Fine-tune LLM for Language Translation

Fangyu Jiang Matric No.: 429844
RPTU Kaiserslautern, Department of Computer Science

Note: *This report contains a project documentation and reflection on the portfolio task submitted for the lecture Engineering with Generative AI in WiSe 2024-25. This report is an original work and will be scrutinised for plagiarism and potential LLM use.*

1 Portfolio documentation

Compile a comprehensive documentation of your project, including all the project phases. You will need to explain every choice you made during the project and your thoughts about the results you get. You will introduce the results in suitable visualisation. Furthermore, you will need to explain which criteria you follow to build your prompts and how they affect the results.

Students write the entire documentation with sections, sub-sections, diagrams, etc in this section. Please write as comprehensively as possible. Head to the document 1_documentation.tex. You are free to use as many subsections as required. We will not provide a template for documentation.

1.1 Research Phase

In the research phase, there are two selections I'm going to make: the dataset selection and the model selection.

With regards to the dataset of this language translation task, the basic requirement is that it should come with at least 1,000 pairs of German-French translations. So the selected dataset should include german and french languages, at least 1,000 rows and preferably be used for translation tasks. I applied these 3 filters to huggingface datasets and it gave me 88 datasets. I went through some of them and tried to get to know 1)how the data format looks like; 2)what the dataset originally is used for; 3)how the translation quality is.

First, the dataset must be a translation between German and French. The google/wmt24pp dataset is a translation between German and English, and French and English, so I excluded this type of data and focused on finding a dataset of translation between German and French. This can filter out many datasets, and finally I focused on the Helsinki-NLP/opus-100 dataset, the Helsinki-NLP/opus_books dataset, and the Helsinki-NLP/europarl dataset.

Second, this task does not limit translations to specific fields. The Helsinki-NLP/opus-100 dataset is an English-centered multilingual corpus. All training pairs contain English on the source or target side, and the official does not specify a specific field. The Helsinki-NLP/opus_books dataset is taken from a collection of copyright free books aligned by Andras Farkas. The Helsinki-NLP/europarl dataset comes from the European Parliament, and the content mainly involves law and politics. Because the task is not limited, it makes me a little entangled. So, I decided to refer to the third point and randomly check the quality of the data.

Third, after comparison, I found that the Helsinki-NLP/opus_books dataset contains strange semantic translations, such as one data "de": "Den ich wandern muß, arms Waisenkind! Weshalb

sandten sie mich so weit, so weit," "fr": "«Pourquoi m'ont-ils envoyé si seul et si loin, là où s'étendent les marécages, là où sont amoncelés les sombres rochers?" Even if I used Google Translate, I couldn't understand it, which would inevitably affect fine-tuning, and the dataset is biased towards the storyline, which I don't think is suitable for this fine-tuning, so I excluded the Helsinki-NLP/opus_books dataset. Most of the data in Helsinki-NLP/opus-100 are short, while the data in Helsinki-NLP/europarl are mostly long sentences. Considering the performance of Colab T4 GPU, the short data in Helsinki-NLP/opus-100 is more suitable. In addition, the content of the Helsinki-NLP/europarl dataset focuses on parliament and is more inclined towards politics and law. Helsinki-NLP/opus-100 is more suitable for the translation of general content. So I finally decided to use Helsinki-NLP/opus-100.

1.2 Design Phase

With respective to the approach of fine-tuning, I chose the

With respective to the split ratio between the training dataset and testing dataset, I chose

1.3 Implementation Phase

https://stackoverflow.com/questions/69609401/suppress-huggingface-logging-warning-setting-pad-token-id-to-eos-token-id

2 Reflection

In 3-5 pages, 1500-2000 words

The purpose of the reflection is to show that you can reflect and assess your own work and learning process critically.

This section needs to be adjusted to align with the reflection requirements specified in the selected task.

Note: You should address all the questions from your selected task. Please list each question and provide your answers in the following enumeration.

For example:

1. What was the most interesting thing that you learnt while working on the portfolio? What aspects did you find interesting or surprising?

   Answer: This is my answer...

2. Which part of the portfolio are you (most) proud of? Why? What were the challenges you faced, and how did you overcome them?

   Answer: This is my answer...

3. What adjustments to your design and implementation were necessary during the implementation phase? What would you change or do differently if you had to do the portfolio task a second time? What would be potential areas for future improvement?

   Answer: This is my answer...

4. Include a brief section on ethical considerations when using these models on language translation tasks.

   Answer: See Section 3 on page 3.

5. From the lecture/course including guest lectures, what topic excited you the most? Why? What would you like to learn more about and why?

   Answer: This is my answer...

6. How did you find working with DIFY platform during the course work? Would you recommend using DIFY in learning Generative AI technologies and why? What is the best start for learning Generative AI either by Python code or No-code platforms and why?

   Answer: This is my answer...

7. How did you find the assignments and exercises in the course and how they help you in portfolio exam?

   Answer: This is my answer...

3 Ethical Considerations

A brief section on ethical considerations when using these models on language translation tasks. Readings:

- ACL Ethics Policy (2023)
- EU AI Act (Article 10 on Translation Systems)
- ISO/IEC 24028:2020 (AI Trustworthiness)

All of the resources used by the student to complete the portfolio task should be organised in the references section. **Note that the Reference section does not count towards the number of pages of the report.** Example references are given below [1] [2] [3]. **If you are using a reference manager like Zotero, you can export your Zotero library as a .bib file and use it on Overleaf. As you cite the article/technology/library in your main text, the References section will automatically update accordingly.** Please include a full list of references found. If students are using Zotero for their research paper management, a bibTeX will help them during citation which automatically adds references to the report.

References

[1] Albert Einstein. Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]. *Annalen der Physik*, 322(10):891–921, 1905.

[2] Donald Knuth. Knuth: Computers and typesetting.

[3] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.