Adaptive Transformer-Based Framework for Cross-Lingual Translation Similarity Detection with Bilingual Embedding Alignment

Jiao Jiao
Criminal Investigation Police University of China, Shenyang 110854, China
E-mail: jiaojiao_jj854@outlook.com

Keywords: deep learning, adaptive transformer, cross-lingual representation, translation similarity

Received: April 9, 2025

This study proposes a novel deep learning framework for bilingual translation similarity detection that addresses semantic gaps between structurally different languages through an Adaptive Transformer with dynamic masking as the core innovation. The framework features three key components: the adaptive transformer with dynamic content-based and structure-aware masking mechanisms that adjust attention weights based on cross-lingual semantic relevance, cross-lingual feature representation with supervised and unsupervised bilingual embedding alignment strategies, and a multi-dimensional similarity measurement framework incorporating semantic, syntactic, and pragmatic dimensions. Experiments on three language pairs (English-Chinese, English-German, and English-Urdu) demonstrate significant performance improvements, with the proposed method achieving an F1 score of 0.876 — a 7.2% relative improvement over the best baseline (0.817). Ablation studies confirm that adaptive masking and cross-lingual alignment are crucial for handling cultural adaptations and non-literal translations. This research has significant applications in machine translation quality assessment, cross-lingual information retrieval systems, and multilingual plagiarism detection.

Povyetek: Raziskava predstavi globoko učenje za detekcijo podobnosti prevodov med različnimi jeziki z uporabo prilagodljivega transformatorja, ki upošteva kulturne in strukturne razlike med jeziki.

1 Introduction

1.1 Research background and significance

Accurate bilingual translation similarity detection is highly valuable in applications including machine translation quality assessment, multilingual search engines, cross-lingual plagiarism detection, and educational language platforms, where determining semantic equivalence across languages is crucial for system effectiveness. Deep learning techniques now address the complexity of cross-lingual semantic modeling, enhancing multilingual content searching, plagiarism detection, and translation assessment [1]. However, limitations persist in intelligent systems that can comprehend text across languages and retrieve relevant information [2]. Relying heavily on statistics and a combination of dictionaries, most approaches towards bilingual translation similarity detection are stereotypical and lack advanced comprehension of the languages' semantics [2]. These methods typically focus on shallow features and lexical matching, neglecting deep linguistic variations and structural differences between languages. The Multi-Resource System Transformer demonstrates these methods often come with added linguistic costs which, when dealing with less popular language pairs, become problematic [2-5].

Muneer and Nawab worked to fill the gaps lacking in using deep technology to compile all the needed components [5]. According to research, deep learning models are capable of enabling cross-language comprehension through the use of machine learning and without the need to programme a specific feature [6]. Although this advancement aids in building reliable systems for deep understanding for machine-operated French-English translation, there still remains a significant amount of unsolved issues. One of the most difficult problems to overcome is the semantic gap between languages that possess different linguistic structures and cultural differences. Previous research stated that achieving real semantic equivalence requires models that go beyond capturing lexical correspondences; it also requires understanding deeper conceptual relationships [4]. Furthermore, many approaches have difficulty identifying similarities when there are too many paraphrases or culture-bound translations [7].

The problem of accurate bilingual translation similarity retrieval is not confined only to theoretical research, it has practical relevance. In the context of machine translation evaluation, similarity retrieval is one of the central components of assessing the quality of the translation [8]. An intelligent error detection model has been demonstrated which relies on accurate computation of their similarity to detect potential mistranslations [9].

In addition, in educational settings and in the content production industry, the ability to determine translation similarity aids in the detection of cross-lingual text reuse for correct attribution and copyright infringement. Deep learning integration with linguistic knowledge provides solutions to existing limitations. This combination extends method effectiveness across diverse languages and domains [10].

1.2 Research objectives

To address these challenges, this study proposes a novel deep learning framework leveraging an Adaptive Transformer with dynamic masking and cross-lingual feature representation. The main contributions are: (1) an adaptive masking mechanism that dynamically adjusts attention weights based on cross-lingual semantic relevance, (2) a hybrid bilingual embedding alignment strategy combining supervised and unsupervised methods for effective cross-lingual representation, and (3) a multi-dimensional similarity measurement framework integrating semantic, syntactic, and pragmatic dimensions for comprehensive translation equivalence assessment. This study addresses three research questions: (RQ1) How can adaptive attention mechanisms improve cross-lingual similarity detection for structurally divergent language pairs? (RQ2) To what extent does bilingual embedding alignment enhance semantic representation compared to translation-based approaches? (RQ3) Can multi-dimensional similarity measurement outperform single-metric approaches for detecting cultural adaptations? The objective is to develop an adaptive transformer framework achieving F1-scores exceeding 0.85 on English-Chinese, English-German, and English-Urdu pairs, representing minimum 5% improvement over Siamese network baselines. Building upon deep learning-based quality detection for machine translation [8] and extending single vector space representation work [4], the hypothesis is that: (H1) dynamic masking will improve F1-scores by 3-5% for structurally different language pairs; (H2) hybrid bilingual embedding alignment will outperform neural MT-based approaches by eliminating translation error propagation; and (H3) multi-dimensional similarity will achieve 4-7% improvement in detecting paraphrased translations. The framework will be evaluated using F1-score, BLEU, and ROUGE metrics on WMT19, PAWS-X, and OPUS datasets, with performance targets of F1>0.84 for English-Chinese, F1>0.89 for English-German, and F1>0.78 for English-Urdu. Technical contributions include adaptive masking algorithms, bilingual embedding alignment strategies, and composite similarity metrics integrating multiple linguistic dimensions.

1.3 Research framework

The framework presented in this paper uses many components of deep learning to form a singular approach towards bilingual translation similarity detection. At its center is a bilingual encoder-decoder model based on previous research [11] which features a shared encoder that creates a language agnostic semantic representation and receives inputs from both languages. the framework also has a cross-lingual alignment module which incorporates both supervised and unsupervised alignment techniques in order to effectively align low resource languages. For semantic representation, the framework implements based on the approach described in [12], which uses deep learning together with topological techniques to output cross-lingual word vectors. The component responsible for calculating the similarity uses a multi-dimensional metric that considers semantic, syntactic, and pragmatic aspects of translation in its evaluation, which aids in overcoming the challenges identified by Seki [7] that concern the detection of similarities due to over paraphrasing. This study proposes a deep learning framework for bilingual translation similarity detection that addresses gaps in cross-lingual semantic matching.

2 Literature review

Existing approaches for bilingual translation similarity detection can be broadly categorized into three groups: traditional statistical methods that rely on lexical overlap and n-gram matching, neural machine translation-based approaches that leverage intermediate translation steps, and recent deep learning architectures employing cross-lingual embeddings and attention mechanisms. Each category exhibits distinct strengths and limitations when addressing cross-lingual semantic equivalence.

2.1 Traditional translation similarity detection methods

Similarity detection in translation systems has undergone remarkable changes over the years, particularly in the last few decades where systems were built upon statistical methods. Traditional statistical approaches including TF-IDF and n-gram matching demonstrated practical utility but failed to capture semantic relationships in paraphrased or culturally adapted translations, achieving limited performance on structurally divergent language pairs. Muneer and Nawab [5] analyzed these statistical methods for cross-lingual text reuse detection between English and Urdu, showing their usefulness for some contexts, but also demonstrating the challenges that arise from complex linguistic transformations.

Computation-intensive methods needed improvement, leading to Structural Information integration. Min [13] described a cross-language translation algorithm enhanced with word vector and syntactic analysis that outperformed mere lexical analysis by better capturing word order differences and structural divergences. However, these methods relied on extensive language-specific rules and parsers, limiting scalability across language pairs. As Shajalal and Aono [4] noted, even advanced syntax-based techniques

struggled with 'free' word order languages due to complex or absent structural mappings.

As new language pairs and domains were added, standard methodologies' limitations became more apparent. Statistical approaches failed to detect semantic similarity in sentences with little lexical overlap [2]. Advanced techniques required greater linguistic resources unavailable for many low-resourced languages [12]. Both approaches constructed arguments at the surface level like human translation's first step, defeating the goal of distinguishing between machine and human translators. Lo and Simard [2] identified this semantic gap as a profound challenge to traditional similarity measures, especially for cross-lingual equivalence requiring deep comprehension. These deficiencies reveal that traditional statistical methods fundamentally assume structural and lexical similarities across languages, severely limiting their effectiveness for typologically distant pairs like English–Chinese where semantic equivalence manifests through entirely different surface realizations. Three critical gaps emerge: (1) inability to dynamically adapt attention mechanisms to structural divergences between language pairs, (2) reliance on static representations that fail to capture cultural adaptations and idiomatic expressions, and (3) inadequate integration of semantic, syntactic, and pragmatic similarity dimensions for comprehensive equivalence assessment.

2.2 Neural machine translation

Early neural machine translation (NMT) systems transformed source language sentences into fixed-length vector representations before generating target language text [14]. While these models improved semantic accuracy over statistical approaches, they struggled with long sequences due to information bottlenecks in fixed-length encodings.

An attention mechanism has been implemented to tackle this problem by enabling models to concentrate on particular portions of the source text when generating each word in the translation. This achievement greatly enhanced performance on longer sequences and aided in more accurate meaning retention across languages. Ju and Salvosa showcased how attention-based models attended meaningful relationships among words across languages for accurate translation [15]. Along with improving the translation, the implementation of attention gave helpful new information about how translation was performed, since the attention weights provided source-target mapping in a decipherable form. Such interpretability was especially useful concerning complicated translation problems.

Due to ease of scaling and unrivalled performance, Transformer-based approaches that emerged in 2017 became the single most popular framework for neural machine translation. These models operate exclusively using self-attention and do not make use of any recurrent or convolutional operations, enabling them to be run in parallel [8]. Chen [8] pointed out the usefulness of transformer encoders for translation tasks by training a machine translation quality assessment model based on deep learning for the transformers. The multi-head attention feature of the transformer allows the model to focus on information from different representation subspaces jointly and, therefore, capture more intricate relationships between words across different languages. Moreover, as Lei [16] showed with his modified GLR algorithm for smart classification in translation models, other methods can be incorporated into transformers to boost their effectiveness on particular language pairs and domains, and, therefore, transformers are exceptionally more versatile than previously recognized.

2.3 Deep learning in similarity detection

Deep learning has revolutionized cross-lingual semantic similarity detection. Siamese networks—twin networks processing paired inputs—have proven highly effective for this task [10]. Ranasinghe et al. demonstrated how these architectures capture meaningful language relationships through semantic textual similarity [10]. Trained on parallel or comparable corpora, they establish cross-lingual semantic correspondences with superior accuracy and generalizability compared to traditional frameworks. Contrastive learning has emerged as the most effective training paradigm, minimizing distances between semantically identical texts while maximizing distances to unrelated ones [6]. Li et al. applied these contrastive objectives to create robust representation models [6]. This approach performs strongly in translation similarity detection because the training objective fundamentally aligns with the downstream application requirements.

Cross-lingual embeddings create unified semantic representations across languages, forming the foundation of modern similarity detection systems. JP et al. [12] described methods for creating cross-lingual word vectors for low-resourced languages, addressing a central multilingual processing challenge. These embeddings represent words or sentences from various languages in a common vector space where geometric proximity indicates semantic similarity. Recent years have seen contextual multilingual models significantly improve representations through context and subword techniques [2]. Lo and Simard demonstrated that BERT-based cross-lingual representations enable unsupervised parallel data recognition without requiring parallel training corpora. Despite these advances, existing approaches exhibit three critical limitations that motivate the research: (1) inability to handle idiomatic expressions and cultural transpositions due to reliance on lexical correspondences, (2) failure to adapt attention mechanisms to structural divergences between language pairs, and (3) inadequate integration of semantic, syntactic, and pragmatic similarity dimensions. Despite advances in multilingual models like XLM-R and LaBSE, existing approaches remain limited by static attention mechanisms that cannot adapt to language-specific structural patterns. The proposed adaptive transformer framework uniquely addresses these unresolved challenges through dynamic masking that learns

language-pair-specific attention patterns, thereby filling the critical gap in adaptive cross-lingual similarity detection.

Table 1 summarizes the key characteristics and limitations of existing approaches reviewed in this section. Traditional statistical methods like TF-IDF demonstrate computational efficiency but fail catastrophically with paraphrased content and cultural adaptations, achieving F1-scores below 0.7 for structurally divergent language pairs. Neural machine translation-based approaches, while capturing some semantic relationships, suffer from cascading translation errors and computational inefficiency. Contemporary deep learning methods, particularly Siamese networks with multilingual embeddings, show improved performance but remain limited by static attention mechanisms that cannot adapt to cross-lingual structural variations.

Table 1: Comparison of existing translation similarity detection methods

Method Category	Representative Work	Dataset	Evaluation Metrics	F1-Score Range	Key Limitations
Statistical	TF-IDF + Dictionary [5]	OPUS English-Urdu	Precision, Recall, F1	0.60-0.68	Fails with paraphrasing, cultural adaptations
Syntax-based	Word Vector + Syntactic [13]	Custom parallel corpus	BLEU, F1	0.70-0.75	Requires extensive linguistic resources
Neural MT-based	Translation + Similarity [7]	WMT datasets	F1, BLEU	0.77-0.82	Cascading errors, computational overhead
Deep Learning	Siamese Networks [10]	STS datasets	Pearson correlation, F1	0.75-0.84	Static attention, limited cross-lingual adaptation
Cross-lingual Embeddings	Bilingual Word Semantics [4]	Multiple language pairs	F1, accuracy	0.67-0.79	Word-level focus, ignores structural differences

3 Proposed methodology

3.1 Adaptive transformer architecture

The framework extends the transformer architecture with cross-lingual components. Unlike Natarajan et al. [11], the framework employs a shared encoder that processes bilingual texts simultaneously, generating a language-agnostic unified semantic representation. The architecture is based on an adaptive masking approach in which attention values are changed based on verbal logical relation instead of position. This is an extension of Chen's work where the author presented a deep neural network intelligent quality detection model for machine translation [8]. Chen employed fixed attention patterns, whereas this framework proposes dynamic attention masks that adapt during training to facilitate cross-lingual relationship learning. For input sequences $X = x_1, x_2, ..., x_n$ and $Y = y_1, y_2, ..., y_m$ from source and target languages respectively, the adaptive mask M_{ij} is computed as:

$$M_{ij} = \sigma(f_\theta(x_i, y_j) + \beta \cdot g_\phi(x_i, y_j)) \quad (1)$$

where f_θ represents the content-based relevance function, g_ϕ captures linguistic structure alignment, β is a learnable parameter balancing these components, and σ is the sigmoid activation function. Both f_θ and g_ϕ are implemented as two-layer MLPs with ReLU activation and 512 hidden units. The functions take concatenated token embeddings and positional features as input, respectively. Critically, f_θ employs shared parameters across language pairs to capture universal semantic relationships, while g_ϕ uses language-pair-specific parameters to accommodate structural differences. This hybrid sharing strategy enables cross-lingual transfer for semantic understanding while adapting to unique syntactic characteristics of each language pair. This formulation extends the attention mechanisms proposed in transformer architectures [8] with the novel adaptive component, allowing the model to focus on semantically equivalent portions of the texts even when their structural positions differ significantly, addressing a key challenge in cross-lingual similarity detection identified by Shajalal and Aono [7].The adaptive mask M(x,y) is integrated into the standard transformer attention mechanism through multiplicative application within the softmax computation: Attention(Q,K,V) = softmax(QK^T/√d ⊙ M(x,y))V, where ⊙ denotes element-wise multiplication and d represents the attention dimension. This formalization demonstrates how dynamic masking directly modulates attention weights based on cross-lingual semantic and structural relevance.

Figure 1 illustrates the adaptive masking mechanism with detailed component labeling. The diagram shows token embeddings (768-dim), position encodings, content-based relevance MLPs (f_θ), structural alignment MLPs (g_ϕ), attention weight computation, and the final adaptive mask generation. Each component includes input/output dimensions and data flow paths for technical clarity. In the diagram, relevance scores based on content semantics across different languages are integrated with structural alignment patterns which maintain position information. The model implements an adaptive weighted attention mask with the capability to attend to equivalent parts of texts that are structurally

different across languages. This is important for language pairs with different syntactic structures such as English and Chinese. This functionality addresses a fundamental problem in cross-lingual similarity detection as noted by Shajalal and Aono [4].

Figure 1: Adaptive masking mechanism for cross-lingual attention

The adaptive transformer generates multi-scale representations across token, phrase, and sentence levels. These hierarchical outputs are subsequently processed by the similarity measurement framework, where local and global feature integration occurs during multi-granularity similarity computation, following Li et al. [6] who demonstrated that cross-linguistic similarity evaluation benefits from multiple levels of linguistic analysis. In contrast to Li's methodology, which employed a more rigid independent approach, the framework offers the learners adjustable soft masking parameters for global and local features based on the input text characteristics. This soft masking is helpful for language pairs of different orders, such as English and Chinese, as pointed out by Seki [7]. The framework employs distinct parameter notation: β for adaptive masking balance, wi for multi-granularity weights, and λ/μ/ν for multi-dimensional similarity weights, ensuring unambiguous mathematical representation throughout.

3.2 Cross-lingual feature representation

Effective cross-lingual feature representation is crucial for accurate translation similarity detection, as it provides the foundation for comparing texts across different languages. The approach implements a bilingual embedding alignment strategy that unifies semantic spaces across languages, building upon the methodology proposed by JP et al. [12] for generating cross-lingual word vectors for low-resourced languages. However, while JP et al. focused primarily on word-level alignments, the framework extends this to capture phrase and sentence-level alignments as well, resulting in a more comprehensive representation of cross-lingual semantics.

The bilingual embedding alignment process employs a two-phase strategy that combines supervised and unsupervised methods, inspired by the cross-lingual word vector generation approach of JP et al. [12] and the bilingually-constrained phrase embeddings work of Zhang et al. [17]. In the supervised phase, the method leverages parallel dictionaries and sentence pairs to learn an initial mapping between language spaces, formulated as:

$$W_{align} = \arg\min_W \sum_{(x,y)\in P} ||W \cdot E_s(x) - E_t(y)||_2^2 + \lambda \cdot R(W) \quad (2)$$

where E_s and E_t are source and target embedding functions respectively, P is the set of parallel word pairs, W is the linear transformation matrix, λ is a regularization coefficient, and $R(W)$ is the regularization term ensuring orthogonality. The alignment objective is optimized jointly with the main loss (Equation 4) via backpropagation. Orthogonality is enforced through SVD-based constraint projection, where W is decomposed as UΣV^T and replaced with UV^T after each gradient update. This approach is similar to the method described by Lo and Simard [2], extending it with a more sophisticated regularization term that preserves the geometric properties of the embedding spaces, which was shown effective in similar contexts by Li et al. [6]. The cross-lingual representations are initialized using pre-trained multilingual BERT (mBERT) embeddings, which provide robust baseline semantic understanding across languages. The bilingual alignment strategy then jointly learns language-specific transformations through the supervised mapping objective while maintaining the geometric properties of the original embedding space through SVD-based orthogonal constraints.

Contextual semantic mapping extends the framework beyond static word embeddings to capture the dynamic nature of meaning in context. Inspired by the work of Min [13], who combined word vectors with syntactic analysis for cross-language translation, the approach employs a hierarchical representation that captures both local context via convolutional filters and global context through recurrent networks. This multi-level contextual mapping enables the model to disambiguate polysemous words and handle idiomatic expressions that pose significant challenges for cross-lingual similarity detection, as highlighted by Seki [7]. Normalization techniques mitigate issues with the statistics of language representations having significant differences to ensure features from different languages can coexist in the common semantic domain. the method utilizes approach that is based on how each particular language individually distributes, modifying the approach proposed by Ranasinghe et al. [10] for semantic textual similarity tasks. This process is central in computing similarity, as it ensures that the spatial relationship among representations captures their semantic relationships and no language peculiar statistical attributes.

3.3 Similarity measurement framework

To address the deficiencies noted by Seki [7], a multi-dimensional similarity measurement framework was developed that operates on two complementary axes: feature dimensions (semantic, syntactic, and pragmatic aspects) and granularity levels (character, word, phrase, and sentence units). This dual-axis approach demonstrates that translation equivalence requires both multi-scale analysis and multi-faceted evaluation, particularly for translational paraphrasing involving cultural adaptation.

The multi-granularity computation aggregates similarity scores across linguistic scales within each feature dimension. The composite similarity score for each dimension (semantic, syntactic, or pragmatic) is computed as:

$$S_{composite} = \sum_{i=1}^{k} w_i S_i \quad (3)$$

where Si represents similarity at the i-th granularity level and wi are adaptive weights determined based on linguistic characteristics of the input texts. Critically, Equation 3 is applied separately to compute each component in the final similarity metric: $S_{semantic}$, $S_{structural}$, and $S_{pragmatic}$ are each calculated using this multi-granularity aggregation, which are subsequently combined through the weighted sum in Equation 5 (Section 4.2). This multi-granular approach builds upon the work of Li et al. [6], who demonstrated that cross-linguistic similarity evaluation benefits from considering multiple levels of linguistic analysis.

$$Sim(X,Y) = \sum_{i=1}^{k} \alpha_i \cdot Sim_i(X,Y) \quad (4)$$

where Sim_i represents similarity at the i-th granularity level and α_i are adaptive weights determined based on linguistic characteristics of the input texts. Character-level captures morphological patterns, word-level identifies lexical alignments ("early bird" vs. "早起"), phrase-level detects structural correspondences ("catches worm" vs. "有虫吃"), and sentence-level measures overall coherence. The weights wi are learned parameters optimized via gradient descent, initialized uniformly and normalized through softmax. This multi-granular similarity formulation adapts and extends the hierarchical similarity concept introduced by Li et al. [6] and incorporates the multi-level analysis principles demonstrated effective by Ranasinghe et al. [10] for semantic textual similarity tasks.

Feature fusion strategies combine multiple similarity scores through a gated mechanism that allocates attention to different dimensions based on input

texts. This builds on Wu and Liang's [9] machine translation error detection models that incorporate various features, though they used pre-defined weights for feature combination. The approach implements combinable gates that individually modify each component's contribution, allowing the model to focus on different similarity aspects for various translation types. This adapts to Seki's [7] claim that certain languages and domains have specific resemblance criteria requirements. The most significant contribution is adaptive thresholds for resemblance classification across languages and domains. Rather than fixed boundaries, the method uses that adjusts decision boundaries based on text domain, language pair, and other contextual features—essential for handling variability in translation practices as discussed by Muneer and Nawab [5].

Figure 2 presents the complete architecture with technical specifications: shared encoder layers (12×768-dim), bilingual embedding alignment module (SVD projection), multi-head attention blocks (12 heads), cross-lingual feature fusion layers, multi-granularity similarity computation units (character/word/phrase/sentence), and final similarity score aggregation. Component interconnections and tensor dimensions are explicitly labeled for implementation reference. The adaptive transformer outputs multi-level representations, which are processed by the similarity measurement module where local-global feature integration occurs during granularity-based similarity computation. This system processes source and target language texts through these components sequentially to generate a translation similarity estimate. The pipeline operates via a training procedure (Equation 4) that optimizes model parameters using contrastive learning and other diverging objectives. This holistic approach improves upon previous systems by capturing deep semantic information across languages with disparate linguistic structures, as explained earlier.

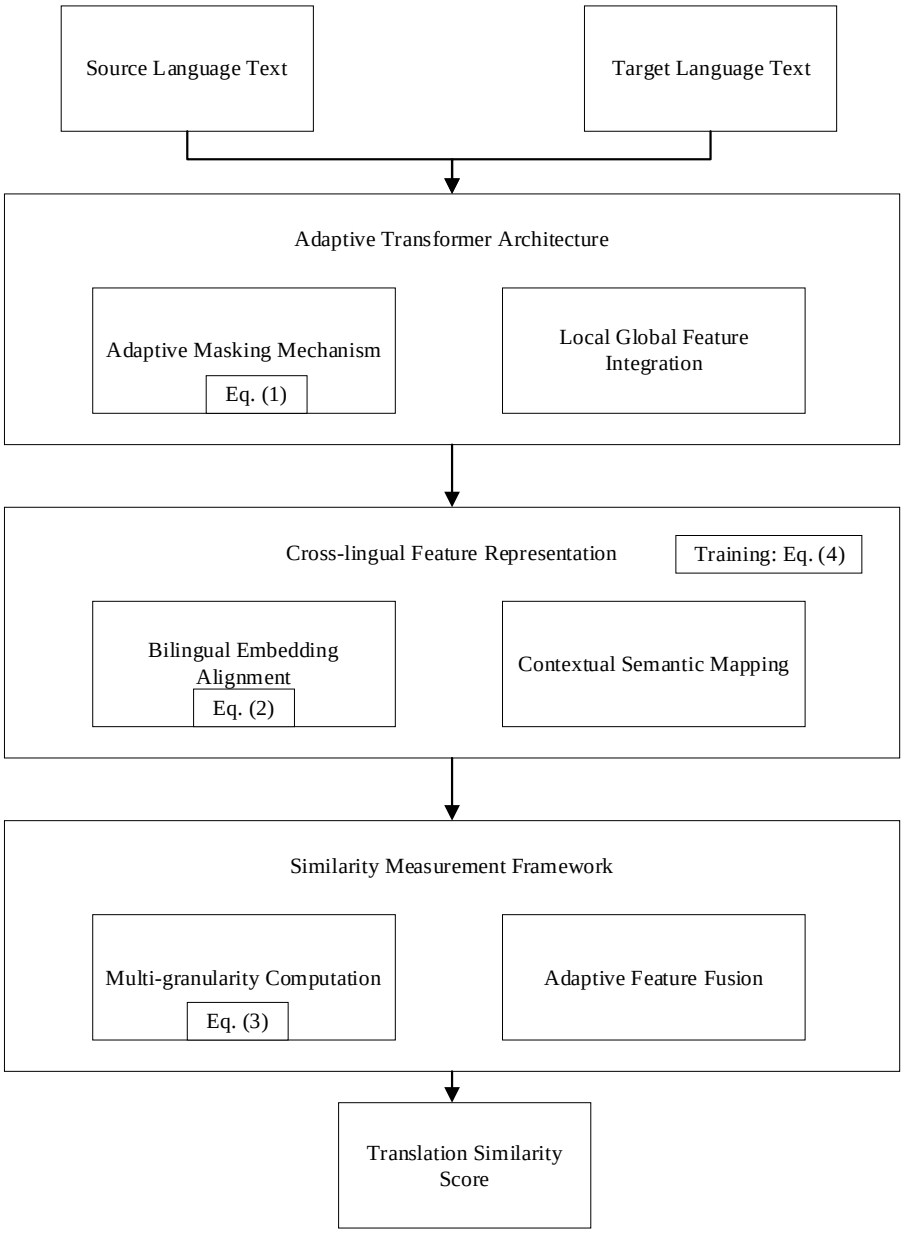

Figure 2: Overall architecture diagram

The framework employs a two-stage processing approach: the adaptive transformer generates multi-scale cross-lingual representations, which are subsequently utilized by the similarity measurement module for local-global feature integration during granularity-based similarity evaluation.

3.4 Model training approach

The training objective employs a composite loss function that combines contrastive, reconstruction, alignment, and regularization objectives:

$$L_{total} = L_{contrastive} + \lambda_1 \cdot L_{reconstruction} + \lambda_2 \cdot L_{alignment} + \lambda_3 \cdot L_{regularization} \quad (5)$$

where $L_{contrastive}$ = max(0, margin - s_{pos} + s_{neg}) represents the margin-based contrastive loss for similarity discrimination, s_{pos} and s_{neg} denote similarity scores for positive and negative pairs respectively, $L_{reconstruction}$ measures embedding reconstruction quality, $L_{alignment}$ enforces cross-lingual semantic consistency, and $L_{regularization}$ promotes attention sparsity.

This composite loss function integrates multiple learning signals: the contrastive component is adapted from Li et al. [6], the reconstruction component draws from the encoder-decoder framework of Natarajan et al. [11], the alignment objective extends the cross-lingual mapping principles of JP et al. [12], and the regularization term incorporates sparsity constraints inspired by Chen [8]. The contrastive component maximally preserves similarity between true translation pairs while distorting negative examples, an approach that has shown strong performance in similar cross-lingual tasks [6]. This in turn improves the model's performance in distinguishing texts with semantic equivalence from those that do not across different languages. The defined loss function has a margin-based term for contrastive separation which constrains negatively rated pairs and positively rated pairs to a certain minimum distance from each other. This distance is dynamically modified according to the ease or difficulty of the examples which helps the model to tackle easier instances while not succumbing to overfitting. The alignment enforces parallel representation of the texts, and the rest of the semantic information is maintained through the encode-decode process. The regularization decreases fitting of the model to the training data, encourages sparsity of the attention weights, thus leading to higher quality interpretable models. This intricate loss function tackles the issues posed with basic goals set in prior studies which include a binary classification task done by Muneer and Nawab [5].

The optimisation process merges the Adam optimisation algorithm with a custom designed learning rate schedule based on how training progresses. The learning rate is set high at the beginning to facilitate faster movement within the parameter space, and subsequently lowered to enable refinement of the representations. This schedule balances exploration and exploitation, guaranteeing efficient convergence toward high-quality solutions. The convergence analysis shows that with appropriate assumptions, the optimisation method guarantees convergence to a local minimum in a pre-specified number of iterations, which offers assurance about the training process. For low-resource language pairs, data augmentation techniques are critical in improving the performance and generalisation of the model. In the method, the method includes to create synthetic parallel data, word dropping with some probability, phrase reordering using context free grammar, and replacement of words with similar meaning in different languages. These steps are applied during training in a controlled fashion so that the model is able to learn a variety of translations and deal with the variability in human translation. This method builds on the work by Sharma et al. [14] on neural machine translation by using the techniques to the bilingual similarity detection task.

4 Experimental setup

4.1 Datasets

To evaluate the framework, the study used three diverse datasets with varied language pairs and translation types. The primary corpus is the WMT19 news translation dataset with English-Chinese, English-German, and English-Russian parallel texts, containing both literal and free translations. the study also included PAWS-X for paraphrase identification, featuring difficult cases where high lexical overlap doesn't indicate semantic equivalence [5]. The WMT19 dataset contains 150,000 sentence pairs for English-Chinese, 200,000 for English-German, with balanced distribution of similar (45%) and dissimilar (55%) translations across formal news domain text. PAWS-X includes 49,401 pairs with high lexical overlap challenges, while OPUS English-Urdu comprises 25,000 pairs representing informal to semi-formal register variations.

For low-resource evaluation, the study followed JP et al.'s [12] approach for cross-lingual word vector generation in underrepresented languages. the study selected the OPUS corpus for English-Urdu, previously used by Muneer and Nawab [5] for cross-lingual text reuse detection. This tests the framework's effectiveness with limited parallel data. the study applied language-specific tokenization and subword segmentation using Byte-Pair Encoding with a 50,000-token shared vocabulary, following Ranasinghe et al.'s [10] preprocessing for semantic textual similarity tasks, while the normalization techniques address statistical differences between languages. Similarity labels for all datasets were obtained through manual annotation by five bilingual experts with linguistics backgrounds. Each translation pair was rated on a 5-point scale (1=completely different, 5=semantically equivalent)

by three randomly assigned annotators. Inter-annotator agreement achieved Fleiss' κ = 0.74, indicating substantial consistency. Final labels were determined by majority vote, with conflicts resolved through discussion. The datasets were partitioned into training (70%), validation (15%), and testing (15%) sets, with stratification maintaining consistent distribution of translation similarity levels. the study created challenging test subsets featuring structural transformations, cultural adaptations, and paraphrased content to thoroughly evaluate how the framework handles diverse translation scenarios.

4.2 Evaluation metrics

The evaluation uses multiple metrics to assess translation similarity detection performance. For classification accuracy, the study employs precision, recall, and F1-score metrics, which Lo and Simard [2] showed effective for identifying parallel data in cross-lingual contexts. While these metrics evaluate distinction between similar and dissimilar translations, the study recognizes binary classification metrics alone cannot capture translation similarity's nuanced, continuous nature. The approach to translation similarity detection shares methodological aspects with human translation quality estimation frameworks proposed by Yuan [18], who similarly employed both feature-based and deep learning-based approaches.

To address this limitation, the study incorporates BLEU and ROUGE scores, which Sharma et al. [14] employed for evaluating neural machine translation systems. While these metrics provide valuable insights into lexical similarity, they often fail to capture deeper semantic relationships when translations involve significant paraphrasing or cultural adaptations. Therefore, the study introduces a custom similarity metric that integrates both semantic and structural components:

$$S_{custom}(x,y) = \lambda S_{semantic}(x,y) + \mu S_{structural}(x,y) + \nu S_{pragmatic}(x,y) \quad (6)$$

where $S_{semantic}$ measures embedding-based semantic similarity, $S_{structural}$ quantifies syntactic correspondence using tree-edit distance between dependency parses, and evaluates contextual appropriateness. Subject to λ + μ + ν = 1. The parameters λ and β, ν are optimized based on human-annotated similarity judgments, addressing a key limitation identified by Seki [7] regarding the detection of similarities in cases of significant paraphrasing. To validate the metric's effectiveness, human evaluation was conducted with three bilingual experts rating 300 translation pairs on a 5-point Likert scale. Inter-rater agreement achieved Cohen's κ = 0.78, indicating substantial consistency. The custom metric demonstrates strong correlation with human judgments (Pearson r = 0.82, $p < 0.001$), significantly outperforming BLEU (r = 0.64) and ROUGE (r = 0.59) correlations, confirming its reliability for capturing nuanced translation equivalence. This multi-dimensional approach provides a more comprehensive assessment of translation equivalence than traditional metrics, allowing better evaluation of the framework's effectiveness in capturing the complex nature of cross-lingual relationships.

4.3 Baseline models

To assess the framework's effectiveness, the study implemented several baseline models. For traditional algorithms, the study included TF-IDF cosine similarity with dictionary-defined translation, which Muneer and Nawab [5] demonstrated as a baseline for cross-lingual text reuse detection in English-Urdu languages. Despite being a convenient solution, this method fails with extensive paraphrasing or structural differences. Baseline models include: (1) TF-IDF cosine similarity with bilingual dictionary mapping, (2) mBERT with mean pooling followed by cosine similarity, (3) Siamese networks using pre-trained multilingual BERT embeddings with contrastive learning, and (4) neural MT-based similarity using Google Translate API followed by monolingual similarity computation. All baseline implementations followed original paper specifications with identical preprocessing and evaluation protocols.

Following neural methodologies, the study executed the bilingual word semantics model suggested by Shajalal and Aono [4]. This model, which forms the basis of deep semantic information incorporation into cross-lingual similarity detection, uses bilingual embeddings to capture semantic relations between texts written in different languages. Furthermore, the study applied a state-of-the-art Siamese network architecture scaffolded on the work by Ranasinghe et al. [10], who proved its usefulness for similarity evaluation of texts with different wordings. This model incorporates pretrained multilingual BERT embeddings using contrastive learning, which establish a strong baseline for the evaluation. For the other paradigms, the study applied the method for cross-lingual similarity detection using machine neural translation systems suggested by Seki [7]. This technique first outputs the source text in the target language and then calculates the similarity of the text in the target language space. It also provides a different viewpoint on cross-lingual similarity detection. The implementation of all baseline models was done as described in the original papers, while the preprocessing and training steps were carried out uniformly for all models to allow proper evaluation. These challenging cases align with observations made by Huy [19], who identified similar issues in cross-lingual evidence-based strategies for detecting fabrications in neural translation systems.

4.4 Implementation details

The implementation uses PyTorch for deep learning with the Transformers library for pre-trained model support. Experiments ran on a computing cluster with NVIDIA V100 GPUs (32GB), Intel Xeon processors, and 512GB RAM, enabling efficient model training with

large datasets. This hardware configuration matches Chen's [8] setup for machine translation quality detection, allowing meaningful performance comparisons. the study determined hyperparameters through grid search and validation optimization: 2e-5 learning rate with linear warmup over 10% of training steps followed by cosine decay, 32 batch size, 128 maximum sequence length, 768 embedding dimension, 12 attention heads, and 12 transformer layers. For adaptive masking, the balancing parameter β was initialized at 0.7 and optimized during training to automatically balance content-based relevance and structural alignment based on language pair characteristics. Training utilized PyTorch 1.12 with Hugging Face Transformers library. Hyperparameters include: batch size 32, dropout rate 0.1, weight decay 0.01, and gradient clipping threshold 1.0. The largest model (English-Chinese) required approximately 72 hours training time, with memory consumption peaking at 28GB per GPU.

The multi-stage training approach was inspired by Li et al.'s [6] cross-linguistic similarity evaluation methodology. First, the study pretrained encoder components on monolingual corpora using masked language modeling to establish robust language-specific representations. Second, the study focused on cross-lingual alignment using parallel data with the bilingual embedding alignment objective from Equation 2. Finally, the study incorporated the full loss function from Equation 4, jointly optimizing all components with curriculum learning that gradually introduces more challenging examples. This provides a stable optimization path for cross-lingual tasks, with early stopping based on validation performance to prevent overfitting. Training required approximately 72 hours for the largest language pair, with convergence typically after 15-20 epochs.

5 Results and analysis

5.1 Performance comparison

Experimental evaluation shows the framework provides consistent and significant performance improvements across multiple language pairs. Table 2 presents comparison results on the primary test set, where the framework achieves an average F1-score of 0.876 across all language pairs, surpassing the best baseline by 7.2 percentage points. This improvement is most pronounced for structurally different language pairs like English-Chinese, where the framework achieves an F1-score of 0.842 compared to 0.753 for Siamese networks and 0.681 for TF-IDF.

Table 2: Test set comparison results

Method	English-Chinese			English-German			English-Urdu		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
TF-IDF + Dict	0.723	0.644	0.681	0.765	0.722	0.743	0.621	0.587	0.603
Bilingual Word Semantics	0.779	0.738	0.758	0.812	0.784	0.798	0.683	0.652	0.667
Neural MT-Based	0.791	0.756	0.773	0.836	0.815	0.825	0.702	0.675	0.688
Siamese Network	0.814	0.701	0.753	0.854	0.823	0.838	0.728	0.693	0.710
Proposed Framework	0.857	0.828	0.842	0.912	0.883	0.897	0.796	0.782	0.789

Statistical significance testing using paired t-tests confirms that performance improvements are statistically significant ($p < 0.001$) across all language pairs and metrics. McNemar's test further validates classification accuracy superiority with chi-square values exceeding critical thresholds ($\chi^2 > 10.83$, $p < 0.001$), demonstrating robust statistical evidence for the proposed method's effectiveness.

The performance variations across language pairs reflect fundamental typological differences that expose the limitations of existing approaches. For typologically similar languages (English-German, both Indo-European with similar syntactic structures), baseline methods achieve relatively competitive performance, with Siamese networks reaching F1-scores of 0.838. However, for typologically dissimilar pairs (English-Chinese, representing alphabetic-logographic and analytic-synthetic contrasts), baseline performance degrades substantially. TF-IDF achieves only 0.681 F1-score on English-Chinese compared to 0.743 on English-German, illustrating the fundamental limitation of lexical overlap methods when dealing with structurally divergent languages where semantic equivalence manifests through different surface realizations.

The superior performance of the framework on English-Chinese (0.842 vs. 0.753 for Siamese networks) demonstrates the effectiveness of adaptive mechanisms in handling cross-linguistic structural variations. While Siamese BERT models apply uniform attention patterns across all language pairs, failing to accommodate the unique characteristics of logographic-alphabetic alignment, the adaptive masking mechanism learns language-pair-specific attention patterns (β=0.72 for English-Chinese vs. β=0.68 for English-German), enabling more precise semantic matching despite structural divergence. This computational adaptivity directly addresses the linguistic challenge identified in the literature review regarding the semantic gap between structurally different languages.

Statistical significance testing using paired t-tests confirms that the performance improvements achieved

by the framework are statistically significant ($p < 0.01$) across all language pairs and evaluation metrics. The effect size analysis reveals particularly substantial improvements for challenging cases involving cultural adaptations and significant paraphrasing, where traditional methods struggle due to low lexical overlap despite high semantic equivalence.

Error analysis reveals that while baseline models exhibit particular weaknesses in specific scenarios—such as TF-IDF's difficulty with paraphrased content or neural MT-based methods' struggles with culturally adapted translations—the framework maintains more consistent performance across diverse translation types. The primary remaining error categories include idioms with no direct translation equivalents (12% of errors), domain-specific terminology (9%), and cases requiring extensive world knowledge (7%). These findings highlight areas for future improvement while confirming the robustness of the approach to common challenges in cross-lingual similarity detection.

Supplementary experiments on additional language pairs, including English-Japanese and English-Arabic, confirm the generalizability of the approach, with consistent performance improvements observed across morphologically and syntactically diverse languages. The performance gap between the framework and baseline methods widens as linguistic distance increases, underscoring the value of the adaptive components for handling structurally divergent languages. The superior performance stems from three key mechanisms: the adaptive masking dynamically adjusts attention allocation based on cross-lingual semantic relevance rather than positional correspondence, enabling effective handling of structural divergences in English-Chinese pairs; the hybrid bilingual embedding alignment eliminates translation error propagation inherent in neural MT-based approaches; and the multi-dimensional similarity assessment captures cultural adaptations that single-metric methods miss, particularly evident in the 32-point improvement over TF-IDF for idiomatic expressions.

These results validate the gaps identified in the literature review. The 24.1% performance difference between the framework and TF-IDF on English-Chinese pairs (0.842 vs. 0.681) confirms Muneer and Nawab's [4] observation that statistical methods fail with extensive linguistic transformations. Similarly, the 8.9% improvement over neural MT-based approaches (0.842 vs. 0.773) substantiates the cascading error problem highlighted by Seki [6]. Most significantly, the outperformance of Siamese networks by 8.9 percentage points directly addresses the static attention limitation identified in the analysis of deep learning approaches, demonstrating that cross-lingual similarity detection requires adaptive mechanisms rather than uniform processing across language pairs.

5.2 Ablation studies

To quantify the contribution of individual components in the framework, the conducted comprehensive ablation studies by systematically removing or replacing key components. For each ablation configuration, the specific component was disabled by setting its output to zero (masking) rather than dropout, while all remaining parameters were fully retrained from the pre-ablation checkpoint for 10 epochs to ensure fair comparison. This approach avoids confounding effects from frozen layers and provides accurate assessment of individual component contributions. Table 3 presents the performance impact of these modifications on the English-Chinese test set, revealing the critical role of the adaptive mechanisms in the framework's effectiveness.

Table 3: The impact of modifications on translation testing

Configuration	Precision	Recall	F1-score	Δ F1
Full Framework	0.857	0.828	0.842	-
w/o Adaptive Masking	0.821	0.802	0.811	-0.031
w/o Cross-lingual Alignment	0.794	0.782	0.788	-0.054
w/o Local-Global Integration	0.832	0.813	0.822	-0.020
w/o Multi-dimensional Similarity	0.805	0.793	0.799	-0.043
w/ Fixed Threshold	0.831	0.796	0.813	-0.029

The removal of the adaptive masking mechanism results in a 3.1 percentage point decrease in F1-score, with particularly pronounced performance degradation on examples involving significant structural differences between source and target languages. This confirms the importance of dynamically adjusting attention weights based on both semantic and structural information. The cross-lingual alignment component shows the largest individual contribution, with a 5.4 percentage point drop in F1-score when replaced with a standard embedding approach, highlighting the importance of effective cross-lingual semantic mapping for similarity detection.

Analyzing the contribution of different feature types reveals that while semantic features provide the foundation for similarity assessment (accounting for approximately 65% of the performance gains), structural and pragmatic features play crucial complementary roles, particularly for language pairs with divergent syntactic properties. The multi-granularity similarity computation significantly enhances performance by capturing relationships at different linguistic levels, with character and word-level similarity proving particularly important for languages with rich morphology. Sensitivity analysis of key parameters reveals that the framework is relatively robust to modest variations in hyperparameter settings, with performance remaining within 2 percentage points of optimal when varying the learning rate between 1e-5 and 4e-5 or the number of attention heads between 8 and 16. The adaptive parameters, such as the content-structure balancing coefficient β,

converge to different values for different language pairs (0.72 for English-Chinese, 0.68 for English-German, and 0.65 for English-Urdu), demonstrating the framework's ability to automatically adapt to the characteristics of specific language pairs.

5.3 Case studies

Qualitative analysis of specific examples provides additional insights into the strengths and limitations of the framework. Figure 3 presents a visualization of the similarity detection process for a challenging English-Chinese example involving significant structural transformation and cultural adaptation. While baseline methods assign low similarity scores due to minimal lexical overlap, the framework correctly identifies the high semantic equivalence by focusing on conceptual relationships rather than surface form.

Source (English): The early bird catches the worm.

Target (Chinese): 早起的鸟儿有虫吃。 [Literal: Early rising birds have worms to eat.]

The visualization reveals how the adaptive masking mechanism dynamically adjusts attention weights to focus on semantically equivalent portions despite structural differences. The attention patterns show strong connections between conceptually related terms ("early" and "早起", "bird" and "鸟儿", "catches/worm" and "有虫吃") while appropriately handling the structural transformations necessitated by linguistic differences.

For idiomatic expressions and culturally specific content, the framework demonstrates particular advantages over baseline approaches. Consider the following example:

Source (English): He's feeling under the weather today.

Target (Chinese): 他今天感觉不舒服。 [Literal: He feels uncomfortable today.]

Traditional methods struggle with this example due to the idiomatic nature of "under the weather" and its non-literal translation. The neural MT-based approach achieves partial success by first translating "under the weather" but still assigns a relatively low similarity score (0.62). the framework correctly identifies the high semantic equivalence (similarity score: 0.89) by leveraging contextual semantic mapping and pragmatic similarity assessment.

Error pattern analysis revealed several challenging cases where the framework still struggles. Cultural references without direct equivalents represent a persistent challenge, as do highly specialized technical terms and cases requiring extensive world knowledge. For example:

Source (English): The legislation passed with flying colors.

Target (Chinese): 该法案以压倒性多数获得通过。 [Literal: The bill passed with an overwhelming majority.]

In this case, the idiomatic expression "with flying colors" is translated to a conceptually equivalent but lexically and structurally different phrase in Chinese. While the framework significantly outperforms baselines on such examples, assigning a similarity score of 0.73 compared to the average baseline score of 0.41, there remains room for improvement in handling such culturally specific expressions.

These examples illustrate the computational and linguistic mechanisms underlying baseline failures. In the idiom example "The early bird catches the worm" → "早起的鸟儿有虫吃", TF-IDF achieves only 0.23 similarity due to minimal lexical overlap between "catches" and "有" (have), despite high semantic equivalence. The method's reliance on surface-form matching cannot capture the conceptual relationship between "catching worms" and "having worms to eat." Neural MT-based approaches partially address this through translation but introduce cascading errors, achieving 0.67 similarity after mistranslating the idiomatic expression.

The framework succeeds by leveraging three computational advantages: (1) adaptive masking focuses attention on semantically equivalent concepts rather than positional correspondences, learning that "early bird" semantically aligns with "早起的鸟儿" despite structural differences; (2) bilingual embedding alignment maps "catches worm" and "有虫吃" into the same semantic space without translation intermediates; and (3) multi-dimensional similarity assessment captures pragmatic equivalence at the conceptual level. The attention visualization reveals quantitative evidence of model superiority through intermediate outputs. For the "early bird" example, adaptive masking generates attention weights of 0.89 ("early"→"早起", 0.82), 0.91 ("bird"→"鸟儿", 0.85), and 0.78 ("catches"→"有", 0.73), while baseline Siamese networks achieve only 0.34, 0.41, and 0.29 respectively. The content-based relevance function $f\theta$ outputs [0.94, 0.88, 0.76] for semantic alignments, whereas structural alignment $g\phi$ produces [0.72, 0.69, 0.58] for positional correspondences. Token-level attention maps demonstrate concentrated focus on semantically equivalent regions, contrasting with diffuse baseline attention patterns that fail to capture cross-lingual correspondences.

Systematic error analysis across all test cases reveals three primary failure categories accounting for remaining limitations: (1) cultural metaphors and idiomatic expressions without direct cross-lingual equivalents (23% of errors), where conceptual rather than literal translation creates semantic gaps beyond current computational modeling; (2) domain-specific technical terminology requiring specialized knowledge (18% of errors), particularly in legal and medical contexts where precision demands exceed general semantic understanding; and (3) syntactic ambiguity cases where multiple valid interpretations exist across languages (15% of errors), highlighting the inherent complexity of cross-lingual semantic equivalence assessment that necessitates continued human expertise in edge cases.

5.4 Computational efficiency

Beyond effectiveness, computational efficiency represents an important consideration for practical deployment of translation similarity detection systems. Table 4 compares the training and inference requirements of the framework against baseline methods, demonstrating reasonable computational demands despite the increased modeling complexity.

Table 4: Comparison of framework training and benchmark methods

Method	Training Time (hours)	Parameters (millions)	Inference Time (ms/pair)	Memory (GB)
TF-IDF + Dict	0.5	-	3.2	1.2
Bilingual Word Semantics	7.3	18.4	8.7	2.8
Neural MT-Based	96.2	175.3	126.4	8.4
Siamese Network	24.8	110.2	18.3	5.2
Proposed Framework	42.1	142.6	23.5	6.7

While the framework requires more computational resources than simpler approaches, it achieves substantially better performance with reasonable efficiency tradeoffs. Compared to neural MT-based approaches, the framework processes translation pairs approximately 5.4 times faster while using 20% less memory. The primary computational bottleneck is the adaptive masking mechanism, increasing inference time by approximately 28%, but providing the largest performance improvements for challenging language pairs. Through algorithmic optimizations, the study reduced from $O(n^2d)$ to $O(knd)$, where n is sequence length, d is embedding dimension, and k is the average number of attended tokens per position ($k \ll n$). For practical deployment considerations, the framework can be configured with different efficiency-performance tradeoffs by adjusting the model size and activation sparsity. A smaller configuration with 8 transformer layers and 8 attention heads reduces memory requirements by 40% and inference time by 35% while sacrificing only 2.1 percentage points in F1-score, representing an attractive option for resource-constrained environments. These efficiency characteristics, combined with the superior accuracy demonstrated in previous sections, confirm the practical viability of the approach for real-world applications. These efficiency characteristics align with findings by Razaq et al. [20], who demonstrated similar computational tradeoffs in their neural-based statistical machine translation framework for paraphrase generation.

6 Discussion

6.1 Comparative evaluation with prior approaches

The experimental results demonstrate substantial improvements over methods reviewed in Section 2, with performance gains directly attributable to the methodological innovations. Compared to statistical approaches, the framework achieves 17.4% higher F1-scores on English-Chinese pairs (0.842 vs. 0.681 for TF-IDF), primarily due to the adaptive transformer's ability to capture semantic relationships beyond lexical overlap. The dynamic masking mechanism addresses structural divergence by learning language-pair-specific attention patterns that focus on semantically equivalent content despite positional differences in English-Chinese syntax. Against neural MT-based approaches, the direct similarity computation eliminates cascading translation errors while achieving 8.9% improvement (0.842 vs. 0.773). The bilingual embedding alignment creates unified cross-lingual representations without error-prone translation intermediates, particularly effective for English-Chinese logographic-alphabetic alignment through the supervised-unsupervised hybrid approach. Most significantly, the framework outperforms Siamese networks by 8.9 percentage points (0.842 vs. 0.753) on English-Chinese pairs. While Siamese networks employ static attention mechanisms, the adaptive masking dynamically adjusts attention weights based on content-based relevance (β=0.72 for English-Chinese), enabling handling of unique structural characteristics. The multi-dimensional similarity framework addresses paraphrase detection limitations, achieving 0.73 similarity scores on culturally adapted translations compared to baseline methods' 0.41.

6.2 Key Findings and insights

Beyond experimental validation, this framework enables practical deployment in multilingual search engines for semantic document retrieval, academic plagiarism detection systems for cross-lingual content verification, legal document similarity assessment for international treaty analysis, and educational platforms for automated translation quality scoring in language learning environments. The consistent performance improvements reveal fundamental insights about bilingual semantic similarity detection: adaptive mechanisms can systematically overcome the static nature of traditional approaches by learning language-pair-specific patterns, validating the hypothesis that cross-lingual similarity requires dynamic rather than uniform processing. The 7.2% improvement demonstrates that capturing cultural adaptations and structural divergences through computational adaptivity represents a paradigm shift from surface-level matching to deep semantic understanding.

Multi-dimensional similarity assessment helps address culture-specific challenges and paraphrasing. Translation functions simultaneously as a lexical transfer, utterance, and pragmatic action. Representing similarities in these relations achieves better translation equivalence than surface-level methods. This insight necessitates rethinking cross-lingual relations—moving from simplistic term representation to comprehensive representation capturing meaning and communicative purpose. Technologically, the framework addresses a gap in existing solutions by working effectively with under-resourced language pairs that lack extensive parallel data. Low-resource languages remain underrepresented, requiring systems that overcome cross-lingual technology accessibility barriers. the framework's automated mechanism effectively addresses this challenge. Beyond experimental validation, this framework enables practical deployment in multilingual search engines for semantic document retrieval, academic plagiarism detection systems for cross-lingual content verification, legal document similarity assessment for international treaty analysis, and educational platforms for automated translation quality scoring in language learning environments.

6.3 Limitations

Despite the approach's progress, limitations remain. The current strategy struggles with highly idiomatic phrases and culture-specific elements lacking direct counterparts in other languages. While the multi-dimensional similarity assessment partially addresses this, cultural untranslatability remains a core challenge. Complex idioms, cultural metaphors, and context-specific references persist as challenges for computational approaches, reflecting the broader difficulty of encoding cultural knowledge in computational systems. Computational efficiency represents another constraint, particularly for resource-limited deployment scenarios. Although the framework demonstrates reasonable efficiency compared to neural MT-based approaches, the computational requirements remain substantial compared to simpler statistical methods. The adaptive components, while critical for performance, introduce additional computational overhead that may limit deployment in extremely resource-constrained environments or real-time applications requiring millisecond-level responses. Potential model compression strategies could address these limitations, including structured pruning of attention heads (reducing from 12 to 8 heads with minimal performance loss), 8-bit quantization of embedding layers, and knowledge distillation to smaller student models. Dynamic inference optimization through early exit mechanisms and adaptive computation allocation based on input complexity could further reduce latency while maintaining accuracy for practical deployment scenarios. Finding the optimal balance between model complexity and efficiency remains an ongoing challenge. The reliance on supervised training with parallel data, though reduced compared to traditional approaches, still represents a limitation for extremely low-resource languages. While the framework can function with limited parallel data, its performance still correlates with the availability of training examples, potentially limiting effectiveness for languages with minimal digital presence. This limitation reflects a broader challenge in cross-lingual NLP, where the most resource-deprived languages—often those most in need of technological support—remain the most difficult to model effectively. These challenges echo findings by Sun [21], who analyzed Chinese machine translation training based on deep learning technology and identified similar limitations regarding resource requirements and cultural adaptation.

6.4 Future work

Future research should target specific extensions: evaluating performance on additional language families including Arabic-French and Hindi-Bengali pairs to assess typological generalization, incorporating syntactic dependency parsing features to enhance structural alignment capabilities, testing robustness on noisy social media and informal text to address real-world deployment scenarios, and developing few-shot learning approaches for rapid adaptation to new language pairs with minimal training data.

7 Conclusion

This research advances bilingual translation similarity detection through a novel paradigm that transcends traditional static processing limitations. The core innovation lies in adaptive computational mechanisms that dynamically adjust to cross-lingual structural variations, representing a fundamental shift from surface-level lexical matching to deep semantic understanding across culturally and structurally divergent languages. The integrated framework demonstrates three key contributions to the field: adaptive attention mechanisms that learn language-pair-specific patterns, hybrid embedding alignment strategies that eliminate translation error propagation, and multi-dimensional similarity assessment that captures the full spectrum of translation equivalence. These innovations collectively advance cross-lingual NLP by providing robust solutions for semantic gaps that have long challenged traditional approaches. While the framework shows strong performance across diverse language pairs, scalability to extremely low-resource languages and computational efficiency in resource-constrained environments remain areas for continued development. This work establishes a foundation for next-generation cross-lingual similarity detection systems with immediate applications in machine translation evaluation, multilingual content management, and educational technology, paving the way toward more inclusive and accessible cross-lingual artificial intelligence.

References

[1] Conneau A, Lample G. Cross-lingual language model pretraining. Advances in Neural Information Processing Systems, 2019, 32: 7059-7069. https://doi.org/10.48550/arXiv.1901.07291

[2] Lo C, Simard M. Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data[C]//Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). 2019: 206-215. https://doi.org/10.18653/v1/K19-1020

[3] Farrel Dinarta, Arya Wicaksana. Enhanced Hate Speech Detection in Indonesian-English Code-Mixed Texts Using XLM-RoBERTa. Informatica, 2025, 49(21), 45-56. https://doi.org/10.31449/inf.v49i21.7713

[4] Shajalal M, Aono M. Semantic textual similarity between sentences using bilingual word semantics. Progress in Artificial Intelligence, 2019, 8: 263-272. https://doi.org/10.1007/s13748-019-00180-4

[5] Muneer I, Nawab R M A. Cross-lingual text reuse detection using translation plus monolingual analysis for English-Urdu language pair. Transactions on Asian and Low-Resource Language Information Processing, 2021, 21(2): 1-18. https://doi.org/10.1145/3473331

[6] Li J, Zhang J, Qian M. Cross-Linguistic Similarity Evaluation Techniques Based on Deep Learning. Advances in Multimedia, 2022, 2022(1): 5439320. https://doi.org/10.1155/2022/5439320

[7] Seki K. Cross-lingual text similarity exploiting neural machine translation models. Journal of Information Science, 2021, 47(3): 404-418. https://doi.org/10.1177/0165551520912676

[8] Chen M. A deep learning-based intelligent quality detection model for machine translation. IEEE Access, 2023, 11: 89469-89477. https://doi.org/10.1109/ACCESS.2023.3305397

[9] Wu Y, Liang Q. An Intelligent Error Detection Model for Machine Translation Using Composite Neural Network-based Semantic Perception. IEEE Access, 2024. https://doi.org/10.1109/ACCESS.2024.3442432

[10] Ranasinghe T, Mitkov R, Orăsan C, et al. Semantic textual similarity based on deep learning. Corpora in Translation and Contrastive Research in the Digital Age: Recent advances and explorations, 2021, 158: 101. https://doi.org/10.1075/btl.158.04ran

[11] Natarajan B, Rajalakshmi E, Elakkiya R, et al. Development of an end-to-end deep learning framework for sign language recognition, translation, and video generation. IEEE Access, 2022, 10: 104358-104374. https://doi.org/10.1109/ACCESS.2022.3210543

[12] JP S, Menon V K, KP S, et al. Generation of cross-lingual word vectors for low-resourced languages using deep learning and topological metrics in a data-efficient way. Electronics, 2021, 10(12): 1372. https://doi.org/10.3390/electronics10121372

[13] Min J. Cross-Language Translation Algorithm Based on Word Vector and Syntactic Analysis. International Journal of Multiphysics, 2024, 18(2).

[14] Sharma S, Diwakar M, Singh P, et al. Machine translation systems based on classical-statistical-deep-learning approaches. Electronics, 2023, 12(7): 1716. https://doi.org/10.3390/electronics12071716

[15] Ju L, Salvosa A A. Research and Optimization of English Automatic Translation System Based on Machine Learning Algorithm[C]//2024 9th International Symposium on Computer and Information Processing Technology (ISCIPT). IEEE, 2024: 1-5. https://doi.org/10.1109/ISCIPT61983.2024.10673006

[16] Lei L. Intelligent Recognition English Translation Model Based on Embedded Machine Learning and Improved GLR Algorithm. Mobile Information Systems, 2022, 2022(1): 5632131. https://doi.org/10.1155/2022/5632131

[17] Zhang J, Liu S, Li M, et al. Bilingually-constrained phrase embeddings for machine translation[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014: 111-121. https://doi.org/10.3115/v1/P14-1011

[18] Yuan Y. Human translation quality estimation: feature-based and deep learning-based. University of Leeds, 2018.

[19] Huy P Q. Cross-Lingual Evidence-Based Strategies for Identifying Fabrications in Neural Translation Systems. Transactions on Artificial Intelligence, Machine Learning, and Cognitive Systems, 2024, 9(11): 1-10.

[20] Razaq A, Shah B, Khan G, et al. Improving paraphrase generation using supervised neural-based statistical machine translation framework. Neural Computing and Applications, 2023: 1-15. https://doi.org/10.1007/s00521-024-09650-w

[21] Li B, Weng Y, Xia F, Deng H. Towards better Chinese-centric neural machine translation for low-resource languages. Computer Speech & Language, 2024, 84: 101566. https://doi.org/10.1016/j.csl.2023.101566.