Evaluating Text Generation: Comparative Analysis and Entropy-Weighted BLEU

Boran Lucas
ENSAE Paris
boran.lucas@ensae.fr

Corentin Domergue
ENSAE Paris
corentin.domergue@ensae.fr

Abstract

Automated Story Generation (ASG) is a vital area in Natural Language Processing (NLP) that requires reliable evaluation methods. In this article, we examine various techniques and metrics for evaluating automated text generation quality, such as pre-trained language models and associated metrics like BLEU, ROUGE, and BERTScore. We analyze the performance of different Automatic Evaluation Metrics (AEM) on the MANS and HANNA datasets, considering various text generators and human judgments. We introduce a new variant of the BLEU metric, called Entropy-Weighted BLEU, which is particularly useful for shorter texts but has limitations. Our study highlights the importance of selecting the right metrics for evaluating text generation quality and emphasizes the need for continuous exploration of new evaluation methods. Our code is available on Github. [1]

1 Introduction

Automatic story generation (ASG) (Guan and Huang, 2020) has made significant progress in recent years, as seen in the success of systems like ChatGPT among other (Colombo* et al., 2019; Colombo et al., 2021a; Jalalzai* et al., 2020). These systems use language models like GPT, BertGeneration, Fusion, or TD-VAE to generate narratives from short prompts or sentences (Gregor and Besse, 2018). However, there are still challenges to overcome, such as ensuring controllability, incorporating common knowledge, and promoting creativity.

Evaluating the quality of ASG outputs is a crucial task, and metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang* et al., 2020) have been developed to measure this. However, evaluating these metrics against human judgment can be expensive and time-consuming, leading to the development of automatic evaluation metrics (AEM) that serve as a proxy for human judgment (Colombo et al., 2022a; Colombo, 2021). Recent research suggests that new metrics should complement existing ones, rather than just improving their correlation with human judgment.(Chhun et al., 2022)

One approach to evaluating ASG models involves comparing the generated text with a reference text. For example, the number of deletions and additions made by the author can be used to assess the quality of the model's proposed continuation of a story. While this approach is not universally applicable, it provides an innovative way to evaluate the performance of ASG models.

Another important aspect of ASG research is the evaluation of different correlation coefficients used to measure the performance of evaluation metrics. Although tools have been proposed to aggregate different correlation scores, studying the correlation between different correlation coefficients is an area that needs more attention. To address this, we propose studying the correlation using ROCStories, WritingPrompts and the HANNA datasets, comparing the scores and rankings provided by the Pearson coefficients.

2 Litterature review

In the field of reference-based metrics, automated evaluation metrics (AEM) can be classified into three categories: string-based metrics, embedding-based metrics, and pre-trained template-based metrics.

String-based metrics evaluate the similarity of two texts by analyzing the raw text, including n-gram co-occurrences. Famous examples of such metrics are BLEU (Papineni et al., 2002) and

[1] https://github.com/Boran-lucas/NLP_project

ROUGE (Lin, 2004). However, this approach has limitations, as it cannot take into account language complexity, such as synonyms.

Metrics based on embeddings are calculated using the embeddings of the words, not the words themselves. There are two types of embeddings: (i) simple word embeddings, obtained for example with word2vec, where each word is linked to a single embedding, and (ii) contextualized word embeddings, obtained for example with BERT, where the embedding of each word depends on its context.

Metrics based on pre-trained models use the language representation contained in these models to evaluate the similarity between texts.

There are many AEMs, and it is impossible to describe them all. The development of new metrics is a constantly evolving field. Metrics based on popular strings, such as BLEU and ROUGE, are almost two decades old, but since then many other metrics have been proposed. In 2005, METEOR (Banerjee and Lavie, 2005) was introduced to overcome the limitations of BLEU, but it was still based on the philosophy of n-gram matching.

With the advent of embeddings, new metrics were designed to take advantage of the representation they offer, such as BERTSCORE (Zhang* et al., 2020) which uses BERT embeddings which are constructed using a transformer-based architecture, a type of neural network that incorporates an attention mechanism (Vaswani et al., 2017), BERTSCORE (Zhao et al., 2019) which aggregates information from different layers via power averaging, BARYSCORE (Colombo et al., 2021b) which uses the Wasserstein barycenter from optimal transport theory, and DEPTHSCORE (Staerman et al., 2022) which relies on a pseudo-metric based on the depth of data.

Finally, another line of research has been developed where the metrics rely on pre-trained models, giving rise to InfoLM (Colombo et al., 2022c) which uses a pre-trained masked language model to represent texts.

3 Experiments Protocol

3.1 Dataset

Focusing on the the MANS (Manually Annotated Stories) dataset, which was introduced by the OpenMEVA paper (Guan et al., 2021), builds upon the ROCStories (shortened as ROC) (Mostafazadeh et al., 2016) and WritingPrompts (shortened as WP) (Fan et al., 2018) datasets, we focused our efforts on story generation and evaluation. The WP dataset contains 303,358 pairs of prompts and stories, with no specific restrictions on writing topics, while the ROC dataset includes 98,162 commonsense stories, each composed of five sentences and around 50 words. We also worked with HANNA (Chhun et al., 2022), which offers annotations for 1,056 stories originating from 96 prompts in the WritingPrompts dataset. For each story, three raters provided annotations based on six distinct criteria: Relevance, Coherence, Empathy, Surprise, Engagement, and Complexity, totaling 19,008 annotations.

3.2 Evaluated Metrics

We conducted experiments using a variety of existing metrics, which can be presented as follows: (a) BLEU score (geometric mean from 1-gram to 4-gram); (b) ROUGE Metrics: ROUGE-1, ROUGE-2, and ROUGE-L, with recall, precision, and F1 score; (c) BERTSCORE Metrics: BERTSCORE precision, recall and F1 ; (d) Barycentric Score Metrics: BARYSCORE with varying weights and standard deviations ; (e) DEPTHSCORE Metric. In addition to these metrics, we have also introduced a variant of the BLEU metric to further enhance the evaluation process for generated text: (f) ENTROPY-WEIGHTED BLEU.

3.3 Expanding Evaluation Techniques: The ENTROPY-WEIGHTED BLEU Metric

The ENTROPY-WEIGHTED BLEU metric is a variant of the BLEU score that assigns weights to n-grams based on their entropy, thus giving greater importance to informative n-grams and reducing the importance of frequent n-grams. The intuition behind this approach is that less frequent and more informative n-grams should have a greater impact on the BLUE score. The standard BLEU score is computed as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log(p_n)\right)$$

Where:

- BP is the Brevity Penalty, calculated as $\text{BP} = \exp\left(1 - \frac{\text{length of reference}}{\text{length of candidate}}\right)$ if the length of the candidate is less than the length of the reference, and 1 otherwise

- w_n is the weight for each n-gram size (typically uniform weights, such as $1/4$ for 4-grams)

- p_n is the modified precision for n-grams

In the case of Entropy-Weighted BLEU, we modify the w_n term by incorporating the weights based on the entropy of the n-grams:

$$w_n = \text{entropy}_{n\text{-gram}_i}$$

Here, $\text{entropy}_{n\text{-gram}_i}$ is the entropy of the n-gram calculated as follows:

$$\text{entropy}_{n\text{-gram}_i} = -P_{n\text{-gram}_i} \cdot \log_2(P_{n\text{-gram}_i})$$

Where $P_{n\text{-gram}_i}$ is the probability of the n-gram in a large reference corpus.

By incorporating these $\text{entropy}_{n\text{-gram}_i}$ values into the w_n calculation, the Entropy-Weighted BLEU score gives more importance to the n-grams with higher entropy, and thus potentially more meaningful.

Since entropy is computed from a reference corpus, we used the whole ”gold responses” for ROCStories and the whole ”human responses” for HANNA as reference corpus for the evaluation.

3.4 Correlations Computations

The average scores for each story were determined by aggregating the ratings from human evaluators. Following that, the metrics detailed in the previous section were calculated for every story produced by an automated system, using the corresponding human-generated story as the gold standard. Correlations were then computed between each metric pair, specifically for the MANS dataset, between the sole human metric and the 23 automated metrics for each text generator, and for the HANNA dataset, between the six human metrics and the 23 automated metrics. We chose to compute these correlations using Pearson’s method. As our focus lies in evaluating the strength of the associations, only the absolute values of these correlations are taken into account in the interpretation.

4 Results

4.1 ROCStories & WP Datasets

As described in the previous paragraph, in the MANS datasets we compute the similarity metrics between the ”gold responses” and the texts generated in response to a prompt for each text generator and we correlate the scores obtained with the average of the scores given by the humans. Results are shown in Figure 1 for the ROCStories dataset and in Figure 2 for the WP dataset.

In these figures, among the metrics class having several possible compositions with precision, recall and f1 for ROUGE and Bert as well as different weights and standard deviation for BaryScore score, we have chosen to present only one metric per metric class by choosing the metric with the strongest correlations. For ROCStories Rouge1-p, BarySD0.001, and Bertscore-p were the best. For WP Rouge1-f, BarySD10, and Bertscore-p were the best. You can find the computations of all correlations and boxplots on selected metrics in appendix A for ROCStories and appendix B for WP dataset. The correlations between each metric and the human scores are reported for the following story generators: *gpt*, *planwrite*, *s2s*, *gptkg*, and *fusion*.

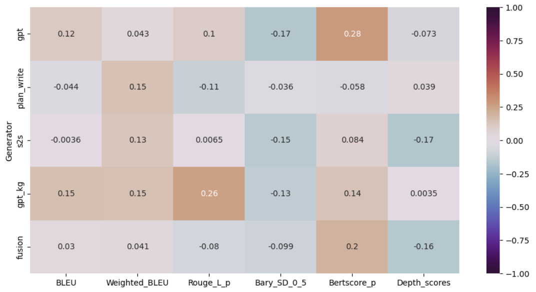

Figure 1: ROC Dataset : Correlations for different generators of selected metrics

Looking at the correlations for each generator, it can be observed that *gptkg* has the highest correlation with the human scores for most of the metrics, followed by *fusion* and *gpt*. On the other hand, *planwrite* and *s2s* have low or negative correlations with most of the metrics.

Overall, the figure suggests that the *gptkg* generator allows the metrics to be more efficient and closer to human-generated metrics than the other generators considered in the study. Moreover, this figure allows us to understand that depending on the chosen generator, the similarity metric has to be adapted accordingly, this table allowing us to have a slight overview of the possible alloys. Another remark that we can underline is that our contribution to the BLUE metric, Entropy-Weighted BLEU , allows us to gain in performance compared to human scores on all generators except *gpt*.

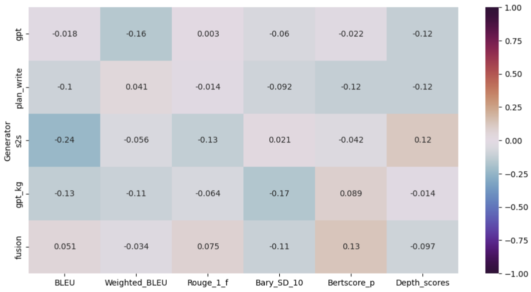

Figure 2: WP Dataset : Correlations for different generators of selected metrics

On this figure, first we notice that the correlations are weaker, it is certainly due to the fact that the WP dataset contains longer stories than ROC and we point then one of the limits of the similarity metrics. Here there is not really a text generator that stands out from the others, and we can confirm that depending on the generator it will be preferable to use certain metrics.

Contrary to the results on ROCStories, here Entropy-Weighted BLEU is less performing than BLEU so we can say that Entropy-Weighted BLEU is better applied to rather short data.

4.2 HANNA Dataset

Similarly, in the HANNA dataset, we compute the similarity metric between the ”human responses” and the text generated in response to a prompt for each text generator and correlate the scores obtained with the different scores given by humans, namely relevance, coherence, empathy, surprise, engagement and complexity. We obtain a correlation matrix between the automatic metrics and the metrics annotated by humans as shown in Figure 3

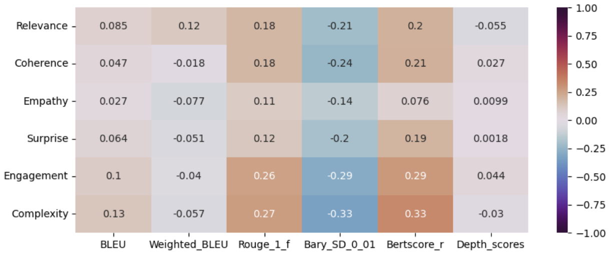

Figure 3: HANNA Dataset : Pearson correlations with best metrics of their category

Again, in this table, among the metrics class having several possible compositions, we have chosen to present only one metric per metric class by choosing the metric with the strongest correlations. You can find the computations of all correlations in appendix C.

The correlation matrix exposes shows the Pearson correlations between different metrics and six human judgments (Relevance, Coherence, Empathy, Surprise, Engagement and Complexity) for generated text.

The correlations show that relevance has a moderate positive correlation with Entropy-Weighted BLEU, Rouge1-f, and Bertscore-r. Coherence is weakly positively correlated with BLEU and Bertscore-r. Empathy has a weak positive correlation with BLEU, but a negative correlation with Entropy-Weighted BLEU. Surprise has a weak positive correlation with BLEU and Bertscore-r, while engagement has a moderate positive correlation with all metrics except Entropy-Weighted BLEU. Interestingly, complexity has a moderate positive correlation with all metrics except for Entropy-Weighted BLEU, which has a weak negative correlation with complexity. These results suggest that the metrics are generally positively correlated with the human judgments, but the strength of the correlations varies depending on the judgment and the metric. The negative correlation between empathy and Entropy-Weighted BLEU may also indicate that this metric does not fully capture the empathetic quality of generated text.

Reading this correlation matrix by column, we can see that the automatic metrics that have the best correlations with the human metrics are: Rouge-1-f, Bertscore-r and Bary-SD-0-001. We can notice that here Entropy-Weighted BLEU has no real interest compared to BLEU, we find the same conclusions as for WP which is logical given that HANNA is based on WP Dataset.

5 Conclusion

In this article, we have examined different techniques and metrics for evaluating the quality of automated text generation. We have discussed pre-trained language models, such as Word2Vec and BERT, as well as different metrics based on these models, such as BLEU, ROUGE, and BERTScore.

We tried to evaluate the quality of text generation on the MANS and HANNA datasets, considering differences between text generators for different text generators such as *gpt*, *planwrite*, *s2s*, *gptkg*, and *fusion*, and considering differences between human judgments for different human judgments such as Relevance, Coherence, Empathy, Surprise, Engagement and Complexity.

Finally, we have introduced a new variant of the BLEU metric, called ENTROPY-WEIGHTED BLEU, which takes into account the entropy of n-grams to give more weight to less frequent and more informative n-grams. We have shown that ENTROPY-WEIGHTED BLEU has a real interest compared to BLEU on datasets with shorter texts. Obviously, ENTROPY-WEIGHTED BLEU has limitations, notably its dependence on the entropy of the n-grams of the reference corpus, its inability to consider semantics and sentence structure, and its sensitivity to the quality and representativeness of the corpus used to calculate the entropies.

In conclusion, this study highlights the importance of choosing the right metrics to evaluate the quality of text generation. It is also important to continue exploring new metrics and methods for evaluating the quality of automated text generation (Colombo et al., 2022b).

References

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Karol Gregor and Frederic Besse. 2018. Temporal difference variational auto-encoder. *CoRR*, abs/1806.03107.

Pierre Colombo*, Wojciech Witon*, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *NAACL 2019*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Hamid Jalalzai*, Pierre Colombo*, Chloé Clavel, Éric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. *NeurIPS 2020*.

Jian Guan and Minlie Huang. 2020. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166, Online. Association for Computational Linguistics.

Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021b. Automatic text evaluation through the lens of wasserstein barycenters. *CoRR*, abs/2108.12463.

Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, (PhD thesis) Institut polytechnique de Paris.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. 2021a. A novel estimator of mutual information for learning to disentangle textual representations. *ACL 2021*.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics. *CoRR*, abs/2105.08920.

Pierre Colombo, Maxime Peyrard, Nathan Noiry, Robert West, and Pablo Piantanida. 2022b. The glass ceiling of automatic evaluation in natural language generation.

Pierre Jean A. Colombo, Chloé Clavel, and Pablo Piantanida. 2022c. Infolm: A new metric to evaluate summarization amp; data2text generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10554–10562.

Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M. Suchanek. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation.

Guillaume Staerman, Pavlo Mozharovskyi, Pierre Colombo, Stéphan Clémençon, and Florence d'Alché Buc. 2022. A pseudo-metric between probability distributions based on depth-trimmed regions.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. 2022a. What are the best systems? new perspectives on nlp benchmarking. *NeurIPS 2022*.

A Results for ROCStories

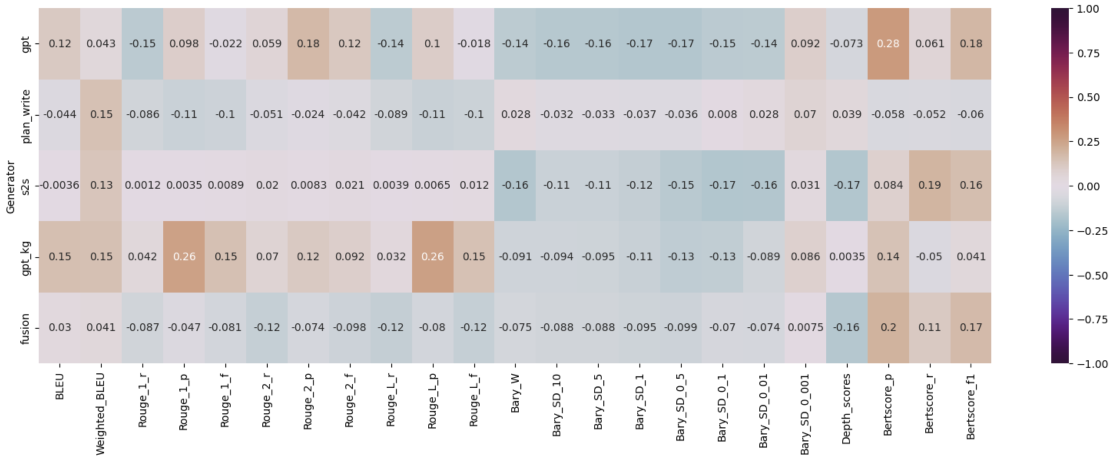

Figure 4: ROC Dataset : Correlations for different generators

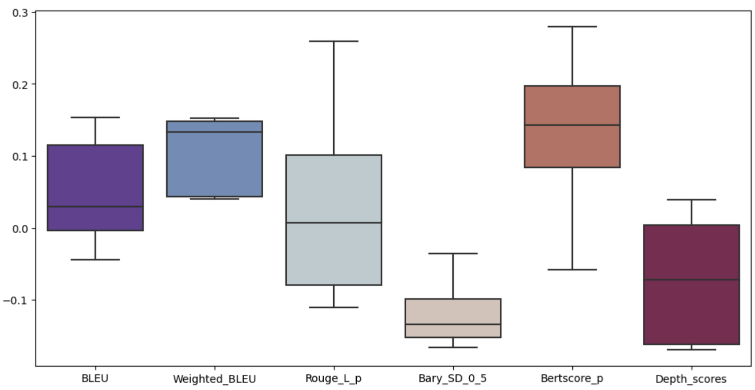

Figure 5: ROC Dataset : Boxplot of distributions on selected metrics for different generators

B Results for WritingPrompts

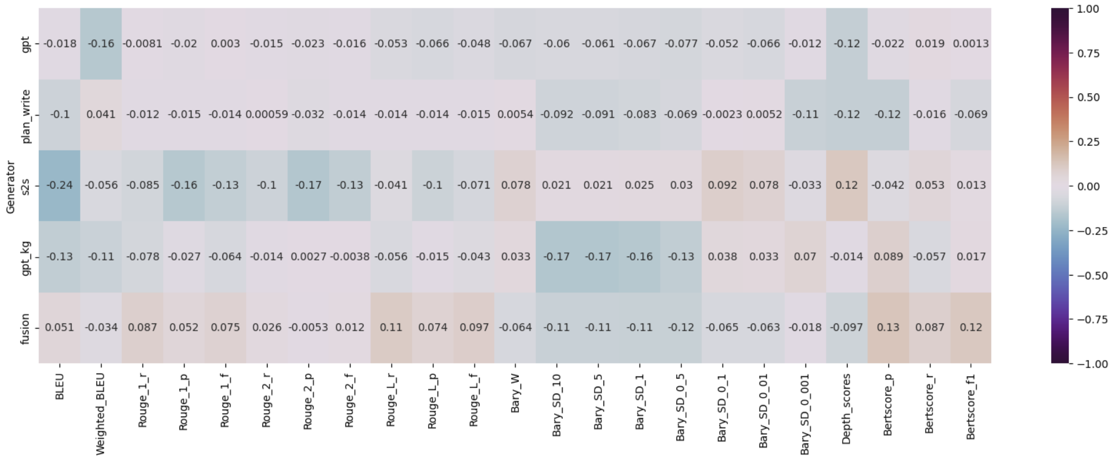

Figure 6: WP Dataset : Correlations for different generators

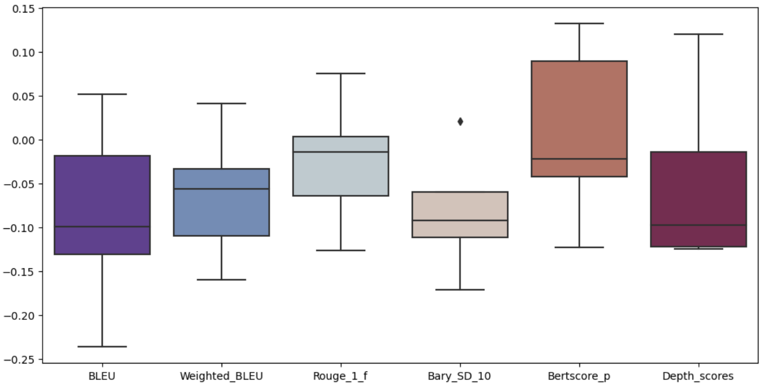

Figure 7: WP Dataset : Boxplot of distributions on selected metrics for different generators

C Results for HANNA

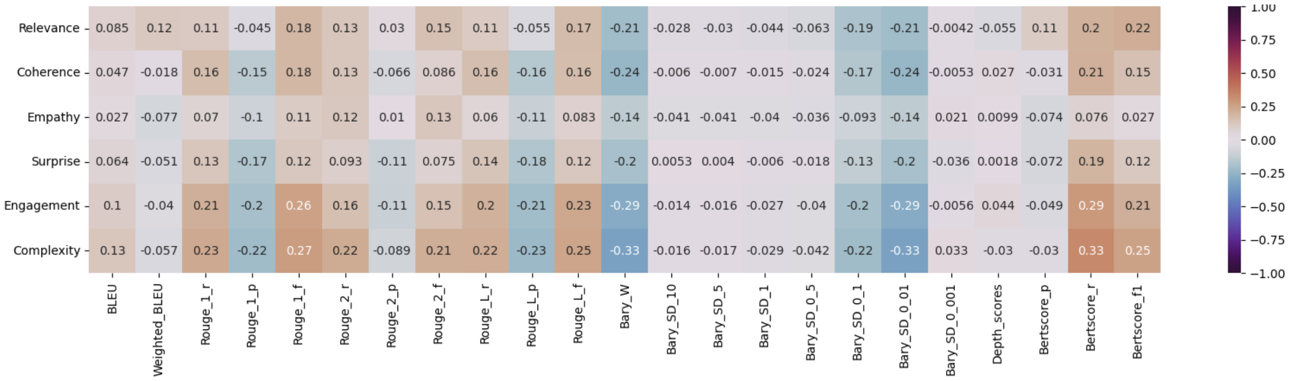

Figure 8: HANNA Dataset : Correlations matrix between automated metrics & human metrics