

ABSTRACT

This dissertation explores the potential that decoder-only Large Language Models (LLMs) have shown in the field of Machine Translation (MT) and leverages it for the task of translating Diagnostic Trouble Codes (DTCs), which contain highly technical automotive language. To do so, an initial evaluation is conducted on five top-performing moderate-sized decoder-only LLMs to understand how they fare against state-of-the-art MT translation models and services such as DeepL and Google Translate. Subsequently, the capabilities of one of the best-performing models, Llama2, is enhanced through few-shot prompting and fine-tuning. The former is conducted using examples extracted from three pools of varying quality and domain, with the in-domain pools showcasing the best results. Fine-tuning was first carried out solely for one language, Portuguese, and subsequently for three languages, Portuguese, French, and Spanish. The fine-tuned models based on Llama2 far surpassed the base model performance and reached near state-of-the-art results. Additionally, although not the main objective of this dissertation, it demonstrated the benefits of using multiple and distinct evaluation metrics.

Keywords: Large Language Models, Machine Translation, Natural Language Processing, Diagnostic Trouble Codes, Automotive Field

RESUMO

Esta dissertação explora o potencial que Grandes Modelos de Linguagem (LLMs) têm demonstrado na área de Tradução Automática (MT) canalizando-o para a tarefa de tradução de Códigos de Diagnóstico (DTC), que contêm linguagem automóvel altamente técnica. Para o efeito, é efectuada uma avaliação inicial em cinco dos melhores LLMs de tamanho médio, para compreender como se comparam em relação a modelos e serviços de tradução automática de última geração, tais como Deepl e o Google Tradutor. Subsequentemente, a capacidade de um dos modelos com melhor desempenho, o Llama2, é ampliada através de *few-shot prompting* e *fine-tuning*. O primeiro é realizado utilizando exemplos extraídos de três conjuntos de qualidade e domínio variáveis, apresentando os conjuntos no domínio os melhores resultados. O *fine-tuning* foi efectuado inicialmente apenas para uma língua, o Português, e depois para três línguas, Português, Francês e Espanhol. Os modelos aperfeiçoados baseados em Llama2 ultrapassaram largamente o desempenho do modelo base e alcançaram resultados próximos do melhor que existe na área. Além de mais, embora não fosse o objetivo principal desta dissertação, foi possível demonstrar os benefícios da utilização de várias métricas de avaliação distintas.

Palavras-chave: Grandes Modelos de Linguagem, Tradução Automática, Processamento de Linguagem Natural, Código de Diagnóstico, Área Automóvel