




Analysis of Automatic Evaluation Metric on Low-Resourced Language: BERTScore vs BLEU Score

Goutam Datta^{1,2} , Nisheeth Joshi¹, and Kusum Gupta¹

¹ School of Mathematical and Computer Science, Banasthali Vidyapeeth, Rajasthan, India
gdattal@yahoo.com, jnisheeth@banasthali.in, gupta_kusum@yahoo.com

² Informatics, School of Computer Science, University of Petroleum and Energy Studies,
Dehradun, India

Abstract. The accurate evaluation of machine translation (MT) is a difficult task. Human evaluation (judgment) is considered to be the best, but it is time-consuming. Hence, the importance of developing an automatic evaluation metric got researchers' attention. In this paper, we have done an in-depth analysis of the performance of the MT engine on low-resourced Bengali to English translations. We analyzed the scores generated by automatic metrics such as BLEU and BERTscore. We have computed the scores of the translation engine manually also based on the parameters used in the human evaluation. Finally, we have measured the correlation of BLEU and BERTScore with human judgment and found that BERTScore has a higher correlation with human judgment for our English to Bangla language pair.

Keywords: BLEU · BERT score · Human evaluation · Machine translation

1 Introduction

MT which automates the conversion of one natural language to other with the help of a sufficient parallel corpus has witnessed a tremendous paradigm shift.

MT started its journey from a dictionary-based, rule-based, statistical MT, phrase-based and most recently MT industry exploits artificial neural network (ANN) in its implementation called Neural Machine Translation (NMT). NMT has its various frameworks with their own merits and demerits [1–4].

MT evaluation is a challenging task when designing a translation system [3, 5, 6]. An evaluation is essential for determining how effective the current model is, estimating how much post-editing is required, and accordingly, the model can be improved during its design phase. MT evaluation is a challenging task since natural language is highly ambiguous. The same sentence can be interpreted differently by two different persons. In MT evaluation it compares translated text i.e. candidate text sometimes also called hypothesis text with the gold standard reference text. There may be single or multiple reference texts that can be produced by human or translation systems. When evaluating MT systems, it can either be done manually or automatically. Sometimes it demands

both. Human evaluation is best but it is time-consuming, costly, and can't be reused. In human evaluation, a quality measure scale of 1 to 5 is given accordingly the translated text is scored based on its adequacy and fluency. Adequacy refers to the completeness of the translated text. Fluency ensures the grammatical correctness of the translated text.

There are numerous automatic evaluation metrics available in the MT evaluation process. Bilingual Evaluation Understudy (BLEU) is one of the popular evaluation metrics based on precision [7]. There is another metric METEOR (Metric for Evaluation of Translation with Explicit ORDERing) which is based on both precision and recall. However, more weightage is given to recall than precision. Some other automatic metrics such as precision, recall, F-measure, ROUGE (Recall-Oriented Understudy for Gisting Evaluation), etc. are available. The most recent automatic metric is BERTScore which captures semantic similarity between reference and translated text.

In this paper, we have attempted to evaluate the accuracy of BERTScore and BLEU scores with the help of a gold standard human score while translating Bangla to English sentences.

The rest of the paper is structured as follows: Sect. 2 highlights some previous work on MT evaluation. Section 3 briefs about our methodology and experimentation. We have analyzed and discussed the results in Sect. 4. Finally, we have presented a brief conclusion and future direction in Sect. 5.

2 Some Previous Work in MT Evaluation

Human evaluation is assumed to be the best in MT evaluation but sometimes it lacks agreement among inter annotators. Also, reusability is a challenge in human evaluation. The authors addressed these two problems with human evaluation in their paper [8].

BLEU is one of the popular automatic evaluation metrics based on precision. BLEU's precision-based computation is based on token matching between a hypothesis text and one or more reference texts. Depending on how many tokens are considered i.e. $n = 1, 2, \text{ or } 3$ it is called uni-gram, bi-gram, or tri-gram. It has been found that lower grams always have a higher score than a higher gram due to their exact token matching criteria.

Another popular automatic evaluation metric is METEOR. METEOR also exploits unigram matching criteria between candidate and reference text at their surface level, and semantic level and it is based on the combination of precision and recall [9].

In Chrf, which is a language-independent, n -gram-based automatic evaluation metric where character level n -gram is exploited to compute F-score to evaluate MT performance. Chrf has shown a better correlation with a human score [10].

ROUGE (Recall-oriented Understudy for Gisting Evaluation) is a recall-based automatic evaluation metric. ROUGE has also different variants. ROUGE-N is like BLEU with multiple n -grams [11].

BERTScore is an embedded-based automatic evaluation metric. BERTScore generates a score with the help of semantic similarity between candidate and reference text, hence its accuracy during evaluation is higher than n -gram-based metrics [12]. BERTScore metric exploits BERT which is a pretrained language model [13, 14].

3 Methodology and Experimentation

In this section, we discuss our methodology used to measure the effectiveness of two popular automatic evaluation metrics: BLEU which is n-gram based, and BERTScore which is embedded based in Bangla to English translation. Our primary objective is to evaluate how well these two automatic evaluation metrics correlate with gold standard human evaluation (human score). The better one will be having a higher correlation with the human score (human judgment). To find the correlation we have used one of the commonly used correlation metrics i.e. Pearson correlation. The Pearson correlation coefficient measures the linear relationship between two variables.

Its value ranges from -1 to $+1$. -1 indicates there is a complete negative correlation and $+1$ indicates a complete positive correlation. 0 indicates no correlation. The values 0.8 and 0.6 indicate strong and moderate positive correlations respectively. The values -0.8 and -0.6 represent strong and moderate negative correlations. The methodology is represented in Fig. 1.

We used the English to Bangla tourism data set collected from TDIL (<https://www.tdil-dc.in/index.php?lang=en>). It contains English to Bangla Parallel sentences total of 11976. We have randomly picked a couple of Bangla sentences from this data set. The corresponding English sentences have been considered reference texts. These selected Bangla sentences are passed to Google translate to translate them into English.

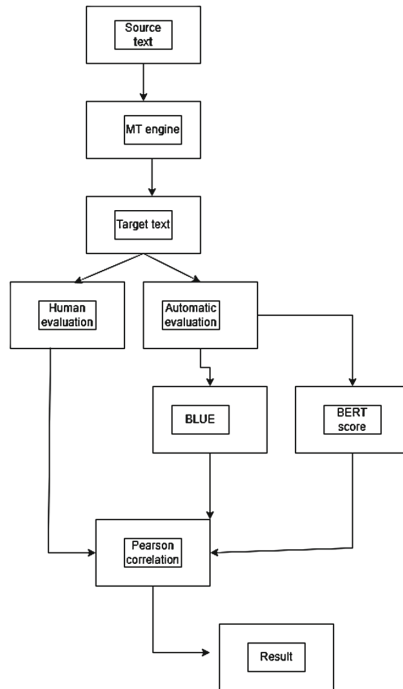


Fig. 1. Block diagram of our methodology.

The randomly picked sentences from 1 to 5 are represented in Table 1.

Table 1. Randomly picked Bangla sentences, their ground truth, and translated texts.

No.	Source Sentence	Reference Sentence (Ground Truth)	Hypothesis Sentence (Translated sentence)
1	সে মে মাসের প্রথম সপ্তাহ থেকে আমাদের সাথে দেখা করেনি।	He has not visited us since the first week of May.	He has not seen us since the first week of May.
2	তিনশো বছরেরও কম পুরনো নবীন শহর জয়পুর ঘুরে দেখ।	Take a tour of Jaipur to know the city which is fairly young, less than three centuries old.	Explore the new city of Jaipur, which is less than three hundred years old.
3	পান্ডারপুর শোলাপুর থেকে ৬৫ কিমি দূরে ভীমরথী নদীর তীরে অবস্থিত।	Pandharpur is located in a place, which is 65 km away from Sholapur, on the banks of river Bhimarathi.	Pandharpur is located on the banks of the Bhimrathi River at a distance of 65 km from Sholapur.
4	হিমাচল প্রদেশে পর্যটনের সেরা সময় অক্টোবর মাস, যখন বহুল প্রচলিত কুল্লু দাশের অনুষ্ঠান পালিত হয় সমগ্র রাজ্যে।	The best time to take up Himachal Pradesh tours is during the month of October, when the popular Kullu Dussehra is being celebrated in the state.	The best time to visit Himachal Pradesh is the month of October, when the popular festival of Kullu Das is celebrated across the state.
5	কেরালার সবথেকে আধুনিক শহর হল কোচি যেখানে সবথেকে ভাল দোকান,বাজার অবস্থিত।	Kochi is the most modern city of Kerala where the best shopping, markets and bazaars are located.	The most modern city in Kerala is Kochi where the best shops and markets are located.

3.1 Manual Score (Human Judgment)

We computed the BLEU score and BERTScore of all these translated texts. For manual score generation, had created a questionnaire that asks for some predefined questions having scales ranging from 0 to 5 to capture the adequacy and fluency of the translated sentences that we had supplied to 10 different human experts having linguistic expertise in these two languages such as Bangla and English. The human experts were given translated versions and reference texts to assign their scores. Finally, we have taken the average of all these scores given by ten different human judges. The adequacy scale has a value of 5 if all meaning is correct, most meaning has a value of 4, and much meaning, a little meaning, and none have the values 3,2 and 1 respectively. Adequacy is

used to ensure the completeness of the translated text. Fluency ensures the grammatical correctness of the translated text. The fluency scale is as follows: the highest score of 5 is assigned to flawless English, good English has a score of 4, and non-native, disfluent, and incomprehensible English have scores of 3, 2, and 1 respectively [15].

3.2 BERTScore

BERTScore is computed by feeding ground truth (reference sentence) and candidate sentence into the pre-trained BERT model. The BERT model has words that are contextually embedded. It tries to match tokens of hypothesis and reference texts with cosine similarity. The BERTScore produces the following output values: precision, recall, and F1-score whose range varies from 0.0 to 1.0.

BLEU

BLEU is a precision-based metric since during its computation it does not consider whether all the words in the reference texts are covered in the hypothesis text or not. BLEU tries to match the MT engine-generated text with one or more reference texts based on how many tokens are considered at a time. That is based on the number of tokens selected for matching it can be 1-g, 2-g, etc.

The computed automatic and manual scores are presented in Table 2. Its diagrammatic representation is given in Fig. 2. The BERT score and human judgment (human score) correlation is given in Table 3.

The BLEU score and human score (human judgment) correlation is presented in Table 4. The observed pattern between two different automatic evaluation metrics and Human scores is discussed in Sect. 4 (Result Analysis and Discussion).

Table 2. Automatic and human scores of the translated texts.

Sentence no	MT Engine	BLEU score(n = 3)	BERTScore	Human judgment
1	Google Translate	0.66	0.93	0.95
2	Google Translate	0.16	0.66	0.85
3	Google Translate	0.17	0.68	0.95
4	Google Translate	0.29	0.73	0.96
5	Google Translate	0.13	0.78	0.94

Table 3. Pearson Correlation between BERTScore and Human Score

BERTScore	Human judgment	Pearson correlation
0.93	0.95	0.46
0.66	0.85	0.46
0.68	0.95	0.46
0.73	0.96	0.46
0.78	0.94	0.46

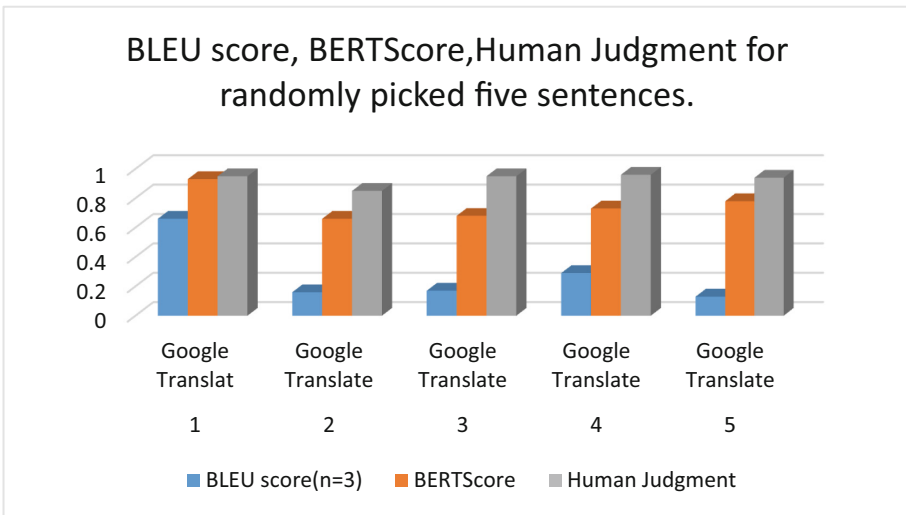


Fig. 2. Various automatic and Human scores for randomly picked five translated sentences.

Table 4. Pearson correlation between BLEU score and human judgment (human score)

BLEU Score	Human judgment	Pearson correlation
0.66	0.95	0.34
0.16	0.85	0.34
0.17	0.95	0.34
0.29	0.96	0.34
0.13	0.94	0.34

4 Result Analysis and Discussion

Analyzing the results, we obtained in Sect. 3, we can see that the automatic evaluation metric BERTScore exhibits a higher correlation with human **judgment** compared to the

n-gram-based BLEU metric (Table 3). The Pearson correlation coefficient has been computed separately between these two automatic metrics and human **judgment** (Human score), i.e., the BERT score vs human score and the BLEU score vs human score. As per the correlation values in Tables 3 and 4, BERTScore always has a higher correlation value with human scores because of its ability to capture a contextual representation of reference and hypothesis texts. Since BLEU tries to match exact tokens between candidate and reference sentences, it fails to generate an authentic score since the word may have its synonym. Further, when we analyze the BLEU score and BERTScore of all these five sentences, it has been found that sentence 1 has the highest score in both the automatic metrics compared to the rest of the sentences (Table 2). The reason for this is reference text and hypothesis text both have maximum token matching in this sentence (Table 1). Hence BLEU and BERTScore both have generated the highest score for this sentence based on their own measuring criteria.

5 Conclusion and Future Work

MT is a fast-growing field. Researchers are continuously working in this domain to upgrade the model to achieve higher accuracy. However, during model design, automatic performance evaluation of the model plays a vital role. Designing an automatic evaluation metric is a challenging task because of its inherent linguistic, syntactic, and semantic intricacies that need to be checked in the hypothesis and reference texts while evaluating the generated (hypothesis) text. Hence, using appropriate evaluation metrics is important. We have understood the patterns of BLEU and BERTScore with gold standard human judgment. For this, we have used appropriate correlation i.e. Pearson correlation. Pearson correlation is suitable when we want to find a linear correlation between the two variables. Based on the patterns of correlation with the human score one can select the appropriate evaluation metric.

However, based on this study, we can say that we still have to go far in the MT automatic evaluation metric. Designing interpretable automatic evaluation metrics which is context-oriented is essential to achieving higher accuracy. However, to design such an evaluation metric creating a domain-specific reference corpus is equally important to achieve the task.

References

1. Stahlberg, F.: Neural machine translation: a review. *J. Artif. Intell. Res.* **69**, 343–418 (2020)
2. Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. 1–15 (2015)
3. Ramesh, A., Parthasarathy, V.B., Haque, R., Way, A.: Comparing statistical and neural machine translation performance on Hindi-To-Tamil and English-To-Tamil. *Digital* **1**, 86–102 (2021)
4. Zhang, Z., Liu, S., Li, M., Zhou, M., Chen, E.: Bidirectional generative adversarial networks for neural machine translation. *CoNLL 2018 - 22nd Conf. Comput. Nat. Lang. Learn. Proc.* 190–199 (2018) <https://doi.org/10.18653/v1/k18-1019>
5. Lankford, S., Afli, H., Way, A.: Human evaluation of english-irish transform-er-based NMT. *Information* **13**, 309 (2022)

6. Fomicheva, M., Specia, L.: Taking MT evaluation metrics to extremes: beyond correlation with human judgments. *Comput. Linguist.* **45**, 515–558 (2019)
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* 311–318 (Association for Computational Linguistics) (2002). <https://doi.org/10.3115/1073083.1073135>
8. Joshi, N., Mathur, I., Darbari, H., Kumar, A.: HEVAL: yet another human evaluation metric. *Int. J. Nat. Lang. Comput.* **2** (2013)
9. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. *Intrinsic Extrinsic Eval. Meas. Mach. Transl. and/or Summ. Proc. Work. ACL 2005* 65–72 (2005)
10. Popović, M.: CHRF: character n-gram f-score for automatic MT evaluation. 10th Work. Stat. Mach. Transl. WMT 2015 2015 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2015 - Proc. 392–395 (2015) <https://doi.org/10.18653/v1/w15-3049>
11. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out* 74–81 (Association for Computational Linguistics, 2004) (2004)
12. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: evaluating text generation with BERT. 1–43 (2019)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bi-directional transformers for language understanding. *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.* **1**, 4171–4186 (2019)
14. Hanna, M., Bojar, O.: A Fine-Grained Analysis of BERTScore. *WMT 2021 - 6th Conf. Mach. Transl. Proc.* 507–517 (2021)
15. Koehn, P., Och, F.J., Marcu, D.: *Statistical Phrase-Based Translation*. 48–54 (2003)