MZ Journals Computing Journal

The Role of Contextual Embeddings in Improving Zero-Shot Translation Quality Estimation

Anita Mishra
Department of Artificial Intelligence, Tribhuvan University, Nepal

Abstract

Zero-shot translation quality estimation (QE) aims to evaluate the quality of translations without reference translations, a critical task for machine translation (MT) systems, particularly in low-resource settings. Contextual embeddings, generated by advanced language models such as BERT, GPT, and their variants, have shown remarkable performance in various natural language processing (NLP) tasks. This paper explores the role of contextual embeddings in enhancing zero-shot QE by leveraging the rich semantic information encapsulated in these embeddings. We present a comprehensive analysis of different contextual embedding models, their integration into QE frameworks, and their impact on QE performance. Our findings indicate that contextual embeddings significantly improve zero-shot QE accuracy, providing a robust foundation for future research in this domain.

***Keywords*:** Contextual Embeddings, Translation Quality Estimation (QE), Zero-Shot Translation, BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT Pretraining Approach), GPT-3 (Generative Pre-trained Transformer 3), Machine Translation (MT), Feature Extraction, Feature Fusion, Prediction Model, Pearson Correlation Coefficient (PCC), Mean Absolute Error (MAE), Accuracy, Natural Language Processing (NLP).

1. Introduction

Translation quality estimation (QE) is essential for assessing the reliability of machine translation (MT) outputs, especially when reference translations are unavailable. Traditional QE methods often rely on features derived from the source and target texts, but these approaches can struggle with zero-shot scenarios where no labeled QE data exists. The advent of contextual embeddings, produced by models like BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and their derivatives, offers a promising avenue for enhancing QE.

These embeddings capture rich contextual information, enabling more accurate semantic understanding and, consequently, better QE. Translation quality estimation (QE) is essential for assessing the reliability of machine translation (MT) outputs, especially when reference translations are unavailable. Traditional QE methods often rely on features derived from the source and target texts, but these approaches can struggle with zero-shot scenarios where no labeled QE data exists. The advent of contextual embeddings, produced by models like BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and their derivatives, offers a promising avenue for enhancing QE. These embeddings capture rich contextual information, enabling more accurate semantic understanding and, consequently, better QE.

With the advent of advanced language models such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and their variants, there has been a significant leap in capturing rich contextual information in embeddings. These contextual embeddings offer a promising solution to improve zero-shot QE by providing a more sophisticated understanding of language semantics and context. This paper explores the integration of contextual embeddings into QE frameworks, evaluating their impact on QE performance and highlighting their potential to address the challenges inherent in zero-shot translation scenarios.

Translation quality estimation (QE) plays a pivotal role in the deployment of machine translation (MT) systems, particularly in assessing the reliability of translations without reference texts. This task becomes more challenging in zero-shot scenarios, where no parallel training data exists for the specific language pair. Traditional QE methods rely on engineered features, such as fluency and adequacy metrics, derived from both source and target texts. However, these methods often fall short in capturing the intricate semantic relationships and contextual dependencies required for accurate evaluation[1]. The development of contextual embeddings, generated by models like BERT, RoBERTa, and GPT-3, has revolutionized natural language processing (NLP). These models pre-train on extensive corpora to capture rich semantic and syntactic information, which can be fine-tuned for specific tasks. Contextual embeddings thus offer a powerful tool for improving QE by providing deeper language understanding and more accurate quality assessments. This study investigates the application of these advanced embeddings in QE, aiming to enhance performance in zero-shot translation scenarios.

2. Contextual Embeddings

Contextual embeddings have emerged as a groundbreaking advancement in natural language processing (NLP), offering a sophisticated representation of words and phrases in context. Unlike static embeddings such as Word2Vec and GloVe, which assign a single vector to each word regardless of its usage, contextual embeddings dynamically generate vectors based on the surrounding text. This allows them to capture nuanced meanings and relationships, significantly enhancing the understanding of language semantics.

Models like BERT (Bidirectional Encoder Representations from Transformers) and its optimized variants such as RoBERTa, as well as generative models like GPT-3, pre-train on massive text corpora using deep learning techniques to learn these embeddings. During pre-training, these models learn to predict words within a given context, thereby acquiring a deep comprehension of linguistic patterns and context[2]. Fine-tuning these pre-trained models on specific tasks further refines their embeddings to suit particular applications. The rich semantic information encoded in contextual embeddings makes them particularly effective for tasks requiring high levels of language understanding, such as translation quality estimation (QE). By leveraging these embeddings, QE systems can achieve more accurate and reliable assessments, especially in zero-shot scenarios where traditional methods struggle.

Our approach to improving zero-shot translation quality estimation (QE) using contextual embeddings involves several key steps. First, we collect and preprocess data from publicly available QE datasets, such as those provided by the WMT QE shared tasks. These datasets include source sentences, their translations, and human-annotated quality scores, which serve as the ground truth for model training and evaluation. Next, we employ state-of-the-art contextual embedding models, including BERT, RoBERTa, and GPT-3, to generate rich, context-aware representations of both source and translated sentences. These embeddings capture the nuanced semantic relationships essential for accurate quality estimation. We integrate these embeddings into a neural network architecture tailored for QE tasks. This architecture comprises three main components: feature extraction, feature fusion, and a prediction model. In the feature extraction phase, we use the chosen embedding models to generate contextual embeddings for each sentence[3]. During feature fusion, we combine the source and target embeddings using techniques such as concatenation, attention mechanisms, or additional transformer layers to enhance the interaction between the two sets of embeddings. Finally, the prediction model, which can be a feedforward neural network or a more complex transformer-based architecture, is trained to predict quality scores or classify translations based on the fused embeddings. We evaluate the performance of our QE framework using metrics such as Pearson correlation coefficient (PCC), mean absolute error (MAE), and accuracy, demonstrating the significant improvements achieved through the integration of contextual embeddings.

3. Embedding Models

BERT (Bidirectional Encoder Representations from Transformers) represents a significant leap forward in the field of natural language processing (NLP). Unlike traditional models that read text sequentially, BERT employs a bidirectional approach, considering the context from both the left and right of each token simultaneously[4]. This bidirectional training allows BERT to capture deeper and more comprehensive semantic and syntactic relationships within the text. BERT is pre-trained on vast amounts of data through two main tasks: masked language modeling (MLM), where random words in a sentence are masked and the model learns to predict them, and next sentence prediction (NSP), which trains the model to understand the relationship between paired sentences. Once pre-trained, BERT can be fine-tuned on specific tasks, such as translation quality estimation (QE), by adjusting the pre-trained model weights with task-specific labeled data. This fine-tuning process allows BERT to adapt its deep contextual understanding to the nuances of the QE task, improving its ability to predict translation quality accurately. The richness and flexibility of

BERT's contextual embeddings make it a powerful tool for zero-shot QE, providing robust performance improvements over traditional QE methods.

RoBERTa (Robustly Optimized BERT Pre-training Approach) is an enhancement of the BERT model, designed to maximize the performance of pre-trained language representations. Building on the strengths of BERT, RoBERTa implements several key optimizations that significantly improve its ability to understand and generate human language. These improvements include training on more data, using larger batch sizes, and removing the next sentence prediction (NSP) objective, which allows the model to focus solely on masked language modeling (MLM). RoBERTa also leverages dynamic masking, where the masked tokens are selected at random during each epoch rather than being fixed, enhancing the model's learning capability[5]. As a result, RoBERTa captures even richer and more nuanced contextual information compared to BERT.

This makes RoBERTa particularly effective for tasks like translation quality estimation (QE), where understanding the intricate relationships within and between sentences is crucial. By fine-tuning RoBERTa on QE tasks, the model can leverage its robust pre-training to deliver highly accurate quality predictions, even in zero-shot scenarios where no labeled training data for the target language pair is available. The superior performance of RoBERTa's contextual embeddings thus provides a substantial advantage in improving the reliability and accuracy of QE systems.

4. QE Framework

Our translation quality estimation (QE) framework is designed to harness the power of contextual embeddings to improve accuracy, especially in zero-shot scenarios. The framework comprises three primary components: feature extraction, feature fusion, and a prediction model. Initially, we use advanced contextual embedding models such as BERT, RoBERTa, and GPT-3 to generate rich embeddings for both source and target sentences. These embeddings capture the semantic and syntactic nuances necessary for accurate quality assessment. In the feature fusion phase, we integrate the source and target embeddings using techniques like concatenation, attention mechanisms, or additional transformer layers[6]. This integration enhances the interaction between the embeddings, allowing the model to better understand the relationship between the original text and its translation.

The final component is the prediction model, which can be a feedforward neural network or a more sophisticated transformer-based architecture. This model is trained to predict translation quality scores or classify the translations based on the fused embeddings. We evaluate the framework's performance using metrics such as Pearson correlation coefficient (PCC), mean absolute error (MAE), and accuracy, demonstrating the significant improvements in QE achieved through the integration of contextual embeddings. This robust framework thus provides a scalable and effective solution for enhancing translation quality estimation in various language pairs and settings.

In our translation quality estimation (QE) framework, feature extraction is a crucial step that leverages advanced contextual embedding models to generate rich, informative representations of both source and translated sentences. We employ state-of-the-art models such as BERT, RoBERTa, and GPT-3, known for their ability to capture deep semantic and syntactic nuances through pre-training on extensive text corpora. During feature extraction, these models process each sentence to produce contextual embeddings, which are dynamic and sensitive to the surrounding context[7]. For the source sentence, the embedding encapsulates its linguistic structure and meaning, while for the translated sentence, the embedding reflects its fidelity and fluency relative to the source.

By extracting these high-dimensional vectors, we obtain a comprehensive and contextually aware representation of the text, which forms the foundation for subsequent phases in the QE framework. The effectiveness of this step is critical, as the quality and depth of the extracted features directly impact the performance of the overall QE system, enabling more accurate and reliable translation quality assessments.

The prediction model is the final and most crucial component of our translation quality estimation (QE) framework, responsible for interpreting the fused contextual embeddings and producing quality predictions. After the feature extraction and fusion phases, the model receives a comprehensive representation of the source and translated sentences. Depending on the complexity and requirements of the task, the prediction model can range from a straightforward feedforward neural network to a more sophisticated transformer-based architecture. This model is designed to handle either regression tasks, where it predicts continuous quality scores, or classification tasks, where it categorizes translations as acceptable or not. During training, the prediction model learns to map the fused embeddings to the target quality metrics by minimizing the loss between the predicted and actual quality scores. This training process benefits from the rich, context-aware features provided by the contextual embeddings, allowing the model to make nuanced and accurate predictions. Evaluated using metrics like Pearson correlation coefficient (PCC), mean absolute error (MAE), and accuracy, the prediction model demonstrates significant improvements in QE performance, particularly in zero-shot scenarios. Its ability to effectively leverage the deep semantic understanding captured by the embeddings underscores its critical role in enhancing the reliability and precision of translation quality estimation.

5. Feature Fusion

Feature fusion is a pivotal step in our translation quality estimation (QE) framework, where the contextual embeddings of source and target sentences are integrated to form a cohesive representation that captures the relationship between the original text and its translation. This process involves combining the extracted embeddings using advanced techniques such as concatenation, attention mechanisms, or additional transformer layers[8]. Concatenation simply joins the embeddings from both sentences into a single vector, providing a straightforward approach to feature integration.

Attention mechanisms, on the other hand, allow the model to weigh the importance of different parts of the sentences dynamically, enhancing the interaction between the source and target embeddings by focusing on the most relevant contextual information. Transformer layers further refine this integration by leveraging their multi-head attention capabilities to capture complex dependencies and relationships across the entire sentence pair. The fused feature vector, enriched with detailed semantic and syntactic nuances from both sentences, is then fed into the prediction model. This fusion process ensures that the model has a comprehensive and contextually aware input, significantly improving its ability to accurately estimate translation quality. By effectively merging the contextual embeddings, feature fusion plays a critical role in enhancing the robustness and accuracy of our QE framework.

6. Evaluation Metrics

Evaluation metrics are essential for assessing the performance of our translation quality estimation (QE) framework, providing quantitative measures of how well the system predicts translation quality. We use several key metrics to evaluate both regression and classification tasks. The Pearson Correlation Coefficient (PCC) measures the linear relationship between predicted quality scores and human-annotated scores, indicating how well the model's predictions align with actual quality assessments[9]. A higher PCC value signifies a stronger correlation and better performance. Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors by calculating the absolute differences between predicted and actual scores. A lower MAE indicates more accurate predictions. For classification tasks, accuracy is used to determine the proportion of correctly classified translations, reflecting the model's effectiveness in categorizing translations into quality levels. These metrics collectively offer a comprehensive evaluation of the QE framework's performance, enabling us to gauge its effectiveness in zero-shot scenarios and identify areas for further improvement. By leveraging these metrics, we ensure that the QE system provides reliable and precise quality assessments, crucial for enhancing machine translation systems.

The Pearson Correlation Coefficient (PCC) is a critical metric for evaluating the performance of translation quality estimation (QE) systems, measuring the strength and direction of the linear relationship between predicted and actual quality scores. PCC ranges from -1 to 1, where a value of 1 indicates a perfect positive correlation, meaning that as the predicted scores increase, the actual scores also increase proportionally. A value of -1 represents a perfect negative correlation, where increases in predicted scores correspond to decreases in actual scores. A value of 0 suggests no linear relationship between the predictions and the true scores. In the context of QE, a higher PCC indicates that the model's predictions closely align with human-annotated quality assessments, reflecting its ability to accurately estimate translation quality. This metric is particularly valuable for regression tasks, providing insights into how well the QE system captures the nuances of translation quality and its overall reliability. By utilizing PCC, we can effectively gauge the performance of our QE framework and ensure that it delivers meaningful and accurate quality predictions.

Mean Absolute Error (MAE) is a key metric for assessing the accuracy of translation quality estimation (QE) systems, quantifying the average magnitude of errors between predicted and actual quality scores. MAE is calculated by taking the average of the absolute differences between each predicted score and its corresponding true score. Unlike metrics that focus on relative or percentage-based errors, MAE provides a direct measure of prediction accuracy in the same units as the quality scores, making it easily interpretable[10]. A lower MAE indicates that the QE model's predictions are closer to the actual quality assessments, reflecting better performance. MAE is particularly useful in evaluating regression tasks where exact score values are critical, as it highlights the magnitude of prediction errors without being affected by the direction of those errors. By employing MAE, we can effectively measure and compare the precision of different QE systems, ensuring that our framework delivers reliable and accurate quality evaluations for machine translation outputs.

7. Accuracy

Accuracy is a fundamental metric for evaluating the performance of translation quality estimation (QE) systems, particularly in classification tasks where translations are categorized into distinct quality levels. It measures the proportion of correctly classified translations out of the total number of translations assessed. Specifically, accuracy is calculated as the ratio of the number of correct predictions (true positives and true negatives) to the total number of predictions, expressed as a percentage.

High accuracy indicates that the QE system effectively distinguishes between different quality categories, providing reliable assessments of translation quality[11]. This metric is crucial in scenarios where translations need to be categorized into quality classes, such as acceptable or not acceptable. By assessing accuracy, we gain insight into the model's effectiveness in classifying translations accurately, which is essential for guiding post-editing efforts and ensuring the overall quality of machine translation outputs.

8. Experiments and Results

In our experiments, we evaluate the effectiveness of contextual embeddings in enhancing zero-shot translation quality estimation (QE) using several state-of-the-art models, including BERT, RoBERTa, and GPT-3. We conduct experiments on multiple publicly available QE datasets, such as those from the WMT QE shared tasks, to ensure comprehensive evaluation across different language pairs and scenarios. Each model is fine-tuned on the QE task using these datasets, and we compare their performance using metrics such as Pearson Correlation Coefficient (PCC), Mean Absolute Error (MAE), and accuracy. Our results reveal that contextual embeddings significantly improve QE performance, with RoBERTa achieving the highest scores in both PCC and MAE, indicating superior ability in capturing translation quality nuances. BERT also shows notable improvements over traditional QE methods, while GPT-3 performs well but with higher computational costs[12]. These findings demonstrate that integrating contextual embeddings into QE frameworks enhances the accuracy and reliability of quality predictions, particularly in zero-shot scenarios where traditional approaches fall short. The experimental results underscore the potential of advanced contextual models in advancing translation quality estimation and provide a robust foundation for further research and development in this field.

This results in a more robust and nuanced understanding of text, as reflected in superior performance metrics, including higher Pearson Correlation Coefficient (PCC) and lower Mean Absolute Error (MAE). GPT-3, with its generative capabilities and extensive training on diverse data, also shows strong performance, particularly in capturing complex linguistic patterns. However, its larger size and computational requirements make it less practical for some applications compared to BERT and RoBERTa. Overall, while all models provide notable improvements in QE, RoBERTa stands out as the most effective due to its optimized architecture and training, offering the best balance of performance and efficiency[13].

The integration of contextual embeddings has a profound impact on translation quality estimation (QE), transforming the accuracy and reliability of quality assessments. Unlike traditional methods that rely on static or surface-level features, contextual embeddings provide a rich, dynamic representation of text, capturing the nuanced semantic and syntactic information necessary for precise evaluation. Models such as BERT, RoBERTa, and GPT-3 excel at understanding the context in which words and phrases appear, allowing them to more effectively assess the quality of translations by recognizing subtle differences and complex relationships between source and target texts[14]. This enhanced contextual understanding significantly improves QE performance, particularly in zero-shot scenarios where no direct training data is available. The ability of contextual embeddings to capture intricate language patterns and dependencies leads to more accurate predictions of translation quality, providing valuable insights for improving machine translation systems and guiding post-editing efforts. Overall, the impact of contextual embeddings represents a major advancement in QE, offering a powerful tool for addressing the challenges of translation assessment and driving progress in the field of machine translation[15].

9. Conclusion

In conclusion, the integration of contextual embeddings into translation quality estimation (QE) frameworks represents a significant advancement in evaluating machine translation outputs, especially in zero-shot scenarios. Our study demonstrates that models such as BERT, RoBERTa, and GPT-3 significantly enhance QE performance by providing deeper semantic understanding and capturing intricate contextual relationships between source and translated sentences. Among these models, RoBERTa shows the most promising results, achieving superior accuracy and reliability in predicting translation quality due to its optimized training and robust embeddings. These findings highlight the effectiveness of leveraging advanced contextual models to overcome the limitations of traditional QE methods, offering more accurate and reliable quality assessments. As machine translation continues to evolve, the insights gained from this research pave the way for further innovations in QE, promising more effective solutions for assessing translation quality in diverse and low-resource settings. Future work can build on these results by exploring additional contextual signals and refining embedding models to further enhance QE performance.

References

[1] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Panda: Prompt transfer meets knowledge distillation for efficient model adaptation," *IEEE Transactions on Knowledge and Data Engineering,* 2024.

[2] L. Zhou, L. Ding, and K. Takeda, "Zero-shot translation quality estimation with explicit cross-lingual patterns," *arXiv preprint arXiv:2010.04989,* 2020.

[3] K. Peng *et al.*, "Towards making the most of chatgpt for machine translation," *arXiv preprint arXiv:2303.13780,* 2023.

[4] L. Ding, L. Wang, X. Liu, D. F. Wong, D. Tao, and Z. Tu, "Progressive multi-granularity training for non-autoregressive translation," *arXiv preprint arXiv:2106.05546,* 2021.

[5] C. Zan, L. Ding, L. Shen, Y. Cao, W. Liu, and D. Tao, "On the complementarity between pre-training and random-initialization for resource-rich machine translation," *arXiv preprint arXiv:2209.03316,* 2022.

[6] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert," *arXiv preprint arXiv:2302.10198,* 2023.

[7] L. Ding, D. Wu, and D. Tao, "The USYD-JD Speech Translation System for IWSLT 2021," *arXiv preprint arXiv:2107.11572,* 2021.

[8] W. Huang *et al.*, "The Xiaomi AI Lab's Speech Translation Systems for IWSLT 2023 Offline Task, Simultaneous Task and Speech-to-Speech Task," in *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, 2023, pp. 411-419.

[9] M. Federico, A. Waibel, M. R. Costa-jussà, J. Niehues, S. Stüker, and E. Salesky, "Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)," in *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, 2021.

[10] Y. Xiao *et al.*, "A survey on non-autoregressive generation for neural machine translation and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11407-11427, 2023.

[11] N.-Q. Pham, J. Niehues, T.-L. Ha, and A. Waibel, "Improving zero-shot translation with language-independent constraints," *arXiv preprint arXiv:1906.08584,* 2019.

[12] T. Schuster, O. Ram, R. Barzilay, and A. Globerson, "Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing," *arXiv preprint arXiv:1902.09492,* 2019.

[13] B. Thompson and M. Post, "Automatic machine translation evaluation in many languages via zero-shot paraphrasing," *arXiv preprint arXiv:2004.14564,* 2020.

[14] H. Xu and P. Koehn, "Zero-Shot Cross-Lingual Dependency Parsing through Contextual Embedding Transformation," *arXiv preprint arXiv:2103.02212*, 2021.

[15] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the association for computational linguistics*, vol. 7, pp. 597-610, 2019.