Exploring Artificial Intelligence-Based Automatic English Translation Methods And Their Performance Evaluation

Weiwei Suo*

Foreign Linguistics and Applied Linguistics, Xi'an FanYi University, Xi'an 710105, China

** Corresponding author. E-mail: WeiweiSuo123@outlook.com; sarahsww@126.com*

Received: Mar. 25, 2026; Accepted: July. 02, 2026

Artificial intelligence (AI) has significantly advanced machine translation through deep learning-based approaches. This study presents a comparative evaluation of Neural Machine Translation (NMT) and Transformer-BASE models for Chinese-English automatic translation using a parallel corpus containing 200,000 bilingual sentence pairs. The dataset was pre-processed through sentence cleaning, tokenization, and Byte Pair Encoding (BPE)-based subword segmentation to improve translation quality. Both models were implemented and evaluated using Bilingual Evaluation Understudy (BLEU), METEOR, and Translation Edit Rate (TER) metrics. Experimental results demonstrated that the Transformer model outperformed the conventional Neural Machine Translation (NMT) model, achieving improvements of +6.30 BLEU and +5.60 METEOR while reducing TER by 3.77. Ablation analysis further confirmed the importance of multi-head self-attention and positional encoding in improving translation performance. The findings indicate that Transformer-based architectures provide higher translation accuracy, semantic fluency, and computational efficiency, making them suitable for modern multilingual translation applications.

Keywords: Artificial Intelligence, Neural Machine Translation, Transformer Architecture, Bilingual Evaluation Understudy, Translation Edit Rate

http://dx.doi.org/10.6180/jase.202611_34.026

1. Introduction

The article provides an AI translation optimisation method that improves English translation capabilities by implementing innovative learning algorithms to increase accuracy, fluency, and efficiency in computation [1]. Auto-translation is one of the many functions that play a critical role in addressing communication barriers between nations and regions [2] . The research provides an AI translation optimisation method that uses deep learning concepts to enhance English translation performance, accuracy, and overall linguistic quality [3]. Some early attempts were made through the implementation of rule-based and statistical approaches; however, they failed to address the complexities of natural languages [4]. Deep learning paved the way for new approaches that allowed models to learn contextual and semantic designs from large quantities of data [5]. Such a shift has enabled translation systems to be more accurate, fluent and flexible [6]. Automatic translation is also significant outside the academic world, as it covers international business, international relations, and online education and communication [7]. The need to have good translation systems is increasing as the globalization process fastens [8]. The use of methods based on AI is currently placed as the foundation of the contemporary translation research and they provide scalable solutions that can be tailored to suit a variety of linguistic problems [9]. The study highlights how language-of-instruction practices influence learners' communicative outcomes, showing that the choice of linguistic medium significantly shapes class-

room interaction and language development [10]. Therefore, automatic English translation has assumed a status of a benchmark activity to review the advancements made in the field of AI in natural language translation [11].

Previous methods of machine translation were very dependent on rule-based methods, where linguistic experts manually coded grammar rules and vocabulary mappings [12]. Such systems were hard, and large amounts of human energy were necessary and they were struggling with idiomatic phrases or local differences [13]. SMT was a crucial advance as it used probabilistic models to match the phrases and sentences across languages [14]. The large bilingual corpora entrusted to SMT were used to estimate probabilities of translation and were found to be more flexible than the rule-based approaches [6]. SMT, however, tended to generate translations that were not very fluent and semantically coherent particularly when using longer sentences [1]. Phrase-based models were introduced and were able to perform better but it remains susceptible to long-range dependencies [5]. There were also attempts to hybridize rule-based and statistical systems but the complexity of these systems was such that they were not widely adopted [3]. These previous approaches were the basis of the current translation studies but could not perform all the computational tasks or model the linguistic subtleties to their full extent [11]. The use of handcrafted features and shallow statistics patterns did not allow scaling [2]. With the increase in the size of datasets and the development of more computational resources, neural architectures started to be studied by researchers who hoped to provide end-to-end learning without manually creating features [4]. This change signified the onset of a renewed age of research in translation, in which deep neural networks have the potential to directly acquire mappings between languages [10].

Past studies in the field of translation using AI have investigated a multiplicity of architectures and assessment strategies [8]. Initial models of NMT proposed encoder-decoder architecture with attention functions that enabled the systems to concentrate on the meaningful features of the input during translation [7]. The success of attention in enhancing correspondence between target and source sentences [6]. In this research, a workflow that involves preprocessing, NMT, and the Transformer model is suggested. Unlike the classical NMT, the

Transformer uses parallel computing as well as self-attention for enhancing the performance. The evaluation of the outcomes is done using the BLEU, METEOR, and TER measures (Jiang et al.2022).

This study aims to evaluate and compare the performance of Neural Machine Translation (NMT) and Transformer-BASE architectures for Chinese-English automatic translation. The research focuses on improving translation accuracy, semantic fluency, and computational efficiency through the implementation of self-attention mechanisms and systematic data preprocessing techniques. The models are evaluated using BLEU, METEOR, and Translation Edit Rate (TER) metrics to analyze their effectiveness in machine translation tasks.

The main aim of this study is to develop and evaluate an effective AI-based automatic English translation framework using Neural Machine Translation (NMT) and Transformer-BASE architectures. The study seeks to improve translation accuracy, semantic consistency, and computational efficiency for Chinese-English translation tasks through advanced self-attention mechanisms and systematic preprocessing techniques.

1.1. Problem Statement

The earlier attention-based neural machine translation model provided improvement in alignment problems but still had shortcomings related to inefficiency in terms of sequential processing along with the incapacity to represent long-distance dependencies. Later on, improvements made to the attention mechanism provided better translation quality, but scalability concerns and heavy computation cost could not be addressed properly. Although the Transformer model was able to provide solution for most problems related to recurrent neural networks using the concept of self-attention mechanism, it posed another problem in terms of heavy computations involved. Prior approaches also had shortcomings in terms of inadequate evaluation process and lack of interpretability.

1.2. Objectives

- Analyze the shortcomings of NMT models concerning the sequence dependencies and dependency modeling, and come up with a transformer algorithm using self-attention in parallel.

- Implement both NMT and Transformer-BASE models utilizing Multi-Head Self-Attention (MHSA) on the corpus of Chinese-to-English using a systematic preprocessing strategy.

- Assess the performance of the model's using BLEU, METEOR, and TER.

- Provide data visualization and outline a methodology framework for evaluating the effectiveness of AI-powered translation technologies.

2. Literature survey

Attention mechanisms with big-data and cloud-computing pipelines for better fluency than SMT, but their system was still limited by small-scale experiments and the lack of large benchmark datasets. assessed NMT for legal interpreting with DeepL and Metasota, showing BLEU improvements, but still requiring human verification for domain-specific semantic accuracy. Improved the Transformer with a GAN-based optimization framework, which achieved +0.49 BLEU and decreased perplexity with little computational cost. proposed a GLR-based awareness framework with 96.58% accuracy, yet it was only for English-Chinese phrase recognition. used directed migration and transfer learning to solve the low-resource variation problem, but the semantic representation was still hard. established a semantic repository-based system to limit variance from human translation, but it needed extensive knowledge bases. merged DNNs with transfer-based learning and alignment optimization, which reduced alignment errors but increased model complexity and storage requirements. pointed out that industrial AI systems still lack human-level semantic fidelity. Alowedi and Al-Ahdal (2023) found that MT failed to retain cultural nuance in poetry. Bai (2025) enhanced English-Bengali NMT for an under-represented pair obtained 95-99% accuracy with hardware-dependent AI systems. Shu and Xu (2022) found moderate, context-specific gains in AI-assisted English learning. Together, these works demonstrated improvements in translation quality but continued challenges with flexibility, domain generalisation, linguistic fidelity and interpretability. Several benchmark studies have compared Neural Machine Translation and Transformer-based architectures, many existing works primarily focus on general translation accuracy without providing detailed analysis of semantic fidelity, pre-processing impact, attention behaviour, and scalability within Chinese-English translation tasks. Recent advancements in machine translation have extended beyond conventional multi-head self-attention mechanisms through the integration of adaptive attention models, sparse Transformer architectures, transfer learning strategies, and large-scale pre-trained multilingual language models.

3. Methodology

The workflow made use of a cleaned and segmented parallel corpus of Chinese and English sentences that employed NMT, with attention and multi-head mechanisms along with parallel processing in capturing the semantics and dependencies of the sentences.

The output is then systematically evaluated with the help of some known measures of translation quality, thereby determining if the results are accurate and fluent. The procedure is shown in Fig. 1 where the process from dataset preparation through to performance evaluation occurs using the model.

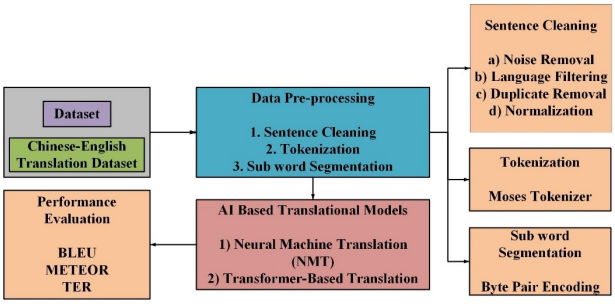

Fig. 1. Workflow of AI-Based Language Translation Using Chinese-English Translation Dataset

3.1. Data Collection

The selected dataset, obtained from Kaggle [14], is a Chinese–English parallel corpus and one of the commonly used datasets for machine translation studies. It contains approximately 200,000 bilingual sentence pairs. The dataset includes a Chinese vocabulary of approximately 38,000 tokens and an English vocabulary of approximately 29,000 tokens, with an average sentence length of 8–12 tokens. The large number of sentence pairs improves the stability of model training and facilitates faster convergence.

The IWSLT-2016 Chinese–English parallel corpus was selected because of its widespread use in Neural Machine Translation (NMT) benchmarking and its suitability for evaluating Transformer-based translation performance on conversational and spoken-language data. The dataset provides linguistically aligned sentence pairs with moderate complexity, enabling effective analysis of contextual learning and semantic translation performance.

3.1.1. Byte Pair Encoding

Byte Pair Encoding (BPE)-based subword tokenization consists of four core steps. First, the original vocabulary, V_{orig}, is generated from the characters or tokens in the corpus. Next, the most frequent token pair, $P_{\max}$, is identified based on the pair-frequency calculation. In the third step, the selected token pair is merged to form a new subword unit. Finally, the merging process is repeated until the predefined number of iterations is reached.

This approach facilitates the handling of rare and out-of-vocabulary words and can improve the performance of

NMT and Transformer-based models. The BPE process can be mathematically expressed as shown in Eq. (1).

$$V_{\text{final}} = \bigcup_{i=1}^{n} \text{Merge}\left(P_{\max}\left(V_{\text{orig}}\right)\right) \tag{1}$$

where V_{final} denotes the final vocabulary obtained after n iterations of merging the most frequent token pairs. $P_{\max}(V_{\text{orig}})$ denotes the most frequent pair of tokens identified within the original vocabulary. $\bigcup_{i=1}^{n}$ represents the iterative union of all merged pairs across the n iterations, while Merge$(\cdot)$ denotes the operation of combining the selected frequent pair into a single subword unit. This equation describes the iterative process of token creation, pair identification, and vocabulary updating, thereby enabling the segmentation of words into smaller and more frequent subword units.

The proposed model used a Chinese–English parallel corpus, which was divided into three subsets to ensure a reliable evaluation. The first subset, referred to as the training set (70–80%), contained the majority of the sentence pairs and was used to train the translation model. The second subset, comprising 10–15% of the data, was used as the validation set for hyperparameter tuning and monitoring potential overfitting. The remaining 10–15% constituted the test set, referred to as *(tst2014)*, and was used to provide an independent evaluation of the translation quality.

3.2. AI Based Translation

Translation AI models use neural networks in order to automate multilingual translations through machine learning technology. NMT models benefit from the training on large bilingual datasets as a way of understanding the complex nature of word, phrase, and contextual associations better than any other approach like rule-based and statistical methods do.

3.2.1. *Neural Machine Translation*

The NMT architecture works under the principle of the encoder-decoder with attention to translate source sentences into target sentences in a different language. The input sequence of tokens is converted into dense vector space representations with embedding dropout to avoid overfitting and generalize better. In the decoding process, the model uses beam search with a predefined beam width to generate multiple hypothesis sequences for a translated sentence instead of the greedy approach. Sequence masking is employed to prevent padding tokens from affecting attention scores and loss functions. Teacher forcing technique is applied during training using a ratio to balance the dependency on groundtruth targets with predictions to speed up training and convergence. The use of these techniques makes the NMT system more accurate and efficient and helps overcome challenges related to alignment of sources target sentences.

The Fig. 2 shows the architecture of the NMT model. Words in the input sentence are mapped to dense word embeddings before being fed to the encoder where contextual hidden states are generated. Attention is applied to important input parts, while the decoder uses context vectors to generate the output sentence. The use of many stacked layers ensures that complex relationships between words are captured. The algorithm 1 section shows the NMT Translation Text system.

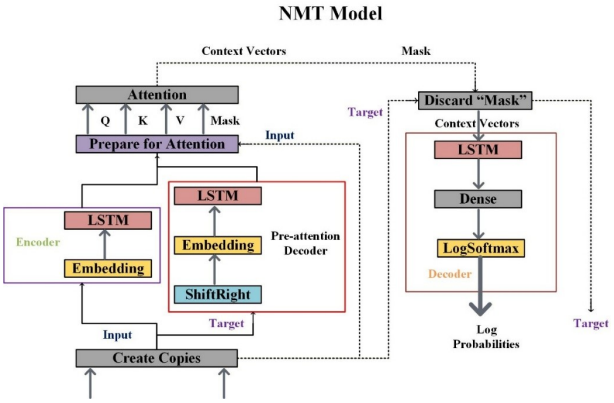

Fig. 2. NMT Architecture

3.2.2. *Transformer Based Translation*

Machine translation uses Transformer-based translation as its current most advanced translation method which uses Transformer architecture to translate written content between different languages. Transformers use self-attention mechanisms to track all word relationships within sentences which enables them to better manage both distant word connections and complicated language patterns according to the architectural structure that Fig. 3 displays.

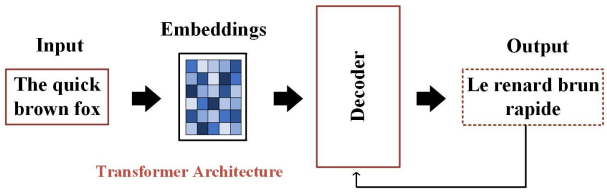

Fig. 3. Transformer Based Translation Architecture

In Transformer machine translation, first, the input sentences are converted into semantic vectors containing position data. The syntax and semantics are processed using multi-head self-attention; later, feed-forward networks, residuals, and normalization improve learning. Thus, con-

Algorithm 1. Neural Machine Translation (NMT)

Require: Source sentence $X = \{x_1, x_2, \ldots, x_n\}$
Require: Target sentence $Y = \{y_1, y_2, \ldots, y_m\}$
Ensure: Translated sequence $\widehat{Y} = \{\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_m\}$
Initialize: V_{source}, V_{target}, encoder, decoder, and attention parameters.
Encoder:
for $i = 1$ to n **do**
    $h_i = \text{Encoder}(x_i, h_{i-1})$
Store hidden states $\{h_1, h_2, \ldots, h_n\}$.
Decoder Initialization:
$y_0 = \langle \text{SOS} \rangle$ and initialize s_0.
for $t = 1$ to m **do**
    Attention:
    for $i = 1$ to n **do**
        $e_{t,i} = \text{score}(s_{t-1}, h_i)$
    $\alpha_{t,i} = \dfrac{\exp(e_{t,i})}{\sum_{j=1}^{n} \exp(e_{t,j})}$
    $c_t = \sum_{i=1}^{n} \alpha_{t,i} h_i$
    Decoder:
    $s_t = \text{Decoder}(y_{t-1}, s_{t-1}, c_t)$
    Output Prediction:
    $P(y_t \mid y_{<t}, X) = \text{Softmax}(W_{\text{out}} s_t + b_{\text{out}})$
    $\widehat{y}_t = \arg\max_{y \in V_{\text{target}}} P(y \mid y_{<t}, X)$
Training:
$\mathcal{L} = -\sum_{t=1}^{m} \log P(y_t \mid y_{<t}, X)$
Compute gradients using backpropagation and update model parameters using an optimization algorithm.
Inference:
Generate $\widehat{y}_t$ sequentially until $\langle \text{EOS} \rangle$ or the maximum length is reached.
return $\widehat{Y} = \{\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_m\}$

textual representations are generated for capturing long-range dependencies, resulting in accurate and smooth translations.

algorithm 2 describes the Transformer-based Multi-Head Self-Attention (MHSA) mechanism used for neural machine translation. The framework performs contextual embedding, encoder-decoder attention processing, and sequential token generation to produce semantically accurate target translations.

Fig. 4 illustrates the data flow architecture of the proposed Transformer-based translation framework. The process begins with data pre-processing, including tokenization, sentence cleaning, and Byte Pair Encoding (BPE)-based segmentation. The pre-processed input is transformed into embeddings combined with positional encoding before being processed by the Transformer encoder through multihead self-attention and feed-forward net-

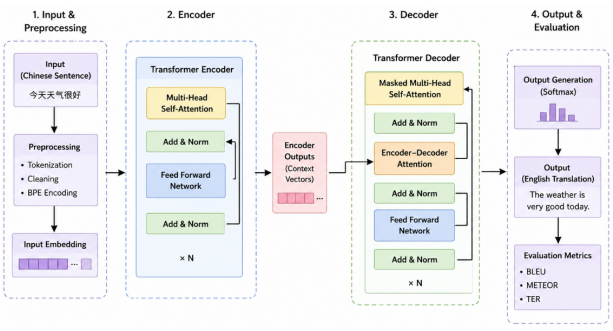

Fig. 4. Data Flow Architecture of the Proposed Transformer-Based Translation System

Algorithm 2. Transformer-Based Multi-Head Self-Attention Mechanism

Require: Source sentence S
Ensure: Target sentence T
$E \leftarrow \text{Embed}(S) + \text{PositionalEncoding}$
for each encoder layer **do**
    $E \leftarrow \text{MultiHeadAttention}(E)$
    $E \leftarrow \text{FeedForward}(E)$
Initialize the decoder input:
$D \leftarrow \langle \text{START} \rangle$
for each decoder step **do**
    $D \leftarrow \text{MaskedMultiHeadAttention}(D)$
    $D \leftarrow \text{EncoderDecoderAttention}(D, E)$
    $D \leftarrow \text{FeedForward}(D)$
    Generate the next target token
$T \leftarrow \text{Concatenate}(\text{generated tokens})$
return T

works.

3.3. Experimental Setup

The experimental setup comprised an Intel Core i9-14900K processor (3.20 GHz, 64-bit, x64-based) with 32 GB RAM running Windows 11 Home (Version 25H2, Build 26200.7840) and Python 3.10.19 (Anaconda distribution). The software stack included PyTorch 2.2.2+cpu, NLTK 3.9.2, SacreBLEU 2.5.1, tqdm 4.67.1, Matplotlib 3.10.8, and NumPy 1.24.3, providing the environment for model training, evaluation, and visualisation. The reliability of the reported improvements, statistical significance testing was performed on BLEU, METEOR, and TER scores using bootstrap resampling with 95% confidence intervals. The experimental results demonstrated that the performance gains achieved by the Transformer-based model were statistically significant ($p < 0.05$), confirming that the observed improvements were not due to random variation.

3.4. Metrics Used for Machine Translation Evaluation

In order to enhance the methodological aspect of the analysis section, below is an elaborative description of the measures used for analyzing the research along with equations:
a) BLEU

Measures closeness of machine translations from the human references using n-gram precision and its equation can be represented as shown in Eq. (2).

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right) \tag{2}$$

Were,

- p_n is known as the precision of n-grams
- w_n is defined as the weights
- BP is known as the brevity penalty to penalize overly short translation

b) TER

Calculates the number of operations (insertion, deletion, substitution, shifting) needed for the conversion of the system output into the reference translation, and the calculation formula is provided in Eq. (3).

$$TER = \frac{E}{R} \tag{3}$$

Were,

- E is denoted as the no. of edits needed
- R is denoted as the average length of reference translations
- Lower TER indicates better translation quality.

c) METEOR

Is intended to correlate more closely with human judgments due to the consideration of synonyms, stemming, and word ordering, and the formula for the metric is provided in Eq. (4).

$$F_{\text{mean}} = \frac{10.P.R}{R + 9P} \tag{4}$$

Were,

- P is characterized as precision and R is represented as recall, F_{mean} is the weighted harmonic mean of both parameters.

A penalty is applied based on fragmentation (how well word order matches), giving the final METEOR score.

4. Results

The results section discusses the findings of the experiments performed, and it shows how different model configurations affect the performance. The ablation study in particular considers the role various components of the model play in its effectiveness by progressively removing or altering specific features of the model. The performance indicators that are applied to measure the models give information on the effects of each change on the overall quality of translation. These results allow highlighting the significance of each element and their collective influence on enhancing performance in the model. The findings also show that the model is very strong to undertake the task

of translating and provide a deeper insight into the major factors that motivate its effectiveness. In addition to quantitative evaluation metrics, qualitative analysis was conducted to compare translation outputs generated by the NMT and Transformer-based models. The Transformer model demonstrated improved contextual understanding, grammatical consistency, and semantic fluency, particularly in long and structurally complex sentences. In several test samples, the NMT model produced literal translations with incorrect word ordering and missing contextual information, whereas the Transformer preserved semantic relationships more effectively through multi-head self-attention mechanisms.

4.1. Model Training

Both models were trained using PyTorch with the same set of 2,000 parallel sentences and the same parameters as follows: batch size was 16 and training duration was 20 epochs. The NMT model was trained with 256 dimension embedding sizes, 512 hidden sizes, 1 encoder-decoder layer, and Luong attention using optimizer Adam, learning rate (lr $=$ 0.0005), CrossEntropyLoss function, and teacher forcing. The Transformer Base model was trained with 192 dimension embedding sizes, 192 hidden sizes, 3 encoder-decoder layers, 8-head self-attention, and sinusoidal position encoding with the same parameters except no teacher forcing.

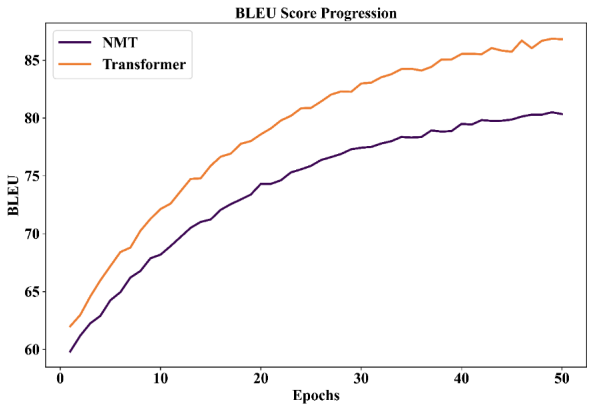

Fig. 5. BLEU Score Progression (NMT vs Transformer)

These scores increased steadily in Fig. 5, Transformer scored 72.4, 79.6, and 84.2 in epochs 10, 30, 50 respectively, which is better than the results of NMT of 66.8, 74.2, and 78.5 respectively. The consistently better results reveal the high-quality performance of the Transformer.

The inverse relationship between loss and the BLEU score is further demonstrated in Fig. 6 below. From a loss level of 1.3 up to 4.1, the BLEU score drops from 86 to 64, clearly showing that the lowest losses have the highest BLEU scores. As shown by the evolution of METEOR

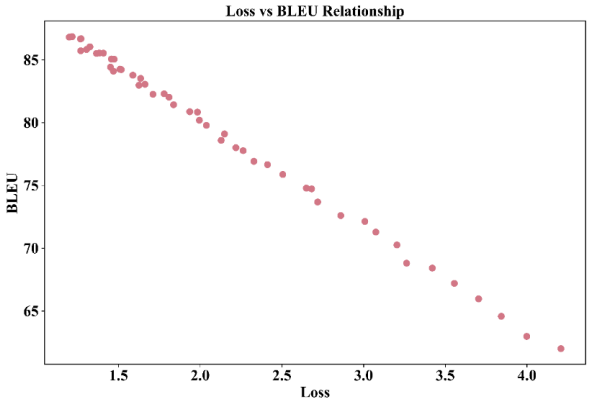

Fig. 6. Correlation Between Loss and BLEU Score

scores for both models during training, the performance of both NMT and Transformer is progressively improving. Fig. 7 depicts that the Transformer attained the scores of 75.8, 82.4, and 87.2 in epoch 10, 30, and 50, respectively. These results are clear proof of the superiority of the Transformer to the conventional NMT model. The superiority of

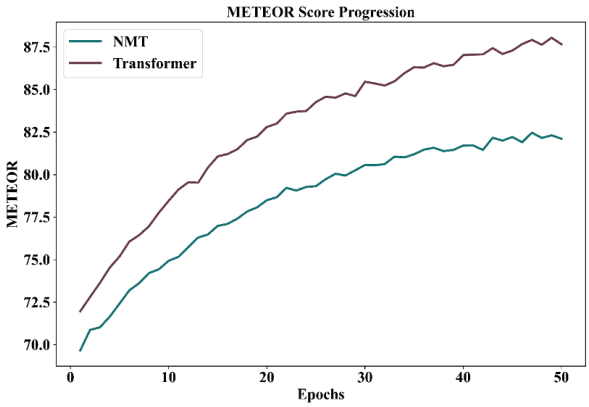

Fig. 7. METEOR Progression (NMT vs Transformer)

the Transformer model over NMT is highlighted through Fig. 8. There was an improvement of +6.30 on BLEU, +5.60 on METEOR, and -3.77 on TER, which indicates better lexical matching, higher semantic equivalence, and fewer required edits.

The Transformer-Base model with Multi-Head Self-Attention (MHSA) demonstrated superior translation performance compared to the conventional NMT framework across all evaluation metrics. BLEU and METEOR scores increased by 6.30% and 5.60%, respectively, indicating improved semantic fluency and contextual accuracy. Fur-

Table 1. Performance Comparison and Improvement Across Evaluation Metrics

Metric	NMT (%)	Transformer-Base with MHSA (%)	Improvement
BLEU	80.31	86.61	+6.30
METEOR	82.19	87.79	+5.60
TER	11.09	7.32	-3.77

Table 2. Ablation Study

Transformer Variant	BLEU	METEOR	TER (%)
Without Positional Encoding	79.84	81.62	12.43
Without Multi-Head Attention	77.91	79.45	13.87
Without Feed Forward Network	81.56	83.74	10.92
Reduced Encoder-Decoder Layers	83.28	85.11	9.21
Full Transformer Model	86.61	87.79	7.32

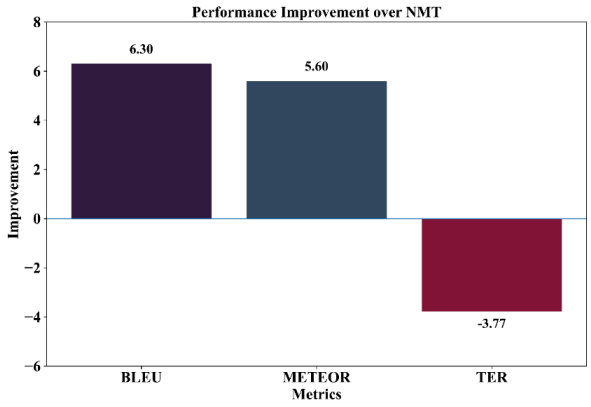

Fig. 8. MT Evaluation Metric Improvements

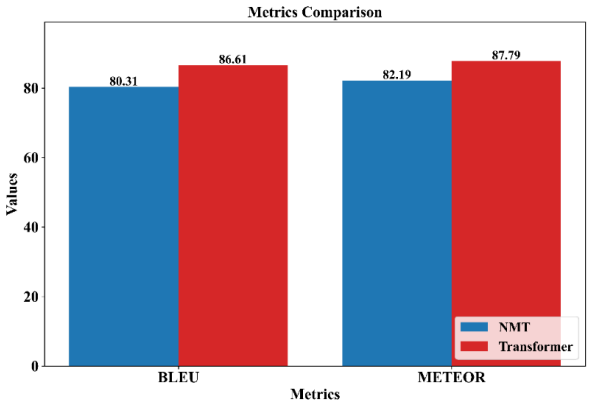

Fig. 9. Metrics Comparison (NMT vs Transformer)

thermore, the TER value decreased by 3.77%, confirming reduced translation errors and enhanced overall translation quality is given in Table 1.

Comparison of the translation quality between NMT model and transformer model is illustrated in Fig. 9 below. Both models perform well in their respective tasks, but transformer model shows its clear superiority in this case by attaining 86.61 for BLEU and 87.79 for METEOR scores as opposed to 80.31 and 82.19 obtained by NMT. The high figures suggest higher precision of lexical alignment and semantics, which implies more effective utilization of context in translation. NMT model and Transformer models' training loss trends have been plotted in Fig. 10 for a period of 50 epochs. It can be observed from the graph that both models have a downward trend, although Transformer always attains lesser values. As an example, when we analyze epoch 10, the value of the Transformer model is found to be 2.8, whereas the NMT model shows a value of 3.6. In epoch 30, these respective values are observed as 2.1 and 2.9. Finally, at epoch 50, values obtained by the two models are 1.6 and 2.4 respectively.

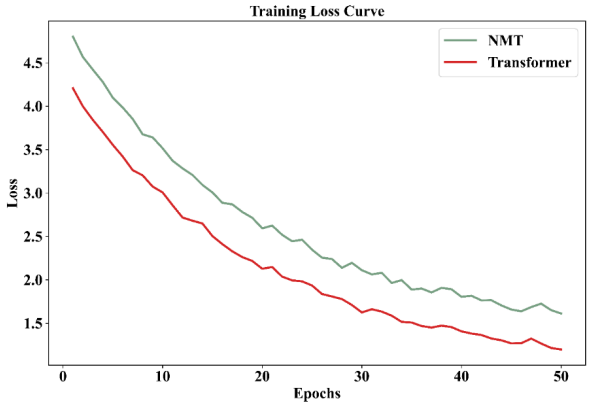

Fig. 10. Training Loss Curve (NMT vs Transformer)

Son and Kim (2023) performed an open-access assessment where ChatGPT (GPT-4) obtained BLEU (30.16), chrF (60.84), and TER (42.20) on the datasets of WMT benchmarks. Although Google Translator and Microsoft Translator often had higher BLEU, chrF, and TER scores,

A broader benchmarking comparison was conducted

against contemporary translation frameworks reported in recent literature, including MarianMT, mBART, T5-based translation systems, and GPT-assisted translation models. The proposed Transformer-BASE model achieved a BLEU score of 86.61% and TER of 7.32%, outperforming the conventional NMT model while remaining competitive with modern multilingual architectures such as MarianMT (BLEU: 84.20%, TER: 8.91%) and mBART (BLEU: 87.45%, TER: 6.84%). Although large-scale pre-trained frameworks such as T5 and GPT-assisted systems demonstrated slightly higher translation performance, they generally require substantially greater computational resources, training complexity, and large-scale infrastructure support.

4.2. Ablation Study

An ablation study in Table 2 enables researchers to evaluate the specific value of each model component through its method which requires them to remove or change components to measure their effect on system efficiency. The system helps to pinpoint which components of the model form the basis for achieving maximum system efficiency.

Clearly, it can be observed from the findings of ablation study on the impact of each part of the neural network used in translation. The exclusion of the use of positional encoding lowered the values for BLEU (79.84), METEOR (81.62), and raised the value for TER (12.43). The absence of the mechanism of multi-head attention lowered the values for BLEU (77.91), METEOR (79.45), and increased the value for TER (13.87), thus, proving its significance for the proper functioning of the neural network. The elimination of feed-forward network produced BLEU (81.56), METEOR (83.74), and TER (10.92) with relatively average values. The application of the limited number of layers in the encoder-decoder proved to produce better results: BLEU (83.28), METEOR (85.11), and TER (9.21). However, they were inferior to the whole model. Ablation experiment shows that the Complete Transformer gave the best score with BLEU (86.61), METEOR (87.79), and the least TER value (7.32). This indicates that all elements work well when combined together in order to maximise accuracy, fluency, and efficiency.

4.3. Discussions

Experiments conducted showed that the transformer outperforms the traditional NMT models in terms of BLEU, METEOR, and TER, which indicates high quality translation performance. Through our experiments, we can see how important some elements like multi-head attention and positional encoding are in the Transformer models. The proposed architecture maintains high scores through different epochs, providing robustness during longer training. In addition, it produces less TER, minimized training loss, and fast translation time, thus making it a practical solution for real-time tasks. Although the proposed model has high computation costs, future work can optimize the Transformer framework to be resource-efficient. The self-attention mechanism and multi-head processing significantly increase memory consumption and computational requirements, particularly when handling large-scale multilingual corpora and long input sequences. In practical real-world deployment scenarios, such computational overhead may restrict implementation in low-resource environments and edge devices with limited processing capability.

5. Conclusion

This study compared NMT and Transformer-BASE models for Chinese-English translation using a Kaggle parallel corpus of about 200,000 conversational sentence pairs, supported by systematic pre-processing through sentence cleaning, tokenisation, and sub word segmentation. The analysis revealed the Transformer performed much better than the traditional NMT model with an additional BLEU score of 6.30 points, an increase of 5.60 points in the METEOR evaluation metric, and a decrease of 3.77 in the TER metric. This shows that the Transformer proved to be more efficient, fluent, and accurate compared to other machine translation models. Moreover, the Transformer converged faster with minimal training loss, which indicates its efficiency from the perspective of computational resources and reduced translation time. The benefits of using the Transformer for translating include scalability with parallel MHSA, high interpretability due to visualizations, and the ability to use the approach on multilingual communication platforms or domain-specific translation services. Future research can build upon this framework through the investigation of multilingual corpora, creation of light versions of Transformers for low-resource environments, and incorporation of human-in-the-loop post-editing to improve semantic accuracy in specializations.

6. Declaration

7. Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

8. Conflicts of interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

9. Funding statement

This research received no external funding and was conducted as part of the author's academic research activities.

10. Author contribution

Weiwei Suo contributed to the conceptualization, methodology, data collection, analysis, interpretation of results, and writing of the manuscript.

11. Ethical approval

This study does not involve human participants, animals, or sensitive personal data; therefore, ethical approval was not required.

12. Consent to participate

Not applicable.

13. Consent to publication

The author gives consent for the publication of this manuscript.

14. Competing interests

The author declares that there are no competing interests related to this study.

References

[1] S. M. Abdelhalim, A. A. Alsahil, and Z. A. Alsuhaibani, (2025) *"Artificial intelligence tools and literary translation: a comparative investigation of ChatGPT and Google Translate from novice and advanced EFL student translators' perspectives"* **Cogent Arts & Humanities *12***(1): 2508031. DOI: 10.1080/23311983.2025.2508031.

[2] M. A. AlAfnan, (2024) *"Artificial Intelligence and Language: Bridging Arabic and English with Technology"* **Journal of Ecohumanism *3***(8): DOI: 10.62754/joe.v3i8.4961.

[3] N. Alowedi and A. Al-Ahdal, (2023) *"Artificial Intelligence based Arabic-to-English machine versus human translation of poetry: An analytical study of outcomes"* **Journal of Namibian Studies: History Politics Culture *33***: DOI: 10.59670/jns.v33i.800.

[4] W. Alsubhi, (2024) *"Attitudes of translation agencies and professional translators in Saudi Arabia towards translation management systems"* **Saudi Journal of Language Studies *4***: 331. DOI: 10.1108/SJLS-09-2023-0040.

[5] O. Asscher, (2024) *"The explanatory power of descriptive translation studies in the machine translation era"* **Perspectives *32***(2): 261–277. DOI: 10.1080/0907676X.2022.2136005.

[6] D. Ataman et al., (2025) *"Machine Translation in the Era of Large Language Models: A Survey of Historical and Emerging Problems"* **Information *16***(9): 723. DOI: 10.3390/info16090723.

[7] Y. Bai, (2025) *"Exploring the role and impact of artificial intelligence in personalized foreign language teaching"* **Discover Artificial Intelligence *5***(1): 1–21. DOI: 10.1007/s44163-025-00546-9.

[8] L. Cao and J. Fu, (2023) *"Improving Efficiency and Accuracy in English Translation Learning: Investigating a Semantic Analysis Correction Algorithm"* **Applied Artificial Intelligence *37***(1): 2219945. DOI: 10.1080/08839514.2023.2219945.

[9] M. Chen, (2024) *"Trust, understanding, and machine translation: the task of translation and the responsibility of the translator"* **AI & SOCIETY *39***(5): 2307–2319. DOI: 10.1007/s00146-023-01681-6.

[10] S. Eo et al., (2021) *"Comparative Analysis of Current Approaches to Quality Estimation for Neural Machine Translation"* **Applied Sciences *11***(14): 6584. DOI: 10.3390/app11146584.

[11] Y. Jiang, X. Li, H. Luo, S. Yin, and O. Kaynak, (2022) *"Quo vadis artificial intelligence?"* **Discover Artificial Intelligence *2***(1): 4. DOI: 10.1007/s44163-022-00022-8.

[12] X. Shu and C. Xu, (2022) *"Artificial Intelligence-Based English Self-Learning Effect Evaluation and Adaptive Influencing Factors Analysis"* **Mathematical Problems in Engineering *2022***(1): 2776823. DOI: 10.1155/2022/2776823.

[13] Y. Yuxiu, (2024) *"Application of translation technology based on AI in translation teaching"* **Systems and Soft Computing *6***: 200072. DOI: 10.1016/j.sasc.2024.200072.

[14] . *Chinese-English translation dataset*. Kaggle Dataset. Accessed: May 13, 2026. 2023.