Automatic Evaluation Metrics for Enhancing the Quality of Automatic Story Generation in NLP

Louisa Camadini
ENSAE Paris
louisa.camadini@ensae.fr

Abstract

In recent years, Automatic Story Generation (ASG) has become a popular subfield of Natural Language Processing (NLP). However, to ensure that ASG models produce high-quality outputs, reliable evaluation methods are necessary. Human evaluation, though effective, can be expensive and time-consuming. As a result, researchers have shifted their focus to developing Automatic Evaluation Metrics (AEM) that can accurately assess the quality of ASG outputs and also align with human judgement. This research area is essential for the progression of ASG and has become a prominent topic in NLP research.

1 Introduction

Automatic story generation is a rapidly evolving field of artificial intelligence that involves using algorithms and machine learning techniques to automatically generate stories, either entirely from scratch or by adapting existing narratives. This technology has the potential to revolutionize the way we produce and consume creative content, as it enables us to quickly generate large volumes of high-quality stories that are tailored to specific audiences and contexts. However, to quantify progress, it is important to carefully evaluate the quality of the systems being developed.

Research into the field of Automatic Story Generation (ASG) heavily relies on both human and automatic evaluation [1, 2, 3, 4, 5, 6]. However, there is currently no agreement on which human evaluation criteria to use, and there has been no examination of how well automatic evaluation criteria align with them [7]. This paper proposes a reevaluation of the some of the ASG evaluation methods, as in [8]. The purpose of this reevaluation is to quantitatively assess the correlations between some automatic metrics and six human criteria. The code is available in our Github repository[1].

Our research in this paper involves utilizing the HANNA dataset[2] introduced in [8] to evaluate the performance of various AEM. Specifically, we reproduce some experiments from this dataset. Consequently, the structure of our paper and the nature of our research is largely derived from this previous work.

2 Related Work

Natural Language Generation (NLG) is the process of generating text or speech from structured data. Automatic evaluation of NLG is a critical research topic that aims to assess the quality of the generated text automatically, without human intervention. The evaluation can be used to compare different NLG systems, improve the quality of the generated output, and provide feedback to the developers.

Several approaches have been proposed for the automatic evaluation of NLG, including *string-based, embedding-based,* or *model-based.*

String-based metrics evaluate the similarity of two texts by analyzing the raw text by various means, including n-gram co-occurrences. For example, the famous ROUGE [9] and BLEU [10] metrics. But this approach has recently been criticized as limited because it cannot take into account language complexity, such as synonyms.

Other measures have been introduced, based on embedding and computed from word embeddings (and not from the words themselves). For example, the embeddings obtained by Word2Vec are single word embeddings (i.e. each word is linked to a single embedding); while those obtained by

[1] https://github.com/louisacamadini/NLP_Project

[2] https://github.com/dig-team/hanna-benchmark-asg

BERT are contextualized embeddings (each word depends on its context).

Finally, model-based measures use the linguistic representation contained in the pre-trained language models.

Here is a non-exhaustive list of the different existing metrics:

1. BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation): string-based metrics, they both rely on n-gram matching to evaluate the similarity between two pieces of text. BLEU is commonly used in machine translation, while ROUGE is often used in text summarization.

2. METEOR [11] (Metric for Evaluation of Translation with Explicit ORdering): This metric was introduced in [11] and aims to overcome some of the shortcomings of BLEU. It still relies on n-gram matching, but also takes into account other factors like word order and synonymy.

3. BERTScore: This metric was introduced in [12] and leverages the embeddings provided by BERT. It calculates the cosine similarity between the embeddings of two pieces of text.

4. MoverScore: This metric, introduced in [13], aggregates information from different layers of a pre-trained language model (specifically, the GPT-2 model) to evaluate text similarity. It uses a power mean to combine the information from different layers.

5. BaryScore [14]: This metric, introduced in 2021, uses the Wasserstein barycenter (a concept from optimal transport theory) to evaluate the similarity between two pieces of text.

6. DepthScore [15]: This metric, introduced in 2022, relies on a pseudo-metric based on data-depth to evaluate the similarity between two pieces of text.

7. BARTScore: Introduced in 2021, this metric relies on a pre-trained sequence-to-sequence model called BART to evaluate text similarity. It calculates the probability that one piece of text would be generated by the model, given the other piece.

8. InfoLM: Introduced in [16], it uses a pre-trained masked language model (specifically, the RoBERTa model) to represent text and evaluate text similarity.

3 Problem Framing

Dataset In a previous study by [8], a set of six comprehensive and independent human evaluation criteria (relevance, coherence, empathy, surprise, engagement and complexity) was introduced for assessing the quality of narratives generated by ASG systems. Additionally, the annotated dataset called HANNA was created, which consists of 1,056 stories produced by 10 different ASG systems, each evaluated by three human raters using the six proposed criteria.

The main part of our study is based on `hanna_metric_scores.csv`[3], file that contains the evaluation scores for ASG systems using the HANNA dataset[4]. This file has 1,056 rows, each corresponding to a story produced by an ASG system, and 9 columns, each corresponding to a different evaluation metric. Each evaluation metric is scored on a scale of 1-5, where a higher score indicates better performance. These scores can be used to evaluate and compare the performance of different ASG systems on the HANNA dataset across multiple metrics.

Metrics The mathematical framework used in this study is based on the one presented in [16], which is described in this paragraph. Specifically, they examine a set of evaluation criteria for ASG, and analyze their correlation with different AEM using the dataset $\mathcal{D}$. The authors of this study previously conducted a similar analysis. Formally, $\mathcal{D} = \{x_i, \{y_i^s, h(x_i, y_i^s)\}_{s=1\ldots S}\}_{i=1\ldots N}$ where x_i is the i-th reference text; y_s^i is the i-th candidate text generated by the s-th NLG system; N is the number of texts in the dataset and S the number of systems available. In the HANNA dataset, N = 96 and S = 10. The vector $x_i = (x_1, ..., x_M)$ is composed of M tokens and $y_i^s = (y_1^s, ..., y_L^s)$ is composed of L tokens. $h(x_i, y_i^s) \in \mathbf{R}^+$ is the score associated by a human annotator to the candidate

[3] https://github.com/dig-team/hanna-benchmark-asg/raw/main/hanna_metric_scores.csv

[4] We only choose the score column, because it provides a balance between recall and precision (see code).

text y_i^s when comparing it with the reference text x_i. We aim at evaluating an AEM f, such that $f(x_i, y_i^s) \in \mathbf{R}^+$.

Evaluation AEM are typically assessed by measuring their correlation with human judgment using one of three correlation coefficients:

- Pearson, which quantifies linear relationships,
- Spearman, which considers the rank of the observations,
- Kendall, which measures similarity in ranking of the points along two axes.

These metrics can also be evaluated at two levels:

1. Text-level correlation, which measures the correlation between the metric and human judgment for each individual text,
2. System-level correlation, which measures the correlation across multiple texts generated by the same ASG system.

4 Experimental settings

We have decided to focus on a handful of metrics: BLEUE, METEOR [11], BERTScore, but our code is easily adapted to the study of any metric. By studying various metrics such as BLEU and ROUGE, METEOR, and BERTScore, we can gain a deeper understanding of their effectiveness and potential drawbacks. By comparing these metrics, we can determine if simple indicators like ROUGE and BLEU have been rightfully criticized, and whether they can be improved with the use of more advanced techniques such as METEOR and BERTScore. Exploring these metrics will enhance our knowledge and help us to make more informed decisions in our work.

BERTScore uses BERT (Bidirectional Encoder Representations from Transformers) to calculate a similarity score between two pieces of text. This has earned it a reputation as a state-of-the-art measure for evaluating the quality of generated text.

We computed each metrics at two distinct levels, text-level correlation and system-level correlation, and for three different correlation coefficients: Pearson, Kendall [17], Spearman.

The function "my_func" was used for creating text-level correlation plots, where each story was considered as a single data point, while the "my_func_sys" function is used for creating system-level correlation plots, where the metrics are aggregated across all the stories in the system and each metric is considered as a single data point. The system-level plot can be used to analyze the overall correlation between the two sets of metrics at the system level, whereas the text-level plot focuses on the correlation between metrics for each individual story.

Text-level You will find in appendix 2, 3, 4, different graphs of the correlation between the human criteria, for the three types of suggested coefficients. In appendix 5, 6, 7, graphs of the correlation between AEM. Overall, the Pearson correlation coefficients are higher than other.

Finally, Figure 8 focus on the Kendall correlation between human criteria and AEM. Note that all correlation coefficients are less than 50%.

In general, Pearson correlation is used to measure the linear relationship between two continuous variables. For the sake of readability, we choose to continue by focusing on the Kendall correlation.

System-level In appendix 9, 10, you can observe graphs of the correlation between the human criteria, and AEM. Figure 1 focus on the Kendall correlation between human criteria and AEM, at the system-level. the coefficients are globally higher at the system-level than at the text-level, where they did not exceed 50%.

We notice that BERTScore (last column of 1) is highly correlated with the human criteria, unlike ROUGE-3 and -4. The objective of the next section is to quantify this.

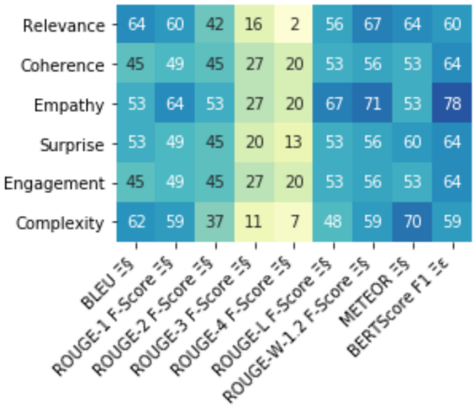


Figure 1: System-level, Absolute Kendall correlations between AEM ans human criteria (%)

5 Results

Here we are going to compare the AEM to determine which are the best, from the human evaluation perspective. This corresponds to the section *"Best metrics for human evaluation"* in the code. The `my_func_sys_top` function was modified to return the x AEM most correlated to all six human criteria.

For $x = 3$, the results are as follows:

- Pearson: ROUGE-1, BERTScore, METEOR
- Kendall: BERTScore, ROUGE-W, METEOR
- Spearman: BERTScore, ROUGE-W, METEOR

For the Pearson correlation: we clearly observe on the graph 11 that ROUGE-1 stands out very clearly from the others, at all levels.
For the Kendall correlation: graph 12 shows that BERTScore and ROUGE-W are competing for the first position.
For the Spearman correlation: see in 13 that BERTScore scores better overall than the others.

Borda's count Now we will suggest a ranking of the metrics using the Borda count (at system-level). We use the `my_func_sys_borda` function and obtain the following table 1.

In this function, Borda's count is used to rank the metrics based on their correlation scores with other criteria. Specifically, the function calculates the correlation scores between each pair of AEM using three different correlation measures (Kendall, Pearson, and Spearman) and constructs a correlation matrix. The function then applies Borda's count to the correlation matrix to rank the AEM based on their total score across the three correlation measures. Finally, the function returns a list of the top AEM based on their Borda's count scores.

6 Conclusion

The first three metrics (ROUGE-1, BERTScore, METEOR) along with ROUGE-W, exhibit strong global rankings. However, BERTScore stands out as the top AEM with two first place and one second place rank.

Table 1: AEMs ranking

	Pearson	Kendall	Spearman
ROUGE-1	1	5	4
BERTScore	2	1	1
METEOR	3	3	3
ROUGE-L	4	4	6
ROUGE-W	5	2	2
ROUGE-2	6	7	7
BLEU	7	6	5
ROUGE-4	8	9	9
ROUGE-3	9	8	8

BERTScore distinguishes itself from other metrics that rely solely on exact word matches or n-gram overlap. It considers semantic meaning and sentence structure by utilizing a pre-trained BERT model to encode the input sentences into high-dimensional vectors. These vectors are then compared using a cosine similarity measure, producing a score that reflects the similarity between the two sentences in terms of both their meaning and structure.

Benchmarking studies have shown that BERTScore outperforms other commonly used metrics such as BLEU and ROUGE. It has also demonstrated a high degree of correlation with human evaluations of text quality. Because of its superior performance and ability to capture semantic meaning, BERTScore has become a favored choice for evaluating the performance of language models and other NLP applications.

In addition to evaluating the quality of outputs generated by ASG systems, it is essential to assess the presence of bias in these systems [18, 19, 20]. Bias can manifest in many ways, including the propagation of harmful stereotypes and the perpetuation of societal inequalities. It is crucial to ensure that ASG systems are designed and evaluated with diversity, equity, and inclusion in mind, and that they do not reinforce or perpetuate negative biases.

References

[1] Brian Thompson and Matt Post. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. *arXiv preprint arXiv:2004.14564*, 2020.

[2] Florian Böhm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214*, 2019.

[3] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5804–5812, 2018.

[4] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[5] Daniel Deutsch, Rotem Dror, and Dan Roth. A statistical analysis of summarization evaluation metrics using resampling methods. *arXiv preprint arXiv:2104.00054*, 2021.

[6] Maja Popović. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618, 2017.

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

[8] Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M Suchanek. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. *arXiv preprint arXiv:2208.11646*, 2022.

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

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, July 2002. Association for Computational Linguistics.

[11] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.*, 2005.

[12] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[13] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*, 2019.

[14] Pierre Colombo, Guillaume Staerman, Chloe Clavel, and Pablo Piantanida. Automatic text evaluation through the lens of wasserstein barycenters. *EMNLP 2021*, 2021.

[15] Guillaume Staerman, Pavlo Mozharovskyi, Pierre Colombo, Stéphan Clémençon, and Florence d'Alché Buc. A pseudo-metric between probability distributions based on depth-trimmed regions. *arXiv e-prints*, pages arXiv–2103, 2021.

[16] P. Piantanida P. Colombo, C. Clavel. Info lm: A new metric to evaluate summarization and data2text generation. student outstanding paper award. *AAAI*, 2022.

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

[18] Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. A differential entropy estimator for training neural networks. In *ICML 2022*, 2022.

[19] Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. Learning disentangled textual representations via statistical measures of similarity. *ACL 2022*, 2022.

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

Appendix

Text-level

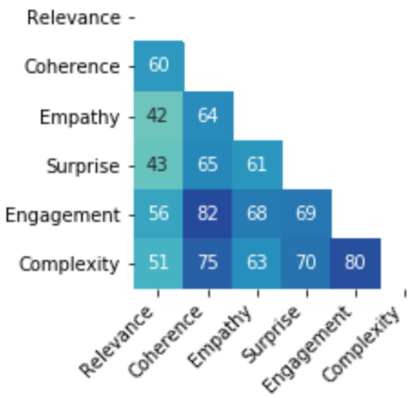


Figure 2: Text-level, Absolute Pearson correlations between human evaluations (%)

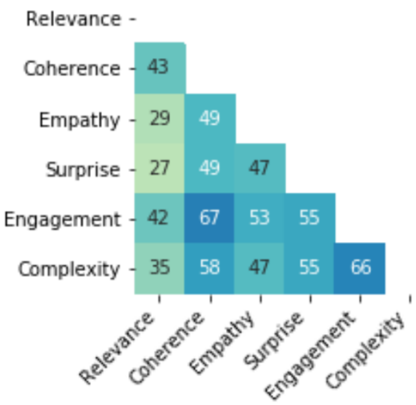


Figure 3: Text-level, Absolute Kendall correlations between human evaluations (%)

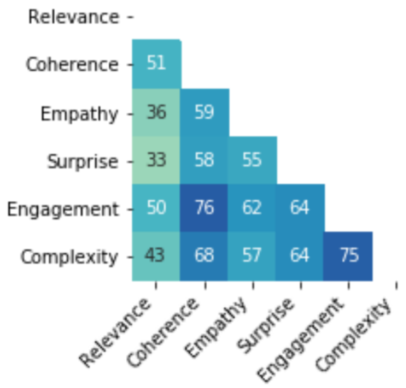


Figure 4: Text-level, Absolute Spearman correlations between human evaluations (%)

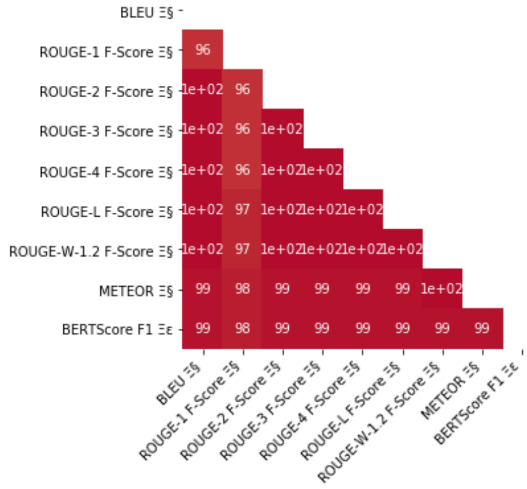


Figure 5: Text-level, Absolute Pearson correlations between AEM (%)

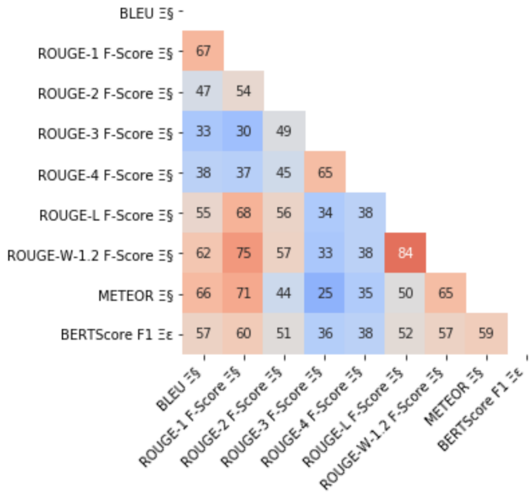


Figure 6: Text-level, Absolute Kendall correlations between AEM (%)

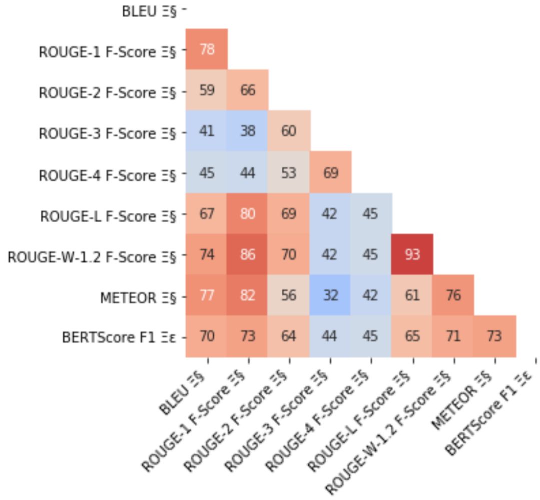
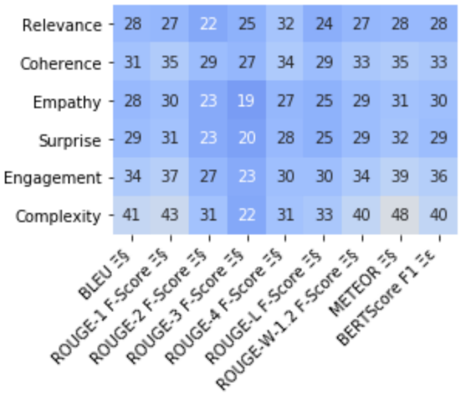


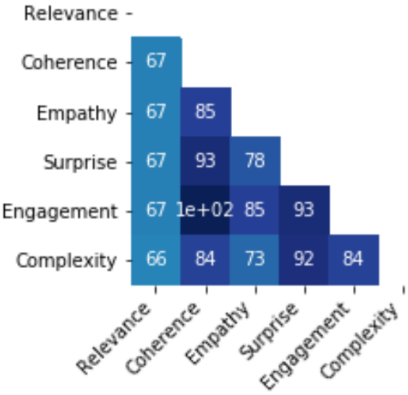
Figure 7: Text-level, Absolute Spearman correlations between AEM (%)



	BLEU Ξ§	ROUGE-1 F-Score Ξ§	ROUGE-2 F-Score Ξ§	ROUGE-3 F-Score Ξ§	ROUGE-4 F-Score Ξ§	ROUGE-L F-Score Ξ§	ROUGE-W-1.2 F-Score Ξ§	METEOR Ξ§	BERTScore F1 Ξε
Relevance	28	27	22	25	32	24	27	28	28
Coherence	31	35	29	27	34	29	33	35	33
Empathy	28	30	23	19	27	25	29	31	30
Surprise	29	31	23	20	28	25	29	32	29
Engagement	34	37	27	23	30	30	34	39	36
Complexity	41	43	31	22	31	33	40	48	40

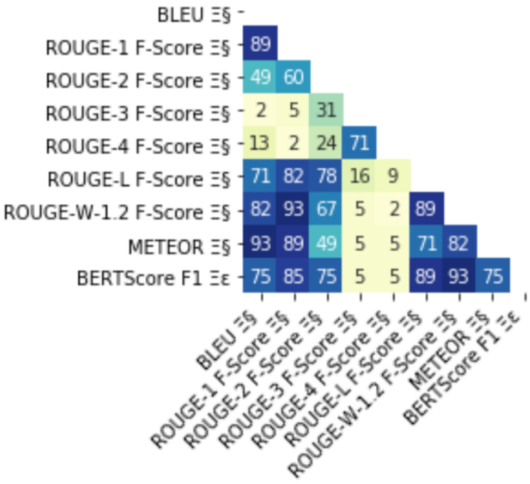
Figure 8: Text-level, Absolute Kendall correlations between AEM ans human criteria (%)

System-level



	Relevance	Coherence	Empathy	Surprise	Engagement	Complexity
Relevance						
Coherence	67					
Empathy	67	85				
Surprise	67	93	78			
Engagement	67	1e+02	85	93		
Complexity	66	84	73	92	84	

Figure 9: System-level, Absolute Kendall correlations between human evaluations (%)



	BLEU Ξ§	ROUGE-1 F-Score Ξ§	ROUGE-2 F-Score Ξ§	ROUGE-3 F-Score Ξ§	ROUGE-4 F-Score Ξ§	ROUGE-L F-Score Ξ§	ROUGE-W-1.2 F-Score Ξ§	METEOR Ξ§	BERTScore F1 Ξε
BLEU Ξ§									
ROUGE-1 F-Score Ξ§	89								
ROUGE-2 F-Score Ξ§	49	60							
ROUGE-3 F-Score Ξ§	2	5	31						
ROUGE-4 F-Score Ξ§	13	2	24	71					
ROUGE-L F-Score Ξ§	71	82	78	16	9				
ROUGE-W-1.2 F-Score Ξ§	82	93	67	5	2	89			
METEOR Ξ§	93	89	49	5	5	71	82		
BERTScore F1 Ξε	75	85	75	5	5	89	93	75	

Figure 10: System-level, Absolute Kendall correlations between AEM (%)

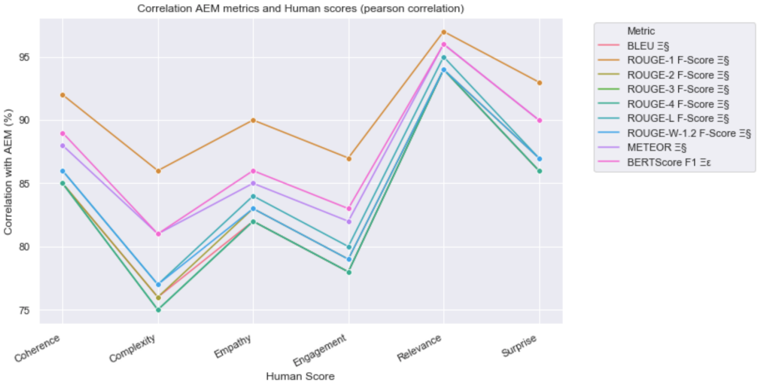


Figure 11

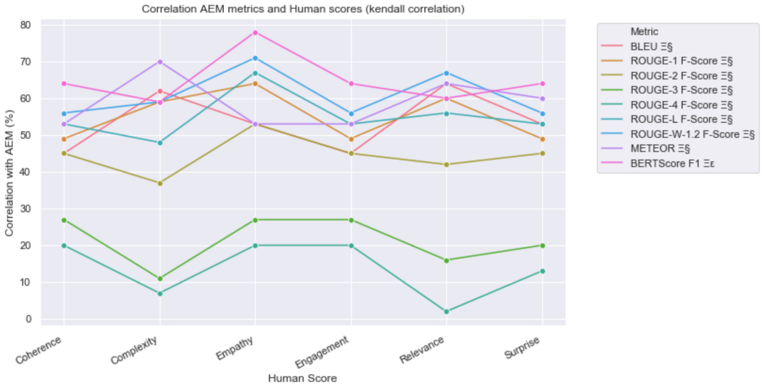


Figure 12

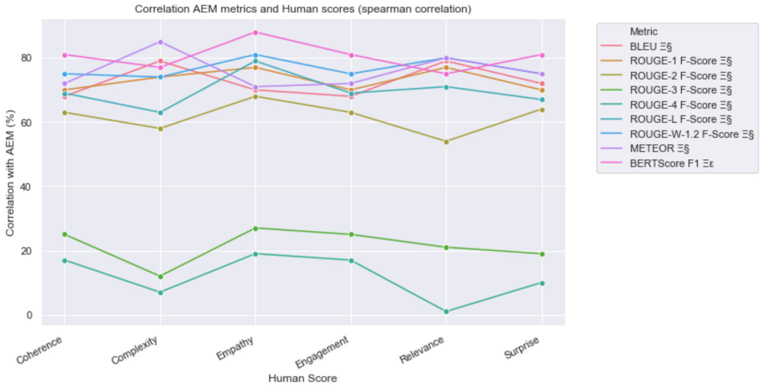


Figure 13