Parallel Corpora Synthesis and Analysis with Large Language Models

Ivan Loginov
University of Colorado
Boulder, USA
ivan.loginov@colorado.edu

ABSTRACT

This study explores the use of Large Language Models (LLMs) to generate and fine-tune parallel corpora, focusing on translation performance. Leveraging Mistral-7B for English text generation and T5-base for translation to German, the research compares three models — T5-base, T5-base fine-tuned on Opus-books collection, and T5-base fine-tuned on synthetic data — on the Flores dataset using SacreBLEU scores.

The findings reveal that while T5-base Opus-books initially performed better, it ultimately achieved a lower SacreBLEU score than the original T5-base. On the other hand, T5-base Synthetic consistently outperformed both models during training, showcasing the promising role of synthetic data in advancing LLM training practices.

1 INTRODUCTION

Current Large Language Models (LLMs) have exhibited impressive performance in machine translation, as noted by Hendy et al. [3]. Their versatility, spanning tasks like text generation, summarization, and question answering, positions them as superior candidates for multilingual conversations compared to Neural Machine Translation (NMT) models.

Despite the advancements introduced by LLMs, their ability to translate non-English languages is not yet optimal[12]. This limitation is largely attributed to the scarcity and constraints of real-world multilingual data[9], posing significant hurdles in training models for cross-language scenarios.

This project aims to conduct a comparative study between real and synthetic text collections to test the hypothesis that an LLM fine-tuned on synthetic data can yield comparable results to one fine-tuned on real-world data.

2 RELATED WORK

This study draws insights from recent research papers in the field of:

2.1 Multilingual Data Synthesis

[9] explores the use of LLMs for synthesizing training data across languages, with a focus on applications in multilingual dense retrieval.

2.2 Multilingual Machine Translation with LLMs

In [11, 12], the authors present empirical results and analysis of multilingual machine translation using LLMs, providing insights into their effectiveness across diverse languages.

This project differs from the aforementioned studies by generating text data with direct and not summarise-then-ask prompting without additional human validation and content filtering[9]. The focus is on comparing SacreBLEU [6] scores between a fine-tuned LLM on both real-world and synthetic data, deviating from the base open-source LLMs comparison [12].

3 PROPOSED WORK

To assess the equivalence of synthetic data and real-world data in fine-tuning LLMs, the following steps will be undertaken:

(1) Acquire two distinct real-world datasets: one for benchmarking translation quality and another for the actual fine-tuning of the model.
(2) Choose a high-performance open-source LLM with fast inference capabilities to generate coherent English sentences.
(3) Generate and store a set of language pair samples, matching the quantity of samples present in the real-world dataset designated for fine-tuning.
(4) Preprocess the generated data to optimize its suitability for subsequent model training.
(5) Fine-tune a lightweight LLM on both real-world and synthetic data independently.
(6) Evaluate three different models: a base model, a model fine-tuned on real-world data, and a model fine-tuned on synthetic data.

3.1 Real-World Data Collection

Our study leverages the recently updated Flores dataset [1, 2, 8], a resource utilized in the benchmarking of LLM translation capabilities in the work by Zhu et al. [12].

For further fine-tuning of the LLM, we incorporate the base Opus dataset [10], comprising diverse sentence pairs from various languages.

For a focused and meaningful comparison, we specifically choose English-German sentence pairs from these datasets, aligning with the language pairs used in the synthetic data generation process.

3.2 LLM Selection

For English sentence generation, our choice is the Mistral-7B open-source model, selected for its superior performance over other fast-inference models across various benchmarks [4].

For translating into German, we opted for the T5-base transformer model [7] owing to its generative and efficient inference capabilities.

3.3 Parallel Corpora Synthesis

To generate a set of translated English-German sentence pairs, we utilized Mistral-7B and T5-base models. Mistral-7B was prompted as follows:

> PROMPT: Create unique and diverse sentences covering a wide range of topics, emotions, and styles.

Subsequently, T5-base took on the task of translating these sentences using the prefix:

> PREFIX: translate English to German: [generated sentence].

Following additional preprocessing, this approach resulted in approximately 30.2 thousand samples. The synthetic data was produced using a computing setup with a 4-core CPU, 30 GB RAM, and two T4 GPUs, each equipped with 16 GB of memory. The code for generating synthetic parallel corpora is available in the GitHub repository[1].

3.4 Data Preprocessing and Warehousing

To refine our dataset, we filter out sentences with two or fewer words and eliminate redundant characters introduced by Mistral-7B. The text generated by Mistral-7B is split into individual sentences, translated using T5-base, and formed into language pair sentences. The resulting processed data is stored in the Apache Parquet file format, enhancing subsequent read operations. For convenient access, the finalized dataset is made available on the Hugging Face platform[2] .

3.5 LLM Fine-Tuning

To train the T5-base model on Opus-books and Synthetic datasets, we aligned the synthetic data with the randomly selected segment of 30.2 thousand samples from the real-world Opus dataset. This alignment ensured a meaningful basis for comparing results. Training in both cases lasted 1.5 hours on the same setup used for generating synthetic English-German pairs. The code for fine-tuning the T5 model is accessible in the GitHub repository[3].

4 EVALUATION

To evaluate the fine-tuned T5-base models, we employed the SacreBLEU metric [6] to measure translation quality, replacing BLEU [5]. BLEU's limitation lies in its requirement for pre-tokenized text, which complicates comparisons across models with different tokenizers. SacreBLEU addresses this issue by standardizing the tokenization process, ensuring more reliable comparisons [6]. We assessed three models: the base T5-base model, the T5-base fine-tuned on the Opus-books dataset (T5-base Opus-books), and the T5-base fine-tuned on synthetic data (T5-base Synthetic).

We conducted the final evaluation of English-to-German translation using the Flores benchmark [1, 2, 8], as also employed in the study by Zhu et al. [12]. Our benchmark comprises 2,009 English-German sentence pairs for this assessment.

[1]Generate data with Mistral-7B and T5-base at https://github.com/jaymanvirk/synthetic_parallel_corpora

[2]Synthetic parallel corpora, English-German (EN-DE) at https://huggingface.co/datasets/jaymanvirk/synthetic_parallel_corpora

[3]Fine-tuning the T5-base at https://github.com/jaymanvirk/synthetic_parallel_corpora

4.1 Fine-tuning T5-base Model

The training progression of the T5-base model, as illustrated in Figures 1a and 1b, reveals a noteworthy pattern. Initially, the training loss experiences a significant drop after the first epoch, subsequently stabilizing for the remainder of the 5-epoch training process. In contrast, the test loss remains relatively constant throughout the entire training duration.

Notably, when fine-tuning on synthetic data, the test loss is consistently lower, measuring at 0.3 compared to 0.6 when fine-tuned on the Opus dataset. This discrepancy suggests that the T5-base model fine-tuned on synthetic data demonstrates improved generalization of translation mechanics.

This observation aligns with the SacreBLEU scores depicted in Figure 1c. The T5-base model fine-tuned on synthetic data consistently achieves higher SacreBLEU scores across every epoch compared to its counterpart fine-tuned on the Opus-books dataset. However, it is essential to highlight that both scenarios exhibit a decline in performance over the training course, with the Opus fine-tuned model showing a more pronounced decrease, resulting in a lower score than the original T5-base model.

4.2 Comparison of T5 Model Performances on Flores English-German Dataset

In Table 1, we present the SacreBLEU scores for three T5 models:

- T5-base
- T5-base Opus-books
- T5-base Synthetic

The initial assessment of the T5-base model on the Flores English-German dataset revealed a baseline SacreBLEU score of 31.85, showcasing the model's inherent translation capabilities between English and German.

Interestingly, fine-tuning T5-base on the Opus dataset resulted in a slightly lower score of 30.5 compared to the baseline. This suggests that the model might be memorizing Opus data rather than enhancing its generalization abilities for translation tasks.

In contrast, fine-tuning the T5-base model on synthetic data led to the highest SacreBLEU score among the three models, reaching 34.67. This indicates that training on synthetic data not only matches the performance of models fine-tuned on real-world data but even surpasses it, showcasing the potential benefits of synthetic data in training.

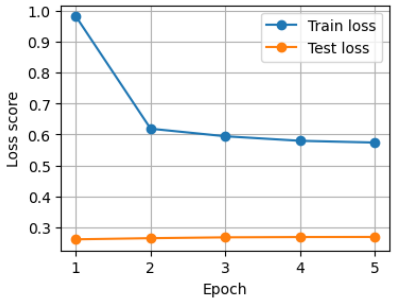
Model	SacreBLEU
T5-base Synthetic	34.674710
T5-base	31.855084
T5-base Opus-books	30.545728

Table 1: SacreBLEU scores for T5 models.

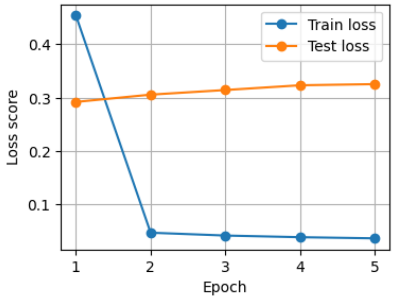
5 DISCUSSION

5.1 Parallel Corpora Synthesis

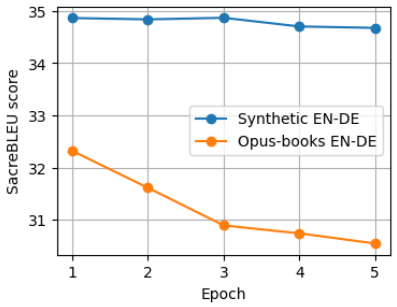
The primary challenge in creating English-German pairs was the inference speed of Mistral-7B and T5-base. With the aforementioned computing setup, the optimal outcome was approximately 1,000



(a) Train and test loss scores of T5-base fine-tuned on Opus-books EN-DE dataset



(b) Train and test loss scores of T5-base fine-tuned on Synthetic EN-DE dataset



(c) SacreBLEU scores of T5-base fine-tuned on Opus-books EN-DE and Synthetic EN-DE datasets

Figure 1: Comparison of T5-base Model Performances on English-German Translation Tasks.

language pairs per hour. This suggests potential for improvement, as the code in this study didn't include additional acceleration methods found in the Hugging Face library[4].

5.2 Parallel Corpora Analysis

The training didn't encounter challenges, except for fine-tuning model hyperparameters to achieve the best balance between speed and quality, which wasn't unexpected. However, the results presented an intriguing scenario. In both scenarios—training on Opus-books and Synthetic data—the fine-tuned models exhibited an increase in SacreBLEU score in the initial epoch, followed by a consistent decline. Notably, the T5-base Opus-books even scored lower than the baseline after completing training.

These findings indicate a need for adjustments in the training process. One potential solution is to explore additional data augmentation or introduce extra dropout techniques to enhance model regularization. This could help maintain and improve the performance observed in the initial stages of training.

5.3 Extended Research

Further research can concentrate on:

- Creating sentence pairs in other languages
- Training other LLMs for performance comparison
- Utilizing additional benchmark datasets

6 CONCLUSION

This study introduces a project focused on leveraging LLMs to generate parallel corpora and analyze model performance fine-tuned on the synthetic data. The research employs Mistral-7B to generate a English text collection and T5-base to translate the generated corpora.

Our main focus is comparing translation performance on the Flores dataset across three models: T5-base, T5-base fine-tuned on the real-world Opus dataset (T5-base Opus-books), and T5-base fine-tuned on synthetic data (T5-base Synthetic), using the SacreBLEU score.

The original T5-base model performed well on the Flores English-German pairs, scoring 31.85 in SacreBLEU. However, when fine-tuning on Opus-books, the T5-base Opus-books initially scored higher in the first epoch but ended up with a lower SacreBLEU of 30.54 after completing training. This drop could be due to the model memorizing new data, leading to less effective word selection for translation.

On the other hand, the T5-base Synthetic exhibited a similar pattern with a high score of 34.86 in the first epoch, followed by a decline in subsequent epochs. Despite this, across all five epochs, T5-base Synthetic consistently outperformed both the original T5-base and T5-base Opus-books, finishing with a SacreBLEU of 34.67. This indicates that refining T5-base with synthetic data not only improved its translation skills compared to the original T5-base but also outperformed the T5-base fine-tuned on the real-world Opus-books dataset. It highlights synthetic data as a promising choice for future LLM training.

For future research, one could explore data augmentation or incorporate extra dropout techniques to improve model regularization and address the issue of SacreBLEU decline during training. Additionally, further steps in synthetic data analysis might involve creating sentence pairs in different languages, training other LLMs for performance comparison, and using additional benchmark datasets to broaden the scope of the study.

[4]Hugging Face Accelerate Library at https://huggingface.co/docs/accelerate/index

7 CHANGELOG

- Charts illustrating T5 models' training loss and SacreBLEU scores loss per epoch
- SacreBLEU comparison table featuring three T5 models
- Further elaboration on the methodology with additional details

REFERENCES

[1] Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F., and Fan, A. The flores-101 evaluation benchmark for low-resource and multilingual machine translation, 2021.

[2] Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english, 2019.

[3] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. How good are gpt models at machine translation? a comprehensive evaluation, 2023.

[4] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

[5] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (Philadelphia, Pennsylvania, USA, July 2002), P. Isabelle, E. Charniak, and D. Lin, Eds., Association for Computational Linguistics, pp. 311–318.

[6] Post, M. A call for clarity in reporting bleu scores, 2018.

[7] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[8] Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. No language left behind: Scaling human-centered machine translation, 2022.

[9] Thakur, N., Ni, J., Ábrego, G. H., Wieting, J., Lin, J., and Cer, D. Leveraging llms for synthesizing training data across many languages in multilingual dense retrieval, 2023.

[10] Tiedemann, J. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (Istanbul, Turkey, May 2012), N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds., European Language Resources Association (ELRA), pp. 2214–2218.

[11] Zhang, B., Haddow, B., and Birch, A. Prompting large language model for machine translation: A case study, 2023.

[12] Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., and Li, L. Multilingual machine translation with large language models: Empirical results and analysis, 2023.