Beyond ROUGE: Applying an ELO algorithm to rank model performances in summarization

Romain Harang[1]
[1]The University of Tokyo
romain-harang@g.ecc.u-tokyo.ac.jp

Abstract

We apply an ELO-based algorithm to evaluate the performances of Language models in text generation by generating one-on-one encounters in which a judge function determines the outcome. We decided to apply this to the text summarization task on the CNN-Dailymail Dataset using state-of-the-art models and GPT3 and GPT4 in 0-shot as the judge function and the difference in ROUGE scores and compare with ROUGE.

We found that this approach is weakly correlated to ROUGE; a correlation can only be found after assuming a significant random noise and giving a radically different ranking of the models on the task. In particular, GPT4 is preferred as a summarizer. We also found that our results contradict the assumption that the ground truth is best for the CNN-Dailymail Dataset, comforting previous findings. This research opens avenues for a more comprehensive understanding of language model performance in text summarization tasks.

1 Introduction

With the advent of Language Models [1], tasks related to text generation have known tremendous developments. These models were proposed on Translation and achieved state-of-the-art performance on the WMT 2014 shared task; on abstractive summarization, this new approach offered similar results to RNN [2] on its first iteration. Nowadays, LLMs are used and preferred in all cases.

Evaluating the performance of generative algorithms for text has always been a crucial yet sometimes overlooked aspect of the process. Usually, metrics are not updated throughout the lifetime of a Dataset. For text Summarizing and Question Answering, it will be ROUGE [3], and for Translation, Captions will be BLEU [4]. Many other variants also incorporating language modeling and word embeddings have been proposed, showing each time better correlation to human judgment than ROUGE and BLEU in a limited scope; those metrics have limited actual use, however (apart from BERTscore [5], that is sometimes used along ROUGE). Later re-evaluation efforts [6] have shown no significant difference between these new metrics and ROUGE and BLEU in their correlation with human scoring nor in identifying the best model given a specific task.

Until recently, evaluation was centered around scoring the similarity of the generated text to a ground truth. But LLMs outputs are now preferred by humans in many cases [7]. For that reason, using humans to compare model outputs is gaining traction [8], utilizing ELO-based algorithms. A high correlation to human judgment was also found using LLMs as judges. However, no comparisons to other metrics were made nor to a ground truth when available; this work addresses that.

2 Related Works

ELO rating: Up until recently, there have been minimal attempts at using any ELO algorithm while evaluating the performances of generated text. Human evaluations based on pair-wise direct comparison of generated text using human annotators have become popular with, for example, the release of ChatBot arena that uses an ELO algorithm to judge the comparative performances of LLMs [8] [9]. LLMs have started to be used; however, when our experiments were conducted, no public data was available on the usage of LLMs in that context, but it has since been increasingly used as well [10]. At the time of these findings, there have been no attempts at comparing an ELO algorithm to currently used metrics or the ground truth in any datasets, to the best of our knowledge.

Text Summarization evaluation: There has been ample review and evaluations of the different metrics and their inner correlations and correlation to the human Judgment [6]. Coupled with the fact that it has been shown [7] that, for example, in summarization, Humans will prefer GPT-3 summaries to Ground truth in the CNN Dailymail Dataset. Therefore, using metrics based on the sentence similarity to Ground truth has shown its limits. Furthermore, the usage of ROUGE has also been criticized for longer texts; it has yielded luster lacking results in [11].

3 Pair-wise comparison method and ELO algorithm

3.1 Direct comparison

We are given a pair of texts $(text\ 1, text\ 2)$; we use a judge to decide whichever is more fitting for the task, in our case, the better summary. We use LLMs (GPT-4 or GPT-3) as judges and ROUGE (by taking the text with higher ROUGE). When using the LLMs, we also use the article text as a reference for judgment, the reference for ROUGE being the ground truth. This allows us to evaluate the ground truth as well with the LLMs.

3.2 The ELO algorithm

We used the previous implementation of the ELO algorithm [12] to build our own. The goal of this algorithm is to, given two arbitrary models m_i and m_j and the generated output $x = (t_i, t_j)$ estimate the probability that t_j is preferred to t_i by our comparison model. Each model is attributed a fitness or ELO score; then this probability is given by $p = \sigma(\frac{ELO(j)-ELO(i)}{b})$ where $\sigma(x) = \frac{1}{1+10^x}$ and b is an interpretation hyper-parameter; ie, a difference in ELO of b corresponds to a ratio of probabilities of 10 ($\frac{\sigma(1)}{\sigma(-1)} = \frac{\sigma(1)}{1-\sigma(1)} = 10$).

Our working hypothesis is that such a fitness score exists. The algorithm will try to estimate it, similar to previous method assumptions. Still, contrarily to them, such a score is not computed directly (at least not necessarily). Instead, such scoring is implicitly inferred by estimating the ELO.

The iterative algorithm to determine the ELO will be based on the maximization of the Log-likelihood function of our problem. We use the update function based on Stochastic gradient descent from [12] (details in Annex)

4 Applying the ELO algorithm to benchmark datasets

4.1 Datasets

We decided to limit ourselves to the summarization task on the CNN-Dailymail Dataset [2]. We will restrict ourselves to using models on the test split (11.5k pairs of articles/summaries are included). An important caveat with this dataset is that although it is the most popular for abstractive text summarization, the summaries were not constructed as such but instead as a concatenation of highlights from the article. From this, it is reasonable to assume that a human or a generative algorithm can find a better option than the ground truth.

4.2 Models

State-of-the-art models for the CNN-Dailymail were selected for our set of experiments:

- BART [13]
- DistilBART (obtain using the principles in [14] by S. Shleifer)
- PEGASUS [15]
- BRIO [16]
- GPT-4 [17]

All models apart from GPT-4 have been trained explicitly and specifically for this task in the CNN-Dailymail train split as well as other compilations of news articles or summarization datasets such as X-Sum [18] and GigaWord. GPT-4 will be used on its 0-shot form, which means the input will be of that form: <prompt><article>. Below in Fig. 1 is an evaluation of the models used based on ROUGE-1,2 and L.

Model	R-1	R-2	R-L
BART	0.366	0.163	0.344
DistilBART	0.384	0.173	0.361
PEGASUS	0.395	0.181	0.372
BRIO	**0.438**	**0.196**	**0.413**
GPT-4	0.275	0.085	0.251

Figure 1 ROUGE-1, 2 and L scores on the test split of the chosen models

4.3 Using ROUGE as the comparison's model

ROUGE provides us with a straightforward way of comparing two generated summaries. We can select the win-

ning party with the highest ROUGE (1,2 or l). For our specific experiment, we decided to use ROUGE-l, take $b = 400$, $\theta_0 = (1000, ..., 1000)$, and $\beta = 50$ with scheduled decrease until 1, below in fig. 2 we can see that the algorithm converges.

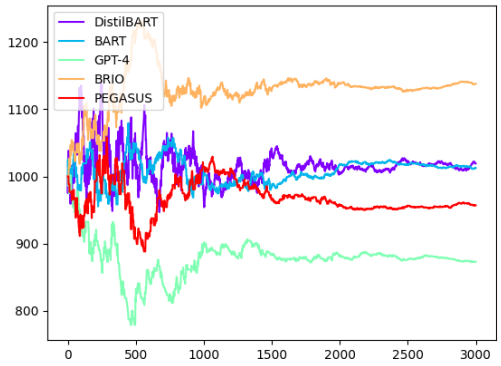

Figure 2 Convergence of the ELO algorithm for ROUGE-l

By taking the final values for the ELO and comparing them with ROUGE scores, we get:

Model	R-1	R-2	R-L	ELO
BART	0.366	0.163	0.344	1010
DistilBART	0.384	0.173	0.361	1020
PEGASUS	0.395	0.181	0.372	960
BRIO	**0.438**	**0.196**	**0.413**	**1140**
GPT-4	0.275	0.085	0.251	873

Figure 3 Scores for ROUGE compared to ELO R-L

Apart from PEGASUS, there is a good consistency between the two aggregation methods for the ROUGE score.

4.4 Using LLMs as the comparison's model

Using the same hyperparameters for the ELO algorithm, we applied different judgment methods. Based on a 0-shot, Ground-truth-free approach. Our prompt becomes then <Instructions><Article><Candidate_0><Candidate_1> then the output will be $y \in [0, 1]$ depending on the preferred candidate. The models used are GPT-3 and GPT-4. We also decide to include the ground truth as a possible candidate. To account for positional biases, the candidate orders are systematically randomized.

If we compare the final ELO scores to the Rouge-l score, we get Fig. 4.

The first observation from this experiment is that GPT-4 is preferred in this context. Especially in the case where the model used to judge is GPT-4. Even though text generated

Model	ELO-4	ELO-3	R-L
GPT-4	**1617**	**1204**	0.251
BART	1037	1024	0.344
BRIO	931	966	**0.413**
Baseline	816	988	-
PEGASUS	798	850	0.372
DistilBART	798	966	0.361

Figure 4 Final scores for GPT-4 and 3 named ELO-4 and 3 r. and Rouge-L

with GPT-4 will get a higher confidence score according to GPT-4, it doesn't automatically translate to GPT-4 having a stronger preference for itself in our task, at least in theory. We see, however, that this is the case. Also, the ground truth is not always preferred to machine-generated text.

Notably, the ELO value has meaning only in a relative sense; an essential difference in ELO means a more significant comparative advantage. For example, with ELO GPT-4, we have an offset of around 600 between GPT-4 and BART, translating to a comparative advantage of 32 times likelihood. But with ELO GPT-3, the difference is only 200, so it is an advantage of only three times.

4.5 Validation: Analysis of the correlation between GPT-3 and 4 judgment outcomes and ROUGE scores

We have seen that regardless of the aggregation method, ROUGE and judgment scores based on GPT with 0 shots give very different results. On the granular level, however, we have not looked if there is any correlation between these scores.

We decided first to calculate the Cohen's Kappa score between GPT-3,4 and ROUGE pair-wise. We can't compare judgment outcomes to ROUGE scores directly, so we use the outcomes based on the ROUGE difference we introduced in ELO-ROUGE. We use Cohen's kappa in a 2-label format (classified as [0,1] for the first and second models, respectively). We aggregate the results for each pair of models below:

Method 1	Method 2	Cohen Kappa
GPT-3	ROUGE	-0.021
GPT-3	GPT-4	0.32
GPT-4	ROUGE	-0.035

Therefore, there is no agreement between GPT-3 or 4 and ROUGE (performs as bad as random). Concerning GPT-3 and GPT-4, they only have a weak, limited agree-

ment with one another. (Good agreement starts at Cohen's kappa of 0.6 according to common heuristics). This tends to show no correlation between ROUGE and GPTs. However, we didn't use the full range of the ROUGE score. It is common practice to evaluate the correlation of 2 metrics of that type by using Spearman's ρ, used, for example, in [6]. We can't here as one of them is not continuous but ordinal. The workaround is to observe the distribution of ROUGE's score for a given label, either model 1 or 2; we get two distributions to compare. One way is to compare the mean of each of the two distributions similarly to Cohen's kappa results. We don't get any different significative results that way.

While manipulating the data, it came to light that if we looked at the histogram of the proportions of classified model 1 at different ROUGE scores, there seemed to be an inverse correlation between these two quantities. As a proposed modelization of that phenomenon, we assume that the probability given the rouge score difference $R = R_2 - R_1$ is going to be 1/2 at the point of ROUGE difference $m = 2(\rho - \frac{1}{2})$ where ρ is the ratio of labels model 2. This probability will then vary linearly until the ROUGE difference hits its boundaries (-1 and 1) or the probability saturates at 0 or 1. Given that we have $P(Label = Model_1|R) = m + aR$, in this definition, we expect the slope (a) to be negative, as we presume that there is a negative correlation between ROUGE difference and the probability that model 1 is chosen as a victor. We compare GPT-4 (ELO) and ROUGE-L only for pairs of models where, in the aggregate, either model's selection ratio is not lower than 10% (not enough data points). We approximate this distribution using linear regression in Fig. 5:

model 1	model 2	slope	r^2
DistilBART	PEGASUS	-0.96	0.78
BART	baseline	-0.31	0.68
BRIO	PEGASUS	-0.66	0.65
BART	BRIO	-0.63	0.61
DistilBART	baseline	-0.44	0.46
PEGASUS	baseline	-0.54	0.42
BRIO	baseline	-0.33	0.40
DistilBART	BRIO	-0.38	0.32
DistilBART	BART	0.67	0.20

Figure 5 Regression's results

As hypothesized, we get a negative slope (the lower, the better) consistently apart from one case where the coefficient of determination is very low, concluding that this is an outliner.

5 Discussion

Our proposed method shows a higher score for some LLMs in this work against the human-made ground truth. Generally, it is best practice to always assume the human-generated text as the best possible or at least a sufficient baseline for which the model should aim. In the case of the CNN-Dailymail Dataset, the summaries proposed are a concatenation of highlights from a news article; it is not surprising to find here and in previous research that it could be improved.

Instead of the ELO algorithm, we could have used win rates as the aggregation method. Both methods could be valid in our particular case, but ELO allows for more meaningful results at a lower cost (i.e., number of pair-wise encounters), as shown in [19].

We argue that even if the Cohen kappa results for GPT-4 and ROUGE suggest that trying to infer the label given by GPT-4 with the sign of the ROUGE score for individual instances of scoring is going to perform similarly to random (in fact, slightly worse); it is still possible at the meso-level to identify a link between the difference in ROUGE score and the probability of being chosen with GPT-4.

6 Conclusion

We were able to show the following by using our newly proposed ELO-based scoring algorithm:

In agreement with past results, this scoring method prefers GPT-4 0-shots summaries, while the ROUGE score of these instances is meager. It also likes some other machine translations to the Ground truth, further suggesting the inadequacy of using it as the sole base to evaluate, at least in the CNN-Dailymail Dataset's case.

By looking at correlations with ROUGE scores, we find only limited noisy correlations to the point of having a slightly negative Cohen Kappa.

Further research is necessary to integrate human annotations to compare them with GPT-4 and use the method on other text generation tasks such as Translation.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," **Advances in neural information processing systems**, vol. 30, 2017.

[2] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, **et al.**, "Abstractive text summarization using sequence-to-sequence rnns and beyond," **arXiv preprint arXiv:1602.06023**, 2016.

[3] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in **Text summarization branches out**, pp. 74–81, 2004.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[5] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," **arXiv preprint arXiv:1904.09675**, 2019.

[6] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," **Transactions of the Association for Computational Linguistics**, vol. 9, pp. 391–409, 2021.

[7] T. Goyal, J. J. Li, and G. Durrett, "News summarization and evaluation in the era of gpt-3," **arXiv preprint arXiv:2209.12356**, 2022.

[8] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, **et al.**, "Judging llm-as-a-judge with mt-bench and chatbot arena," **arXiv preprint arXiv:2306.05685**, 2023.

[9] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan, "A general language assistant as a laboratory for alignment," 2021.

[10] M. Boubdir, E. Kim, B. Ermis, S. Hooker, and M. Fadaee, "Elo uncovered: Robustness and best practices in language model evaluation," 2023.

[11] K. Krishna, A. Roy, and M. Iyyer, "Hurdles to progress in long-form question answering," 2021.

[12] D. G. de Pinho Zanco, L. Szczecinski, E. V. Kuhn, and R. Seara, "A comprehensive analysis of the elo rating algorithm: Stochastic model, convergence characteristics, design guidelines, and experimental results," 2022.

[13] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," **arXiv preprint arXiv:1910.13461**, 2019.

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[15] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in **International Conference on Machine Learning**, pp. 11328–11339, PMLR, 2020.

[16] Y. Liu, P. Liu, D. Radev, and G. Neubig, "Brio: Bringing order to abstractive summarization," **arXiv preprint arXiv:2203.16804**, 2022.

[17] OpenAI, "Gpt-4 technical report," 2023.

[18] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," **arXiv preprint arXiv:1808.08745**, 2018.

[19] C. Olsson, S. Bhupatiraju, T. Brown, A. Odena, and I. Goodfellow, "Skill rating for generative models," 2018.

Annex

ELO algorithm details

For this we define $M = (m_0, ..., m_{n-1})$ the list of n models that are used, then $x_i = (x_{i,0}, ..., x_{i,n-1})$ the ith input where

$$x_{i,j} = \begin{cases} 1 & \text{if j is the second member of our pair} \\ -1 & \text{if it is the first} \\ 0 & \text{else} \end{cases}$$

The complete input is then $X = (x_0, ..., x_{N-1})$ where N is the total number of samples, and then $Y = (y_0, ..., y_{N-1})$ where $y_i = \mathbb{1}_{\text{second team won}}$. Finally, we can define our estimator as $\theta = (\theta_0, ..., \theta_{n-1})$ the estimator for Y will become $\tilde{Y} = \sigma(X\theta^T/b)$ (where the division and σ operator are applied to each member one by one). The stochastic Gradient descent iterative algorithm, in this case, will be

$$\theta_k = \theta_{k-1} + \beta[y_{k-1} - \sigma(x_{k-1}\theta_k^T/b)]x_{k-1}$$

Convergence of the Algorithm

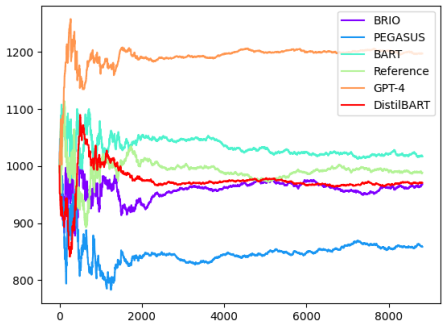

Figure 6 Convergence of the ELO algorithm for GPT-3

Regression process details

The method used to evaluate if this modelization works is the following:

- We convert the distributions into bins, the total number of bins being N
- for each bin, we have the proportion of either model being selected and the total number of data entries in this bin; for our case, we chose 40 bins.

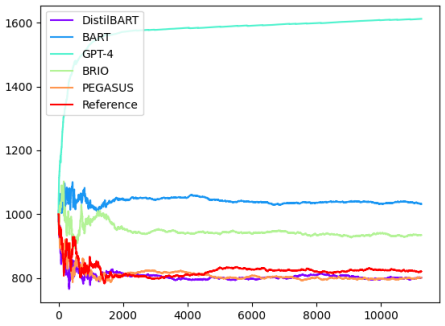

Figure 7 Convergence of the ELO algorithm for GPT-4

- We consider each bin a data point of value x the ROUGE difference corresponding, and value y the proportion of model 1 selected.
- We compute the linear regression of these data points using the number of entries per bin as weights. We get the slope a and the coefficient of determination r^2