Evaluation of Untrained Metrics through Correlation with Human Judgment: the Case of Translation

Alexandrine Lanson*
ENSAE, France
alexandrine.lanson@ensae.fr

Alex Tordjman*
INSEE, France
alex.tordjman@ensae.fr

Abstract

Motivated by the recent development of Natural Laguage Generation (NLG), we give an overview of several untrained metrics used to evaluate NLG algorithms' performances, for a translation task. We use a dataset (WMT16, de-en) composed of pairs of sentences, each of which being labelled with a human score reflecting how similar both sentences are according to human judgment. We compute the correlation between each metric's score and the human reference score, as well as correlations among metrics. Our results show that embedding-based metrics are more correlated with human judgment than string-based metrics; the highest correlation coefficients being obtained for BERTScore. Among metrics, embedding-based metrics are the most correlated with each other.

1 Introduction

The fast development of Natural Language Processing (NLP) algorithms comes with a key component of any machine learning model: performance evaluation. As text's human annotation is expensive and time-consuming (Sai et al., 2020), researchers rely on automatic metrics as a proxy of quality. In Natural Language Generation (NLG) (Colombo* et al., 2019; Jalalzai* et al., 2020; Colombo et al., 2021b), building metrics remains a challenge as it requires evaluating text similarity between the output given by text generating systems and one or several gold-standard reference texts. Each text generation task (translation, story generation, data2text generation...) requires its own performance criteria. For example, key criteria for translation may be fidelity or fluency (King et al., 1999; White et al., 1994).

In order to better assess the progress of new methods, we go through several metrics and compare their performance by evaluating their correlation with human judgment, as the latter is considered to be one of the most important performance criteria (Chatzikoumi, 2019; Specia et al., 2010; Koehn, 2009; Lavie and Agarwal, 2007).

In the following, we consider only untrained metrics (using just the model's output and calculating the metric from a pre-defined algorithm) applied to translation tasks (Doddington, 2002; Popović, 2015). The metrics we focus on are listed in Table 1, classified into two categories: string-based metrics evaluate the textual representation of the inputs, while the chosen embedding-based metrics rely on contextualized embeddings that are continuous representations (Devlin et al., 2018; Wang et al., 2020).

String-based	BLEU (Papineni et al., 2002)
	ROUGE (Lin, 2004)
	chrF (Popović, 2015)
Embedding-based	BERTScore (Zhang et al., 2020)
	MoverScore (Zhao et al., 2019)
	BaryScore (Colombo et al., 2021c)

Table 1: Classification of metrics considered in our study.

String-based metrics rely on string representation and are therefore known to fail to capture nuances when reference and candidate carry the same meaning but with different forms, for example with synonyms and paraphrases (Reiter and Belz, 2009). Embedding-based metrics overcome this limit by measuring semantic similarity rather than lexical overlap. While the latter achieves stronger correlation with human judgment, the former are still the most widely used in Machine Translation (MT) papers (Marie et al., 2021), some reasons being the explainability of the scores (string-based metrics are mostly highly transparent while embedding-based metrics rely on black-box language models such as BERT) or

the computational inefficiency to run expensive new metrics at large scale (Leiter et al., 2022).

Hence, choosing a NLG metric is far from being obvious, even for a given task such as translation; reviews of existing evaluation metrics for NLG tasks have multiplied lately (Marie et al., 2021; Chatzikoumi, 2019; Sai et al., 2020; Reiter and Belz, 2009; Chhun et al., 2022; Staerman et al., 2021; Colombo, 2021). We propose a simple comparison of metrics, in the context of translation, in order to understand what characterizes and differentiates them.

We provide a theoretical framework as well as an implementation of all considered methods than can be found in a github repository: `https://github.com/AlexLsn/Text_similarity_metrics`.

2 Problem Framing

Given a dataset

$$D = \{x_i,\ y_i,\ h(x_i, y_i)\}_{i=1}^{N}$$

where x_i is the i-th reference text, y_i is the i-th associated candidate text; N is the number of texts in the dataset. $h(x_i, y_i) \in \mathcal{R}^+$ is the score associated by a human annotator to the candidate text y_i when comparing it with the reference text. We will consider several evaluation metrics m (see Table 1) and compare them to h.

More precisely, we aim at comparing scores given by metrics prediction and text level human judgment (Chatzikoumi, 2019). Text-level correlation refers to the evaluation of the ability of a metric to measure the semantic equivalence between a candidate and a reference sentence (Colombo et al., 2021c). Using above-introduced notations, with C a correlation coefficient, it writes:

$$C_{text} = \frac{1}{N}\sum_{i=1}^{N} C((m(y_i, x_i),\ h(y_i, x_i)).$$

As mentioned in introduction (1), we use this correlation as an evaluation measure of the translation task we consider.

3 Experiments and Protocol

Choice of dataset We conduct our experiments on the WMT16 dataset (Bojar et al., 2016). We focus on the de-en pair, that is translation from German to English. This dataset is composed of 500 pairs of sentences and the associated human similarity score (a real number between -1.9 and 1.3).

Text level correlation To measure text-level correlation introduced in Section 2, we consider three common coefficients: Pearson's coefficient (Pearson, 1895; Leusch et al., 2003) measures linear correlation, while Spearman's (Fieller et al., 1957; Melamed et al., 2003) and Kendall's (Kendall, 1938) coefficients compare the ranks of data.

Significance testing To check how trust-worthy the differences between our metric scores are, we follow the recommandations of Marie et al. (2021) and perform statistical significance testing using Williams test (Steiger, 1980) as considered observations are correlated (Graham and Baldwin, 2014; Graham, 2015; Graham et al., 2015).

Choice of model Some metrics (BERTScore, MoverScore, BaryScore) performances are dependent on the choice of the model; thus, we work with one single model in the whole study to be able to compare scores. Due to computational constraints, we choose to use DistilBERT (Sanh et al., 2019). This model reduces the size of a BERT model (Devlin et al., 2018) by 40%, while retaining 97% of its language understanding capabilities and being 60% faster.

4 Results

4.1 Correlation to human judgment

Absolute correlations between metric prediction and text level human judgement on the de-en pair of WMT16 are given in Table 2. Pearson's r, Spearman's ρ and Kendall's τ coefficients are reported. They are all significant at the 1% level. We use ROUGE1 for computing ROUGE scores.

	r	ρ	τ
BLEU	0.54	0.48	0.34
ROUGE	0.57	0.53	0.38
chrF	0.60	0.56	0.41
BERTScore	**0.72**	**0.70**	**0.62**
MoverScore	0.70	0.68	0.50
BaryScore	0.69	0.65	0.48

Table 2: Absolute correlations (Pearson - r, Spearman -ρ, and Kendall - τ) between metric prediction and text level human judgment on the de-en pair of WMT16.

As reported in Table 2, the absolute correlation coefficient with human judgment is higher for embeddings-based metrics than for edit-based

metrics. These results are not surprising: for example, metrics relying on contextualized embeddings, as opposed to string-based metrics, are able to handle synonyms, and these should not be penalized by humans for a task such as translation. BERTScore gives the highest values, with BaryScore and MoverScore being close to the former.

Figures 1 and 2 show the distribution of BLEU scores -respectively BERTScores- compared to human scores. One can notice that at low values of BLEU score, there is high a level of uncorrelation with human judgment, which can take whatever values in the admitted interval.

BERTScore and human judgment display a stronger correlation, with a linear relationship pattern arising. Interestingly enough, lower values of human judgment are associated with a higher variance in BERTScore while the distribution is more concentrated for higher values. Plots for other metrics can be found in Appendix 6.1.

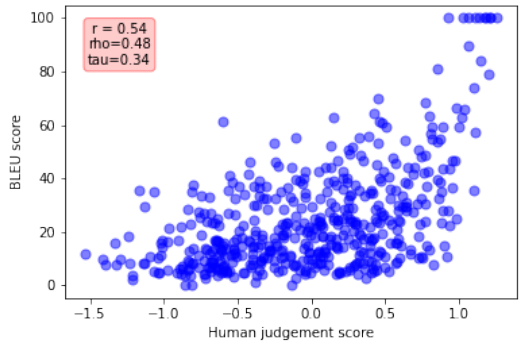

Figure 1: BLEU score vs human judgement score.

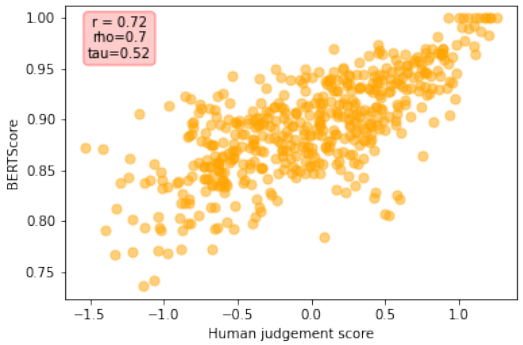

Figure 2: BERT score vs human judgement score.

We test the significance of the increase in correlation between each pair of considered metrics for the Pearson correlation coefficient, using William's test as our observations are correlated (Graham and Baldwin, 2014). Results can be found in Figure 3. One can notice the less-significant results are concerning string-based metrics on one side (BLEU-ROUGE, chrF-ROUGE) and embedding-based on the other side (MoverScore-BERTScore, MoverScore-BaryScore), showing once again that this categorization is relevant.

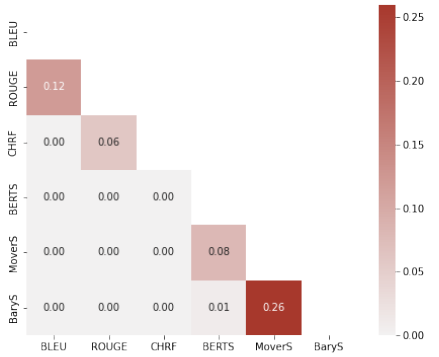

Figure 3: Significance testing on de-en for WMT16. In the matrix is reported the p-value of the Williams Test.

Despite the fact that embedding-based metrics give significant better results than string-based metrics, almost 99% of the research papers in machine translation rely on BLEU to evaluate translation quality, while more than 100 other metrics have been proposed since 2010 (Marie et al., 2021; Novikova et al., 2017). An understanding would come from the fact that BLEU is explainable, quick, inexpensive, language independent, and that due to its use the comparison to previous research is easier (Leiter et al., 2022). Still, knowing that such a metric cannot measure semantic similarity, this calls into question the credibility of the NGL evaluation field.

4.2 Correlation across metrics

Figure 4 displays the intercorrelation across metrics based on Pearson's r (results for Spearman's ρ and Kendall's τ can be found in Appendix 6.2). For all three measures, the highest levels of correlation are observed for metrics based on BERT (BertScore, MoverScore and BaryScore) (Devlin et al., 2018), which is consistent with similar findings in the literature (Colombo et al., 2021c). This high level of correlation is due to the common pre-trained pattern of BERTScore, BaryScore and MoverScore metrics, that have been trained on

general text generation tasks, without specific fit to translation (Zhang et al., 2020; Zhao et al., 2019; Colombo et al., 2021c; Devlin et al., 2018). Our result is consistent with Gupta et al. (2019), who are able to group metrics by correlation pattern. The authors also investigate the possibility of allowing for more complex relationship (non-linear) between correlated metrics, but their results are similar to the Pearson's measure.

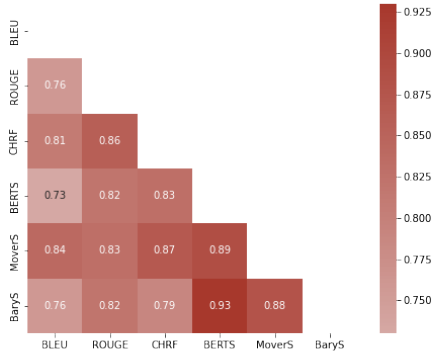

Figure 4: Absolute intercorrelation among metrics according to Pearson's r.

5 Conclusion

Our results show clear better performance of embedding-based metrics for measuring text similarity on the WMT16 de-en dataset. We provide ways of visualizing correlations between metrics and human judgment that support these findings. However, the latter must be contextualized: a recent study by Colombo et al. (2022) reveals that automatic metrics, old (string-based) and new (embedding-based), show stronger similarity between each other than with humans. The authors thus encourage further research to focus on minimizing similarity with existing metrics rather than maximizing correlation with human judgment, in order to capture new aspects and ensure that the field is progressing. In Belouadi and Eger (2022), the authors advocate for the use of fully untrained metrics so as to be able to apply metrics to cases were supervision is infeasible (rare languages translation).

With more time, and to support our results, we would perform evaluation on other datasets. It would also be interesting to include more string-based metrics such as METEOR (Banerjee and Lavie, 2005) or embedding-based metrics such as InfoML (Colombo et al., 2021a). Moreover, a more comprehensive comparison of our results with the literature would be a definite avenue for improvement. Finally, comparing untrained metrics on other tasks such as data2textgeneration (Perez-Beltrachini et al., 2016; Castro Ferreira et al., 2020; Gardent et al., 2017) or summary generation (Nenkova and Passonneau, 2004; Bhandari et al., 2020; Nallapati et al., 2016; Colombo et al., 2022) would also provide a better understanding of metrics' specificities.

6 Appendix

6.1 Correlation plots: metrics vs human judgement

We report on Figures 5, 6, 7, 8 the correlation between metrics and human score as well as each correlation coefficient on WMT16 for the de-en pair. Results for BLEU and BERT can be found in the main content.

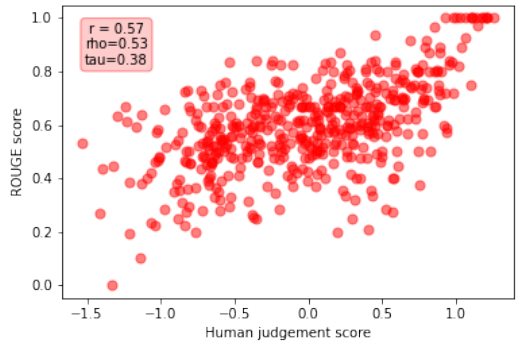

Figure 5: ROUGE score vs human judgment score. The values of the correlation coefficients r, ρ and τ are given.

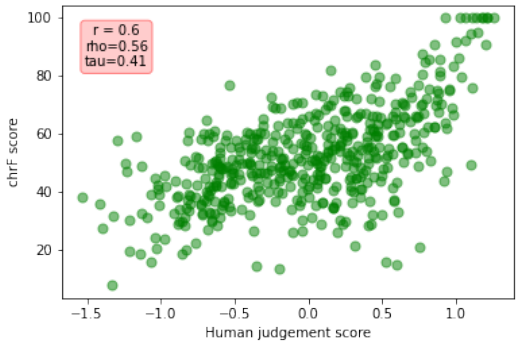

Figure 6: chrF score vs human judgment score. The values of the correlation coefficients r, ρ and τ are given.

6.2 Correlation among metrics

Correlation between metrics' prediction WMT16 (de-en) can be found in Figure 9 for Spearman's ρ and Figure 10 for Kendall's τ. Results for Pearson's coefficients can be found in the main content.

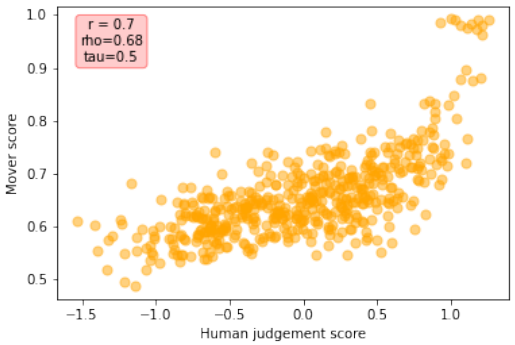

Figure 7: MoverScore vs human judgment score. The values of the correlation coefficients r, ρ and τ are given.

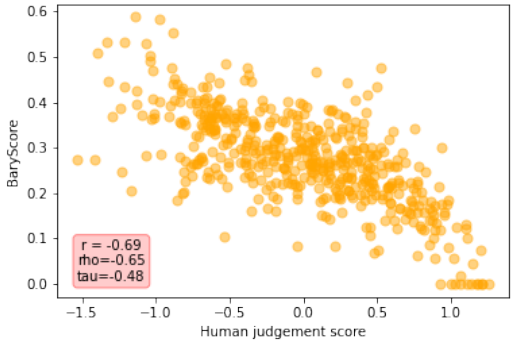

Figure 8: BaryScore vs human judgment score. The values of the correlation coefficients r, ρ and τ are given.

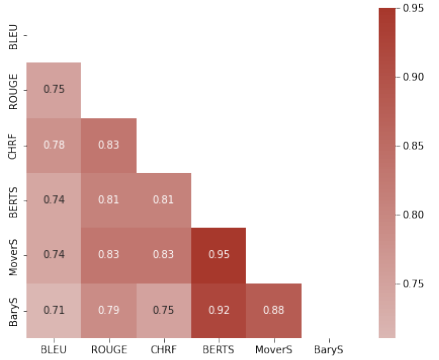

Figure 9: Absolute intercorrelation among metrics according to Spearman's ρ.

.

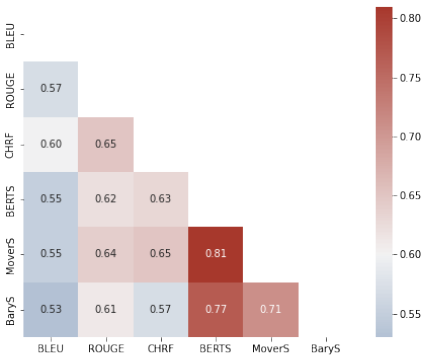

Figure 10: Absolute intercorrelation among metrics according to Kendall's τ.

References

Karl Pearson. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

E. C. Fieller, H. O. Hartley, and E. S. Pearson. 1957. TESTS FOR RANK CORRELATION COEFFICIENTS. I. *Biometrika*, 44(3-4):470–481.

James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87:245–251.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

Margaret King, Eduard Hovy, Benjamin K. Tsou, John White, and Yusoff Zaharin. 1999. MT evaluation. In *Proceedings of Machine Translation Summit VII*, pages 197–207, Singapore, Singapore.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. pages 138–145.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 61–63.

Ani Nenkova and Rebecca Passonneau. 2004. 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. 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.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. pages 228–231.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24:39–50.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813, Beijing, China. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Laura Perez-Beltrachini, Rania Sayed, and Claire Gardent. 2016. Building RDF content for data-to-text generation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1493–1502, Osaka, Japan. The COLING 2016 Organizing Committee.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

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.

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

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

Soumyajit Gupta, Mucahid Kutlu, Vivek Khetan, and Matthew Lease. 2019. *Correlation, Prediction and Ranking of Evaluation Metrics in Information Retrieval*, pages 636–651.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Eirini Chatzikoumi. 2019. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26:1–25.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuxuan Wang, Yutai Hou, Wanxiang Che, and Ting Liu. 2020. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*, 11:1611–1630.

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*.

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

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020. A survey of evaluation metrics used for nlg systems.

Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021c. Automatic text evaluation through the lens of Wasserstein barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10466.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. 2021a. Infolm: A new metric to evaluate summarization & data2text generation. *arXiv preprint arXiv:2112.01589*.

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. 2021b. A novel estimator of mutual information for learning to disentangle textual representations. *ACL 2021*.

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

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

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

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. Towards explainable evaluation metrics for natural language generation.

Jonas Belouadi and Steffen Eger. 2022. Uscore: An effective approach to fully unsupervised evaluation metrics for machine translation.