Text Similarity as An Evaluation Measure of Machine Translation

GitHubCode

Overleaf

Arthur THOMAS
3A ENSAE
arthur.thomas.eleve@ensae.fr

Doulo SOW
3A ENSAE
doulo.sow@ensae.fr

Abstract

This paper is aimed at benchmarking the correlation of some existing metrics with human scores, in the context of Machine Translation related tasks.
Do to so, we focus on untrained measures, in particular BLEU (string-based), InfoLM, TER, DepthScore, METEOR, and BaryScore (embedding-based, text as probability distributions) introduced in [Colombo et al., 2021b]. The data used for this task comes from the Tenth Workshop on Machine Translation (WMT15) and includes pairwise generated sentences and reference sentences associated with human annotation.
To assess the relevance of each aforementioned evaluation metrics, we rely on it closeness to the human annotation in terms of correlation: Spearman, Pearson and Kendall.
The main conclusion to be drawn from our numerical results is that Baryscore is the most correlated with human annotation.

1 Introduction

Natural Language Processing (NLP) computer programs are getting more and more used in real life. From translating text from one language to another, through responding to spoken commands to summarizing large volumes of text rapidly, their application occur in several fields. Although its importance and usefulness, the quality of natural language generation systems must be evaluated using objective and precise criteria in order to have confidence in their results.
This paper is aimed at focusing on the Machine Translation task in order to benchmark some existing criteria to evaluate the accuracy or similitude of a natural language generation system to a human annotation or gold-reference. Indeed, due to both high annotation costs and time, researchers tend to rely on automatic evaluation to compare the outputs of such systems. Regarding how the used metric is obtained, two categories are distinguished: pre-trained metrics and untrained metrics. This work focuses on the later which is split into three subgroups [Colombo et al., 2021b]: (1) string matching; (2) edit based and (3) embedding based metrics. (1) string matching; (2) edit based and (3) embedding based metrics. For each of them, we propose at least one type of metric that can be used to assess the evaluation criteria. In particular, we describe three particular exemples of each type of metrics : BLEU (for string-based), BaryScore (for embedding based metrics), TER (edit based) and infoLM (for a mix) metrics. The remainder of the paper is structured as follows. Section 2 presents the general and theoretical framework of the Machine Translation task and provides the main findings of the related literature. Section 3 focuses on the chosen metrics and introduces core theoretical considerations and the main steps for the computation tasks. Section 3 presents and discusses the numerical results of such metrics applied on data.

Contributions : (1) we introduce brief summaries and explanations of the existing metrics in order that the intuition of this metrics could be understood by non-experts ; (2) we benchmark the correlation of existing metrics with human scores for the data2text generation task.

2 Problem Framing and State of the Art

Although we decide to focus on Machine Translation[Colombo et al., 2021b], benchamarking existing metrics could also be done in several tasks, e.g., Data2Text generation([Colombo et al., 2021a];) or story generation as in [Chhun et al., 2022]. Regardless, the task one is interested in, the general framework is set as following:

2.1 Problem Framing

Given a dataset $\mathcal{D} = \{r_i, \{c_i^s, h(r_i, c_i^s)\}_{s=1}^{S}\}_{i=1}^{N}$, where:

- $r_i = (r_1, \ldots, r_M)$ is the i-th reference text composed of M tokens (e.g., words or subwords);
- $c_i^s = (c_1, \ldots, c_L)$ is the i-th candidate text composed of L tokens generated by the s-th NLG (Natural Language Generation) system;
- $h(r_i, c_i^s) \in \mathbb{R}^+$ is the score associated by a human annotation to the candidate text compared to the reference one,

the main goal is then to define an evaluator metric such that $f(r_i, c_i^s) \in \mathbb{R}^+$ which ideally should be correlated to the human annotation $h(r_i, c_i^s)$. This correlation is measured by correlation coefficients such as the Pearson, Spearman or Kendhal coefficients.

2.2 Data

The data we use in this paper are part of the resources provided for the Tenth Workshop on Machine Translation (WMT15) and include pairwise generated English sentences and reference sentences associated with human annotation. TrueSkill [Sakaguchi et al., 2014], originally developed by Microsoft Research for the Xbox Live gaming community, was adapted to Machine Translation Evaluation task and used as the human evaluation ranking for all translation shared tasks in this workshop.

2.3 State of the Art

As precised above, this paper focuses in the particular task of Machine Translation. In the literature, several papers, departing from the framework described above with specific applications such as Data2Text or story generation in addition to Machine Translation, are published in order to shed light on how different metrics perform compared to human annotation task which is indeed time consuming. As Machine Translation task, for instance, [Colombo et al., 2021b] and [Leusch et al., 2003] use respectively BaryScore metric and a string-to-string distance measure to evaluate how well a Machine Translator perform. [Perez-Beltrachini et al., 2016], and again [Colombo et al., 2021b] focus on Data2Text task to evaluate, regarding correlation metric, to evaluate metrics compared to human annotation. From the story generator point of view, [Xu et al., 2018] propose skeleton-based model to promote the coherence of generated stories. In the same sense, [Chhun et al., 2022] introduce a set of 6 orthogonal and comprehensive human criteria, carefully motivated by the social sciences literature, which could serve as a guideline for evaluating metrics.

3 Presentation of the metrics

3.1 BLEU Score

We will introduce first a string-based metric : BLEU[Papineni et al., 2002]. This metric is language-independent, and can be then used to evaluate a lot of models.

BLEU is computed as follows : $BLEU_N = Bp * exp(\sum_{n=1}^{N} w_n * log(p_n))$

where Bp is the Brevity Penalty, p_n the n-gram clipped precision, w_n the weight associated to the n-gram clipped precision and N the bound of the set of integers m over which we want the geometric mean of the m-gram clipped precisions to be calculated, a n-gram being a set of 'n' consecutive words, taken in order.

The clipped precision is a way to solve the problem of repetitions in the predicted sentence : for example, "The the the the", prediction of "The apple is red", has a precision 1-gram of 1 if the precision is equal to the number of correct predicted words (1-gram) over the number of total predicted words (as "The" is in the list of 1-gram of the sentence of the reference text). Here, the number of correct predicted n-grams is the number of n-grams from the predicted sentence that matches with all of the reference text ; however, the count for each correct n-gram is limited to the maximum number of times that a n-gram occurs in the reference text.

Nevertheless, if "the" is the prediction of "The apple is red", the clipped precision will be 1. To have a significant metric, we need a penalty for too short predictions. The Brevity penalty is a mean to solve this issue. Brevity penalty either is equal to 1 if the number of words in the predicted sentence is strictly superior to the number of words in the sentence of the reference text, or to $exp(\frac{1 - nb_p}{1 - nb_r})$, if the number of words in the predicted sentence (nb_p) is inferior or equal to the number of words in the sentence of the reference text (nb_r).

However, this metric has some drawbacks : The main issue is that this score doesn't take into account synonyms. What is more, the importance of words is ignored (incorrect words like "to" have the same weight as words that contributes significantly to the meaning), and finally it does not take into account the order of words (mainly for $BLEU_1$).

3.2 TER

We will introduce second an edit based metric : TER[Snover et al., 2006]. The TER score measures the minimum number of edits required to modify a hypothesis so that it exactly matches one of the references. This is then normalized by the average length of the references. The TER score is specifically focused on determining the minimum number of edits needed to modify the hypothesis to match with the closest reference.

$$TER = \frac{number\ of\ edits}{average\ length\ of\ the\ setence}$$

The TER score takes into account 4 types of edits that can be made to a hypothesis in order to match a reference. These edits are : the insertion, the deletion, the substitution of individual words and the shifting of word sequences within the hypothesis. It is important to note that all edits, including shifts of any length and distance count as the same value, so that they are considered equally in calculating the TER score. Furthermore, punctuation tokens are treated as normal words, and any mistakes in capitalization are also counted as edits.

3.3 Baryscore

The BaryScore introduced in [Colombo et al., 2021b] is based on deep contextualized embeddings (e.g., BERT, Roberta, ELMo) and mostly relies on Wasserstein barycentric distributions[1] of contextual encoder layers for x_i and y_i^s. The metric can be summarized in two steps: (1)-compute the Wasserstein barycentric distributions [Agueh and Carlier, 2011] of contextual encoder layers for x_i and y_i^s; (2)-Evaluate these barycentric distributions using the Wasserstein distance $\mathcal{W}$. In brief, the algorithm (the notation is adapted to our framework described above) proposed by [Colombo et al., 2021b] to compute the BaryScore is as following:

INPUT: r_i, c_i^s, $(\phi_k, \ldots, \phi_K)$ pre-trained layers from BERT or ELMo:
For k from 1 to K:

1. compute layers embeddings: $\phi_k(r_i)$ and $\phi_k(c_i^s)$;
2. compute $\hat{\mu}_{r_i,k} = \sum_{m=1}^{M} \alpha_m \delta_{\phi_k(r_m)}$ and $\hat{\mu}_{c_i^s,k} = \sum_{l=1}^{L} \beta_l \delta_{\phi_k(c_l)}$, where: α_m and β_l are repectively the inverse document frequencies of each token r_m and c_l and $\delta_{(\cdot)}$ the Dirac's measure.
3. Compute the two Wasserstein barycenters: $\hat{\mu}_{r_i} = \arg\min_{\hat{u}} \sum_{k=1}^{K} \mathcal{W}(\hat{\mu}_{r_i,k}, \hat{u})$ and $\hat{\mu}_{c_i^s} = \arg\min_{\hat{u}} \sum_{k=1}^{K} \mathcal{W}(\hat{\mu}_{c_i,k}, \hat{u})$

OUTPUT: Compute the Wasserstein distance of these two barycenters which correponds to the Baryscore: $\mathcal{W}(\hat{\mu}_{r_i}, \hat{\mu}_{c_i^s})$

3.4 InfoLM

For this metric, we rely on the work of COLOMBO [Colombo et al., 2021a]. For tha calculation of the infoLM, we need two different elements.

First this metric uses a pretrained language model (PLM). The PLM processes a given input sentence x, with a masked position i ($[x]^i$), and generates a discrete probability distribution ($p_\Omega(.|[x]^i)$) across the vocabulary (Ω).

Additionally, this metric uses an information measure ($J : [0;1]^{=|\Omega|} \times [0;1]^{=|\Omega|}$) determines the similarity between the accumulated distributions. The information measure can be a divergence measure (α divergence,γ divergence, AB Divergence) or a distance measure (L1 distance, L2 distance, L∞ distance, Fisher-Rao distance)

The calculation of InfoLM involves three critical steps: first, compute the individual distributions for both the candidate C and reference R. Second, aggregate these individual distributions using a weighted sum : $p_\Omega(.|x) = \sum_{i=1}^{N} \gamma_i \times p_\Omega(.|[x]^i)$ where γ_i is the importance of the i-th token in the sentence. Finally, the similarity between the aggregated distributions is determined using J.

[1] See [Colombo et al., 2021b], for details.

4 Numerical results

We will compute the correlation between human annotations and automatic metrics. We put us in the framework of Colombo et al.[Colombo et al., 2021a]. We take human annotation into account because it is agreed that measuring how well an automatic metric correlates with human judgment is crucial. There is an ongoing discussion about which type of correlation (such as Pearson, Spearman, or Kendall) is the most useful for evaluating automated metrics. Two main strategies are often used for evaluation: (1) text-level correlation and (2) system-level correlation. In this case, we will use the text-level correlation approach as we evaluate the WMT15 model [Bojar et al., 2015]

Formally, the text level correlation $C_{t,f}$ is computed as follows:

$$C_{t,f} = \frac{1}{N} \sum_{i=1}^{N} K(F_i, H_i) \quad (1)$$

where $F_i = [f(c_i, r_i^1), ..., f(c_i, r_i^S)]$ and $H_i = [h(c_i, r_i^1), ..., h(c_i, r_i^S)]$ are the vectors composed of scores provided respectively by the automatic metric f and the human annotation h. $K : R^N \times R^N \rightarrow [-1, 1]$ is the chosen correlation measure.

The different correlation measure are : Pearson [Leusch et al., 2003], Spearman [Melamed et al., 2003] and Kendall [KENDALL, 1938]. We compute the correlations between human annotation and automatic metrics. Further more, we compute the correlation between several automatic metrics. The results are in Figure 1. As we are mainly interested in the Spearman correlations between human annotation and automatic metrics, the values can be found in the first columns. We observe that the best correlation achieved is 0.76 by Baryscore. METEOR has a correlation superior to 0.5, but all the other metrics are below this threshold. All the p-values associated to these correlations are below the usual significance thresholds (cf. Table 1).

We remark that the correlations between human annotation and automatic metrics are similar to the correlations of automatic metrics between themselves.

5 Discussion/Conclusion

We computed the correlations between human annotation and automatic metrics. We observe that the best correlation achieved is 0.76 by Baryscore. METEOR has a correlation superior to 0.5, but all the other metrics are below this threshold. All the p-values associated to these correlations are below the usual significance thresholds.

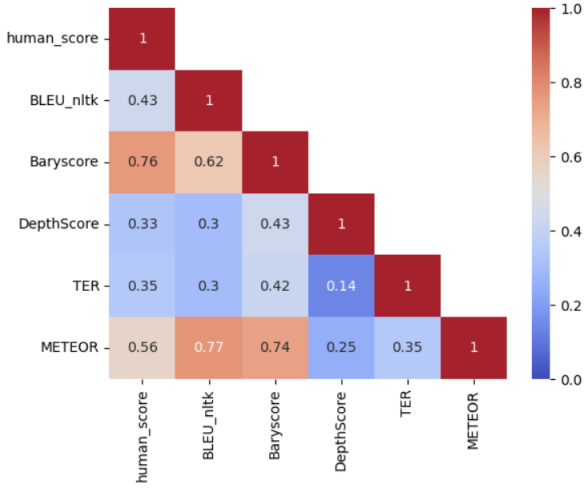

Figure 1: Matrix of the Spearman correlations between human annotation and automatic metrics

Table 1: Matrix of the Spearman p values between annotations and automatic metrics

human_score	BLEU_nltk	Baryscore	DepthScore	TER	METEOR
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00

Our results are plausible. Indeed, we have correlation between human metrics and automatic metrics similar to other work that study it, e.g. Moramarco [Moramarco et al., 2022] Shimorina [Shimorina, 2021]. It could be either due to our data (ML translation) or to the quantity of data studied.

A limitation to our work is that we have only look for correlation between human metrics and automatic metrics. However, another important point is the complementarity between human metrics and automatic metrics. It will be a way to see if "automatic metrics are similar to each other much more than they are to human metric", therefore the new automatic metrics should look for both correlation with human ones and complementary with existing metrics. The lack of complementarity between existing metrics and human metrics has be shown in Colombo et al. [Colombo et al., 2022].

References

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein Space. *Siam journal on mathematical analysis*, 2011.

OndÅ™ej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors. *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, September 2015. URL http://aclweb.org/anthology/W15-30.

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

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

Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 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, 2021b.

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

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

M. G. KENDALL. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93, 06 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL https://doi.org/10.1093/biomet/30.1-2.81.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA, September 23-27 2003. URL https://aclanthology.org/2003.mtsummit-papers.32.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 61–63, 2003. URL https://aclanthology.org/N03-2021.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. Human evaluation and correlation with automatic metrics in consultation note generation, 2022.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Laura Perez-Beltrachini, Rania Sayed, and Claire Gardent. 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, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1141.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-3301.

Anastasia Shimorina. Human vs automatic metrics: on the importance of correlation design, 2021.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 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, August 8-12 2006. Association for Machine Translation in the Americas. URL https://aclanthology.org/2006.amta-papers.25.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1462. URL https://aclanthology.org/D18-1462.

Table 2: Matrix of the Pearson p values between human annotation and automatic metrics

human_score	BLEU_nltk	Baryscore	DepthScore	TER	METEOR
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00

Appendix

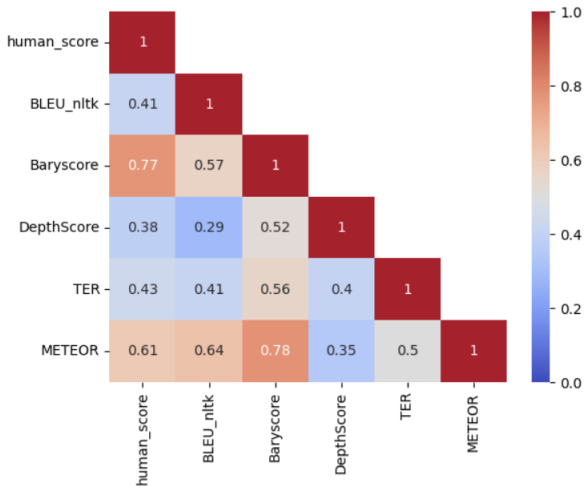

Figure 2: Matrix of the Pearson correlations between human annotation and automatic metrics

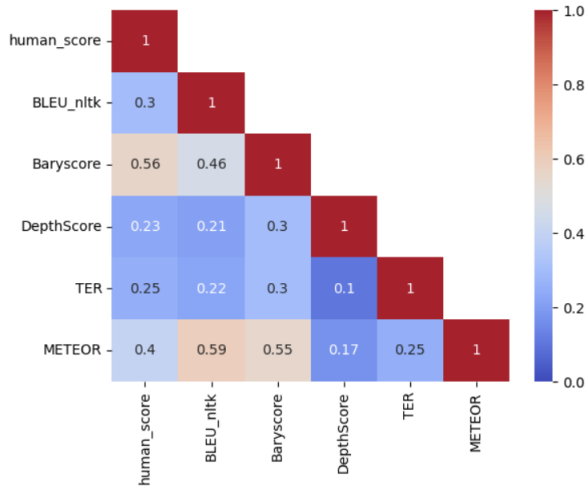

Figure 3: Matrix of the Kendall correlations between human annotation and automatic metrics

Table 3: Matrix of the Kendall p values between human annotation and automatic metrics

human_score	BLEU_nltk	Baryscore	DepthScore	TER	METEOR
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00