# Aggregating similarity metrics for natural language generation

Grégoire Retourné*
ENSAE, France
gregoire.retourne@gmail.com

Hugo Peltier*
ENSAE, France
hugo.peltier@ensae.fr

## Abstract

*This paper conducts a benchmark of most classical natural language generation metrics [11] on a translation task. We evaluated the correlation between values of similarity between a reference translation and a candidate, and a human scoring of this candidate. We then established a ranking of the metrics relatively to this score, which is common to what we could have expected. Finally, we propose a way to aggregate different metrics as a vote of expert through Kemeny [12] consensus, to be able to grasp the best characteristics of each metric, which are to be very good on text-level features (BLEU [21] and ROUGE [17] for instance), and high-level ones (BERTScore [27]). Alas, this ranking is only relevant if the metrics behave differently relatively to another on different tasks, which is not the case here. We made our code available on GitHub at https://github.com/greg2451/aggregating-text-similarity-metrics. It includes a simple way to re-run our experiment on the WMT16 and WMT17 datasets, as well as some code to aggregate metrics with Kemeny [12] consensus.*

## 1. Introduction

The issue of metrics in natural language generation (NLG) is both fundamental and relatively complex [18, 7, 4]. Indeed, unlike other problems, such as classification [22, 9], the lack of target labels makes it difficult to evaluate the performance of an algorithm without human input. Currently, for the evaluation of these specific tasks, the most reliable method of evaluating algorithms is based on human scoring of the generated texts. For a number of practical reasons (slow, costly, not very scalable, etc.), it would be desirable to automate this evaluation phase. The challenge is therefore to create metrics that are as close as possible to the evaluation of a human annotator.

Within NLG tasks, translation evaluation is particularly complex since it requires to keep both high-level information (the meaning of the translated text) and lower-level information (staying as close as possible to the original text). This article therefore aims to explore and benchmark the different automatic metrics currently used for the evaluation of algorithms.

Several approaches are competing in the field of automatic evaluation of NLG algorithms. Some so-called "string-based" approaches, such as ROUGE [17] or BLEU [21] metrics, are rather low-level. Others are high level (BERTScore, Baryscore [8]...), i.e. they capture finer information from the text, such as its meaning.

Finally, this paper aims to experiment with a method of choosing an evaluation metric when the source language is unknown, based on the Kemeny consensus. This method could be used for multilingual translation algorithms.

## 2. Metrics

Text similarity metrics are hard to define by nature. If it is clear that a similarity score should be high on two "similar" texts, and "low" on dissimilar texts, this definition is not quantitative. Overall, there is a common consensus that the GOLD standard is correlation with human judgement [1]. The existing many metrics [11] for measuring text similarity in the literature can be classified in two different categories: discrete (or string-based) and continuous (or embedding-based).

### 2.1. String-based metrics

String-based metrics are based on the string representation of the text, and they are usually used for short texts, such as sentences. Most often they compare the reference text and the candidate relying on surface forms, potentially counting n-grams that are common to both texts, or computing the edit distance between the two texts. Some common string-based metrics are BLEU [21, 24], ROUGE [17], METEOR [14] [15], CHRF [23], TER. These metrics have the principal advantage of being fast and easy to implement, but they are not able to capture the meaning of the text, and they are not able to handle long texts. Since they are based on the string representation of the text, different words with similar meaning will be considered as different, and hence the

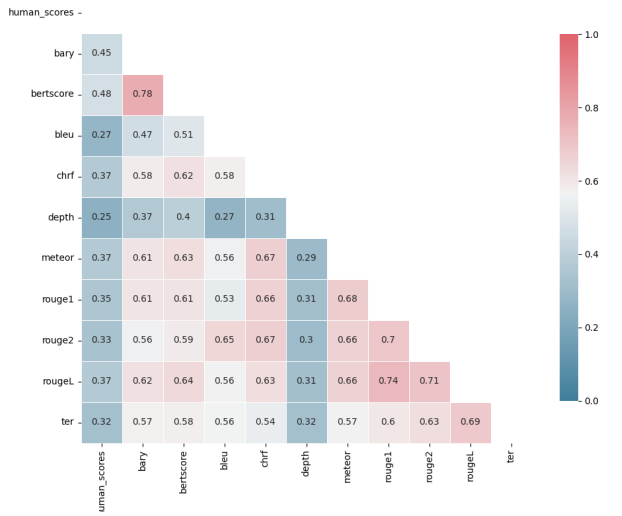


Figure 1. Correlation between human judgments and automatic scores

score will be low. However, these methods are still used in practice in many applications, such as machine translation, and they show good correlation with human ratings [1].

### 2.2. Embedding-based metrics

Embedding-based metrics are based on the representation of the text as a vector in a high-dimensional space. First introduced in the context of machine translation [19], they have been used in many applications, such as text summarization [16], question answering [10], and text classification [10]. The vector representation of the text is obtained by passing the text through a neural network, which is usually pre-trained on a large corpus. Then, having access of many different representations of the text, it is possible to compute the similarity between the reference text and the candidate, by using some similarity measures between vectors. These types of metrics include BERTScore [27], MoverScore [28] BaryScore [8], DepthScore [26], INFOLM [5], and many others. If they have shown much better correlation with human ratings than string-based metrics, it comes with a price: they are much slower and more complex to implement. Moreover, they loose the interpretability of string-based metrics, since the vector representation is not directly interpretable.

### 2.3. Correlation Measures

At this point, we have a column of metrics values, and a column of human scores. Since we want to measure the correlation between the two columns, we need to define a correlation measure. There are three main correlation measures: Pearson's correlation, Spearman's correlation [25], and Kendall's correlation [13]. Since there are no clear consensus on which measure is the best, and since we can have it for free, we will use all of them to measure the correlation between the metrics and the human scores. Intuitively, the ranks of absolute correlation value should not differ too much between the three measures.

## 3. Methodology

Evaluating text similarity measures is a challenging task, since it requires a very particular setup that is not quite common in datasets.

Table 1. Kendall correlation on WMT16

| | Language Pair | | |
|---|---|---|---|
| | **de-en** | **cs-en** | **fi-en** |
| **BARYSCORE** | 48.5 | 49.1 | 43.0 |
| **BERTSCORE** | **52.1** | **52.4** | **46.7** |
| **BLEU** | 27.2 | 28.0 | 21.1 |
| **CHRF** | 40.7 | 37.7 | 31.1 |
| **DEPTHSCORE** | 29.9 | 25.2 | 23.5 |
| **METEOR** | 38.9 | 36.8 | 32.9 |
| **ROUGE1** | 37.8 | 34.8 | 29.4 |
| **ROUGE2** | 36.4 | 33.6 | 28.1 |
| **ROUGEL** | 40.4 | 38.4 | 34.0 |
| **TER** | 36.3 | 32.2 | 26.3 |

### 3.1. Setup

As mentioned before, the goal of text similarity metrics is to best embed the human knowledge in a computation. To evaluate the performance of a metric on that task, we need:

- a set of reference sentences
- a set of sentences to be compared to the reference sentences
- a set of human rating between the reference sentences and the sentences to be compared

If the two first items are quite easy to obtain, the last one is not. Indeed, it requires a lot of time and effort to rate a large number of sentences. Moreover, the very notion of *human rating* is hard to define, since it is not clear what is the best way to rate a sentence, and how to achieve consensus among raters. To overcome this problem, some frameworks have been proposed, such as Pyramids [20], which is a method for evaluating content selection in summarization.

Furthermore, depending on the natural language generation task, we are interested in different types of similarity. For instance, in text summarization, we are interested to retrieve all the information of the original text, while in translation, we are more interested in the overall meaning of the text. We decided to focus on the latter task, since it is more

challenging, and it is also the one that is most relevant to our work. It is by design high-level, and it requires a good understanding of the language, meaning that the similarity metric should be able to capture the meaning of the text, and hence the need for continuous similarity measures.

### 3.2. Datasets

In this section, we present the datasets used in our experiments.

The WMT dataset [3] [2] is a collection of parallel texts in different languages. It is used to evaluate the performance of machine translation systems. We used several pairs of languages, all being translated to English, for a total of 3360 sentences. Each sentence has a reference translation, a translation to be compared to the reference and a human rating between the two.

We present the results of the experiments on the WMT16 dataset in Section 4.

### 3.3. Execution

The experiments were executed on a macOS machine with M1 processor and 16GB of RAM using Python 3.10, with no GPU. The code is available on GitHub at `https://github.com/greg2451/aggregating-text-similarity-metrics`. Follow the instructions in the README to install the dependencies and run the experiments.

Table 2. Pearson correlation on WMT16

| Language Pair | | | |
|---|---|---|---|
| | **de-en** | **cs-en** | **fi-en** |
| **BARYSCORE** | 68.4 | 68.7 | 72.9 |
| **BERTSCORE** | **72.0** | **72.0** | **74.3** |
| **BLEU** | 45.1 | 40.7 | 40.2 |
| **CHRF** | 60.4 | 55.1 | 57.5 |
| **DEPTHSCORE** | 49.7 | 41.9 | 44.1 |
| **METEOR** | 57.2 | 54.3 | 60.6 |
| **ROUGE1** | 56.6 | 51.4 | 56.8 |
| **ROUGE2** | 56.2 | 49.7 | 52.8 |
| **ROUGEL** | 60.6 | 55.6 | 60.9 |
| **TER** | 48.4 | 46.1 | 42.5 |

## 4. Experiments

We present here a part of our results applying different metrics on the WMT16 dataset [3], for three language couples (de-en), (cs-en) and (fi-en). We then computed the correlations with human score (Kendall on Fig.1, Pearson on Fig.2 and Spearman on Fig.3). The rest of the results can be easily obtained by running the released code of the paper.

Overall, it is very satisfying to notice some real correlation between human judgments and the scores. This validates the approach of using automatic scoring in the evaluation of MT systems.

Table 3. Spearman correlation on WMT16

| Language Pair | | | |
|---|---|---|---|
| | **de-en** | **cs-en** | **fi-en** |
| **BARYSCORE** | 65.1 | 67.8 | 71.2 |
| **BERTSCORE** | **70.0** | **71.7** | **72.2** |
| **BLEU** | 35.9 | 37.3 | 34.5 |
| **CHRF** | 56.3 | 53.2 | 51.9 |
| **DEPTHSCORE** | 42.6 | 36.5 | 36.5 |
| **METEOR** | 54.4 | 51.7 | 55.8 |
| **ROUGE1** | 52.7 | 49.1 | 51.8 |
| **ROUGE2** | 51.0 | 47.6 | 45.4 |
| **ROUGEL** | 56.1 | 54.1 | 56.8 |
| **TER** | 50.7 | 46.2 | 43.9 |

As expected, the best correlations are comparably obtained by BERTScore, and BaryScore, since they rely on embeddings, and thus are able to capture the high-level meaning of the sentences, contrary to the other scores that are more sensitive to the surface form of the sentences. The difference is how the different layers are then combined to obtain a single score, which in the case of BaryScore is through Wasserstein Barycenters [8].

Note that we did not expect such results for DepthScore, since it is also an embedding-based score, especially given the high performance announced [26]. Our hypothesis is that using a smaller backbone model required to adapt some hyper-parameters of the metric, which we were not able to do in the limited time we had for this work, and thus that does not invalidate the approach.

Image.1 shows the (kendall) correlation between each score and the human judgments for all the samples in the dataset (5360 pairs of sentences).

It is interesting to note that both baryscore and bertscore have a high correlation between each other: this is expected as they are both based on the same model (in our case, `distilbert-base-uncased`).

Another significant result from this experiment is to show that human judgement appears to be different from any other automatic metrics: every considered metrics have higher correlations to other metrics than to human scores. This indicates that all of them share some characteristics, and are still far away from correctly approximating human judgment. This possibly means that there are still a lot of room to improve in that domain, and that classical modern metrics are all doing similar mistakes in approximating human judgement.

## 5. Kemeny ranking concensus

As we have seen previously, depending on the datasets proposed, some metrics will correspond more or less well

to human judgement. For example, when evaluating the Kendall tau [13] of the CHRF [23] and METEOR [14] [15] metrics, we find that CHRF has a better coefficient than METEOR when translating into Romanian and a worse coefficient when translating into Russian.

This raises the question of how to select a good metric with knowledge of the overall performance of the metrics on a working dataset. Indeed, the challenge of automating the evaluation of NLG algorithms is to be able to apply an automatic method that is as close as possible to human opinion [1]. In our case, we are interested in finding the metric that will best correspond to human opinion in general when evaluating the various tasks proposed (here translation into Czech, German, Finnish or Russian). This case would correspond to the evaluation of a multilingual algorithm for which the input language is not known. It is therefore necessary to establish a ranking of the metrics that is as robust as possible to the different languages used.

We base our choice of metrics on a Kemeny [12] ranking consensus. This ranking method is also used to rank algorithms according to their performance on a number of tasks. Colombo et al. [6] propose a method of ranking algorithms by aggregating their ranks on different tasks. We apply this idea to find a metric that best matches human judgment. We first rank the different metrics for each of these translation tasks from a given language to English (a low ranking corresponds to a high correlation with human opinion). Finally, once these rankings are established (cf Table 4), we can establish an overall ranking of the metrics following the Kemeny ranking consensus. We reverse the classical point of view of using fixed metrics to evaluate algorithms on different tasks and rank the algorithms. We use Kemeny consensus to evaluate the performance of metrics (correlation with human evaluation) on fixed datasets to rank the metrics (cf Table 5).

This classification is based on Kendall's correlation [13] which measures the correlation between two elements of $\mathfrak{S}_n$. This distance $d$ is defined as follows: for $\sigma, \tau \in \mathfrak{S}_n$, $d(\sigma, \tau) = \sum_{1 \leq i,j \leq n} I_{(\sigma_i - \sigma_j)(\tau_i - \tau_j) < 0}$. A Kemeny consensus $\sigma^*$ of $\sigma_1, \sigma_T \in \mathfrak{S}_n$ is a solution of the minimization problem, $\min_{\sigma \in \mathfrak{S}_n} \sum_{1 \leq t \leq T} d(\sigma, \sigma_t)$.

The Kemeny consensus ranking is not very tractable as it is a NP-hard problem.

## 6. Conclusions

We have tested the performance of several automatic scoring methods on the WMT16 and WMT17 datasets, and we have confirmed that they are able to capture a part of the human judgment. As already expected in the literature, the best results are obtained by metrics based on embeddings. However, as noted by Figure.1, all the automatic metrics seem to live in a similar cluster, and human judgment in another one, which indicates that we are still far from being able to fully automate the evaluation of MT systems, and that we are repeating the same errors from one automatic metric to another. Finally, we were able to use Kemeny's consensus ranking to order metrics according to their correlation with human judgment on translations from different languages into English. This process will allow us to choose an optimal evaluation metric when translating from an unknown language. This may be useful for automatic evaluation purposes.

Table 4. Ranking of metrics on different tasks evaluation according to Kendall correlation with human evaluation

| | **cs-en** | **de-en** | **fi-en** | **ro-en** | **ru-en** |
|---|---|---|---|---|---|
| **BARYSCORE** | 2 | 2 | 2 | 2 | 2 |
| **BERTSCORE** | 1 | 1 | 1 | 1 | 1 |
| **BLEU** | 8 | 8 | 8 | 8 | 8 |
| **CHRF** | 4 | 6 | 5 | 3 | 4 |
| **METEOR** | 6 | 4 | 4 | 4 | 3 |
| **ROUGE1** | 5 | 7 | 6 | 6 | 6 |
| **ROUGEL** | 3 | 3 | 3 | 5 | 5 |
| **TER** | 7 | 5 | 7 | 7 | 7 |

Table 5. Ranking of metrics when aggregated by using Kemeny ranking consensus

| | **Bry** | **Brt** | **Ble** | **Chrf** | **Met** | **R1** | **RL** | **Ter** |
|---|---|---|---|---|---|---|---|---|
| **Ranking** | 2 | 1 | 8 | 5 | 4 | 6 | 3 | 7 |

## Future Work

In this study, when working with continuous metrics, we did not change the embedding model. An interesting piece of future work, would be to focus on embedding-based metrics, and try to characterize the effect of the backbone model on the performance of the metric. In particular, we could investigate if the use of a multi-language model, such as mBERT [10], would significantly improve the quality of the metrics, or not. A satisfying result would be to find that multi-language models yield higher quality metrics, and thus would mean that knowing many language gives a better global understanding of high-level semantic meanings.

## References


[1] Abhaya Agarwal and Alon Lavie. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. 06 2008. 1, 2, 4

[2] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, 2016. Association for Computational Linguistics. 3

[3] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, 2015. Association for Computational Linguistics. 3

[4] Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M Suchanek. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. *COLING 2022*, 2022. 1

[5] Pierre Colombo, Chloe Clavel, and Pablo Piantanida. InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation, Mar. 2022. arXiv:2112.01589 [cs]. 2

[6] Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clemencon. What are the best systems? New perspectives on NLP Benchmarking, Oct. 2022. arXiv:2202.03799 [cs]. 4

[7] Pierre Colombo, Maxime Peyrard, Nathan Noiry, Robert West, and Pablo Piantanida. The glass ceiling of automatic evaluation in natural language generation. *arXiv preprint arXiv:2208.14585*, 2022. 1

[8] Pierre Colombo, Guillaume Staerman, Chloe Clavel, and Pablo Piantanida. Automatic Text Evaluation through the Lens of Wasserstein Barycenters, Sept. 2021. arXiv:2108.12463 [cs]. 1, 2, 3

[9] Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. Learning disentangled textual representations via statistical measures of similarity. *ACL 2022*, 2022. 1

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805 [cs]. 2, 4

[11] Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur P. Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. *CoRR*, abs/2102.01672, 2021. 1

[12] John G. Kemeny. Mathematics without Numbers. *Daedalus*, 88(4):577–591, 1959. 1, 4

[13] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 2, 4

[14] Alon Lavie and Abhaya Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics. 1, 4

[15] Alon Lavie and Michael J. Denkowski. The Meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115, Sept. 2009. 1, 4

[16] Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. Scoring Sentence Singletons and Pairs for Abstractive Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy, 2019. Association for Computational Linguistics. 2

[17] 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. 1

[18] C Mellish and R Dale. Evaluation in the context of natural language generation. *Computer Speech & Language*, 12(4):349–373, Oct. 1998. 1

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, Sept. 2013. arXiv:1301.3781 [cs]. 2

[20] Ani Nenkova and Rebecca Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA, May 2004. Association for Computational Linguistics. 2

[21] 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 on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania, 2001. Association for Computational Linguistics. 1

[22] 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. 1

[23] Maja Popović. 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, 2015. Association for Computational Linguistics. 1, 4

[24] Matt Post. A Call for Clarity in Reporting BLEU Scores, Sept. 2018. arXiv:1804.08771 [cs]. 1

[25] C. Spearman. "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904. 2

[26] Guillaume Staerman, Pavlo Mozharovskyi, Pierre Colombo, Stéphan Clémençon, and Florence d'Alché Buc. A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions, Oct. 2022. arXiv:2103.12711 [cs, stat]. 2, 3

[27] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. 1, 2

[28] 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, Sept. 2019. arXiv:1909.02622 [cs]. 2