

Tema:

Aprimoramento de Agente Conversacional Especializado em Amazônia Azul

Introdução e Objetivo

Agentes conversacionais, ou *Chatbots*, são ferramentas geradoras de texto que vêm ganhando muito espaço com os recentes avanços na área de inteligência artificial baseada em aprendizado de máquina. São utilizados em inúmeras aplicações para que os usuários possam realizar buscas, sanar dúvidas, dentre outras atividades.

Neste contexto, o Centro de Inteligência Artificial (C4AI) está desenvolvendo um agente chamado BLAB, um *chatbot* especializado no domínio da Amazônia Azul. Este é o território marítimo brasileiro, de grande importância econômica e ambiental. A motivação por trás do projeto é a de criação de uma ferramenta que facilite o acesso ao imenso arcabouço de conhecimento acerca deste território.

Dentre outras tecnologias, o BLAB é composto por um modelo respondedor, que fornece respostas breves e objetivas a perguntas factuais. Até o momento, o utilizava-se o DEEPAGÉ, um agente respondedor desenvolvido por pesquisadores da USP. Trata-se de um *transformer*, um tipo de modelo de inteligência artificial considerado estado da arte no que diz respeito a geradores de texto.

Todavia, o DEEPAGÉ em sua forma original apresenta uma série de deficiências. Fornece respostas curtas e secas, por vezes incorretas, em decorrência de ter sido treinado com um conjunto de dados limitado e de baixa qualidade.

Assim, o objetivo principal deste projeto é reduzir ou até eliminar estes problemas com a criação do DEEPAGÉ 2, uma versão atualizada do modelo, obtida por meio de um segundo treinamento do agente sobre um *dataset* expandido.



Figura 1 – Ilustração do Território da Amazônia Azul

Desenvolvimento

Para que um agente como o DEEPAGÉ seja treinado, são necessários exemplos de pergunta e resposta. Todavia, o conhecimento disponível da Amazônia Azul é composto por textos corridos, como páginas de Wikipedia e artigos de jornal, sendo assim necessária a conversão destes documentos ao formato adequado.

Para tanto, foi desenvolvido um sistema intermediário, um gerador de *datasets* de pergunta e resposta. Este possui diversos componentes de processamento e geração de texto, como ilustrado na figura 2. Podemos destacar a importância do gerador de perguntas. Até o momento, não havia um modelo deste tipo em língua portuguesa que apresentasse resultados considerados suficientemente bons. Então, como parte deste projeto, uma rede neural *transformer* geradora de perguntas foi treinada.

Com este sistema e um conjunto de documentos sobre a Amazônia Azul em mãos, foi gerado um *dataset* de mais de duas mil perguntas com o qual foi realizado o treinamento do DEEPAGÉ 2.

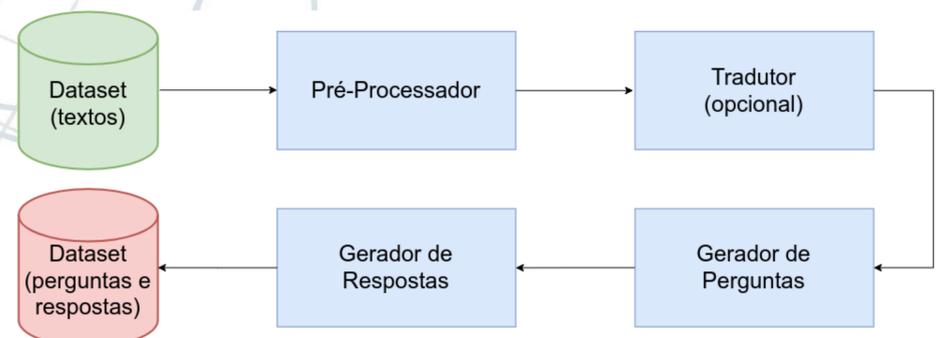


Figura 2 – Arquitetura do Sistema Gerador de Datasets

Resultados e Conclusão

Após o treinamento do DEEPAGÉ 2, este foi avaliado utilizando as métricas automáticas Exact Match, F1 Score e ROUGE-L. Comparando os resultados aos obtidos pelo DEEPAGÉ treinado somente com o *dataset* original, temos um aumento de ao menos 39% na performance em cada uma das métricas.

Podemos concluir que o trabalho foi um sucesso, tendo em vista as claras melhorias apresentadas pelo modelo revisado. Destaca-se também a importância da contribuição intermediária de criação do primeiro modelo *transformer* em língua portuguesa para geração de questões.