How Close are Automated Metrics to Human Judgment in Machine Translation?

Mats BERERD*
ENSAE Paris / ESSEC BS
mats.bererd@ensae.fr

Nicolas JULIEN*
ENSAE Paris / ESSEC BS
nicolas.julien@ensae.fr

Abstract

The field of Machine Translation (MT) has experienced rapid progress in recent years, with significant advancements in neural-based models and parallel corpora. However, the challenge of developing relevant automated metrics to evaluate MT systems remains a significant obstacle. Despite the widespread use of automated Natural Language Processing (NLP) metrics for this purpose, there is growing concern that these metrics do not always align with human judgment, leading to potential inaccuracies in evaluation.

To address this issue, our research paper conducted a benchmark evaluation of various automated NLP metrics at the sentence-level, with a focus on two different approaches: candidate-to-reference and candidate-to-original-sentence, also known as the Quality Estimation (QE) task. Through our evaluation, we found that automated metrics perform well in the former aspect, but there is still significant room for improvement in the latter.

Our research highlights the importance of multilingual QE, as it offers a strategic solution to the challenge of collecting labelled data for each language pair. By overcoming this obstacle, multilingual QE can play a crucial role in improving MT models. However, our findings also underscore the need for further research and development in this area, particularly in developing automated metrics that align more closely with human judgment. Ultimately, improving the accuracy and reliability of automated NLP metrics will be essential to advancing the field of MT and realizing the full potential of machine translation technology.

*These authors contributed equally to this work. Code is available on Github.

1 Introduction

Machine Translation has seen considerable progress since the introduction of Transformers (Vaswani et al., 2017), and more generally of pre-trained language models (Devlin et al., 2018; Radford et al., 2018). However, the evaluation of generated sentences is still a problem in its own right. Evaluating the performance of natural language generation systems using human annotation can be a costly and time-consuming process that requires a significant amount of non-reusable labor (Colombo* et al., 2019; Jalalzai* et al., 2020; Colombo et al., 2021a). To mitigate these issues, researchers often rely on automatic metrics as a substitute measure of quality (Colombo, 2021).

The standard evaluation framework for a metric is to compute its correlation with human judgment, on manually annotated datasets. Translation quality can be assessed at different levels of granularity: word, sentence and document level. Although considerable progress has been made in these areas, automatic metrics often show poor correlation with human judgment, at least at the sentence-level (Liu et al., 2016) On the other hand, at the system level, some metrics can show correlations higher than 0.9 (Ma et al., 2019).

The problem lies essentially in the lack of datasets and the inherent bias of human annotation. Moreover, aggregating the different aspects of a translation according to a metric makes it difficult to account for the aspects that the metric manages to capture or not (Guan et al., 2021).

Today, existing metrics can be broken down into three categories:

1. Edit based (Snover et al., 2006): the metric counts the number of operations required to go from the translated sentence to a reference

sentence. The possible operations are insertion, deletion and substitution.

2. N-gram based (Papineni et al., 2002) : the metric is computed from the overlap of n-gram between the reference and the translated sentence.

3. Embedded based (Kusner et al., 2015; Zhang et al., 2019; Zhao et al., 2019; Clark et al., 2019) : sentence are embedded using a model language and a similarity measure is computed.

We can already see here that Edit based N- gram based metrics seem less suitable to score a translation directly from the original sentence.

In this work, we propose a benchmark of different usual automatic metrics on the test set newstest2020 of WMT2020 (Barrault et al., 2020). We use a dataset that has been re-annotated by professional translators, which gives a more reliable notation than the one done by crowd-workers. In addition, to simplify the benchmark, we used the Scalar Quality Metrics (SQM) as a reference, which uses a scale of 0 to 6 to evaluate the quality of a translation, unlike the WMT's 0 to 100 rating (Freitag et al., 2021).

The benchmark is broken down into two parts: one deals with the correlation with human judgment between a proposed translation and a reference translation, the other between the proposed translation and the original (i.e. foreign language) sentence. The second approach is called Quality Estimation, and is a burning topic in the research community, given that it aims to create metrics that do not need a reference translation to work, and therefore being much more available for use in practice.

2 Experiments Protocol

2.1 Dataset

Here we seek to compare different automatic metrics. In machine translation, datasets are usually composed of the triplets sentence to translate, translated sentence, references. In order to evaluate the correlation with human judgment, we add to this triplet the SQM score which gives the translation quality of the translated sentence.

Two modifications are made on the original dataset:

Table 1: Statistics of the dataset

Number of original sentences (OS)	638
Number of translated sentences (TS)	5816
Average number of references (R) per OS	2.98

Table 2: Example of a quadruplets

Original:	J'aime les chiens
Translated:	I'm playing with a dog
References:	[I like dogs, I love dogs]
SQM Score:	3.5

1. The scores of the different translators for the same translation are averaged to obtain a more continuous score between 0 and 6 (which can be resized between 0 and 1).

2. There is no reference translation: to overcome this, as soon as one of the translators has given the maximum score of 6, the proposed translation is considered a reference.

Table 1 shows the statistics of the resulting dataset and Table 2 gives a dummy example.

2.2 Automatic Metrics for Translated to Reference Correlation

2.2.1 N-Gram Based Metric

1. **BLEU** (Papineni et al., 2002)

   BLEU compares the n-gram of the candidate translation with the n-gram of the reference translations to count the number of matches, independently of the positions they occur.It uses a modified n-gram precision score p_n associated with a weight w_n, as well as a Brevity Penalty BP to get :

   $$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \times \log(p_n)\right)$$

2. **ROUGE** (Lin, 2004)

   ROUGE is actually a set of metrics. In addition to precision, ROUGE also takes into account the recall on the n-gram to form a more accurate F_β-score. ROUGE-L measures the longest common subsequences (between the candidate and the references. Since it does not capture synonymous or topic concepts, ROUGE 2.0 (Ganesan, 2018) address this by providing a synonymy dictionary and Part-Of-Speech tagging.

3. **METEOR** (Banerjee and Lavie, 2005; Guo et al., 2018)

   METEOR is also a F_β-score that use Porter stemming and synonymy matching using WordNet (Miller, 1995). To account for the word order in the candidate a penalty function is introduced, which gives :

$$METEOR = (1 - Penalty) \times F_\beta$$

2.2.2 Embedded Based Metric

1. **BERTScore** (Zhang et al., 2019)

   Instead of using n-grams, BERTScore computes a similarity (usually Cosine) for each token in the candidate sentence with each token in the reference sentence. BERTScore is more robust to paraphrase and capture more easily distant dependencies

2. **BLEURT** (Sellam et al., 2020)

   The aim is to combine expressivity and robustness by pre-training a fully learned metric on large amounts of synthetic data, before fine tuning it on human-ratings. Given two sentences x and $\tilde{x}$ we use BERT for sentence pairs, $BERT(x, \tilde{x})$ to extract the [CLS] token embedding $v_{[CLS]}$. To classify, we add a linear layer on top of the [CLS] vector to predict the rating, where W and b are weight matrix and bias vector :

$$\hat{y} = W v_{[CLS]} + b$$

3. **BaryScore** (Colombo et al., 2021b)

   BaryScore combines the layers of BERT to calculate a similarity score. However, instead of using a vector embedding, BaryScore models the layer output as a probability distribution. This allows BaryScore to combine the different outputs using the Wasserstein space topology. Since BaryScore does not handle multiple references, we take the average of the scores obtained.

4. **DepthScore** (Staerman et al., 2021a)

   Similar to BERTScore, DepthScore use a single layer of a pretrained language model to get a discrete probabilities measure of the candidate $\hat{\mu}^C_{\cdot,l}$and the reference $\hat{\mu}^R_{\cdot,l}$ and compute a similarity score using the pseudo metric of (Staerman et al., 2021b). Since DepthScore does not handle multiple references, we take the average of the scores obtained.

2.3 Automatic Metrics for Translated to Orignal Sentence Correlation

The idea here is to use the previous metrics not between a candidate and the reference, but between the candidate and the original sentence. Of course, only embedded-based models seem to be better adapted. Two infrastructures are analysed.

2.3.1 Multilingual BERT structures

As shown in Figure 1, we use a trained multilingual BERT model on several languages to then test the BERTScore, BaryScore, and DepthScore metrics that do not require training. In addition, we also try cross-encoder model from Sentence-Transformers (Reimers and Gurevych, 2019), which works similarly to BERTScore.

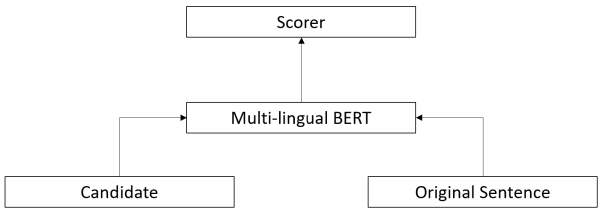

Figure 1: Multilingual BERT structure

2.3.2 Siamese BERT structures

Here we prefer a Siamese structure that encodes the candidate and the original sentence in respective BERTs (see Figure 2) Only the BaryScore and DepthScore metrics seem to be the most suitable for this type of structure.

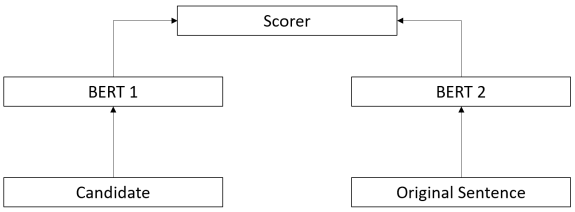

Figure 2: Siamese BERT structure

2.4 Correlation Evaluation

We use the standard Pearson, Spearman and Kendall correlation coefficient on a segment level.

2.4.1 Pearson Correlation Coefficient (PCC)

The PCC, or r is a measure of linear correlation between two variables :

$$r_{X,Y} = \frac{\mathbb{C}ov(X,Y)}{\sqrt{\mathbb{V}ar(X)\mathbb{V}ar(Y))}}$$

2.4.2 Spearman's Rank Correlation Coefficient ρ

It is a measure of rank correlation between two variables. In other words, it shows the relationship between two variable using a monotonic function. Converting X,Y to ranks $rg(X)$, $rg(Y)$:

$$\rho_{X,Y} = r_{rg(X),rg(Y)} = \frac{\mathbb{C}ov(rg(X),rg(Y))}{\sqrt{\mathbb{V}ar(rg(X))\mathbb{V}ar(rg(Y))}}$$

2.4.3 Kendall Rank Correlation Coefficient τ

It is a measure of the ordinal association between two variables. For any pair (x_i, y_i) and (x_j, y_j), $i < j$, the pair are concordant if both $x_i > x_j$ and $y_i > y_j$ holds or both $x_i < x_j$ and $y_i < y_j$ holds; otherwise the pair are discordant. For n observation :

$$\tau_{X,Y} = \frac{\#\text{Concordant pairs} - \#\text{Discordant pairs}}{\#\text{Number of pairs}}$$

3 Results

Due to lack of computational resources, the coefficients below have been computed with a sample of 500 quadruplets of our final dataset. Indeed, embedded based metrics are resource intensive.

3.1 Translated to References

Metric	Pearson	Spearman	Kendall
bleu	0.659636	0.784567	0.611053
meteor	0.644352	0.775174	0.576915
bertscore	0.671346	0.793900	0.608994
bleurt	**0.701576**	0.796326	0.604044
rouge1	0.658301	**0.798812**	**0.624249**
rouge2	0.677770	0.793060	0.616871
rougeL	0.655965	0.794507	0.618991
rougeLsum	0.655965	0.794507	0.618991
baryscore_W	-0.561375	-0.603565	-0.438494
baryscore_SD_10	-0.277762	-0.276258	-0.195089
baryscore_SD_1	-0.349625	-0.366585	-0.261753
baryscore_SD_5	-0.284509	-0.284985	-0.201229
baryscore_SD_0.1	-0.564299	-0.605902	-0.441264
baryscore_SD_0.5	-0.443399	-0.479060	-0.348217
baryscore_SD_0.01	-0.561362	-0.603572	-0.438450
baryscore_SD_0.001	0.098422	0.199817	0.139307
depth_score	-0.500916	-0.606358	-0.436903

Table 3: Evaluation of correlation coefficient (Pearson, Spearman and Kendall) between automated metric and human judgment for Translation-Original similarity evaluation

3.2 Translated to Original

Metric	Pearson	Spearman
baryscore_W	-0.137279	-0.142643
baryscore_SD_10	-0.194361	-0.212121
baryscore_SD_1	-0.193026	-0.211178
baryscore_SD_5	-0.194259	-0.212340
baryscore_SD_0.1	-0.147889	-0.169007
baryscore_SD_0.5	-0.190128	-0.208709
baryscore_SD_0.01	-0.137188	-0.142703
baryscore_SD_0.001	0.047921	-0.020487
depth_score	-0.037193	0.003231
siam_baryscore_W	-0.050786	-0.063746
siam_baryscore_SD_10	-0.043611	-0.061870
siam_baryscore_SD_1	-0.043827	-0.062036
siam_baryscore_SD_5	-0.043634	-0.061885
siam_baryscore_SD_0.1	-0.046168	-0.063744
siam_baryscore_SD_0.5	-0.044071	-0.062197
siam_baryscore_SD_0.01	-0.055554	-0.068986
siam_depth_score	0.098776	0.105129
Sentence Transformer	**0.233240**	**0.231915**
bertscore_multi	0.152020	0.144655

Metric	Kendall
baryscore_W	-0.100542
baryscore_SD_10	-0.152267
baryscore_SD_1	-0.151465
baryscore_SD_5	-0.152642
baryscore_SD_0.1	-0.119284
baryscore_SD_0.5	-0.149606
baryscore_SD_0.01	-0.100525
baryscore_SD_0.001	-0.014129
depth_score	0.001253
siam_baryscore_W	-0.044443
siam_baryscore_SD_10	-0.041373
siam_baryscore_SD_1	-0.041714
siam_baryscore_SD_5	-0.041476
siam_baryscore_SD_0.1	-0.043607
siam_baryscore_SD_0.5	-0.042004
siam_baryscore_SD_0.01	-0.047223
siam_depth_score	0.073179
Sentence Transformer	**0.163062**
bertscore_multi	0.101557

Table 4: Evaluation of correlation coefficient (Pearson, Spearman and Kendall) between automated metric and human judgment for Translation-Original similarity evaluation

4 Conclusion

In the first part of our study, the sentence-level evaluation of various automated metrics using Spearman's and Kendall's correlations showed that ROUGE performed the best, followed by BLEURT for Spearman's correlation. However, BaryScore and DepthScore, despite obtaining respectable scores, failed to capture human judgment effectively.

It is important to note that the dataset used in this study initially had no references, and the references were taken from the proposed translations. Additionally, the texts were taken from the WMT, which is the dataset on which BLEURT was fine-tuned.

In the second part of the study, the results were unexpected, as none of the metrics achieved high correlations. Instead, the model based on Sentence Transformers, which was originally trained for Information Retrieval, achieved the best performance.

For future research, it would be interesting to evaluate the same metrics at the word-level and system-level for our dataset and other task (Chhun et al., 2022). This would provide a more comprehensive evaluation of automated metrics and their ability to capture human judgment accurately.

In order to advance research in machine translation, it is crucial to identify the factors that prevent automatic metrics from accurately approximating human judgments. It is important to understand and characterize the unobserved variance between the two evaluations. For instance, it is necessary to determine whether the discrepancies arise solely from the biases of human evaluators. Some studies suggest that automatic metrics are subject to a glass ceiling that limits their ability to approximate human judgment (Colombo et al., 2022b,a). It remains uncertain whether this ceiling can be surpassed in the future. Therefore, further research is required to investigate the limitations of automatic metrics and identify potential ways to address them.

.

References

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

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 Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

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.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Michael D. Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *CoRR*, abs/1603.08023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Yinuo Guo, Chong Ruan, and Junfeng Hu. 2018. Meteor++: Incorporating copy knowledge into machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 740–745, Belgium, Brussels. Association for Computational Linguistics.

Kavita Ganesan. 2018. ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks. *CoRR*, abs/1803.01937.

Pierre Colombo*, Wojciech Witon*, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *NAACL 2019*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

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. *CoRR*, abs/1909.02622.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. *CoRR*, abs/2004.04696.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Hamid Jalalzai*, Pierre Colombo*, Chloé Clavel, Éric Gaussier, Giovanna Varni, Emmanuel Vignon, and

Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. *NeurIPS 2020*.

Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021b. Automatic text evaluation through the lens of wasserstein barycenters. *CoRR*, abs/2108.12463.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation.

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

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

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics. *CoRR*, abs/2105.08920.

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

Guillaume Staerman, Pavlo Mozharovskyi, Pierre Colombo, Stéphan Clémençon, and Florence d'Alché Buc. 2021b. A pseudo-metric between probability distributions based on depth-trimmed regions.

Pierre Colombo, Maxime Peyrard, Nathan Noiry, Robert West, and Pablo Piantanida. 2022b. The glass ceiling of automatic evaluation in natural language generation.

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

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