ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

An Intelligent Comparative Study of Multilingual Corpora of English in ASEAN Driven by Neural Machine Translation

Jianghong Kuang[1], *Haopeng Song*[2,*] *and Yanfu Zhang*[1]

[1] School of Languages, Cultures and International Education, Liuzhou Institute of Technology Liuzhou 545000, Guangxi, China
[2] School of General Education and Pedagogy, Guangzhou Huashang Vocational College, Guangzhou 511300, Guangdong, China

SUMMARY: *In the context of the continuous growth in cross-border digital communication and multilingual information processing demands, the modeling and difference identification of ASEAN English multilingual corpora have become an important topic in intelligent translation research. This paper constructs a technical framework that integrates corpus collection, language identification, sentence-level alignment, subword segmentation, Transformer modeling, and intelligent comparative analysis. It introduces multi-head self-attention, difference scoring function, vector representation learning, and feature fusion mechanisms, and uses joint loss to achieve collaborative optimization of translation generation and difference discrimination. Experiments show that the model in this paper achieves a BLEU score of 39.63, a TER of 0.347, a semantic similarity of 0.879, and a difference identification accuracy of 92.3%; compared with Transformer, BLEU increases by 7.3% and the accuracy improves by 3.6 percentage points. This method can effectively reveal the translation differences between ASEAN English and Thai, Vietnamese, Indonesian, and Malay, and has practical significance for regional English variant computing research and intelligent translation analysis.*

Povzetek: This study focuses on the intelligent comparison of multilingual corpora of ASEAN English, and constructs a neural machine translation framework that integrates Transformer semantic modeling, difference scoring, feature fusion, and joint optimization. Based on experiments with four types of corpora, the model's BLEU score reaches 39.63, TER drops to 0.347, and the accuracy of difference identification reaches 92.3%, which can well reveal the differences in cross-language vocabulary, structure, and semantics.

KEYWORDS: *Neural Machine Translation; ASEAN English;Multilingual Corpus; Intelligent Comparison;Corpus-based Translation Studies*

1 Introduction

In the current context where cross-border digital dissemination is accelerating, intelligent translation services are continuously expanding, and multilingual information processing demands are increasing, conducting intelligent analysis of language corpora related to regional language ecology has gradually become an important direction in natural language processing research [1]. The languages in the ASEAN region are diverse, and the actual usage scenarios of English also have strong regional and situational characteristics. Influenced by national

*haopeng_song@163.com
https://doi.org/10.65102/is2026087
Published: 30 April, 2026

language policies, social cultural environments, and communication habits, ASEAN English in different contexts shows significant differences in vocabulary selection, syntactic arrangement, semantic emphasis, and pragmatic expression [2, 3]. We believe that this difference not only increases the difficulty of cross-language information conversion and machine translation modeling, but also provides a practical observation object for multilingual corpus computing research. Previous translation studies have relied more on manual induction, small-scale sample comparison, and empirical analysis. These methods have certain value in revealing local language phenomena, but when dealing with large-scale corpus processing, cross-language correspondence relationship mining, and dynamic difference identification, they often have problems such as low processing efficiency, subjective analysis, and difficulty in extending the research process, and have become difficult to meet the needs of current intelligent translation research and automatic language comparison analysis [4].

With the rapid development of deep learning methods and pre-trained language models, neural machine translation has demonstrated increasingly prominent advantages in semantic representation, context modeling, and cross-language mapping [5, 6]. Especially after the widespread application of encoder-decoder structures, attention mechanisms, and Transformer architectures, machine translation systems can learn relatively stable semantic alignment relationships from large-scale parallel corpora and have achieved good results in multilingual transfer, low-resource compensation, and translation quality optimization. At the same time, technologies such as text cleaning, sentence alignment, subword segmentation, vector representation, and similarity calculation involved in building multilingual corpora have gradually matured, providing more solid technical support for corpus-based translation research. When reviewing related studies, we found that combining neural machine translation with intelligent corpus comparison not only applies translation models to generate translations, but also enables unified modeling of the structural characteristics of ASEAN English multilingual corpora from a computational perspective, and realizes quantitative identification of difference patterns through automatic evaluation indicators and feature analysis methods, thereby improving the objectivity, repeatability, and interpretability of translation research [7].

However, based on the current research, although there has been an emerging trend of cross-disciplinary interaction among regional English varieties, multilingual translation models, and corpus-based translation studies, the related achievements still have some obvious shortcomings. We have noticed that some studies have insufficient attention to the systematic collection and standardized processing of multilingual corpora in the context of ASEAN English, resulting in a lack of a stable and unified data foundation for model training and subsequent comparative analysis; some studies focus more on the display of translation results, but do not discuss the semantic alignment mechanism, feature extraction process, and intelligent comparison path driven by neural machine translation adequately; and some studies do not closely connect the analysis of language differences with the evaluation of computational models, and have not yet formed a complete technical chain from corpus processing, model construction to translation validation [9, 10]. Based on the above understanding, we aim to address the task of intelligent comparison of multilingual corpora in ASEAN English, and attempt to construct a research framework that integrates corpus collection, alignment preprocessing, neural machine translation modeling, and difference feature analysis, in order to explore the computational realization path for the study of regional English varieties.

This study takes the multilingual texts related to ASEAN English as the research object, combining neural machine translation technology and intelligent comparison methods, and focuses on discussing the strategies for collecting and constructing multilingual corpora as well as the pre-processing of alignment, the design ideas of the intelligent comparison model for the corpora, and the correlation between the translation output quality and the differences of the

corpora. We hope that through this research, the automation level of processing multilingual corpora related to ASEAN English can be improved, the accuracy and interpretability of difference identification driven by machine translation can be enhanced, and a more computer technology-supported analytical framework for translation research based on corpora can be provided.

2 ASEAN English Multilingual Corpus Intelligent Comparison Research Foundation

2.1 Language Characteristics and Research Value of ASEAN English Multilingual Corpora

The multilingual corpora of ASEAN English have strong regional attributes and also exhibit distinct cross-language contact features and continuous evolution characteristics. In the social communication, government affairs exchanges, educational applications, and digital media practices of ASEAN countries, English is often not a standalone language system but rather coexists and influences various languages such as Malay, Thai, Vietnamese, Indonesian, Filipino, and Chinese over a long period of time. During this process, phenomena such as lexical borrowing, word order migration, semantic extension, and pragmatic adjustment frequently occur. We believe that it is precisely through this continuous contact and constant reorganization that ASEAN English gradually formed regionalized expression characteristics distinct from traditional British and American English, and made the multilingual corpora exhibit stronger heterogeneity and more complex correspondences. From the phonetic perspective, ASEAN English has fixed features such as the voicing of voiceless consonants, the simplification of dental sounds, and the weakening of diphthongs. It often adds tone particles like "lah" and "ah" at the end of sentences, forming a unified regional phonetic pattern (Zhang Hailin, 2014). At the lexical level, this variant systematically absorbs local language vocabulary such as Malay and Chinese dialects, for example, borrowed words like "sarong" and "kampong", while some English words have developed exclusive regional semantics, and the flexible use of word classes has also formed a fixed pattern, with clear lexical formation rules. In terms of syntax, ASEAN English shows a simplified tense, flexible use of articles, and a topic-prioritizing sentence structure. Relying on the "least effort principle" of language acquisition, it has formed stable syntactic norms rather than random grammatical mistakes.[8] For the neural machine translation task, this complexity undoubtedly increases the difficulty of cross-language representation learning and semantic alignment, but it also provides more abundant and valuable analysis objects for intelligent comparative research.

From the perspective of specific language manifestations, the multilingual corpora of ASEAN English usually have more localized vocabulary, frequent mixed expressions, obvious syntactic compression, and strong context dependence. When some local cultural concepts enter English expressions, they gradually form relatively stable but not yet fully standardized lexical forms; influenced by the migration of the mother tongue, some texts will also adopt different expression styles from conventional English writing in terms of modifier order, tense usage, and subject-predicate structure arrangement. At the same time, the cross-platform communication environment further strengthens colloquial, omitting, and abbreviation expressions, making the processing difficulty of corpus cleaning, segmentation, and sentence alignment significantly increase. Relying solely on manual reading and experience induction, it is often difficult to continuously and stably identify these differences under large-scale data conditions. In contrast, by using computational methods such as corpus encoding, feature extraction, sentence alignment, and semantic similarity calculation, we can conduct more

systematic and repeatable comparative analyses of multilingual corpora within a unified framework.

To measure the significance of a certain language feature in multilingual corpora, this paper defines the feature strength as:

$$F_i = \frac{n_i}{N} \times \log(1 + m_i) \quad (1)$$

Among them, F_i represents the intensity value of the i-th language feature, n_i indicates the occurrence frequency of this feature in the target corpus, N represents the total number of words in the corpus, and mi represents the cross-lingual mapping times of this feature in the multilingual aligned corpus. This formula takes into account both the surface frequency of the feature and the cross-lingual correlation degree, and can provide quantitative basis for feature selection, difficulty location, and difference weight allocation in the subsequent neural machine translation modeling.

From the perspective of research value, the multilingual corpus of ASEAN English is not only an important window for observing the formation mechanism of regional English variants, but also a key resource for verifying the cross-lingual transfer ability of neural machine translation and the effectiveness of intelligent comparison algorithms. On the one hand, multilingual parallel or quasi-parallel corpora can support the encoder's joint learning of regional variant words, non-typical syntax, and semantic drift phenomena, improving the model's ability to recognize low-frequency expressions and mixed structures; on the other hand, the difference comparison based on the corpus library can reveal the expression patterns of ASEAN English under different language contact conditions through word vector distribution, sentence vector similarity, translation deviation statistics, and clustering results analysis. Thus, this type of corpus not only has obvious linguistic research significance but also has high natural language processing application value, and is an important data foundation connecting translation research and intelligent computing. The main language features and research value of the ASEAN English multilingual corpus are shown in Table 1.

Table 1: Main Language Features and Research Value of ASEAN English Multilingual Corpus

Language Feature Type	Concrete Manifestation	Computational Processing Difficulties	Technological Approaches Available	Research Value
Vocabulary Variants	More local loanwords, mixed words, and abbreviated expressions	High word table discreteness, difficult standardization	Subword segmentation, dynamic word table expansion, word vector modeling	Enhance regional vocabulary recognition and cross-lingual mapping ability
Syntactic Variation	Modification order changes, omission structures, and non-typical combinations frequently	Unstable boundary identification and structure alignment	Dependency analysis, Transformer encoding, syntactic enhanced representation	Support complex sentence translation and structure difference analysis
Semantic Drift	The same expression has different meanings in different national contexts	High semantic alignment error	Context representation learning, attention mechanism, semantic similarity calculation	Improve semantic layer translation quality and comparison accuracy
Pragmatic Differences	Politeness strategies, administrative expressions, and media styles are significantly different	Implicit semantics are difficult to model explicitly	Multi-task learning, paragraph-level modeling, context feature fusion	Support translation theory interpretation and cross-cultural communication analysis
Textual Mix	English and local languages alternate, code	Complex word segmentation, sentence	Language recognition, sequence labeling, mixed	Enhance complex corpus processing and system

	conversion is frequent	segmentation, and alignment	corpus cleaning	adaptability

2.2 Theoretical Support for Neural Machine Translation and Intelligent Comparison Technology

Neural machine translation is an important technical path for multilingual text processing at present. Its core idea is to complete source language representation learning, cross-lingual semantic mapping, and target language sequence generation within a unified neural network framework. Compared with traditional translation methods based on rules or statistics, neural machine translation can use the end-to-end training mechanism to model the context-dependent relationships as a whole, and is more conducive to capturing vocabulary variants, non-typical syntactic structures, and implicit semantic relationships when dealing with complex texts such as ASEAN English, which is strongly influenced by regional language contact. Especially the combination of the encoder-decoder structure and the attention mechanism gives the model significant advantages in long-distance dependency preservation, local semantic alignment, and multilingual transfer learning, laying the foundation for subsequent intelligent comparison of corpora.

In the neural machine translation framework, given the source language sentence $X = (x_1, x_2, \ldots, x_n)$ and the target language sentence $Y = (y_1, y_2, \ldots, y_m)$, the model's goal is to maximize the conditional probability of the target sequence, and its expression is:

$$P(Y|X) = \prod_{t=1}^{m} P\left(y_t | y < t, X\right) \tag{2}$$

Among them, y < t indicates the target word sequence that has been generated before the t-th moment. This equation indicates that the generation of the target language is not an isolated prediction, but is based on the joint condition of the overall semantic representation of the source language and the historical output of the target end. For the multilingual English corpus of ASEAN, this modeling method can better adapt to the context disturbances caused by cross-language expression differences and improve the model's semantic retention ability for regional English varieties.

On this basis, the intelligent comparison technology further undertakes the functions of multilingual difference identification, translation result analysis, and feature quantification of the corpus. Its essence is to map different language texts to a unified vector space, and through semantic similarity, distance measurement, and clustering analysis, etc., to automatically discriminate the corresponding degree of cross-language expressions. At the vector representation level, the source sentence vector is denoted as h_s, and the target sentence vector is denoted as ht, then the semantic similarity between the two can be expressed as:

$$\mathrm{Sim}(h_s, h_t) = \frac{h_s \cdot h_t}{\|h_s\| \|h_t\|} \tag{3}$$

Among them, $\mathrm{Sim}(h_s, h_t)$ represents the cosine similarity value, with a higher value indicating a stronger correspondence between the two sentences in the semantic space. Through this metric, quantitative analysis can be conducted at the computational level to assess the semantic alignment quality, translation deviation, and expression differences of the multi-language corpora of ASEAN English. This avoids the subjective bias that may arise from relying solely on manual judgment.

The relationship between neural machine translation and intelligent comparison is not a

simple superimposition but rather a collaborative mechanism combining generation and analysis. The former is responsible for constructing cross-language conversion channels and outputting translation results with context-dependent features; the latter is responsible for vectorizing comparison, structuring evaluation, and explaining differences of the translation results and the original corpus. After combining the two, not only can the automation level of processing the multi-language corpora of ASEAN English be improved, but also a quantifiable and reproducible analytical basis can be provided for translation research based on the corpus. Therefore, the technical system centered on neural machine translation and supported by intelligent comparison constitutes the main theoretical basis for conducting research on the multi-language corpora of ASEAN English in this paper.

3 Neural Machine Translation-driven Intelligent Comparative Method for ASEAN English Multilingual Corpora

3.1 Collection, Construction, and Alignment Preprocessing of ASEAN English Multilingual Corpora

The quality of ASEAN English multilingual corpora directly affects the effectiveness of subsequent neural machine translation modeling and intelligent comparative analysis. Due to the complex text sources, frequent language contact, and significant differences in expression styles in the ASEAN region, the original corpora usually have problems such as excessive web noise, mixed languages, unclear sentence and paragraph boundaries, and loose cross-language correspondence. Without a systematic data engineering processing procedure, the neural machine translation model is prone to being disturbed by low-quality samples during the training stage, thereby reducing the accuracy of semantic representation and difference identification. Therefore, in the corpus construction stage of this study, a preprocessing mechanism for natural language processing tasks is introduced, forming a technical process of "multi-source collection - quality screening - language identification - sentence-level alignment - encoding representation". The collection, construction, and alignment preprocessing process of ASEAN English multilingual corpora is shown in Figure 1.

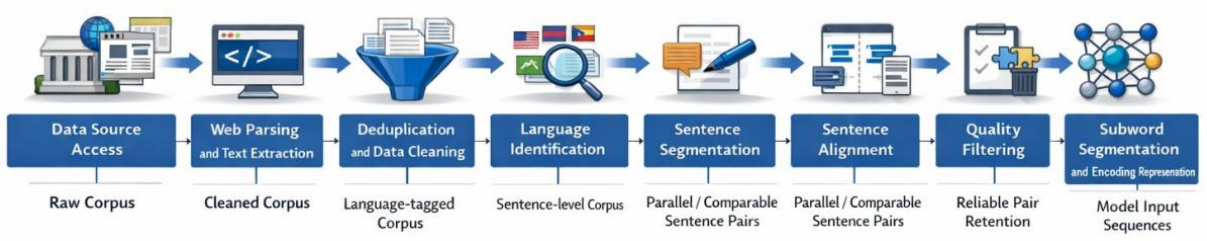

Figure 1: Process for collecting and aligning multilingual corpora of ASEAN English and other languages

(1) Multi-source corpus collection and document-level quality screening

To ensure the coverage and regional representativeness of the corpora, this paper collects English and related language data from news portals, policy announcements, educational information, public service texts, regional cooperation materials, and multilingual online texts in the ASEAN region. Technically, a Python web crawler framework, HTML parsing tools, and text extraction rules are used to batch obtain web page texts, and metadata fields such as source

tags, time tags, language tags, and topic tags are established. Since ASEAN English texts often co-occur with Malay, Vietnamese, Thai, Indonesian, and Filipino, this paper retains relevant language texts during the corpus collection stage to facilitate the subsequent construction of cross-language parallel or quasi-parallel corpora.

Considering the significant differences in the availability of different source texts, this paper introduces a document quality scoring mechanism to conduct preliminary screening of the original corpora. Let the quality score of the kth document be Q_k, then:

$$Q_k = \lambda_1 \cdot \frac{L_k}{L_{max}} + \lambda_2 \cdot (1 - R_k) + \lambda_3 \cdot C_k \tag{4}$$

where L_k represents the effective length of the document, L_{max} represents the maximum effective length in the current batch of documents, R_k represents the repetition rate, Ck represents the text integrity coefficient, λ_1, λ_2, and λ_3 are weight parameters, and they satisfy $\lambda_1 + \lambda_2 + \lambda_3 = 1$. When Q_k is below the set threshold, the document is determined to be a low-quality sample and is removed. This formula comprehensively considers text length, redundancy degree, and structural integrity, and can complete effective filtering before the data source enters sentence-level processing, thereby reducing the interference of noisy corpora on subsequent training.

(2)Corpus cleaning and language identification preprocessing

After the initial collection, the original corpora still contain advertisement language, navigation characters, HTML residual tags, abnormal spaces, and duplicate paragraphs, etc. For this, this paper adopts a hierarchical cleaning strategy: at the document level, hash fingerprints and near-duplicate matching are used to remove duplicate texts; at the sentence segment level, regular rules are used to delete special symbols, meaningless links, and garbled characters; at the encoding level, the character set, punctuation format, and capitalization form are unified to improve the consistency of subsequent sentence segmentation and word segmentation. For the mixed language phenomenon in the ASEAN English corpora, relying solely on rule-based identification is difficult to stably distinguish the language boundaries. Therefore, this paper combines a statistical model to perform language identification on sentence-level texts.

Let the confidence of a sentence segment s belonging to language l be P(l|s), then it can be expressed as:

$$P(l|s) = \frac{\exp(z_l)}{\sum_{j=1}^{M} \exp(z_j)} \tag{5}$$

Among them, z_l represents the classification score of sentence segment s in language *l*, and M represents the total number of candidate languages. This formula essentially uses Softmax normalization to convert the multilingual discrimination results into a probability distribution, thereby achieving the automatic distinction of English, Malay, Vietnamese, Thai, and other text. When the maximum confidence is lower than the threshold, this sentence segment is marked as a mixed or low-confidence text, and its participation priority is reduced in the subsequent alignment stage. Through this process, the problem of language drift caused by frequent code conversion and dense local loanwords in ASEAN English corpora can be effectively solved. The design of pre-processing and alignment rules for ASEAN English multilingual corpora is shown in Table 2.

Table 2: Design of Pre-processing and Alignment Rules for ASEAN English Multilingual Corpora

Processing Step	Primary Task	Technical Method	Output Result
Data Collection	Scraping of ASEAN English and related language texts	Web scraping, API access, text extraction	Original corpus
Data Cleaning	Deleting noise, duplicates, and invalid paragraphs	Hashing de-duplication, regular filtering, format standardization	Cleaned corpus
Language Identification	Determining the language of sentence segments	langid/fastText, rule verification	Language annotation corpus
Sentence Segmentation	Constructing sentence-level processing units	Sentence splitting, rule correction	Sentence-level corpus
Sentence Alignment	Constructing cross-language sentence relationships	Length constraints, vector similarity matching	Parallel/parallel-like sentence pairs
Subword Segmentation	Relieving the influence of low-frequency words and spelling variants	BPE, SentencePiece	Sequence input to model

Table 2 indicates that the processing of ASEAN English multilingual corpora is not a simple text organization process, but a systematic engineering that covers quality control, language discrimination, sentence decomposition and representation standardization. Only by establishing stable data pre-processing rules can the subsequent training of neural machine translation models and intelligent comparative analysis have a unified input.

(3)Sentence-level Alignment and Cross-language Matching Strategies

In the construction of multilingual corpora, sentence-level alignment is a key link connecting the pre-processing stage and the model modeling stage. Traditional alignment methods based on sentence length ratio or document position are prone to errors due to free translation, sentence order changes, and omission structures in the ASEAN English scenario, resulting in a high error rate. To improve the accuracy of sentence alignment, this paper adopts a joint scoring mechanism of "semantic similarity + length constraint + translation consistency". Let the English sentence vector be h_i^{en}, and the target sentence vector be h_j^{tg}, then the comprehensive alignment score of sentence pair (i, j) is defined as:

$$S_{ij} = \alpha \cdot \frac{h_i^{en} \cdot h_j^{tg}}{\|h_i^{en}\| \|h_j^{tg}\|} + \beta \cdot \left(1 - \frac{|l_i - l_j|}{\max(l_i, l_j)}\right) + \gamma \cdot p_{ij} \quad (6)$$

where α, β, and γ are weight coefficients, satisfying α + β + γ = 1; the first term represents the cosine similarity of the source language and the target sentence vectors, the second term represents the sentence length compatibility, l_i and l_j are the lengths of the source and target sentences respectively, and the third term p_{ij} represents the translation consistency score based on dictionary hit rate or back-translation consistency. If $S_{ij} \geq \tau$, then this sentence pair is retained and enters the parallel or quasi-parallel corpus set. This mechanism takes into account both surface length information and deep semantic information, and can better adapt to the regional

variations, loanword phenomena, and asymmetry in structure of ASEAN English texts.

(4)Subword Segmentation and Model Input Representation Construction

After sentence-level alignment is completed, the text needs to be converted into a standardized input form that can be directly processed by the neural machine translation model. Considering that there are a large number of low-frequency words, local proper names, mixed spellings, and unregistered words in the ASEAN English multilingual corpora, if modeling is directly based on words, it is likely to cause sparse word tables and unstable representations. Therefore, in this paper, the BPE or SentencePiece method is adopted to perform subword segmentation on the sentence pairs sequence, mapping the original word sequence to finer-grained subword units. On this basis, the discrete subword sequence is converted into vector representations and the position information is superimposed to form the model input.

Let the input sentence after segmentation be the subword sequence $X = (x_1, x_2, \ldots, x_n)$, then the input representation at the t-th position can be written as:

$$e_t = E(x_t) + Pos(t) \tag{7}$$

Among them, E(xt) represents the embedding vector of the subword xt, and Pos(t) represents the position encoding vector. This formula integrates the lexical semantic information with the sequence position information, providing a unified representation basis for the subsequent encoder modeling. Through subword-level input construction, it not only alleviates the problem of unregistered words caused by regional variations in the ASEAN English language variants, but also enhances the learning ability of multi-language shared word segments and improves the adaptability of the neural machine translation model to complex texts.

In summary, this paper has established a basic data processing system for the intelligent comparison task of ASEAN English multilingual corpora through multi-source collection, quality scoring, language identification, joint alignment, and subword representation. This system not only improves the structuring and computability of the corpora, but also provides stable data support for the subsequent semantic mapping and difference detection driven by neural machine translation.

3.2 Design of the Neural Machine Translation-driven Intelligent Comparison Model for Corpora

After completing the collection and construction of ASEAN English multilingual corpora, sentence-level alignment and preprocessing, this paper further constructs a neural machine translation-driven intelligent comparison model to achieve semantic mapping, difference identification, and feature quantification of cross-language texts. Considering the problems such as dense vocabulary variations, significant regional expression differences, and incomplete symmetry of cross-language semantic mapping in the ASEAN English corpora, this paper adopts the overall design idea of "input representation - cross-language modeling - difference calculation - result output", coupling the deep semantic learning ability of neural machine translation with the quantitative analysis ability of the intelligent comparison module to form a unified computational framework. The structure of the neural machine translation-driven intelligent comparison model for corpora is shown in Figure 2. Unlike traditional translation models that only focus on generating the target language, this paper places greater emphasis on the model's reutilization of "translation process information", meaning it not only considers whether the target language is generated, but also pays attention to how the source language and the target language correspond, deviate, and reorganize in the semantic space. Such a processing approach is more suitable for the research on multilingual corpora of ASEAN

English, because in such corpora, a large number of differences do not directly manifest in whether "translation can be achieved", but rather in whether the expression level, structural direction, and semantic focus of the translation have changed after the translation. Based on this understanding, in the model design, this paper integrates translation modeling and difference discrimination into the same framework, enabling the model to undertake both cross-linguistic mapping tasks and comparative analysis tasks.

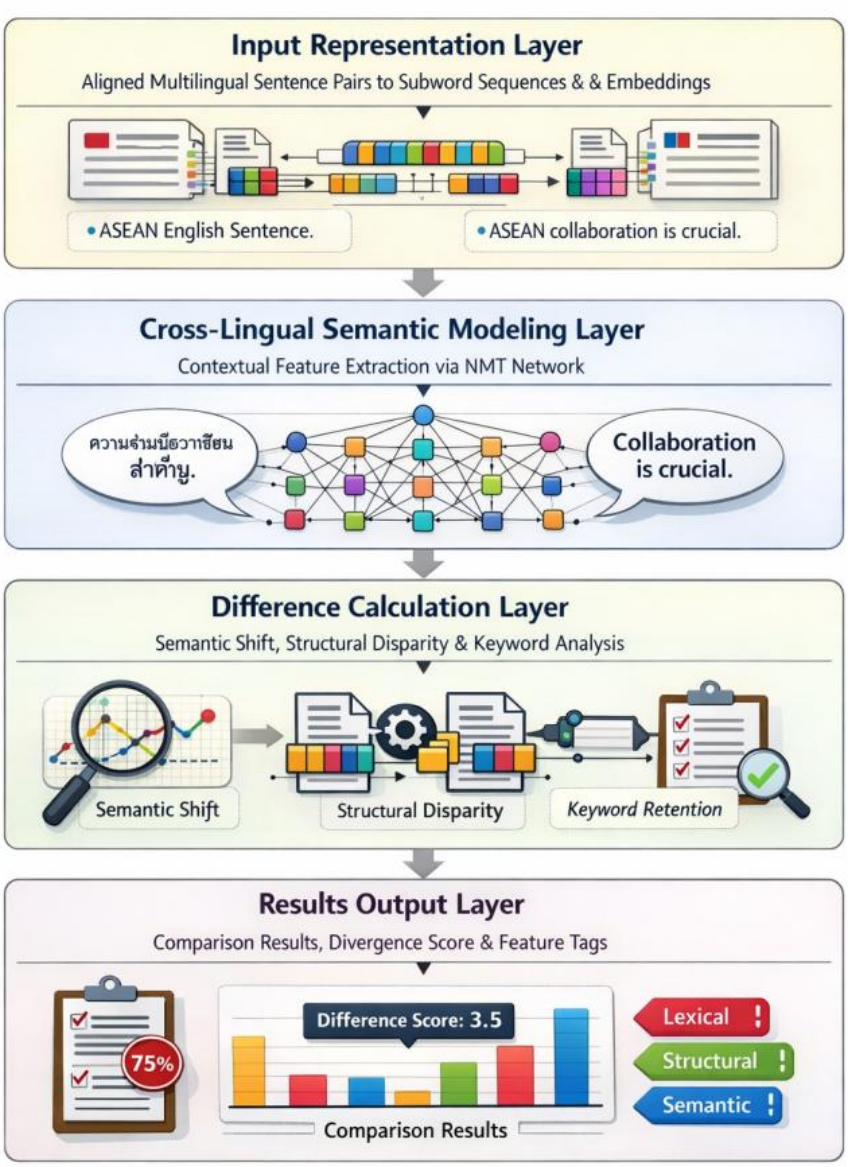

Figure 2: Structure of the intelligent comparison model driven by neural machine translation

(1) Overall architecture design

The model in this paper consists of four parts: the input representation layer, the cross-lingual semantic modeling layer, the difference calculation layer, and the result output layer. The input representation layer is responsible for converting the aligned preprocessed multilingual sentence pairs of ASEAN English into unified subword sequences and vector inputs; the cross-lingual semantic modeling layer relies on the neural machine translation network to extract the contextual correlation features between the source language and the target language; the difference calculation layer jointly analyzes the semantic shift, structural

differences, and keyword preservation in the shared representation space; the result output layer outputs the comparison results of the corpus, difference scores, and feature labels. This architecture extends neural machine translation from a simple translation generation tool to a cross-lingual feature extractor, enabling it to directly serve the intelligent comparison task of multilingual corpora of ASEAN English.

From the perspective of system mapping, let the source language input be Xs and the target language input be Xt, then the overall processing process of the model can be expressed as:

$$C = \Phi\left(F_{cmp}(F_{nmt}(X_s, X_t))\right) \tag{8}$$

where, $F_{nmt}(\cdot)$ represents the neural machine translation-driven cross-lingual semantic modeling module, $F_{cmp}(\cdot)$ represents the intelligent comparison module, $\Phi(\cdot)$ represents the output mapping function, and C represents the final intelligent comparison result of the corpus. This formula reflects the task logic of "first modeling, then comparing, and then outputting" of the model in this paper, and also demonstrates the core driving role of neural machine translation in the entire system.At the input stage, this paper does not simply treat different languages as separate collections of texts. Instead, by unifying the sub-word space and sharing the semantic embedding method, it attempts to retain as much cross-language transferable information as possible. The practical significance of this approach lies in the fact that some regional loanwords, abbreviations, and mixed expressions in ASEAN English, although having differences in their surface forms, often share similar semantic cores in cross-language contexts. If such information is overly fragmented at the input stage, subsequent modeling is prone to mistakenly interpret "transferable differences" as "completely heterogeneous differences", thereby affecting the translation output and comparison results.

(2) Cross-lingual semantic modeling module

To enhance the model's adaptability to ASEAN English regional variants, mixed expressions, and atypical syntactic structures,this paper adopts a Transformer-based encoding representation mechanism in the semantic modeling layer. After embedding and position encoding, the input sequence first enters the multi-head self-attention unit, extracting context-related features in different semantic subspaces. The attention calculation form is:

$$\mathrm{Att}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

Among them, Q, K, and V represent the query matrix, key matrix, and value matrix respectively, and d_k represents the dimension of the key vector. This formula can establish the dependency relationships between terms on a global scale, and has a good ability to capture the long-distance associations caused by loanwords, sentence order changes, and regionalized expressions in the ASEAN English corpus.

Under the multi-head mechanism, different attention heads respectively learn different semantic association patterns, and their aggregated results can be written as:

$$H = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h)W^o \tag{10}$$

Among them, head_i represents the output of the i-th attention head, h represents the number of attention heads, Wo is the output transformation matrix, and H is the encoded context representation. This process can enhance the model's ability to extract multi-granularity semantic features, enabling the source language and target language texts to obtain more stable context representations in a unified representation space.Considering that there are numerous

omissions, short sentence concatenations and partial code conversions in the multilingual corpora of ASEAN English, this paper, in addition to the Transformer encoding representation, also emphasizes the role of residual connections, layer normalization and feedforward networks in stabilizing the propagation of features. Multi-head attention is responsible for identifying "where it is relevant", while the feedforward transformation further compresses and reorganizes "how the relevant information is represented". After the collaboration of these two, it can reduce the interference of regional expression fluctuations on the main semantic content. For language combinations with relatively significant structural differences such as English-Tai and English-Vietnamese, this mechanism is particularly important because the model needs to maintain the cross-language semantic mainline without excessive drift even when there are significant differences in word order.

Furthermore, in the modeling process of this paper, the encoder output is not only used as the input for the decoding end, but also the intermediate semantic states are retained simultaneously for the subsequent feature extraction of the difference modules. Thus, the model's output is no longer just "translation results", but also includes "semantic trajectory information formed during the translation process". These intermediate representations can provide a more detailed judgment basis for difference scoring than the terminal text, and also help explain certain phenomena where the surface is similar but the semantic focus has changed in the corpora.

(3) Intelligent Comparison and Difference Scoring Module

After obtaining the cross-language hidden representations, this paper further constructs an intelligent comparison module to jointly analyze the source language representation, the target language representation, and the intermediate translation features. Considering that using only a single semantic similarity is insufficient to fully reflect the complex differences in the multilingual corpus of ASEAN English, this paper incorporates semantic deviation, structural deviation, and keyword retention rate into a unified scoring framework. Let the global representation of the source language be gs, and the global representation of the target language be gt. Then, the semantic deviation term is defined as:

$$\Delta_{\text{sem}} = \|g_s - g_t\|_2 \tag{11}$$

where $\|\cdot\|_2$ represents the Euclidean norm. This formula measures the deviation of the source language and the target language in the shared semantic space through vector distance. The larger the value, the more obvious the cross-language expression difference.

Based on this, this paper constructs a comprehensive difference scoring function:

$$D = \mu_1 \Delta_{\text{sem}} + \mu_2 \Delta_{\text{str}} + \mu_3 (1 - R_k) \tag{12}$$

where Δ_{str} represents syntactic or length structure deviation, R_k represents keyword retention rate, μ_1, μ_2, and μ_3 are weight coefficients, and they satisfy $\mu_1 + \mu_2 + \mu_3 = 1$. This formula not only focuses on the distance changes in the semantic space, but also takes into account the retention of text form and key information levels, making it more suitable for complex comparison scenarios in the multilingual corpus of ASEAN English where "surface differences but semantic proximity" or "form proximity but semantic deviation" exist.

To further enhance the interpretability of the results, this paper fuses the encoder's global representation, the decoder's global representation, and the difference score to form a unified comparison feature vector:

$$Z = W_f[H_s; H_t; D] + b_f \tag{13}$$

Among them, H_s represents the global features on the source side, Ht represents the global features on the target side, D represents the difference scoring vector, [;] represents the vector concatenation operation, W_f and bf respectively represent the weights and biases of the fusion layer. This formula enables the model to comprehensively utilize the translation modeling features and the difference analysis features within the same vector space, thereby improving the stability of the intelligent comparison results of the corpus. The core modules and functions of the neural machine translation-driven intelligent corpus comparison model are shown in Table 3.

Table 3: Core Modules and Functions of the Neural Machine Translation-Driven Intelligent Corpus Comparison Model

Module Name	Primary Input	Core Function	Primary Output
Input Representation Module	Multilingual sentence pairs sequence	Subword segmentation, embedding mapping, position encoding	Standardized vector sequence
Semantic Modeling Module	Source language and target language vector sequences	Context-dependent learning and cross-language feature extraction	Hidden layer representation matrix
Difference Calculation Module	Source-target global representation	Semantic deviation, structural deviation and keyword retention rate calculation	Comprehensive difference score
Feature Fusion Module	Coding features, decoding features, difference features	Multidimensional feature integration	Comparison feature vector
Output Discrimination Module	Combined vector	Generate corpus comparison results and difference labels	Intelligent comparison output

Table 3 indicates that the model proposed in this paper does not simply apply the neural machine translation results for posterior analysis, but builds a continuous processing chain from cross-language representation learning to difference quantification output within the model, thereby improving the automation and interpretability of the ASEAN English multilingual corpus comparison task.

(4) Joint Optimization and Output Mechanism

To balance the translation modeling quality and the intelligent comparison effect, this paper adopts a joint loss function to optimize the model end-to-end. Among them, the neural machine translation sub-task is responsible for constraining cross-language representation learning and hidden state update, and the intelligent comparison sub-task is responsible for constraining the fitting degree between the difference score and the true label. The joint objective function is expressed as:

$$\mathcal{L} = \lambda \mathcal{L}_{nmt} + (1 - \lambda)\mathcal{L}_{cmp} \tag{14}$$

where $\mathcal{L}_{nmt}$ represents the translation modeling loss, $\mathcal{L}_{cmp}$ represents the intelligent comparison loss, and λ is the balance coefficient. Through joint optimization, the model can maintain the cross-language semantic modeling ability while improving the recognition accuracy of the difference patterns of ASEAN English multilingual corpora.

In summary, the neural machine translation-driven intelligent corpus comparison model

constructed in this study, with Transformer-style cross-language semantic modeling as the core, supplemented by the difference scoring and feature fusion mechanism, and achieved through joint loss function for multi-task collaborative optimization. This model not only can complete the automated comparison of ASEAN English multilingual corpora, but also provides stable computational support for the quality assessment of translation output, difference feature identification, and translation theory verification in subsequent experiments.

4 Experimental and Result Analysis

4.1 Experimental Corpus Source and Evaluation Index Settings

To verify the effectiveness of the neural machine translation-driven intelligent corpus comparison model for cross-language semantic mapping, translation output quality, and difference identification in the ASEAN English multilingual corpus, this paper selects ASEAN regional news and information, public service texts, policy announcements, and cooperation and exchange materials as the experimental corpus sources, and constructs parallel or quasi-parallel corpus sets of English-Taiwanese,•English-Vietnamese,•English-Indonesian, and English-Malay language combinations. All corpora are processed for duplicate removal, language identification, sentence-level alignment, and subword segmentation before entering the experiment, and are divided into training sets, validation sets, and test sets in a 8:1:1 ratio to ensure the stability of the model training and testing process. The composition of the experimental corpora and main statistical information are shown in Table 4.

Table 4: Composition of Experimental Corpora and Main Statistical Information

Language Pair	Training Set / Sentence Pairs	Validation Set / Sentence Pairs	Test Set / Sentence Pairs	Average Sentence Length / Words	Vocabulary Size
English–Thai	18,240	2,280	2,280	21.6	18,450
English–Vietnamese	19,680	2,460	2,460	23.1	20,130
English–Indonesian	17,440	2,180	2,180	20.8	17,690
English–Malay	16,800	2,100	2,100	19.7	16,940

Table 4 shows that the experimental corpus in this paper has good balance in terms of language coverage, sample size and sentence length distribution, which can provide relatively stable data support for subsequent model training, performance comparison and difference analysis.

In terms of evaluation indicators, this paper comprehensively uses BLEU, TER, semantic similarity and difference recognition accuracy to evaluate the performance of the model. Among them, BLEU is mainly used to measure the n-gram matching degree between the machine translation output and the reference translation, and its expression is:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{15}$$

Among them, BP represents the length penalty term, wn represents the weight of the nth-

order n-gram, and pn represents the corresponding accuracy rate. To further measure the editing cost of the translation and the output deviation, this paper introduces TER as an auxiliary evaluation indicator. Its calculation formula is:

$$\mathrm{TER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I} + \mathrm{Sh}}{\mathrm{N}_{\mathrm{ref}}} \tag{16}$$

Among them, S represents the number of replacements, D represents the number of deletions, I represents the number of insertions, Sh represents the number of block movements, and Nref indicates the number of words in the reference translation. The lower the TER value, the closer the model output is to the reference translation. The accuracy of difference identification is used to measure the discriminative effect of the intelligent comparison sub-task on the difference type labels, and the relevant labels are generated based on manual review rules and the annotation results of the corpus, corresponding to the intelligent comparison loss term in Section 3.2. These indicators can evaluate the model performance from three aspects: translation quality, editing cost, and difference identification effect, providing a unified basis for the analysis of subsequent experimental results.

4.2 Analysis of Output Quality of Neural Machine Translation and Intelligent Comparison Effect

To verify the practical effect of the model in the multi-language corpus task of ASEAN English, this paper selects four methods: Seq2Seq+Attention, BiLSTM-Attention, Transformer, and the model proposed in this research for comparative experiments, and comprehensively evaluates the results from four dimensions: BLEU, TER, semantic similarity, and difference identification accuracy. The output quality and intelligent comparison effect of different models on the test set are shown in Table 5.

Table 5: Comparison of Output Quality and Intelligent Comparison Effect of Different Models

Model	BLEU	TER	Semantic Similarity	Difference Identification Accuracy/%
Seq2Seq+Attention	31.42	0.462	0.781	82.6
BiLSTM-Attention	33.87	0.428	0.806	85.4
Transformer	36.95	0.391	0.842	88.7
The model proposed in this paper	39.63	0.347	0.879	92.3

As shown in Table 5, the model proposed in this paper outperforms the comparison methods in all indicators, indicating that the joint design of neural machine translation and intelligent comparison module can simultaneously improve the translation quality and difference identification ability. Among them, the BLEU value of the model proposed in this paper reaches 39.63, which is 2.68 higher than that of Transformer; TER drops to 0.347, which is 0.115 lower than that of Seq2Seq+Attention; the semantic similarity increases to 0.879, and the difference identification accuracy reaches 92.3%, indicating that the model has stronger stability in cross-language semantic preservation and difference discrimination.From the overall trend of changes, the four types of models exhibit a relatively clear progressive relationship in terms of capabilities. Seq2Seq + Attention, as an earlier encoding-decoding structure, has already been able to complete basic cross-language mapping. However, in the task of this ASEAN English corpus which contains regional variations and multiple contextual factors, it still tends to have

unstable local correspondences, insufficient long-distance dependencies, and loss of translation details. Therefore, its BLEU score is only 31.42, and the difference recognition accuracy rate is 82.6%. BiLSTM-Attention has stronger sequence modeling capabilities than the former, with a BLEU score of 33.87, an increased semantic similarity to 0.806, indicating that bidirectional context modeling has significantly helped in understanding complex sentence pairs. The advantage of Transformer is further demonstrated in global dependency capture and multi-head attention mechanism. Its BLEU score reaches 36.95, and the TER drops to 0.391, indicating that the model has become more mature in handling cross-language structure rearrangement and semantic focus.

To further observe the convergence characteristics during the model training process, this paper counts the BLEU change trend of different models on the validation set. The BLEU change situation of different models in the training rounds is shown in Figure 3.

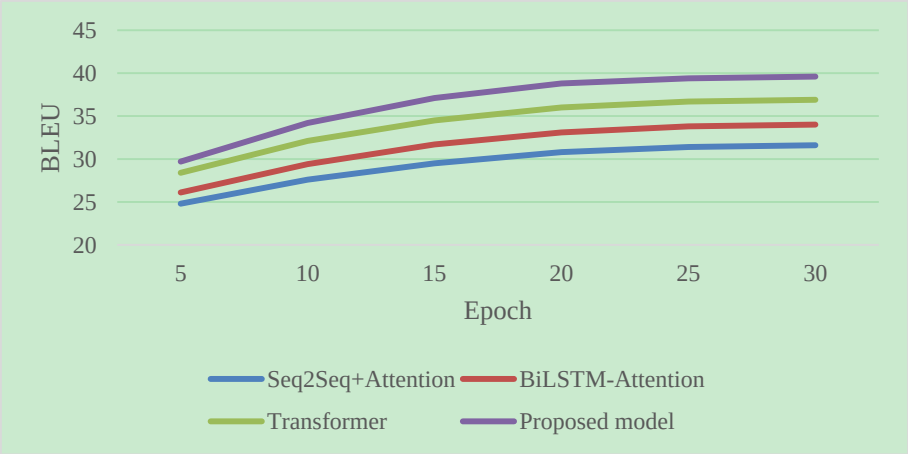

Figure 3: Line chart showing the changes in BLEU values of different models in the validation set

From Figure 3, it can be seen that all models showed a rapid upward trend in the first 10 training rounds. However, as the training progressed, the performance differences among different methods gradually widened. The model in this paper has significantly outperformed the other models after the 15th round and has stabilized at 39.6 in the 30th round, demonstrating better convergence speed and later stability. Based on the results in Table 5 and Figure 3, it can be concluded that this model performs optimally in terms of translation output quality and intelligent comparison effect compared to Transformer. The BLEU score has increased by approximately 7.3%, the accuracy of difference identification has increased by 3.6 percentage points, and TER has decreased by 0.044. This indicates that the constructed joint model can more effectively adapt to the complex expression differences in the multilingual corpora of ASEAN English.

4.3 Translation-based verification of the differences characteristics in ASEAN English multilingual corpora

After completing the analysis of translation output quality and intelligent comparison effect, this paper further validates the difference characteristics in ASEAN English multilingual corpora from a translation perspective. ASEAN English is influenced by local language contact, institutional expression habits, and cultural context during its long-term regional dissemination, often exhibiting phenomena such as lexical borrowing, structural adjustment, and semantic shift. To test whether the output results of this model have translation interpretive validity,

representative samples from different language combinations were selected, and manual review and result comparison were conducted from four aspects: lexical differences, structural differences, semantic shift, and keyword retention. The verification results of typical language case examples are shown in Table 6.

Table 6: Verification results of difference characteristics of typical language cases

Case No.	Language Pair	Difference Dimension	Typical Manifestation	Model Decision Result	Manual Verification
C1	English–Thai	Lexical Difference	Local administrative loanwords replace general English expressions	Lexical Difference	Consistent
C2	English–Vietnamese	Structural Difference	Adjustment of modifier position leads to changes in syntactic order	Structural Difference	Consistent
C3	English–Indonesian	Semantic Shift	Synonymous expressions show changes in semantic intensity in policy contexts	Semantic Shift	Consistent
C4	English–Malay	Keyword Retention	Core terms are stably preserved, with only partial compression of local expressions	Low Difference	Consistent

Table 6 shows that the identification results of the model in this paper for the difference types in typical language cases are consistent with the manual judgment, indicating that the intelligent comparison mechanism constructed can not only complete the quantitative analysis of differences at the numerical level but also support the interpretation of phenomena at the translation level. Especially in the English - Thai and English - Vietnamese samples, the model can accurately identify the loanword phenomenon and structural reorganization features in regional expressions, demonstrating better cross-language interpretation ability.

To further present the overall distribution characteristics of different language combinations in multi-dimensional difference indicators, this paper draws a bar chart based on the test set statistics results and compares the difference characteristic indicators. The comparison of difference characteristic indicators for different language combinations is shown in Figure 4.

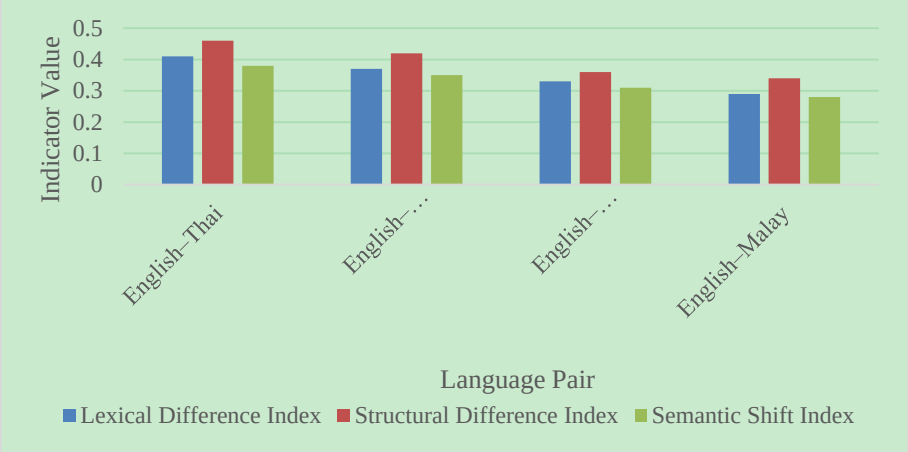

Figure 4: Comparison Bar Chart of Difference Feature Indicators for Different Language Combinations

From Figure 4, it can be seen that the English - Thai language corpus has the highest values in all three difference indicators. Among them, the structural difference index reaches 0.46, indicating that this language combination has more obvious translation conversion pressure in terms of expression sequence adjustment and syntactic mapping; the English - Malay language corpus has the lowest values for all three indicators overall. The vocabulary difference index and semantic offset index are 0.29 and 0.28 respectively, indicating that its cross-language mapping stability is relatively stronger. By combining the results in Table 6 and Figure 4, it can be concluded that the model in this paper not only can accurately determine the types of differences at the typical case level, but also can effectively distinguish the translation differences of different language combinations at the overall distribution level. Among them, the English - Thai language combination has a 0.12 higher structural difference index and a 0.10 higher semantic offset index than the English - Malay language combination, indicating that the method proposed in this paper can clearly reveal the translation difference characteristics and their hierarchical distribution in the ASEAN English multilingual language corpus.

5 Conclusion

This paper focuses on the intelligent comparison task of multilingual corpora of ASEAN English. It has constructed a technical route consisting of corpus collection, quality screening, language identification, sentence-level alignment, subword segmentation, Transformer cross-lingual semantic modeling, difference scoring, feature fusion, and joint loss optimization, achieving integrated processing of translation generation and difference discrimination. Experimental results show that the model in this paper achieves superior performance on English-Tai, English-Vietnamese, English-Indonesian, and English-Malay corpora, with BLEU reaching 39.63, TER dropping to 0.347, semantic similarity reaching 0.879, and difference recognition accuracy reaching 92.3%. Compared with Transformer, it has improved BLEU and accuracy by 7.3% and 3.6 percentage points respectively. Further case verification and difference feature statistics indicate that the English-Tai combination has a structural difference index of 0.46, and the semantic offset index of English-Malay is only 0.28, demonstrating that this method can accurately distinguish translation differences at the lexical, structural, and semantic levels of different language combinations. The research results show that the coupling of neural machine translation and intelligent comparison mechanism not only enhances the automation and refinement level of multilingual corpus processing of ASEAN English, but also provides a quantifiable and reproducible computational framework for regional English variant research and translation analysis based on corpora. Future research can further introduce larger-scale real-scenario corpora, paragraph-level context modeling, and multimodal semantic information to continuously improve the model's generalization ability in low-resource languages.

Furthermore, from a linguistic perspective, this research has expanded traditional study on regional English varieties, providing a new insight into the analysis of essential characteristics of ASEAN English, and achieving a deep integration of theoretical linguistics and computational linguistics. As a typical English variety, ASEAN English is influenced by the underlying transfer of various languages such as Thai, Malay, and Vietnamese, possessing unique systematic rules in terms of vocabulary, syntax, and pragmatics. Based on the methodological foundation of corpus linguistics, this research not only quantitatively explore

the differences in the use of words, sentence structures, lexical collocations, and modal expressions of ASEAN English under different native language backgrounds, but also reveal the transfer mechanisms of isolating and agglutinative languages on English varieties from a language typology perspective, verifying that ASEAN English is not a corrupted form of standard English but an independent variety with a stable system. Meanwhile, by analyzing the adaptation deviations of NMT models in different language pair translations, the core obstacles in cross-language communication can be located from a linguistic perspective, so as to help clarify the conflicts between ASEAN English and standard English in aspects of culture-loaded words, pragmatic habits, and grammar. Besides, it also lays a theoretical foundation for the standardization of translation in cross-border communication, government affairs, education, and other scenarios, possessing dual values of linguistic theoretical innovation and practical application.

Funding

This work was supported by the following projects:1.The Horizontal Project with Guangxi Maijie Information Technology Co., Ltd: Research on Product Copy Translation for ASEAN Cross-border E-commerce Platforms(2026); 2.The Horizontal Project with Beijing Keyi Culture & Technology Co., Ltd: Construction of Cross-border E-commerce English Terminology Database(2025); 3.The Horizontal Project, Shenzhen MarsHub Co., Ltd.:Construction of Automotive English Terminology Database(2025); 4.The Horizontal Project, Yunlian Network (Wuhan) Information Technology Co., Ltd. (2025); 5.The Special Project on Foreign Languages of Guangxi Philosophy and Social Sciences:A Study on the Digital Teaching Competence of English Teachers in Guangxi Universities (Project No.: 23WYL003); 6.The Industry-University Collaboration & Collaborative Education Project of the Ministry of Education: Construction and Research of an Intelligent Translation Laboratory for China-ASEAN Advanced Manufacturing (Project No.: 231100630125510)

About the Author

Author: Jianghong Kuang, born in Guilin, Guangxi P.R. China, in 1974, obtained a bachelor's degree from Zhongnan Financial and Economic University in China and a master’s degree in Guangxi Normal University. Currently teach at the School of Languages, Cultures, and International Education, Liuzhou Institute of Technology. Major research direction is Pragmatic Translation and Cross-cultural communication.

Corresponding author: Haopeng, Song, born in Jinhua, Zhejiang, P.R. China, in 1999, obtained a bachelor's degree from Liuzhou Institute of Technology in China, and a master’s degree from Guangxi University of Science and Technology. Currently teach at the School of General Education and Pedagogy, Guangzhou Huashang Vocational College. Major research field are L2 learning and translation studies.

Other author: Yanfu, Zhang, born in Baise, Guangxi, P.R. China, in 2005. Currently is a junior student majoring in Translation at Liuzhou Institute of Technology.

References

[1] San M E, Usanavasin S, Thu Y K,et al.A Study for Enhancing Low-resource Thai-Myanmar-English Neural Machine Translation[J].ACM Transactions on Asian and Low-Resource Language Information Processing, 2024, 23(4):24.DOI:10.1145/3645111.

[2] Stoyanova-Georgieva I.NATURAL VS ARTIFICIAL INTELLIGENCE AND NEURAL MACHINE TRANSLATION IN SPECIALISED TRANSLATION: A COMPARATIVE STUDY[J].Studies in Linguistics, Culture & FLT, 2025, 13(3).DOI:10.46687/XZWW6164.

[3] Balashov Y.The boundaries of meaning: a case study in neural machine translation[J]. Inquiry, 2025,68(7):1651-1684.DOI:10.1080/0020174X.2022.2113429.

[4] Yazar B K, Kili E.Using Kolmogorov–Arnold Networks in Transformer Model: A Study on Low-Resource Neural Machine Translation[J].IEEE Access, 2025, 13(13): 147034-147053. DOI:10.1109/ACCESS.2025.3601069.

[5] Escolano C,Marta R. Costa mg ussà,José A. R. Fonollosa.From bilingual to multilingual neural-based machine translation by incremental training[J].Journal of the Association for Information Science & Technology, 2021, 72.DOI:10.1002/ASI.24395.

[6] Wu Y, Qin Y.Machine translation of English speech: Comparison of multiple algorithms[J]. Journal of Intelligent Systems, 2022, 31(1):159-167. DOI:10.1515/jisys-2022-0005.

[7] Yuen J C.Embracing Artificial Intelligence-Assisted Machine Translation to Broaden Citations in Journal Articles.[J].Plastic and Reconstructive Surgery, 2025, 155(3):2. DOI:10.1097/PRS.0000000000011719.

[8] Zhang Hailin. A Cross-cultural Perspective on the Variants of ASEAN English [J]. Journal of Nanning Vocational and Technical College, 2014, 19(03): 30-32.

[9] Tian E, Zhu Z, Liu F,et al.Multimodal Neural Machine Translation Based on Knowledge Distillation and Anti-Noise Interaction[J].Computers, Materials & Continua, 2025, 83(2).DOI:10.32604/cmc.2025.061145.

[10] Yang S, Yang Q.Joint pairwise learning and masked language models for neural machine translation of English[J]. Artificial Life and Robotics, 2025, 30(2): 342-353. DOI:10.1007/s10015-025-01008-2.

[11] Suo W.Adaptive neural machine translation with attention mechanisms for English texts[J].International Journal of Information and Communication Technology, 2025, 26(15):25-40.DOI:10.1504/IJICT.2025.146372.

[12] Zhu S, Jian D, Xiong D.A Survey of Multilingual Neural Machine Translation Based on Sparse Models[J]. Tsinghua Science and Technology, 2025, 30(6):2399-2418. DOI:10.26599/TST.2023.9010097.

[13] Sun K, Tian Y, Zheng X,et al.Document-level Mongolian-Chinese Neural Machine Translation Incorporating Target-Side Data Augmentation and Topic Information[J].2024 International Conference on Asian Language Processing (IALP), 2024:204-209. DOI:10.1109/ialp63756.2024.10661113.

[14] Quoc T N, Thanh H L, Van H P.Khmer-Vietnamese Neural Machine Translation Improvement Using Data Augmentation Strategies[J].Informatica: An International Journal of Computing and Informatics, 2023, 47(3):349-360.

[15] Chen X, Wu L, Zhang Y.Enhancing use of BERT information in neural machine translation with masking-BERT attention[J].Proceedings of SPIE, 2023, 12717(000): 15.DOI:10.1117/12.2684653.

[16] Gong L, Li Y, Guo J,et al.Enhancing low-resource neural machine translation with syntax-graph guided self-attention[J].Knowledge-based systems, 2022(Jun.21):246. DOI:10.1016/j.knosys.2022.108615.

[17] Tay C.The Impact of Artificial Intelligence on International Trade: Evidence From Google Neural Machine Translation[J].Journal of Technological Advancements (JTA), 2021, 1(1).DOI:10.4018/JTA.20210101.oa6.

[18] Santhosh L, Asha K N, Potdar V,et al.Translating Iconic Indian Speeches from English to Kannada Using Neural Machine Translation[J].2024 Fourth International Conference on Multimedia Processing, Communication & Information Technology (MPCIT), 2024:349-357.DOI:10.1109/mpcit62449.2024.10892724.

[19] Wang T, Yu Z, Yu W,et al.Improving Chinese-Vietnamese Prototype Neural Machine Translation with Irrelevant Word Detection Denoising[J].2024 7th International Conference on Machine Learning and Natural Language Processing (MLNLP), 2024:1-6. DOI:10.1109/mlnlp63328.2024.10800263.

[20] Wang X, Wang T, Ricardo Muoz Martín, et al.Investigating usability in postediting neural machine translation: Evidence from translation trainees' self-perception and performance[J]. Across Languages and Cultures, 2021, 22(1):100-123. DOI:10.1556/084.2021.00006.

[21] Du Y.Construction of a Neural Machine Translation Model based on Cloud LM Algorithm[J].2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), 2024:1-5.DOI:10.1109/iacis61494.2024.10721713.

[22] Igarashi R, Miyagawa S.Enhancing Neural Machine Translation for Ainu-Japanese: A Comprehensive Study on the Impact of Domain and Dialect Integration[J].Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, 2024:413-422.DOI:10.18653/v1/2024.nlp4dh-1.40.

[23] Pramodya A, Mahima K T Y, Pushpananda R,et al.Enhancing Neural Machine Translation for the Sinhala-Tamil Language Pair with Limited Resources[J].International Journal on Advances in ICT for Emerging Regions (ICTer), 2024, 17(1):24-33. DOI:10.4038/icter.v17i1.7274.

[24] Iyer V, Barba E, Birch A,et al.Code-Switching with Word Senses for Pretraining in Neural Machine Translation[J].Findings of the Association for Computational Linguistics: EMNLP 2023, 2023:12889-12901. DOI:10.18653/v1/2023.findings-emnlp. 859.

[25] Cao H, Han D, Chu Y,et al.Multi-mechanism neural machine translation framework for automatic program repair[J].Journal of Intelligent & Fuzzy Systems, 2024, 46(4):7859–7873.DOI:10.3233/JIFS-234037.