Study of automatic evaluation metrics applied to story generation in relation to human metrics

DESBOIS Vinciane
ENSAE IPParis
vinciane.desbois@ensae.fr

MILLET Clémence
ENSAE IPParis
clemence.millet@ensae.fr

Link Github

Abstract

Automatic story generation is a complex branch of NLP whose evaluation techniques have been less studied than for summarization or data-to-text. In this analysis (see our source code[1]), we will focus on the relevance of the different existing automatic metrics, both traditional and more recent, to evaluate this type of task. With the help of a dataset annotated by human evaluators, we compare automatic metrics to human metrics, look for correlations between them and observe the performance of automatic metrics in predicting some human metrics. Our results mainly show a high similarity between all automatic metrics and their difficulty in predicting human metrics, even when combined.

1 Problem Framing

1.1 Problems and related work

The question of automatic evaluation of text generation is crucial considering that language models are multiplying and that human evaluation is long and costly. Many studies have been conducted to propose and test the relevance of evaluation metrics. Meta-evaluation has been done in recent works especially for summarization (Bandhari and al., 2020 (9)) or question answering (Chen and al., 2019 (10), but less in the area of story generation that we decided to study here. This generation task is particular since the human evaluation criteria can be multiple: grammatical correctness of the text, coherence with the prompt (short instruction on the story to be generated) but also imagination, generated emotion or complexity of the story.

These particularities can make the task of story generation particularly difficult to evaluate. Not only can automatic metrics have difficulty taking into account so many subtle facets, but it is also complex to select the important human criteria to consider. Yet these criteria are crucial to assess the relevance of different automatic metrics. Based on the study of Chhun et al. (2022) (11), we propose to analyze the correlations of different automatic metrics with 6 human metrics adapted to story generation, and to test their relevance. Using the criteria put forward in this paper (11), we will base our analysis on the following human metrics: relevance, coherence, empathy, surprise, engagement and complexity. As one particular addition compared to this paper, we will focus on the MENLI metric, recently proposed by Chen and al. (2022 (12)). Being based on Natural Language Inference and particularly designed to be robust to adversarial attacks, MENLI could be able to give a complementary and renewed type of automatic evaluation.

Our goal is to determine whether automatic metrics are relevant for story generation, i.e., are correlated or predictive of human metrics, and potentially identify the most relevant ones. We will also look at a possible complementarity or similarity between the different automatic metrics, with a methodology similar to Colombo and al. (2022 (13)), to assess the relevance of combining metrics.

1.2 Dataset, generation systems and metrics

To perform our analysis, we rely on Hanna, a large dataset of Human-ANnotated NArratives for Automatic Story Generation (ASG) evaluation, proposed by Colombo et al (2022 (11)). It is an annotated dataset of 1056 stories produced by 10 different ASG (Automatic Story Generation) systems. Each of the 96 prompts is subject to a story generation by each of the 10 ASGs. And each generated story is evaluated by 72 existing automatic metrics, and annotated by three human per-

[1] https://github.com/nom-utilisateur/nom-depot

sons according to the criteria presented previously and a methodology made explicit in the paper by Colombo et al (11).

We studied 8 of the 10 models grouped in the dataset: BertGeneration (Rothe et al., 2020 (17)), CTRL (Keskar et al., 2019 (4)), RoBERTa (Liu et al., 2019 (2)), XLNet (Yang et al., 2019 (8)), GPT-2 (Radford et al., 2019 (3)), Fusion (Fan et al., 2018 (7)), HINT (Guan et al., 2021a (5)), and TD-VAE (Wilmot and Keller, 2021 (6)). We leave aside the 2 other versions of GPT included in HANNA.

As far as automatic metrics are concerned, we decided to simplify the analysis by choosing 19 metrics out of the 72 of HANNA, among which:

- 5 Reference-based string-based metrics (BLEU (14), ROUGE-1 F-Score (15), METEOR (26), chrF (24), CIDEr (25))
- 5 Reference-based embedding-based metrics (ROUGE-WE-3 F-Score (16), BERTScore F1 (17), MoverScore (18), DepthScore (19), BaryScore-W (21))
- 3 Reference-based model-based metrics (S3-Pyramid (30), SummaQA (29), InfoLM-FisherRao (28))
- 2 Reference-free string-based metrics (Novelty-1 (27), Repetition-1 (27))
- 2 Reference-free embedding-based metric (SUPERT-Golden (23), SUPERT-PS (23))
- 2 Reference-free model-based metrics (BARTScore-SH (20), BLANC-Golden (22))

We found it interesting to add to our analysis a recently developed metric that was based on slightly different principles than the others. We considered the MENLI metric proposed by Chen et al. (2022 (12)) whose principle is based on Natural Language Inference and which is particularly designed to be robust to adversarial attacks (introduction of a small perturbation in the input text that can distort the model prediction). This metric is also interesting because its source code directly foresees its association with other existing metrics in order to benefit from the complementary information brought by each of the two metrics.

Therefore, we selected two versions of this metric: one coupled with the BERTScore-F1 metric and the other associated with MoverScore. We then applied these two metrics to the whole dataset, increasing our number of evaluation metrics to 21. This gives us our final dataset on which we conducted our study.

2 Experiments Protocol

2.1 Correlation between metrics

In the first part of the study, we attempt to compare the evaluation of stories by the different metrics. We first analyze the correlation coefficients between the metrics, human and automatic, according to the Pearson formula. But following the reasoning of Colombo and al. (13), we consider that the primary role of the metrics is to discriminate between different models. The evaluation values given by the metrics are therefore less relevant than the classification that they propose between the different generation systems. Thus, we then study the correlations between model rankings using Kendall's τ measure. We compute the complementarity (or inverse similarity) between two metrics by averaging the distance between their generation model rankings for each prompt. It gives formally (as explained by Colombo and al. (13)) :

$$C(m_0, m_1) = \frac{1}{K}\sum_{k=1}^{K} d_\tau(\sigma_k^{m_0}, \sigma_k^{m_1})$$

where C is the complementarity coefficient, m_0, m_1 are the two metrics, K is the number of prompts, $\sigma_k^{m_0}$ and $\sigma_k^{m_1}$ are the rankings of the models for prompt k by each metric and d_τ is the normalized Kendall's distance (linked to τ the following way: $d_\tau = \frac{1-\tau}{2}$).
This measure allows us to see how the rankings of systems by different metrics differ from each other.

2.2 Predictions of human metrics with different types of metrics

In a second step, we study how some human metrics can be predicted by automatic metrics and by the other human metrics. To do so, we select two 'target' human metrics that we will successively try to predict: the Relevance variable and the Coherence variable. These two criteria seem to us to be the most general among the 6 human evaluation criteria and therefore potentially the most likely to be predicted by the automatic metrics. We use LightGBM (light gradient-boosting machine) models to conduct these explorations since they are optimized and fast models, adequate for our problem.

Following the methodology of Colombo et al (2022 (13)), we train three models to study three elements for each of the two target metrics chosen:

1. the performance of the automatic metrics in predicting a human metric
2. the performance of other human metrics to predict the human target metric
3. the performance of the combined automatic and human metrics to predict the human target metric

3 Results

3.1 Correlation between automatic metrics and human metrics

The correlation matrix directly shows that the human scores are highly correlated with each other and the automatic metrics are highly correlated with each other. On the contrary, the correlations between human scores and automatic metrics are quite low, which is disappointing and indicates that automatic metrics struggle to capture the essence of human evaluation criteria.

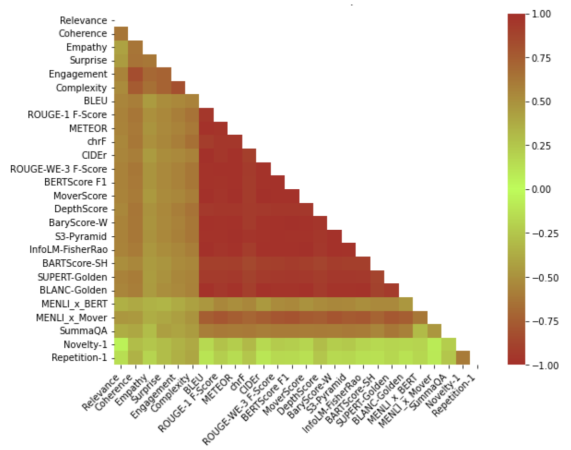


Figure 1: Correlation matrix (Pearson) between the different metrics

In Figure 1 (see Appendix 4.2 for a larger version), the red triangle corresponds to the correlations between the automatic metrics, we see that these are very close to 1 or -1 in most cases. On the other hand, the green vertical rectangle shows that the automatic metrics are poorly correlated with the human criteria. Finally, we notice that the reference-free variables 'Novelty' and 'Repetition' are poorly correlated to the others: they could therefore bring additional information, which can be evaluated with the complementarity measures based on the model rankings.

If we look at the ranking of the different models by the metrics, which is more meaningful for our analysis, we observe more or less the same phenomenon: the human scores tend to rank the models in the same way, which often differs from the ranking done by the automatic metrics.

On figure 2, we observe that even if there are patterns that are often verified (the GPT-2 model very well rated and the HINT model often at the end of the ranking for example), there are important differences between the automatic metrics and the human scores (this is particularly visible for the RoBERTa and XLNet models. Moreover, some metrics score very differently from human scores. CIDEr and 'MENLI x Mover' in particular present a quite different ranking. The measure of complementarity between metrics (via Kendall's distance, see Annexe 4.2) also shows little complementarity between metrics except for the metric Repetition which is very different from the others. BLANC and CIDEr also show more complementarity but they are not any more similar to the human metrics.

3.2 Prediction of human metrics

By training models on different metrics to try to predict successively the variable 'Relevance' and the variable 'Coherence', we find similar results to those of Colombo et al. (13). Indeed, surprisingly we manage to predict better these human scores with the remaining human scores, yet supposed to focus on different qualities of the generation, than with automatic metrics. And when we combine automatic metrics and human scores, we obtain only slightly better results compared to human metrics.

We also observe that the consistency variable is much better predicted than the relevance variable, both by human and automatic metrics, which shows that automatic metrics would focus more on the correctness and form of the text than on its consistency with the input text. Finally, the MENLI metrics (coupled with BERTScore and then with MoverScore) slightly improve the predictive performance of the models, thus providing some additional information without being decisive in the ranking of the importance variables (see Appendix 4.2). The best ranked metrics are often met-

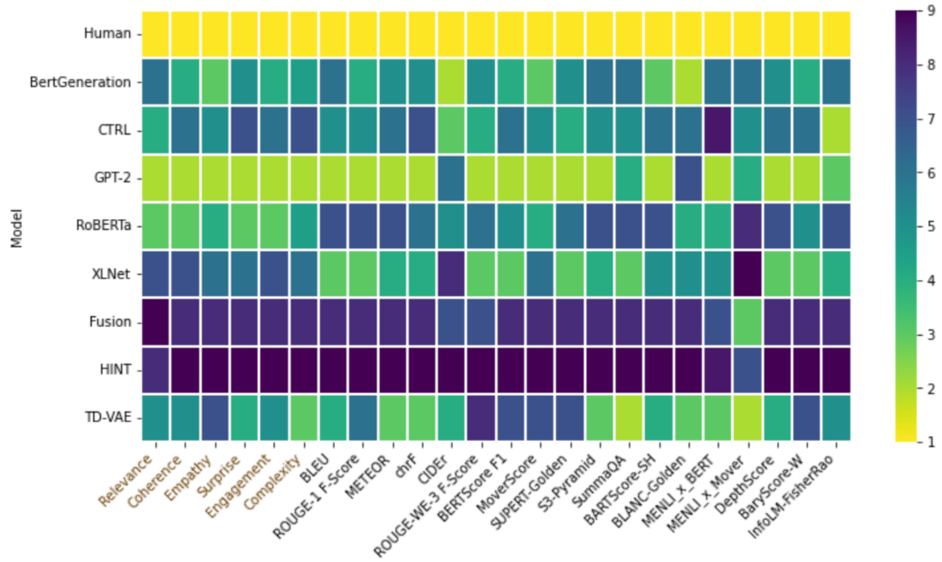


Figure 2: Ranking of the generation systems by the different metrics

Y =	**Relevance**	**Coherence**
AEM	0.855	0.618
AEM-M	0.848	0.625
H	0.806	0.447
AEM_H	0.748	0.425
AEM-M_H	0.756	0.419

Table 1: RMSE scores of LGBM models trained for different X variables (rows) and different target variables (columns)[2]

rics that we noticed as being a little less similar to the others, such as Repetition, Novelty or SummaQA. The model would therefore put them forward because they carry new information compared to the other metrics that are quite similar to each other. Some metrics have a very asymmetrical importance depending on the target score chosen: 'BARTScore-SH' for example takes a great importance in the prediction of the 'Coherence' score.

4 Discussion

4.1 Overall results

The suitability of automatic metrics for story generation was investigated and three main points were revealed:

- automatic metrics are highly correlated with each other and poorly correlated with human scoring
- human metrics are better predicted by other human criteria than by automatic metrics, even if they still provide information that allows for slightly better performance when coupled with human scores, especially those that are constructed completely differently (Repetition or Novelty for example)
- recent metrics such as BLANC, MENLI or InfoLM manage to bring some new information, sometimes according to a precise criterion, but remain highly correlated to the usual automatic metrics

4.2 Extension

We have focused here on a very specific task of text generation: story generation. This task is complex because of the creative freedom it leaves to the generation system and to all the aspects to be taken into account to evaluate it. Thus, this study does not necessarily generalize to all NLP tasks and it would be interesting to extend it on different datasets for a more exhaustive analysis of the performance of the automatic metrics.

Moreover, we have started to look at the different human criteria that each metric may be better able to predict. We could extend this by analyzing precisely which aspects of human evaluation are more accurately captured by each metric and then combine automatic metrics that would be close to different human criteria.

[2] The columns codes are AEM: Automatic metrics, AEM-M: Automatic metrics with MENLI, H: Human, AEM_H: Automatic and human metrics combined, AEM-M_H: Automatic and human metrics combined with MENLI

References

[1] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. *Leveraging pre-trained checkpoints for sequence generation tasks.* Transactions of the Association for Computational Linguistics, 8:264–280.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. *Language models are unsupervised multitask learners.* OpenAI blog, 1(8):9.

[4] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. *Ctrl: A conditional transformer language model for controllable generation.* arXiv preprint arXiv:1909.0585

[5] Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021a. *Long text generation by modeling sentence-level and discourse level coherence.* In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6379–6393, Online. Association for Computational Linguistics.

[6] David Wilmot and Frank Keller. 2021. *A temporal variational model for story generation.* arXiv preprint arXiv:2109.06807.

[7] 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.

[8] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *Xlnet: Generalized autoregressive pretraining for language understanding.* In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 5754–5764.

[9] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. *Re-evaluating evaluation in text summarization.* In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9347–9359, Online. Association for Computational Linguistics. 2020.

[10] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. *Evaluating question answering evaluation.* In Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pages 119–124, Hong Kong, China. Association for Computational Linguistics. 2019.

[11] Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. *Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation.* In Proceedings of the 29th International Conference on Computational Linguistics, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. 2022.

[12] Chen, Yanran Eger, Steffen. *MENLI: Robust Evaluation Metrics from Natural Language Inference.* 2022.

[13] Colombo, Pierre Peyrard, Maxime Noiry, Nathan West, Robert Piantanida, Pablo. *The Glass Ceiling of Automatic Evaluation in Natural Language Generation.* 2022.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing 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.

[15] 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.

[16] 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.

[17] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. *Bertscore: Evaluating text generation with BERT.* In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

[18] 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.* In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 563–578, Hong-Kong, China. Association for Computational Linguistics.

[19] Guillaume Staerman, Pavlo Mozharovskyi, Stéphan Clémençon, and Florence d'Alché Buc. 2021. *A pseudo-metric between probability distributions based on depth-trimmed regions.* arXiv preprint arXiv:2103.12711.

[20] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. *Bartscore: Evaluating generated text as text generation.* Advances in Neural Information Processing Systems, 34.

[21] Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021d. *Automatic text evaluation through the lens of Wasserstein barycenters.* In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10450–10466, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[22] Cleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. *Fill in the BLANC: Human-free quality estimation of document summaries.* In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, pages 11–20, Online. Association for Computational Linguistics.

[23] 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.

[24] Maja Popovic. 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.

[25] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. *Cider: Consensus-based image description evaluation.* In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 4566–4575. IEEE Computer Society.

[26] 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.

[27] Alexander R Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. *Summeval: Re-evaluating summarization evaluation.* Transactions of the Association for Computational Linguistics, 9:391–409.

[28] Pierre Jean A Colombo, Chloé Clavel, and Pablo Piantanida. 2022c. *Infolm: A new metric to evaluate summarization data2text generation.* In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 10554–10562

[29] Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. *Answers unite! unsupervised metrics for reinforced summarization models.* In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

[30] Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. *Learning to score system summaries for better content selection evaluation.* In Proceedings of the Workshop on New Frontiers in Summarization, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistic.

Annexe A

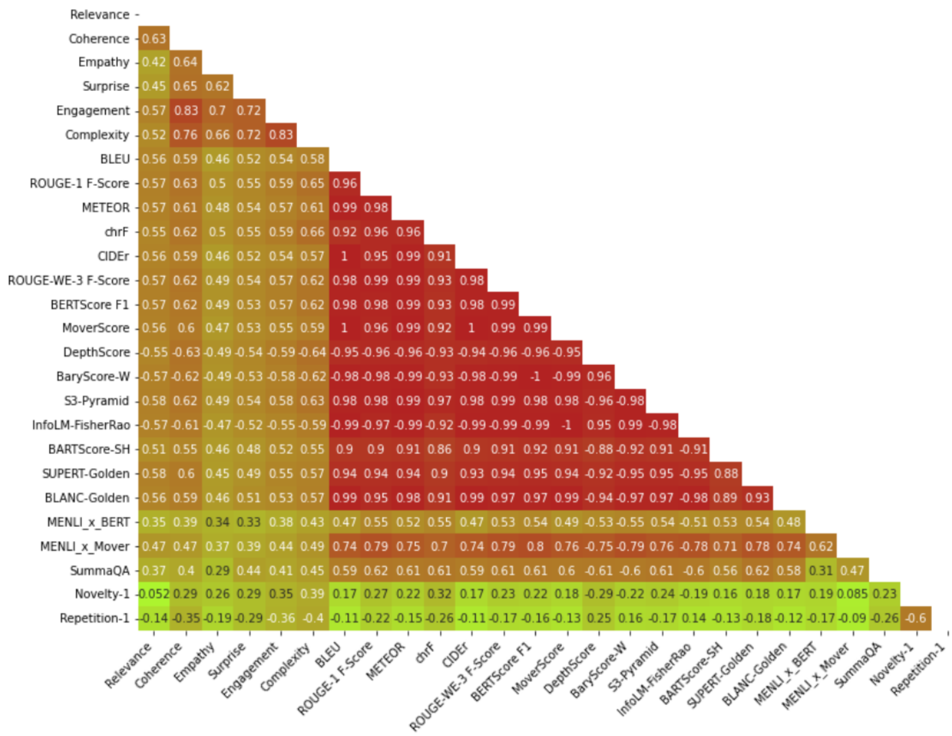


Figure 3: Correlations between metrics (Pearson)

Annexe B

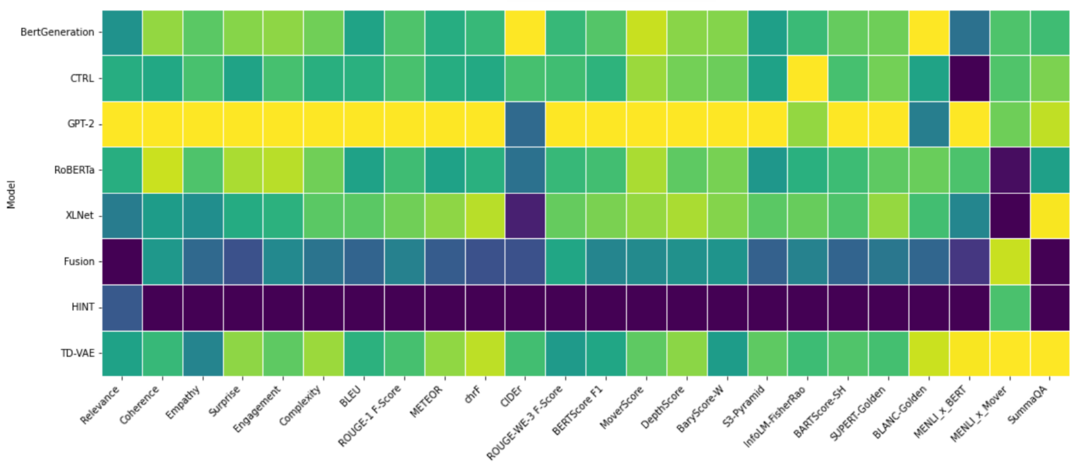


Figure 4: Scoring of the generation systems by the different metrics

Annexe C

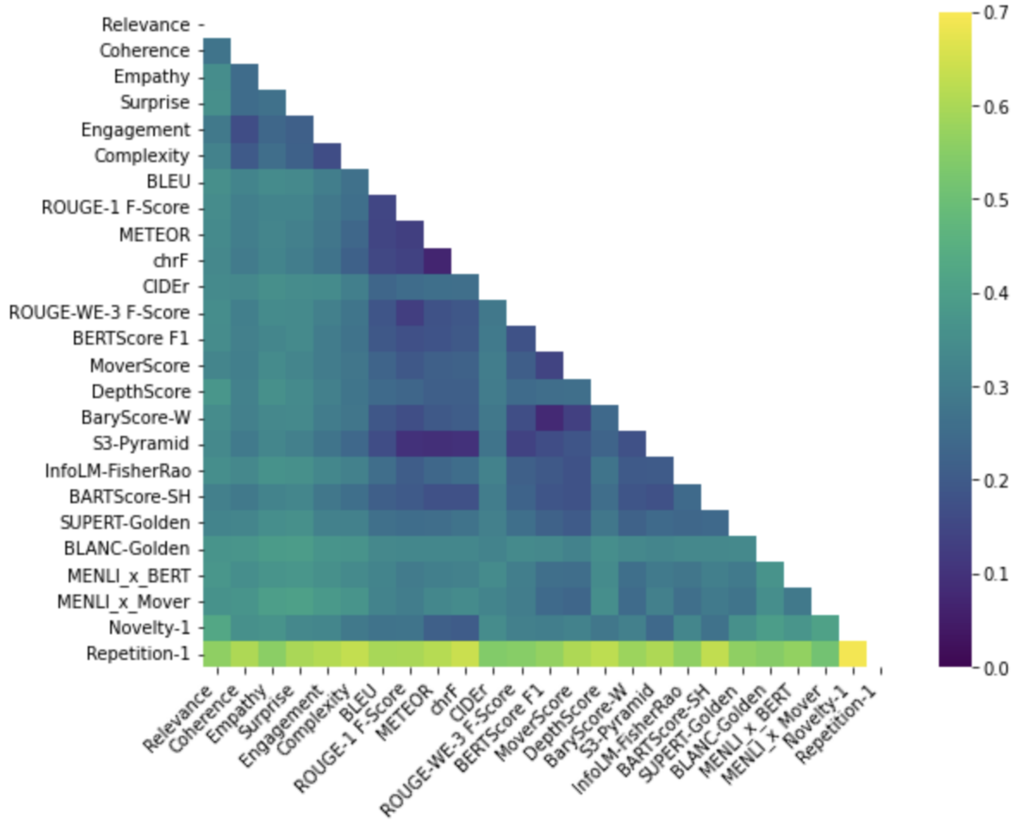
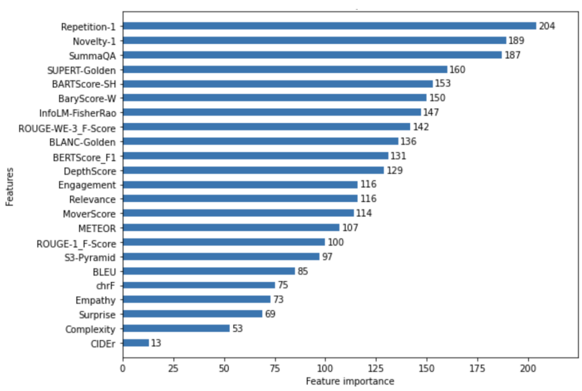
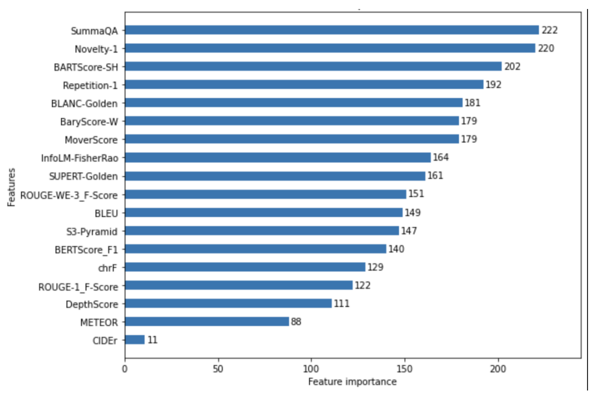


Figure 5: Complementarities between metrics (with Kendall's distance)

Annexe D



Prediction for the 'Relevance' metric



Prediction for the 'Coherence' metric

Figure 6: Feature importance for the human metrics prediction