The Glass Ceiling of Automatic Evaluation in Natural Language Generation

Pierre Colombo[1,2], **Maxime Peyrard**[3], **Nathan Noiry**[4] **Robert West**[3] **Pablo Piantanida**[5]
[1]MICS - CentraleSupélec, [2] Equall,
[3] EPFL, [4] Telecom Paris
[5] ILLS, CNRS - CentraleSupélec
colombo.pierre@centralesupelec.fr

Abstract

Automatic evaluation metrics capable of replacing human judgments are critical to allowing fast development of new methods. Thus, numerous research efforts have focused on crafting such metrics. In this work, we take a step back and analyze recent progress by comparing the body of existing automatic metrics and human metrics altogether. As metrics are used based on how they rank systems, we compare metrics in the space of system rankings. Our extensive statistical analysis reveals surprising findings: automatic metrics – old and new – are much more similar to each other than to humans. Automatic metrics are not complementary and rank systems similarly. Strikingly, human metrics predict each other much better than the combination of all automatic metrics used to predict a human metric. It is surprising because human metrics are often designed to be independent, to capture different aspects of quality, e.g. *content fidelity* or *readability*. We provide a discussion of these findings and recommendations for future work in the field of evaluation.

1 Introduction

Crafting automatic evaluation metrics (AEM) able to replace human judgments is critical to guide progress in natural language generation (NLG), as such automatic metrics allow for cheap, fast, and large-scale development of new ideas. The NLG fields are then heavily influenced by the set of AEM used to decide which systems are valuable. Therefore, a large body of work has focused on improving the ability of AEM to predict human judgments.

Human judgment data is typically employed to decide which metric to select based on correlation analysis with human annotations (Rankel et al., 2013; Owczarzak et al., 2012; Graham, 2015). In this work, we take a step back and investigate the relationship between existing AEM and human judgments globally. We do not make metric recommendation but reflect upon the global progress in the field of automatic evaluation. Our work is motivated by the findings of Fig. 1. It depicts the improvement over time, when new metrics were introduced, in the ability to fit human judgments when using all existing metrics as features. The fit is measured by the correlation with humans of a trained classifier in a 5-fold cross-validation setup. Remarkably, our observations indicate that there have been only minor incremental improvements, and the progress in recent years appears to be reaching a saturation point.

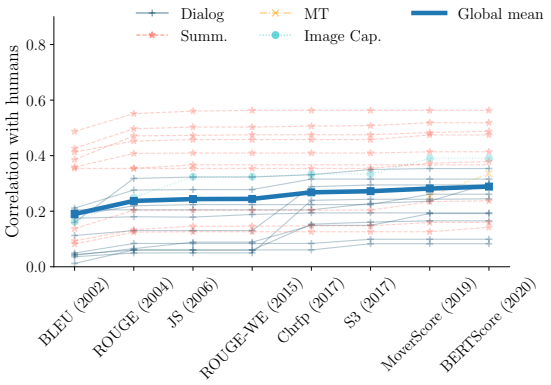

Figure 1: **Correlation with humans over time considering all existing metrics combined.** On the x-axis: evaluation metrics ordered by their release time; y-axis: utterance-level Kendall's τ with human when training a model to fit human judgments with all metrics available at the time (5-Fold cross-validation with XGBoost regressor). The dotted lines represent different human annotations and datasets. Different variants of the same metrics (like ROUGE-1 and ROUGE-2) are averaged. The datasets and metrics are described in Sec. 2.

Recent works emphasized the importance of viewing metrics in terms of how they rank systems instead of just comparing score values (Novikova et al., 2018; Peyrard et al., 2021; Colombo et al., 2022). Indeed, not only ranking is a more robust framework of comparison, it is also more aligned with the way metrics are used: identifying and extracting the "best system". Thus, we perform

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Findings Papers), pages 178–183
November 1–4, 2023.

our analysis in the space of rankings. i.e., how do metrics rank systems? By analyzing 9 datasets covering 4 tasks and 270k scores, we made the following observations:

Findings. (i) Automatic metrics are much more similar to each other, in terms of how they rank systems, than they are to human metrics. It means that AEM, even the more recent transformer-based ones are similar to the older ones when used in practice (ROUGE and BLEU). (ii) This lack of complementarity results in the inability to fit human judgments even when all these metrics are taken together as features for a classifier predicting humans. (iii) Quite surprisingly, different human dimensions – different annotations guidelines such as *readability*, or *content fidelity* – are very predictive of each other, whereas AEM are much less predictive of humans. This finding is striking because human metrics are designed to capture different and independent aspects of quality whereas AEM have been selected precisely for their ability to match humans. We would expect human metrics to be uncorrelated and automatic metric to be highly correlated with humans but we observe the opposite. First, it casts serious doubt about the ability of AEM to replace human judgments. Then, the correlation between independent human annotations of quality hints at some latent inherent *goodness* of systems: good systems are good in different aspect whereas bad systems are bad across all aspects.

Our findings have several consequences that can inform future research. Newly introduced metrics are not complementary to previous ones, resulting in small global improvements. As a way forward, we propose that research, instead of crafting metrics that maximize correlation with humans, focus on making metrics that **also** aim to be explicitly complementary to the set of existing metrics. This would enforce maximal marginal gain and ensure that the field, as a whole, makes progress towards capturing the complexity of human annotations.

For practitioners, it is common practice to report several AEM in the hope to get a better view of system performances. However, reporting several metrics that all produce similar rankings does not bring useful additional information. With our proposal, reporting a set of complementary metrics would better serve the intended purpose.

To help research build upon our work and use our measure of complementarity, we make our code available at `github`.

2 Methodology

Terminology. Let $\mathcal{X}$ be the space of possible outputs for an NLG task. An NLG metric is a function $m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ which, from a given textual candidate $C \in \mathcal{X}$ and corresponding reference $R \in \mathcal{X}$, computes a score $m(C,R)$ reflecting the properties that C should satisfy (*e.g.* fluency, fidelity...). Of course, it is illusory to summarize subtle semantic properties by a single scalar and one is rather seeking for metrics that are able to discriminate between different systems. In fact, crafted AEM are evaluated by comparison to human judgments: one usually computes ranking correlations such as the Kendall's τ. Higher correlations indicate that the AEM is a better replacement for the human metrics.

Encoding metrics with rankings. Since the usage of NLG metrics is to rank systems, we choose to represent an NLG metric, automatic or human, by the ranking it induces on a set of systems or utterances. More formally, for $S \geq 1$ NLG systems evaluated on a dataset made of $U \geq 1$ utterances, there exists a natural ranking representations of m:

Each utterance $u \in \{1,\dots,U\}$ induces a ranking $\sigma^m_u \in \mathbb{R}^S$ of the S systems seen as a vector σ^m_u, where $\sigma^m_u[s]$ is the rank of system $s \in \{1,\dots,S\}$. For a system s, the representation of a metric m, noted $\sigma^m[s]$, is sum of *rankings over the utterances*:

$$\sigma^m[s] := \sum_{u=1}^{U} \sigma^m_u(s) \in \mathbb{R}^N. \tag{1}$$

We call this **System level representation.**
Symmetrically, each system $s \in \{1,\dots,S\}$ induces a ranking $\rho^m_s \in \mathbb{R}^U$ of the U utterances, where $\rho^m_s[u]$ is the rank of utterance u. The **Utterance level representation** of m is sum of *rankings over the systems*:

$$\rho^m[u] := \sum_{s=1}^{S} \rho^n_s \in \mathbb{R}^K. \tag{2}$$

Using the space of rankings has been shown to be more robust than the raw scores as it is less sensitive to outliers and statistical variations (Novikova et al., 2017; Peyrard et al., 2021; Colombo et al., 2022). Furthermore, this representation is closely tied to Borda counts, which enjoys theoretical properties: the ranking induced by $\sigma^{m,S}$ is a 5-approximation of the Kemeny-consensus which is a good notion of average in the symmetric group

(Kemeny, 1959; Young and Levenglick, 1978; Coppersmith et al., 2006). It is moreover the fastest approximation of the Kemeny-consensus whose computation is NP-hard (Ali and Meilă, 2012).

Complementarity. We measure the *complementarity* between two metrics – humans or automatic – by the average over utterances of the distance between their rankings of systems. Formally, for two metrics m_0 and m_1, complementarity is given by:

$$C(m_0, m_1) := \frac{1}{U} \sum_{u=1}^{U} d_\tau(\sigma_u^{m_0}, \sigma_u^{m_1}), \qquad (3)$$

where d_τ is the normalized Kendall's distance between the vectors of rank. It is related to the Kendall's rank correlation τ by: $\tau = 1 - 2d_\tau$.

Similarly, we define the complementarity between a metric m_0 and a set of other metrics $\boldsymbol{m} := \{m_i\}_{i=1,\dots,l}$, as the average pairwise complementarity:

$$C(m_0, \boldsymbol{m}) = \frac{1}{l} \sum_{i=1,\dots,l} C(m_0, m_i). \qquad (4)$$

Complementarity measures the extent to which a metric ranks systems differently than another metrics or a set of other metrics. Whether comparing two metrics or a metric with set, it is a number between 0 and 1 where 0 indicates that the metrics rank systems in the exact same order and 1 indicates the exact opposite order. In between, it counts the number of inversions between the two rank lists normalized by the number of possible pairs of systems.

2.1 Dataset description

To ensure a wide coverage of NLG we focus on four different problems *i.e.*, dialogue generation (using PersonaChat (PC) and TopicalChat (TC) (Mehri and Eskenazi, 2020)), image description (relying on FLICKR (Young et al., 2014)), summary evaluation (via TAC08 (Dang and Owczarzak, 2008), TAC10, TAC11 (Owczarzak and Dang, 2011), RSUM (Bhandari et al., 2020) and SEVAL (Fabbri et al., 2021)), and translation (focusing on multilingual quality estimation (MLQE) Ranasinghe et al. (2021)).

For each task, we gather datasets and rely on AEM such as JS [1-2] (Lin et al., 2006), BLEU (Papineni et al., 2002; Post, 2018), Chrfpp (Popović, 2017), S3 (both variant pyr/resp) (Peyrard et al., 2017), ROUGE (Lin, 2004) (including 5 of its variants (Ng and Abrecht, 2015)), BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019). For MLQE we solely consider several version of BERTScore, MoverScore and ContrastScore. The human evalutions criterion are specific to each dataset and will be identified by starting with an H:. Overall, our final datasets gather over 270k scores.[1]

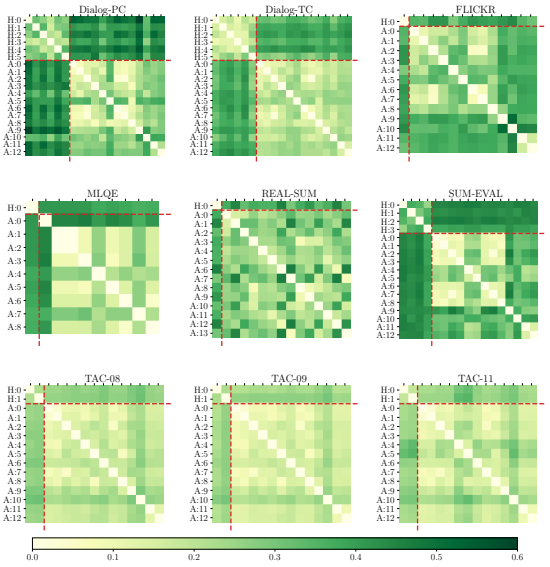

Figure 2: **Complementarity**: For each dataset, the pairwise complementarity between each pair of metrics as computed by Eq. 3 both human and automatic. In these matrix plot, symmetric by design, we ordered metrics to have the human one first and the automatic ones after, the red lines trace the limit between humans and AEM.

3 Experiments

Finding 1: Automatic metrics are similar to each other much more than they are to human metric. In Fig. 2, we report the pairwise complementarity between each pair of metrics as computed by Eq. 3 for both human and AEM. When aggregated over pairs and over datasets, we obtain an average complementarity between: (i) two human metrics of $.16 \pm .01$, (ii) two AEM of $.20 \pm .01$ and (iii) a human and an automatic metric of $.35 \pm .02$.

Importantly, we observe across datasets low complementarity, i.e., strong similarity, between AEM, low complementarity between human metrics but

[1]The selection of these metrics was driven by their widespread usage and recognition, as reported in numerous research papers. In order to streamline our analysis and address practical considerations, we opted to exclude recent metrics, such as those based on GPT-3/4, due to their expensive evaluation requirements on large benchmarks and reliance on proprietary models with undisclosed datasets.

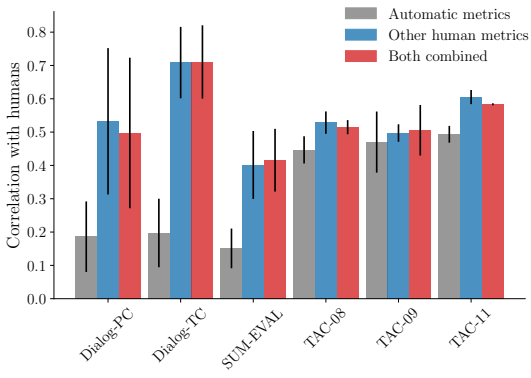

Figure 3: **Human metrics are significantly more predictive of each other than AEM.** On this plot, we report the 5-fold cross-validated result of fitting an XGBoost regressor on various feature sets: (i) all available AEM, (ii) other human metrics when available, and (iii) both automatic and human metrics. The fit is measured as the average instance-level correlation in the test set.

high complementarity, i.e., low similarity, between automatic and human metrics.

We draw two conclusions from this analysis: (i) AEM rank systems similarly but (ii) differently than humans. There is some nuances across datasets. The effect described above is particularly strong in the Dialog, MLQE and SUM-Eval datasets. In particular, we notice that TAC datasets, from the summarization task, have lower complementarity in general, meaning that all metrics, human and automatic, are more similar. Indeed, a lot of works have relied on these datasets to develop new metrics. The more recent REAL-SUM and SUM-Eval reveal much lower metric similarity.

Finding 2: Automatic metrics even all combined do not explain human metrics. If AEM are rather different than human metrics, we might wonder whether it is possible to get a good approximation of human judgments by combining existing AEM together. To account for possible correlations, we rely on XGBoost regressors with 5-fold cross-validation to predict human judgments. The training is performed on three different features space: *(i)* AEM only, *(ii)* other human metrics only and *(iii)* both sets of metrics combined. We compute the Kendall's τ between predictions and ground truths and report the results in Fig. 3.

The plot confirms that AEM struggle to capture human judgment subtlety: correlation rarely exceeds .4 on held-out data. In contrast, human metrics are much more predictive of each others, even if they are often supposed to capture different concepts. Finally, it is worth noting that adding AEM to human ones do not marginally improve the prediction power. These findings cast shadows over recent progress in the field.

4 Discussion

Our analysis reveals that studied automatic metrics are not complementary, and recent automatic metrics actually capture the same properties of human judgments as older ones. Furthermore, the studied metrics are not strong predictors of human judgments. Quite surprisingly, other human metrics which are often designed to be independent of each other end-up being more predictive of each other than automatic metrics. This predictability of human metrics from one another can be explained due to the available datasets: when a system is good at extracting content, it is also often good at making the content readable, when a system is bad it is often bad across the board in all human metrics. However, the fact the considered automatic metrics are less predictive than other human dimensions casts some shadow over recent progress in the field. It shows that the current strategy of crafting metrics with slightly better correlation than baselines with one of the human metrics has reached its limit and some qualitative change would be needed.A promising strategy to address the limitations of automatic metrics is to report several of them, hoping that they will together give a more robust overview of system performance. However, this makes sense only if automatic metrics measure different aspects of human judgments, i.e., if they are complementary. In this work, we have seen that metrics are in fact not complementary, as they produce similar rankings of systems.

Proposition for future work To foster meaningful progress in the field of automatic evaluation, we propose that future research craft new metrics not only to maximize correlation with human judgments but also to minimize the similarity with the body of existing automatic metrics. This would ensure that the field progresses as whole by focusing on capturing aspects of human judgments that are not already captured by existing metrics. Furthermore, the reporting of several metrics that have been demonstrated to be complementary could become again a valid heuristic to get a robust overview of model performance. In practice, researchers could re-use our code and analysis to enforce complementarity by, for example, enforcing new metrics to have low complementarity as measured by Eq. 3.

5 Limitations

Even though we have considered a representative set of automatic evaluation metrics, new ones are constantly introduced and could be added to such an analysis. Similarly, new datasets could be added to the analysis and impact the results. In an effort to make our findings relevant in the long run, we release an easy-to-use code base to replicate our analysis with new metrics and datasets.

Like the majority of analysis on automatic evaluation metrics, ours rely on the assumption that human judgments are valid and meaningful. However, some works have questioned the quality of human judgments in standard datasets.

References

Alnur Ali and Marina Meilă. 2012. Experiments with kemeny ranking: What works when? *Mathematical Social Sciences*, 64(1):28–40.

John J Bartholdi, Craig A Tovey, and Michael A Trick. 1989. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241.

Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. 2022. What are the best systems? new perspectives on nlp benchmarking. *arXiv preprint arXiv:2202.03799*.

Don Coppersmith, Lisa Fleischer, and Atri Rudra. 2006. Ordering by weighted number of wins gives a good ranking for weighted tournaments. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 776–782.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. In *TAC*.

Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622.

Alexander R Fabbri, Wojciech Kryściński, 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.

Yvette Graham. 2015. Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137. Association for Computational Linguistics.

Ian T Jolliffe and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

John G Kemeny. 1959. Mathematics without numbers. *Daedalus*, 88(4):577–591.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 463–470.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9. Association for Computational Linguistics.

Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the tac 2011 summarization track: Guided task and aesop task. In *Proceedings of the Text Analysis Conference (TAC 2011), Gaithersburg, Maryland, USA, November*.

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 ACL*, pages 311–318, Philadelphia, Pennsylvania, USA. ACL.

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.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of nlp systems. *arXiv preprint arXiv:2110.10746*.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second WMT*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2021. An exploratory analysis of multilingual word-level quality estimation with cross-lingual transformers. *arXiv preprint arXiv:2106.00143*.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 131–136, Sofia, Bulgaria. Association for Computational Linguistics.

H Peyton Young and Arthur Levenglick. 1978. A consistent extension of condorcet's election principle. *SIAM Journal on applied Mathematics*, 35(2):285–300.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert.

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.