Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 2, 2023

A Review of Machine Translation Quality Assessment Methods

Wenbo Wang *

School of computing, Central China Normal University, Wuhan, Hubei 430079, China
*** Corresponding author Email:** xiaobo.0412@outlook.com

Abstract: Machine translation quality assessment plays a crucial role in enhancing the performance of machine translation systems. This review aims to survey and outline the current major methods used for assessing machine translation quality, which can be broadly categorized into manual and automatic assessment methods. Upon analyzing the existing literature, it becomes evident that while manual evaluation methods yield high-quality results, they are time-consuming and labor-intensive. On the other hand, automatic evaluation methods are cost-effective and rapid, but their evaluation results do not match the expertise of human evaluators. The objective of this paper is to furnish researchers with a comprehensive overview of machine translation quality assessment methods, enabling them to select appropriate approaches based on their specific experimental requirements. Additionally, we aspire to offer valuable insights and novel perspectives for further advancements in the field of machine translation quality assessment methods.

Keywords: Machine Translation; Automatic Evaluation Metric; Human Evaluation Methods.

1. Introduction

As the demand for translation continues to rise, traditional human translation alone is no longer sufficient, leading to the emergence of machine translation (MT). The evaluation of machine translation quality has played a pivotal role in advancing the development of MT systems. Over time, assessment methods have transitioned from traditional manual assessment to current automatic assessment methods. However, the existing literature on MT quality assessment methods is lacking, necessitating a comprehensive summary of this field.

The purpose of this review is to examine and summarize the currently available methods for assessing the quality of MT, while discussing their respective strengths and weaknesses. The review begins by exploring traditional manual assessment methods and then delves into supervised and unsupervised learning-based assessment methods. By presenting a comprehensive overview, this review aims to assist researchers in selecting the most suitable MT quality assessment methods for their future work. Additionally, this paper will address the challenges associated with MT quality assessment and propose directions for future research, providing researchers with new perspectives and insights to enhance their techniques in this domain.

2. Classification of MT Quality Assessment Methods

MT Quality Assessment (MTQA) methods can be broadly categorized into manual assessment methods and automatic assessment methods. Manual assessment methods involve human assessors, while automatic assessment methods utilize mathematical formulas or deep learning algorithms to quantify and measure translation quality.

Manual evaluation methods are effective in assessing translation quality. Bilingual assessors, with their deep understanding of language, can evaluate translations based on accuracy, fluency, expressive integrity, and the ability to capture nuances in complex texts. However, manual assessment methods have drawbacks. They are time-consuming, labor-intensive, and reliant on finding trustworthy bilingual evaluators. Additionally, subjective differences among evaluators can lead to varying assessment results for the same translation. To enhance reliability, averaging scores from multiple assessors is commonly employed.

To overcome the limitations of manual assessment methods, researchers have turned to automated assessment methods. They employ objective evaluation metrics through mathematical formulas or deep learning algorithms. However, these metrics may not fully capture human understanding of translation. Traditional automatic evaluation metrics, such as BLEU, focus on lexical similarity and often overlook contextual information, resulting in inaccurate evaluations [1]. To address this, researchers have proposed automatic evaluation methods based on transformer-based language models [2]. These methods consider contextual relationships and semantic features, making them closer to human evaluation [3].

3. Manual Assessment Methods

Human evaluation methods are mainly divided into methods based on direct judgement and methods not based on direct judgement [4].

The direct judgement-based assessment method was still used as the main method for evaluating English translations at the MT Society 2022 [5]. The evaluator reads and evaluates the translation output by the system and compares its meaning with an English reference translation manually translated by a human translator. The evaluator is asked to rate the MT result on a scale such as 0-5 or 0-100 based on its accuracy, fluency, and similarity to the reference translation. The advantage of a direct judgement based assessment method is that it is possible to see from the scoring results not only how one translation result is better than another, but also how far that result is better than the other [6].

The non-direct judgement based approach assesses the translation quality by asking the assessor to perform tasks related to the MT results, such as reading comprehension, information retrieval or gap filling [7], and assesses the translation quality based on the results of the tasks. This assessment method can assess the MT results more comprehensively and is more objective than the direct judgement-based assessment method, but it requires more time and manpower costs and has higher requirements for the assessors.

4. Automated Assessment Methodology

So far automatic evaluation methods can be mainly classified into three main categories: methods based on reference translation, methods based on unsupervised learning, and methods based on supervised learning. Reference translation-based methods mainly assess translation quality by comparing MT results with reference translations, and measure the degree of similarity between MT results and reference translations through mathematical formulas [8], etc. Unsupervised learning-based approaches learn the criteria for assessing translation quality from the data itself by using datasets that do not require human labelling to learn models or compute metrics. Supervised learning-based approaches learn models or compute metrics by training with human labelled datasets, which are usually human labelled reference translations.

4.1. Methodology based on Reference Translations

4.1.1. BLEU

The BLEU metric (Bilingual Evaluation Understudy) was proposed by Kishore Papineni et al. BLEU uses an N-gram matching rule to compare the percentage of similarity of n groups of words between MT results and reference translations. 1-gram results tend to represent adequacy, while 2-gram and above matching results represent fluency and text readability. [1] The advantages of this method are: (1) fast calculation speed; (2) low resource languages are also applicable; (3) the calculation method is easy to understand. But it also has disadvantages, in the process of matching the translation result must match the reference translation exactly ignoring the syntactic semantics, and even manual translation is impossible to do this.

4.1.2. METEOR

To address the shortcomings of BLEU, the researchers proposed the METEOR (Metric for Evaluation of Translation with Explicit ORdering) method. METEOR takes into account both accuracy and recall and proposes an automatic evaluation based on the weighted reconciled mean of the single-precision and single-word recall metrics [9]. Meanwhile, it expands the synonym set with knowledge sources such as wordnet, while considering the lexical properties of words. In summary, compared with BLEU and NIST, METEOR considers more semantic and lexical information and provides a more accurate assessment of MT quality.

4.1.3. Other Methods based on Reference Translation

In WMT19, which is also a reference translation-based method, Character n-gram F-score (CHRF) demonstrates a higher correlation with human judgement than BLEU.CHRF is based on character level n-gram matching to calculate and evaluate the quality of MT results. Firstly, the MT results and reference translations are segmented at the character level, and then the n-gram match is calculated. Finally, the precision rate, recall rate and F-score are calculated based on the matching degree. Compared with BLEU, CHRF is more suitable for evaluating some non-standardized texts, such as the spoken language domain.

There are many other methods based on reference translation, which are not listed here. The central idea of the automatic evaluation method based on reference translation is to evaluate the quality of MT by comparing the result of MT with the reference translation, which cannot be used if there is no reference translation.

4.2. Unsupervised Learning Based Approach

YiSi is an automatic evaluation framework that can be applied to different resource languages, which utilises two measurement models and provides the option to fall back to a lexically-based evaluation method such as BLEU when a semantic model of the evaluated language is not available. YiSi was first used in the WMT18 Shared Tasks and performed well and consistently in tests related to human judgement. There are three versions of YiSi: YiSi-0, YiSi-1, and YiSi-2.

YiSi-0 is a resource-free variant of YiSi that uses the longest common character substring accuracy to evaluate the lexical similarity between the MT result and the reference translation. YiSi-1 is a monolingual variant of YiSi that uses an embedding model to evaluate lexical semantic similarity and a semantic role tagger to evaluate structural semantic similarity. YiSi-2 is a cross-linguistic variant of YiSi, which uses a cross-linguistic embedding model to evaluate cross-linguistic lexical semantic similarity, along with a semantic role tagger to evaluate structural semantic similarity.

4.3. Supervised Learning based Approach

4.3.1. BEER

BEER (BEtter Evaluation as Ranking) is a supervised learning-based method for automatically evaluating the quality of MT that can combine a large number of features into a linear model. BEER uses character-level n-grams to compute similarity and alignment trees to compute fluency of input sentences. The model is also trained using labelled data to increase the similarity with human rankings.

4.3.2. COMET

Cross-lingual optimised metric for evaluation of translation (COMET) is a neural network based on pytorch framework for training multilingual MT evaluation models. COMET supports two different architectures: the Estimator model and the Translation Ranking model. The fundamental difference between these two models lies in the training objective, which is to regress the quality scores directly, while the training objective of the latter is to minimise the distance between the "good" MT results and the reference translation. Both models are composed of a cross-lingual encoder and a pooling layer, and COMET has the advantage of being easily adapted to and optimised for different types of human judgements on the quality of MTs [9].

4.3.3. Other Supervised Learning based Methods

With the development of deep learning and neural networks, there are many supervised learning based automatic methods for evaluating the quality of MT such as RUSE, BLEND, BLEURT and so on in addition to BEER and COMET. The main idea of these methods is to assess the quality of MT results by using manually labelled translation

scores or datasets for prediction.

5. Challenge

A large number of researchers have designed experiments to prove that neural network-based methods for automatically assessing the quality of MT perform better than traditional assessment methods. However, the most commonly used evaluation methods nowadays are still the most traditional methods such as BLEU. So the challenge now is to find out why the use of higher performance automated evaluation methods is not widespread and why people prefer to go for traditional automated evaluation metrics. There is also the fact that human evaluation is still used as a benchmark for evaluating translation results for different tasks in WMT22, and how to improve the automatic evaluation method so that it can be aligned with the quality of human evaluation should also be a key area of focus for researchers in the future.

6. Conclusion and Recommendations

Finally, in this paper we review common MT quality assessment methods regarding manual and automatic assessment. Traditional manual evaluation methods evaluate MT results by human experts, which can get high quality evaluation results, but they are time-consuming and costly with a certain degree of subjectivity. The automatic evaluation method evaluates the MT results through mathematical formulae or computer algorithms, which is fast and low-cost, but cannot provide high-quality evaluation results like the manual evaluation method. In order to further develop the MT quality assessment methods, it is suggested to combine the advantages of manual assessment and automatic assessment. It is also suggested that the development of automatic assessment methods should focus on improving neural network-based assessment methods so that such high-performance assessment methods can be used universally.

References

[1] Papineni K., Roukos S., Ward T. and Zhu W.J. (2002). BLEU: A method for automatic evaluation of machine translation. In: Proceedings of ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, pp.311-318.

[2] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. Bert. (2018) Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

[3] Freitag, M., Rei, R., Mathur, N., Lo, C.k., Stewart, C., Foster, G., Lavie, A., Bojar, O. (2021) Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain. In Proceedings of the Sixth Conference on Machine Translation, Online, pp. 733-774.

[4] Chatzikoumi E. (2020) How to evaluate machine translation: a review of automated and human metrics. Natural Language Engineering. 26(2):137-161.

[5] Zerva, C., Blain, F., Rei, R., Lertvittayakumjorn, P., De Souza, J. G., Eger, S., ... & Specia, L. (2022) Findings of the wmt 2022 shared task on quality estimation. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp.69-99.

[6] Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013) Continuous measurement scales in human evaluation of machine translation. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. pp.33-41.

[7] Ageeva E., Tyers F., Forcada M. and Perez-Ortiz J. (2015) Evaluating machine translation for assimilation via a gap-filling task. In: Proceedings of the Conference of the European Association for Machine Translation, Antalya, Turkey, pp. 137-144.

[8] Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., ... & Lim, H. (2023). A survey on evaluation metrics for machine translation. Mathematics, 11(4): 1006.

[9] Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp.65-72.