CISC7021 Applied Natural Language Processing

Final Report (Group 37)

Comparison of Alternative Evaluation Approaches for Natural Language Applications

CHAN Chi Weng **KE Siyun** **LEONG Ka Hou**

University of Macau

{mc35171, garfield.ke, mc35254}@connect.um.edu.mo

Abstract

This study demonstrates a review of the capacity of several evaluation approaches for the quality of machine-generated texts. A significant finding from our research is that these scoring algorithms do not adequately account for the substitution of synonyms. Despite producing ostensibly correct translations, the algorithms often failed to recognize synonym substitutions as valid, leading to an underestimation of translation quality.

1 Introduction

While evaluating the performance of natural language processing techniques like machine translation systems, some alternative evaluation methods are declared to be more robust, accurate, and practical than traditional methods like BLEU score etc. This project aims to investigate and verify a few novel approaches for evaluating NLP applications (e.g. SMT/NMT models or similar text generation applications) and attempts to analyze the advantages or limitations of the novel methods.

Generally, this project refers to a "*BERTScore*" evaluation metric (Tianyi, 2020), a "*BARTScore*" algorithm (Weizhe, 2020), for the evaluation of *text generation* models, it also considers **a neural evaluation framework** "*COMET*" (Ricardo, 2020) evaluating the performance of *multilingual MT models*.

Additionally, this project also observed some potential well-performed models like an evaluation framework "*GPT Score*"(Jinlan, 2023)

2 Methodologies description

The project conducts a series of experiments to test and compare both traditional evaluation methods (e.g. BLEU), and some new evaluation methods, to understand why these methods are controversial and discuss their pros and cons in machine translation or other NLP applications. Briefly, the experiments consider the performance of evaluation methods on actual machine-generated parallel texts obtained through well-known translation solutions.

2.1 Obtaining datasets

To obtain parallel datasets containing both the reference texts and machine-generated texts as candidate texts. We referred to the Chinese and English parallel sets retrieved from the International Workshop on Spoken Language Translation (IWSLT) 2014 evaluation campaign. The Chinese texts are processed through a third-party translation solution so as to obtain machine-generated English texts as candidate texts, and the original English texts in IWSLT2014 datasets are regarded as reference texts in our experiments.

The dataset is released on GitHub (stated in the Acknowledgement) for reproduction. The dataset used in our research comprised a substantial total of 184475 lines of Chinese text. Given the large volume of data, we explored various translation APIs for our task, including options from Hugging Face and Deep Translation. However, these APIs did not meet our expectations in terms of translation quality, performance, or cost-effectiveness.

Therefore, we ultimately chose to utilize the Google Translation API for our task since the cloud translation solution from Google is regarded as a gold standard in the industry. Despite some limitations, the Google Translation API still provides a stable and free solution that could handle our extensive dataset while maintaining a reasonable level of translation quality, which made it an ideal choice for our research.

2.2 Dataset pre-processing and experiment categories

In the study, 6 experiments (categories *a-f*) with varied arrangements of reference and candidate sets are performed. Among these reference and candidate sets, in order to better understand the differences among various evaluation algorithms, we applied five distinct treatments to the original MT results. The treatments were designed to represent a range of potential translation scenarios, thus enabling us to test the robustness of the evaluation algorithms across diverse situations. In detail, the following treatments can be observed in the study,

i. **Original.** These original MT results were directly sourced from the Google Translation API and were not subjected to any further processing or modifications. The rationale behind using these unaltered results was to establish a reference point or a baseline against which the treated results could be compared.

ii. **Word shuffled.** One such method involved randomly shuffling the words within each translated sentence while ensuring that this shuffling was confined to each sentence. In other words, no word was relocated from one sentence to another. The primary objective of employing this shuffling method was to create a comparison benchmark with the original results. This would allow us to measure the extent of the scoring deviation introduced by the disruption of sentence structure.

iii. **4-gram reordered.** This is a treatment like the previous shuffling treatment, the words in sentences are reordered by every 4 words, which promises there is a significant change in the actual meaning and readability of these English sentences, but they reserved some 4-gram similarities compared to the original texts. A good evaluation method is expected to observe the loss in meaning and not to be confused by the false similarities.

iv.**Synonym** The purpose of implementing this synonym substitution was to assess whether the various scoring algorithms could effectively understand and account for semantic equivalence. By replacing words with their synonyms, we aimed to maintain the semantic integrity of the sentences while altering their lexical composition. This method allowed us to test the sensitivity of the evaluation algorithms to semantic variations in the text, providing valuable insights into their ability to evaluate translation quality beyond mere lexical matching.

Meanwhile, the following categories of experiments are conducted on the evaluation methods with those pre-processed datasets,

a) **Original standard English texts vs. Original translated English texts.** The evaluation methods are expected to generate a relatively high score on this pair to acknowledge a good translation.

b) **Original standard English texts vs. Shuffled translated English texts.** The evaluation methods are expected to return a low score on this pair to represent the loss of meaning in sentences.

c) **Original standard English texts vs. 4-gram reordered translated English texts.** The result shall be similar to the last pair but the score could be a little higher.

d) **Original translated English texts vs. 4-gram reordered translated English texts.** The evaluation methods are expected to report a relatively low score on this pair to represent the loss of meaning in sentences.

e) **Original standard English texts vs. 4-gram reordered standard English texts.** The result shall be similar to the result in experiments in *d)*.

f) **Original standard English texts vs. Standard English texts with synonyms replaced.** A good evaluation method is expected to provide a high rating similar to the original dataset.

3 Experiments and analysis

3.1 Evaluation approaches to be tested

To ensure comparability and visibility in our experiment, we employed four different algorithms to evaluate the quality of the translations.

Using multiple evaluation algorithms allowed us to cross-verify our findings and ensured that our conclusions were not biased towards the peculiarities of any particular evaluation method. Each algorithm was applied to the original MT results as well as to the results of each of the six treatments, providing a comprehensive view of their performance under various translation scenarios. The following approaches are considered in the experiments,

BLEU. The first of these evaluation methods was the Bilingual Evaluation Understudy (BLEU) score. The BLEU score is a widely used metric for machine translation evaluation. It compares a candidate translation to one or more reference translations, measuring the precision of n-grams in the translated text. This metric is valued for its simplicity and correlation with human judgment, although it has certain limitations, particularly concerning semantic equivalence.

BARTScore. The BARTScore is an evaluation model based on BART, an advanced sequence-to-sequence model. It evaluates the quality of generated text by measuring how likely the model can reconstruct the generated text.

The idea behind BARTScore is that if the quality of the generated text is high, the model will be able to convert the generated text back to the reference output or source text with a higher possibility.

BARTScore has a shorter evaluation time compared to metrics like BLEU and can be computed fully automatically without human references. This enables it to evaluate a broader range of generated text, not limited to tasks with references like summarization and translation.

BERTScore. Following the use of the Bilingual Evaluation Understudy (BLEU) score, the second evaluation method we employed was the BERTScore. BERTScore is a metric for evaluating text generation which leverages the pre-trained contextual embeddings from BERT (Bidirectional Encoder Representations from Transformers). Unlike BLEU, which relies on n-gram precision, BERTScore measures the similarity between the BERT embeddings of the candidate and reference sentences. This allows for a more nuanced understanding of the semantic and syntactic similarities between the compared texts.

There are 3 scores in BERTscore, which are

- Precision, abbreviated as P, measures the percentage of predicted positive samples that are correctly classified as positive. It is defined as the ratio of true positives and overall predicted positives. High precision indicates a low false positive rate.
- Recall, abbreviated as R, measures the percentage of actual positive samples that are correctly predicted as positive. It is defined as the ratio of true positives and overall actual positives. High recall indicates a low false negative rate.
- F1-score, abbreviated as F1, provides a harmonic mean of Precision and Recall, balancing both metrics. It is defined as:

  *F1 = 2 * (Precision * Recall) / (Precision + Recall)*

  F1-score reaches its best value at 1 and worst at 0. So it is commonly used as a summary metric to evaluate classifiers, with a higher F1 score indicating better performance in terms of both precision and recall.

Unified BERTScore uses the F1 score in our experimental results

COMET The third evaluation method we utilized was COMET. COMET (Crosslingual Optimized Metric for Evaluation of Translation) is a machine learning-based framework for the automatic evaluation of machine translation. It was trained on a large-scale dataset of human translation rankings, making it capable of capturing more nuanced and complex aspects of translation quality compared to simpler metrics like BLEU. COMET leverages transformer-based models to understand the context and semantics of the text, providing a more holistic evaluation.

We selected the Unbabel/wmt22-comet-da model for the COMET evaluation. In the scoring process, we provided three documents to the model:

1. The source document, which was in Chinese and had not been translated.
2. The MT result, which comprised the machine-translated sentences from the dataset described in the previous section.

3. The reference (ref) document, which was a professionally translated version of the source document, serves as a high-quality translation reference.

We calculated the average score for each document by summing the scores assigned by the model to every sentence in the document and then dividing by the total number of sentences. This average score represented the overall translation quality of the document according to each evaluation method.

Average Score = (Sum of all sentence scores) / (Total number of sentences)

3.2 Experiment result

As stated previously in methodology, 6 categories of experiments are conducted. The results are displayed in the following tables, which are the system-level scores from the evaluation approaches. The original scoring data (by sentences) are kept in GitHub (stated in the Acknowledgement) for review.

BLEU

#	Ref	Can	Score
a	*1*	*3*	*0.184*
b	*1*	*3.1*	*0.004*
c	*1*	*3.2*	*0.113*
d	*3*	*3.2*	*0.6598*
e	*1*	*1.2*	*0.6704*
f	*1*	*1.3*	*0.004*

Note- The list of dataset #s used in the experiments:
(1) Original standard English texts
(1.2) 4-grams reordered standard English texts
(1.3) Synonyms replaced standard English texts
(3) Original translation texts
(3.1) Shuffled translation texts
(3.2) 4-grams reordered translation texts
The highlighted scores represent cautioned results (with partial errors), and the red scores represent an unacceptable mismatch with the expectations.

From the results, we proved the weakness of BLEU in some aspects. E.g., the result in experiment *a* represents the overall quality of Google Cloud translation API. However, it only reports a very low score because of corpus similarity that cannot represent the exact quality of the translated texts.

Moreover, from the result of experiments *d* and *e*, we observe that BLEU does not correctly catch the loss of meaning in re-ordered texts, in the opposite, it reports a very high score for these unreadable texts.

Meanwhile, BLEU does not correctly react to the synonyms replacement in which it returns a very low score.

BART(Model: facebook/bart-large-cnn)

#	Ref	Can	Score
a	*1*	*3*	*-3.174*
b	*1*	*3.1*	*-7.389*
c	*1*	*3.2*	*-5.298*
d	*3*	*3.2*	*-3.915*
e	*1*	*1.2*	*-3.817*
f	*1*	*1.3*	*-7.016*

Note- The list of dataset #s used in the experiments:
(1) Original standard English texts
(1.2) 4-grams reordered standard English texts
(1.3) Synonyms replaced standard English texts
(3) Original translation texts
(3.1) Shuffled translation texts
(3.2) 4-grams reordered translation texts
The highlighted scores represent cautioned results (with partial errors), and the red scores represent an unacceptable mismatch with the expectations.

Note that BART requires us to select a pre-trained language model, the model `bart-large-cnn` from Facebook is selected in this experiment. The BART model correctly reported a high score for the original translation result.

However, BART still focuses on the comparison of the corpus that does not correctly respond to the 4-gram reordering and synonyms replacement process.

BERT

#	Ref	Can	Score (F1)
a	*1*	*3*	*0.932*
b	*1*	*3.1*	*0.855*
c	*1*	*3.2*	*0.882*
d	*3*	*3.2*	*0.913*
e	*1*	*1.2*	*0.911*
f	*1*	*1.3*	*0.853*

Note- The list of dataset #s used in the experiments:
(1) Original standard English texts
(1.2) 4-grams reordered standard English texts
(1.3) Synonyms replaced standard English texts
(3) Original translation texts
(3.1) Shuffled translation texts
(3.2) 4-grams reordered translation texts
The highlighted scores represent cautioned results (with partial errors), and the red scores represent an unacceptable mismatch with the expectations.

Similar to BART, BERT correctly reported a high score for the original translation result. It still has a bad response to the 4-gram reordered texts and synonyms replacement process. Interestingly, it seems to have better capacity on shuffled/4-gram reordered texts because it correctly rated the two categories a low score at the same level.

COMET (Model: Unbabel/wmt22-comet-da)

#	Ref	Can	Score (F1)
a	*1*	*3*	*0.825*
b	*1*	*3.1*	*0.541*
c	*1*	*3.2*	*0.653*
d	*3*	*3.2*	*0.751*
e	*1*	*1.2*	*0.732*
f	*1*	*1.3*	*0.488*

Note- The list of dataset #s used in the experiments:
(1) Original standard English texts
(1.2) 4-grams reordered standard English texts
(1.3) Synonyms replaced standard English texts
(3) Original translation texts
(3.1) Shuffled translation texts
(3.2) 4-grams reordered translation texts
The highlighted scores represent cautioned results (with partial errors), and the red scores represent an unacceptable mismatch with the expectations.

COMET correctly responded to almost all categories of the experiments; it only gave a wrong score for synonym replacement with a very low score.

As stated previously, we used the model `wmt22-comet-da` from Unbabel AI in the experiment.

4 Conclusion

The capacities of the tested evaluation approaches are summarized in the following table.

	BLEU	BART	BERT	COMET
Capacity to evaluate translation quality	*B*	*A*	*A*	*A*
Capacity to distinguish shuffled corpus	*C*	*B*	*C*	*A*
Capacity to rate replaced synonyms	*C*	*C*	*B*	*C*
Time of calculation	*A*	*C*	*C*	*C*

A-Good, B- Medium, C-Bad

Overall, despite the unfortunate fact that all these models did not show acceptable capacity for replaced synonyms. We suppose that all three novel methods have better performance than BLEU in all capacities we tested. The BLEU had bad performance either on normally evaluating translation quality or distinguishing generated texts with good corpus but incorrect grammar.

BARTScore and BERTScore had good performance on rating translation quality, and BART partially distinguished the generated texts with good corpus but shuffled order (i.e. incorrect grammar).

COMET had good performance on both evaluating translation quality and distinguishing generated texts with incorrect grammar.

However, these new methods all require a device with GPU support. The BLEU might be a practical alternative for teams without GPU access.

4.1 Future work

Despite the evaluation approaches we tested in the project; we also noticed some new methods published recently. One of the competitors is GPTScore, which is a novel evaluation framework that leverages pre-trained models (like GPT) to perform an as-desired evaluation of generated text via natural language instructions.

In experiments across multiple text generation tasks, evaluation aspects, and corresponding datasets, GPTScore has shown superior performance and addresses long-standing challenges in text evaluation—how to achieve customized, multi-faceted evaluation without the need for annotated samples.

It is expected that GPTScore might be a possible solution to deal with the synonym replacement issue witnessed in our experiment.

Acknowledgement

Source of dataset:

https://workshop2014.iwslt.org

Translation API provider:

translated by Google
http://translate.google.com

GitHub repository for replication:

https://github.com/garfieldw33/Comparison-of-Alternative-Evaluation-Approaches-for-Natural-Language-Applications/tree/main

Reference

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. (2020). BERTScore: Evaluating Text Generation with BERT. In Proceedings of ICLR 2020.

Ricardo Rei, Craig Stewart, Ana C Farinha, Alon Lavie. (2020). COMET: A Neural Framework for MT Evaluation. In Proceedings of EMNLP 2020.

Weizhe Yuan, Graham Neubig, Pengfei Liu. (2020). BARTScore: Evaluating Generated Text as Text Generation. In Proceedings of NeurIPS 2021.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, Pengfei Liu. (2023). GPTScore: Evaluate as You Desire. arXiv:2302.04166.

Alexander Clark, Chris Fox and Shalom Lappin(2010). The Handbook of Computational Linguistics and Natural Language Processing. p574

Pawade, Dipti, Avani Sakhapara, Mansi Jain, Neha Jain, and Krushi Gada. (2018) Story scrambler-automatic text generation using word level RNN-LSTM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, 2019.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. arXiv preprint arXiv:1903.08855, 2019a.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R. (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 7871–7880. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.703. URL https://doi.org/10.18653/v1/2020.acl-main.703.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155, 2022.