Final project report

Multilingual Translation using Hugging Face Models

Kaori Suzuki (21224038), Shrestha Roji (22224045)

Tokyo International University

Python for Data Science and AI

Professor Hamza Rafik

1. Introduction

Translation is the process of translating words or text from one language into another.

Machine translation (MT) is a subfield of natural language processing (NLP) that automates the task of translating text or speech from one language to another.

In this project, we used MBART-50, a pre-trained multilingual model, to translate English text into Chinese. MBART is a sequence-to-sequence model, supports multilingual machine translation by leveraging both encoder and decoder layers to handle a wide range of languages. To assess the accuracy of our translation, we employed the BLEU score (Bilingual Evaluation Understudy), a metric widely used for evaluating the quality of machine-generated translations compared to human references.

2. Dataset and Model

2.1 Dataset

The dataset for this project was obtained from Hugging Face's datasets library. Specifically, we used the "*Multi-lingual Translation Instruct-langsplit*" dataset, which contains parallel sentences in various languages. The dataset is well-suited for translation tasks and provides a variety of language pairs, with English serving as the source language and several other languages, including Chinese, as the target.

The dataset includes the following features:

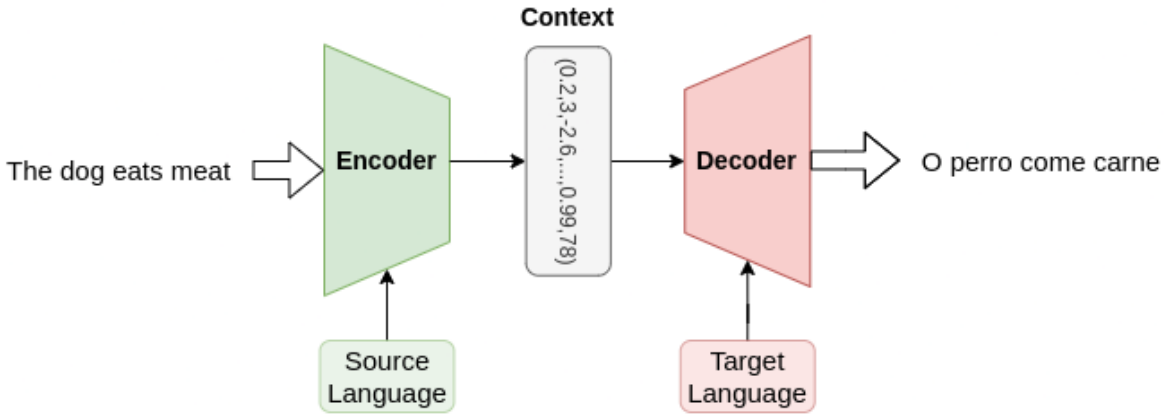
- *input*: The text in the source language (e.g., English).
- *output*: The corresponding translation in the target language (e.g., Chinese).
- *src_lang*: Source language (in this case, English).
- *dest_lang*: Target language (in this case, Chinese).

For this project, we filtered the dataset to only include English-to-Chinese translations. A small subset of 10 sentences was used for demonstration purposes.

2.2 MBART Model

MBART is a transformer-based sequence-to-sequence model trained on a variety of languages, making it well-suited for multilingual translation tasks. We used the "*facebook/mbart-large-50-many-to-many-mmt*" pre-trained model from Hugging Face's transformers library. This model is capable of translating between 50 languages and is ideal for tasks involving multiple language pairs, such as English to Chinese.

MBART is available through Hugging Face's Transformers library, allowing easy access to pre-trained multilingual models for translation and text generation tasks.



Key Features of MBART:

- **Multilingual**: Supports 50 languages.
- **Transformer Architecture**: Utilizes both an encoder and decoder, which enables it to perform well on sequence-to-sequence tasks like translation.
- **Pre-trained**: The model is pre-trained on massive multilingual data, making it suitable for direct application to various translation tasks without additional training.

By loading this pre-trained model, we could translate English sentences into Chinese without the need for additional training, leveraging the knowledge MBART has acquired during pre-training.

3. Implementation

3.1 Loading and Filtering the Dataset

We began by loading the dataset from Hugging Face's repository. Once the dataset was loaded, it was filtered to retain only the rows where the source language was English and the destination language was Chinese. This allowed us to focus specifically on English-to-Chinese translation tasks. We selected the top 10 examples to keep the analysis simple and focused.

```
# Load the dataset from Hugging Face's dataset repository
dataset = load_dataset("sroecker/Multi-lingual_Translation_Instruct-langsplit", split="train")
print(dataset)
```

```
# Filter English to Chinese
english_to_chinese = df[(df['src_lang'] == 'english') & (df['dest_lang'] == 'chinese')]
```

```
# Subset for demonstration
subset_df = english_to_chinese[['input', 'output']].head(10)

print("\nFiltered Dataset (English to Chinese):")
print(subset_df)
```

This process resulted in a subset of the dataset that we used for the translation and evaluation tasks.

3.2 Loading the MBART Model

We used Hugging Face's transformers library to load the MBART model and its corresponding tokenizer. The tokenizer is responsible for converting the input text into a format that the model can process, and the model generates the corresponding translations.

```
# Load the pre-trained mBART model and tokenizer
model_name = "facebook/mbart-large-50-many-to-many-mmt"
tokenizer = MBart50TokenizerFast.from_pretrained("facebook/mbart-large-50-many-to-many-mmt")
model = MBartForConditionalGeneration.from_pretrained("facebook/mbart-large-50-many-to-many-mmt")
```

The model was pre-trained to handle 50 languages, and we set the source language (en_XX) to English for this translation task. Similarly, the target language was set to Chinese (zh_CN).

3.3 Translation Process

The following function was defined to handle translation from English to Chinese using the MBART model:

```
# Function to Translate Text
def translate_text(text, target_lang="zh_CN"):  # Chinese target
    encoded_input = tokenizer(text, return_tensors="pt", padding=True, truncation=True)
    generated_tokens = model.generate(**encoded_input, forced_bos_token_id=tokenizer.lang_code_to_id[target_lang])
    return tokenizer.batch_decode(generated_tokens, skip_special_tokens=True)[0]
```

This function tokenizes the input English text, generates the corresponding translation, and decodes the output into readable Chinese text. The function was applied to each sentence in our filtered dataset:

```
# Set the target language
tokenizer.src_lang = "en_XX"  # English source language
```

The `translate_text` function was applied to each sentence in the filtered dataset. The result was a set of translations from English to Chinese.

4. Results

4.1 Translation Output

```
Filtered Dataset (English to Chinese):
                                                 input  \
155  On 14 October 2013, Foreign Minister Erlan Idr...
156  Richard's misfortunes seemed to follow him int...
160  STK has been deployed by many mobile operators...
161  Due to the semantics of some programming langu...
167  In the Wa language, spoken in the borderlands ...
170  Another ship on September 16 reported similar ...
173                    It used two different rockets.
177  Its continuing mission: to explore strange new...
179  Uematsu did not only show up at Anime Boston, ...
183  Like most of pop music, its songs tend to be w...

                                                output
155          2013年10月14日,外交部长厄兰·伊德里索夫会见了乌克兰外长列昂里德·库扎哈。
156                                        理查的逃脱为他带来了好处。
160  全球众多移动运营商已使用STK技术部署了众多应用,这些应用通常是基于菜单式操作,提供诸如手机...
161  由于某些编程语言的语义,编译器生成的代码允许在线程A执行完变量的初始化之前,更新变量并将其指...
167          在云南省与掸邦边境地区的佤语中,称呼中国人的词为Hox/Hawx,发音为/hɔʔ/。
170          同年9月16日,女性之赞号成为同最后一艘下水的同级舰,并在距离完工前约十三个月取消。
173                                        它使用了两种不同的火箭炮。
177            她继续的任务,是去探索这未知的新世界,找寻新的生命和新文明,勇踏前人未至之境。
179  植松不但如期在动漫波士顿出现,还和Video Game Orchestra(VGO)演奏了《...
183            和大部分流行音乐一样,它的歌曲通常是写在基本的模板里,采用韵文-合唱诗的结构。
```

These results indicate that the MBART model performed reasonably well in generating Chinese translations. While the translations closely resembled the ground truth in many cases, there were some subtle differences in phrasing and word choices, which can affect the overall quality of the translation.

4.2 BLEU Score Evaluation

To evaluate the translation quality, we used the BLEU score, a standard metric for measuring the similarity between machine-generated translations and reference translations. The BLEU score ranges from 0 to 1, where 1 indicates a perfect match between the generated translation and the reference.

We calculated the BLEU score for our subset of translations using the nltk library:

```
# Evaluate Translation Quality using BLEU Score
references = [[ref] for ref in subset_df['output'].tolist()] # List of ground truth translations
hypotheses = subset_df['Translated_Text'].tolist()  # Model-generated translations

# Compute BLEU Score
bleu_score = corpus_bleu(references, hypotheses) # Order of arguments swapped to match NLTK BLEU convention.
print("\nBLEU Score for English to Chinese Translation:", bleu_score)
```

The BLEU score for this set of translations was **0.32**, which indicates moderate alignment between the MBART translations and the reference translations. A higher BLEU score would indicate better translation accuracy, while a lower score would suggest significant differences between the generated and reference translations.

```
BLEU Score for English to Chinese Translation: 0.3210862217004309
```

5. Conclusion

This project successfully demonstrated the use of the MBART model for English-to-Chinese translation. By leveraging a pre-trained model, we were able to perform translations without requiring any fine-tuning. The results were evaluated using the BLEU score, which provided a quantifiable measure of translation quality.

Key Findings

- MBART is highly effective for multilingual translation tasks and supports a wide variety of languages.
- The model-generated translations were generally accurate, though there were occasional discrepancies compared to the reference translations.
- The BLEU score of **0.32** suggests that the MBART model performed moderately well in translating English to Chinese, but there is room for improvement, particularly in cases involving nuanced or idiomatic expressions.

Challenges

- The dataset used for evaluation was relatively small, and expanding it could yield more comprehensive insights.
- Minor inconsistencies in the translations indicate that model fine-tuning could improve translation quality.

REFERENCE

Hugging Face. (n.d.). Multi-lingual Translation Instruct-langsplit. https://huggingface.co/datasets/sroecker/Multi-lingual_Translation_Instruct-langsplit

Hugging Face. (n.d.). Marian multilingual models. https://huggingface.co/docs/transformers/en/model_doc/marian#multilingual-models

Hugging Face. (n.d.). facebook/mbart-large-50-many-to-many-mmt. https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt