Assessing and Evaluating the Performance of Sequence to Sequence Models in Natural Language Generation

Emilio Medina

UWC Maastricht, Maastricht, Netherlands

Supervised by: Joe Xiao

Abstract

Sequence-to-sequence models are a type of machine learning encoder-decoder architecture designed for tasks involving sequential data. This data type is vast and of great significance, yet, little research is available about the performance comparison between different sequence-to-sequence models. This paper aims to give a quantitative and qualitative analysis and comparison of an RNN, GRU, LSTM, and Transformer Model. This was achieved using the most well-known sequence-to-sequence metrics: Rougescore, BLEU, and BERTscore. The analysis was done for the task of text generation of Homer's writing style using a small corpus of data. It was observed that for these conditions, the automated scores (Rougescore and BLEU score) are futile since they reward the mimicking of a sentence rather than the similarity to the writing style. Additionally, it was noted that the lack of data impacted the performance of the more complex models, supporting the claim that when little data is available less complex models proved to be more efficient. These findings are relevant since they offered a comparison between models for text generation tasks and suggested the need for more and different sequence-to-sequence evaluation metrics.

Introduction

Machine learning has become and is still on its way to being one of the revolutionary inventions of our lifetime. This discipline encapsulates a variety of models that simplify and even improve a number of tasks when compared to human performance. It includes a broad range of architectures that allow for numerous different potential applications; nevertheless, the models that commonly hold the most prestige cannot be applied to a great number of situations (Moses, 2021). The architectures I am talking about are deep neural networks, which are multi-level convolutional networks that perform non-linear operations with, so far, one of the best efficiencies (S. Jothilakshmi, 2016). This is the case because deep neural networks are effective whenever there are large labeled training sets that map input and output in a one-to-one manner, but they are not as good at mapping sequential data, which is not necessarily labeled

(Sutskever, 2014). Additionally, another of the main deep neural networks' flaws is the amount of complexity required, which oftentimes cannot be handled by the freely available GPUs, and require software and hardware that not everybody possesses, making it impractical in some situations. To solve these issues, an architecture known as RNN (Recurrent Neural Network) is used. It was brought up in 1986 (Yanhui, 2021) and consists of a relatively simple API combined with deep neural networks where a fixed-size input vector produces a fixed-sized output vector, carrying out several steps during the process where the vectors are repeatedly mapped. This type of approach is very useful for sequences since it can analyze the data through a holistic approach, and depending on the arguments, it can generate another sequence or just a value, but the output is generated with the consideration of the whole sequence, instead of the individual parts of it. Moreover, as mentioned, it is applicable for sequential data, such as language, sound, or some specific vectors, but it can also analyze non-sequential data with a sequential approach to enhance the effectivity of the algorithm, as it is sometimes done in computer vision (Karpathy, 2015). All of these possibilities are encapsulated in the sequence spectrum, which labels models from one-to-one (non-sequential) to many-to-many (sequence to sequence).

The specific fields where these encoder-decoder architectures have been historically and recently most applied are numerous and varied. One of the main ones, which is also explored in this research paper is Natural Language Processing, including dialog, translation, summarization, paraphrasis, and recognition, among many more. For instance, most translator algorithms use some sort of sequence-to-sequence model which allows them to yield more than one result and identify the different definitions of a word in the different contexts. Most recently, transformer models have been used for this task. Additionally, several of these algorithms include an attention feature, which allows them to focus and pay more emphasis on a specific part of the sample data, being able to increase the importance that things such as keywords have on the outcome. Moreover, there have been many recent implementations that separate from the usual natural language processing, such as Jupiter Networks Inc.'s efforts to implement sequence-to-sequence models to enhance technical support automation and provide better quality support in a more profitable manner (Allipour, 2018).

Furthermore, as expressed, sequence-to-sequence models are a reality and are used in a wide range of fields. Several papers and studies are available and cover both the technical and practical side of the use of this architecture, including examples such as “Sequence to Sequence Learning with Neural Networks” from Sustkever, where their difference with non-sequential neural networks is discussed, or “Evaluation of Text Generation: A survey”, from Celikyilmaz, which analyzes the performance of different automated evaluation metrics for this kind of models. Nonetheless, despite the thorough understanding and research available in the different architectures, there is little quantitative analysis that compares the performance of the diverse sequence to sequence models in a controlled environment. There is information about the methodology and performance of a specific architecture, but comparisons with other sequence-to-sequence models under the same task are not available. Additionally, due to the lack of comparison between different sequence-to-sequence models, the effectiveness of the existing metrics to assess this type of architecture is unknown, and a discussion about which metric is better in what situation or whether automated metrics are truly useful in generative tasks is unavailable.

Due to the gap of knowledge in this field, the motivation of this research paper is to compare and contrast the best-known sequence-to-sequence models and provide both a quantitative and qualitative analysis and comparison of these models in order to draw conclusions about which one is more efficient in generative language tasks, and which evaluation metrics proved to be useful to draw these conclusions. The aim of this paper will be to compare 4 of the best-known encoder-decoder models to analyze sequential data in a generative language task. More precisely, this paper will compare the effectiveness of an RNN, LSTM, GRU, and a GPT-3 powered transformer model in a text generation task, where the objective is to mimic Homer's writing style using a small training corpus. The conclusions will be drawn using Rouge, BLEU, and Bertscores; thus, the usefulness of these metrics in assessing the effectiveness of a model in a generation task will be assessed and, if necessary, compared with other alternatives such as human-fed feedback. The completion of the proposed aim is important since the decision of choosing which sequence-to-sequence algorithm to use, just as with any other model, is dependent on a great number of variables, including training corpus size, the complexity of the algorithm, and the task at stake, among many others. Hence, the research paper will contribute by providing quantitative and qualitative data on what algorithm is best in a specific situation and will promote further analysis of these types of models under diverse conditions. It will give a rough understanding of how each model is unique, and in what type of tasks each of the models excels on.

Methodology

In order to achieve a meaningful analysis and comparison of sequence-to-sequence models, an RNN, LSTM, GRU, and Transformer Model powered by a GPT-3 will be compared in a controlled environment. This environment consists of a set of data which is based on Homer's books: The Odyssey (800 B.C.E) and The Iliad (800 B.C.E). The used version will be the English translation of Samuel Butler, available in the MIT Classics digital library. For this experiment, a total of 5 books, which translates to 25,431 words, or 2,591 lines of code, are rawly fed into the models. The selected books are The Iliad: book 1, book 7, book 12, and The Odyssey: book 9, book 14. The sentences were manually selected, attempting to pick random samples that included a variety of length sentences and topics, aiming to test the model in as many contexts as possible. The data was divided into sets of training vs test data in a 95:5 ratio. The test data consists of sentence pairs, whose first part is the sample reference sentence that is common to every model, and the second part will be predicted by every architecture and then compared to the reference data. Additionally, all of the models will be trained for a total of 100 epochs to ensure fair conditions. These epochs are trained on the Google Colab environment using Python 3 and the T4 GPU.

Firstly, an RNN will be assessed. RNNs use previous data to create output. They do so by giving a weight and a bias to every value of the sequential vector, usually with a "tanh" activation function. RNNs give the same weight to all the parts of the sequence, producing short-term memory due to the vanishing gradient problem. The vanishing gradient problem occurs because to train the network the algorithm back propagates to time, calculating the gradient on every step. Since the weights are the same, older gradients become less significant

and impactful, causing short-term memory since they do not affect the final output that has back-propagated other values before (Pedamallu, 2020). Regarding the RNN used in the experiment, it is a Simple RNN with 128 layers from the Keras library. It also includes a dense layer with “softmax” activation, a “categorical_crossentropy” loss function, and an “adam” optimizer. It will predict the same number of words that are present in the reference set of data.

Then, a GRU (Deng, 2019) will be analyzed. GRUs have a similar architecture and purpose as RNNs, yet, they are more modern and attempt to solve the vanishing gradient problem. They do this by storing the activation value in a memory cell so it can be taken into account in further iterations. The weight of this stored activation value depends on the gates of the architecture, which are neural networks. In GRUs, there are two gates, the reset gate and the update gate. The update gate is in charge of deciding whether the current cell should be updated with the activation value or not. The reset gate is in charge of deciding the importance of the previous cell. The GRU used for the experiment is a GRU with 128 layers from the Keras library. It has the same activation, dense layer, loss function, and optimizer as the RNN.

Also, an LSTM (Yong, 2019) will be evaluated. Just as GRUs, LSTMs have a very similar architecture to RNNs, being the presence of gates what differentiates them the most and prevents them from having a short-term memory. Aside from the already mentioned gates present in GRUs, LSTMs have two extra gates. Firstly, the forget gate, which decides what and how much information from the previous cell should be forgotten. Second, the output gate, which selects what part of each cell is output to the activation function, also known as the hidden state. In the case of the experiment, an LSTM with 128 layers from the Keras library was used. It shares the activation, dense layer, loss function, and optimizer with the foregoing two architectures. The RNN, GRU, and LSTM, apart from some similarities in architecture and purpose, have two main things in common. One of them is that the next sentence prediction is done through a next word prediction or next token prediction process, where the amount of tokens that are selected in each iteration corresponds to the amount of tokens in the reference data. This connects with the next similarity, which is that unlike the last model used, these 3 architectures can be classified as many-to-one models, given that they turn a sequence of vectors and tokens into a single token. This operation is repeated numerous times, with a sequence-to-sequence approach to come up with the predicted sentences that are intended to be as similar as possible to Homer’s style.

Additionally, a Transformer model (Fayyaz, 2022) will be assessed. Transformer models are architectures that use the characteristic encoder-decoder approach of sequence-to-sequence algorithms, plus an internalized use of the attention mechanism. As previously introduced, the attention mechanism aids the models to highlight and emphasize important information which has a higher impact on the weights of the hidden states. In order to do this it uses dot products, where the dot product of all the tokens present in a sequence is calculated. This dot product indicates the similarity and correlation between the different vectors of the sequence, and with this information, the attention mechanism can emphasize those vectors with a correlation after the vectors with a higher score which is calculated as a final probability distribution, are given a higher weight on their activation function. Transformer models employ this type of algorithm in every encoder layer of the architecture. This approach is known as a self-attention mechanism, and it does not only yield the improved results seen in

normal attention mechanisms, but it also supports the encoder layers to achieve a more effective encoded sequence (Yasar, 2021). For the experiment, a GPT-3 pre-trained transformer model, which is about the size of GPT-Neo and is available in the 'happytransformer' (Fillion, 2023) library, was used. It was additionally trained with the same data as the other algorithms. Also, it was trained with a top k of 50 and a temperature of 0.7. The top k is the number of most likely tokens that the model considers for each subsequent token: higher values allow a higher pool of tokens, and lower values provide safer results. Alternatively, in the context of a transformer model, the temperature determines the probability mass function, which is in charge of distributing the predicted output, influencing the likelihood of lower probability outputs (Aws, 2024). Unlike the rest of the architectures, the transformer model is a many-to-many sequence-to-sequence model. In this case, the amount of predicted tokens was not eligible, causing a sequence of data to be encoded and decoded to another sequence whose only technical limitation is the maximum amount of predicted tokens, yet, the actual length varies, and it is up to the model to decide what sequence fits the previous sentence the best.

Apart from the used models that produce the predicted sentences, it is important to mention the metrics that are used to evaluate and draw predictions on the experiment. Firstly, a Rouge score for every architecture was calculated. Celikyilmaz classified the Rouge score as an n-gram recall metric. This means that it is able to measure the overlap between reference vectors and predicted vectors in a sequence. This type of metric is usually used to evaluate summarization tasks; nonetheless, it lacks the ability to recognize the semantic importance of a predicted set (Mudadla, 2023). In addition, a BLEU score was calculated, which is classified by Celikyilmaz as an n-gram precision metric. The BLEU score attempts to capture the similarity between human-generated sequences and predicted sequences depending not only on the overlap of tokens but also their precision, thus, their position within the overall sequence. This metric is usually used to evaluate translation models, yet, it does not capture many external factors, such as grammatical correctness or alternate semantically correct choices. Finally, the BERTscore of every model will be calculated as well. BERTscore attempts to compute similarity by calculating a score comparing each token in the candidate sequence with each token in the reference sequence. This score is calculated using contextual embeddings, achieving semantical significance and qualitative data that is closer to human judgment. These 3 metrics were chosen strategically to make sure that the main aspects of Natural Language Processing are quantified, intending to provide a meaningful comparison and complete the objective of the research.

Results

The experiment's results include the text generated by the different sequence-to-sequence algorithms (from which a sample will be presented) and a quantitative analysis using the evaluation metrics for sequence-to-sequence models. Additionally, the mean was calculated, and a boxplot was generated for every model and metric to visualize the quantitative results visually.

Base Sentence vs Reference Sentence Comparison

Base Sentence	Reference Sentence
When Hector heard this he was glad, and went about among the Trojan ranks holding his spear by the middle to keep them back	and they all sat down at his bidding
Ulysses went back to his own place, and Eumaeus strewed some green brushwood on the floor and threw a sheepskin on top of it for Telemachus to sit upon.	then the swineherd brought them platters of cold meat the remains from what they had eaten the day before
There are twelve chief men among you, and counting myself there are thirteen	contribute each of you a clean cloak a shirt and a talent of fine gold
His comrades then lifted him off the ground and bore him away from the battle to the place where his horses stood waiting for him at the rear of the fight with their driver and the chariot	these then took him towards the city groaning and in great pain
Even so, O Melanippus, did stalwart Antilochus spring upon you to strip you of your armour	but noble Hector marked him and came running up to him through the thick of the battle
Meanwhile great Ajax kept on trying to drive a spear into Hector	but hector was so skilful that he held his broad shoulders well under cover of his ox hide shield
Jove's daughter Venus answered, Juno, august queen of goddesses, daughter of mighty Saturn	say what you want and I will do it for at once
Hear me, O King, whoever you may be, and save me from the anger of the sea-god Neptune, for I approach you prayerfully	any one who has lost his way has at all times a claim even upon the gods
When the god had accomplished the deed of love, he took her hand in his own and said	tyro rejoice in all good will the embraces of the gods are not fruitless
When the child of morning, rosy-fingered Dawn, appeared, the sons of Autolycus went out with their hounds hunting, and Ulysses went too	they climbed the wooded slopes of Parnassus and soon reached its breezy upland valleys

Table 1. Sample comparison between the base sentence and the reference sentence extracted from Homer's books. Made by Emilio Medina.

The foregoing table shows a sample of the organization of the test data. For the train data, there was not any formatting involved, all the data was directly extracted from Samuel

Butler's translation, whereas for the test data, every 2 sentences fed were divided into a base sentence and a reference sentence. As mentioned in the methodology, the division was made manually, and a criteria that tried to ensure randomness and a great variety of sentences was applied. The reference sentence was fed as ground truth data on each of the metrics employed, and it was used to achieve quantitative data for each of the models' predictions.

Reference Sentence vs Models' Predictions Comparison

Reference Sentence	RNN Prediction	LSTM Prediction	GRU Prediction	GPT3 Prediction
and they all sat down at his bidding	to fight on this the son of atreus	from the sun went vulcan near them in	to try and pacify his mother juno the	Hesitiments in the son of Atreus were high, and his fathers was great. He was a strong man, and had that look in his eyes. He was strong with his left hand and with his right.
then the swineherd brought them platters of cold meat the remains from what they had eaten the day before	you back for she had said he be it in your tent and at the ransom for you are	them giving every man his drink offering out of his feast most his mast and made fast the wide	the achaeans are minerva had fight not first to fight nor gnawed but let you a prize instead or	So it is, he cried, and so it is, and he sat down again. But now, he added, there is another way. How? There is another way. But how? What
contribute each of you a clean cloak a shirt and a talent of fine gold	a hecatomb to whom the god and will with him of all the city of	i shall sent me to bring you back your child and to offer sacrifice to	and his mother heard his shoulder and the arrows rattled on his back with the	Dread me? I have followed you. Dread me? I mean it. I mean it. Dreads brows were set in thought. You mean you believe in ghosts, don
these then took him towards the city groaning and in great pain	son of peleus mighty warrior and at the trojans for the host	rest of his sword roasted them till they were done and drew	tents and rowed the ship to the place where they would have	-riding. You men have got it hard, Sir Battista, when we have got to go back and free him of his people. odynam But we have got to go back and free him of his people

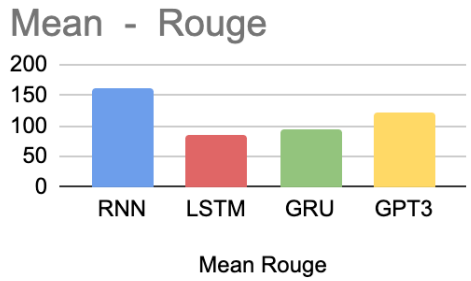
but noble Hector marked him and came running up to him through the thick of the battle	heart in them for they had had way in rage for the present shore and they heard	prayer and do a ransom for you but i shall come to look and get for we	king and by offer sacrifice and if you ever help your prayer and let your arrows avenge	and give you the sceptre. But you have a sceptre. Now, I shall have to ask you to come with me, if I shall come without it. I shall come without you, if you will.
but hector was so skilful that he held his broad shoulders well under cover of his ox hide shield	the trojans for you know my daughter nor in for you would shall not with you in all your	the old and they been see the mast ashore high and dry upon the sands set her strong props	jove right she was live that you must find me a prize instead or i alone among the argives	Therefore, I will go round and round in youraspberry but my son is still young, he will go round and round in yourrussian he will go round and round in yourrussian he will not go round and round in yourr
say what you want and I will do it for at once	taking a man were i will to the people do that been	got out him and hear your anger of dreams born and bred	therefore and the son of atreus who has neither the son of	of Saturn, seer-at-nibbled with Atreus, had fashioned the heaven from the ground up, and from its depthsrawling in his anger, he set the house of Olympus on this errand, toihil
any one who has lost his way has at all times a claim even upon the gods	have her offered laid him for you are to olympus and on their hands as he was	not not how to help you is help it is no railing at me with the side	to their sacrifice from the ground and chose a crew of twenty oarsmen as he sped onward	I will not take that which you have done. I will not ask for the ransom that you will pay me. I will not ask for what you have done. I will not ask for what you have done.
tyro rejoice in all good will the embraces of the gods are not fruitless	you in their own mist him girl his bow and his quiver upon his	out and of the cunning workman vulcan began to try and pacify his mother	of the argives and the woman was loth to go and in the house	Wife, child, help me get him a drink of water. I can't, I've got to go to the store. You are always going to go back there yourself. I will

they climbed the wooded slopes of Parnassus and soon reached its breezy upland valleys	people men not take about this girl for you are for any a prize	good for the host of the achaeans have hold him a second time as	had must for him the son of atreus or to restrain himself and check	robed in thought, by the grace of the god Silt,ettle down, and take care of it.-The Odyssey, Odyssey (3ds. by Vsevolus, 2013) Dawn, son of morning

Table 2. Sample comparison between the ground truth data and each of the models' predictions. Made by Emilio Medina.

The past table portrays the sample predictions for a portion of the test data that was analyzed during the experiment. The first and most visible difference among them is the fluctuation in outputs between the many-to-one models against the many-to-many transformer model, given that this last one offers predicted sentences of varying sentence length. Furthermore, despite the fact that most of the used architectures are similar to an extent, the slight differences in the type and amount of gates produce absolutely different outputs when predicting the same sentence. Additionally, it is important to mention that the training time for the many-to-one models was about 30 minutes per model, unlike the transformer model, which trained the fed data in a matter of a few seconds. Nonetheless, this is a pre-trained model, and the previous training had to take a lot of time and a big database, which is expensive and time-consuming.

Regarding the quality of the predictions, using a human-based metric (or qualitative observations), the transformer model produced an overall higher quality of predictions. Unlike the other models, the predicted sentences of the transformer model respect English grammar and writing conventions to a greater extent and include more features that align with Homer's style, such as an archaic vocabulary and a high degree of detail and explanation within the imagery.



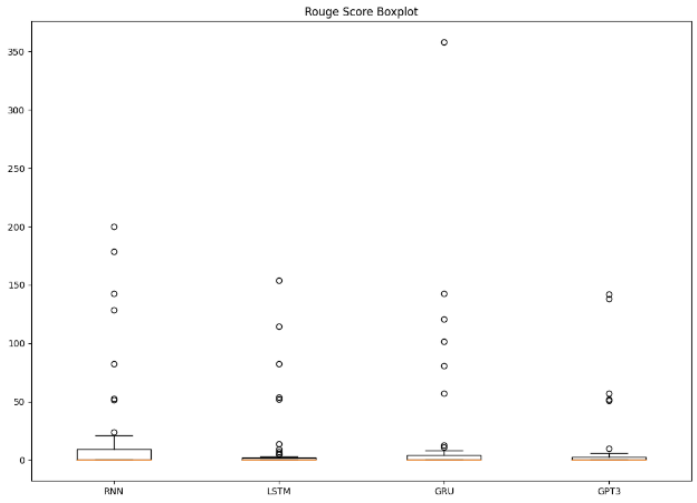
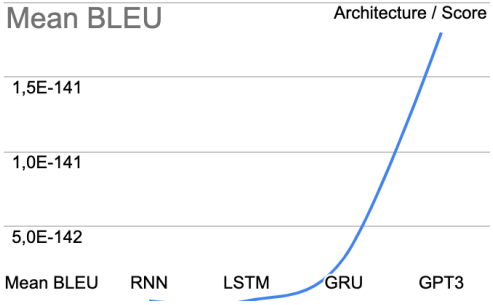


Figure 1. Mean Rouge score graph for every model.
Made by Emilio Medina

Figure 2. Rouge Score Boxplot for every model.
Made by Emilio Medina

These graphs show the difference in Rouge score between the 4 models. As seen, the model with the highest average on the test data is the RNN, with a mean score of 161.48. The model with the lowest mean is the LSTM, with a mean score of 85.11. The boxplot also indicates that the RNN had the highest mean but exposes the imprecision of the predicted scores for each of the models, especially for the RNN.



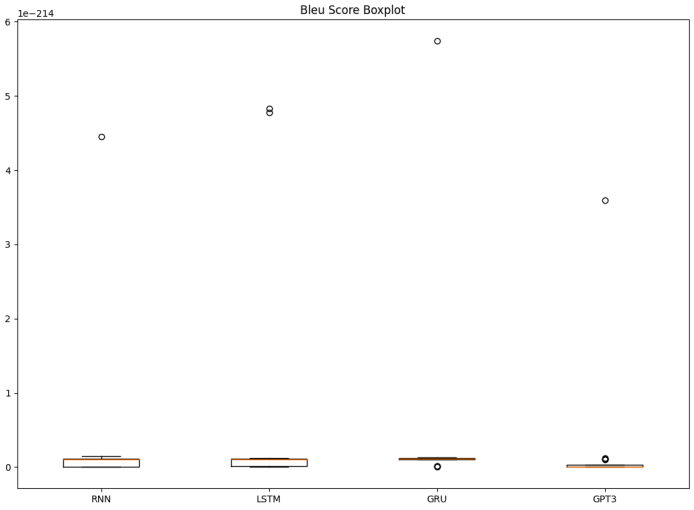


Figure 3. Mean BLEU score graph for every model.
Made by Emilio Medina

Figure 4. BLEU score Boxplot for every model.
Made by Emilio Medina

These graphs portray the difference in BLEU score for every analyzed model. As seen in the mean graph, the GPT-3 model has the highest mean BLEU score, while the RNN has the lowest mean. Nonetheless, the shown scores are negligible, as they are all extremely close to 0, meaning that all the models performed poorly regarding n-gram precision, and the observed trends only show very slight differences between the models. Additionally, there are very big outliers in the data, as shown in the boxplots, expressing how only a minimal amount of the sample predicted sentences had a decent BLEU score.

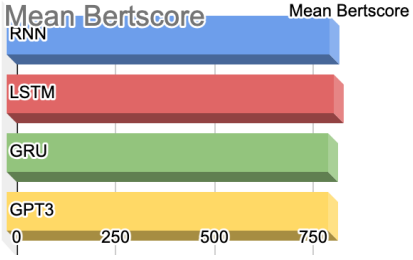


Figure 5. Mean Bertscore graph for every model.
Made by Emilio Medina

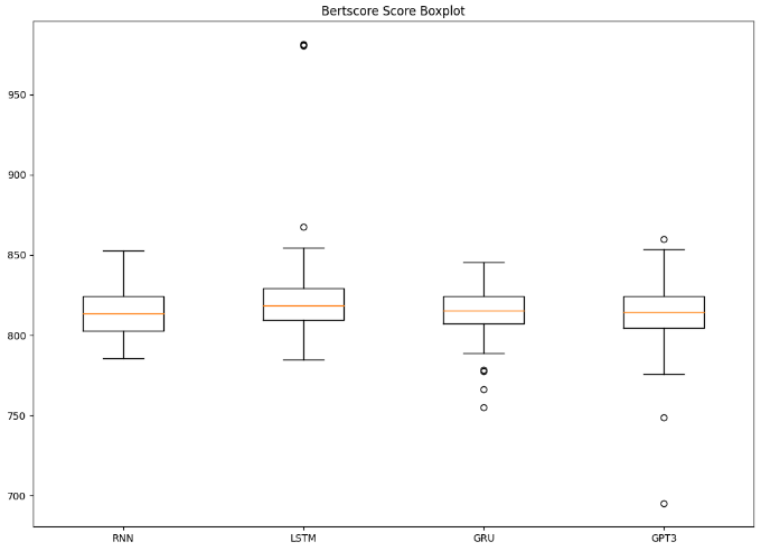


Figure 6. Bertscore Boxplot for every model.
Made by Emilio Medina

The past graphs show the differences in Bertscore between the models. As seen, all the models performed very similarly to each other, given that the mean scores are very near and the box plots are relatively in the same position Nonetheless, the model with the highest mean Bertscore is the LSTM with 827.965 points, and the model with the lowest mean Bertscore is the GPT3 with an 813.027 score (less than 1 point away from the GRU mean Bertscore). The Boxplot graph expresses the consistency, accuracy, and precision of the results, given that they are all quite high and close to each other, suggesting that they constantly offer a contextually correct prediction.

Discussion

The results from the experiment suggest several interesting findings that portray the performance of sequence-to-sequence models but also highlight the limitations of the experiment and the available quantitative metrics to evaluate this kind of task. Firstly, for the Rouge score, the model with the lowest complexity achieved the highest score. As mentioned in the methodology, the Rouge score is useful to know how consistent is the token recall within a sequence. This means that the most simple algorithm, the RNN: that one without any gates and which usually suffers from the vanishing gradient problem (or a short-term memory), achieved the highest score. Although this sounds counterintuitive, it is the nature of the task and of the metric that causes these unexpected results. In essence, for the many-to-one models, the whole experiment consists of which model can predict the most common words more consistently, a task whose importance is increased when the Rouge score is measured since it increases as the number of equal words also increases. This happens because the test data is sentences that the models have never encountered, and given that they do not count with any contextual

embedding or any form of achieving an understanding of the training data, their training process just helps them to know which words are the most optimal in all of the presented situations. Nonetheless, these unencountered sentences are situations that are not present in the training data, causing the models that you would expect to perform the best, which are the most complex (in this case, the LSTM), to not perform as great, given that they try to predict more unexpected words in their attempt to maximize efficiency. If the task were to predict the next sentence of data that has been already trained, similar to a supervised learning situation, then these complex algorithms would indeed very likely be the ones with the highest rouge score; nonetheless, this is not the situation, and when encountered with unknown sequences it is the least complex algorithms that perform the best because they learn the common words, such as prepositions or pronouns which are likely to appear in any sentence of the same author with the same writing style. Therefore, even though the RNN predictions are not the most sophisticated or accurate, they are the ones with the highest Rouge score because the inclusion of the most common words is more prominent in that sequence-to-sequence model. This interpretation of the results can be confirmed with the comparison of sample sentences. For instance, for the reference sentence of “and they all sat down at his bidding,” the RNN predicted “to fight on this the son of atreus”, the GRU predicted “to try and pacify his mother juno the” and the LSTM predicted, “from the sun went vulcan near them in”. Even though none of the models achieved a high rouge score for this sentence, it is clear that the RNN goes for a more conservative approach which maximizes this score with the prediction of common words such as “to”, “on” or “of”, common verbs as “fight” instead of “pacify” and more present characters in the story, which in this case is “atreus” instead of “vulcan” or “juno,” who are less present in Homer’s pieces.

Even though the RNN was the model with the highest rouge score, the GPT-3 had a considerably higher score than the two other models. These scores have to do with the fact that the Transformer model had pre-trained information, and there were occasions where it was able to identify the context of the sentence and give a very accurate usage of words which maximized the rouge score since it knew where the sentence came from. This happened on several occasions, and that increased the rougescore of the GPT-3; nonetheless, it did not happen every time, showing that the conservative approach from a simple algorithm was better than the “all or nothing” strategy of the GPT-3, which was better than the blind word selection of the complex many-to-one architectures. This approach from the GPT-3 can be confirmed with one of the reference sentences, where the prediction identified the origin of the reference sentence: “Wife, child, help me get him a drink of water. I can't, I've got to go to the store. You are always going to go back there yourself. I will robed in thought, by the grace of the god Silt,ettle down, and take care of it.-The Odyssey, Odyssey (3ds. by Vsevolus, 2013) Dawn, son of morning”, nonetheless, in those occasions were it did not identify the context it predicted a style which was not close to Homer, and in some occasions even nonsensical: “Therefore, I will go round and round in youraspberry but my son is still young, he will go round and round in yourrussian he will go round and round in yourrussian he will not go round and round in yourr”. Overall, these results show that the rouge score is useful to identify the efficiency in supervised learning styled tasks, where the context or the style of the writer is not really important, and it is rather the repetition of common tokens that matters in these tasks, making the least complex models and those with prior knowledge on the context the most effective. In a real-world context, it suggests that although RNNs may not be the most precise for NLP tasks, sequence-to-sequence models, and especially RNNs, can be applied to other tasks, including

classification tasks, where non-complex models can excel or tasks that include a small training corpus.

As for the BLEU score, there was a general trend of increasing score as complexity increased, being the GPT-3 the model with the highest BLEU score. As mentioned previously, the BLEU score measures the precision of the vectors, meaning that it prioritizes not only the presence of a token but also its position in the sequence. This approach of giving a score is what explains the true meaning obtained from the scores, which is that all the models performed poorly regarding token precision. Even though there seems to be a trend related to complexity, it can be neglected, given that all the values are virtually 0. What this shows is firstly an underwhelming performance of the models in predicting exact sequences; nonetheless, it is unfair to expect high results and a perfect sequence prediction when the models have never seen the test data before, and they cannot exactly predict something they have not seen. The only model that could have achieved this is the GPT-3, and although its score is low it is not a surprise that it is the architecture with the highest score as it is the only model which could have had access to the reference sentences. The negligible score evidences that for the evaluation of a text generation task, an automated metric, which also considers precision is not an optima metric. It suggests that both Celikyilmaz in “Evaluation of Text Generation: A Survey ", and Turing in 1950, with his statement that gave birth to Computer Science, proposed that a machine can be considered intelligent if it can fool a human of whether it is a machine or not. Both of these authors claimed that a human-centered evaluation technique is useful to assess sequence-to-sequence tasks, and according to the results in the case of text generation, it proved to be true. It is not fair to give 4 different models a score close to 0, and almost incomparable because they do not match with a reference sentence, given that the task of the architecture is not to copy the author but the style of the author, and potentially imitate it creating pieces which the author never even thought about but thanks to AI now have his essence. Thus, the BLEU score is a futile tool for the evaluation of text generation tasks, and it should be replaced with human-centered evaluation techniques that have proven to be effective in algorithms such as OpenAI’s Chat GPT.

Additionally, the results of the Machine-Learning metric, BertScore, showed very similar results between the 4 models, placing LSTM as the best model and the GPT3 as the worst one. Bertscore is a metric that uses contextual embeddings to assess the semantic significance of each generated sentence compared to the reference sentence. Given the nature of the task, the first thing that can be noted is that it is the most useful metric employed given that, unlike the previous metrics, it does not assess new sentences based on a zero-sum technique, promoting the imitation of the reference sentence; instead, it looks for semantic similarity, which is a way of imitating a writing style. There is a positive correlation between the many-to-one models, portraying the expected results given that the most complex model (LSTM) was the architecture that portrayed the writing style the best since it had the most semantic significance. This differs from the results obtained with the Rouge score since this metric does not reward the repetition of common words and instead prioritizes the relationship that is sought after – a logical one. Nevertheless, if complexity and semantic significance are rewarded, the Transformer model was expected to have the greatest Bertscore, yet, it has the lowest one. Firstly, it is important to note that this result is influenced by outliers since, as seen in the Boxplot, 2 values negatively exceed the minimum value. This influences the lower Bertscore; nonetheless, these outliers are caused

by the nature of the experiment and are related to the reason why the RNN performed so well with the Rouge score, which is the fact that a small corpus of data was used. As mentioned in the introduction and methodology, this experiment used a relatively small amount of data since the GPU that trained the models could not handle larger quantities of raw data due to the complexity of some of the architectures. Thus, a model that on paper should have performed better than the rest of the models regarding semantic similarity was not able to do so because it was "undertrained," and as seen in some of the sample results in some occasions, it produced outputs that have little to no similarity with Homer's writing style. This is an important consideration when choosing which architecture to choose because if the outliers are ignored and more complex algorithms are chosen for text generation tasks if the data is not enough, as with this experiment, there will be negative effects since the architectures will not have enough context to make predictions and these will be totally unrelated.

Overall, the experiment considered a wide range of variables which made the results inconsistent and unexpected, but this lack of quality of results is useful for drawing alternate conclusions regarding the metrics of sequence-to-sequence models To begin with, as said when analyzing the Rouge score and BLEU score, although less complex models seemed to be more effective this was due to two main reasons. Firstly, the usage of a small corpus of data negatively influenced the performance of the more complex models and enhanced the simpler ones. Secondly, the fact that these two scores, despite being two of the most popular, are not useful to evaluate the relationship to a writing style. The Rouge and BLEU scores reward similarity with a reference sentence, giving the most points to sentences that mimic what is expected. Nonetheless, text generation involves the creation of something new, something unseen in the training data, and thus, it should not copy the reference sentence because what it tries to mimic is its style or semantic significance, not the actual tokens of the sentence. Since Rouge and BLEU do not reward new sentences, their result is misleading when evaluating text generation, and it is fundamentally flawed to assess this kind of task using Rouge and BLEU scores. Moreover, the need for human-centered metrics proved to be required due to the lack of useful automated metrics. These metrics could be further developed for specific tasks to decrease bias and combined with AI-powered metrics such as Bertscore, which are the best available metrics for a quantitative score. This statement shows a lack of available metrics and an opportunity for research where metrics that use different methods to reward a similarity, in this case, writing style, can be developed since it is necessary to have quantitative data to determine which models work best in different situations.

Limitations

After analyzing the results, it is safe to say that there is one very specific decision that affected the outcome of the experiment and the evidence to comfortably draw conclusions based on the data. This decision is the use of a small corpus of data for both the training and testing steps, given that this negatively influenced the quality of the more complex models. Although the data helped to disprove the effectiveness of automated metrics for text generation tasks, it is not totally clear whether more complex models, such as the Transformer model, are better in text generation tasks. To draw this conclusion, a human-based metric had to be employed, which is undesirable given that the research aimed to achieve quantitative results.

Thus, to repeat the experiment and achieve clear evidence that can prove or disprove the hypothesis, a bigger corpus of data that enhances the quality of the complex models is required, yet, it can only be attained with the use of better technology, such as a GPU that can bare with the complexity of the models.

Conclusion

In conclusion, sequence-to-sequence models proved to be effective in text-generation tasks, yet, it is complicated and ambiguous to draw conclusions based on the available quantitative metrics given that they reward what should not be assessed in text-generation tasks. Automated metrics present a fundamental discrepancy given that generative models are supposed to produce new things; nonetheless, automated metrics measure the similarity with the ground truth instead of the generative capabilities. Furthermore, the decision of what algorithm is best depends on the context of the situation, and for this paper, it was shown that using a small corpus of data the less complex algorithms, especially the RNN, achieve a higher consistency of quality results. Yet, if metrics that value semantic significance, such as AI-powered metrics or human-based metrics, are used, the more complex models proved to provide higher-quality results. Nonetheless, to confirm this assumption with quantitative data a better model with the appropriate amount of data must be used, and the fact that contextual embeddings are an indicator of correlation with a writing style must be confirmed.

Acknowledgments
Joe Xiao

Bibliography

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. “Evaluation of Text Generation: A Survey.” *arXiv* 2 (June 26, 2020). https://arxiv.org/pdf/2006.14799.

G. Aalipour, P. Kumar, S. Aditham, T. Nguyen and A. Sood, "Applications of Sequence to Sequence Models for Technical Support Automation," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4861-4869, doi: 10.1109/BigData.2018.8622395

GeeksforGeeks. “Box Plot in Python Using Matplotlib.” GeeksforGeeks, March 8, 2022. https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib/.

———. “Next Word Prediction With Deep Learning in NLP.” GeeksforGeeks, March 13, 2024. https://www.geeksforgeeks.org/next-word-prediction-with-deep-learning-in-nlp/.

“Inference Parameters - Amazon Bedrock,” n.d. https://docs.aws.amazon.com/bedrock/latest/userguide/inference-parameters.html.

Karpathy, Andrej. “The Unreasonable Effectiveness of Recurrent Neural Networks.” Andrej Karpathy Blog, May 21, 2015. https://karpathy.github.io/2015/05/21/rnn-effectiveness/.

M, Siddharth. “Long Short Term Memory: Predict the Next Word.” Analytics Vidhya, February 26, 2024. https://www.analyticsvidhya.com/blog/2021/08/predict-the-next-word-of-your-text-using-long-short-term-memory-lstm/.

Moses, Kriz. “Encoder-Decoder Seq2Seq Models, Clearly Explained!!” *Medium*, January 7, 2022. https://medium.com/analytics-vidhya/encoder-decoder-seq2seq-models-clearly-explained-c34186fbf49b.

Mudadla, Sujatha. “NLP Model Metrics - Sujatha Mudadla - Medium.” *Medium*, December 14, 2023. https://medium.com/@sujathamudadla1213/nlp-model-metrics-b3fa32373269.

Sutskever, Ilya, Oriol Vinyals, and Quoc Le. “Sequence to Sequence Learning With Neural Networks.” *arXiv* 3 (September 10, 2014). https://arxiv.org/abs/1409.3215.

Team, AIContentfy, and AIContentfy Team. “Exploring Text Generation Models: A Comprehensive Overview.” AIContentfy, September 1, 2023. https://aicontentfy.com/en/blog/exploring-text-generation-models-comprehensive-overview.

Classics MIT. “The Internet Classics Archive | Works by Homer,” n.d. https://classics.mit.edu/Browse/browse-Homer.html.

Yanhui, Chen. “A Battle Against Amnesia: A Brief History and Introduction of Recurrent Neural Networks.” *Medium*, January 7, 2022. https://towardsdatascience.com/a-battle-against-amnesia-a-brief-history-and-introduction-of-recurrent-neural-networks-50496aae6740.

Y. Deng, L. Wang, H. Jia, X. Tong and F. Li, "A Sequence-to-Sequence Deep Learning Architecture Based on Bidirectional GRU for Type Recognition and Time Location of Combined Power Quality Disturbance," in IEEE Transactions on Industrial Informatics, vol. 15, no. 8, pp. 4481-4493, Aug. 2019, doi: 10.1109/TII.2019.2895054.

Yong Yu, Xiaosheng Si, Changhua Hu, Jianxun Zhang; A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Comput 2019; 31 (7): 1235–1270. doi: https://doi.org/10.1162/neco_a_01199

Ysthehurricane. “Next Word Prediction BI-LSTM Tutorial Easy Way.” Kaggle, August 15, 2021. https://www.kaggle.com/code/ysthehurricane/next-word-prediction-bi-lstm-tutorial-easy-way/input.

Zahra Fayyaz, Aya Altamimi, Carina Zoellner, Nicole Klein, Oliver T. Wolf, Sen Cheng, Laurenz Wiskott; A Model of Semantic Completion in Generative Episodic Memory. Neural Comput 2022; 34 (9): 1841–1870. doi: https://doi.org/10.1162/neco_a_01520

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "BERTScore: Evaluating Text Generation With BERT." *arXiv (Cornell University)* 3 (April 21, 2019). https://arxiv.org/pdf/1904.09675.pdf.