Benchmark of automatic metrics on Automatic Story Generation : do results depend on correlation coefficients ?

MENDRAS Pauline *
Ensae
pauline.mendras@ensae.fr

NIETGE Clotilde *
Ensae
clotilde.nietge@ensae.fr

Abstract

Accurately estimating the correlation between metrics and human judgment for ASG models is crucial to evaluate the effectiveness of these metrics and enhance the models. To this end, Kendall, Pearson, and Spearman correlation coefficients offer various correlation measures and rankings of metrics. Our research paper proposes to assess the discrepancies in metric rankings among these correlation coefficients, emphasizing the significance of considering the peculiarities of each correlation coefficient before potential aggregation. By analyzing these differences, we can gain a better understanding of the strengths and limitations of each correlation coefficient and develop more reliable evaluation strategies for ASG models. Additionally, our findings highlight the need for further research on the effectiveness of different correlation measures in evaluating ASG models. Code for our research is available on Github [1].

1 Introduction

The increasing popularity of natural language generation (NLG) systems (Jalalzai* et al., 2020; Colombo et al., 2021a), such as ChatGPT, has led to a growing interest in automatic evaluation methods. These systems have a wide range of potential applications, including customer service, virtual assistants, and content creation. As such, it is essential to have reliable and efficient methods to evaluate the performance of NLG systems. Automated evaluation metrics provide a way to measure the quality of generated text by comparing it to a reference text or to human judgments.

Automatic evaluation is particularly important in NLG because the quality of the generated text is often subjective and difficult to quantify. Additionally, manual evaluation is time-consuming and expensive, and it may not always be feasible to obtain human judgments for large datasets (Colombo, 2021; Colombo et al., 2021c). Thus, automatic evaluation metrics can be used to quickly and efficiently assess the quality of generated text, which can help developers and researchers to improve NLG systems.

In this paper, we examine the metrics under the lens of Automatic Story Generation a specific subfield of NLG. Automatic story generation is a growing area of interest in natural language processing, which utilizes machine learning techniques to generate fictional narratives (Chhun et al., 2022). These systems have the potential to produce engaging and original stories that are tailored to specific audiences or contexts. The applications of this technology are diverse, ranging from entertainment and education to marketing and advertising.

Automatic story generation (ASG) takes as input a short sentence (a *prompt*) and aims at generating a narrative from it. In addition to GPT, the family of languages models from which ChatGPT is originated, other ASG systems exist such as `BERTGeneration` (Faidon Mitzalis, 2021), `Fusion` (Fan et al., 2018) or `TD-VAE` (David Wilmot, 2021). In order to assess the relevance of the generated stories and to improve ASG models, evaluation metrics (like `BLEU` (Papineni et al., 2002), `BertScore` (Tianyi Zhang and Artzi, 2020) or `ROUGE` (Lin, 2004)) have been developed. These metrics evaluate the text generated by ASG models.

But is the evaluation of these metrics satisfactory? In other words, do they come close to what a human judgement might produce? To answer these questions, work has been done to evaluate these metrics by measuring their correlation with human judgements (Chhun et al., 2022). To do

[1] https://github.com/PaulineMendras/NLP-ENSAE-3A-MENDRAS-NIETGE-Topic4.git

this, they have used correlation coefficients, such as Kendall (Kendall, 1938), Pearson or Spearman coefficients. Depending on the criteria selected, correlation coefficients can result in different scores and rankings for the different metrics. Tools to aggregate these different correlation scores have been proposed using, for instance, the Kemeny consensus and the Borda count (Colombo et al., 2022b). However, the choice of correlation coefficient (Kendall, Pearson, Spearman) might alter the choice of metrics too, in which case the three correlation coefficients should be aggregated.

Contribution: In this paper, we propose to study the stage before the aggregation of the different correlation scores. We believe that the study of the correlation between the different correlation coefficients is a missing piece in the ASG literature. For this purpose, we have used the HANNA dataset (Chhun et al., 2022) and reproduced the work previously done by the authors by using two other correlation coefficients. We have then studied in details the scores and compared the rankings of the metrics provided by three correlation measures: the Kendall, Pearson and Spearman coefficients. The results show that there could be important differences in metrics ranking between correlation coefficients. In particular, our work highlights the importance of the choice of correlation coefficients and the usefulness of aggregating them to produce a final ranking.

2 Related work

The development of ASG models has been massive in recent years. BERTGeneration (Faidon Mitzalis, 2021), Fusion (Fan et al., 2018) and GPT (Nema and Khapra, 2018) are among the best known. A recent advance is the consideration of human emotions. The EMOTICONS system (Colombo et al., 2019) for instance is an affect-driven dialog system, which generates emotional responses in a controlled manner using a continuous representation of emotions. However, the quantitative (BLEU score) and qualitative performance of this system is not entirely satisfactory, as EMOTICONS does not generate different emotions equally well.

The evaluation of ASG systems is indeed essential to measure and improve the accuracy of human text generators. Several metrics have therefore been developed, such as GRUEN (Zhu and Bhat, 2020), BLEU (Papineni et al., 2002), BertScore (Tianyi Zhang and Artzi, 2020) or ROUGE (Lin, 2004) scores. They are based on different criteria to assess a text generated by ASG models, such as quality content, emotion faithfulness, grammar, logicality or creativity. As an example, more recently, InfoML (Colombo et al., 2022a) is a metric using the criteria of data coverage, relevance, correctness, structure and fluency to evaluate a text generation. It is an automatic pre-trained model which is robust to synonyms. Another example is the BaryScore (Colombo et al., 2021b) metric which is based on optimal transport tools: the Wasserstein distance and barycentres, and on a probabilistic distribution rather than on embedding vectors. The BaryScore metric performs better than BERT metrics and is more consistent for text summaries.

The emergence of several metrics has therefore made it necessary to compare them and in particular to detect correlation with human judgement. It has been proved for instance that InfoML (Colombo et al., 2022a) has a better system-level and text-level correlations with human judgement than other metrics. However, there is a key difference between *automatic metrics* (AEM) and human metrics (Colombo et al., 2022c). AEM rank systems similarly but differently than humans. More surprisingly, human metrics predict each other much better than the combination of all automatic metrics. This casts serious doubt about the ability of AEM to replace human judgements.

Measuring the correlation between AEM and human judgement is the goal of the HANNA project (Chhun et al., 2022) on which our work is based. HANNA contains annotations for 1,056 stories generated from 96 ASG models. Each story is annotated by three raters on six criteria (relevance, coherence, empathy, surprise, engagement and complexity). Additionally, those 1,056 stories are evaluated by 72 automatic metrics. The use of coefficient correlations enables them to compare human scores and AEM scores and to detect system-level or text-level correlations. The aggregation of the different correlation scores on criteria is done using the Kemeny consensus method and the Borda count. These methods are proved to be more reliable and more robust than the mean aggregation (Colombo et al., 2022b). The results of the HANNA projet reveal that most AEM have moderate story-level correlations with hu-

man metrics, while system-level correlations are a bite better. GPT-2 (Radford et al., 2019) has the better scores for all the six criteria and ROUGE (Lin, 2004), Bary-Score (Colombo et al., 2021b), DepthScore (Guillaume Staerman, 2022), BARTScore (Weizhe Yuan, 2021) have the higher system-level aggregated correlation score.

However, the aggregation process can be viewed as a "black box" process, and it seems important to highlight the original rankings that have led to the final scores. In this way, we propose to study in details the scores and rankings of the metrics based on three correlation coefficients: the Kendall, Pearson and Spearman coefficients . This will allow us to analyse whether there are any consequential differences in rankings between metrics that could influence the final ranking.

3 Methods

To best evaluate automatic metrics, they are compared to human judgments by calculating the correlations between the two. For this, there are three correlation coefficients: Pearson, Spearman and Kendall that can be calculated on two levels of granularity: text-correlation and system-correlation.

Let y_i^j be the story generated by system $j \in \{1, ..., S\}$ for prompt $i \in \{1, ..., N\}$, $m(y_i^j)$ is the score associated by the metric m and $h(y_i^j)$ the human judgement of the generated story y_i^j.

Text-level Correlation. For each prompt, we compute the correlation between the automatic metric and the human judgment. Then we average these correlations over all the prompts.

$K = \frac{1}{N} \sum_{i=1}^{N} Corr(M^{text}, H^{text})$
with • $M^{text} = [m(y_i^1), ..., m(y_i^S)]$
• $H^{text} = [h(y_i^1), ..., h(y_i^S)]$

System-level Correlation. We compute the mean of the metric scores for each model and the average of the human judgments. Then we calculate the correlation between these averages. This correlation allows us to compare the ASG systems between them. Formally:

$K = Corr(M^{sys}, H^{sys})$ with
• $M^{sys} = [\frac{1}{N} \sum_{i=1}^{N} m(y_i^1), ..., \frac{1}{N} \sum_{i=1}^{N} m(y_i^S)]$
• $H^{sys} = [\frac{1}{N} \sum_{i=1}^{N} h(y_i^1, ..., \frac{1}{N} \sum_{i=1}^{N} h(y_i^S)]$

The common method is to remove outliers (Mathur et al., 2020) because Pearson absolute value correlations are sensitive to outliers.

Human judgments, which may contain noise, are also normalized (Banerjee and Lavie, 2005). However, the Pearson, Spearman and Kendall coefficients are calculated in different ways. Pearson considers linear dependencies, while the other coefficients are based on rank dependencies. Therefore, the ranking of the best metrics may depend on the correlation coefficient considered.

To see if it is the case, we implemented Wilcoxon test. Let the metric $i \in \{1, ..., N\}$, X_i its associated Pearson coefficient, Y_i its Spearman coefficient and $R_i = |X_i - Y_i|$ its associated rank of the coefficients difference. The statistic of the test is $T_i = \sum_{i=1}^{n} R_i \mathbb{1}_{X_i - Y_i > 0}$ where the differences $(X_i - Y_i)_i$ are supposed independent. Higher the T-statistic, more different are the correlation coefficient.

4 Experimental setting

Our experiment setting will be based on *Automated Story Generation* via a prompt. As we need human judgments on the systems, we used an existing database: HANNA (Human-ANnotated NArratives for ASG evaluation) (Chhun et al., 2022). This database, coming from WritingPrompts dataset, is composed of 1,056 stories generated by 10 systems. WritingPrompts (Fan et al., 2018) dataset is composed of short sentences called prompt and has been largely used for *Automatic Story Generation*. On one hand, the advantages of the database HANNA are the quality and quantity of human evaluations on six criteria relevant to the evaluation of text generation: relevance, coherence, empathy, surprise, engagement and complexity. Having several human criteria will make more precise the evaluation of the metrics. Moreover, for each prompt/story combination, we have three different evaluations which we have averaged. On the other hand, this database provides the scores of 72 automatic metrics associated to each generated story.

We removed the outliers to improve the efficiency of the correlation coefficients. This was equivalent to remove the texts generated by humans. It is on this database that we applied the correlations detailed in section 3 both at the text-level and system-level.

Finally, for each criteria, we ranked the different metrics. The rankings might be different for each

correlation coefficient. Therefore, the Wilcoxon test allow us to assess if two correlation coefficients rank the metrics in the same order.

5 Results

For each of the six criteria (relevance, coherence, surprise, empathy, engagement and complexity), the correlation between the scores given by each of the 72 metrics and human judgement is calculated using three correlation coefficients: Kendall, Pearson and Spearson (5.1). Three different rankings of the top five metrics can then be made (5.2). The challenge is then to study the similarity and correlation of these different rankings using the Wilcoxon test (5.3). This will allow us to conclude whether the results differ or not by using different correlation coefficients.

5.1 Correlation between automatic metrics and human judgements

The figure 1 gathers on a same heatmap the system-level correlation measures between human judgement and the different metrics (on the ordinate) for each of the six criteria (on the abscissa), for respectively the Kendall coefficient. The heatmaps for the Pearson and Spearman coefficients as well as for text-level correlations are in the appendix (figures 2, 3, 4, 5 and 6).

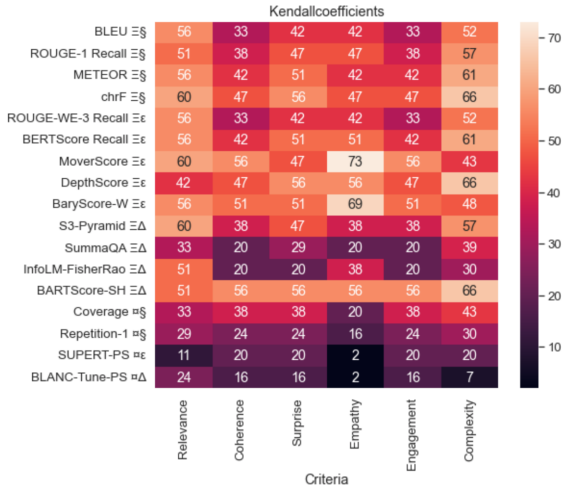

Figure 1: Story-level Kendall correlations (%) between human criteria

We can rank the automatic metrics with their coefficient correlation. Higher the coefficient, better the metric is. For instance, Kendall correlations range from 2 % (for the metric BLANC (Vasilyev et al., 2020) in empathy) to 73 % (for the metric MoverScore (Zhao et al., 2019)). By looking at these heatmaps, we can quickly see differences in correlation measures between the three coefficients. As an example, considering the complexity criterion, the SUPERT-PS (Gao et al., 2020) metric is better ranked than the BLANC metric by the Kendall coefficient (20 % vs. 7 %), while it is the BLANC metric which is better ranked by the Pearson coefficient (33 % vs. 68 %).

5.2 What is the Top-5 metrics based on criteria ?

These difference become more apparent when considering the ranking of the top-5 metrics. The table 1 shows the ranking of the top 5 metrics by the three correlation coefficients, concerning the relevance criterion and at the system-level. The rankings for the coherence and surprise criteria at the system-level are in the appendix (tables 3 and 4).

Rank	Pearson	Spearman	Kendall
1	BLEU	MoverScore	S3-Pyramid
2	ROUGE-1 Recall	S3-Pyramid	chrF
3	S3-Pyramid	chrF	MoverScore
4	METEOR	METEOR	BLEU
5	BARTScore-SH	BLEU	BERTScore Recall

Table 1: Top-5 best metrics at system-level for relevance criterion

By looking at the table, the differences between the rankings are quite obvious. For instance, if the BLEU metrics appear in the three rankings, it is ranked first in Pearson ranking while it is fourth and fifth in Kendall and Spearman respectively. The MoverScore metric is ranked second in the Spearman ranking while it does not appear in the Pearson'one. Similarly, the ROUGE metric is ranked second by the Pearson correlation coefficient and is not present in the Spearman and Kendall top-5 best metrics rankings.

5.3 Is the automatic metrics ranking the same through the different correlation coefficients ?

Measuring the correlation between correlation coefficients seems then necessary. The Wilcoxon test enables us to analyse the differences between the rankings of metrics by the three correlation coefficients. The higher the Wilcoxon test score between two correlation coefficients, the greater the differences in rankings of metrics between these two coefficients. The table 2 shows the Wilcoxon

scores between the Kendall, Pearson and Spearman coefficients for the six criteria.

Coefficient correlation	Relevance	Coherence	Surprise	Empathy	Engagement	Complexity
Pearson - Spearman	845.0	335.0	256.0	319.0	248.0	254.0
Kendall - Spearman	6.0	41.0	9.0	48.0	41.0	1.0
Kendall - Pearson	86.0	90.0	43.0	103.0	60.0	49.0

Table 2: Wilcoxon statistic for relevance criterion for each pair of correlation coefficient

The Wilcoxon test confirms the above intuitions. Differences in rankings exist between the correlation coefficients. In particular, for all criteria, there are large differences between the Pearson and Spearman rankings, whereas these differences are smaller between the Kendall and Pearson rankings. The Wilcoxon test seems to suggest that the differences are very small between the Kendall and Spearman rankings, especially for the complexity and surprise criteria.

These results highlight the differences in measurement and ranking that can exist between the different correlation coefficients. This emphasises the need to study in detail the correlation coefficients used and their specificity. Using different correlation coefficients and aggregating their results with the Kemeny consenus and Borda count as done in Colombo et al. (2022b) is a consistent solution to overcome these differences and produce a global ranking.

6 Discussion and conclusion

Assessing the performance of ASG metrics requires measuring their correlation with human judgment, but different correlation coefficients have varying specificities. Our study investigates the correlation between these coefficients and, specifically, the differences in metric rankings they produce. Through the Wilcoxon test, we find that significant differences in rankings can exist between coefficients. Notably, we observe larger discrepancies between Pearson and Spearman coefficients than between Kendall and Spearman. These findings emphasize the importance of selecting appropriate correlation coefficients and aggregating them for more accurate metric rankings.

Futur Works: There is still room for improvement, as we have not analyzed what could explain these differences. This may be related to the structure of the dataset itself and the choice of the criteria studied. Some criteria, such as surprise, are highly subjective and can be assessed very differently by different humans and therefore show a high degree of variance. This can be difficult to measure by metrics. Indeed, according to Deutsch et al. (2021), for summarization, confidence intervals are rather wide, which state a high uncertainty in the reliability of automatic metrics. Therefore the correlation measure can be biased, which can explain differences in scores between different correlation coefficients.

In addition, we did not take into account the unlabeled part of the testset of `WritingPrompt`. As human judgements are expensive, only a subset of the overall dataset is evalualed. An idea from Deutsch et al. (2022) would be to use the unlabelled stories to improve the computation of correlation at system-level.

References

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. pages 74–81.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. pages 889–898.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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. pages 563–578.

Felix Wu Kilian Q. Weinberger Tianyi Zhang, Varsha Kishore and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. pages 11–20.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. pages 94–108.

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*.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. page 4984–4997.

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

Pengfei Liu Weizhe Yuan, Graham Neubig. 2021. Bartscore: Evaluating generated text as text generation.

Pierre Colombo, Chouchang Yang, Giovanna Varni, and Chloé Clavel. 2021c. Beam search with bidirectional strategies for neural response generation. *ICNLSP 2021*.

Pranava Madhyastha Lucia Specia Faidon Mitzalis, Ozan Caglayan. 2021. Bertgen: Multi-task generation through bert.

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*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods.

Frank Keller David Wilmot. 2021. A temporal variational model for story generation.

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

Pierre Colombo, Chloé Clavel, and Pablo Piantanida. 2022a. A new metric to evaluate summarization data2text generation.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. pages 5794–5836.

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

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clemencon. 2022b. What are the best systems? new perspectives on nlp benchmarking.

Appendix

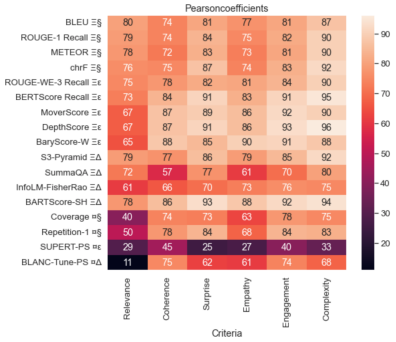

Figure 2: Story-level Pearson correlations (%) between human criteria

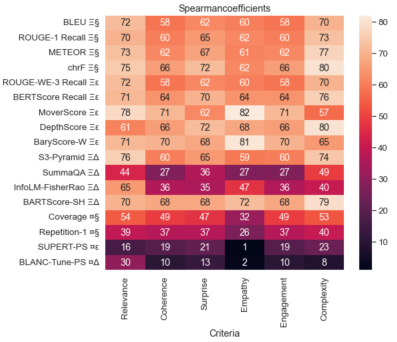

Figure 3: Story-level Spearman correlations (%) between human criteria

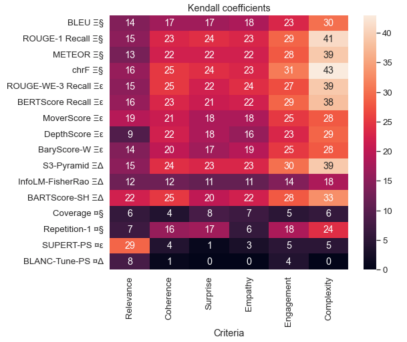

Figure 4: Text-level Kendall correlations (%) between human criteria

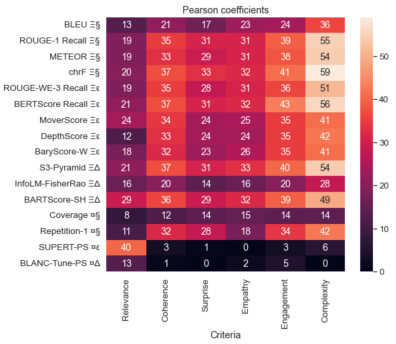

Figure 5: Text-level Pearson correlations (%) between human criteria

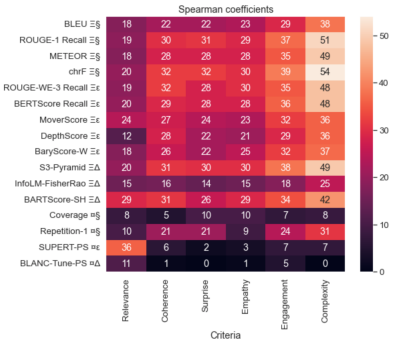

Figure 6: Text-level Spearman correlations (%) between human criteria

Rank	Spearman	Pearson	Kendall
1	MoverScore	BaryScore-W	BARTScore-SH
2	BaryScore-W	DepthScore	MoverScore
3	BARTScore-SH	MoverScore	BaryScore-W
4	chrF §	BARTScore-SH	chrF §
5	DepthScore	BERTScore Recall	DepthScore

Table 3: Top-5 best metrics at system-level for coherence criterion

Rank	Spearman	Pearson	Kendall
1	chrF §	BARTScore-SH	chrF §
2	DepthScore	BERTScore Recall	BARTScore-SH
3	BERTScore Recall	DepthScore	DepthScore
4	BaryScore-W	MoverScore	BaryScore-W
5	BARTScore-SH	chrF §	METEOR §

Table 4: Top-5 best metrics at system-level for surprise criterion