da média do intervalo proposto. Nota-se também que este foi o indicador com maior desvio padrão (1.41 pontos). Assim, o que provavelmente diminuiu a média da métrica foi uma parcela de usuários que teve alguma sequência de perguntas não respondidas e deu uma nota extremamente baixa. Com isso, as ações de melhoria do *EM-score* listadas na subseção anterior, provavelmente seriam capazes de melhorar este indicador.

Com relação à confiabilidade dos usuários no *chatbot*, foi obtida a maior média dentre as métricas qualitativas com 4.45 pontos. Alguns fatores podem ter influenciado positivamente na métrica, como ao fornecimento da fonte caso solicitada, as respostas fornecidas possuírem boa riqueza de detalhes ou o fato de o *chatbot* fazer parte de um projeto

acadêmico.

Por fim, a última pergunta, relacionada ao aprendizado dos usuários, trouxe uma nota média de 4.05 pontos, sendo assim, a grande maioria dos usuários finalizou a interação com algum sentimento de aprendizado. Um dos motivos disso seria fato de o tema não ser muito difundido popularmente. pode ser também devido ao fato de os usuários perguntarem coisas que genuinamente não sabem ou gostariam de saber. Assim, quando recebem uma resposta as pessoas podem sentir que algum conhecimento foi obtido durante a experiência.

As perguntas do formulário relacionadas a probabilidade de indicação, frequência em que o *chatbot* respondeu às perguntas, confiabilidade das respostas e de sentimento de aprendizado dos usuários também foram avaliadas nos resultados qualitativos de [5]. Para todas essas quatro perguntas foram obtidas notas médias maiores do que às de [5].

Vemos assim que o método utilizado neste trabalho, focando na experiência do usuário trouxe resultados satisfatórios para a fluidez da experiência bem como no aprendizado dos usuários, cumprindo assim os requisitos qualitativos definidos em 5.

8 CONCLUSÕES

Agentes conversacionais inteligentes que utilizam conceitos de IA como NLP para interagir com usuários sem a necessidade de intervenção humana, também conhecidos como *chatbots*, vêm ganhado relevância no cenário tecnológico mundial. É comprovada na literatura sua capacidade de servir propósitos diversos, e, mais especificamente, no âmbito educacional, de melhorar o processo de ensinar e aprender de forma personalizada e eficiente. Contudo, o desenvolvimento e a implementação dessa tecnologia de modo a resultar em aceitação dos usuários pode ser desafiador.

Em paralelo, a Amazônia azul, área costeira que compreende 4,5 milhões de km^2, apresenta notável importância econômica, científica e ambiental, mas ainda demanda conscientização da população à respeito. Nela, realiza-se desde pesca, turismo e transporte marítimo até exploração de energias renováveis e extrações de petróleo e gás natural, demonstrando a grande riqueza de recursos vivos, minerais e energéticos.

Pensando nisso, no presente trabalho foi desenvolvido um *chatbot* com capacidade de responder a perguntas diversas a respeito da Amazônia Azul de forma suficientemente fluida, confiável, recomendável e proporcionando aprendizados valiosos acerca do tema aos usuários a partir da utilização de ferramentas de aplicação de IA como o IBM Watson.

Por fim, o trabalho representa também uma contribuição para o avanço de conhecimentos relacionados à sistemas QA, reunindo temas e estudos recentes, em especial os conteúdos em português, que ainda têm considerável espaço para desenvolvimentos na literatura. Além disso, este trabalho pode ter seus indicadores e valores alcançados vistos como uma referência para trabalhos futuros relacionados, assim como o procedimento desenvolvido para geração extensiva de pares pergunta-resposta pode servir de inspiração para outros métodos ou mesmo utilizado em conjunto a aplicações de *machine learning*, para automações em sua realização.

8.1 Considerações para trabalhos futuros

Para trabalhos futuros, existem diversos pontos em que o *chatbot* aqui desenvolvido poderia ser aprimorado. A maior oportunidade de melhoria a ser abordada está relacionada a quantidade de perguntas não respondidas pelo sistema. Na seção 7 foram levantadas ações que poderiam resultar em melhorias nesse aspecto, como por exemplo liberar o sistema para usuários e endereçar perguntas feitas pelos mesmos que não foram respondidas de maneira iterativa, em seguida aplicando o procedimento abordado na seção 5.2.3 para geração extensiva de pares pergunta-resposta a partir das respostas iniciais.

Outra possibilidade de melhoria relacionada a adicionar *intents* seria a adição de pares pergunta-resposta contidos em bases QA como as citadas no decorrer deste trabalho, muitas das quais já estão em português e são sobre temas relacionados, mas não foram utilizadas nesse sentido.

Além disso, vale tomar ações para aumentar o *recall* do sistema, utilizando de elocuções para diminuir a incidência de falsos negativos. Como citado na pergunta aberta do formulário, também seria interessante para a experiência do usuário dicas mais direcionadas a perguntas anteriores ou dicas mais genéricas como "curiosidades" ou "informações interessantes" sobre a Amazônia Azul.

REFERÊNCIAS

[1] Kingsley Nketia Acheampong e Wenhong Tian. "Advancement of Textual Answer Triggering: Cognitive Boosting". Em: *IEEE Transactions on Emerging Topics in Computing* 10.1 (2022), pp. 361–372. DOI: 10.1109/TETC.2020.3022731 (ver p. 23).

[2] Jafar Alzubi, Anand Nayyar e Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview". Em: *Journal of Physics: Conference Series* 1142 (nov. de 2018), p. 012012. DOI: 10.1088/1742-6596/1142/1/012012. URL: https://doi.org/10.1088/1742-6596/1142/1/012012 (ver p. 13).

[3] Petter Bae Brandtzaeg e Asbjørn Følstad. "Why people use chatbots". Em: *International conference on internet science*. Springer. 2017, pp. 377–392 (ver pp. 14, 27).

[4] F. N. Cação et al. *DEEPAGÉ: Answering Questions in Portuguese About the Brazilian Environment*. English. Vol. 13074 LNAI. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2021, pp. 419–433. URL: www.scopus.com (ver pp. 20, 21).

[5] Gabriel Okamoto Carlos. "Desenvolvimento de um chatbot sobre a Amazônia Azul". Em: *Escola Politécnica da Universidade de São Paulo* (2021) (ver pp. 21, 33, 36, 47, 50).

[6] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Em: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, jun. de 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (ver p. 22).

[7] Eduardo Tannuri Fabio Cozman. *Aprendizado de Máquina Enriquecido por Conhecimento para Raciocínio sobre Dados Oceânicos*. https://c4ai.inova.usp.br/pt/pesquisas/#KEML_port (ver p. 13).

[8] IBM. *How your dialog is processed*. 2021. URL: https://cloud.ibm.com/docs/assistant?topic=assistant-dialog-build (acesso em 16/10/2022) (ver p. 32).

[9] *IBM Watson Assistant documentation*. https://cloud.ibm.com/docs/assistant?topic=assistant-intents#:~:text=Intents%20ar%20purposes%20or%20goals,flow%20for%20reponding%20to%0it.. Accessed: 2022-12-14 (ver p. 15).

[10] Antje Janssen et al. "How to Make chatbots productive – A user-oriented implementation framework". Em: *International Journal of Human-Computer Studies* 168 (2022), p. 102921. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2022.102921. URL: https://www.sciencedirect.com/science/article/pii/S1071581922001410 (ver pp. 26, 27).

[11] Sweta P. Lende e M. M. Raghuwanshi. "Question answering system on education acts using NLP techniques". Em: *2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*. 2016, pp. 1–6. DOI: 10.1109/STARTUP.2016.7583963 (ver pp. 14, 15).

[12] Patrick Lewis et al. "PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them". Em: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 1098–1115. DOI: 10.1162/tacl_a_00415. URL: https://aclanthology.org/2021.tacl-1.65 (ver pp. 20–22).

[13] Elizabeth D Liddy. "Natural language processing". Em: (2001) (ver p. 14).

[14] Stefano Magistretti, Claudio Dell'Era e Antonio Messeni Petruzzelli. "How intelligent is Watson? Enabling digital transformation through artificial intelligence". Em: *Business Horizons* 62.6 (2019). Digital Transformation & Disruption, pp. 819–829. ISSN: 0007-6813. DOI: https://doi.org/10.1016/j.bushor.2019.08.004. URL: https://www.sciencedirect.com/science/article/pii/S0007681319301168 (ver p. 29).

[15] Robert Nisbet, Gary Miner e Ken Yale. "Chapter 22 - IBM Watson". Em: *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)*. Ed. por Robert Nisbet, Gary Miner e Ken Yale. Second Edition. Boston: Academic Press, 2018, pp. 773–781. ISBN: 978-0-12-416632-5. DOI: https://doi.org/10.1016/B978-0-12-416632-5.00022-0. URL: https://www.sciencedirect.com/science/article/pii/B9780124166325000220 (ver p. 28).

[16] *O que é amazônia azul e quais seus recursos?* https://www.ecycle.com.br/amazonia-azul/. Accessed: 2022-04-19 (ver p. 12).

[17] Chinedu Wilfred Okonkwo e Abejide Ade-Ibijola. "Chatbots applications in education: A systematic review". Em: *Computers and Education: Artificial Intelligence* 2 (2021), p. 100033. ISSN: 2666-920X. DOI: https://doi.org/10.1016/j.caeai.

2021.100033. URL: https://www.sciencedirect.com/science/article/pii/S2666920X21000278 (ver pp. 14, 27).

[18] A. F. A. Paschoal et al. "Pirá: A Bilingual Portuguese-English Dataset for Question-Answering about the Ocean". English. Em: *International Conference on Information and Knowledge Management, Proceedings*. Cited By :3. 2021, pp. 4544–4553. URL: www.scopus.com (ver pp. 11, 23).

[19] *Portal Amazônia responde: O que é Amazônia Azul?* https://portalamazonia.com/amazonia/portal-amazonia-responde-o-que-e-amazonia-azul. Accessed: 2022-04-19 (ver p. 12).

[20] P. Rajpurkar, R. Jia e P. Liang. "Know what you don't know: Unanswerable questions for SQuAD". English. Em: *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. Vol. 2. Cited By :461. 2018, pp. 784–789. URL: www.scopus.com (ver p. 23).

[21] Michael Shumanov e Lester Johnson. "Making conversations with chatbots more personalized". Em: *Computers in Human Behavior* 117 (2021), p. 106627. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2020.106627. URL: https://www.sciencedirect.com/science/article/pii/S0747563220303745 (ver p. 26).

[22] Zeeshan Haque Syed et al. "Question Answering Chatbot for Troubleshooting Queries based on Transfer Learning". Em: *Procedia Computer Science* 192 (2021). Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021, pp. 941–950. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.08.097. URL: https://www.sciencedirect.com/science/article/pii/S1877050921015854 (ver pp. 21, 24).

[23] Jatin Karthik Tripathy et al. "Comprehensive analysis of embeddings and pretraining in NLP". Em: *Computer Science Review* 42 (2021), p. 100433. ISSN: 1574-0137. DOI: https://doi.org/10.1016/j.cosrev.2021.100433. URL: https://www.sciencedirect.com/science/article/pii/S1574013721000733 (ver pp. 23, 25, 26).