LEVERAGING LLMS ON MACHINE TRANSLATION WITH DOMAIN-SPECIFIC CONTEXT

REPORT

Mohamed HANNANI

12 Septembre 2023

Contents

1 Introduction

The ever-expanding landscape of natural language processing and machine translation has introduced in a new era of communication, bridging linguistic divides and facilitating cross-cultural understanding. Among the groundbreaking developments in this field, large language models have emerged as powerful tools capable of handling a myriad of language-related tasks. In particular, models like GPT, Llama 2, and Falcon have garnered significant attention due to their remarkable ability to generate coherent and contextually relevant text across multiple languages.

However, machine translation extends beyond word conversion; It necessitates the preservation of nuances, idioms, and the unique stylistic attributes that characterize human language. Enter the concept of adaptive translation, a paradigm that seeks to refine machine translation by tailoring it to specific domains, genres, or styles. Importantly, to avoid the resource-intensive processes of training and fine-tuning models, adaptive translation techniques are employed, providing a more efficient and effective means of achieving domain-specific translation goals. In this endeavor, we present a comprehensive exploration of "Adaptive Machine Translation with Large Language Models" paper focusing on its application to the translation of English to Arabic, with a particular emphasis on leveraging GPT-3.5 Turbo with some improvements on the workflow and fuzzy matches selection.

This report examines the subtleties of adapting machine translation to domain-specific requirements. To do so, we employed a corpus of approximately 1,500 movie subtitles, carefully translated from English to Arabic. These subtitles encapsulate the essence of cinematic language and offer a rich source of context for our translation model. Prior to inference, we harnessed the power of the Sentence-Transformer model to compute embeddings for these subtitles, facilitating the efficient retrieval of similar sentences through the use of the FAISS indexing system developed by Facebook. This approach paved the way for the composition of contextually rich prompts, allowing GPT-3.5 Turbo to follow the stylistic cues present in the domain-specific examples.

In the subsequent sections, we detail our methodology, experimental setup, results, and discuss the broader implications of adaptive translation in machine learning. This journey showcases GPT-3.5 Turbo's potential to bridge linguistic gaps while preserving linguistic richness in English to Arabic translation.

2 Methodology

In this section, we provide a detailed account of the methodology used in our project to implement "Adaptive Machine Translation with Large Language Models" for English to Arabic translation with the assistance of GPT-3.5 Turbo. The methodology encompasses data collection, data preprocessing, sentence embedding generation, FAISS indexing, and prompt composition.

You can find all the implementation in our GitHub Repository.

2.1 Data Collection

To build a robust translation model with domain-specific knowledge, we collected a dataset of approximately 1,500 movie subtitles that had been meticulously translated from English to Arabic. These subtitles were selected to represent a diverse range of cinematic language styles and contexts.

2.1.1 Data Preprocessing

Prior to any model training or embedding generation, the collected dataset underwent rigorous preprocessing. This included the removal of duplicates, noise, and any formatting inconsistencies to ensure a clean and coherent dataset.

2.1.2 Sentence Embedding Generation

To facilitate efficient retrieval and context-aware prompts for GPT-3.5 Turbo, we employed the Sentence-Transformer model. This model was used to compute embeddings for each sentence in our preprocessed dataset. Sentence embeddings capture semantic information and contextual nuances, which is crucial for generating accurate translations.

2.1.3 FAISS Indexing

The generated sentence embeddings were indexed using the FAISS (Facebook AI Similarity Search) system. FAISS provides fast and memory-efficient similarity search capabilities, enabling quick retrieval of sentences with similar embeddings. This indexing system streamlined the process of finding contextually relevant examples for prompt composition during inference.

2.1.4 Prompt Composition

For each translation request, we utilized the FAISS index to retrieve the top-5 closest sentence embeddings from the dataset. These retrieved sentences were then used to compose contextually rich prompts for GPT-3.5 Turbo. By incorporating examples from the domain-specific dataset, we aimed to guide the model in following the desired style and context while translating sentences.

3 Examples of citations, figures, tables, references

3.1 Citations

Citations use `natbib`. The documentation may be found at

http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf

Here is an example usage of the two main commands (`citet` and `citep`): Some people thought a thing (Kour and Saabne, 2014b; Hadash et al., 2018) but other people thought something else (Kour and Saabne, 2014a). Many people have speculated that if we knew exactly why Kour and Saabne (2014a) thought this...

3.2 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure 1. Here is how you add footnotes. [1] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

3.3 Tables

See awesome Table 1.

[1] Sample of the first footnote.

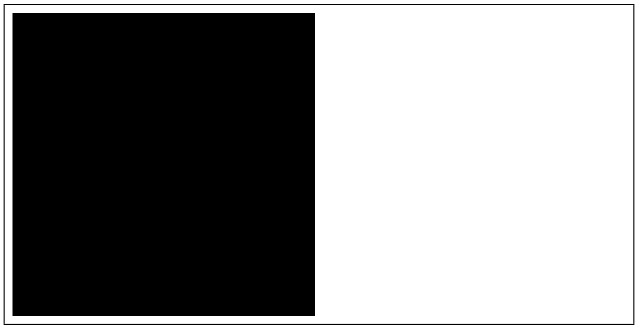

Figure 1: Sample figure caption.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

The documentation for `booktabs` ('Publication quality tables in LaTeX') is available from:

`https://www.ctan.org/pkg/booktabs`

3.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

References

Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. 2018. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*.

George Kour and Raid Saabne. 2014a. Fast classification of handwritten on-line arabic characters. In *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, pages 312–318. IEEE.

George Kour and Raid Saabne. 2014b. Real-time segmentation of on-line handwritten arabic script. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 417–422. IEEE.