

UNIDOMBOSCO
CURSO DE CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

LUCAS KAMINSKI SCHIMIDT DA SILVEIRA

**PROJETO DO CURSO CIÊNCIAS DE DADOS I:
DESENVOLVIMENTO DE UMA PESQUISA CIENTÍFICA**

CURITIBA
2023

SUMÁRIO

1. ESCOLHA DO TEMA	3
2. DEFINIÇÃO DO PROBLEMA	3
3. HIPÓTESES	3
3.1 Formulações de prompt específicos geram traduções mais precisas	3
3.2 Variações no prompt afetam a precisão da tradução	3
3.3 Estruturas de prompt influenciam na fluidez das traduções	4
4. OBJETIVOS	4
4.1 Avaliar a sensibilidade dos LLMs em diferentes estruturas de prompt	4
4.2 Identificar estratégias para promover traduções mais precisas	4
5. METODOLOGIA	4
5.1 Etapas	4
5.1.1 Revisão bibliográfica	4
5.1.2 Coleta dos dados	5
5.1.2 Formulação dos prompts	5
5.1.3 Execução dos experimentos	5
5.1.4 Análise e interpretação dos resultados	5
5.2 Ferramentas	5
5.2.1 Linguagem de programação	5
5.2.2 LLM	6
5.2.3 Métricas de avaliação	6

1. ESCOLHA DO TEMA

O tema selecionado para a pesquisa foca na utilização de Large Language Models (LLMs) como tradutores, com ênfase no impacto da construção do prompt na qualidade da tradução final. A relevância desse estudo está na compreensão do funcionamento dos modelos linguísticos de grande escala e sua influência direta na precisão das traduções, seja em processos totalmente automáticos ou assistidos por humanos. Este estudo surge em um momento em que tais modelos desempenham um papel fundamental na tradução em diversos sistemas e aplicações, resultado de sua crescente popularização global.

2. DEFINIÇÃO DO PROBLEMA

O problema nesta pesquisa se concentra em compreender de que maneira a variação na construção do prompt influencia a qualidade das traduções realizadas por Large Language Models (LLMs). O foco está na necessidade de identificar e analisar como ajustes na formulação do input fornecido aos LLMs impactam a precisão, fluidez e adequação semântica das traduções geradas, investigando a sensibilidade dos LLMs à formulação do prompt e explorando como diferentes abordagens na construção desse input podem resultar em variações significativas no desempenho de tradução. Dessa forma, o problema central aborda não apenas a influência da construção do prompt na qualidade das traduções, mas também as possíveis estratégias para otimizar essa formulação, aprimorando a eficácia dos LLMs como ferramentas de tradução confiáveis e precisas.

3. HIPÓTESES

3.1 Formulações de prompt específicos geram traduções mais precisas

Acredita-se que prompts mais específicos e que sejam direcionados ao objetivo final, com contexto prévio e informações estruturais, fornece um maior entendimento da intenção do texto e portanto, gerará uma tradução de maior qualidade.

3.2 Variações no prompt afetam a precisão da tradução

Diferentes formulações e estruturações de prompt têm impacto direto na precisão das traduções realizadas, podendo resultar em variações significativas na qualidade do resultado final e na adequação semântica da tradução.

3.3 Estruturas de prompt influenciam na fluidez das traduções

Estruturas de prompt mais coerentes e alinhadas com a gramática e o contexto linguístico tendem a resultar em traduções mais fluidas, evitando quebras ou ambiguidades e gerando um resultado mais natural e próximo ao texto original.

4. OBJETIVOS

4.1 Avaliar a sensibilidade dos LLMs em diferentes estruturas de prompt

Pretende-se investigar e compreender a sensibilidade dos Large Language Models (LLMs) em relação à variação na estrutura do prompt fornecido para a tradução. A partir de experimentos e análises comparativas, utilizando diferentes estruturas de prompts, pode-se verificar como as variações influenciam a qualidade e a precisão das traduções realizadas pelos LLMs.

4.2 Identificar estratégias para promover traduções mais precisas

Visa-se identificar estratégias específicas na construção do prompt que possam aprimorar a precisão das traduções realizadas pelos LLMs. Serão analisadas abordagens e técnicas que, ao serem aplicadas na formulação do prompt, tendem a resultar em traduções mais fiéis ao texto original, minimizando possíveis distorções ou erros.

5. METODOLOGIA

Para a realização desta pesquisa será adotado uma abordagem experimental e analítica, com a realização de experimentos controlados para avaliar o impacto da variação na formulação do prompt nos resultados das traduções feitas pelos Large Language Models (LLMs). Posteriormente, serão realizadas análises comparativas entre diferentes estratégias de formulação de prompts para identificar padrões que influenciam positivamente na qualidade das traduções.

5.1 Etapas

5.1.1 Revisão bibliográfica

Será realizada uma revisão bibliográfica para identificar e analisar trabalhos relacionados ao tema, com o objetivo de compreender o estado da arte e as principais abordagens e técnicas utilizadas na construção de prompts para tradução com LLMs.

5.1.2 Coleta dos dados

Processo de coleta manual de dados de tradução de textos em espanhol para o português, disponibilizados pelo projeto OPUS-MT, disponíveis no link <http://opus.nlpl.eu/>, que contém um conjunto de corpora de tradução de código aberto, com textos em diversos idiomas. Nesta etapa serão selecionados textos de diferentes tamanho e complexidade, com o objetivo de avaliar o impacto das variações em diferentes contextos.

5.1.2 Formulação dos prompts

Serão definidas diferentes estratégias para a formulação dos prompts, considerando diferentes aspectos como a inclusão de informações contextuais, a especificidade do pedido de tradução e a estrutura gramatical do prompt. Serão elaborados prompts que variam desde instruções genéricas até prompts mais específicos e detalhados, visando cobrir uma ampla gama de abordagens.

5.1.3 Execução dos experimentos

Os experimentos serão executados utilizando um LLM estruturado modelo GPT a ser definido, com a utilização de diversos prompts definidos na etapa anterior pedindo a mesma tradução de um dos textos selecionados. Os resultados serão registrados para uma posterior análise comparativa.

5.1.4 Análise e interpretação dos resultados

Ocorrerá uma análise detalhada das traduções geradas pelos LLMs com o auxílio de métricas de avaliação de qualidade de tradução, como BLEU, para comparar e avaliar a precisão, fluidez e adequação semântica das traduções produzidas por cada estratégia. Com base nos resultados obtidos, serão identificados padrões e tendências que possam indicar estratégias mais eficazes para a formulação de prompts requerentes de tradução.

5.2 Ferramentas

5.2.1 Linguagem de programação

A linguagem de programação utilizada para a realização dos experimentos será Python, devido à sua ampla utilização em projetos de Inteligência Artificial e Processamento de Linguagem Natural, além de sua facilidade de uso e grande disponibilidade de bibliotecas e frameworks.

5.2.2 LLM

O modelo de Large Language Model (LLM) utilizado para a realização dos experimentos será o GPT-3.5 da OpenAI, devido a sua capacidade de tradução entre diversos idiomas. O modelo será utilizado através da própria interface gráfica da OpenAI, disponível no link <https://chat.openai.com/>.

5.2.3 Métricas de avaliação

As métricas de avaliação de qualidade de tradução utilizadas serão o BLEU, uma métrica de avaliação de qualidade de tradução que compara uma tradução candidata com uma ou mais traduções de referência e disponível através da biblioteca NLTK do ambiente Python, disponível no link <https://www.nltk.org/>.