Transformer Models for Brazilian Portuguese Question Generation: An Experimental Study

Júlia da Rocha Junqueira, Ulisses B. Corrêa, Larissa A. Freitas
Center for Technological Advancement (CDTec),
Federal University of Pelotas (UFPel)
96010-610 - Brazil
{julia.rjunqueira, ulisses, larissa}@inf.ufpel.edu.br

Abstract

Unlike tasks such as translation or summarization, generating meaningful questions necessitates a profound understanding of context, semantics, and syntax. This complexity arises from the need to not only comprehend the given text comprehensively but also infer information gaps, identify relevant entities, and construct syntactically and semantically correct interrogative sentences. We address this challenge by proposing an experimental fine-tuning approach for encoder-decoder models (T5, FLAN-T5, and BART-PT) tailored explicitly for Brazilian Portuguese question generation. Our study involves fine-tuning these models on the SQUAD-v1.1 dataset and subsequent evaluation, also on SQUAD-v1.1. Through our experimental endeavors, BART returned a higher result in all the ROUGE metrics, as ROUGE-1 0.46, ROUGE-2 0.24, and ROUGE-L 0.43, suggesting a higher lexical similarity in the questions generated, and it is comparable to the results of the question generation task for the English language. We explored how these advancements can significantly enhance the precision and quality of the question generation task in Brazilian Portuguese, bridging the gap between training data and the intricacies of interrogative sentence construction.

Introduction

In recent years, Natural Language Processing (NLP) technologies have undergone remarkable advancements, driven mainly by the advent of Transformer models. The ability of these models to process and understand language in a more human-like manner has opened up new possibilities and challenges in NLP. Characterized by their parallel multi-head attention mechanisms, Transformers have revolutionized tasks such as translation, summarization, and question-answering (LeCun, Bengio, and Hinton, 2015).

In NLP, the Question Generation (QG) task involves creating meaningful and contextually relevant questions from a given set of textual information (Rus et al., 2010). It requires a deep understanding of language structures and semantics by the model since it needs to know *what to ask* and *how to ask* (Pan et al., 2019).

Generating meaningful questions requires a deeper level of understanding, unlike other tasks such as translation or summarization, where the output is derived from the input with a specific transformation. The model must not only comprehend the given text thoroughly but also infer the information gaps, identify relevant entities, and construct syntactically and semantically correct interrogative sentences (Pan et al., 2019). The complexity of this task is amplified by the diverse contexts in which questions are generated.

One of the primary reasons question generation poses a significant challenge for these Transformers models is the discrepancy between the training data and the target task. Pre-trained language models are typically trained on large corpora of declarative sentences, which constitute the majority of textual data available (Park, Hong, and Park, 2022). Declarative sentences, used to make statements or convey information, differ significantly from interrogative sentences, which are employed to ask questions.

This work proposes an experimental fine-tuning of encoder-decoder models BART[1] (Lewis et al., 2019), T5[2] (Raffel et al., 2020), and FLAN-T5 (Chung et al., 2022) on the question generation task for Brazilian Portuguese. For this, we do the fine-tuning of the models using SQUAD-v1.1 dataset (Rajpurkar et al., 2016), followed by the evaluation of these models, examining how the capabilities of this work can be utilized to enhance the precision and quality of this task. Futhermore, we will be experimenting with the generation of factoid questions, which is a type of question that seeks a specific piece of factual information, for example: *"Who was the first president of the United States?"*.

The structure of the remainder of this work is as follows: **Related Works** examines relevant literature published previously; **Methodology** outlines the procedures we employed to conduct the experiments. This includes details about the datasets we used, the fine-tuning process, and the data flow between different tasks. **Experiments** presents the configuration, metrics evaluated, and hyperparameters used to approach each task; **Final Remarks** summarizes the work and discusses potential future studies.

[1]Bidirectional and Auto-Regressive Transformers
[2]Text-to-Text Transfer Transformer

Related Works

The field of neural ranking and question generation has seen significant advancements in recent years. There is some notable contributions in these areas, such as:

Park, Hong, and Park (2022) proposed a novel post-training method of KoBART [3] to enhance it for Korean question generation. Their method introduced a question-infilling objective to KoBART, augmented training data for question generation, and introduced a Korean spacing objective. They also proposed KorQuAD-QG, a new dataset for Korean Question Generation, to verify the performance of the proposed post-training.

One of the primary challenges in QG is the availability of high-quality Question-Answering (QA) datasets, especially in languages other than English. Lower-resourced languages, such as Portuguese, often lack large-scale QA datasets, making it difficult to explore and experiment with the latest neural techniques in QG.

The approach taken by Leite and Lopes Cardoso (2022) in this study involves framing the QG problem as a sequence-to-sequence task. The researchers fine-tune a pre-trained language model, T5 (Text-To-Text Transfer Transformer), for the specific task of generating factoid or wh-questions in Portuguese. By doing so, they leverage the capabilities of pre-trained language models to adapt them to the QG task.

Despite the potential challenges associated with using a machine-translated dataset for training neural models, the automatic evaluation of the Portuguese neural QG models yields results comparable to those obtained for English. This suggests that the proposed approach by Leite and Lopes Cardoso (2022) is effective in generating questions in Portuguese, even when faced with the limitations of a translated dataset.

These works highlight the potential of fine-tuning and post-training methods in improving the performance and robustness of neural models in various tasks. Our study aims to build upon these findings and explore new avenues for enhancing neural models.

Methodology

In order to meet our primary goal, our work's methodology comprises four main steps. Initially, we selected the models, especially the ones that already were trained in (or could accept) the Brazilian Portuguese. After a bibliographic review, selecting the category of encoder-decoder seq2seq large language models. After this, we fine-tuned the models using the SQUAD v1.1 Brazilian Portuguese (Rajpurkar et al., 2016) data set based on Wikipedia data. Thereon, we tested and evaluated the results using the ROUGE Metric (Lin, 2004).

Dataset (Step 1)

The chosen dataset was SQUADv1.1-PT, created by automatically translating the SQUAD content using the Google Cloud API[4], which contains 87,599 rows of paragraphs for train and 10,570 for validation. The data itself includes columns with titles, contexts, questions, and answers based on Wikipedia articles, where the answer to every question is a segment of text from the corresponding reading context (Rajpurkar et al., 2016). The dataset is based in factoid questions.

As an example:

- **Question:** *A quem a Virgem Maria supostamente apareceu em 1858 em Lourdes, na França? (Who did the Virgin Mary supposedly appear in 1858 in Lourdes, France?)*
- **Answer:** *Saint Bernadette Soubirous*

This dataset was chosen because of the high recurrence of his appearance in researches, so we could analyse better the results, and the data was also broken in columns that were more useful to this work.

Fine-tuning (Step 2)

For the fine-tuning step, we pre-processed the data using the Hugging Face v4.35.0 default input for ConditionalGeneration. This way, the data assumes a new format, being written two new columns in the dataset, *input* and *target*. The input contains a replica of the context and answer columns, and the target has the a content of the question columns. By organizing the data into these two specific columns, we create a structured input format for models.

After that, we set up the train and validation tokenized datasets into data loaders using PyTorch's DataLoader [5] class, with a batch size of 16 and shuffling the data during each epoch, which is a common practice in training deep learning models to introduce randomness and prevent the model from memorizing the order of the training samples.

We used the PyTorch's TrainerAPI [6] for feature-complete training since it offers a streamlined and efficient approach to fine-tuning and evaluating the models for the task. It follows best practices and conventions, making it easier to collaborate and share research findings within our group.

Test (Step 3)

For the experiments, we choose to use the configuration shown in the table bellow, as it returns computational efficiency and a reasonable results. The process of fine-tuning the models is customized to balance computational constraints and the necessity for satisfactory model performance. It is crucial to understand that those limitations influenced the selection of these hyperparameters and may not necessarily reflect the best setup for the problem at hand.

A weight decay of 0.01 was applied, which helps in preventing overfitting by adding a penalty term to the loss function. We also used a mini batch size of 8 samples, with a low number of epochs, set to 2, to balance computational efficiency and model stability. CrossEntropy is set as the loss function, measuring the dissimilarity between predicted and actual class distributions. These configurations are equivalent for all the three experimented techniques.

[3] Available at: https://huggingface.co/hyunwoongko/kobart

[4] Available at: https://cloud.google.com/translate/docs/reference/rest

[5] Available at: https://pytorch.org/docs/stable/data.html\#torch.utils.data.DataLoader

[6] Available at: https://huggingface.co/docs/transformers/main_classes/trainer

Results (Step 4)

To evaluate the results, we used the ROUGE Metric since these metrics compare the reference and predicted questions, providing a high score when there is a considerable similarity in wording and meaning between the predicted and reference questions (Mohammadshahi et al., 2022). As described below:

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{RefSum}} \sum_{gram \in S} \text{Count}_{match}(gram_n)}{\sum_{S \in \text{RefSum}} \sum_{gram \in S} \text{Count}(gram_n)} \quad (1)$$

Where RefSum stands for reference summaries, n stands for the length of the n-gram, $gram_n$, and $\text{Count}_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries (Lin, 2004).

$$\text{ROUGE-L Precision} = \frac{\text{—LCS—}}{\text{Number of unigrams in C}} \quad (2)$$

$$\text{ROUGE-L Recall} = \frac{\text{—LCS—}}{\text{Number of unigrams in R}} \quad (3)$$

$$\text{ROUGE-L F1-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (4)$$

$$\text{ROUGE-Lsum} = \sum_{\text{each sentence}} \text{ROUGE-L F1-score} \quad (5)$$

Where:

- *R* stands for Reference Text
- *C* stands for Candidate (Predicted Text)
- *LCS* stands for Longest Common Subsequence between *R* and *C*.
- *—LCS—* stands for the length of the *LCS*

As for ROUGE-Lsum, it is a variant of ROUGE-L. It calculates the longest matching sequence for each sentence in the reference and candidate text and adds them up. The equations for precision, recall, and F1-score remain the same but are applied at the sentence level and then summed (See, Liu, and Manning, 2017).

However, this evaluation approach can sometimes penalize the returned question even if it is valid because it does not show high lexical or semantic similarity to the reference questions (Mohammadshahi et al., 2022). In this work, we used the ROUGE-1, ROUGE-2 and ROUGE-LSUM metrics, where ROUGE-1 and ROUGE-2 stands for the overlap of uni-grams and bi-grams, respectively, between the candidate summary and reference summary in ROUGE-N.

Experiments

In Table 1, we present the results provided by the experiments executed in BART, T5 and FLAN-T5 (F-T5), comparing them by ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-LSUM (RLSum) metrics.

Before fine-tuning, all the three models returned relatively low scores in all the metrics. FLAN-T5 revealed the worst ROUGE scores between the other two, but showed the best Loss score. The lower loss in FLAN-T5, both in original and fine-tuned models could be associated with the specific architecture and training based on the T5 model, since it occurred before and after our fine-tuning. BART revealed slightly better results when comparing the ROUGE scores, being the best RLSum between the other two models. Still, showed the lowest Loss score of all the experiments.

However, after the fine-tuning process, there is a noticeable improvement in all metrics for all the models, specially RLSum. This indicates that the fine-tuning process has effectively enhanced the performance of the models in the question generation task. It is possible to observe that T5 returned the lowest result based on RLSum, 0.33, but returned a lower Loss than BART. FLAN-T5 had the best results when comparing Loss, returned a RLSum lower than BART too, but better than T5, suggesting that the modifications in FLAN-T5 have led to an improvement in performance.

In Table 2, we present the results of our experiments, organized into columns representing the generated and reference questions, and rows representing the three different models BART, T5 and FLAN-T5 (F-T5). The reference question was randomly picked, to ensure that the sample used for comparison represents the entire population from which it is drawn. This analysis is crucial for understanding whether fine-tuning enhances the quality and relevance of the generated questions.

The context assigned as a input for this question was:

- **Context:** *Super Bowl 50 foi um jogo de futebol americano para determinar o campeão da National Football League (NFL) para a temporada de 2015. O campeão da American Football Conference (AFC), Denver Broncos, derrotou a campeã Carolina Panthers, da National Football Conference (NFC), por 24 a 10, e conquistou seu terceiro título no Super Bowl. O jogo foi disputado em 7 de fevereiro de 2016, no Levi's Stadium, na área da baía de San Francisco, em Santa Clara, Califórnia. Como este foi o 50º Super Bowl, a liga enfatizou o "aniversário de ouro" com várias iniciativas de ouro, bem como a suspensão temporária da tradição de nomear cada jogo do Super Bowl com algarismos romanos (sob os quais o jogo seria conhecido como "Super Bowl L"), para que o logotipo possa destacar os algarismos arábicos 50.*
- **Answer:** *Estádio de Levi*

From the reference question picked, we analyze the generated question corresponding to it. It is possible to observe how the results shown in Table 1 are now more visible, allowing for an assessment of how the models' performance improves or changes after fine-tuning, since it provides a direct reflection of the data in Table 1. As it shows, before

	Model	Loss	R1	R2	RLSum
Baseline	BART	6.39	0.198	0.067	0.16
	T5	3.49	0.025	0.007	0.024
	F-T5	2.83	0.020	0.004	0.018
Fine-tuned	BART	1.26	**0.46**	**0.24**	**0.43**
	T5	1.03	0.34	0.14	0.33
	F-T5	**1.02**	0.36	0.16	0.34

Table 1: Results obtained on the QG task before (Baseline) and after (Fine-tuned) the fine-tuning.

	Model	Generated Question
Baseline	BART	context: Super Bowl 50 foi um jogo de futebol americano para determinar o campeão da
	T5	True
	F-T5	" "
Fine-tuned	BART	Onde estava o jogo do Super Bowl 50?
	T5	Onde está localizado o Levi's Stadium
	F-T5	Onde foi o Super Bowl 50 disputado?

Table 2: Predicted questions generated by the models on the QG task before (Baseline) and after (Fine-tuned) the fine-tuning. The reference equivalent to the question in the dataset is *"Em que local aconteceu o Super Bowl 50?"*

the fine-tuning, the models didn't generate any question. T5 generated "True", F-T5 returned a empty string, and BART just replicated the context.

After the fine-tuning, the models showed a great improvement, all models were able to generate relevant questions for the reference question "Em que local aconteceu o Super Bowl 50?" ["Where was Super Bowl 50 held?"]. The T5 model wrote a more semantically correct question in Brazilian Portuguese, but computed the answer (Levi's Stadium, in English or Estádio de Levi, in Brazilian Portuguese) at the end of the phrase, and also didn't represented the final punctuation "?", not indicating that it is indeed a question.

FLAN-T5 generated a question that is not grammatically correct, as it is built as an English question. "Onde foi o Super Bowl 50 disputado?" can be translated exactly, word for word, as "Where was Super Bowl 50 played?". It is a problem because in Brazilian Portuguese, the correct way to build this question is "Where was played the Super Bowl 50?". When comparing this trait to BART's generated question, it is implied that FLAN-T5 being just a multi-lingual model, not specified only to Brazilian Portuguese, makes a difference in the grammar of the questions. BART returned the best answer, but not the perfect one.

BART returned a higher result in all the ROUGE metrics, suggesting a higher lexical similarity in the questions generated. However, it also shows a higher Loss result, implying a worse performance. Therefore, based on the RLSum values alone, the BART model is the best among these three models. However, it's important to note that the RLSum is just one aspect of model performance, and other metrics, such as the Loss scores, should also be considered for a comprehensive evaluation.

These results hold significant importance to various applications in Brazilian Portuguese. First and foremost, these advancements can significantly enhance the precision and quality of the models, in tasks such as QG. And as Brazilian Portuguese is a widely spoken language with its own linguistic intricacies, advancements in QG for this language open up avenues for innovation across various sectors, ultimately contributing to the broader goal of fostering effective communication and understanding in the landscape.

Final Remarks

In this study, we carried out an extensive analysis about the performance of BART, T5 and FLAN-T5 language models in the Question Generation task for Brazilian Portuguese. Our main evaluation focused on the models' ability to return succinct and correct questions, based on context and an answer.

We know that at times we will not always have an answer, but in this case, we provide the answer precisely to have similar questions in return, for a better comparison of results. We delved into multiple facets, encompassing response hallucination, question complexity, and the influence of context on performance. The obtained results from our experiments also showcase the effectiveness of the fine-tuning of the models for the question generation task, providing valuable insights into the model's performance, validating its utility, and paving the way for further advancements in natural language processing tasks.

While these results are promising, ongoing research and continuous refinement of the strategies are essential to enhance its accuracy further. Additionally, the feedback gained from this work provides valuable contributions to further assignments for these models. Regarding our future research, there are many improving achievements to acquire. Initially, the range could be broadened by incorporating additional datasets, allowing for a more extensive analysis. It would also be advantageous to test other models and their strategies for a larger comparison. Furthermore, it is worth considering alternative hyperparameters and fine-tuning strategies for the task, intending to achieve enhanced outcomes potentially.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We would like to thank the FAPERGS - Brasil for Financial Support, Award Agreement 22/2551-0000598-5. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature* 521(7553):436–444.

Leite, B., and Lopes Cardoso, H. 2022. Neural question generation for the portuguese language: A preliminary study. In Marreiros, G.; Martins, B.; Paiva, A.; Ribeiro, B.; and Sardinha, A., eds., *Progress in Artificial Intelligence*, 780–793. Cham: Springer International Publishing.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Mohammadshahi, A.; Scialom, T.; Yazdani, M.; Yanki, P.; Fan, A.; Henderson, J.; and Saeidi, M. 2022. Rquge: Reference-free metric for evaluating question generation by answering the question. *arXiv preprint arXiv:2211.01482*.

Pan, L.; Lei, W.; Chua, T.-S.; and Kan, M.-Y. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

Park, G.-M.; Hong, S.-E.; and Park, S.-B. 2022. Post-training with interrogative sentences for enhancing bart-based korean question generator. In *Proceedings of the 2nd conference of the Asia-Pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing*, 202–209.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21(1):5485–5551.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints* arXiv:1606.05250.

Rus, V.; Wyse, B.; Piwek, P.; Lintean, M.; Stoyanchev, S.; and Moldovan, C. 2010. The first question generation shared task evaluation challenge.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.