Replication: Longformer and Summarization evaluation
Team: Quirky Nation

Alekhya Pyla
apyla

Vignatha Yenugutala
vyenugut

Sai Gopal Athili
sathili

Abstract

This is a replication study of the Longformer paper(I. Beltagy and Cohan. (2020)). Traditional transformers limit the input tokenization to 512 tokens. Longformer solves this problem, enabling long document seq-to-seq tasks by modifying positional embeddings. In this paper, we experimented with Longformer Encoder Decoder, a variant of LED. An analysis of the performance on the summarization task of the arXiv dataset(Arman Cohan and Goharian (2018)) was done. Furthermore, analysis of summarization performance is done using automatic metrics like ROUGE, Chrf Bleu, Rouge-we, and quality dimension metrics like coherence, relevance, fluency, and consistency. The quality dimensions are derived for the results on Arxiv data by building a regressor on the results from the SummEval(Fabbri and Radev (2021)) paper. The metrics that were considered for the regression are Rouge, Rouge-we, Chrf, Bleu, and Meteor. The important metrics that match with high-quality dimension values were discussed. Finally, a qualitative analysis of the LED-generated summaries was detailed to show how the automatic metrics can be misleading in terms of summarization tasks.

1 Introduction

Transformers have achieved state-of-the-art results primarily due to the self-attention mechanism. Though very powerful, the self-attention mechanism has a computational scales quadratically in time and memory with the sequence length, inhibiting the transformer-based models to the process long sequences. To address this limitation, the Longformer with attention, a mechanism is introduced that scales linearly with sequence length, making it is easy to process documents of thousands of tokens or longer. This is an advantage for natural language tasks such as long document summarizations where existing approaches partition or shorten the long context into smaller sequences. This could potentially result in a loss of important cross-partition information(solved by targeted complex architectures).

For supporting long document generative sequence-to-sequence tasks such as summarization, an Encoder-Decoder model variant of Longformer, called LED(Longformer-Encoder-Decoder) was introduced in the paper. In our work, we intend to demonstrate its effectiveness in the summarization of scientific papers dataset of ArXiv to generate concise and coherent abstracts and aim to replicate the results in this paper. Also, the scarcity of comprehensive up-to-date studies on evaluation metrics for text summarization and the lack of consensus regarding evaluation protocols continue to inhibit progress. We try to demonstrate these shortcomings using analysis done in SummEval (Fabbri and Radev, 2021). We aim to prove that the currently available set of summarization metrics does not ensure high-quality summaries.

We start our evaluation by generating test data arxiv summaries(6.4k) using LED. We evaluated these summaries using Rouge-1,2 and L scores to verify our replication. We extended this evaluation set to Rouge-3,4, Rouge-we-1,2,3, Meteor, and Bleu metrics. We then implemented a regressor to produce quality dimension metrics to further evaluate these longformer summaries. Finally, we produce a qualitative analysis showing that neither human-annotated nor automatic metrics can perfectly evaluate a summary.

2 Related Work

There have been earlier models which tried to implement some form of long transformers like TXL(Zihang Dai and Salakhutdinov (2019)) Transformer XL. But these models have only been successful in autoregressive language modeling and are not bidirectional. The model with the most similar attention pattern to that of the Long Transformer is the Sparse Transformer(Rewon Child and

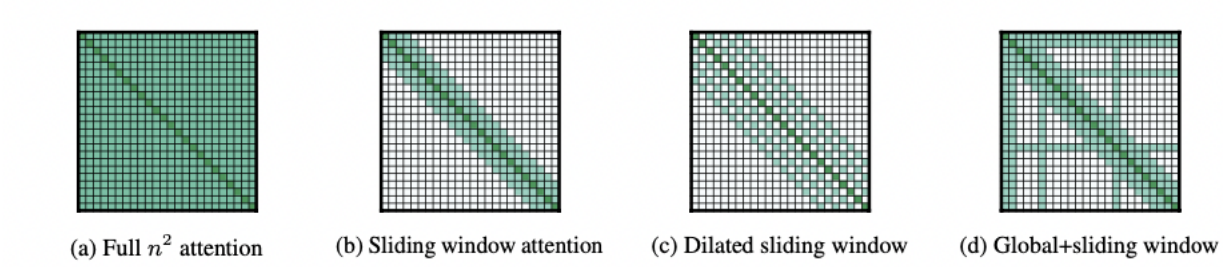

Figure 1: Figure from I. Beltagy and Cohan. (2020) - Comparing the full self-attention pattern with the configurable attention patterns in Longformer.

Sutskever (2019)), which uses a form of the dilated sliding window of blocks of size 8x8. Another contemporary work similar to the Long Transformers is Big Bird: Transformers for Longer Sequences(Manzil Zaheer and Ahmed (2020)) which also uses a sparse attention mechanism to handle contexts of up to 8x of conventional transformers.

To cater to the analysis of the summarization, SummvEval(Fabbri and Radev (2021)) paper is used as a reference that focuses on developing evaluation metrics that better correlate with human judgments. We follow the human evaluation protocol from Sebastian Gehrmann and Rush (2018) involving relevance, consistency, fluency, and coherence. There are also previous similar works that also focus on human evaluation on summarizations like Wojciech Kryscin (2019).

Taking inspiration from this work, we have decided to take the results of the LED model on Arxiv data and evaluate them using a regression model on the data provided by the Summeval paper.

3 Approach

3.1 LongFormer

The original Transformer consists of an encoder-decoder architecture which is used for the sequence to sequence tasks such as summarization and translation. It has a self-attention component with $O(n^2)$ time and memory complexity where n is the input sequence length. To address this, the full self-attention matrix is sparsified according to an attention pattern that scales linearly with the input sequence, making it efficient for longer sequences, unlike the full self-attention. For these, sliding window and dilated sliding window attention were proposed.

The windowed and dilated attention are not flexible enough to learn task-specific representations. Hence, global attention to a few pre-selected input locations was added to the paper. Figure 1 shows these attention patterns, also an example of sliding window attention with global attention at a few tokens at custom locations. Since the number of such tokens is small relative to and independent of n the complexity of the combined local and global attention is still O(n).

Rouge-1	Rouge-2	Rouge-L
0.4663	0.1962	0.4183

Table 1: Rouge Scores of Longformer model as mentioned in paper

For modeling long sequences for the sequence to sequence learning tasks, a variant of this long former which has both encoder and decoder was proposed in the paper. The encoder reads the document and its decoder generates the output summary. This model is Longformer - Encoder-Decoder (LED) and it scales linearly with the input. Since pretraining LED is expensive, we initialize LED parameters from BART, and BART's exact architecture in terms of the number of layers and hidden sizes. The results of this LED model on Arxiv data as mentioned in the paper are given in table 1 .

3.2 SummEval

SummEval paper provides a comprehensive, up-to-date study on the evaluation protocols by performing a summarization analysis on 23 different models. The dataset used for summarization is CNN Daily Mail. The outputs are evaluated using-automatic metrics like ROUGE, BLUE, etc, and human evaluation. The paper highlights the correlations between the both.

3.2.1 Automatic Metrics

Many automatic metrics have been proposed in the paper for evaluating both summarization and other text generation models. The metric focused in our

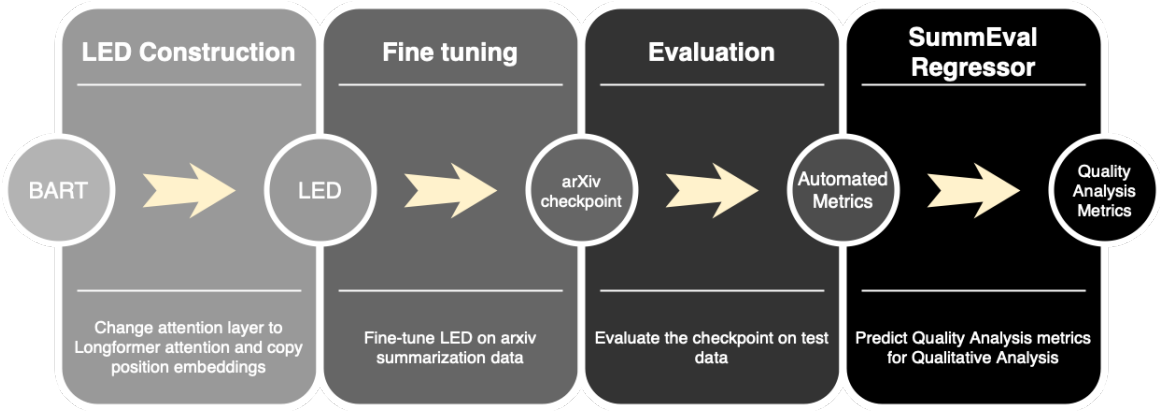

Figure 2: Complete flow diagram of LED analysis.

experiments is ROUGE, METEOR, BLUE, and CHRF as these metrics correlated positively with human evaluation in the SummEval paper.

ROUGE (Lin, 2004), measures the number of overlapping textual units (n-grams, word sequences) between the generated summary and a set of gold reference summaries. ROUGE-L measures the longest common subsequence (LCS) between our model output and reference. ROUGE-WE (Ng and Abrecht, 2015) is an extension of ROUGE where soft lexical matching based on the cosine similarity of Word2Vec embeddings is used.

Further, CHRF (character n-gram F-score) (Popović, 2015) calculates the character-based n-gram overlap between model outputs and reference documents. METEOR Lavie and Agarwal (2007) calculates the alignment (harmonic mean) of unigrams in the generated summary to 0 or 1 unigrams in the reference, based on stemming, synonyms, and paraphrastic matches. BLEU (Papineni et al., 2002) is a corpus-level metric that calculates n-gram overlap between generated and reference summary by including a brevity penalty.

3.2.2 Human Evaluation

The Human Evaluation protocol for Evaluating Summarization from Sebastian Gehrmann and Rush (2018) consists of four metrics:
1.Relevance - Selection of important content from the source.
2.Consistency - Factual alignment between the summary and the source.
3.Fluency - Quality of individual sentences.
4.Coherence - Collective quality of all sentences.

For every generated summary, a mixture of experts and crowd-sourced annotations are collected for each of the four metrics shown above, to ensure high-quality annotations. We will call these metrics quality analysis metrics from now on as these will not be generated manually in our work.

3.2.3 Correlation between Human and Automatic Evaluations

Kendall's tau rank correlations between automatic metrics and human judgments are calculated following (Louis and Nenkova, 2013). Out of all 4 metrics, it is observed that correlation within the coherence dimension is the minimum. This could be because metrics like ROUGE, often rely on the subsequence of elements like ngrams and do not measure the interdependence between consecutive sentences. Similarly, low and moderate correlation scores were found for the relevance dimension. This could be due to the subjectivity of defining relevance, which can vary for human annotations. The highest correlations are observed for consistency and fluency. The strong correlation with consistency could be attributed to the low abstractiveness of most neural models. For coherence, the highest correlation is observed with Rouge-4, Rouge-we-1, and CHRF and for consistency, it is with Rouge-3 and Meteor. For fluency, the highest correlation is with Rouge-1, Rouge-4, Rouge-we-1, and Meteor and for relevance, it is with rouge-we-1, chrf, and meteor. Thus with these observations, we have decided to use the rouge, rouge-we-1, bleu(most common), chrf, and meteor metrics for qualitative analysis.

4 Experiment Setup

4.1 Data Set

Evaluation of LED is performed on the arXiv summarization dataset (Cohan et al., 2018). The data set is a collection of scientific papers varying across

multiple domains like physics, medicine, etc. Every data point consists of the paper, the corresponding abstract, and section names. The task aims to summarize the given scientific paper and generate the abstract. The evaluation is performed on the test set, which consists of 6440 articles. The 90th percentile document size in terms of the number of tokens(words) in the test set is 10500, therefore making this a good data set to measure the effectiveness of long document summarization using LED.

4.2 Replication on arxiv data

We used the longformer encoder-decoder for this replication with the encoder having an attention window size of 1024 tokens. Using the pretrained model, publicly available as allenai/led-large-16384-arxiv, the summarization was performed. The summarization task along with the metric computations required a GPU compute of close to 24 hours for all the 6400 test samples. Considering such high computational requirements, fine-tuning of the model cannot be achieved with the resources at hand. The automatic metrics used to evaluate summaries are ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, CHRF, METEOR, BLEU, ROUGE-WE-1, ROUGE-WE-2, and ROUGE-WE-3.

4.3 Training Regressor for SummEval human metrics

Metrics	Coherence	Consistency	Fluency	Relevance
Rouge-1	**0.0788**	0.0197	0.0125	**0.0640**
Rouge-2	0.0222	0.0037	0.0123	0.0256
Rouge-3	0.0191	**0.1386**	**0.1386**	**0.0758**
Rouge-4	0.0754	**0.3931**	**0.4611**	**0.0513**
Rouge-L	**0.1793**	0.0384	0.0336	0.0216
Bleu	0.0710	0.0204	0.0289	0.0338
Chrf	**0.2305**	**0.1235**	**0.1150**	**0.3507**
Meteor	**0.1409**	**0.0801**	**0.0811**	**0.2905**
Rouge_we_1	**0.1092**	**0.1142**	**0.0673**	0.0372
Rouge_we_2	0.0430	0.0311	0.0138	0.0244
Rouge_we_3	0.0299	0.0365	0.0353	0.0243

Table 2: The features importance values for Random Forest Regressor

We take the data of Automatic and Quality Analysis metrics for 23 model summaries(originally used for drawing correlations in the SummEval paper) to construct a RandomForestRegressor. This regressor takes in Automatic Summarization metrics and outputs quality analysis metrics. This decreases the overhead of manually annotating these metrics. Since this data extends across a variety of models and articles we assume the generality can be extended to Longformer too. The main reason for this implementation is to provide human-understandable metrics for qualitative analysis.

Metric	Value	Metric	Value
Rouge1	0.4549	CHRF	0.4763
Rouge2	0.1929	METEOR	0.3583
Rouge3	0.1028	Rouge_we_1	0.4622
Rouge4	0.0638	Rouge_we_2	0.3095
RougeL	0.2658	Rouge_we_3	0.3440
BLEU	13.5528		

Table 3: Performance of Longformer-Encoder-Decoder on Arxiv dataset

Coherence	Consistency	Fluency	Relevance
3.7	4.5	4.6	3.9

Table 4: Quality Analysis Metrics from Regressor

5 Evaluation

5.1 LED arxiv results

The longformer-encoder-decoder was tested on 6400 samples of Arxiv data and its performance is mentioned in Table 3. The Rouge2 F1 score of 0.1937 is closer to the performance of 6000 articles mentioned in the paper 1. The rouge scores obtained on 6000 articles are given in the paper as rouge1 - 0.4663, rouge2 - 0.1962, rougeL - 0.4183. The results obtained from our replication are rouge1 - 0.4549, rouge2 - 0.1929, and rougeL - 0.2658. RougeL was observed to be different from the results mentioned in the paper.

The quality analysis metrics are shown in the table 4, where each metric can take values in the range of 1 to 5. It is observed that the model-generated summaries have high consistency and fluency scores. The high consistency score for long document summarization indicates that the Longformer can effectively summarize the document covering the entire span, including tokens past the 512 limit compared to general transformers. The lowest coherence score suggests that the inter-dependency between sentences is not sufficient.

To test the validity of the regressor, we withheld 10% of the data(170 out of 1700 total data points). The mean square error of the regressor calculated using cross-validation was found to be 0.1829. This

Paper 1 Summary
we describe a spectroscopic survey designed to uncover an estimated @xmath040 am cvn stars hiding in the photometric database of the sloan digital sky survey (sdss) .
Paper 2 Summary
we study the detectability of circular polarization in a ... we find that the circular polarization can not be detected for an isotropic background .... gravitational waves .
Combined Summary
we characterize sgwb by the so called stokes @xmath0 parameter and calculate generalized overlap reduction functions (orfs) so that we can probe the circular polarization of the sgwb.

Table 5: Example Summarization where longformer is relevant as it captures context from both sources

Generated Summary	**Gold Summary**
the spectrometer is a very compact magnetic spectrometer suitable especially for the detection of kaons. the spectrometer was recently dismantled at the sis facility at gsi and re - installed in the spectrometer hall at mami. the spectrometer was recently dismantled at the sis facility at gsi and re - installed in the spectrometer hall at mami. the spectrometer was recently dismantled at the sis facility at gsi and re - installed in the spectrometer hall at mami. the spectrometer is a very compact magnetic spectrometer suitable especially for the detection of kaons.	at the institut fr kernphysik in mainz , germany , the microtron mami has been upgraded to 1.5gev electron beam energy . the magnetic spectrometer is now operated by the a1 collaboration to study strangeness electro - production . its compact design and its capability to detect negative and positive charged particles simultaneously under forward scattering angles complements the existing spectrometers . in 2008 kaon production off a liquid hydrogen target was measured at @xmath0 0.050(gev@xmath1)@xmath2 and 0.036(gev@xmath1)@xmath2 . associated @xmath3 and @xmath4 hyperons were identified in the missing mass spectra ...

Table 6: Example Summarization where fluency is high but summary is not appropriate

Generated Summary	**Gold Summary**
we present the results of a population synthesis study of white dwarfs with @xmath0 k. we estimated the masses of all da white dwarfs found by @xcite, @xcite and @xcite among the 4.5 million spectra acquired by the sloan digital sky survey data release 12...	we present the mass distribution for all s / n@xmath015 pure da white dwarfs detected in the sloan digital sky survey up to data release 12 , fitted with koester models for ml2/@xmath1 , and with @xmath2 k ...

Table 7: Example Summarization where relevance, coherence is high but summary is about different star

Generated Summary	**Gold Summary**
we study the three - state - dependent (that is, the state - set contains three states) quantum copying more carefully and generalize the method to the multi - state - dependent cloning process. some upper bounds on the multi - state - dependent quantum cloning process are given.	due to the no - cloning theorem , the unknown quantum state can only be cloned approximately or exactly with some probability . there are two types of cloners : universal and state - dependent cloner . the optimal universal cloner has been found and could be viewed as a special state - dependent quantum cloner ...

Table 8: Example Summarization with low relevance

value is low w.r.t to the quality dimension values used in the regressor which are in the range of 1-5. Hence, the regressor performed decently well.

The feature importance values of the Random Forest Regressor for each of Coherence, Consistency, Fluency, and Relevance were mentioned in the table 2. Since these are feature importance values of Random forest regressor, they are all positive.

5.2 SummEval inferences from Regressor

We have highlighted the top 5 metrics for each of the quality dimension models. The metric, Chrf is observed as important for all quality dimensions. This could be because of the reason that it uses a character level n-gram model, thus less constrained and there would be more such overlaps between the gold and generated summary. Meteor also was observed to be important, this could be because the reason that it uses unigrams which are less constrained than higher ngrams, and it uses stemming, synonyms, and paraphrastic matches. These top metrics agreed with the Summeval paper to a good extent. The difference in these results with the Summeval paper could be the error of the regressor or the interdependency of these metrics with each other.

5.3 Qualitative Analysis

To understand the effectiveness of the LED in summarizing long documents, we experimented by trying to summarize two different papers and observe if LED can coherently capture concepts from both. Both document1 and document2 have sizes of 7k tokens each, with the combined document size of close to 15k tokens. It is observed from Table 5, the combined summary generated by LED consists of components from both the documents, which are sufficiently large size documents.

Table 6 shows the case where the generated summary has repetitive sentences. The regressor gives a fluency score of 4.54 which is a higher value. This shows an inherent flaw in the fluency metric where two generated sentences are fluent but the repetition of the same sentences gives a higher value of fluency. Thus, the coherence, in this case, is 3.3 which indicates that this sample doesn't capture the concise summary well.

Table 7 shows the example where the generated summary and gold summary have high relevance 4.1. In this case, both texts talk about dwarf stars but the stark difference is in the type of dwarf star talked about which is different. Therefore, in such examples, one fundamental fact like the star name, in this case, brings a vast difference to the summary but relevance scores will still be high in such cases. Similarly, the metrics which calculate lexical overlap like ROUGE scores, chrf, and METEOR will also show higher values. This could be one of the main fallbacks of such metrics for evaluating summaries.

Table 8 depicts the case where the summary generated has no relation with the gold summary overall and the relevance value generated of 3.3 depicts this accurately.

6 Conclusions and Future Work

The original paper included creating the variant of transformer for long documents by changing the attention mechanisms from self-attention to local and global attention. One of the main challenges in this work was the high computation requirement of Longformer. The training of this long transformer required 4-8 RTX8000 GPUs for 13-16 days and hence, we were not able to train it with the resources at hand. Also, fine-tuning the model to a downstream task such as arXiv requires training for 4 days. Hence, we utilized the model fine-tuned on arxiv data by the authors and evaluated the results. We implemented text summarization for arxiv articles using this model and obtained rouge scores similar to the ones mentioned in the paper.

Further, we utilized the results obtained on arxiv data to evaluate the summary using quality analysis metrics like Coherence, consistency, fluency, and relevance. We utilized these annotated data from (Fabbri and Radev, 2021) and modeled a regressor on arxiv articles' results and showed how the correlation from (Fabbri and Radev, 2021) is analogous to the feature importance of our trained regressor.

If given more time and resources, we would love to dig more into the attention mechanisms of longformers(trying and tweaking the model without using the fine-tuned checkpoints). Also, we would be interested in incrementing the SummEval (Fabbri and Radev, 2021) work by including longformer as another model for analysis.

References

Doo Soon Kim Trung Bui-Seokhwan Kim Walter Chang Arman Cohan, Franck Dernoncourt and Nazli Goharian. 2018. A discourse-aware attention model for ab-

stractive summarization of long documents. *NAACL-HLT 2018*.

Kryściński W. McCann B. Xiong C. Socher-R. Fabbri, A. and D. Radev. 2021. Summeval: Re-evaluating summarization evaluation. arXiv:2007.12626.

M. E. Peters I. Beltagy and A. Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint*, arXiv:2004.05150.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Kumar Avinava Dubey Joshua Ainslie-C. Alberti S. Ontan o n Philip Pham Anirudh Ravula Qifan Wang L. Yang Manzil Zaheer, Guru Guruganesh and A. Ahmed. 2020. Big bird: Transformers for longer sequences. *ArXiv*, abs/2007.14062.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford Rewon Child, Scott Gray and Ilya Sutskever. 2019. Generating long se- quences with sparse transformers. *arXiv preprint*, abs/1904.10509.

Yuntian Deng Sebastian Gehrmann and Alexander Rush. 2018. Bottom-up abstractive summarization. *2018 Conference on Empirical Methods in Natural Language Processing pages 4098–4109, Brussels, Belgium.*

Bryan McCann Caiming Xiong-Richard Socher Wojciech Kryscin, Nitish Shirish Keskar. 2019. Neural text summarization: A critical evaluation.

Yiming Yang Jaime G. Car-bonell Quoc V. Le Zihang Dai, Zhilin Yang and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv:2007.12626.

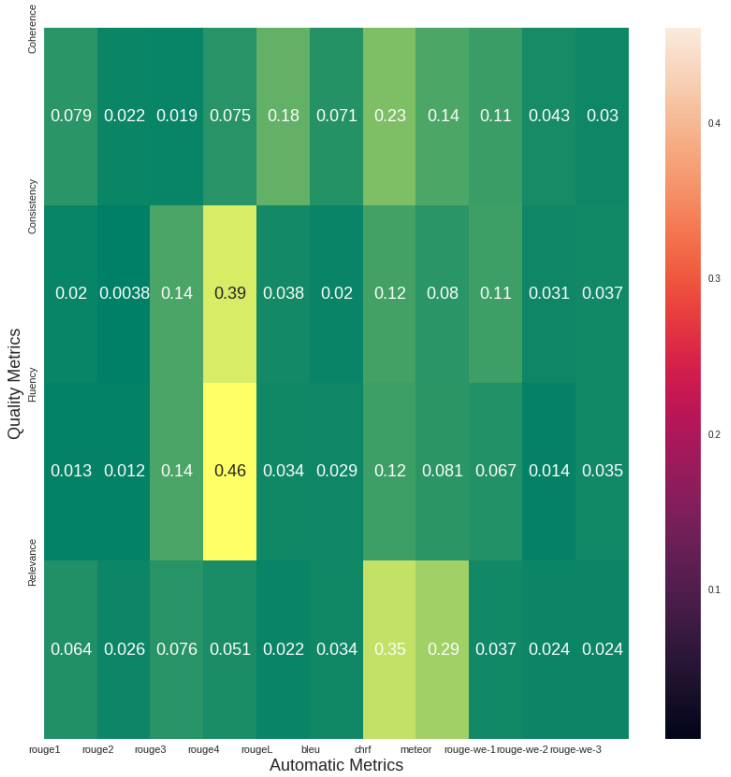

Figure 3: Correlations between quality and automatic metrics

7 Appendix

The code is available at : Project Link