Neural Metrics for Machine Translation Suck - But So Does Everything Else

Courtney Mazzulla
cleemazzulla@berkeley.edu

Tynan Prasad
prasadtyrnan@berkeley.edu

Julian Rippert
jrippert@berkeley.edu

Abstract

This document provides an in-depth analysis of machine translation (MT) metrics, their techniques, and the trade-offs associated with their usage. Evaluating the quality of machine translation output is crucial for assessing translation system performance. However, different evaluation metrics have distinct characteristics and considerations, leading to trade-offs regarding their strengths and limitations. Subsequently, this paper proposes a neural network-based approach to replicate features that human raters use during their assessments. This approach achieves incredible performance, considering it is built using a lightweight model architecture with just a few features. This paper provides the foundation for future work toward leveraging feature engineering to develop metric models that produce machine translation scores that highly correlate with human raters.

Please see our Code.

1 Introduction

This paper presents the intended contribution of the UC Berkeley School of Information to the WMT 2023 Metrics Shared Task (Freitag et al., 2022).

In neural network architectures and methodologies, advances have led to increasingly robust machine translation (MT) task performance. Such advancements have outgrown previous MT metrics such as BLEU (Papineni et al., 2002) and ROUGUE (Lin, 2004). These rule-based metrics concentrated on lexical matches and could not decipher ambiguity and complex meaning within human language. Although expert human raters are still considered the gold standard, neural (or learned) metrics have shown stronger correlations to human judgment.

The WMT Metrics Shared Task is an annual benchmark for comparing translation performance metrics and their ability to approximate human assessments. Among the WMT Metrics Shared Task, top performers are metrics such as BLEURT-20 and COMET-22 which continue to improve on correlation with human raters even across text domains and translation challenges. (Freitag et al., 2022)

The WMT 2022 results (Freitag et al., 2022) noted that there had been a slowing of new neural metrics achieving state-of-the-art (SOA). Withstanding any novel model architectural changes, the top performers have resulted in simply training larger models or developing ensemble-based metrics. As such, we recommend to improve upon BLEURT-20 by:

1. Further fine-tuning the top performing model checkpoint on MQM data, which spans three language pairs across domains such as News, Social, and eCommerce.

2. Performing augmentation techniques on the training data to improve performance.

The two augmentation techniques implemented on the training data are as follows:

1. Employment of lexical (synonym) substitution by replacing a randomly selected token with semantically similar words to compensate for the scarcity of labeled data, ultimately helping our model avoid over-fitting. This approach combats the words not covered by the classification model - a.k.a. out of vocabulary (OOV) words, by using semantically similar words to improve robustness and generalization of the metric automatically. (Elekes et al., 2019)

2. Reordering and shuffling sentences when applicable. We can leverage this naive technique to create further augmented language sentence pairs.

Finally, we present a novel approach which leverages feature engineering to extract explicit lexical features from the language pairs. Using the explicit features we demonstrate the ability to outperform the fine-tuned BLEURT model through using a lightweight feed-forward neural network.

2 Corpora

Three distinct data sets are provided as part of the annual WMT task competition. Each data set consists of MT language pairs in various languages and unique human rater scores:

1. Directs Assessments (DA) - For each input, humans rate the output from each system with an absolute score or label. The sequence-level ratings can then be used to calculate system-level ranking. DA is the official ranking for the translation shared task since the WMT17 metrics challenge.

2. Scalar Quality Metric (SMQ) - This metric collects segment-level scalar ratings with document context. SQM uses a 0-6 scale for translation quality assessment. Another difference is that the sentences were rated by professional translators instead of crowd workers or researchers.

3. Multidimensional Quality Metric (MQM) - Metric based on an error hierarchy consisting of the following dimensions: Accuracy, Fluency, Terminology, Style, and Locale. Each criterion is rated an error severity independent of category and then weighted to develop a composite score.

3 Background

BLEURT was a novel machine translation evaluation breakthrough. Fundamentally, BLEURT is a regression model with a trained linear layer built upon the [CLS] token of the underlying BERT model. At the time of creation, human-labeled rating data was scarce. As such, before fine-tuning the 2017 WMT available data, the BLEURT team 'warmed up' the neural evaluation model by pretraining it on synthetic data (reference-candidate pairs) from Wikipedia. These sentences were 'perturbed' in three ways:

1. Mask Filling with BERT - Words within sentences were randomly selected to be masked. Up to 15 words per sentence could be masked and were selected by the language model using beam search of size 8.

2. Back-Translation - Using a transformer model for the English-German language pairs

3. Word Dropping - Words randomly dropped in a uniform distribution up until the entire length of the sentence, applied to roughly a third of all back-translated sentence pairs.

3.1 Baseline

Before conducting our experiments, we ran analyses to establish a baseline of neural models and their correlations with human ratings. Given computational restraints, we used the BLEURT-20-D6 model checkpoint to gauge BLEURT's performance. BLEURT-20-D6 is a smaller model than the top BLEURT-20 checkpoint with 45M and 579M parameters, respectively. Although BLEURT-20-D6 is a lower-performing model, it showed a significantly higher correlation with human ratings than the BLEU metric. See Figure 1 below.

Figure 1: Baseline Results on MQM Data

Baseline Pearson Correlation with Human Rating			
Model Name	zh-en	en-de	en-ru
BLEU	0.0939	0.1025	-0.0014
BLEURT-20-D6	0.3226	0.2678	-0.1369

4 Experiments

4.1 Data Augmentation

In order to increase the robustness of our metric model, we implemented two data augmentation techniques to augment the language pairs. It is important to note that while the text of the translations were altered, the human rater score remained the same. This brings up a risk that by augmenting the sentences, the scores may not reflect the new augmented text appropriately. However, an analysis conducted (Wei et el., 2019) found that when comparing language pairs and their augmented counterparts in a two-dimensional space, the embedding of the augmented pairs closely mirrored those of their original sentences. Additionally, there is minimal concern that such augmentation would decrease the performance of any deep model considering “deep learning is robust to massive label noise" (Rolnick et al., 2017).

4.1.1 Synonym Substitution

In order to help increase the robustness and 'generalizability' of our neural metric, exposing the model to a broader vocabulary is crucial. The model randomly picks one token from the machine-translated sentence and uses a sizeable linguistic database, Wordnet toolkit, to find synonyms. Knowing that there may be no synonyms or many depending on the word, the model will always pick the first synonym if available.

4.1.2 Sentence Re-ordering

We further introduced noise into our augmented language pairs by sentence reordering. The reordering is a naive approach to further help with overfitting and help increase the robustness of the learned metric. Additionally, since only 11.3% of the machine translations have more than one sentence, there is little risk that the quality of the labels will be damaged overall.

After fine-tuning the BLEURT-20-D6 model using the augmented data set, we noted a tremendous drop in performance compared to the non-finetuned checkpoint. We presume the performance drop was because the implemented synonym substitution method was not robust enough to comprehend the ambiguity of human language. Implementing a BERT-based synonym substitution may improve this challenge in future training. Ideally, BERT would help find the embedding representation of another token closest to the embedding of the randomly selected token to replace. Additionally, the synonym substitution method's effectiveness would be enhanced through hard-coded rules. For instance, inserting synonyms for the words 'the' or 'I' did not seem to provide much value.

4.2 Abstractive Approach

In addition to fine-tuning BLEURT, we were also interested in pursuing an alternate approach, focusing on extracting features instead of utilizing large transformers. This involved calculating things like the overlap of engrams of varying sizes, the count of tokens in both the reference and candidate string, and something we termed the 'mutation string', which categorized errors into three categories and assigned them varying magnitudes. Once those features were generated, they were used in a lightweight neural network consisting of one hidden layer with 8 neurons and an output layer targeting a normalized version of the MQM assessment scores. The theory behind the feature engineering driven model approach is that when the human evaluators score translations using the MQM rubric, they are looking for a handful of expressible things using explicit rules. If we can identify some of those assessment rules, we can produce them as explicit features and use them in a model.

4.2.1 Mutation Strings

Using DNA mutations as an analog, we can categorize mutations between the candidate and reference strings as either deletions, insertions, or substitutions. A deletion occurs when a token or set of tokens is present in the reference and not in the candidate in a particular position. For example, the mutation string for reference "The man went on the boat" and candidate "The man went on boat" is "The man went on DEL boat". An insertion is the inverse of a deletion; a token present in the candidate at a position is absent from that position in the reference. For example, reference "The man went on the boat" and candidate "The man went on the big boat" becomes "The man went on the INS boat". A substitution is when the tokens present at a position in the reference and candidate are different. For example, candidate "The man went on a boat" and reference "The man went on the boat" produce mutation string "The man went on SUB boat". Once the mutation strings were calculated, we collected the count of mutations of each type for each reference and candidate pair.

4.2.2 Feature Analysis

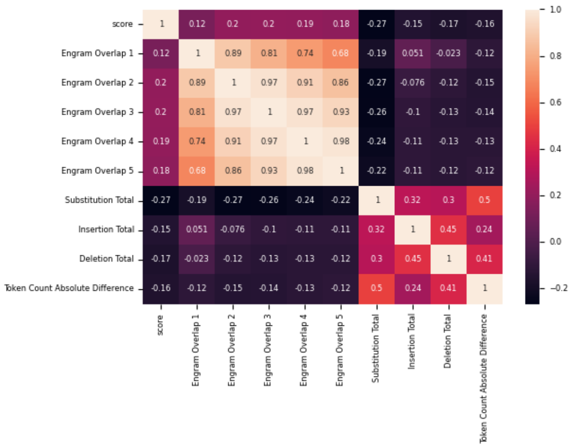
Once all of our features had been calculated, we performed some analysis to get a sense of their predictive power before actually training a model on the data.

From this, we saw positive correlations between the engram overlaps of different sizes with the MQM scores and negative correlations with the count of mutations and the absolute difference in token counts between candidate and reference. Of note are the high correlations among the variable engram overlaps, which is to be expected, and the high correlations between the counts of different mutations and the absolute difference in token counts.

4.2.3 Models

Fully connected models with one hidden layer of size eight with a RELU activation function and an output layer of size one with a sigmoid activa-

Figure 2: A heatmap of the correlations of the features and scores



tion function were trained for three epochs on the features and normalized scores using a variety of feature sets. Using only the token counts, the model was able to outperform BLEU and BLEURT-20-D6 on zh-en data, achieving a correlation of 0.47605, but under-performing compared to BLEURT-20-D6 on the en-de data, with a correlation of 0.20148. A model trained only using the engram overlaps resulted in correlations of 0.17797 and 0.20453 on the zh-en and en-de data, respectively. Only once we trained a model using only the counts of each mutation type that we beat BLEURT-20-D6 on all measures, achieving correlations of 0.50302 and 0.30395 on zh-en and en-de data, respectively. Combining all the features under one model does not yield significant performance increases, resulting in a correlation of 0.49897 for zh-en and 0.28019 for en-de. Given disparities in the human evaluations of German and English translations, it is possible to yield increased performance by training two models, one for each language. However, to keep the results analogous to those of the BLEURT-20-D6 model, we are only reporting those of models trained on both. See Figure 3 below for a summary of the results.

5 Results

After running baseline analyses and initial experiments, we quickly noted that the English-Russian language pair performed poorly. Further EDA revealed that the MQM scores for this language pair were non-standardized. Dropping this language pair from the data set provided an immediate improvement for the Chinese-English and the English-German language pairs.

Figure 3: Final Results with MQM Data

Pearson Correlation with Human Rating		
Model Name	zh-en	en-de
BLEURT-20-D6	0.3833	0.2861
BLEURT-20-D6 (Augmented)	0.2291	0.1678
FNN (Engram Overlaps)	0.1778	0.2015
FNN (Token Counts)	0.4761	0.2015
FNN (Mutation Counts)	0.5030	0.3040

Some of the models are working off feature schemes that have obvious flaws when exposed to certain inputs. For example, working from token counts, a perfect translation, and a random combination of tokens that happens to have the same length as the reference look the same. Clearly, those should receive a different evaluation. However, because the data in this set is limited to relatively good translations resulting from submissions to the machine translation competition, it is possible to achieve a decent degree of accuracy working from this flawed scheme. Mutation counts communicate some of the same information (a pair with no deletions or insertions must be the same length) but do not have the aforementioned weakness of token counts. As such, we feel comfortable positing mutation strings as a plausible candidate solution to translation quality assessment.

6 Conclusions

Calculating a standardized, human assessment of the quality of translation is a complicated task. With sufficient guidelines, assessments can be generalized between assessors, and the judgment ceases to be nuanced with too many rules. Many of the scores in the MQM data set are based on one person's opinion, making it hard to determine if that score is accurate to the general understanding of the quality of that translation. The requirement for expertise in evaluators renders it even more challenging to get a large number of assessments for each translation. Showing that the scores can be predicted with some accuracy using something as simple as the number of tokens present in each string shows that either this data is too skewed towards high-quality translations or the rubric being used is too rigid. It calls into question whether any neural metric assessed in this way can actually have its quality determined.

That said, we do believe that derived features have the potential to be helpful in future neural metrics. Even without guidelines to follow, readers of translations will rely on some heuristics to determine if the translation is a good one. By making some of those possible strategies explicit features for our models, we may empower them to make more human decisions.

Limitations

We expect that our method works mostly for languages with limited morphology, like English, In addition, our model is limited to language pairs and text sizes of few sentences at maximum. We would expect translations of paragraphs for instance to introduce much more complexity. Domain generalization (or the sustained increase of correlation with human judgement) across text of various topics is only possible for languages within the scope of the model as well as learned context. For instance, our neural metric would not be an appropriate tool to be used for machine translation in the medical field barring further fine-tuning with appropriate data.

Additionally, computational limitations like limited GPU resources hindered us from implementing the BERT based synonym substitution which proved to be a very memory heavy method.

Acknowledgements

This research paper was created as a part of the Natural Language Processing course (W266) in UC Berkeley's Masters in Data Science program. Special thanks to our instructors Mike Tamir, PhD and Paul Spiegelhalter, PhD for their consultation, proof-reading, feedback and continued support.

7 References

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In Proceedings of the Seventh Conference on Machine Translation, Abu Dhabi. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7881–7892, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1 /2020.acl-main.704

David Rolnick, Andreas Veit, Serge J. Belongie, and Nir Shavit. 2017. Deep learning is robust to massive label noise. CoRR, abs/1705.10694.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of ACL.

Abel Elekes, Simone Di Stefano, Martin Schaler, Klemens Bohm, and Matthias Keller. 2019. Learning from Few Samples: Lexical Substitution with Word Embeddings for Short Text Classification

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Workshop on Text Summarization Branches Out.