Nova IMS MT Metrics Shared Task

Catarina Natário Moreira, David Sotto-Mayor Machado, Maikel Sousa
20201034, 20201023, 20200735

1 Introduction

The conference on Statistical Machine Translation (WMT) is a prestigious conference in the field of NLP and Machine Translations that builds on a series of annual workshops and conferences in Machine Translation, going back to 2006. One of those competitions is the Metrics shared task whose goal is to create a metric that correlates with human assessments of quality. During this work, we had the pleasure to participate in a simulated version of this shared task.

In recent years, machine translation (MT) has already become part of our everyday life and it has changed dramatically as it has for all text handling technologies. Thus, machine translation is a subfield of computational linguistics that draws ideas from linguistics, computer science, information theory, artificial intelligence and statistics. With the use of this fields, it investigates the different approaches to translating text from one natural language to another. For a long time, it had a bad reputation for being perceived as low quality. However, in the last two decades, there has been a great progress in the quality of MT, which has also made it interesting for use in the translation industry. Although the quality has increased, it is still far worse than human translation. However, this does not mean that there are no good results when applied and that it is not a good practice to use. It should be noted that integration of human and machine translation is a promising workflow for the future, however it will not replace human translations. It will help them and increase the productive in the translation process.

Although in past, translation agencies and other professional translators were the ones that worked in the translation industry, we are currently facing a large and rapid growth in the range of machine translation solutions that are quite useful for practice. Besides translation, it is important to have the right evaluation metrics for MT. Nowadays, these evaluation metrics for ML have increasing however they are not very good on distinguishing good from merely fair translations. We believe that the main issue can be the fact that they have inability to properly capture the meaning of the sentences. In this work, with help of different evolution metrics, we will evaluate the difference between the reference text and the translation text. After knowing the score of each metric, we will apply the *Pearson* and *Kendall* correlation as an assessment criterion to choose the metric with the best score.

2 Method/Approach

2.1 Analysis and Pre- processing

Initially, we started by importing from the corpus the 6 different corpora that the professor made available. After that, and after some analysis of each corpora, we decided to remove rows with missing translations or references. We then started with the pre-processing part for each corpora presented. First, we did a small inspection of each corpora. Next, we calculated the number of annotators and after that, we calculate the mean of the avg-score also for each corpora. Finally, still in this inspection part, we calculated and presented the 10 most frequent words in each corpora. As expected, in the corpora in which the translation was made into English, the words presented were almost the same despite being in different corpora and in different documents.

With this initial analysis, it was possible to discover the next steps to be taken. In this way, we started by doing an initial preprocessing.

For the corpora where the translation was done for Finnish and for the corpora where the translations were done for English, we started by putting all the text in a lower case. After that, we removed numerical data and the punctuation. Here, we had to keep in mind since Finnish has letters that English does not cover. After that, we remove the stop words from each corpora and, taking into account the different languages. Finally, we calculated the lemmatizes and the stemmers for each corpora and for each document. However, we only apply the lemmatizes on corpora.

For the Chinese translation, we created a function to remove numerical data, punctuation and also the Chinese stop words.

2.2 Bag-of-Words

After this preprocessing, we decided to apply the bag-of-words. In this case, we created a function that receives a list of documents - corpus - and extract the top k most frequent n-grams for that corpus.

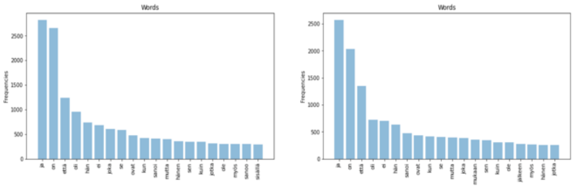


Figure 1 – The 10 most popular unigrams in the reference and translation columns, respectively (corpora ***en-fi***)

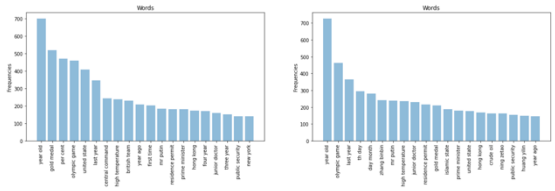


Figure 2 – The 10 most popular bigrams in the reference and translation columns, respectively (corpora ***zh-en***)

2.3 Evolution Methods

Evolution metrics shape the direction of research. They are used to compare experiments, to decide what gets published, to identify weakness and determine what to work on and to decide which model we want to deploy. However, there are some challenges that MT brings. Firstly, one human reference isn't the only correct target translation. Secondly, human evaluations are the gold standard but are costly and time consuming.

We can group existing metrics into:

- **Lexical Metrics** (BLEU, METEOR and chrF)
- **Embedding base** Metrics (BERTScore)
- **Learnable Metrics** (COMET, BLEURT)

The aim of this project was to create a model capable of calculate the score between the translation and reference column. In this way, it is possible to know if the translation was made correctly or not. For this work, after doing the pre-processing, we apply:

BLEU Scores – Bi-lingual Evaluation Understudy

This method measures the n-gram precision of a translation against a reference and Bleu score is computed at the corpus level, and it combines the unigram, bigram, trigram and 4-gram precision into a single score between 0 and 1. However, we only apply BLEU-4 (cumulative 4-gram BLEU score). Although the algorithm does not take a long time to process, both the results (score) and the correlation are low.

ROUGE-N (1 and 2) and ROUGE-L

ROUGE-N measures the number of matching n-grams between our model generated text (translation) and the reference. Here, we opt to use ROUGE-1 and ROUGE-2, and, with this method, we are able to measure the match-rate of unigrams and bigrams, respectively, between our model output(translation) and reference.

After that, we applied ROUGE-L, and this one measure the longest common subsequence between our model output and reference. All this means is that we count the longest sequence of tokens that is shared between translation and reference.

After choosing the ROUGE-N we used and ROUGE-L, we calculate ROUGE recall, precision and F1-Score. This algorithm presents both better F1-Score and correlation values.

WMD – World Movers Distance

Initially, Word Movers Distance was not proposed specifically for Machine Translation, but it has been used for it.

World Mover's Distance is based on recent results in word embeddings that learn semantically meaningful representations for words from local co-occurences in sentences.

This method suggests that distances between embedded word vectors are, to some degree, semantically meaningful. It utilizes this property of word vector embeddings and treats text documents as a weighted point cloud of embedded words.

Although we have low values, the correlation is higher with this method.

BERTScore – Bidirectional Encoder Representations from Transformers

This method is a Transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. BERT was created and published in 2018 by Jacob Devlin and his colleagues from Google. As of 2019, Google has been leveraging BERT to better understand user searches.

The original English-language BERT has two models: **the BERT$_{\text{BASE}}$**: 12 Encoders with 12 bidirectional self-attention heads, and **the BERT$_{\text{LARGE}}$**: 24 Encoders with 24 bidirectional self-attention heads.

Here we tried to apply this method with the initial pre-processing and without any pre-processing. After evaluating, they presented similar results.

This method takes a long time to process. However, it presents a median correlation.

BLEURT

BLEURT was developed from BERT providing state-of-the-art results for three consecutive years of the WMT Metrics shared task.
This method demands the same computing resources as BERTScore and we couldn't find considerable differences from this last metric.

COMET

COMET is an open-source framework for MT evaluation that can be used to evaluate MT systems with our currently available high-performing metrics and to train and develop new metrics. COMET can capture semantic similarities even where there is lexical disparity.

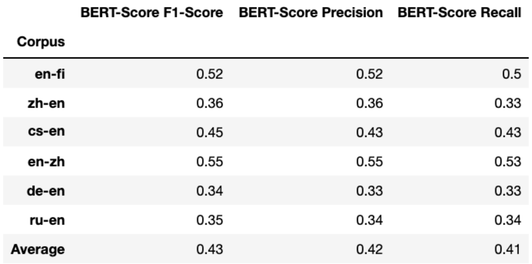
For each method, we assess its performance using *Kendall* and *Pearson* correlation.

Comparing this metric to the previous ones, it was much more demanding in computing resources and achieved very similar results.

3 Results and Discussion

As we mentioned before, we apply different evaluation methods to evaluate the translation in relation to the reference.

After thoroughly analyzing the results obtained and exploring the *Pearson* correlation of the different methods, we decided that the BERTScore would be the metric to use against the test set.



	BERT-Score F1-Score	BERT-Score Precision	BERT-Score Recall
Corpus			
en-fi	0.52	0.52	0.5
zh-en	0.36	0.36	0.33
cs-en	0.45	0.43	0.43
en-zh	0.55	0.55	0.53
de-en	0.34	0.33	0.33
ru-en	0.35	0.34	0.34
Average	0.43	0.42	0.41

Table 1 – Pearson results for BERT-Score

4 Conclusion

BERTScore was the metric with the best performance in our initial corpus in terms of correlations with z-score.
According to the documentation, this method is the state-of-the-art metric for evaluation of Machine Translation task, and it is used as base for BLEURT. COMET was trained with the initial corpus, and we found some computing limitations hence we considered that it was not the best option for this project.

References

[1] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, e Y. Artzi, «BERTScore: Evaluating Text Generation with BERT», arXiv:1904.09675 [cs], Fev. 2020

[2] «Issue #109 - COMET- the Crosslingual Optimised Metric for Evaluation of Translation | Iconic Translation Machines».