Fala-me sobre

Prémio Arquivo.pt 2023

1 Identificação

- Título: Fala-me sobre
- Área temática: Informática; Linguística Computacional; Inteligência Artificial
- Candidatos: João Figueira e Pedro Mendes
- Emails: joaoperfig@gmail.com e pedro.neto.mendes29@gmail.com

2 Descrição do Trabalho

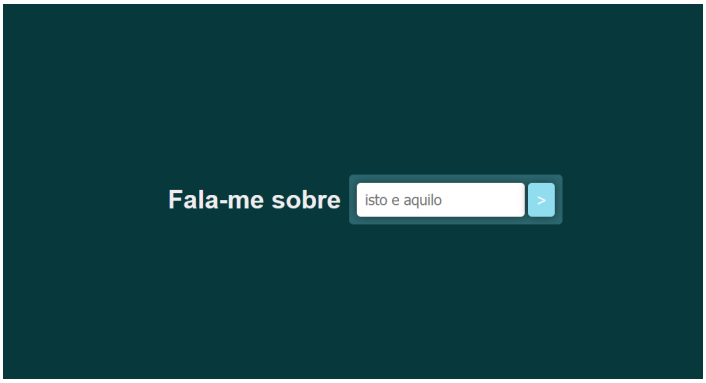

Figura 1: Frontpage da página "Fala-me sobre".

O mundo da linguística computacional tem vindo a ser frequentemente revolucionado por modelos neuronais "deep"como o ELMo [1] ou o BERT [2]. Recentemente, o aparecimento do ChatGPT e outros modelos de linguagem generativos grandes (LLMs) [3] tem contribuído para avanços em várias áreas. Uma importante função destes tem sido para chatbots capazes de ter conversas sobre uma grande variedade de temas.

Estes modelos, porém, apresentam algumas limitações. Para os treinar, são necessários muitos recursos, como grandes volumes de dados e ferramentas computacionais especializadas(como GPUs de treino). Estes requisitos apresentam desafios para uso individual ou de organizações mais pequenas.

Como modelos treinados para objetivos mais gerais, não são especializados num tópico e mostrando dificuldade a responder a perguntas mais específicas. Além disso, as conversas podem-se basear em respostas genéricas ou irrelevantes, o que pode afetar negativamente a qualidade da conversa. Modelos generativos de texto também são conhecidos por poder apresentar o problema coloquialmente chamado de "alucinações" [4], em que o modelo inventa espontaneamente informação falsa.

Para ultrapassar estas limitações, o "Fala-me sobre"usa as paginas arquivadas pelo Arquivo.pt, em específico, os conteúdos em texto destas, para criar modelos de linguagem especializados no tópico pesquisado. Para tal, experimentámos duas alternativas:

1. Usar modelos já existentes, nomeadamente o ChatGPT, com o auxílio um texto de informação pertinente para o tópico desejado, obtido através do arquivo.pt, e processado por um modelo de linguagem, para informar e contextualizar o agente.

2. Fazer finetune de modelos GPT open-source, em grandes corpus de texto sobre tópicos niche, obtidos no arquivo.pt. Para tal utilizar LoRAs [5] para um treino mais rápido dos e para impedir que estes percam informação já adquirida.

Descrevemos, primeiramente, os passos da primeira abordagem:

1. Recolha de texto sobre o tópico pesquisado dos sites guardados pelo arquivo.pt, separadamente para vários períodos do passado.

2. Envio destes textos e do tópico pesquisado para a API do GPT 3.5 para filtrar conteúdo não relevante e resumir a restante informação.

3. Compilação dos vários resumos e organização em timeline de eventos e informação, subsequentemente enviada como background para informar e contextualizar o modelo de linguagem.

4. Inicialização do chatbot capaz de usar a informação geral disponível, complementado pela informação mais específica sobre o tema pesquisado.

Para a segunda abordagem, são necessários os seguintes passos:

1. Recolha de corpus grande de texto sobre o tópico pesquisado dos sites guardados pelo arquivo.pt.

2. Download de modelos GPT pré-treinados. Devido às nossas limitações de hardware usámos o relativamente pequeno modelo opt_1.3b[6] da Facebook.

3. Finetune de um adaptador LoRA, que adapte o modelo aos textos recolhidos, usando a plataforma de finetuning "oobabooga".

4. Inicialização do chatbot capaz de usar a informação geral disponível, complementado pela informação mais específica sobre o tema pesquisado.

3 Objetivos

O objetivo principal deste trabalho é disponibilizar um chatbot com conhecimentos mais específicos sobre um tópico selecionado e permitir ao utilizador conversar com este. Para tal, a utilização da ferramenta Arquivo.pt simplifica em tempo e complexidade o treino do modelo de linguagem, pois permite o acesso aos dados relevantes em tempo real sem recursos adicionais. Assim, este projeto visa a criar um parceiro de conversa de maior qualidade para aprofundar os conhecimentos num tema selecionado.

4 Resultados Atingidos

O "Fala-me sobre"consegue reunir informação usando o Arquivo.pt e otimizar um modelo de linguagem para criar um chatbot especializado, capaz de ter conversas mais completas neste tema.

Para a primeira abordagem, criámos um sistema com frontend html e uma API em python a gerir os dados e a comunicação entre as várias APIs e o frontend. Os resultados que mostramos nesta secção foram obtidos nesta plataforma. Um screenshot da página inicial é mostrado na figura 1.

Quando aplicado a um tema mais geral, um que o chatGPT também tem conhecimento sobre, pode não existir grande diferença entre a qualidade das respostas dadas. Isto pode ser observado na figura 2.

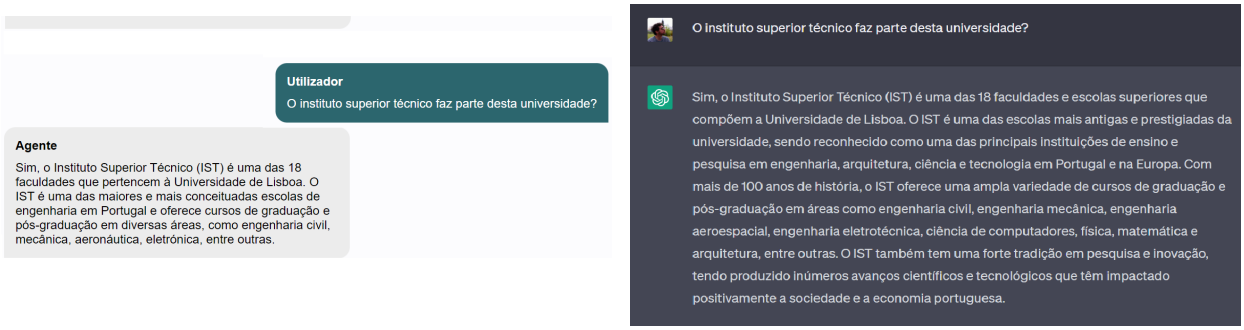

Figura 2: Exemplo de uma pergunta com resultados semelhantes para os dois chatbots.

No entanto, ao fazer perguntas mais específicas, começa-se a ver que as respostas do chatGPT ficam muito gerais sem oferecer conteúdo relevante. Na figura 3, pode-se observar isto mesmo. O chatGPT responde dizendo que há esforços a ser feitos para aumentar a opção de alojamento para os estudantes, mas não detalha nenhuma solução específica. O chatbot do "Fala-me sobre", por outro lado, responde dizendo que um investimento de 30 milhões está a ser feito para esse mesmo problema e que se espera a oferta de 1000 camas nos próximos dois anos. Comprova-se assim que o chatbot do "Fala-me sobre"consegue ser mais detalhado e apresentar factos específicos à questão.

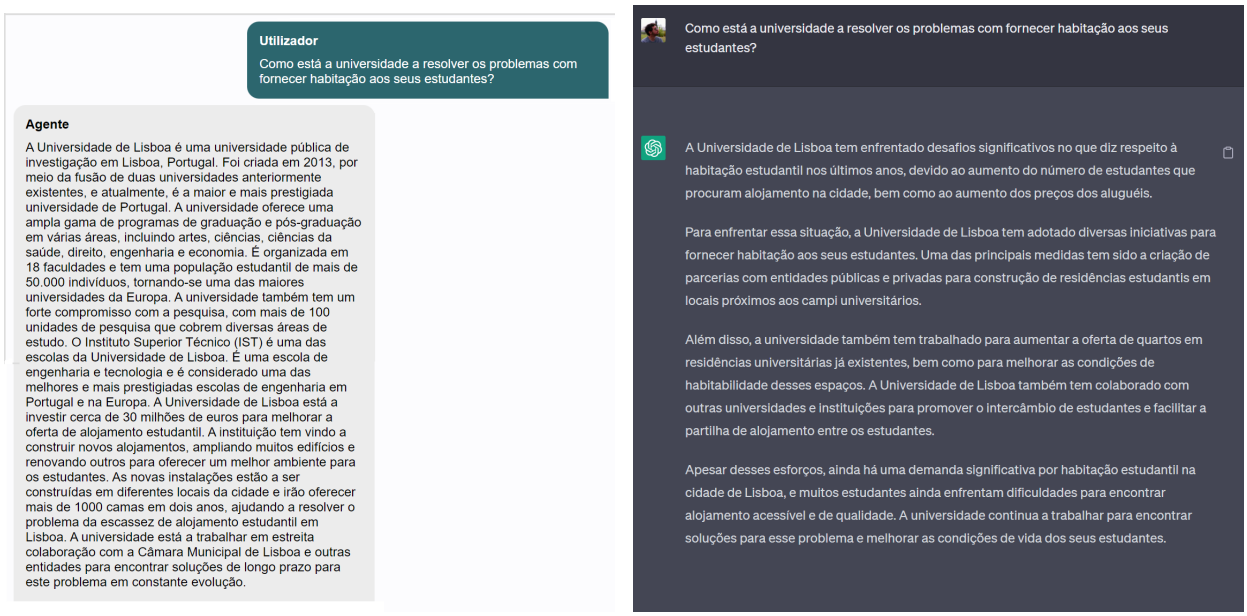

Figura 3: Exemplo de uma pergunta com resposta mais completa e específica para o chatbot do "Fala-me sobre".

A verdadeira vantagem revela-se quando este modelo é usado para temas menos populares. Nestes casos, o chatGPT reconhece que não consegue responder à pergunta, pois não possui conhecimento, como pode ser visto na figura 4

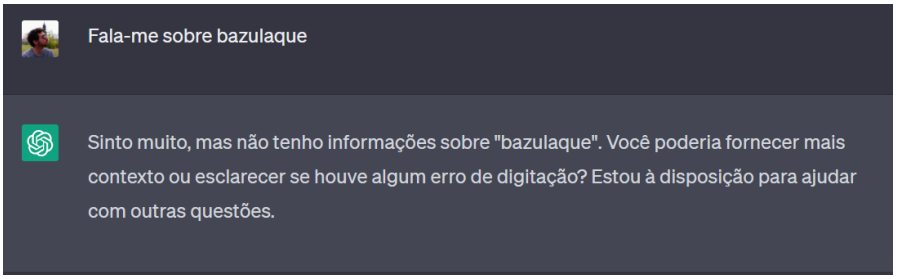

Figura 4: Exemplo do chatGPT a não conseguir responder a uma pergunta.

No entanto, o chatbot do "Fala-me sobre"não só mostra conhecer o tema, como consegue usar os recursos usuais do chatGPT. Consegue assim integrar o novo conhecimento adquirido através do Arquivo.pt nas funcionalidades normais do chatGPT.

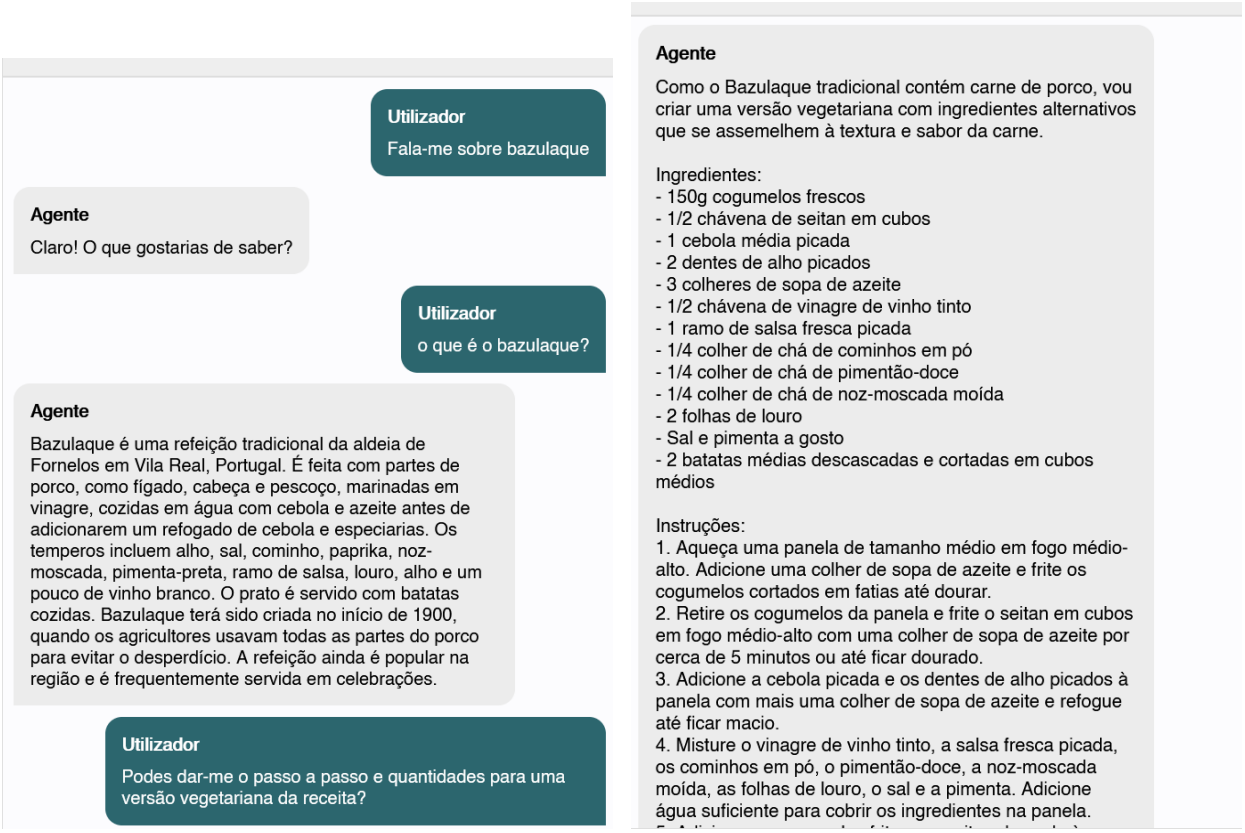

Figura 5: Exemplo do chatbot do "Fala-me sobre"a ter uma conversa sobre o mesmo assunto em que o chatGPT falhou.

Adicionalmente, o ChatGPT pode mesmo chegar a dar respostas erradas para perguntas acerca de temas mais específicos. Na figura 6 mostramos um exemplo de o ChatGPT a responder erradamente ao número de jogadores no jogo "sueca italiana". Isto é corrigido quando o modelo tem o contexto correto como mostramos no "Fala-me sobre".

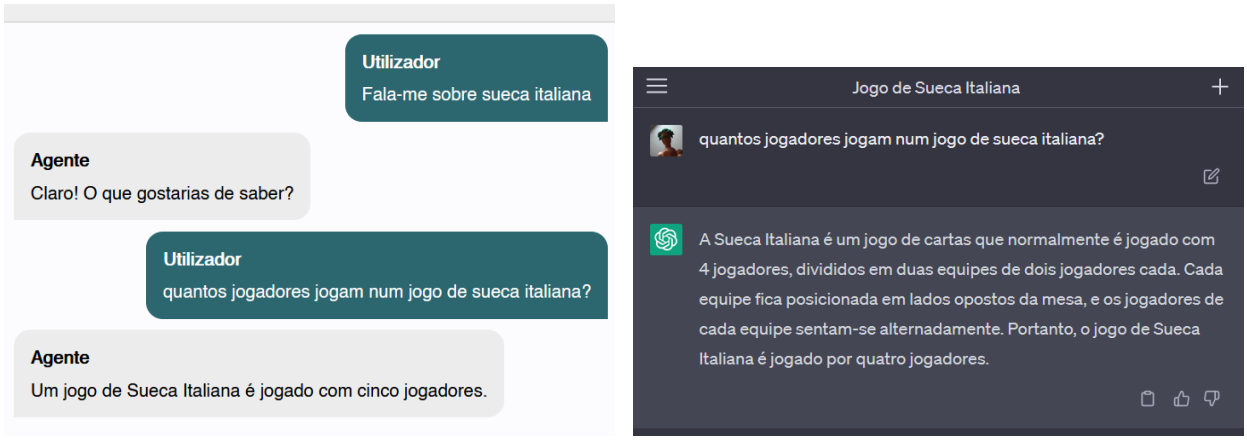

Figura 6: Comparação de uma pergunta corretamente respondida pelo "Fala-me sobre", a que o chatGPT responde erradamente.

Para a segunda abordagem, concluímos que as limitações de hardware não nos permitiram treinar um modelo suficientemente grande para poder ter uma desempenho equiparável ao ChatGPT. Na nossa experiência treinámos um modelo em artigos recolhidos acerca do código da estrada português. Infelizmente o modelo mostrou grandes quantidades de alucinação e desvio do tópico, como pode ser visto na imagem **??**. Esperamos no futuro ter acesso a recursos que nos permitam revisitar esta experiência e obter melhores resultados.

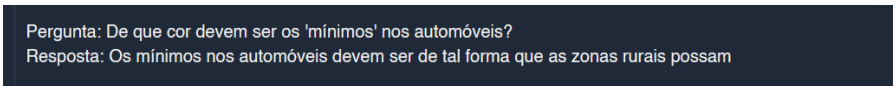

Figura 7: Exemplo do OPT_1.3 treinado em código da estrada a falhar na resposta a uma pergunta de código.

Resumidamente, para temas mais gerais, o nosso chatbot oferece alguma vantagem ao responder de forma menos geral e apresentado mais factos concretos que o chatGPT. Em temas mais específicos, chatbots como o chatGPT falham em conseguir ter conversas sobre os tópicos escolhidos, podendo admitir que não têm qualquer informação sobre este tema, nestes casos o chatbot do "Fala-me sobre"não só reconhece o tópico, como consegue conversar sobre este, dando a conhecer a informação que recolheu anteriormente.

5 Originalidade e caráter inovador

O "Fala-me sobre"é um projeto inovador que usa a vasta quantidade de informação arquivada pelo Arquivo.pt para otimizar modelos de linguagem em tópicos específicos pesquisados. Esta abordagem permite criar chatbots com conhecimentos mais profundos no assunto escolhido. Este trabalho é pioneiro no uso do Arquivo.pt para recolher dados que podem ser usados nestes modelos. Mas será apenas um primeiro passo nesta área.

6 Impacto social (aplicação e utilidade social)

Devido ao facto de o "Fala-me sobre"expandir o conhecimento do ChatGPT para incluir informações mais específicas sobre Portugal, esperamos aumentar a utilidade desses modelos para auxiliar em tópicos como gastronomia local, tradições portuguesas e eventos regionais, de forma a ter um impacto significativo na vida de uma variedade mais ampla de utilizadores em Portugal, incluindo minorias e pessoas que vivem em aldeias mais isoladas.

O "Fala-me sobre"foi pensado para ser uma extensão ao uso normal do motor de busca do Arquivo.pt, simplificando a pesquisa de informação. Usando o chatbot, podemos fazer várias perguntas sobre o tema em questão sem ter de procurar extensivamente nos recursos fornecidos pelo Arquivo.pt.

No entanto, é importante mencionar as limitações desta ferramenta. Estes chatbots não são infalíveis agora e a informação deve ser sempre confirmada nas páginas resultantes da pesquisa no Arquivo.pt. No futuro esperamos incluir citações de onde encontrar a informação específica mencionada e com os avanços gerais em modelos de linguagem, ter cada vez mais confiança nos resultados apresentados. Desta forma, o "Fala-me sobre"tem utilidade social relevante ao tornar o processo de pesquisa de informações mais eficiente e acessível para os utilizadores do Arquivo.pt.

7 Impacto científico (aplicação e utilidade científica)

Este projeto demonstra a vantagem em utilizar a API do Arquivo.pt para obter dados que depois são usados no treino de modelos. Esta utilização pode levar à criação de várias e diversas ferramentas baseadas neste principio, usando este projeto como exemplo.

Além disso, a nossa experiência mostra também a importância da dimensão de modelos tipo GPT para conseguirem dar respostas precisas e concisas sobre um tópico. O nosso projeto mostra também que na era dos LLMs o paradigma de treinar modelos em corpos grandes de dados começa a ser gradualmente substituído pelo o prompt engineering e que modelos pre-treinados, aos quais é dada informação e exemplos, podem ter tão boa ou melhor desempenho que modelos finetunned.

8 Relevância da utilização do Arquivo.pt

O Arquivo.pt foi uma das ferramentas principais deste projeto e fundamental para o sucesso deste. Usando a API do Arquivo.pt, a procura de dados foi simplificada em complexidade, tempo e ferramentas necessárias (ferramentas de web scrapping). Além disso, como esta informação está guardada, permite a utilização deste chatbot em tópicos onde a informação já não está disponível noutro lugar. Assim, em certos tópicos, a utilização do Arquivo.pt permite-nos adicionar informação relevante que seria impossível de outra forma.

9 Comentários adicionais

Como referido anteriormente, pensamos que o Arquivo.pt pode ser muito útil para fazer um ajuste fino a modelos de inteligência artificial. Trabalhamos neste objetivo, mas com resultados ainda não satisfatórios.

Pensamos que estes resultados vêm da falta de acesso a recursos mais poderosos, como GPUs capazes de treinar modelos maiores e à filtragem da informação recolhida que tem de ser feita com maior rigor.

Para filtrar documentos com melhor qualidade, filtramos documentos em jornais, usando a lista de jornais centenários disponibilizada pela AMCC, à qual adicionamos mais jornais contemporâneos.

Adicionalmente, o sistema atual demora cerca de 40 segundos a criar um agente. No futuro, este tempo pode ser otimizado paralelizando os pedidos às APIs do chatGPT e do Arquivo.pt.

10 Recursos complementares

- Website temporário, http://188.37.68.45:22125/: Site onde se pode testar o "Fala-me sobre"
- Github, https://github.com/joaoperfig/falamesobre: Github com acesso a todos os códigos e instruções de instalação do projeto localmente
- API ChatGPT, https://openai.com/blog/introducing-chatgpt-and-whisper-apis: Página com informação sobre utilização da API do chatGPT
- Modelo OPT, https://huggingface.co/facebook/opt-1.3b: Acesso ao download de modelo GPT para finetune local
- Plataforma oobabooga, https://github.com/oobabooga/text-generation-webui: Plataforma para finetune de LoRAs e modelos GPT

Referências

[1] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[4] S. Das, S. Saha, and R. K. Srihari, "Diving deep into modes of fact hallucinations in dialogue systems," 2023.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.

[6] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "Opt: Open pre-trained transformer language models," 2022.