# Question Generation for the Portuguese Language

RAFAEL ALEXANDRE DA ENCARNAÇÃO GALHOZ, Instituto Superior Técnico, Portugal

Question generation is an important task in the effort to automatically process natural language data. It can be used as a component in the context of many significant problems such as automatic tutoring systems, improving the performance of passage retrieval or question answering models, and enabling chatbots to lead a conversation. Recent approaches leverage sequence-to-sequence models based on Transformers to achieve state-of-art results. However, most of these advances are still within the English language.

With this in mind this work focuses on the study and development different models based on the Transformer T5 architecture using both supervised and self-critical sequence training over a Portuguese translated version of the SQuAD 1.1 dataset. We compare the results obtained with English baselines, using automatic evaluation metrics. In the end it was possible to observe that the Portuguese models generate questions with lower quality and poorer syntax, although with automatic evaluation results comparable to the ones obtained in the English language models.

**Keywords:Question Generation; Transformer; Self-Critical Sequence Training; Deep Learning; Natural Language Processing; Portuguese Language;**

## 1 INTRODUCTION

The goal of question generation is to generate valid and fluent questions according to a given textual paragraph. This is a crucial aspect of the effort to automatically process natural language data, and it can be used in many scenarios such as developing automatic tutoring systems [34], improving the performance of passage retrieval [21] or question answering models [30], and enabling chatbots to lead a conversation.

Recent approaches to question generation have used sequence-to-sequence models based on Transformers, being often trained to generate a plausible question conditioned on an input document and a candidate answer span within that document [18]. Still, most of these approaches have been used only in the context of small experiments with English datasets.

This study advances over previous neural models for question generation in several directions at the same time focusing in the Portuguese language. This includes fine-tuning Transformer models such as T5 with the combined use of supervised and reinforcement learning for model training (i.e., combining the standard teacher forcing approach for maximum likelihood training, with policy gradient techniques to maximize rewards that estimate question quality and answerability) [11, 47], and exploring decoding and/or initialization methods that promote diverse generations [17, 42].

In terms of the experimental evaluation, it should be stressed that even for the English language there are currently no dedicated question generation datasets, and many authors have used the context-question-answer triples available in datasets such as SQuAD [27] and MS MARCO [20]. The main focus of this work will be on question generation for the Portuguese language, resorting to the use of machine translation to convert context-question-answer triples datasets into Portuguese so that the resulting data can be used to inform model training, evaluate these models trained over the machine-translated data, and assess the quality of the questions generated by the model compared to other models trained over English data.

### 1.1 Contributions

This work is based around fine-tuning a state-of-art Transformer T5 model in Portuguese question generation, using a machine translated version of SQuAD v1.1, and evaluating the quality of the generated questions when compared to other English question generation models. We fine-tune our models using the teacher forcing approach for maximum likelihood training using the cross-entropy loss function, and also use self-critical sequence training to fine-tune an already trained question generation model using using three different model-based rewards.

In the end, we compare the results of our fine-tuned Portuguese question generation models with both English baseline models and state-of-art approaches developed by other authors. We observed that the trained Portuguese question generation models obtain scores in the automatic evaluation metrics similar to early English question generation models.

## 2 THE T5 ARCHITECTURE

The exceptional performance of the transformer architecture when first introduced on machine translation [36] made it quickly noticed. Some authors realized that this architecture could be improved and applied to a wider range of tasks. With this several different architectures emerged such as BERT [6], composed only by the encoder blocks of the transformer, GPT [25], which has only decoder blocks, and T5 [26], that follows the traditional encoder-decoder transformer architecture.

T5, also known as Text-to-Text-Transfer-Transformer has the goal of unifying all-natural language tasks into a common text-to-text format, taking the text as input and outputting the new resulting text. The T5 model was pre-trained in the language modeling task on the C4 dataset. This task consisted in masking certain words in a paragraph with a masking token and sending them to the model with the goal of predicting what were the original words that were masked by the masked tokens. An example of this language modeling task can be seen in Figure 1. After the pre-training process, the T5 model is fine-tuned in various different tasks that include summarization, question answering, and text classification.

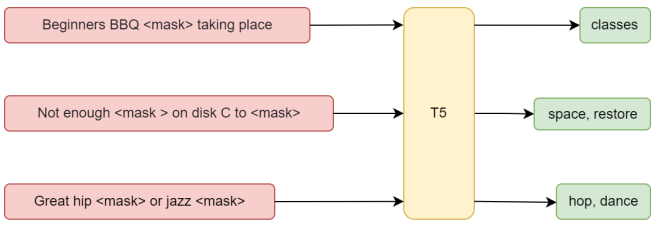


Fig. 1. The unsupervised training process of T5.

## 2.1 PTT5 and mT5

PTT5 is a T5 monolingual model pre-trained in the Brazilian Portuguese language. The training process starts from the original pre-trained T5 checkpoints and follows the same unsupervised pre-training implemented by T5, but with the Brazilian Portuguese BrWac [37] dataset, containing a large corpus of web pages in Brazilian Portuguese.

The PTT5 model was also fine-tuned on some specific tasks. The first two tasks use the ASSIN 2 dataset [29], composed of short Brazilian Portuguese sentence pairs and their respective semantic similarity and entailment relations, to allow the model to predict the semantic similarity and entailment between two different sentences. The last task uses the HAREM dataset [32], containing a collection of Portuguese-named entities, to allow the model to, given a Portuguese sentence, recognize Portuguese-named entities and their corresponding classes.

In contrast, mT5 [41] is a multilingual model, pre-trained with massive amounts of data from 101 languages, including Portuguese. It is only pre-trained through unsupervised language modeling using a massive multilingual version of the original C4 dataset, the mC4 corpus. In contrast to T5 (and PTT5), it is not fine-tuned on any specific tasks, and therefore needs to be fine-tuned before being ready for any specific multi-language task.

## 2.2 Fine-Tuning T5 for Question Generation

To fine-tune these models in the question generation task, we optimize the model parameters with the cross-entropy loss, used extensively in sequence-to-sequence models, with the goal of maximizing the log-likelihood over the training data.

During the question generation training, the model receives a tokenized input consisting of the answer and context to generate the predicted question tokens with a higher likelihood that will be used to compute the loss against the ground-truth. We apply the standard teacher-forcing strategy consisting of using the ground-truth (instead of the output of the previous sequence) when predicting the next sequence. An example of the question generation task can be seen in fig. 2.

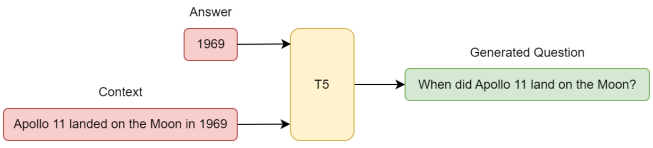


Fig. 2. An example of the question generation task using the T5 transformer.

# 3 SELF-CRITICAL SEQUENCE TRAINING

Transformer based models and other sequence-to-sequence models when using the standard teacher forcing algorithm [2] (i.e. using the ground truth as the input, instead of the model output of a prior time step as an input when training) create mismatches in generating the next action during training and testing. This happens due to the model not having access to the ground truth data during the prediction.

This problem is regarded as the exposure bias problem [28] and leads to an error accumulation during prediction, conditioning the generated words to the ground truth instead of the previously generated words. An example of the difference between the training and testing process during teacher forcing can be seen in Figure 3.

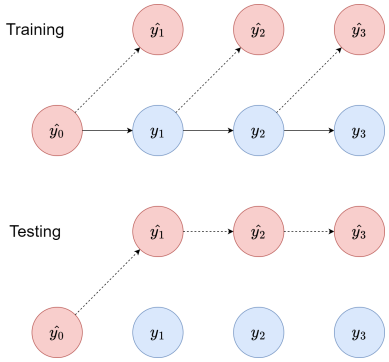


Fig. 3. The difference between training and prediction during teacher forcing.

Multiples proposals were suggested to solve this problem including scheduled sampling [28] and the use of adversarial generative models [5, 10, 35], however, our focus will revolve around the use of reinforcement learning on sequence to sequence models [12, 23]. For this reason, we resort to mixed strategy training where, after using teacher forcing, we train the model again using reinforcement learning via policy gradient, exposing the model to its own predictions and making the ground-truth available only for the reward calculation. There are multiple approaches to achieve this on sequence-to-sequence models. The one we considered relies on a policy-based reinforcement learning method using the REINFORCE algorithm [12] to solve the training/testing evaluation mismatch problem.

In reinforcement learning, an agent chooses an action based on a specific policy $\pi$. On sequence to sequence models a parameterized policy $\pi_\theta$ can be represented as $\pi_\theta(y_t|\hat{y}_{t-1}, s_t, c_{t-1})$, where $s_t$ represents the decoder and $c_t$ the context at a time step $t$. When choosing its actions regarding a current policy the model observes the rewards only at the end when comparing the sequence of predicted actions $\hat{y}_t$ (using the current policy) against the ground-truth actions $y_t$ using an evaluation metric. With this, the goal of the training process consists in finding the right parameters of the agent capable of maximizing the expected reward and we can define this loss as the negative expected reward:

$$L_\theta = -\mathbb{E}_{\hat{y}1,...,\hat{y}T\sim\pi_\theta(\hat{y}_1,...,\hat{y}_T)}[r(\hat{y}_1, ..., \hat{y}_T)], \tag{1}$$

where $r(\hat{y}_1, ..., \hat{y}_T)$ represents the reward associated.

The derivative of this loss can be calculated using a single sample from the distribution of actions of the sequence-to-sequence model as:

$$\nabla_\theta L_\theta = -\mathbb{E}_{\hat{y}1,...,T\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(\hat{y}_1...T) r(\hat{y}_1...T)], \tag{2}$$

or using the chain rule:

$$\nabla_\theta L_\theta = \frac{\partial L_\theta}{\partial \theta} = \sum_t \frac{\partial L_\theta}{\partial o_t}\frac{\partial o_t}{\partial \theta}, \tag{3}$$

$$\frac{\partial L_\theta}{\partial o_t} = (\pi_\theta(\hat{y}_t|\hat{y}_{t-1}, s_t, c_{t-1}) - \mathbf{1}(\hat{y}_t))(r(\hat{y}_1, ..., \hat{y}_T) - r_b), \tag{4}$$

where $\mathbf{1}(\hat{y}_t)$ constitutes a 1-of-$|A|$ representation of the ground truth output and $r_b$ the baseline reward, with the goal of forcing the model to select actions that result in a higher reward than the baseline reward. One possible way to compute the baseline reward is to use greedy decoding and compute the reward obtained with the result, assuring that the reward obtained by our current model only gives positive rewards if the sample is better than the current output.

This represents the REINFORCE algorithm [38], i.e. a policy gradient algorithm that can be used on sequence-to-sequence problems. However one of the major problems of this algorithm is that it suffers from high variance, since it is calculated every time a new sample is used for training. To minimize this, it is possible to sample a batch of $N$ sequences of actions at the same time in order to update the gradient, computing the average of these actions with:

$$L_\theta = \frac{1}{N}\sum_{i=1}^{N}\sum_t \log \pi_\theta(\hat{y}_{i,t}|\hat{y}_{i,t-1}, s_{i,t}, c_{i,t-1}) \times (r(\hat{y}_{i,1}, ..., \hat{y}_{i,T}) - r_b), \tag{5}$$

where $r_b$ represents the baseline reward.

## 4 EXPERIMENTAL SETUP

This section presents the baselines and the proposed question generation models, providing the implementation details behind the training process of the different models. We also describe the datasets used for both English and Portuguese language evaluation and the metrics used to assess the models.

### 4.1 Datasets

To train the different models proposed we leveraged the Stanford Question Answering Dataset (SQuAD) [27] dataset, containing almost 100,000 diverse and high-quality question-answer pairs generated by human crowdworkers on more than 500 high-quality Wikipedia articles of different topics. The dataset is mainly composed of questions starting with "What" (57k), "Who" (10k), "Which" (7k), "When" and "How many" (6k), requiring responses that are in the form of places, objects, events, persons, dates, or numbers.

In the case of the English baseline models we used the SQuAD v1.1 original dataset. Due due to the lack of dedicated Portuguese question and answering datasets, and also the lack of resources to manually translate one of those datasets, we resorted to machine translation to train the models in Portuguese. For this, we used an already existing Portuguese machine-translated version of SQuAD v1.1, containing the translated context-question-answer triples that can be used for question generation.

We evaluate the performance of each model on the respective SQuAD datasets, by splitting each dataset into a training set with 80%, the validation set with 10%, and the test set with 10% of the data, making sure that an article is only in a set and that there are no different passages extracted from the same article in multiple sets. Also, we ensure that the distribution of the articles used in the splits is the same between both languages. An example extracted from the training split in both languages can be seen in Section 4.1.

**Context:** Independence was unilaterally declared on 24 September 1973. Recognition became universal following the 25 April 1974 socialist-inspired military coup in Portugal, which overthrew Lisbon's Estado Novo regime.
**Question:** Who was overthrown in the coup?
**Answer:** Lisbon's Estado Novo regime

**Context:** A independência foi declarada unilateralmente em 24 de setembro de 1973. O reconhecimento tornou-se universal após o golpe militar de inspiração socialista de 25 de abril de 1974 em Portugal, que derrubou o Regime Estado Novo de Lisboa.
**Question:** Quem foi derrubado no golpe?
**Answer:** Regime Estado Novo de Lisboa

Fig. 4. An entry of the SQuAD v1.1 dataset in English and Portuguese.

### 4.2 Baseline Models

Our approach uses different baselines for each language. For the English language we utilize the original pre-trained T5 model architecture and mT5 [41] trained over the original SQuAD v1.1 dataset. For the Portuguese language we make use of two different pre-trained T5 models, namely PTT5 [3] and mT5 [41] trained over the Portuguese translated SQuAD v1.1 dataset. For both T5, mT5, and the PPT5 models, we use the pre-trained base models available on the Hugging Face platform due to hardware constraints that would limit the use of larger models. We also compare our results with other models developed by previous authors.

### 4.3 Implementation Details

All the models are fine-tuned over the *base* versions of the respective pre-trained models using the same parameters, with the exception of the reinforcement learning models that use a smaller learning rate. For all the models we use a maximum input sequence length of 512 (the answer and the context after tokenization) and a maximum output sequence length of 96 (that corresponds to the question). We use a batch size of 16 over a maximum of 10 epochs, and we choose the model checkpoint that achieves the lowest validation cross-entropy error over the validation set. Regarding the optimization, we use the AdamW optimizer with Adam's epsilon of $1 \times 10^{-6}$, a learning rate of $1 \times 10^{-4}$ for the cross-entropy training, and a learning rate of $1 \times 10^{-7}$ for the reinforcement learning training. The decoding length during the inference process is equal to the maximum output token size during the training process and for the automatic evaluation results decoding we use the beam search strategy with 5 beams.

To train the *T5-base*, *PTT5* and *mT5-base* models (for both languages) the tokenization is done by sending the answer and context to the tokenizer, while for the *T5-base-tokens* and *PTT5-base-tokens* the tokenization of the input is done by first concatenating two extra tokens added to the tokenizer (*<answer>*, *<context>*) with the answer text and context text, i.e. "*<answer>* ..answer text.. *<context>* ..context text.." and then sending it to the tokenizer.

All the reinforcement learning models are trained over the *ptt5-base-tokens* changing the loss function from the cross-entropy to the one represented in Equation 5., where the reward of the sampled action is calculated by computing the reward over the question generated by multinomial sampling and the baseline reward is the reward of the question generated by greedy decoding. The *PTT5-base-bertscore* uses the BERTScore metric as the reward, the *PTT5-base-bleurt* uses the BLEURT metric as the reward, while the *PTT5-base-qa-loss* uses the BERT-qa-loss metric as the reward.

### 4.4 Metrics and Evaluation

To evaluate the models we use the standard automatic evaluation metrics widely-used in question generation models [17, 40, 47], namely BLEU [22], ROUGE-L [16] and METEOR [1]. This allows us to compare our results to previous studies.

Due to the low correlation between these n-gram based metrics and human judgment described [11, 19] we also leverage some model base similarity metrics, specifically BERTScore [45] and BLEURT [33]. For BERTScore we use the default models for each language [45], while for BLEURT metric we opted to use the BLEURT-20 for both languages.

Additionally, we also developed a metric based on answerability QAP (Question Answering Probability) [44] consisting of using a BERT-based question answering model able to predict the probability of a certain token being the start or end of the answer and computing the cross-entropy loss between the predicted and true answer by computing the sum of the cross-entropy for the start and end positions of the answers. The value of this metric is then given by $e^{-loss} \times 100$, where *loss* is the loss returned by the BERT question-answering model. For the Portuguese language, we selected the BERT-based question answering model trained by Guillou[9], and for the English language, we used the model trained by Sanh et al.[31].

## 5 AUTOMATIC EVALUATION RESULTS

The automatic evaluation results on the test sets for the English language are given in Table 1, while the Portuguese results are given in Table 2. Both tables contain the results of some of the prior work on question generation developed by other authors and our trained question generation models.

In terms of general results for both languages, the ACS-aware question generation model developed by Liu et al.[17] obtains the best results in BLEU1, BLEU2, BLEU3, and ROUGE-L. The noise-aware question generation model proposed by Xiao et al.[39] achieves the highest score in the BLEU4 and METEOR metrics.

### 5.1 English Question Generation Results

Our fine-tuned English baseline models performed slightly worse on the different automatic evaluation metrics when compared against the previous state-of-art models. The baseline model with the best results is **T5-base-tokens** and has scores of 50.50 in BLEU1, 35.57 in BLEU2, 27.19 in BLEU3, 21.43 in BLEU4, 50.47 in ROUGE-L, 24.36 in METEOR, 91.97 in BERTScore, 54.72 in BLEURT and 69.64 in the QA-loss metric.

When comparing the **T5-base-tokens** results with the best ones in each metric, it is possible to observe a decrease of 1.80 in BLEU1, 1.13 in BLEU2, 0.81 in BLEU3, 3.97 in BLEU4, 2.77 in ROUGE-L and 2.56 in METEOR.

Table 1. Automatic evaluation of the English question generation models on the English test set. "BL" represents BLEU, "RL" represents ROUGE-L, "MTR" represents METEOR, "BERTS" represents BERTScore, "BLRT" represents BLUERT and QA-loss represents the Question Answerability metric. The values of prior works are removed directly from the original papers. Metric values not reported are displayed by "–". The best value of each metric is shown in bold.

| Model | BL1 | BL2 | BL3 | BL4 | RL | MTR | BERTS | BLRT | QA-loss |
|---|---|---|---|---|---|---|---|---|---|
| [8] | 43.09 | 25.96 | 17.50 | 12.28 | 39.75 | 16.62 | – | – | – |
| [46] | 43.47 | 28.23 | 20.40 | 15.32 | 43.91 | 19.29 | – | – | – |
| [15] | 45.66 | 30.21 | 21.82 | 16.27 | 44.35 | 20.36 | – | – | – |
| [4] | 49.73 | 34.60 | 26.13 | 20.33 | 48.23 | 23.88 | – | – | – |
| [7] | – | – | – | 22.12 | 51.07 | 25.06 | – | – | – |
| [39] | – | – | – | **25.40** | 52.84 | **26.92** | – | – | – |
| [24] | – | – | – | 25.01 | 52.57 | 26.83 | – | – | – |
| [17] | **52.30** | **36.70** | **28.00** | 22.05 | **53.25** | 25.11 | – | – | – |
| [47] | – | – | 25.88 | 20.13 | 47.51 | 22.93 | – | – | – |
| [13] | 48.88 | 34.37 | 26.18 | 20.55 | 49.56 | 24.29 | – | – | – |
| **mT5-base** | 48.45 | 33.67 | 25.51 | 19.97 | 48.94 | 23.24 | 91.64 | 52.83 | 67.16 |
| **T5-base** | 50.01 | 35.32 | 27.01 | 21.26 | 50.35 | 24.23 | 91.95 | 54.54 | 68.73 |
| **T5-base-tokens** | 50.50 | 35.57 | 27.19 | 21.43 | 50.47 | 24.36 | **91.97** | **54.72** | **69.64** |

### 5.2 Portuguese Question Generation Models Results

Our Portuguese baseline models perform slightly better than the PTT5 model trained by Leite and Lopes Cardoso [13], that followed the same training strategy and hyper-parameters (except the batch size which is 32). Our equivalent **PTT5-base** model records an increase of 1.36 in BLEU1, 1.63 in BLEU2, 1.57 in BLEU3, 1.34 in BLEU4 and 1.77 in ROUGE-L. This minor increase in all the metrics can be explained by the use of a different Portuguese translation of the SQuAD v1.1 dataset, where some issues created by the machine translation were fixed.

The Portuguese model with the best results in all the n-gram similarity metrics is **PTT5-base-tokens** with the scores of 45.70 in BLEU1, 32.23 in BLEU2, 24.61 in BLEU3, 19.29 in BLEU4, 45.50 in ROUGE-L, 32.80 in METEOR, 82.07 in BERTScore, 44.80 in BLEURT and 64.58 in the QA-loss metric. In contrast, the **PTT5-base-qa-loss** model performed the best in all the model-based metrics, including both model-based similarity metrics and the answerability metric, having scores of 45.51 in BLEU1, 32.07 in BLEU2, 24.48 in BLEU3, 19.18 in BLEU4, 45.41 in ROUGE-L, 32.66 in METEOR, 82.12 in BERTSCORE, 44.90 in BLEURT and 65.59 in the QA-loss metric.

When analyzing the Portuguese models and our English baseline models results, as excepted, it is possible to observe a considerable decrease in the results on all metrics. When comparing the

Portuguese **PTT5-base-tokens** with the equivalent English **PTT5-base-tokens** models there is a difference of 4.8 in BLEU1, 3.34 in BLEU2, 2.58 in BLEU3, 2.14 in BLEU4 and 4.98 in ROUGE-L. This discrepancy is even higher when compared to the models with state-of-art results. This disparity between the English and Portuguese results can be explained due to the low quality of machine-translated data, where a brief analysis of the Portuguese-translated dataset can reveal multiple syntax errors, grammatical errors, and mismatches between the answer translation and the answer span in the context.

Table 2. Automatic evaluation of the Portuguese question generation models on the Portuguese test set. The values of prior works are removed directly from the original papers. Metric values not reported are displayed by "—". The best value of each metric is shown in bold.

| Model | BL1 | BL2 | BL3 | BL4 | RL | MTR | BERTS | BLRT | QA-loss |
|---|---|---|---|---|---|---|---|---|---|
| [13] | 43.61 | 30.04 | 22.58 | 17.54 | 43.64 | — | — | — | — |
| **mT5-base** | 44.78 | 31.33 | 23.82 | 18.60 | 44.71 | 31.94 | 81.80 | 43.56 | 63.52 |
| **PTT5-base** | 44.97 | 31.67 | 24.15 | 18.88 | 45.41 | 32.42 | 82.08 | 44.74 | 64.03 |
| **PTT5-base-tokens** | **45.70** | **32.23** | **24.61** | **19.29** | **45.50** | 32.80 | 82.07 | 44.80 | 64.58 |
| **PTT5-base-qa-loss** | 45.51 | 32.07 | 24.48 | 19.18 | 45.41 | 32.66 | **82.12** | **44.90** | **65.59** |
| **PTT5-base-bertscore** | 45.36 | 31.99 | 24.41 | 19.12 | 45.40 | 32.60 | 82.07 | 44.77 | 64.42 |
| **PTT5-base-bleurt** | 45.38 | 31.98 | 24.39 | 19.11 | 45.42 | 32.59 | 82.07 | 44.79 | 64.49 |

Overall, we observe that the results of the Portuguese generation models in the automatic evaluation metrics are comparable to the results of some earlier works developed in English question generation. However, when compared to English state-of-art models we can see a substantial decrease in performance in all the metrics. It is also possible to observe that trained unilingual models in Portuguese and English perform better than their corresponding multilingual models trained in the target language. Further analysis with a high-quality Portuguese question-answer dataset is necessary.

Regarding the self-critical sequence training results, in overall, the three models that were trained using this strategy performed slightly worse after the training process, decreasing the scores in almost all the automatic evaluation metrics. The only model that improved in some way was the **PTT5-base-qa-loss** model, which after training registered an improvement in all the model-based metrics, including its own metric used as a reward, the QA-loss. However, this improvement in the model-based metrics came at the cost of the scores in the other n-gram similarity metrics.

Figure 5 represents the automatic evaluation results of the **PTT5-base-qa-loss** model over the test set after each epoch. In there it is possible to observe a slight decrease of the n-gram similarity metrics, in the beginning, stabilizing between epochs five and eight with a small increase, and then decreasing once again more abruptly. It is possible to observe a local maximum in epoch seven on the BLEU4 and METEOR, and in epoch six for ROUGE-L.The QA-loss metric improved roughly constant until epoch seven where it reached a local maximum, it decreased in epoch eight, reached the absolute maximum in epoch nine, and decreased again in the last epoch. The BERTScore and BLEURT metrics behaved very alike during training, maintaining the scores more or less stable until epoch five, where they increased to the highest values during epochs six and seven, and then decreased constantly until the end of the training.

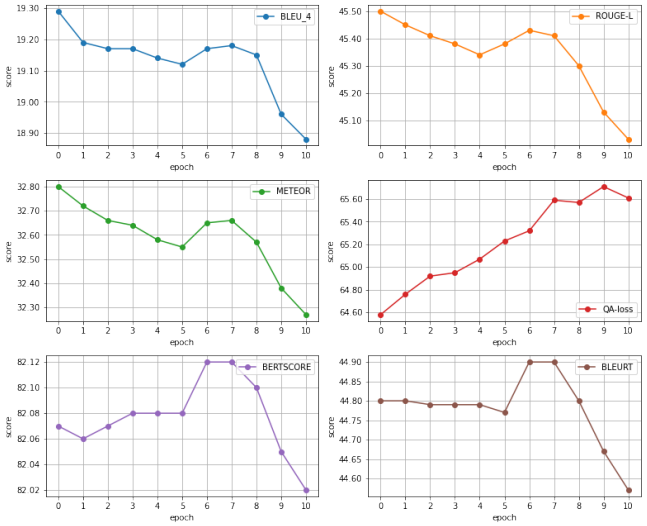


Fig. 5. Automatic evaluation metrics scores of **PTT5-base-qa-loss** over the test set after a set number of epochs. The epoch number zero represents the initial state of the model before the self-critical sequence training.

We consider epoch seven the one with the best results of self-critical sequence training using the QA-loss reward, with scores of 45.51 in BLEU1, 32.07 in BLEU2, 24.48 in BLEU3, 19.18 in BLEU4, 45.41 in ROUGE-L, 32.66 in METEOR, 82.12 in BERTScore, 44.90 in BLEURT and 65.59 in QA-loss. This results in a decrease of 0.09 to 0.19 in the n-gram similarity metrics and an increase of 0.05, 0.10, and 1.01 in BERTScore, BLEURT, and QA-loss respectively

The other two models trained using self-critical sequence training, **PTT5-base-bertscore** or **PTT5-base-bleurt** performed poorly, and any training using the BERTScore and BLEURT as a reward resulted in a constant decrease in all the automatic evaluation metrics, including the metrics used as a reward For this reason, the values registered in Table 2 were recorded after only the first epoch during the training process.

## 6 QUESTIONS GENERATED BY THE PORTUGUESE MODELS

In this section, we show some examples of the question generated by our model together with their results in the automatic evaluation metrics. To simplify we show only questions generated by the model with the best performance in the automatic evaluation, **PTT5-base-tokens**. More examples of generated questions by all the different Portuguese question generation models can be seen in Table 3.

### 6.1 Questions Correctly Generated

Some examples of questions properly generated by the Portuguese model can be seen in Figure 6 with the respective evaluation metrics. We consider a question correctly generated if it follows three criteria: (i) it respects the syntax of the Portuguese language; (ii) it is relevant (i.e. a question which refers to a specific objective fact present in the context and does not allow any rationalization to find the answer); and (iii) it is answerable.

The first example of a generated question in the Figure 6 is almost equal to the ground-truth, changing only part of the predicate more specifically the words "foi iniciada" by "começou". For this reason, this example has a high score in all the reported metrics.

The second generated question is also very similar in meaning to the original question however with more words, adding more constraints on the phrase such as "família Bell" and "Washington DC" instead of only "Bell" and "DC", and also the verb used is different "comprou" instead of "adquiriu". This makes most of the scores considerable lower than in the first example, reporting a BLEU4 score of 0.01.

The last example is the trickiest one because while the generated question is fluent, relevant, and answerable it is completely different from the ground-truth and therefore has almost the lowest possible scores in all the similarity metrics scores. Interestingly enough, since the question is well-formed and the answer provided is correct, the QA-loss score is very high, which helps confirm the value and quality of the generated question. This example helps reinforce that there is a disparity between similarity metrics scores and the quality of a generated question and that one can not fully assess the quality of a question generation model by only observing the automatic evaluation results.

**Context:** Segundo uma tradição relatada pela primeira vez por Sulcard em cerca de 1080, uma igreja foi fundada no local (então conhecida como Thorn Ey (Ilha Thorn)) no século VII, na época de Mellitus, um bispo de Londres. A construção da igreja atual começou em 1245, por ordem do rei Henrique III.
**Answer:** 1245
**Ground-Truth:** Quando foi iniciada a construção da igreja atual?
**Generated:** Quando começou a construção da igreja atual?
**BLEU4**: 62.40 **ROUGE-L**: 81.49 **METEOR**: 70.31 **BERTScore**: 97.49 **BLEURT**: 83.73 **QA-loss**: 98.28

**Context:** A casa da família Bell estava em Cambridge, Massachusetts, até 1880, quando o sogro de Bell comprou uma casa em Washington, DC, e mais tarde em 1882 comprou uma casa na mesma cidade para a família de Bell, para que pudessem ficar com ele enquanto ele participou de inúmeros processos judiciais envolvendo disputas de patentes.
**Answer:** 1882
**Ground-Truth:** Em que ano Bell adquiriu uma casa em DC?
**Generated:** Em que ano a família Bell comprou uma casa em Washington DC?
**BLEU4**: 0.01 **ROUGE-L**: 80.15 **METEOR**: 64.89 **BERTScore**: 86.61 **BLEURT**: 73.59 **QA-loss**: 99.75

**Context:** Em 1858, o imperador francês Napoleão III obteve com sucesso a posse, em nome do governo francês, da Longwood House e das terras ao seu redor, última residência de Napoleão I (que morreu ali em 1821). Ainda é propriedade francesa, administrada por um representante francês e sob a autoridade do Ministério de Relações Exteriores da França.
**Answer:** Napoleão I
**Ground-Truth:** Quem foi o último morador da casa de Longwood antes de Napoleão III assumir o controle?
**Generated:** Quem morreu em 1821?
**BLEU4**: 0.0 **ROUGE-L**: 16.55 **METEOR**: 4.29 **BERTScore**: 73.38 **BLEURT**: 15.59 **QA-loss**: 92.80

Fig. 6. Example of correctly generated questions by the **PTT5-base-tokens** model.

## 6.2 Questions Generated with Semantic Errors

Some examples of questions generated by the Portuguese model with semantic errors that make the questions unanswerable or with an incorrect answer can be seen in Figure 7 with the respective evaluation metrics. In this case, we consider a question unanswerable if it can not be answered with the given context. A question with an incorrect answer is a question that can not be answered correctly with the answer given during the generation but can be answered based on the context.

The first example of a generated question in the Figure 7 is completely different from the ground-truth. However, the issue regarding this question is that the correct answer to "A quem pertence o Museu Presidencial?" is not "Abraham Lincoln" but "estado de Illinois" or "Illinois". For this reason, the QA-loss score is low, because it was calculated using the wrong answer. Also, the results of the similarity metrics are very low due to the lack of similarity between the generated and original questions.

The second generated question represents a question that does not make sense and can not be answered based on the context provided. This makes this question unanswerable and consequently the answer incorrect, which makes the QA-loss almost 0. However, due to having some of the words contained in the ground-truth it has higher scores in the similarity metrics than the first example.

**Context:** Bibliotecas e museus foram estabelecidos para outros presidentes, mas eles não fazem parte do sistema de bibliotecas presidenciais da NARA e são operados por fundações privadas, sociedades históricas ou governos estaduais, incluindo Abraham Lincoln, Rutherford B. Hayes, William McKinley, Woodrow Bibliotecas Wilson e Calvin Coolidge. Por exemplo, a Biblioteca e Museu Presidencial Abraham Lincoln pertence e é operada pelo estado de Illinois.
**Answer:** Abraham Lincoln
**Ground-Truth:** Qual é o primeiro presidente que a NARA não possui registros em seu sistema de bibliotecas presidenciais?
**Generated:** A quem pertence o Museu Presidencial?
**BLEU4**: 0.0 **ROUGE-L**: 14.82 **METEOR**: 5.90 **BERTScore**: 74.48 **BLEURT**: 7.37 **QA-loss**: 4.30

**Context:**A corrupção também pode afetar os vários componentes das atividades de Esportes (árbitros, jogadores, equipe médica e de laboratório envolvidos nos controles antidoping, membros da federação esportiva nacional e comitês internacionais que decidem sobre a alocação de contratos e locais de competição).
**Answer:** Esportes
**Ground-Truth:** A corrupção pode determinar certas coisas em que tipo de atividades?
**Generated:** O que pode afetar a corrupção?
**BLEU4**: 0.0 **ROUGE-L**: 30.15 **METEOR**: 22.69 **BERTScore**: 81.1 **BLEURT**: 47.03 **QA-loss**: 0.04

Fig. 7. Example of generated questions by the **PTT5-base-tokens** model with semantic errors.

## 6.3 Questions Generated with Syntax Errors

Some examples of questions generated by the Portuguese model with syntax errors can be seen in Figure 8 with the respective evaluation metrics. A question with syntax errors is a question that is not well-formed and has some sort of grammatical or lexical issue. A large number of questions generated by the model have syntax problems, so in these examples, we only show questions where the issue was so significant that the question could not be understood by looking at either the context, answer, or the respective ground-truth.

In the first example of the Figure 8, the question generated does not follow the correct syntax of the Portuguese language, making it confusing and almost impossible to answer it correctly. Nonetheless by looking at the beginning of the context and the given answer we can somewhat understand what the model tried to achieve. Due to the low similarity between the generated and ground-truth questions, the similarity metrics results are very low. Also, even with the syntax problems of the question, the QA-loss metric result was decent for this example.

Regarding the second example, the question generated also does not follow the correct syntax of the Portuguese language, and it is practically impossible to understand and answer correctly. Despite this, the QA-loss metric score of this question is very high, and also low scores on all the similarity metrics.

**Context:** Varejistas, fabricantes de artigos esportivos e outras empresas se beneficiam da luz solar adicional da tarde, pois induz os clientes a comprar e a participar de esportes ao ar livre da tarde. Em 1984, a revista Fortune estimou que uma extensão de sete semanas do horário de verão renderia US $ 30 milhões adicionais para as lojas 7-Eleven, e a National Golf Foundation estimou que a extensão aumentaria as receitas da indústria de golfe de US $ 200 milhões para US $ 300 milhões. Um estudo de 1999 estimou que o horário de verão aumenta a receita do setor de lazer da União Europeia em cerca de 3%.
**Answer:** artigos esportivos
**Ground-Truth:** Que categoria de produtos usados em atividades ao ar livre se beneficia da hora extra de luz do dia no horário de verão?
**Generated:** O que os fabricantes de?
**BLEU4**: 0.0 **ROUGE-L**: 18.05 **METEOR**: 4.61 **BERTScore**: 66.40 **BLEURT**: 3.60 **QA-loss**: 72.68

**Context:** Os dois países planejaram uma missão conjunta para atracar a última nave Apollo dos EUA com uma Soyuz, conhecida como Projeto de Teste Apollo-Soyuz (ASTP). Para se preparar, os EUA projetaram um módulo de ancoragem para o Apollo compatível com o sistema de ancoragem soviético, que permitia que qualquer de suas embarcações atracasse com qualquer outro (por exemplo, Soyuz / Soyuz e Soyuz / Salyut). O módulo também era necessário como uma câmara de ar para permitir que os homens visitassem a nave um do outro, que possuía atmosferas incompatíveis na cabine. A URSS usou a missão Soyuz 16 em dezembro de 1974 para se preparar para o ASTP.
**Answer:** Projeto de Teste Apollo-Soyuz
**Ground-Truth:** ASTP significa o que?
**Generated:** ASTP de quê?
**BLEU4**: 0.0 **ROUGE-L**: 21.79 **METEOR**: 12.63 **BERTScore**: 75.48 **BLEURT**: 48.35 **QA-loss**: 96.46

Fig. 8. Example of generated questions by the **PTT5-base-tokens** model with syntax errors.

## 7 CONCLUSION

The goal of this study was to address the current state-of-art of question generation while exploring how to implement these methods for question generation in a different language without a decent amount of natural language processing corpora. We explored fine-tuning a pre-trained state-of-art language model in Portuguese question generation using a machine-translated dataset. We achieved this by developing multiple baselines and models using both supervised and reinforcement learning and evaluating them using automatic evaluation metrics.

In the end, we conclude that the quantitative results obtained on the Portuguese models were comparable with earlier works made by other authors using dedicated high-quality question-answering datasets. Nonetheless, it was possible to observe issues regarding the use of a machine-translated dataset with poor quality to fine-tune a question generation model. For this reason, we emphasize the importance for the natural language processing community of the creation of dedicated Portuguese and other language question-answer datasets. Also, we want to reinforce that the automatic evaluation metrics based on n-grams used in this study have been proven to correlate poorly with human judgment, so a further human evaluation would provide a more accurate understanding of the quality of the generated questions.

### 7.1 Future Work

Although the automatic evaluation results fell under our initial expectations of using a machine-translated dataset, with results comparable to early English question generation models, there are still multiple ways capable of improving the results of question generation in the Portuguese language.

The most straightforward method to achieve this is to implement state-of-art strategies that improved the results of English question generation models. We detected two different approaches that generated great results in English and have the possibility of improving the results in Portuguese: (i) the ACS-aware question generation model proposed by Liu et al.[17] that promotes the generation of more diverse question and (ii) the noise-aware question generation model proposed by Xiao et al. ([39] that implements a noise-aware generation method that strives to reduce the discrepancy between training and inference to minimize the exposure bias problem. Additionally, it could also be interesting to explore other pre-trained sequence-to-sequence models such as PEGASUS [43] and BART [14].

Another idea is to experiment using self-critical sequence training with other model-based text generation metrics or n-gram similarity metrics such as the ones introduced by Nema and Khapra[19] that were not used in this work. Also, the use of other methods to solve the exposure bias problem such as scheduled sampling [28], the use of adversarial generative models [5, 10, 35] or the noise generation method [39] can be more effective than the use of self-critical sequence training.

Other possible and more reliable path to improve Portuguese question generation further is to improve the quality of Portuguese corpora for question generation and question answering, either by developing a rule-based system to clean up the data or manually by hiring human crowdworkers. This would significantly improve the results by reducing the number of errors on the translated dataset.

## REFERENCES


[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909

[2] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada) *(NIPS'15)*. MIT Press, Cambridge, MA, USA, 1171–1179.

[3] Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. arXiv:2008.09144 [cs.CL]

[4] Ying-Hong Chan and Yao-Chung Fan. 2019. A Recurrent BERT-based Model for Question Generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Hong Kong, China, 154–162. https://doi.org/10.18653/v1/D19-5821

[5] Tong Che, Yanran Li, Ruixiang Zhang, R. Devon Hjelm, Wenjie Li, Yangqiu Song, and Yoshua Bengio. 2017. Maximum-Likelihood Augmented Discrete Generative Adversarial Networks. *ArXiv* abs/1702.07983 (2017). arXiv:1702.07983 [cs.AI]

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2019). arXiv:1810.04805 [cs.CL]

[7] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. https://doi.org/10.48550/ARXIV.1905.03197

[8] Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension. , 1342–1352 pages. https://doi.org/10.18653/v1/P17-1123

[9] Pierre Guillou. 2021. Portuguese BERT base cased QA (Question Answering), finetuned on SQUAD v1.1.

[10] Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long Text Generation via Adversarial Training with Leaked Information. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). https://ojs.aaai.org/index.php/AAAI/article/view/11957

[11] Tom Hosking and Sebastian Riedel. 2019. Evaluating Rewards for Question Generation Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2278–2283. https://doi.org/10.18653/v1/N19-1237

[12] Yaser Keneshloo, Tian Shi, Naren Ramakrishnan, and Chandan K. Reddy. 2020. Deep Reinforcement Learning for Sequence-to-Sequence Models. *IEEE Transactions on Neural Networks and Learning Systems* 31, 7 (2020), 2469–2489. https://doi.org/10.1109/TNNLS.2019.2929141

[13] Bernardo Leite and Henrique Lopes Cardoso. 2022. Neural Question Generation for the Portuguese Language: A Preliminary Study. In *Progress in Artificial Intelligence*, Goreti Marreiros, Bruno Martins, Ana Paiva, Bernardete Ribeiro, and Alberto Sardinha (Eds.). Springer International Publishing, Cham, 780–793.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. https://doi.org/10.48550/ARXIV.1910.13461

[15] Jingjing Li, Yifan Gao, Lidong Bing, Irwin King, and Michael R. Lyu. 2019. Improving Question Generation With to the Point Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3216–3226. https://doi.org/10.18653/v1/D19-1317

[16] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[17] Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. 2020. Asking Questions the Human Way: Scalable Question-Answer Generation from Text Corpus. *Proceedings of The Web Conference 2020* (Apr 2020). https://doi.org/10.1145/3366423.3380270

[18] Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2021. Simplifying Paragraph-Level Question Generation via Transformer Language Models. In *PRICAI 2021: Trends in Artificial Intelligence*, Duc Nghia Pham, Thanaruk Theeramunkong, Guido Governatori, and Fenrong Liu (Eds.). Springer International Publishing, Cham, 323–334.

[19] Preksha Nema and Mitesh M. Khapra. 2018. Towards a Better Metric for Evaluating Question Generation Systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3950–3959. https://doi.org/10.18653/v1/D18-1429

[20] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268 (2016). arXiv:1611.09268 http://arxiv.org/abs/1611.09268

[21] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *CoRR* abs/1904.08375 (2019). arXiv:1904.08375 http://arxiv.org/abs/1904.08375

[22] 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 on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[23] Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A Deep Reinforced Model for Abstractive Summarization. arXiv:1705.04304 [cs.CL]

[24] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting Future N-gram for Sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2401–2410. https://doi.org/10.18653/v1/2020.findings-emnlp.217

[25] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[27] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

[28] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. arXiv:1511.06732 [cs.LG]

[29] Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*. Springer, 406–412.

[30] Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2020. Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering. *CoRR* abs/2010.12643 (2020). arXiv:2010.12643 https://arxiv.org/abs/2010.12643

[31] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC$^2$Workshop*.

[32] Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. Harem: An advanced ner evaluation contest for portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*.

[33] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. *CoRR* abs/2004.04696 (2020). arXiv:2004.04696 https://arxiv.org/abs/2004.04696

[34] Riken Shah, Deesha Shah, and Lakshmi Kurup. 2017. Automatic question generation for intelligent tutoring systems. In *2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*. 127–132. https://doi.org/10.1109/CSCITA.2017.8066538

[35] Jinyue Su, Jiacheng Xu, Xipeng Qiu, and Xuanjing Huang. 2018. Incorporating Discriminator in Sentence Generation: a Gibbs Sampling Method. arXiv:1802.08970 [cs.CL]

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[37] Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC Corpus: A New Open Resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1686

[38] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. https://doi.org/10.1023/A:1022672621406

[39] Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-GEN: An Enhanced Multi-Flow Pre-training and Fine-tuning Framework for Natural Language Generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3997–4003. https://doi.org/10.24963/ijcai.2020/553 Main track.

[40] Yuxi Xie, Liangming Pan, Dongzhe Wang, Min-Yen Kan, and Yansong Feng. 2020. Exploring Question-Specific Rewards for Generating Deep Questions. *CoRR* abs/2011.01102 (2020). arXiv:2011.01102 https://arxiv.org/abs/2011.01102

[41] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41

[42] Xiang Yue, Xinliang Frederick Zhang, Ziyu Yao, Simon Lin, and Huan Sun. 2021. CliniQG4QA: Generating Diverse Questions for Domain Adaptation of Clinical Question Answering. arXiv:2010.16021 [cs.CL]

[43] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. https:

//doi.org/10.48550/ARXIV.1912.08777

[44] Shiyue Zhang and Mohit Bansal. 2019. Addressing Semantic Drift in Question Generation for Semi-Supervised Question Answering. *CoRR* abs/1909.06356 (2019). arXiv:1909.06356 http://arxiv.org/abs/1909.06356

[45] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL]

[46] Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3901–3910. https://doi.org/10.18653/v1/D18-1424

[47] Peide Zhu and Claudia Hauff. 2021. Evaluating BERT-Based Rewards for Question Generation with Reinforcement Learning. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (Virtual Event, Canada) *(ICTIR '21)*. Association for Computing Machinery, New York, NY, USA, 261–270. https://doi.org/10.1145/3471158.3472240

Table 3. Examples of generated questions using the developed Portuguese question generation models. The contexts, ground-truths and answers are all extracted from the Portuguese test set.

| **Example 1** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Context** | O primeiro satélite do sistema de segunda geração, o Compass-M1, foi lançado em 2007. Seguiu-se mais nove satélites durante o período 2009-2011, alcançando uma cobertura regional funcional. Um total de 16 satélites foram lançados durante esta fase. | | | | | | |
| **Answer** | 16 | **BL4** | **RL** | **MTR** | **BERTS** | **BLRT** | **QA-loss** |
| **Ground-truth** | Quantos satélites foram lançados desde 2007? | — | — | — | — | — | 0.74 |
| *mT5-base* | Quantos satélites foram lançados durante o período 2009-2011? | **31.56** | **63.94** | **53.73** | **92.86** | **66.11** | **1.02** |
| *PTT5-base* | Quantos satélites foram lançados durante a segunda geração? | **31.56** | **63.94** | **53.73** | 91.09 | 52.51 | 0.25 |
| *PTT5-base-tokens* | Quantos satélites foram lançados durante a segunda geração? | **31.56** | **63.94** | **53.73** | 91.09 | 52.51 | 0.25 |
| *PTT5-base-qa-loss* | Quantos satélites foram lançados durante a segunda geração? | **31.56** | **63.94** | **53.73** | 91.09 | 52.51 | 0.25 |
| *PTT5-base-bertscore* | Quantos satélites foram lançados durante a segunda geração? | **31.56** | **63.94** | **53.73** | 91.09 | 52.51 | 0.25 |
| *PTT5-base-bleurt* | Quantos satélites foram lançados durante a segunda geração? | **31.56** | **63.94** | **53.73** | 91.09 | 52.51 | 0.25 |
| **Answer** | 2007 | **BL4** | **RL** | **MTR** | **BERTS** | **BLRT** | **QA-loss** |
| **Ground-truth** | Quando foi lançado o satélite Compass-M1? | — | — | — | — | — | **76.04** |
| *mT5-base* | Quando o Compass-M1 foi lançado? | 0.0 | 60.7 | 40.68 | 87.68 | **79.36** | 66.66 |
| *PTT5-base* | Quando o Compass-M1 foi lançado? | 0.0 | 60.7 | 40.68 | 87.68 | **79.36** | 66.66 |
| *PTT5-base-tokens* | Quando o satélite experimental do Compass-M1 foi lançado? | 0.0 | **63.94** | **52.54** | **88.98** | 74.18 | 52.42 |
| *PTT5-base-qa-loss* | Quando o satélite experimental do Compass-M1 foi lançado? | 0.0 | **63.94** | **52.54** | **88.98** | 74.18 | 52.42 |
| *PTT5-base-bertscore* | Quando o satélite experimental do Compass-M1 foi lançado? | 0.0 | **63.94** | **52.54** | **88.98** | 74.18 | 52.42 |
| *PTT5-base-bleurt* | Quando o satélite experimental do Compass-M1 foi lançado? | 0.0 | **63.94** | **52.54** | **88.98** | 74.18 | 52.42 |
| **Answer** | cobertura regional funcional | **BL4** | **RL** | **MTR** | **BERTS** | **BLRT** | **QA-loss** |
| **Ground-truth** | O que foi alcançado com o lançamento de 9 satélites adicionais entre 2009 e 2011? | — | — | — | — | — | **83.95** |
| *mT5-base* | O que o Compass-M1 alcançou durante o período 2009-2011? | 0.0 | 29.54 | 15.14 | **72.54** | **31.66** | 82.53 |
| *PTT5-base* | Qual foi a cobertura do Compass-M1? | 0.0 | 16.25 | 4.36 | 65.61 | 11.33 | 28.53 |
| *PTT5-base-tokens* | O que o Compass-M1 alcançou? | 0.0 | **33.61** | **16.26** | 66.11 | 17.17 | 82.21 |
| *PTT5-base-qa-loss* | O que o satélite Compass-M1 alcançou? | 0.0 | 24.37 | 8.73 | 67.06 | 22.06 | 80.07 |
| *PTT5-base-bertscore* | Que tipo de cobertura o Compass-M1 alcançou? | 0.0 | 23.58 | 9.64 | 62.04 | 18.48 | 32.31 |
| *PTT5-base-bleurt* | Que tipo de cobertura o Compass-M1 alcançou? | 0.0 | 23.58 | 9.64 | 62.04 | 18.48 | 32.31 |
| **Example 2** | | | | | | | |
| **Context** | BBC Television é um serviço da British Broadcasting Corporation. A corporação, que opera no Reino Unido sob os termos de uma carta real desde 1927, produz programas de televisão por conta própria desde 1932, embora o início de seu serviço regular de transmissão de televisão seja datado de 2 de novembro de 1936. | | | | | | |
| **Answer** | Reino Unido | **BL4** | **RL** | **MTR** | **BERTS** | **BLRT** | **QA-loss** |
| **Ground-truth** | Em que país a BBC está sediada? | — | — | — | — | — | 98.97 |
| *mT5-base* | Onde a BBC Television opera desde 1927? | 0.0 | 37.5 | 22.62 | 73.89 | 42.24 | 75.29 |
| *PTT5-base* | Em que país a BBC opera? | **55.78** | **79.05** | **65.78** | **91.25** | **67.64** | **99.14** |
| *PTT5-base-tokens* | Em que país a BBC opera? | **55.78** | **79.05** | **65.78** | **91.25** | **67.64** | **99.14** |
| *PTT5-base-qa-loss* | Em que país a BBC opera? | **55.78** | **79.05** | **65.78** | **91.25** | **67.64** | **99.14** |
| *PTT5-base-bertscore* | Em que país a BBC opera? | **55.78** | **79.05** | **65.78** | **91.25** | **67.64** | **99.14** |
| *PTT5-base-bleurt* | Em que país a BBC opera? | **55.78** | **79.05** | **65.78** | **91.25** | **67.64** | **99.14** |
| **Answer** | 2 de novembro de 1936 | **BL4** | **RL** | **MTR** | **BERTS** | **BLRT** | **QA-loss** |
| **Ground-truth** | Em que data a BBC transmitia regularmente pela TV? | — | — | — | — | — | 98.65 |
| *mT5-base* | Qual é o início do serviço regular de transmissão de televisão da BBC? | 0.0 | 17.18 | 5.46 | 75.77 | 55.22 | 98.69 |
| *PTT5-base* | Quando foi o início do serviço regular de transmissão de televisão? | 0.0 | 9.24 | 2.86 | 75.28 | 55.2 | **99.15** |
| *PTT5-base-tokens* | Quando a BBC começou a transmitir programas de televisão? | 0.0 | **30.0** | **18.62** | **78.14** | **68.49** | 28.96 |
| *PTT5-base-qa-loss* | Quando foi o início do serviço regular de transmissão de televisão da BBC? | 0.0 | 17.18 | 5.46 | 76.97 | 63.78 | 98.15 |
| *PTT5-base-bertscore* | Quando foi o início do serviço regular de transmissão de televisão da BBC? | 0.0 | 17.18 | 5.46 | 76.97 | 63.78 | 98.15 |
| *PTT5-base-bleurt* | Quando foi o início do serviço regular de transmissão de televisão da BBC? | 0.0 | 17.18 | 5.46 | 76.97 | 63.78 | 98.15 |