Tradução Neural Customizada: Aplicação de NLP no Domínio Biomédico

Autores:
TIA: 10403109 NOME: Erik Samuel Viana Hsu
TIA: 10400995 NOME: Mateus Kenzo Iochimoto
TIA: 10400764 NOME: Thiago Shihan Cardoso Toma

1. Introdução

A tradução automática tem evoluído significativamente com o avanço dos modelos de aprendizado profundo, especialmente no campo de Processamento de Linguagem Natural (NLP). Modelos de tradução neural, como os baseados em Transformers, têm demonstrado grande eficácia ao capturar padrões complexos em textos e produzir traduções de alta qualidade. No entanto, muitos modelos pré-treinados apresentam limitações ao serem aplicados em domínios específicos, como o biomédico, onde terminologias e contextos especializados são importantes.

Este projeto tem como objetivo desenvolver um sistema de tradução automática neural do inglês para o português, utilizando a biblioteca Hugging Face Transformers. O foco está em adaptar um modelo pré-treinado para tradução (*fine-tuning*) com dados personalizados do EMEA Parallel Corpus, um conjunto de frases paralelas especializado no domínio biomédico. Com essa abordagem, buscamos melhorar a qualidade da tradução em textos relacionados à área médica e farmacêutica, onde a precisão linguística e contextual é crucial.

2. Técnicas Utilizadas

2.1 Dados Utilizados

Os dados para o treinamento e avaliação foram retirados do EMEA Parallel Corpus, que contém frases paralelas em inglês e português, voltadas para o domínio biomédico. Este corpus foi escolhido devido à sua relevância no contexto médico, com terminologias específicas que não são bem representadas.
Os arquivos utilizados foram:

- **EMEA.en-pt.en:** Frases originais em inglês.
- **EMEA.en-pt.pt:** Traduções correspondentes em português.

Para o treinamento, foram usados 2000 pares de frases como conjunto de treino e 500 pares de frases para validação.
Um exemplo de frase que existe nesses arquivos:

EN:

“PHARMACEUTICAL PARTICULARS
6.1 List of excipients
Lactose monohydrate Maize starch Microcrystalline cellulose Hydroxypropyl cellulose Magnesium stearate
Indigo carmine aluminium lake (E132)”

PT:
“INFORMAÇÕES FARMACÊUTICAS
Lista dos excipientes
Lactose mono- hidratada Amido de milho Celulose microcristalina Hidroxipropilcelulose Estearato de magnésio
Laca alumínica de carmim de indigo (E132)”

2.2 Modelo e Biblioteca

O modelo utilizado foi o Helsinki-NLP/opus-mt-en-ROMANCE, disponível na biblioteca Hugging Face Transformers. Este modelo é parte da família OPUS-MT, que contém modelos de tradução pré-treinados para múltiplos pares de idiomas. A escolha por este modelo se deu por sua capacidade de traduzir do inglês para línguas de origem latinas, incluindo o português.

O processo de adaptação (fine-tuning) envolveu:

- Treinamento supervisionado com os pares de frases do EMEA Parallel Corpus.
- Ajuste dos pesos do modelo para melhorar a tradução em um domínio específico (biomédico).

2.3 Ferramentas e Pacotes

As seguintes ferramentas foram utilizadas para a implementação do sistema:

- Transformers: Para carregamento e fine-tuning do modelo.
- Datasets: Para manipulação dos dados e tokenização.
- Evaluate: Para cálculo da métrica de avaliação sacreBLEU.
- PyTorch: Framework backend para o treinamento do modelo, garantindo flexibilidade e eficiência computacional.

2.4 Avaliação

A métrica utilizada para avaliar o desempenho do modelo foi a *sacreBLEU*. Essa métrica mede a similaridade entre a tradução gerada pelo modelo e a tradução de referência, sendo amplamente utilizada em tarefas de tradução automática.

Resultados preliminares indicaram que o modelo ajustado conseguiu traduzir frases biomédicas com maior precisão em comparação ao modelo pré-treinado

original. Por exemplo, ao traduzir "Keep medicine away from children", o modelo produziu a tradução correta: "Manter os medicamentos longe das crianças".

2.5 Testes

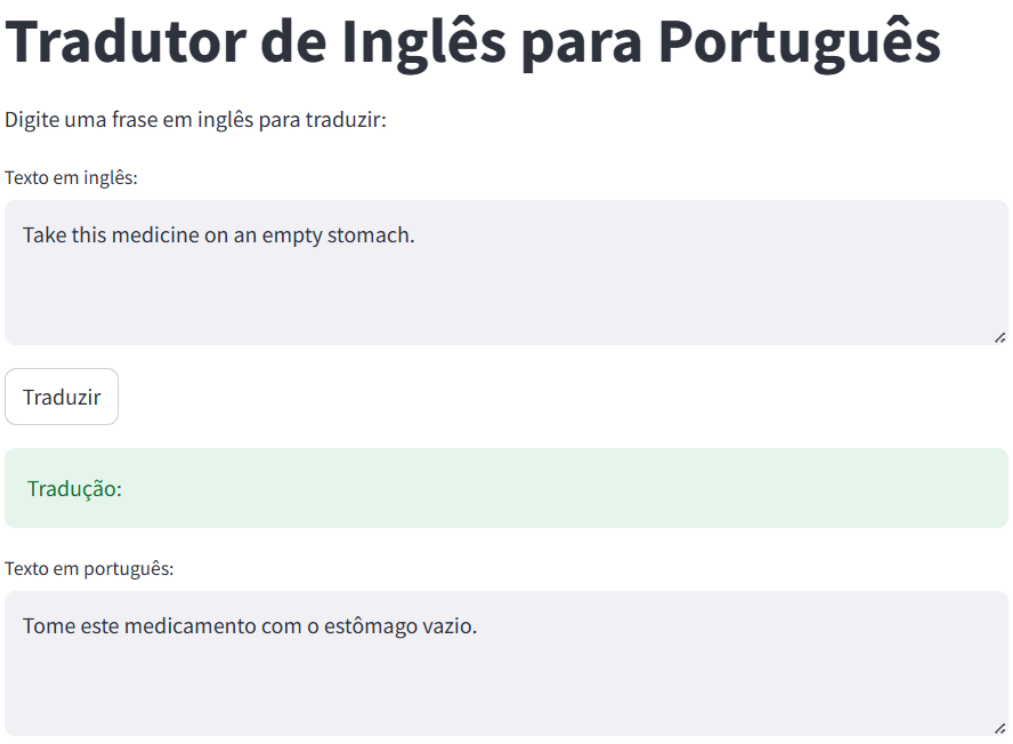
Para esta seção pensamos em criar uns casos de testes que pensamos interessantes para testar o funcionamento desse modelo de tradução, para isso pensamos nos seguintes casos:

a) Frases dentro do contexto biomédico;
   i) frases curtas e diretas.
   ii) frases técnicas.
   iii) frases com termos específicos.
b) Frases com leves erros gramáticos;
c) Frases que possuem palavras com duplo contexto;
d) Frases com palavras extraídas do documento usado para fine-tuning do modelo;

2.5.1 Frases para o domínio biomédico:

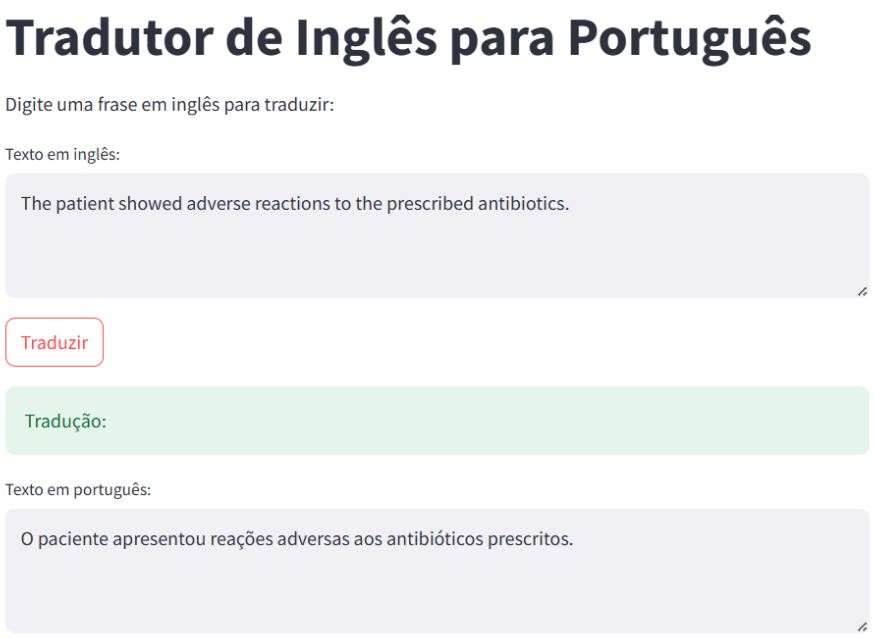
2.5.1.1 Frases curtas e diretas:

- Entrada: *"Take this medicine on an empty stomach."*
- Esperado: *"Tome este medicamento com o estômago vazio."*



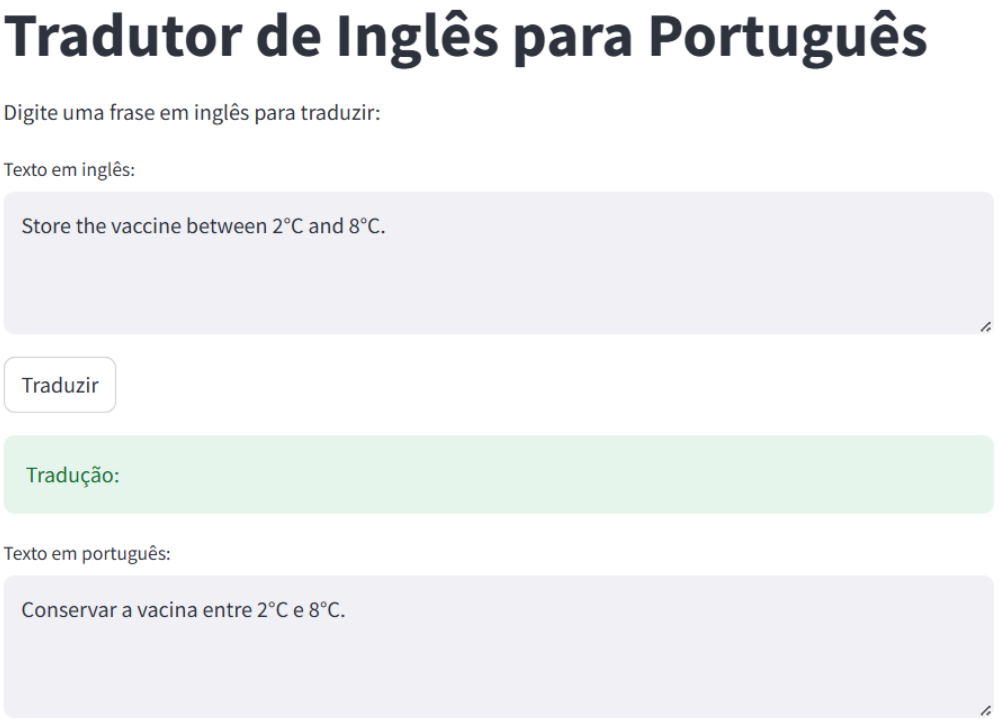
2.5.1.2 Frases técnicas:

- Entrada: *"The patient showed adverse reactions to the prescribed antibiotics."*
- Esperado: *"O paciente apresentou reações adversas aos antibióticos prescritos."*



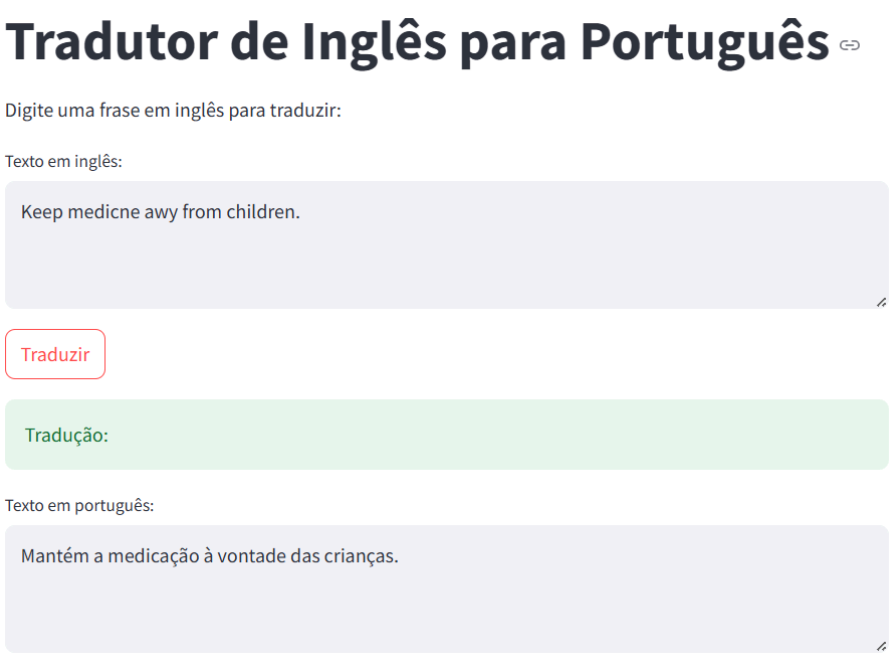
2.5.1.3 Frases com termos específicos:

- Entrada: *"Store the vaccine between 2°C and 8°C."*
- Esperado: *"Armazene a vacina entre 2°C e 8°C."*



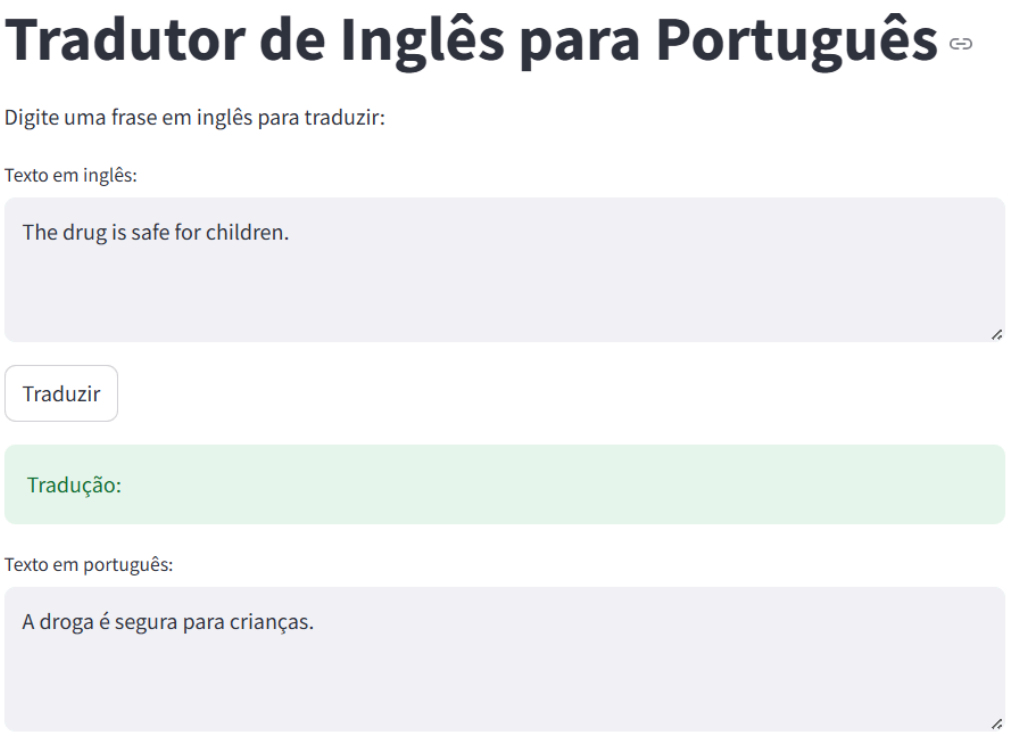
2.5.2 Frases com leves erros gramáticos

- Entrada: *"Keep medicne awy from children."* (com erros de digitação)
- Esperado: *"Mantenha os medicamentos longe das crianças."*



2.5.3 Frases que possuem palavras com duplo contexto

- Entrada: *"The drug is safe for children."*
    a. Pode significar: "O medicamento é seguro para crianças."
    b. Ou: "A droga é segura para crianças." (Dependendo do contexto, "drug" pode ser traduzido como "medicamento" ou "droga").



2.5.4 Frases com palavras extraídas do documento usado para fine-tuning do modelo

EN:
“PHARMACEUTICAL PARTICULARS
6.1 List of excipients
Lactose monohydrate Maize starch Microcrystalline cellulose Hydroxypropyl cellulose Magnesium stearate
Indigo carmine aluminium lake (E132)”

PT:
“INFORMAÇÕES FARMACÊUTICAS
Lista dos excipientes
Lactose mono- hidratada Amido de milho Celulose microcristalina Hidroxipropilcelulose Estearato de magnésio
Laca alumínica de carmim de indigo (E132)”

The formulation includes several excipients chosen for their functional properties. **Lactose monohydrate** serves as a filler and aids in tablet compressibility, while maize starch functions both as a binder and a disintegrant. **Microcrystalline cellulose** contributes to tablet strength and enhances flowability during manufacturing. **Hydroxypropyl cellulose** is included as a binder and to improve the tablet's dissolution profile. **Magnesium stearate** acts as a lubricant to ensure smooth tablet ejection from the press. For coloring purposes, the formulation also contains **indigo carmine aluminium lake (E132)**, providing a distinct visual identity to the product.

A formulação inclui vários excipientes escolhidos pelas suas propriedades funcionais. A **lactose mono-hidratada** serve como um enchimento e ajuda na compressibilidade dos comprimidos, enquanto o amido de milho funciona tanto como um ligante e um desintegrante. **Celulose microcristalina** contribui para a resistência dos comprimidos e aumenta a fluibilidade durante a fabricação. A **hidroxipropilcelulose** é incluída como um ligante e para melhorar o perfil de dissolução do comprimido. **Estearato de magnésio** atua como um lubrificante para garantir a ejeção suave dos comprimidos da prensa. Para efeitos de coloração, a formulação também contém o **lago de alumínio indigo carmine (E132)**, fornecendo uma identidade visual distinta ao produto.

3. Conclusão

O presente trabalho demonstrou que o fine-tuning de um modelo de tradução automática neural com dados especializados pode melhorar de forma significativa a qualidade das traduções em domínios técnicos, como o biomédico. A utilização do

modelo **Helsinki-NLP/opus-mt-en-ROMANCE**, em conjunto com o **EMEA Parallel Corpus**, permitiu que o sistema aprendesse terminologias e expressões específicas do setor médico, resultando em traduções mais precisas e contextualizadas.

Apesar dos avanços alcançados, o modelo ainda apresenta desafios. Foram observadas dificuldades em lidar com termos sensíveis ao contexto, sentenças com erros leves de gramática, e em alguns casos, com a escolha de palavras mais apropriadas ao vocabulário técnico biomédico. Esses aspectos indicam que há espaço para aprimoramentos futuros, especialmente no refinamento da capacidade do modelo em interpretar nuances linguísticas e adaptar-se melhor a estruturas linguísticas menos convencionais.

Para trabalhos futuros, propõe-se a ampliação do conjunto de dados especializados, a inclusão de mecanismos que melhorem a sensibilidade ao contexto e a incorporação de estratégias de pós-processamento linguístico. Além disso, o sistema pode ser adaptado para outros pares de idiomas e domínios técnicos, o que amplia sua aplicabilidade e impacto potencial em diferentes áreas do conhecimento.

4. Referências

[1] OPUS EMEA Corpus: https://opus.nlpl.eu/EMEA.php
[2] MarianMT - Hugging Face: https://huggingface.co/Helsinki-NLP/opus-mt-en-pt
[3] Hugging Face Transformers: https://huggingface.co/docs/transformers/index
[4] Streamlit: https://streamlit.io/
[5] sacreBleu: https://huggingface.co/spaces/evaluate-metric/sacrebleu
[6] fine-tuning de um modelo: https://medium.com/data-hackers/desmistificando-o-fine-tuning-de-llms-na-pr%C3%A1tica-peft-lora-qlora-e-hamb%C3%BArgueres-ca6e6008241f