Recommendation for Expanding the Few-Shot Region-Aware Machine Translation Benchmark (FRMT) via a Temporal Dimensioion

By: Eric Danforth
W266 - Section 001

Abstract:

While human evaluation continues to be the best method for determining the quality of neural machine translation, there has been improvement in both automatic evaluation and model performance. The rise in popularity of large language models that are suited to a broad range of tasks, including machine translation, is only expected to continue for the foreseeable future; however, this is one area that they continue to struggle. Benchmarks are import tools that can accelerate the progress towards improvement of translation quality. In this project, I show the shortcomings of state of the art large language models to appropriately locate translated text in the correct historical timeframe and region as well as the ease of fine tuning T5 models to complete this task. And I make a recommendation to expand a recently released benchmark to include temporality and point to the improvements that may result.

Background:

The FRMT benchmark (Riley et al., 2022) has recently been introduced by Google Research to enable fine-grained comparison across regions for two dialects each of Portuguese and Mandarin Chinese. Its goal is to evaluate the ability of machine translation models to perform few-shot region control, meaning the ability to adapt to specific regional dialects and styles of language with limited training data. To my knowledge it the first example of a benchmark to focus on intra-language differences. It's an important step in the right direction to encourage the introduction of variations in language, but I believe that temporality will play an increasingly important role in the understanding of historical documents and improving the training data for large languages models.

Optical character recognition, or OCR, plays an important role in gathering data from the pre-digital era. Results can vary, especially with handwritten text, but all of the examples I will use are from print materials. The quality of the output text also depends on the input image quality, but here again I will be using documents where the majority appear to be professionally scanned for archival purposes. And finally, while English may have the most support for OCR products, I will use model weights that are specifically adapted to Portuguese, although not specifically to either region or temporality.

Project:

In this project I will show the need for expanding the focus of the FRMT benchmark to include a temporal dimension. Specifically, I focused on English to European Portuguese and English to Brazilian Portuguese language pairs, using examples gathered from scanned books. Due to the potential scale of this endeavor I chose to limit my research to the decade from 1900-1909. This timeframe is convenient because it is in the public domain and it is before the first modern orthograpic reforms of Portuguese in Portugal in 1911 (*"Reforms of Portuguese Orthography"*). An expansion of this project to more recent decades will be particularly interesting due to the restrictions on the press in Brazil during the last century, the rise in literacy over that same time period, as well as continuing orthographic changes in the two major regional variants of the Portuguese language. There is also a rich body of literature dating back centuries for further exploration prior to the modern era.

Motivation:

The temporal dimension can help address a significant challenge in machine translation - the translation of historical documents and literature. This type of writing contains vocabulary, grammar, and orthography that may not be present in modern language, making it difficult for machine translation models to accurately translate them. By creating a benchmark that includes examples from different time periods, we can encourage research in this area and help develop more effective machine translation models that can handle these challenges. This can ultimately lead to better parallel corpora and improved translation quality in future models.

Data:

My primary sources were scanned books from the national libraries of Portugal and Brazil - the Biblioteca Nacional de Portugal and Fundação Biblioteca Nacional. To supplement the Brazilian publications I used the digital library of São Paulo - Biblioteca de São Paulo. I did note that there was a wealth of resources available via the Hathi Trust - many of them scanned from Google's Library Project, but their terms of service prohibit automated downloading and the process to gain access to data for research was prohibitively complex for the scope of this project. The Portuguese and Brazilian books were accessed in PDF form and underwent preprocessing that split each page into a separate image. I experimented in Photoshop with color correction curves that would work well with the yellowed pages of the scanned texts as well as help to remove inclusion of artifacts from bleed through on the opposite side of each page and applied them in python followed by a reduction of red and yellow color channels, grayscale conversion, application of median blur, and segmentation via Otsu's thresholding. I OCRed the preprocessed images with Google's tesseract using the model optimized for Portuguese, which I found to be the most reliable after trying multiple methods.

T5 Model Fine-Tuning:

Starting with my OCRed text, I pre-processed to split into individual phrases which were then paired with their associated English translations via the NLP-Helsinki model for Romance languages to English translation based on OPUS-MT (Tildeman et al. 2020) and using Marian (Junczys-Dowmunt et al. 2018) and hosted on Hugging Face. I chose this method because of the prohibitive high cost of the Google Translate API after testing the results and judging them from my own experience in Portuguese to be very high quality. It's worth mentioning that while Portuguese is one of the 24 official languages of the European Union there is currently only one other open source model on Hugging Face (Lopes et al. 2020) which I didn't find suitable for the task, perhaps because their training data was more recent and limited to Brazilian Portuguese.

Testing:

I've tested the few shot capabilites using the ChatGPT 3.5 API and my fine tuned models. Evaluation metrics will be based on BLEU (Papineni et al., 2002) and BLEURT (Sellam et al. 2020).

European Portguese:

Average ChatGPT BLEU Score: 0.0825
Average ChatGPT BLEURT Score: 0.5959
Average Fine-Tuned BLEU Score: 0.3174
Average Fine-Tuned BLEURT Score: 0.7574

Brazilian Portuguese:

> Average ChatGPT BLEU Score: 0.0891
> Average ChatGPT BLEURT Score: 0.6120
> Average Fine-Tuned BLEU Score: 0.4669
> Average Fine-Tuned BLEURT Score: 0.8020

Here is an example that illustrates the challenges with the OCR process, the high performance of the fine-tuned models and the shortcomings of ChatGPT. This was for European Portuguese and ChatGPT was prompted to localize in the correct decade and given three examples of orthographic differences in the prompt:

> Source: In this unique case, I was able to recognize the specific difference of two nearby types, belonging to two geological epochs
>
> Target: Neste caso unico, pude reconhecer a diflernça especifica de dous typos proximos, pertencentes àás duas epochas geologicas
>
> ChatGPT 3.5: Neste caso único, fui capaz de reconhecer a diferença específica de dois tipos próximos, pertencentes a duas épocas geológicas
>
> Fine-Tuned: Neste caso unico, pude reconhecer a differença especifica de dous typos proximos, pertencentes a duas épochas geologicas

ChatGPT failed four times to provide the correct orthographic variation for the decade in question. While tesseract incorrectly OCRed two of the words in the target phrase during preprocessing steps. And there are several words that are missing accent marks in the fine-tuned model, indicative that they were missed in OCR and are biasing the model to not include them.

Future exploration:

I believe that there can be a positive feedback loop between period and regional appropriate machine translation and optical character recognition. And high quality OCR is important for developing a semantic understanding of the text which will improve the quality and help reduce the size of large language models. This year alone over 60,000 books entered the public domain, and there are a wealth of scanned books going back centuries which are much more suited for use as authoritative text for training data than bulk collections like Common Crawl. But the current challenge is that OCR is best adapted to the modern era where training data has been freely available. Since it will take decades for works entering the public domain to catch up to the internet era, I think this will be a vital area of research. And the usefulness of large language models trained on better data may pave the way to open research to current copyright protected works which has been embroiled in litigation since the first mass-digitizations efforts began. As size and power consumption is reduced and training and fine-tuning becomes more accessible we should see a proliferation of models that can use these period appropriate techniques. One such application in eduction could be training models to a certain time period and fine-tuning them with the personalities of historical figures. It's important to make these resources available in other languages as well and the exploration of regional dialects opens the door to further fine grained divisions such as are often associated with administrative divisions below the national level.

References:

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant FRMT: A Benchmark for Few-Shot Region-Aware Machine Translation. *arXiv preprint arXiv:2210.00193.*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.*

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7881–7892, Online. Association for Computational Linguistics.*

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *arXiv preprint arXiv:2202.11822.*

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. *In Proceedings of ACL 2018, System Demonstrations, pages 116–121.*

Tiedemann, Jörg and Santhosh Thottingal (2020). “OPUS-MT – Building open translation services for the World”. In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT).* Lisboa, Portugal: European Association for Machine Translation, pp. 479–480. URL: https://helda.helsinki.fi/bitstream/handle/10138/327852/2020.eamt_1_499.pdf.

Alexandre Lopes, Rodrigo Nogueira, Roberto Lotufo, and Helio Pedrini. 2020. Lite Training Strategies for Portuguese-English and English-Portuguese Translation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers.*

Reforms of Portuguese orthography (2022) *Wikipedia*. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/Reforms_of_Portuguese_orthography (Accessed: April 16, 2023).