Document-Level Neural Machine Translation via Multi-Scale Wavelet Fusion and G-Meshed Transformer with Attention Alignment

Linru Guo
HeNan Forestry Vocational College, LuoYang 471000, China
E-mail: LinruGuo@outlook.com

Keywords: Document-level neural machine translation, multi-scale wavelet fusion, G-meshed-transformer, attention alignment

Received: July 24, 2025

Document-level neural machine translation (NMT) aims to improve translation coherence by modeling cross-sentence dependencies. However, existing models like the sentence-level Transformer and G-Transformer struggle with capturing global context and produce noisy attention distributions. This paper introduces a novel document-level NMT framework that integrates multi-scale wavelet feature fusion, Balanced Contextual Slicing, a G-Meshed Transformer decoder, and an attention alignment mechanism. The method enhances encoder input using wavelet-derived semantic features, while parity resolution splits documents into overlapping segments to provide richer context without increasing parameters. A mesh structure in the decoder improves feature sharing and weighting across sentences. An attention alignment module further guides the model to focus on semantically relevant context using a lightweight context detector. Experiments on three English-German datasets (TED, News, Europarl) show that our model consistently outperforms strong baselines. In the two-stage training setup, it improves BLEU scores by +0.68 on TED, +0.81 on News, and +1.34 on Europarl over the sentence-level Transformer (average +0.95). With mBART-25 pretraining, it still gains +0.60 BLEU on average over the G-Transformer baseline. The results confirm that our approach significantly improves translation consistency, attention concentration, and handling of discourse phenomena such as deixis and ellipsis.
These results highlight the effectiveness of our framework in enhancing document-level translation consistency, contextual representation, and discourse-level coherence.

Povzetek: Članek predstavi dokumentni NMT, ki z valovno fuzijo značilk, uravnoteženim rezanjem konteksta, G-mesh dekoderjem in poravnavo pozornosti bolje izkorišča medstavčni kontekst ter izboljša konsistenco prevodov.

1 Introduction

Machine Translation refers to an automated linguistic conversion method that leverages computer algorithms to translate written or spoken content from one language into another. This technology significantly reduces the need for human labor and physical resources in translation tasks, thereby enhancing the efficiency and accessibility of cross-cultural communication. As advancements in natural language processing continue to progress, Neural Machine Translation (NMT) has emerged as a dominant approach within this domain. It is now extensively implemented across diverse practical contexts. At present, neural machine translation is commonly implemented in chapter-level NMT and document-level NMT, in which chapter-level NMT deals with the whole paragraph or article and focuses on the relationship between paragraphs, while document-level NMT deals with the whole document or a long article and needs to deal with the structure and context at the document level. Although chapter-level NMT has achieved excellent results in translating whole paragraphs or articles, document-level NMT is more complex than chapter-level NMT in terms of processing content, especially for the translation of long articles, which will lead to the loss of part of the contextual information if each paragraph is translated separately.

In recent years, document-level neural machine translation (NMT) has gained increasing attention due to its ability to capture broader contextual information, enhance translation quality, and improve consistency and coherence across documents. Unlike sentence-level approaches, document-level NMT treats the entire document as input, preserving semantic relationships across sentences and paragraphs. This enables more accurate translation of complex sentence structures. However, existing models often fail to fully utilize global context, largely due to limitations in training data, which typically lacks rich contextual annotations.

To address these challenges, this study proposes a document-level NMT framework based on the G-Meshed Transformer architecture. The core contributions are as follows: 1)Context-Aware Data Augmentation: A parity resolution method is proposed to enrich contextual signals without introducing additional model parameters, particularly benefiting low-resource translation tasks. 2)G-Meshed Transformer Decoder: A mesh-based decoder is designed to enhance context sharing and

filtering. It also incorporates label-based attention alignment to reduce the hypothesis space and mitigate noisy attention distributions. 3)Efficient Pretraining and Fine-Tuning: To lower training costs on large datasets, a two-stage process is used: pretraining with a standard Transformer, followed by fine-tuning with the proposed architecture.

Based on these contributions, this study evaluates three key goals: (1) Reducing attention entropy to improve focus in document-level attention distributions, (2) Improving translation quality as measured by BLEU scores on TED, News, and Europarl datasets, and (3) Enhancing discourse-level translation, particularly for phenomena such as deixis and ellipsis.

2 Related work

2.1 Neural machine translation

Neural machine translation is predominantly implemented through an end-to-end encoder-decoder architecture [1], which generally comprises two separate neural components: an encoder and a decoder. The encoder transforms each token from the source language into a dense vector representation and subsequently encodes the entire sequence into a continuous semantic space. The decoder, operating in a left-to-right manner, utilizes this semantic representation to sequentially produce output tokens in the target language. Notably, Chrisman et al. [2] were the first to introduce an end-to-end trainable model for translation tasks. To further enhance translation quality and processing efficiency, Cho et al. [3] proposed the use of recurrent neural networks (RNNs), thereby contributing a foundational architecture to the evolution of neural machine translation systems. Meanwhile, Sutskever et al [4] proposed LSTM, although the above deep neural networks achieved excellent performance on NMT. Fixed dimensional vectors are the bottleneck to improve the performance of encoder-decoder architectures and to efficiently encode and decode over long sequences.Bahdanau et al [5] proposed new encoder-decoder architectures with an attention mechanism to solve the problems such as long-distance dependency.

Vaswani et al. [6] introduced the Transformer architecture, which has since become the prevailing paradigm in machine translation research. Unlike previous models, the Transformer eliminates the use of both recurrent and convolutional operations, relying solely on the attention mechanism. This design not only facilitates highly efficient parallel processing but also enhances the model’s ability to capture global dependencies through multi-head attention. As a result, the Transformer achieves superior translation performance and offers improved interpretability. Although Transfor mer has pushed the development of neural machine translation, it has high training cost and weak ability to deal with long sequential statements. Therefore, researchers have proposed many improved derivative models, among which Kitaev et al. [7] proposed Reformer which uses locally sensitive hashing instead of dot product notation, and replaces standard residuals with a reversible residual layer. Subsequently, Bao et al. [8] proposed G-Transformer to use group labeling for attention guidance to solve the problem of long-distance dependency in multi-sentence translation.

2.2 Document-level neural machine translation

Document-level neural machine translation is different from traditional machine translation methods in that document-level translation needs to consider a larger inter-sentence context, including back references, lexical articulation and other discourse structures, while contextual information can provide some auxiliary information to the sentence to reduce the ambiguity of the sentence in the translation process. Depending on the type of document information used, document-level neural network translation can be broadly categorized into dynamic translation memories [9], contextual sentences [10], and whole documents [11].Tu et al. [12] proposed a dynamic cache-like memory to maintain a hidden representation of previously translated words. This memory mechanism encourages words in similar contexts to share similar translations, thus enhancing articulation to some extent. The primary distinction between utilizing an entire document and relying on adjacent sentences lies in the extent of contextual information, specifically the number of surrounding sentences incorporated.

2.3 Data enhancement for machine translation

Low-resource languages are those that do not have enough training data or have poor quality training data [13]. In this case, using traditional machine learning methods often leads to poor results. Augmenting data serves as a practical strategy for expanding the training dataset by synthesizing additional examples, thereby enhancing the model's capacity to generalize across unseen instances. In low-resource languages, data augmentation can be implemented in a variety of ways. By using these methods, a large amount of new training data can be generated, thus improving the training effect and generalization ability of the machine learning model.

Wei et al. [14] introduced a context-aware enhancement technique based on the premise that sentence semantics remain stable even when individual words are replaced with alternatives bearing similar paradigmatic relationships. This method enables sentence expansion without violating label constraints; however, it necessitates the integration of a well-constructed thesaurus. In another study, Rico et al. [15] utilized monolingual datasets in conjunction with automated back-translation, treating the results as supplementary parallel corpora. This strategy led to performance improvements, establishing new benchmarks, though the repetition in data incurred greater computational demands. Along similar lines, Sergey et al. [16] put

forward a strategy involving the enrichment of parallel corpora through the back-translation of target-language texts. Their work extended the theoretical and practical scope of back-translation and examined various techniques for generating synthetic source content. Lin et al. [17] adopted a method of pretraining a generalized multilingual neural translation model using randomly aligned lexical substitutions, effectively clustering semantically related terms across languages within the same vector space. Moreover, Pan et al. [18] implemented a contrastive learning approach to enhance a unified multilingual translation model, which proved effective in narrowing semantic divergences across languages and substantially improving performance in zero-shot translation tasks. However, each of the above data enhancement methods has its own advantages, but all of them introduce other parameters, which leads to the increase of training complexity and other problems. Beyond textual augmentation, multimodal approaches have also been explored. For example, Fang [24] proposed incorporating phrase-level universal visual representations into NMT to enrich semantic information.

2.4 Comparison and motivation

Table 1 summarizes recent document-level NMT methods evaluated on the TED, News, and Europarl datasets. These models adopt various strategies—such as dynamic cache memory, group-based attention alignment, and attention making to improve translation consistency. While effective to a degree, these approaches still struggle with key challenges: (1) modeling long-range dependencies across full documents, and (2) mitigating noise or diffuse attention, especially in longer contexts. Our proposed method addresses these limitations by integrating multi-scale wavelet feature fusion to strengthen hierarchical semantic encoding and employing a G-Meshed Transformer decoder to facilitate feature sharing and filtering. Additionally, our attention alignment module explicitly reduces attention entropy by guiding focus toward semantically relevant content—an aspect not directly optimized in prior work.

3 Method

The framework diagram of the algorithm proposed in this paper is shown in Fig. 1, and the main steps include: firstly, the metadata set is enhanced with parity resolution data, and then the input documents are encoded into individual units using the G-Meshed In the Transformer architecture, positional grouping labels such as 1, 2, 3, and so forth are assigned to differentiate sentence positions within the document. During translation, the target-to-source attention is directed by aligning the positional label of each target sentence with its corresponding source sentence label. This guided alignment effectively constrains the attention mechanism's search space, thereby reducing the overall hypothesis space and improving translation focus. The grouping labels are constraints on the attention, which helps to distinguish the current sentence from the preceding and following sentences. The group label is a constraint on attention, which helps to distinguish between the current sentence and the previous and next sentences. In addition, a mesh structure is added to the decoder, through which the model shares the feature information provided by the dataset, and the key features are screened and given weights, which helps to maximize the performance of the model.

Table 1: Comparison of related document-level NMT methods

Method	Main Technique	TED BLEU	News BLEU	Europarl BLEU	Limitation Addressed
Sentence-Level Transformer	Standard self-attention, no context modeling	24.79	25.28	31.33	Ignores document-level context
G-Transformer [8]	Group label alignment, G-attention	25.10	25.58	32.34	Limited context use, no entropy control
Attention Calibration [23]	Attention regularization	24.97	25.20	32.41	Partial improvement, attention still diffuse
Proposed (Ours)	Wavelet fusion + G-Mesh + attention alignment	25.47	26.09	32.67	✓ Handles long-range context✓ Reduces noise in attention

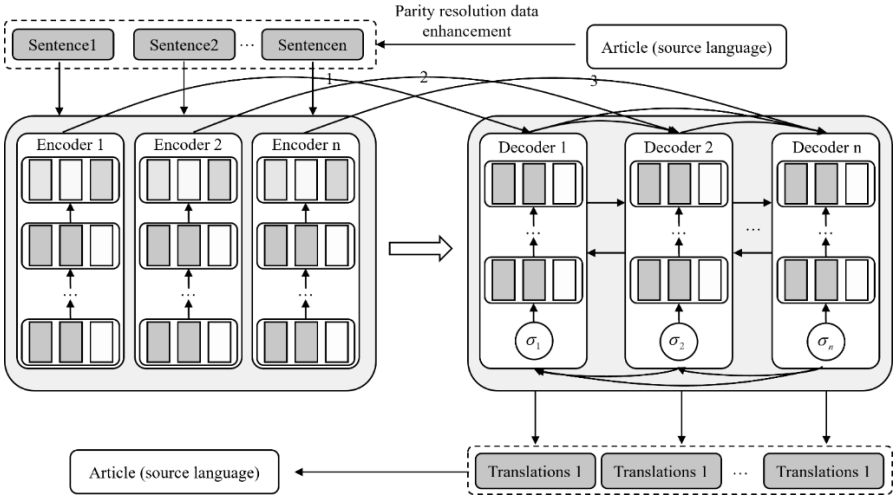


Figure 1: General framework

3.1 Balanced contextual slicing

We propose a balanced contextual slicing (BCS) strategy, which was previously referred to as parity resolution augmentation, to generate additional training samples by splitting each document into near-equal contiguous segments. Data augmentation is a widely adopted strategy in machine learning aimed at expanding the dataset to support improved learning efficiency and enhanced generalization performance. In the context of machine translation, this approach has proven valuable for boosting translation accuracy and the model's adaptability across diverse linguistic inputs. Conventionally, document-level neural machine translation datasets treat each text as a sequence of isolated sentences, where the relationship between input and output is governed by the following formulation:

$$D_S = \{x^1, x^2, \ldots, x^n\} \quad (1)$$

$$D_T = \{y^1, y^2, \ldots, y^n\} \quad (2)$$

where D_S and D_T are the source and target language documents, respectively, and n is defined as the number of utterances contained in the document. Therefore, the traditional document-level neural machine translation process can be defined as:

$$L_{\text{doc}} = -\sum_{i=1}^{N}\sum_{j=1}^{L_y^i} \log p_\theta \left(y_j^i \mid y_{<j}^i, x^i, S^i, T^i \right) \quad (3)$$

where S^i is the utterance of the source document and T^i is the utterance of the target document, which in many document-level translations consists of only two to three sentences, and most of the current work focuses on S^i, by using either layered attention or an additional encoder, the translation data is shown in Figure 2.

Compared with the traditional document-level neural machine translation dataset, this paper enhances the document-level translation dataset with parity resolution data. The defining formula is as follows:

$$L_{n\text{-}doc} = -\sum_{i=1}^{L_y} \log p_\theta \left(y_i \mid y_{<i}, D_x \right) \quad (4)$$

where D_x is the full context of the source document and $y_{<i}$ is each statement of the target document. Specifically, in this paper, each document is divided into k parts equally for multiple splitting, and all the sequences are collected together, $K \in \{1, 2, 3\}$. For each segment, the corresponding source and target sentences are concatenated and used as additional parallel examples. For example, a six-sentence document yields: two 3-sentence segments, three 2-sentence segments, and six 1-sentence segments (in addition to the full 6-sentence document).

Algorithm 1: Parity resolution augmentation (PRA)

Input: Document $D = \{(x^1, y^1), \ldots, (x^n, y^n)\}$, maximum split count K
Output: Augmented set $\mathcal{A}$
1. Initialize $\mathcal{A} \leftarrow \emptyset$.
2. For each $k \in \{1,2, \ldots, K\}$:
a. Partition $\{1, \ldots, n\}$ into k contiguous blocks of size $\lfloor n/k \rfloor$ or $\lceil n/k \rceil$.
b. For each block b:
Concatenate source sentences in b into X_b.
Concatenate target sentences in b into Y_b.
Add (X_b, Y_b) to $\mathcal{A}$.
3. Return $\mathcal{A}$.

To better guide the attention mechanism, each sentence within a document is assigned a group label (e.g., 1, 2, 3, …) indicating its position in the document. During training, these labels are appended as additional features alongside token embeddings for both the source and target sentences. In practice, this is implemented by embedding each group label into a fixed-dimensional vector space and adding it element-wise to the token embeddings, like positional encoding.

During inference, group labels are generated dynamically based on the decoding process. Specifically, when decoding a target sentence, the model assigns it the same group label as its aligned source sentence. This alignment is tracked at the segment level: if the model is translating tokens from the second source sentence, the current target tokens inherit the group label "2." This ensures that target-to-source attention is constrained within the corresponding group, while still allowing soft attention to nearby groups.

Formally, let $g_s(i)$ denote the group label of source token x_i, and $g_t(j)$ the group label of target token y_j. The cross-attention score between y_j and x_i is then modified as:

$$\alpha_{ij} = \frac{\exp\left(\frac{Q_j K_i^T}{\sqrt{d}} + \lambda \cdot \mathbb{I}[g_s(i) = g_t(j)]\right)}{\sum_m \exp\left(\frac{Q_j K_m^T}{\sqrt{d}} + \lambda \cdot \mathbb{I}[g_s(m) = g_t(j)]\right)}, \quad (5)$$

where λ is a learned bias weight and $I[\cdot]$ is an indicator function. This formulation prioritizes attention between source and target tokens that share the same label, thereby reducing noisy attention across unrelated segments.

3.2 G-Meshed-transformer model

Transformer has achieved great success in many natural language processing tasks (e.g., machine translation, language modeling, and text categorization) and computer vision tasks (target detection and image segmentation), and has become a mainstream model for machine translation. However, document-level translation uses long data sequences, and by extending the translation scope to the entire document, supervised training of the Transformer can fail due to the large target-to-source attention complexity. Therefore, G-Transformer introduces localization assumptions into Transformer as a generalization preference to reduce the assumption space of target-to-source attention.

At the same time, this paper further optimizes the characteristics of the document-level dataset enhanced by parity resolution data by adding a mesh structure in the decoder, and the model propagates feature information through the mesh structure in the decoder, allowing it to emphasize key contextual signals and suppress irrelevant noise. It filters the key feature information and assigns weights to it, and at the same time, the model matches the labels of the target and the source to guide the target-to-source attention mechanism, which reduces the assumption space of the attention mechanism and helps the model to This reduces the hypothesis space of the attention mechanism, helps the model to distinguish between the current sentence and the context, and significantly improves the translation performance.

Although the Meshed-Transformer remains an extension of the original Transformer architecture, it continues to employ dot-product attention mechanisms. In this framework, the encoder processes input vectors to derive the query (Q), key (K), and value (V) representations. The resulting output is computed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

Here, d_k denotes the dimensionality of the key vector K. The overall architecture continues to follow the encoder-decoder paradigm, as illustrated in Figure 2.

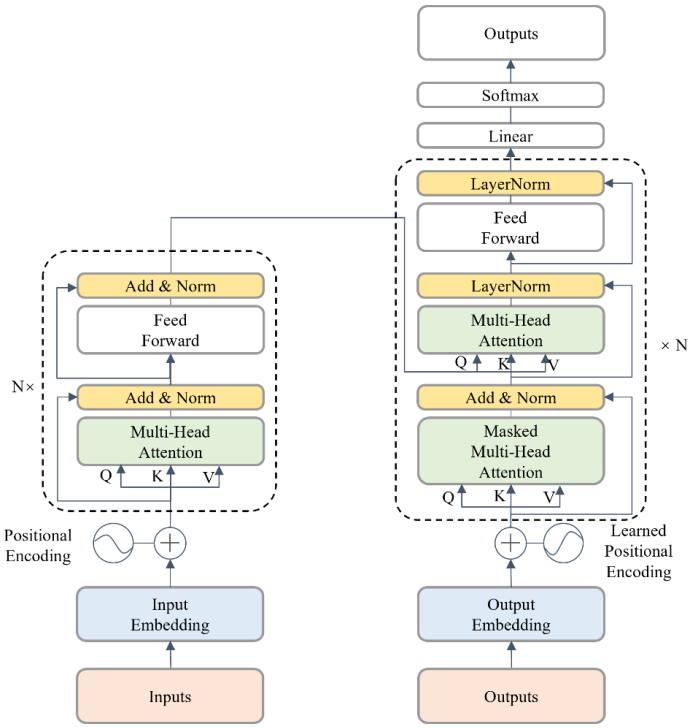


Figure 2: Diagram of G-Meshed-Transformer framework

4 Document-level attention alignment algorithm

Currently, document-level translation models focus on eliciting document-level contextual information through an attention mechanism. However, document-level attention requires longer inputs to process than intra-sentence attention, leading to excessive distraction (high entropy). This inevitably elicits meaningless noise signals and affects the model's ability to effectively utilize document information. To address this problem, this paper proposes a perturbation-detection-based alignment method for document-level attention, which aims to enhance the focus of attention on the relevant contextual words for each word during the translation process, thus improving the focus of attention.

4.1 Attention and word alignment in neural machine translation

The interpretability of the attention mechanism has been a hot research topic, and much work has focused on exploring whether the attention weights faithfully represent the effect of each input signal on the output symbols. In the field of question-answer systems or text categorization, some researchers have found that replacing high attentional weights with lower attentional weights does not affect the predictive performance of the model, possibly because the attentional mechanism tends to assign higher weights to unimportant tokens, such as punctuation and stop words; on the other hand, some studies have shown that the correlation between the attentional weights and the gradient-based feature importance measure is weak. In addition, unlike the direct discussion of attentional weights as an explanation for model decisions, the literature has found that trained attentional mechanisms are equipped to learn some meaningful information about the relationship between the input and output, such as syntactic information.

In the field of machine translation, most of the work suggests that the key contribution of the attention mechanism to the model is to perform word alignment between the output words and the input words, and thus a lot of efforts have been made in word alignment to guide the model's attention weights, or extracting more accurate word alignment from the trained attention weights A recent study has shown that the level of the attention weights is not too far away from human intuition and expectation, and that the neural mistranslations or over-translations in machine translation come from inaccurate attention weights, i.e., the attention model has not been able to learn reliable word alignment information. This conclusion is consistent with previous findings, and Lu et al. provide a Chinese-English sentence-level translation example to illustrate the effect of word alignment on neural machine translation, which is shown in Table 2. After translating the target word “in”, the Transformer focuses its attention on the source word “far suburb”, and then successfully translates “countryside “. After translating the target word “deaths”, the Transformer incorrectly focuses on the terminator “<EOS>”, causing the model's translation to end prematurely, and part of the source content (i.e., “traffic interruption”) to be translated. “After word alignment, the model gives part of its attention to the source content “traffic interruptions”, so that the relevant content can continue to be translated.

Table 2: Case: Impact of word alignment on machine translation

Item	Content
Source language sentence	远郊 连 日 大雪 多 人 死亡 交通 中断
Reference translation	heavy snow in countryside left many deaths and trans-portation disrupted
Translation before word alignment (Transformer):	heavy snow in countryside caused many deaths
Translation after word alignment (Transformer):	heavy snow in countryside has caused many deaths andtrafic interruption

On the task of document translation, some works also hope to improve the performance of the model through word alignment information. For example, Yin et al. use an attention regularization algorithm to direct attention to human-labeled relevant contextual words, which can increase the performance of disambiguation for cross-sentence references.Lei et al. first analyze the relationships such as repetition, co-reference, and subordination in the document using a document parsing tool, and then force the model to pay more attention to these articulatory relationships through an attention mask to reduce the impact of irrelevant contextual information. contextual information. Unlike their work, in this paper, we hope to make the model's attention mechanism more capable of focusing on relevant document-level contexts through self-supervision without inducing additional knowledge base.

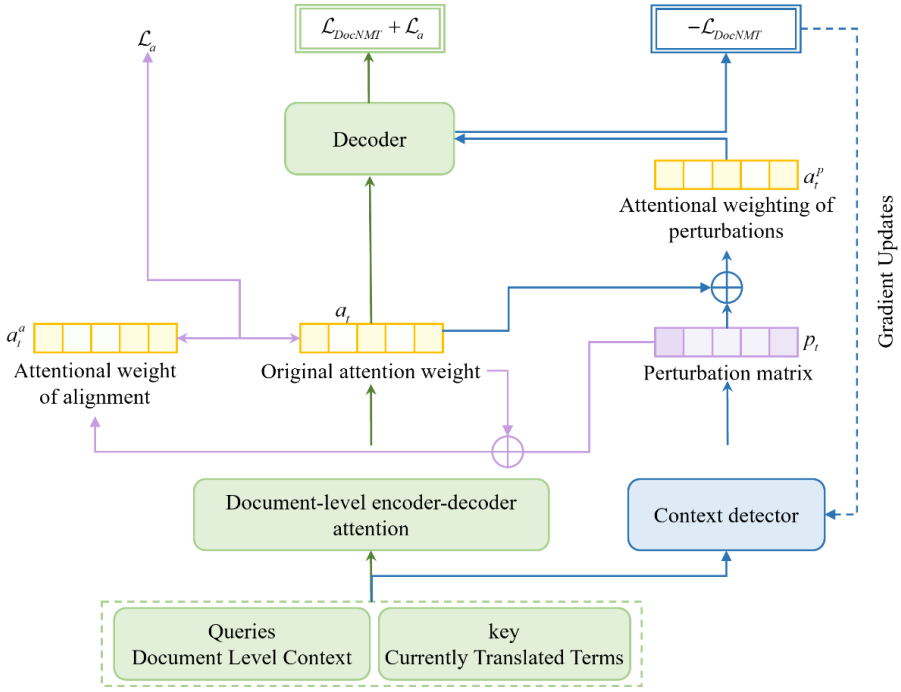


Figure 3: Model structure of the proposed document-level attention alignment method. The figure illustrates the decoder layer, context detector, and how alignment loss is integrated with cross-attention.

4.2 Document-level attention alignment algorithm

The basic structure of the document-level attention alignment method proposed in this paper is shown in Fig. 3. In order to facilitate the elaboration of the figure only shows the structure of a decoder layer in the model. The whole training process can be divided into two main steps:(1) Context Detector Training:First, a context detector (Context DetectorCD) is trained, which is attached to the backbone model as a lightweight network. Its main function is to evaluate the importance of each source-side contextual vocabulary to the model prediction, so as to identify the key contextual information that has a significant impact on the translation results. (2) Attention weight optimization:Based on the evaluation results of CD, an additional attention alignment loss function is induced C. The purpose of this loss function is to optimize the model's attention weights so that attention is more focused on the contextual terms that are practically significant to the translation, while reducing the interference from the noisy context.

4.3 Context detector

This paper proposes a document-level translation consistency enhancement method based on multi-scale wavelet fusion and an improved Transformer architecture. Multi-scale semantic features are extracted via wavelet decomposition and injected into the Transformer to improve context modeling. In our implementation, we use the Daubechies-4 (db4) wavelet with two levels of decomposition; the resulting approximation and detail coefficients are concatenated with the original embeddings and linearly projected before being injected into the Transformer encoder. Specifically, we construct source and target document-level context sets as:

$$D_S = \{x^1, x^2, \ldots, x^n\}, D_T = \{y^1, y^2, \ldots, y^n\} \tag{7}$$

The training objective of the model involves optimizing the conditional probability of the target sequences, guided by supervision that incorporates document-level contextual information:

$$L_{\text{doc}} = -\sum_{i=1}^{N} \sum_{j=1}^{L_y^i} \log p_\theta \left(y_j^i \mid y_{<j}^i, x^i, S^i, T^i \right) \tag{8}$$

Additionally, a non-document-level loss is introduced for comparative evaluation:

$$L_{n\text{-}doc} = -\sum_{i=1}^{L_y} \log p_\theta \left(y_i \mid y_{<i}, D_x \right) \tag{9}$$

The self-attention operation is formulated through a scaled dot-product approach, which serves as the foundation for computing attention scores:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

To enhance alignment and consistency, we incorporate a word alignment module. As illustrated in Fig. 4, alignment behavior after the introduction of contextual learning demonstrates improved focus around key tokens like "in" and "deaths", which significantly enhances coherence in translation.

To formally define the attention alignment objective, we introduce an additional document-level attention loss that guides the model to assign higher weights to contextually relevant tokens identified by the context detector. Let the predicted attention distribution for target token y_j over source tokens be $\alpha_j = \{\alpha_{j1}, \ldots, \alpha_{jn}\}$. The context detector produces a binary mask $m_j = \{m_{j1}, \ldots, m_{jn}\}$, where $m_{ji} = 1$ if source token x_i is marked as relevant for predicting y_j, and $m_{ji} = 0$ otherwise.

The attention alignment loss is defined as:

$$L_{\text{attn}} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n} m_{ji} \log \alpha_{ji}, \quad (11)$$

where N is the total number of target tokens and n the number of source tokens. This loss encourages the attention mechanism to concentrate on positions identified as relevant by the context detector, thereby reducing entropy and noisy attention weights.

The final training objective integrates the standard document-level translation loss with this auxiliary attention alignment loss:

$$L_{\text{total}} = L_{\text{n-doc}} + \alpha L_{\text{attn}}, \quad (12)$$

where $L_{\text{n-doc}}$ is the negative log-likelihood loss for document-level translation, and α is a tunable hyperparameter controlling the strength of attention alignment.

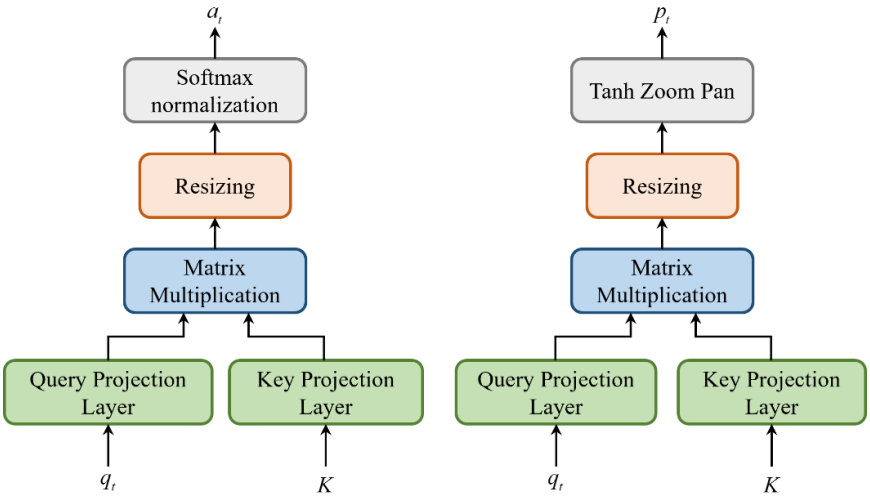


Figure 4 Example of word alignment before and after applying the proposed alignment optimization. The figure shows improved attention focus around key tokens such as ‘in’ and ‘deaths.

5 Results and discussion

5.1 Experimental data

This chapter also conducts experiments on three widely used English-German document translation baseline datasets extracted by Maruf et al. [19] including the domains TED, News and Europarl. In this paper, the proposed model is applied to G-Transformer, a baseline for performing pseudo-document translation, and therefore the data needs to be preprocessed according to its requirements. Firstly, all the documents in the database are sliced into pseudo-documents with a maximum length of 512 tokens, and if a single sentence is longer than 512 tokens, it is treated as a separate pseudo-document. The results of the slicing of the three databases, TED, News and Europarl, are shown in Table 3. During data processing, all texts were tokenized using the Moses toolkit and segmented into sub-words using the Byte Pair Encoding (BPE) model [20]

The constructed vocabulary contains approximately 30K subword units using Byte Pair Encoding (BPE). Across all datasets, the out-of-vocabulary (OOV) rate is below 0.5% after BPE segmentation. The TED corpus has a type–token ratio of ~0.18, the News dataset ~0.22, and Europarl ~0.20, indicating moderate lexical diversity. These statistics provide a clearer picture of the learning environment.

Table 3: English-German translation training set/validation set/test set cut-off results

Dataset	Number of pseudo-documents	Average number of sentences	Average number of words
TED	11K/483/123	18.3/18.5/18.3	436/428/429
News	18.5K/172/263	12.8/12.6/11.3	380/355/321
Europarl	0.16M/346/498	10.3/10.4/10.3	320/326/323

In this paper, the Transformer model is implemented using the Fairseq toolkit and the experiments are performed on two GeForce RTX3090 GPUs. In this paper, we apply the proposed attention alignment method to the document-level encoder-decoder attention of G-Transformer.G-Transformer adopts the setup recommended by Bao et al [8]. i.e., only intra-sentence attention is used in the lower layers of the Transformer,

and both document-level global attention and intra-sentence attention are used in the higher layers (the last two layers).

(1) Two-stage training: First, a basic Transformer model is trained for sentence-level translation using sentence-level data sliced from the dataset, and then the proposed document-level translation model is fine-tuned for sentence-level translation using document data. The sentence-level model is configured as a Transformer-base model with 6 layers of encoder and 6 layers of decoder, 8 attention heads, 512 dimensions of intermediate hidden representations, and 2048 dimensions of feed-forward network hidden representations. For sentence-level model training, we use the Adam optimizer ($\theta_1 = 0.9, \theta_2 = 0.998$) to optimize model parameters. An inverse square root scheduling strategy is utilized to adjust the learning rate dynamically, beginning with an initial rate of 5e-4 and incorporating 4000 warm-up iterations. A dropout rate of 0.3 is employed to mitigate overfitting, alongside a label smoothing factor of 0.1 to enhance generalization. For the fine-tuning phase, pretrained model parameters retain a learning rate of 5e-4, whereas newly introduced parameters are updated using a reduced rate of 1e-4. Word dropout is set to 0.1. For the smaller TED dataset, warm-up steps are set to 2000; for larger datasets such as News and Europarl, warm-up is 4000 and dropout remains 0.3 (0.4 for News). All other hyperparameters are consistent with the sentence-level training setup. The batch size is set to 4096, and model parameters are updated every 8 epochs.

(2) mBART Pretraining: We initialize Transformer parameters using mBART-25 [21] (multilingual Bidirectional and Auto-Regressive Transformers with 25 languages) and fine-tune the document-level translation model on this basis. mBART25 is a large-scale multilingual sequence-to-sequence model pretrained on 25 languages including English, Chinese, German, and Russian. It employs various pretraining strategies such as masked language modeling and next sentence prediction, enabling the model to learn rich linguistic representations and contextual understanding. The large configuration is adopted, with 12 encoder and 12 decoder layers, 16 attention heads, and a hidden size of 1024. Following the fine-tuning strategy recommended by Liu et al. [21], we update parameters using Adam ($\theta_1 = 0.9, \theta_2 = 0.998$), with a learning rate of 3e-5, 2500 warm-up steps, attention dropout of 0.1, general dropout of 0.3, label smoothing of 0.2, and no word dropout.

During the decoding phase, the model that achieves the lowest validation loss is chosen for evaluation. Decoding is performed using the beam search strategy, employing a beam width of five. The case-sensitive 4th order sentence level BLEU [22] value is used as an evaluation metric.

5.2 Experimental configuration

The evaluation of the proposed translation model focuses on three standard metrics: BLEU for translation quality, attention entropy for contextual focus, and micro-average accuracy on discourse phenomena such as deixis and ellipsis. BLEU scores are reported for each dataset (TED, News, Europarl), while entropy reduction is visualized through training curves and final attention distributions. Discourse-level evaluation is conducted using a comparative test set containing distractor translations to assess the model's ability to resolve cross-sentence dependencies.

Two different generation environments are chosen in this experiment: parallel (P) means that the audio content and the input text are the same, while non-parallel (NP) means that the audio content and the input text are not the same. The choice of parallel generation environment reduces the influence of text content on speech generation and focuses more on the importance of speech timbre in speech generation. The comparison models chosen in this paper include the generalized high-quality TTS model VITS, the SRD-VC speech conversion model trained with timbre classifiers and adversarial learning, and the STYLE speech generation model improved with the FastSpeech2 model. In terms of efficiency, our model increases training time and memory usage by about 15–20% compared to G-Transformer, but remains fully trainable on two RTX 3090 GPUs.

5.3 Evaluation metrics

To ensure clarity and reproducibility, we formally define the metrics used to evaluate our proposed method.

(1) BLEU score

The BLEU (Bilingual Evaluation Understudy) metric [2] measures n-gram precision between a candidate translation and one or more reference translations. It is computed as:

$$\text{BLEU} = \text{BP} \cdot \exp(\sum_{n=1}^{N} w_n \log p_n), \quad (12)$$

where p_n is the modified n-gram precision, w_n is the weight assigned to n-grams (usually uniform, w_n = 1/N), and BP is the brevity penalty defined as:

$$BP = \begin{cases} 1 & if\ c > r \\ exp(1 - \frac{r}{c}) & if\ c \le r \end{cases} \quad (13)$$

with c the length of the candidate translation and r the length of the reference.

(2) Attention entropy

To measure the sharpness of attention distributions, we compute the Shannon entropy for each attention vector $a = \{a_1, a_2, \ldots, a_n\}$:

$$H(a) = -\sum_{i=1}^{n} a_i \log a_i, \quad (14)$$

where a_i denotes the normalized attention weight for token i. Lower entropy indicates more focused attention, while higher entropy reflects diffuse or noisy distributions.

(3) Micro-average accuracy for discourse phenomena

We adopt the micro-average accuracy metric to evaluate handling of discourse phenomena such as deixis and ellipsis. Given a set of classification decisions across all instances:

$$\text{Micro-Accuracy} = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} (TP_i + FN_i)}, \quad (15)$$

where TP_i and FN_i are the true positives and false negatives for class i, respectively, and K is the total number of classes. This formulation ensures that each prediction contributes equally, making it suitable for imbalanced test sets.

With these metrics, we evaluate translation quality (BLEU), contextual focus (entropy), and discourse-level accuracy (micro-average).

5.4 Comparative results

In this paper, three other systems are constructed for comparison:

(1) Sentence-level Transformer: A typical sentence-level neural machine translation model, Transformer, is modeled using a full multi-head attention mechanism. For the training setup of mBART pre-training, this paper uses sentence-level data sliced and diced from the dataset to fine-tune it for sentence-level translation.

(2) G-Transformer: constrains the lower-level attention of the transformer model to the current sentence and allows the higher-level model to model the whole pseudo-document and the current sentence separately. 3.

(3) Attention Calibration[23]: Attention calibration algorithm designed for sentence-level neural machine translation. Same as the attention alignment algorithm proposed in this paper, this paper extends and applies it to the document-level encoder-decoder attention of GTransformer.

Table 4: Experimental results for different systems (BLEU)

System	TED BLEU	News BLEU	Euro BLEU	Mean BLEU	Avg. Entropy
Two-stage training					
Sentence-level Transformer [6]	24.79	25.28	31.33	27.13	3.92
G-Transformer [8]	25.10 (+0.31)	25.58 (+0.30)	32.34 (+1.01)	27.67 (+0.54)	3.75
Attention Calibration [23]	24.97 (+0.18)	25.20 (−0.08)	32.41 (+1.08)	27.53 (+0.40)	3.78
G-Transformer + Alignment (no mesh)	25.28 (+0.49)	25.82 (+0.54)	32.45 (+1.12)	27.85 (+0.72)	3.61
Attention Alignment + Mesh (Proposed)	25.47 (+0.68)	26.09 (+0.81)	32.67 (+1.34)	28.08 (+0.95)	3.45
mBART pre-training					
Sentence-level Transformer [6]	27.75	29.91	–	28.83	3.70
G-Transformer [8]	28.11 (+0.36)	30.29 (+0.38)	–	29.20 (+0.37)	3.58
Attention Calibration [23]	28.02 (+0.27)	30.25 (+0.34)	–	29.14 (+0.31)	3.60
Attention alignment + G-Transformer	28.33 (+0.58)	30.53 (+0.62)	–	29.43 (+0.60)	3.47

The results of the four systems under the two-stage training configuration are shown in the upper half of Table 4, while results under the mBART pretraining configuration are shown in the lower half. BLEU improvements over the sentence-level Transformer baseline are indicated in parentheses. Due to memory limitations, we were unable to run the mBART configuration on the Europarl dataset. Future work will address this by either training on a reduced Europarl subset or applying memory-efficient techniques such as gradient checkpointing to enable full-scale mBART evaluation.

Under the two-stage training setup, the G-Transformer improves BLEU scores by an average of +0.54 over the sentence-level Transformer across the three datasets. When Attention Calibration is applied, performance gains are inconsistent and, in some cases, slightly lower. In contrast, integrating our proposed attention alignment method into G-Transformer yields a further average BLEU improvement of +0.95, highlighting the effectiveness of our approach. To verify the robustness of these gains, we conducted paired bootstrap resampling (1,000 samples) between our method and the baselines. The improvements on TED (+0.68), News (+0.81), and Europarl (+1.34) are statistically significant at $p < 0.01$, with 95% confidence intervals excluding zero.

To further isolate the contributions of each component, we introduced an additional baseline: G-Transformer with attention alignment but without the mesh decoder. This variant retains label-guided attention alignment while removing mesh-based feature propagation. On the News dataset, it achieved a BLEU score of 25.82, compared to 26.09 for the full model with the mesh decoder. Similarly, on TED, it reached 25.28 (vs. 25.47 full), and on Europarl 32.45 (vs. 32.67 full). These results suggest that attention alignment accounts for the majority of the performance gain, while the mesh decoder provides a consistent incremental improvement of about +0.2 BLEU across datasets.

For the mBART pretraining setup, the G-Transformer outperforms the sentence-level Transformer

by an average of +0.37 BLEU on TED and News. Attention Calibration yields negligible improvement. Our method further increases the average BLEU to +0.60, though the margin is narrower compared to the two-stage setup. Nevertheless, paired tests show these gains remain statistically significant at $p < 0.05$, indicating the continued utility of our approach even with strong pretrained models.

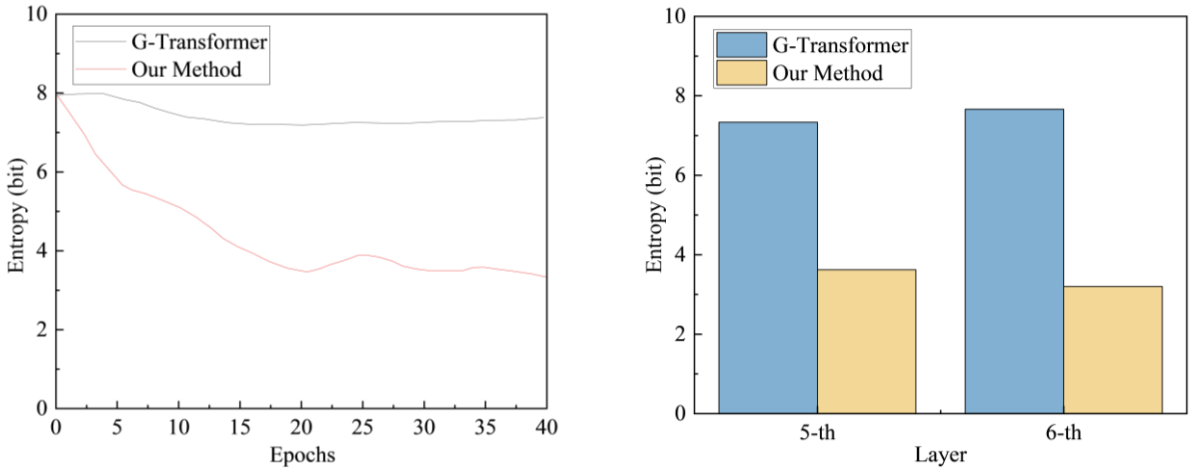
5.5 Ablation experiment

This section uses the two-stage training setup to evaluate three aspects: the impact of hyperparameter α, changes in attention entropy, and performance on discourse-level phenomena.

(1) Effect of α on Entropy and BLEU

On the News dataset, we varied α (Eq. 11) to observe its effect on attention entropy and BLEU score. As shown in Figure 4.4, increasing α initially reduces entropy and raises BLEU, until peaking at $\alpha = 1.5$. Beyond that, BLEU drops as entropy increases. This supports our assumption that emphasizing only a few key contextual tokens improves translation quality.

(2) Attention Distribution



(a) Attention training process (b) Distribution of Results

Figure 5: Comparison of type attention entropy values

Figure 5(a) shows the entropy trend during training. G-Transformer maintains high entropy throughout, indicating noisy context attention. In contrast, our method sharply reduces entropy in the first 20 epochs and remains low. Figure 5(b) confirms our model's more focused attention after convergence, suggesting stronger confidence in selecting relevant context.

(3) Translation of Discourse Phenomena

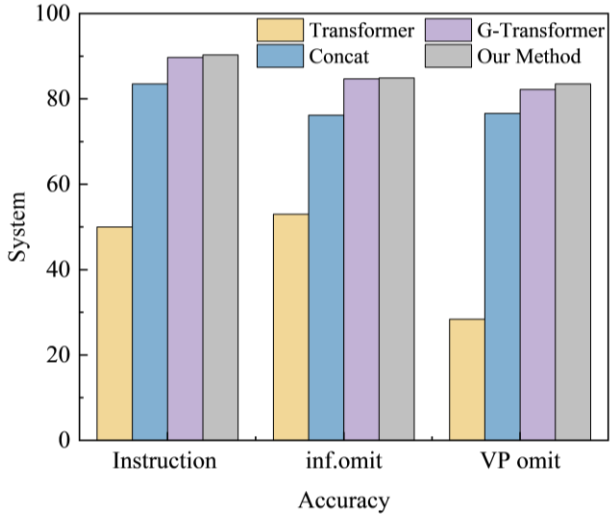


Figure 6: Accuracy of translation of discourse phenomena

We further test how well the model handles discourse-level constructs like deixis and ellipsis using a benchmark evaluation set from Voita et al. The task involves identifying the correct translation among distractors. Models are trained first on 6M sentence-level pairs, then 1.5M document-level samples. Compared to baselines, our method significantly improves accuracy on indicative and VP omission cases, highlighting its strength

in modeling long-range dependencies and sentence relations.

The experimental results and ablation studies presented above confirm that the proposed method offers meaningful improvements over previous document-level NMT approaches. Compared to the sentence-level Transformer and G-Transformer baselines shown in Table 1, our model consistently yields higher BLEU scores: +0.68 on TED, +0.81 on News, and +1.34 on Europarl under two-stage training. Notably, these gains are accompanied by a significant reduction in attention entropy, indicating more focused and semantically meaningful attention distributions.

The attention alignment mechanism is the primary driver of this entropy reduction. By training a lightweight context detector that supervises attention weight adjustments, the model learns to prioritize only the most relevant context words. This targeted supervision improves alignment accuracy and reduces noise from unrelated tokens—issues that prior methods like Attention Calibration only partially addressed. Among the architectural components, the G-Meshed decoder plays a vital role in facilitating cross-sentence information flow, while the label-based positional constraints help guide attention alignment during training. Ablation results indicate that while both components contribute, the attention alignment module has the most direct impact on entropy and BLEU improvements.

Nonetheless, there are limitations. Due to memory constraints, we were unable to train the mBART-based model on the Europarl dataset. This restricts our ability to validate performance on very large corpora and may impact generalizability on high-resource settings. Future work should explore more memory-efficient training strategies or distribute fine-tuning to address this.

6 Conclusion

In this paper, we proposed a novel document-level neural machine translation method that integrates multi-scale wavelet feature fusion with an improved G-Meshed-Transformer architecture. To address the limitations of sentence-level translation in capturing long-range dependencies and document-level context, we introduced a Balanced Contextual Slicing strategy to enhance contextual learning without increasing model complexity. Furthermore, we incorporated a mesh structure in the decoder and designed a document-level attention alignment algorithm that explicitly guides attention to informative contextual tokens. Extensive experiments on three benchmark datasets (TED, News, Europarl) under two training configurations (two-stage and mBART pre-training) demonstrate the effectiveness of our approach. The proposed model consistently outperforms strong baselines in BLEU scores, particularly achieving an average gain of +0.95 BLEU under two-stage training and +0.60 BLEU with mBART pre-training. Ablation studies validate the role of alignment entropy optimization and show that our method enhances attention concentration and improves the translation of discourse phenomena such as deixis and ellipsis.

In summary, our method provides a general and effective framework for improving document-level translation consistency by leveraging both enhanced contextual encoding and guided attention. Future work may explore cross-lingual extensions and dynamic discourse-aware attention mechanisms to further improve robustness across language pairs and domains.

References

[1] Zhang, J., & Zong, C. (2020). Neural machine translation: Challenges, progress and future. Science China Technological Sciences, 63(10), 2028–2050.

[2] Dabre, R., Chu, C., & Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. ACM Computing Surveys (CSUR), 53(5), 1–38.

[3] Cho, K., Van Merriënboer, B., Gulcehre, C., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[4] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems (NeurIPS).

[5] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[6] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 30.

[7] Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451.

[8] Bao, G., Zhang, Y., Teng, Z., et al. (2021). G-transformer for document-level machine translation. arXiv preprint arXiv:2105.14761.

[9] Tu, Z., Liu, Y., Shi, S., et al. (2018). Learning to remember translation history with a continuous cache. Transactions of the Association for Computational Linguistics, 6, 407–420.

[10] Ji, B., Zhang, Z., Duan, X., et al. (2020). Cross-lingual pre-training-based transfer for zero-shot neural machine translation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(1), 115–122.

[11] Maruf, S., Martins, A. F. T., & Haffari, G. (2019). Selective attention for context-aware neural machine translation. arXiv preprint arXiv:1903.08788.

[12] Tu, Z., Liu, Y., Shi, S., et al. (2018). Learning to remember translation history with a continuous cache. Transactions of the Association for Computational Linguistics, 6, 407–420.

[13] Cheng, Y. (2019). Joint training for pivot-based neural machine translation. In Joint Training for Neural Machine Translation (pp. 41–54). Springer, Singapore.

[14] Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201.

[15] Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709.
[16] Edunov, S., Ott, M., Auli, M., et al. (2018). Understanding back-translation at scale. arXiv preprint arXiv:1808.09381.
[17] Lin, Z., Pan, X., Wang, M., et al. (2020). Pre-training multilingual neural machine translation by leveraging alignment information. arXiv preprint arXiv:2010.03142.
[18] Pan, X., Wang, M., Wu, L., et al. (2021). Contrastive learning for many-to-many multilingual neural machine translation. arXiv preprint arXiv:2105.09501.
[19] Maruf, S., Martins, A. F. T., & Haffari, G. (2019). Selective attention for context-aware neural machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers), 3092–3102.
[20] Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1 (Long Papers), 1715–1725.
[21] Liu, Y., Gu, J., Goyal, N., et al. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8, 726–742.
[22] Papineni, K., Roukos, S., Ward, T., et al. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 311–318.
[23] Lu, Y., Zeng, J., Zhang, J., et al. (2021). Attention calibration for Transformer in neural machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) and the 11th International Joint Conference on Natural Language Processing (IJCNLP), Volume 1 (Long Papers), 1288–1298.
[24] Fang, Q., & Feng, Y. (2022). Neural machine translation with phrase-level universal visual representations. arXiv preprint arXiv:2203.10299.