Abstractive Summarization of Portuguese Texts by fine-tuning the Portuguese-Based T5 Model

Caio Villela

July 2021

Abstract

The following project aims to fine-tune and implement the Portuguese Vocabulary Pretrained T5 model [1] on abstractive text summarization tasks in Brazilian Portuguese.

The model was fine-tuned on the XSum Dataset [5], which is English based, in accordance to recent and promising results on Zero-Shot Cross Lingual Summarization [2]. The main metric used to evaluate the model performance was the ROUGE-score [4], over which the final model did not perform as expected.

1 Introduction

With the continuous growth of text corpus and non-structured written data available on the internet, as well as the everlasting demand of pin-pointing and extracting important information from texts throughout history, text summarization has become one of the most sought after Natural Language Processing (NLP) techniques in recent years, in parallel with the evolution of Sequence to Sequence Transformers' performance.

There are two main summarization methods explored throughout the development of NLP models:

- Extractive summarization: identify important sections of the text and generate verbatim producing a subset of the sentences from the original text.
- Abstractive summarization: reproduce important material in a new way after interpretation and examination of the text using advanced natural language techniques to generate a new shorter text that conveys the most critical information from the original one.

In this work we will be focusing on the implementation of abstractive summarization.

The T5 Transformer [6] has been one of the most praised models of its kind, achieving State-of-The-Art results on multiple sequence to sequence tasks, including a ROUGE-1 Score of 43.52 on the CNN/DailyMail Dataset [7] for abstractive text summarization (ASSUM), the best of its kind when it was released. Nevertheless, such performance has never been measured on Portuguese based models, as most of NLP models are English oriented.

While NLP applications in Portuguese mostly tend to rely on multilingual models, their performance is below of what is expected for ASSUM tasks to be implemented on real-world scenarios. The lack of appropriate Portuguese datasets for most NLP tasks is also a recurring and difficult to overcome problem.

In this project we will be fine-tuning the already pretrained PT-T5 model on the XSum dataset in hope its performance is slightly worse, or equal to, the original model's performance on English tasks. After more objective metrics are calculated, we will test the summarization on Portuguese texts of interest, more specifically on informative newsletters made available by Brazil's Health Ministry.

2 Data set

The T5 model has been successful performing multiple NLP tasks, and has been pretrained to do so in the following fashion:

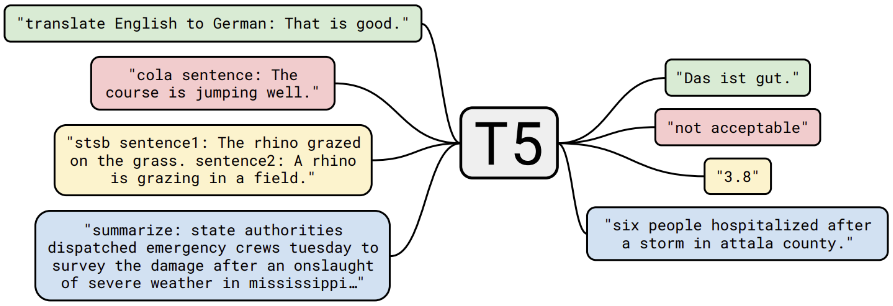
"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 1: T5 prefixes for different tasks.

The XSum (Extreme Summarization) dataset, which has approximately 240k text pairs, was downloaded via the Huggingface API/Transformers Library [8] and preprocessed accordingly. It is composed of news articles and their respective one-sentence summaries.

The ”summarize: ” prefix was added to the beggining of the documents, and the tokenized sentences were padded or truncated to a maximum length of 512 tokens.

3 Methodology

Given the convenience and vast documentation on fine-tuning such models through the Trainer class, from the Transformers Library, it was the first attempted approach for this project.

Even though the official Transformers documentation provides us with multiple examples on fine-tuning the T5 model for summarization, the results obtained were very far from the expected results displayed on the T5 original paper. Even after finetuning the Trainer class instance according to the official forum discussion on the subject, the obtained results were far from expected.

Later a full implementation using the Pytorch Lightning Library [3] was attempted, based on this script, and it was much more successful at the task at hand.

In order to prove that the fine-tuner was working properly without having to go through the full and extensive dataset, it was proposed to train the model on a few text pairs, as to overfit it on the given examples. After that was successfully achieved, the training on the full dataset took place.

4 Experiments

At first, in order to validate the trainer performance, the original T5-base model was finetuned in accordance to the following hyperparameters:

Table 1: Hyperparameters used on the fine-tuning of the model.

Hyperparameter	**Value**
Optimizer	AdamW
Learning Rate	3e-4
Weight Decay	1e-5
AdamW epsilon	1e-8
Batch Size	4
Gradient Accumulation Steps	64
Eval Accumulation Steps	64

As previously mentioned, the first attempts on training the model were made by using the Trainer class and Huggingface official examples. The following figures represent the ROUGE1, ROUGE2 and ROUGEL metric evolution for different runs on said trainer, with minor hyperparameter tweaks in between them.

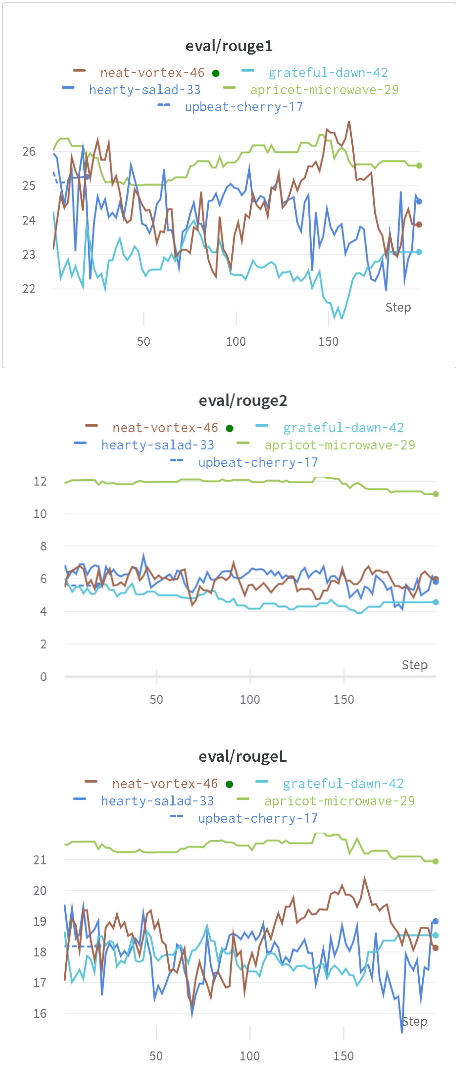

Figure 2: ROUGE scores for different runs, when tuning with Trainer class.

It is noticeable that there was no improvement on Rouge evaluation throughout the training of the previous model, and that the results (25.5 Rouge1 score for the best run), were far from expected.

The second implementation, however, managed to obtain much more interesting results. It is a Pytorch-Lightning based model, which was based on this script. The output generation was more finely controlled, as expected, and the model managed to score closely to the state-of-the-art results presented on Google's T5. The hyperparameters used were the same as before, except for the implementation of a linear scheduler with warmup for the AdamW optimizer.

After the first implementation using the t5-base checkpoint was successful, the Portuguese Vocabulary Pretrained T5 model was fine-tuned in the same fashion. Below, we can see both model's progress throughout the training process. The brown curve represents the **T5-base** run, where as the grey curve represents the **PT-T5** evolution.

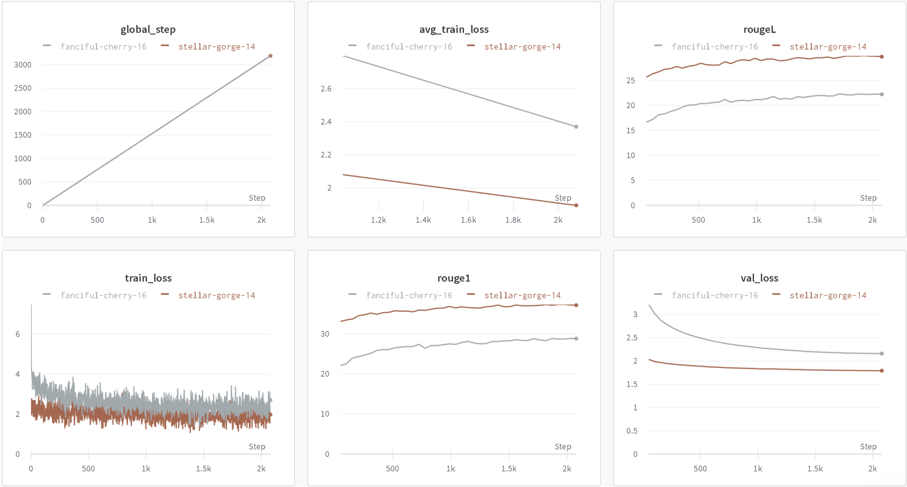

Figure 3: Training losses and evaluation metrics for the PytorchLightning fine-tune approach.

As expected, the English model has an overall better performance on the evaluative metrics, reaching a 37.16 Rouge 1 score on the XSum Dataset. On the other hand, the Portuguese-based model performed better than the T5-base model with no fine tuning on this dataset. It can be observed that both the validation and train losses for the PTT5-base model were decaying much more rapidly than T5-base's throughout the training process, which makes room for questioning if the model performance could be improved even more with more training epochs.

5 Results

Besides the more straight-forward scored metric, text summarization must take into account human-inferred conclusions over the produced summaries. After a few experiments, it was evident that the English-based model was able to produce concise and very closely related to the benchmark summaries.

On the other hand, the Portuguese based fine tuned summarizer produced decent summaries in the English language, similar to the ones produced by the fine-tuned base model. Nevertheless, the model lacked coherence, producing many times what is described as ”delusional” summaries, that gravely distort the original information on the analysed document.

As to the main concern for this project, the quality of summaries in the Portuguese language had interesting results. The model seems to overlap English words on Portuguese summaries, nevertheless it is capable of interpreting the main concepts, and mostly adds directly translated English words on the summarized document.

6 Conclusion

The fine-tune of Portuguese based models for text summarization on English datasets was not successful, but it is proven that the model could perform fairly well on English tasks and create cohesive summaries in Portuguese, if not for the English terms mixed in between them.

7 Future Work

This project could be improved by exploring the Teacher-Student mechanism discussed in [2]. Another possibility for evolving the project would be to create and document a Portuguese based summarization dataset, in order to train the models on the language.

References

[1] Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*, 2020.

[2] Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy, July 2019. Association for Computational Linguistics.

[3] William Falcon et al. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 2019.

[4] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. page 10, 01 2004.

[5] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the

limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[7] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.