A Dual-Channel Transformer Model for Translationese Detection Integrating Syntactic and Lexical Features

QIANQIAN TIAN*
School of Foreign Languages
Zhengzhou Shengda University
Zhengzhou 451191
CHINA

*Corresponding Author

Abstract: - With the growing use of neural machine translation and LLM-based translation, translationese detection has become important for assessing translation naturalness and source-language interference. Existing methods often rely on explicit features or unified pretrained representations, giving limited attention to the joint modeling of lexical deviation, syntactic transfer, and local evidence. This paper proposes a Transformer-based dual-channel framework that separately encodes lexical and syntactic information, models their interaction through cross-channel fusion, and introduces local auxiliary discrimination to support sentence-level prediction. Experiments on the WMT Chinese news test set and the UNPC Chinese formal corpus show that the proposed method improves performance over the strongest syntax-enhanced baseline on WMT and achieves comparable results on UNPC. The framework demonstrates stable cross-domain performance and can support machine translation filtering, post-editing assistance, and generated-text style identification.

Key-Words: - Syntactic Transfer; Lexical Deviation; Dual-Channel Modeling; Cross-Channel Fusion; Local Auxiliary Discrimination; Translationese Detection; Chinese Text Classification; Machine Translation Quality Assessment

Received: May 13, 2026. Revised: July 2, 2026. Accepted: August 6, 2026. Published: November 9, 2026.

1 Introduction

With the spread of neural machine translation (NMT) and large language model (LLM)-based translation, translation quality assessment is no longer limited to adequacy and fluency [1]. Translationese detection has gradually attracted attention from scholars of language and communication worldwide [2]. Translationese refers to the transfer of source-language conventions into the target language, often reflected in lexical selection tendencies and residual syntactic structures [3]. Therefore, computational identification of translationese is needed to analyse the naturalness of machine-generated translations and to support translation quality evaluation, post-editing, and generated text style identification [4]. Although previous studies have shown that translationese has relatively stable and learnable features, how to jointly model lexical deviation and syntactic transfer in the Chinese context remains insufficiently studied [5]. Existing models also provide limited support for linking lexical cues, syntactic structures, and local translationese evidence within the same training framework.

The existing translation quality detection methods initially relied mainly on artificially designed language features, such as word selection and syntactic structure. Although these methods were intuitive and easy to explain, they had difficulty truly describing the relationship between words and syntax, and their effectiveness was unstable when applied in different fields [6]. Later, methods gradually shifted to pre-trained models, using overall text representations for judgment. This led to a significant improvement in effectiveness, but most methods still represented the entire sentence with a single vector, without separately modeling the syntactic-related information [7]. Some research employed continuous scoring and local fragment analysis to more precisely locate the problems. These methods did not effectively integrate the various components, such as word contributions, local fragments, and scoring, lacking a unified explanation path between them [8]. Therefore, there is still a problem: how to deal with syntactic structure, word relationships, and local evidence simultaneously within a model.

To address these issues, this paper has improved the existing translation language detection framework. Firstly, a dual-channel heterogeneous encoding structure is constructed to model lexical

information and syntactic information separately; secondly, a cross-channel interaction and fusion mechanism is introduced to model the relationship between lexical deviations and syntactic transfer; thirdly, a local auxiliary discrimination module is designed to identify high-risk local segments and introduce local evidence into the overall prediction process. On this basis, cross-channel consistency constraints are also added to enable lexical representation, syntactic representation, local prediction, and sentence-level classification to be jointly optimized within the same framework.

This paper proposes a translation sentence detection model that integrates linguistic information. To a certain extent, it compensates for the shortcomings of traditional text classification methods in terms of structural sensitivity and local interpretability. This paper aims to conduct translation quality assessment by integrating various linguistic information. It emphasizes the interaction between vocabulary and syntax, as well as the role of local fragment information in the final judgment. The experimental results show that this method has relatively stable performance on multiple datasets and indicate that this modeling approach, which combines local evidence with overall prediction, is effective. From an application perspective, it can be used for filtering machine translation outputs, assisting post-editing, and identifying the style of generated text.

2 Related Work

2.1 Application Scenarios and Challenges

Translationese detection, which involves identifying non-native features left by the source language migration in the target language text, has become an important topic at the intersection of translation technology and natural language processing [9]. Besides evaluating the quality of machine translation and the effects of post-editing, it can also be used for evaluating the naturalness of generated text, cross-language content review, and style control in multilingual content production [10]. In addition, direct assessment of translationese can broaden the evaluation of unnatural expressions [11].

These studies typically employ synthetic texts generated by different systems, parallel corpora of source and target languages, as well as mixed data of human translation and machine translation [12]. However, during the actual construction and use of the data, there are still some shortcomings. For example, the degree of matching between different data is still not ideal, which is particularly evident in the Chinese context [13]. This is mainly reflected in inconsistent localization processing and a lack of annotation and analysis capabilities for fine-grained segments, which also limits its support for local translation phenomena to some extent [14].

As for evaluation methods, most existing studies still use accuracy, Macro-F1, AUC, and other metrics to evaluate classification performance. Correlation or ranking-based metrics are often employed for continuous prediction tasks. These metrics do not tell us what type of language signal the model is using. The information given is still rather indirect and has little explanatory power for specific language phenomena.

2.2 Overview of Mainstream Approaches

Previous research is currently mainly conducted along the following choices.

One type of approach mainly relies on language features or pre-trained representations for modeling [15]. Traditional feature methods usually construct representations from aspects such as word usage, distribution of function words, word order sequence, and syntactic complexity. The advantage is that they are intuitive and highly interpretable. However, they still have limitations in capturing the interaction between word selection and syntactic structure, and are also relatively sensitive to fluctuations in different genres or data sources.

Another type of approach relies more on pre-trained encoders, making judgments by learning the differences between the original text and the translated text in the representation space [16]. These methods usually employ overall text representations, and syntactic information is mostly incorporated implicitly rather than being modeled separately. Therefore, the specific role of syntactic structure in translation quality assessment has not been fully explored.

In addition, some studies focus on continuous evaluation and local interpretation [17]. These methods usually represent the degree of translation with a continuous numerical value, and through analysis at the paragraph or segment level, observe which parts of the sentences are closer to the features of the target text [18]. This approach has advanced in fine-grained analysis, but different information is still handled separately. Currently, there is no framework that can integrate these pieces of information simultaneously.

2.3 Closely Related Work

The most relevant work related to this study mainly proceeds along the following lines.

The continuous evaluation method estimates the degree of translationese through probability

differences or distribution comparisons, and converts the task into continuous value representation [19]. Such methods provide important references for task modeling. This study concentrates on the modeling at the structural level. For instance, based on sentence-level judgment, it also takes into account the influence of lexical deviations and syntactic changes on the model's decision-making.

Regarding the analysis of explicit features in both Chinese and English contexts, related studies have constructed a multi-level language feature system based on relatively standardized texts [20]. These efforts identify the distinctive features of the translated language and emphasize the importance of syntactic and lexical cues, but their focus remains on the selection and shallow combination of explicit features.

Focusing on interpretable detection and local attribution, post-event explanation methods identify high-contributing terms or key segments, progressing from whether translation orientation can be detected to how to detect it [21]. This kind of analysis occurs more after the prediction, and involves less consideration of how these judgments are formed within the model. Thus, this study proposes a dual-channel representation and fusion framework that integrates lexical and syntactic information and introduces local translation signals into the training process as auxiliary constraints.

2.4 Summary and Research Gap

In short, the existing research has achieved considerable results in areas such as explicit feature design, pre-training representation learning, continuous scoring, and local interpretability analysis. However, these methods often approach the problem from a single perspective, such as relying on surface statistical features, unified text representations, or local explanation signals, and the modeling of translation deviations remains relatively fragmented. How to simultaneously capture syntactic changes, lexical deviations, and local translation evidence within the same framework remains an issue that needs further resolution.

This limitation is more evident in the Chinese context. Syntactic information is rarely modelled explicitly; the relationship between lexical and structural cues remains insufficiently represented, and local translationese phenomena are seldom incorporated into training supervision.

This study integrates syntactic cues, lexical signals, and local translationese states into a unified detection framework. The proposed method preserves lexical-syntactic heterogeneity through dual-channel encoding, models their interaction through cross-channel fusion, and uses local auxiliary signals to support sentence-level prediction.

Accordingly, this study focuses on three linked issues: explicit lexical-syntactic representation, cross-level feature interaction, and local auxiliary learning for Chinese translationese detection.

3 Methodology

3.1 Problem Formulation

We formulate translationese detection as a sentence-level classification problem under joint lexical-syntactic conditions and augment it with local auxiliary prediction to support the global prediction. Let the target-language sentence sample be denoted by x, and both a lexical view and a syntactic view of x are used as input:

$$x=(x^{lex},x^{syn}),x^{lex}=(w_1,\ldots,w_n),x^{syn}=(s_1,\ldots,s_n). \quad (1)$$

Here, x^{lex} denotes the token sequence, capturing contextual semantics and lexical stylistic information, while x^{syn} represents the corresponding syntactic observations, such as part-of-speech tags, dependency relations, or other structural features. After embedding, the lexical and syntactic inputs can be represented as $E^{lex} \in \mathbb{R}^{n\times d_l}$ and $E^{syn} \in \mathbb{R}^{n\times d_s}$, where n is the sequence length and d_l, d_s denote the lexical and syntactic feature dimensions, respectively. This form suggests that translationese is caused not only by deviations in surface vocabulary but also by the combined effects of lexical deviation and syntactic transfer; therefore, the input should be regarded as a joint dual-view representation.

The task output is defined as a joint variable comprising global and local states:

$$y=(y^g,y^l_{1:m}),y^g\in\{0,1\} \text{ or } [0,1],y^l_j\in[0,1],\ j=1,\ldots,m. \quad (2)$$

Here, y^g denotes whether the sentence exhibits translationese (or its intensity), and $y^l_{1:m}$ represents the translationese intensity of each of the m local segments obtained by partitioning the sentence. In practice, local states are typically approximated using weak supervision signals. Thus, the local labels are not treated as independent ground-truth annotations, but as auxiliary signals that guide the model to capture localized translationese evidence.

We assume that samples are independently and identically distributed according to a joint distribution:

$$(x,y)\sim P(X^{lex},X^{syn},Y^g,Y^l_{1:m}), \quad (3)$$

and denote the training set as $D_{train}=\{(x_i,y_i)\}_{i=1}^{N}$. The ideal learning objective can then be expressed as:

$$f^*=\arg\min_f E_{(x,y)\sim P}[L_g(f^g(x),y^g)+\lambda L_l(f^l(x),y^l_{1:m})]. \quad (4)$$

Here, f^g and f^l denote the global and local output functions, respectively; L_g and L_l represent the sentence-level and local-level loss functions; and λ is a balancing coefficient. This objective keeps the main task centred on sentence-level classification while using local supervision to regularize the representation learning process.

Unlike traditional sentence-level classification, this formulation jointly considers the global translationese label and local translationese intensities. The global decision is therefore expected to be supported by both lexical-syntactic evidence and accumulated local patterns. This form provides a general basis for the following dual-channel model, cross-channel interaction, and local auxiliary learning.

3.2 Overall Framework

The overall framework is shown in Figure 1 and mainly consists of three parts: input, encoding and interaction processing, and output. The input includes the word sequence of the target language text and its corresponding syntactic information; the middle part consists of three modules, namely dual-channel representation learning, cross-channel interaction fusion, and a local auxiliary discrimination module; the output includes sentence-level translation prediction results and local segment-level translation distributions, which are used to supplement the overall judgment.

The framework consists of three stages: heterogeneous input encoding, cross-channel interaction and fusion, and joint output generation. At the first stage, lexical and syntactic information are treated as two related but heterogeneous inputs, and their discriminative representations are learned separately. At the second stage, a unified interaction mechanism builds dependency relationships between the two types of features and captures both lexical deviation and syntactic transfer. Finally, the fused representation is used to obtain sentence-level predictions and local auxiliary outputs for global classification and local pattern recognition.

The main advantage of the framework is that it links representation learning, feature interaction, and local evidence modelling within the same detection process. The local auxiliary outputs jointly constrain the learning of the fused representation with the main task. At the same time, the cross-channel interaction mechanism models the bidirectional dependency between lexical and syntactic representations to form joint lexical-syntactic evidence.

The first module encodes lexical sequences and syntactic observations into two parallel representations. The second module models the dependency between the two feature channels through cross-channel interaction and gated fusion. The third module produces segment-level auxiliary predictions and a sentence-level output, thereby helping the model capture localized translationese phenomena. Together, these three parts form the proposed dual-channel translationese detection framework.

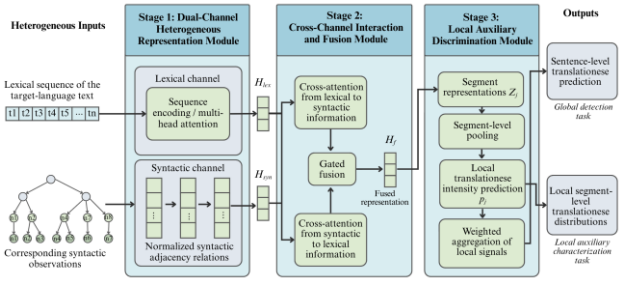


Fig. 1: Overall framework of the proposed dual-channel translationese detection model
Source: created by the authors

3.3 Module Descriptions

3.3.1 Dual-Channel Heterogeneous Representation Module

Previous methods of modelling translationese often treat the input as a unified text sequence, with syntactic information ignored or only implicitly encoded. Therefore, models are more likely to capture lexical traces but fail to adequately recognise word-order transfer and structural residues. This module separates lexical and syntactic cues at the input stage, thereby addressing the imbalance between lexical and syntactic signals.

The two components are a sequence encoder and a structure propagator. The lexical channel models long-range semantic and stylistic dependencies using self-attention, while the syntactic channel propagates structural information according to dependency adjacency relations. Lexical deviation and syntactic transfer occur at different linguistic levels, and early fusion may cause structural bias to be concealed by surface semantic patterns. The structure is shown in Figure 2.

Let E^{lex} and E^{syn} denote the lexical embeddings and initial syntactic representations, respectively. Here, n is the sequence length and d is the hidden representation dimension after linear projection. The two channels are updated as follows:

$$H^{lex}=MHA(E^{lex},E^{lex},E^{lex}),H^{syn}=\sigma(\widetilde{A}E^{syn}W_s), \quad (5)$$

where $\widetilde{A}$ is the normalized syntactic adjacency matrix, W_s is a learnable parameter matrix, and $\sigma(\cdot)$ denotes a nonlinear activation function. The resulting representations H^{lex} and H^{syn} are used as inputs for subsequent cross-channel interaction. The lexical

channel captures contextual and stylistic dependencies, while the syntactic channel preserves dependency-based structural propagation. The lexical self-attention layer has a computational complexity of $O(n^2 d)$, while syntactic propagation has a complexity of $O(|E|d)$, where $|E|$ denotes the number of dependency edges.

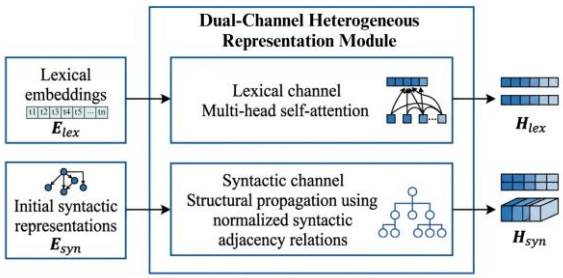


Fig. 2: Dual-channel heterogeneous representation module for lexical–syntactic parallel encoding
Source: created by the authors

3.3.2 Cross-Channel Interaction and Fusion Module

Although the two-channel encoding retains different types of information, it does not automatically associate lexical and syntactic cues. Simple concatenation is insufficient to determine whether a lexical anomaly is related to syntactic transfer. This module models cross-level feature interaction between lexical and syntactic representations.

The module follows a cross-attention and gated-fusion design. Each representation is conditionally reweighted by the other, and adaptive fusion is performed through a gating unit. The modelling motivation is that translationese can be viewed as a cross-level inconsistency between lexical choice and syntactic organization. Conditional interaction therefore helps capture dependencies between lexical deviation and syntactic transfer. This module does not force the two representations to become identical; instead, it preserves useful cross-channel discrepancies. The structure is shown in Figure 3.

Given the outputs H^{lex} and H^{syn} from the previous module, the fused representation is defined as:

$$H^f = G \odot Attn(H^{lex}, H^{syn}, H^{syn}) + (1-G) \odot Attn(H^{syn}, H^{lex}, H^{lex}) \quad (6)$$

$$G = \sigma([H^{lex}; H^{syn}]W_g + b_g) \quad (7)$$

In terms of specific implementation, $Attn(\cdot)$ represents cross-attention calculation, while W_g and b_g are learnable parameters. $[\cdot;\cdot]$ indicates vector concatenation, and $\odot$ represents element-wise multiplication. The core of the overall design lies in enabling information between different channels to interact with each other, rather than simply concatenating them. Therefore, the fusion process is actually an interactive mechanism with weight adjustment, used for dynamic allocation between lexical and syntactic information. The gate G controls the contribution ratio of information from different directions, avoiding one channel dominating the fusion process and making the representation more balanced.

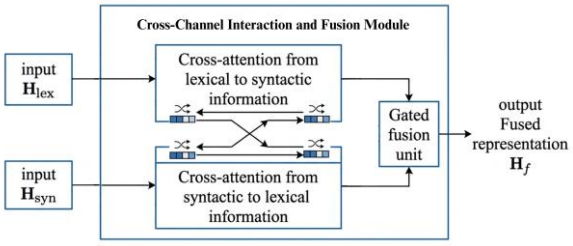


Fig. 3: Cross-channel interaction and fusion module of the proposed method
Source: created by the authors

3.3.3 Local Auxiliary Discrimination Module

Currently, most studies mainly rely on sentence-level classification results or post-hoc attribution analysis after prediction. Therefore, the local translation feature characteristics in the text usually do not directly participate in the model training. This may result in the model being able to recognize the overall translation tendency, but it is difficult to capture the specific language evidence that supports this judgment. This module addresses this issue by introducing local information to assist the model in learning.

This module first assesses the translationese intensity at the segment-level, and then completes the sentence-level prediction by aggregating local signals. This design takes into account that translation phenomena often occur in specific expressions, such as the selection of literal translations, the use of fixed collocations, and residual syntactic patterns. The specific structure is shown in Figure 4.

Let Z_j denote the representation of the j-th segment derived from the fused representation. The local and global outputs are defined as:

$$p_j = \sigma(W_l Pool(Z_j) + b_l), \hat{y}^{loc} = \sum_{j=1}^{m} \alpha_j p_j, y^g = \sigma(w_c \hat{y}^{loc} + b_c), \quad (8)$$

where p_j denotes the translationese intensity of the j-th segment, α_j is the segment weight, W_l, b_l, w_c, and b_c are learnable parameters, and $Pool(\cdot)$ denotes segment-level pooling. The segment weights α_j are normalized across the m segments so that local evidence can be aggregated into a sentence-level signal. Compared with an independent prediction head, this formulation makes the sentence-level

output explicitly dependent on the weighted aggregation of local translationese intensities.

From a probabilistic perspective, this module can be viewed as a lightweight local-global coupling model:

$$P(y, p \mid x) \propto \exp\left(\psi_g(y, H_f) + \sum_{j=1}^{m} \psi_l(p_j, Z_j) - \gamma \sum_{j=1}^{m} \alpha_j (y - p_j)^2\right), \quad (9)$$

where ψ_g and ψ_l are global and local potentials. The last term penalizes inconsistency between the global label and local intensities. Therefore, Eq. (8) can be regarded as a tractable mean-field-style approximation of local-global coupling rather than full CRF inference.

The purpose of this module is not simply to add another prediction head, but to process local translationese information and use it as auxiliary supervision during training. Thus, local predictions provide segment-level evidence, while their weighted aggregation supports the final global decision.

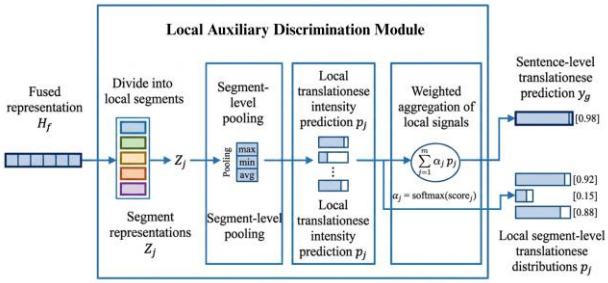


Fig. 4: Local auxiliary discrimination module of the proposed method
Source: created by the authors

3.4 Objective Function and Optimization

To enable joint optimization of the dual-channel heterogeneous representation, cross-channel interaction and fusion, and local auxiliary discrimination modules under a unified objective, we define the overall loss as a weighted combination of a sentence-level supervision term, a local auxiliary term, a cross-channel consistency regularizer, and a local-global consistency term:

$$L = L_g + \lambda_1 L_l + \lambda_2 L_c. \quad (10)$$

Here, L_g denotes the sentence-level translationese detection loss, L_l the local auxiliary loss, L_c the cross-channel consistency regularization term, and L_{lg} the consistency loss between local aggregation and global prediction. The coefficients λ_1, λ_2, and λ_3 control the weights of the auxiliary objectives. They are selected by grid search on the validation set, with $\lambda_1 = 0.5$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.2$.

The sentence-level term optimizes the primary classification task. Let N denote the number of samples, $y_i^g \in \{0,1\}$ the ground-truth label, and $\hat{y}_i^g \in (0,1)$ the predicted probability. The binary cross-entropy loss is:

$$L_g = -\frac{1}{N}\sum_{i=1}^{N} [\, y_i^g \log \hat{y}_i^g + (1-y_i^g)\log(1-\hat{y}_i^g)]. \quad (11)$$

As shown in Eq. (11), this term enforces consistency between predictions and ground-truth labels and serves as the core optimization objective.

The local auxiliary loss introduces the word-level translation signal into the training process. For the i-th sample, the sentence is divided into m segments. Let $p_{ij} \in (0,1)$ represent the predicted translation strength value of the j-th segment, and $z_{ij} \in [0,1]$ represent the corresponding weakly supervised label. Then, the loss function is defined as follows:

$$L_l = \frac{1}{Nm}\sum_{i=1}^{N} \sum_{j=1}^{m} (p_{ij} - z_{ij})^2. \quad (12)$$

Among them, z_{ij} is generated through heuristic rules and is used to approximately represent the local translationese intensity. This auxiliary term merely provides additional local learning signals for the model, enabling it to focus on the translationese features in specific segments while making overall predictions. In the experiment, the number of segments was fixed at m=4 to achieve a balance between the detail of local information and the stability of training.

To further integrate the segment-level prediction results with the sentence-level judgment, the local aggregation prediction is defined as follows:

$$\hat{y}_i^{loc} = \sum_{j=1}^{m} \alpha_{ij} p_{ij}, \quad (13)$$

Among them, α_{ij} represents the segmented weights after normalization. On this basis, the local and global consistency losses are further defined:

$$L_{lg} = \frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i^g - \hat{y}_i^{loc})^2. \quad (14)$$

This term encourages the sentence-level prediction to be supported by accumulated local evidence and makes the relationship between local outputs and global supervision explicit in the objective function.

The cross-channel term constrains lexical and syntactic representations to maintain moderate alignment prior to fusion, without collapsing into identical representations. Let $H_i^{lex} \in R^{n \times d}$ and $H_i^{syn} \in R^{n \times d}$ denote the lexical and syntactic representations of the i-th sample, respectively, and let Pool(·) denote mean pooling. The regularization term is defined as:

$$L_c = \frac{1}{N}\sum_{i=1}^{N} \left(1 - \cos(\mathrm{Pool}(H_i^{lex}), \mathrm{Pool}(H_i^{syn}))\right). \quad (15)$$

Among them, $\cos(\cdot,\cdot)$ represents cosine similarity, and $\rho \in (0,1)$ represents the expected cross-channel similarity level. In this study, ρ is set

to 0.5 based on the performance of the validation set. This parameter can prevent them from being independent of each other or completely redundant. Through Eq. (15), the model can combine the two types of information and retain the characteristics of the lexical and syntactic representations, respectively. We choose cosine similarity as the measurement method because it focuses on the directional relationship between vectors and is not affected by vector scale; thus, eliminating the need to limit the vector norms of lexical representations and syntactic representations to be the same.

We adopt AdamW with an initial learning rate of 2×10^{-5} and a weight decay coefficient of 1×10^{-2}. The batch size is 32, and the model is trained for 20 epochs. Linear warm-up and cosine decay are used, with warm-up accounting for 10% of total training steps. Dropout with a rate of 0.1 is applied after the fusion layer. Model selection is based on validation Macro-F1, and early stopping is applied if validation performance does not improve for three consecutive epochs. The complete objective in Eq. (10) also provides the basis for the following analysis of objective boundedness, optimization stability, and parameter sensitivity.

3.5 Model Properties, Optimization Stability, and Sensitivity Analysis

The objective in Eq. (10) provides a constrained optimization form for the proposed dual-channel model. Since $\hat{y}_i^g$, p_{ij}, and $\hat{y}_i^{loc}$ are generated by sigmoid outputs, they are bounded within $(0, 1)$. With probability clipping $[\epsilon, 1-\epsilon]$, the sentence-level loss is bounded. The local auxiliary and local-global consistency losses are also bounded, while the cosine similarity in L_c lies in $[-1,1]$. Therefore,

$$0 \le L \le -\log \epsilon + \lambda_1 + \lambda_2(1+\rho)^2 + \lambda_3. \quad (16)$$

This indicates that the total objective remains finite when λ_1, λ_2, and λ_3 are finite, so the auxiliary terms do not introduce unbounded penalties.

The update of model parameters is simultaneously influenced by sentence-level supervision, local auxiliary supervision, cross-channel constraints, and local and global consistency constraints. If the four loss terms are smooth with constants β_g, β_l, β_c, and β_{lg}, then the total objective is β-smooth, where $\beta = \beta_g + \lambda_1\beta_l + \lambda_2\beta_c + \lambda_3\beta_{lg}$. For $0 < \eta < 1/\beta$, standard smooth non-convex optimization gives

$$\min_{0 \le t < T} \| \nabla L(\theta_t) \|_2^2 \le \frac{2(L(\theta_0) - L^*)}{\eta(1-\beta\eta/2)T}. \quad (17)$$

As T increases, the optimization process gradually approaches the first-order stable region. This indicates that although a larger auxiliary weight can enhance the local constraint effect, it may also increase the sensitivity of the optimization process. Thus, it is necessary to reasonably set the learning rate.

The local and global consistency terms further connect the local prediction results and the sentence-level judgment, enabling local information to participate in model training through both the local auxiliary loss and the consistency loss. From a generalization perspective, this bounded objective function establishes the relationship between empirical risk, expected risk, sample size, and the complexity of the dual-channel assumption, providing a statistical learning theory-based explanation for model optimization; however, it does not mean that it has universal advantages in all scenarios.

To analyze the model stability with respect to input variations, let pθ(x) represent the sentence-level prediction result. If the Lipschitz constants of the vocabulary encoder, syntactic propagator, fusion layer, and classifier are K_{lex}, K_{syn}, K_f, and K_o respectively, then:

$$\| p_\theta(x + \Delta x) - p_\theta(x) \|_2 \le K_o K_f (K_{lex} + K_{syn}) \| \Delta x \|_2. \quad (18)$$

This shows that prediction variation is impacted by the effective Lipschitz constant of the composed model. The consistency term can reduce the unstable changes between different representations, ensuring a certain degree of coordination of lexical information, syntactic information, and local evidence during the training process.

Among them, the coefficients λ_1, λ_2, and λ_3 are respectively used to adjust the degree of influence of local weak supervision, the consistency of lexical and syntactic information, and the local prediction and overall judgment. If these parameters are set too large, the model may overly focus on local information and neglect the respective roles of lexical and syntactic information. The target similarity ρ is used to control the degree of information consistency maintained between the two channels. These parameters are selected through the validation set to achieve a balance between model performance, representation stability, and the training process.

4 Experiments and Results

4.1 Experimental Setup

This paper executes experiments based on the main dataset and the auxiliary verification dataset. The data division is shown in Table 1. Among them, the WMT Chinese news test set is used as the main evaluation data, and the remaining data is used for the overall performance analysis of the model and some

ablation experiments. The auxiliary dataset is the Chinese part of the UN Parallel Corpus (UNPC); it will be used for cross-genre validation and to test how well explicit syntactic models work on formal texts. The former is general news-style written Chinese and is suitable for investigating lexical correspondence tendencies and local structural deviations in translated texts. The latter is based on official United Nations documents and thus contains more formal and legalistic expressions, longer sentences, nominalized constructions, and standardized syntactic patterns; it is therefore more suitable for testing sensitivity to syntactic transfer.

The WMT Chinese news test set is based on the WMT 2021 News Translation Task and includes the WMT18, WMT19, and WMT20 subsets; there are a total of 3,960 Chinese-English parallel sentences. The UN Parallel Corpus v1.0 is a collection of publicly available official UN records and meeting documents, containing human-translated content from 1990 to 2014, and is aligned at the sentence level. The Chinese-English part has about 91,028 aligned files and 15,886,041 bilingual sentence pairs. Instead of using these resources as pre-packaged datasets for translationese detection directly, we create sentence-level binary classification samples. Translated texts refer to Chinese translations that explicitly show their sources, and original texts refer to native Chinese texts. We have controlled for genre, length, and granularity in the construction of the dataset as much as possible and used a consistent splitting rule to reduce source bias and minimise near-duplicate leakage.

A common pipeline is used to pre-process both datasets for text cleaning, sentence division, character normalization, length filtering and duplicate removal. The lexical channel is generated by a pre-trained Chinese tokenizer in the form of token sequences, and at the same time, the syntactic channel constructs normalized syntactic adjacency matrices through dependency parsing of this text to ensure strict alignment between the two inputs. As UNPC has many long sentences and paragraph-level sections, sentence segmentation will be performed prior to model training.

Table 1. Overview of Experimental Datasets

Dataset	Purpose	Scale	Description
WMT Chinese News Test Set	Core evaluation, main comparison, ablation	3,960 parallel sentence pairs	From WMT 2021 News Translation Task (WMT18–20)
UNPC Chinese Formal Text	Cross-genre validation	~91,028 files; 15.9M sentence pairs	UN official documents (1990–2014), sentence-aligned

Source: created by the authors

Experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU, with Python 3.10, PyTorch 2.1, and CUDA 12.1. All neural baselines are trained under comparable settings: hidden size of 768, maximum sequence length of 256, batch size of 32, and AdamW optimizer with an initial learning rate of 2×10^{-5} and weight decay of 1×10^{-2}. The maximum number of training epochs is set to 20, with early stopping based on validation performance. The overall loss follows Eq. (10), with $\lambda_1 = 0.5$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.2$. The target similarity level in $L_c$is set to $\rho = 0.5$according to validation performance. All neural baselines are trained under the same data split, training budget, optimizer, and early-stopping rule to ensure fair comparison.

The evaluation metrics include Accuracy, Macro-F1, Expected Calibration Error (ECE), and Fusion Gain. Accuracy and Macro-F1 evaluate classification performance, ECE measures calibration quality, and Fusion Gain quantifies the additional error reduction obtained by dual-channel fusion over single-channel variants. The subsequent analysis mainly uses macro F1 as the core evaluation metric and further examines model prediction stability and the impact of the dual-channel design by combining ECE and fusion gain.

4.2 Baselines

To provide an all-around assessment of the proposed method, we compare the proposed method with four representative baselines, such as traditional feature-based methods, classical neural models, pretrained models, and syntax-enhanced architectures.

First, an SVM is used with hand-crafted linguistic features as the traditional baseline [22]. This approach uses explicit linguistic features such as function words, part-of-speech sequences, and lexical richness for classification with an SVM classifier; this is the traditional feature-engineering model for translationese detection. Second, a BiLSTM-based text classification model is used as the classical neural baseline [23]. A bidirectional LSTM is employed to encode the input sequence, followed by pooling and a classification layer; it performs as well as other typical deep sequential models on this task. Third, a fine-tuned RoBERTa model is used as the pre-trained baseline [24]. Based on a pre-trained encoder for Chinese, this model is used to perform end-to-end fine-tuning of sentence-level classification and adopts the widely used unified representation paradigm in Chinese text classification. Finally, a syntax-aware Transformer is added as a syntax-enhanced baseline [25]. This model explicitly adds syntactic structural information to the Transformer and serves as the strongest syntax-enhanced comparison method in the model.

The four representative paradigms of baselines for translationese detection are: explicit feature-based methods, classical deep sequence models, unified pretrained representations, and syntax-enhanced models. All neural networks in the baseline should be trained with the same computational budget and evaluated using the same methods for fair comparison, and the traditional baseline should follow a standard feature construction process. The following experiments will compare the overall classification performance, robustness, and improvements after adding dual-channel fusion.

4.3 Quantitative Results

To assess the effect of the proposed method, it is compared with the four categories of baseline approaches on the WMT Chinese news test set and the UNPC Chinese formal text corpus. The results are as follows (Table 2 (Appendix)). All experiments are conducted with five independent runs, and the mean ± standard deviation is reported.

As shown in Table 2 (Appendix), the proposed method achieves the best performance on WMT. Compared with the Syntax-Aware Transformer, it improves Accuracy from 89.4 to 90.0 and Macro-F1 from 88.8 to 89.6, while reducing ECE from 0.037 to 0.034. The improvement in Macro-F1 is statistically significant, and the Accuracy gain reaches the 0.05 significance level; the ECE reduction shows the same direction but does not reach statistical significance. These results indicate that the proposed method improves news-domain translationese detection and produces more balanced classification performance.

On UNPC, the proposed method remains comparable to the Syntax-Aware Transformer. Its Accuracy is 90.8, slightly lower than the 91.0 of the Syntax-Aware Transformer; its Macro-F1 is 90.8, marginally higher than 90.7; and its ECE decreases from 0.040 to 0.038. None of these differences are statistically significant. This result suggests that the advantage of dual-channel interaction is weaker in formal-text settings, where syntax-enhanced baselines already show strong performance. Overall, the proposed method shows clearer improvement on the main WMT dataset and comparable performance on the auxiliary UNPC dataset. The effect of dual-channel modelling and cross-channel interaction is therefore more pronounced in news-style text than in formal text.

Fusion Gain is introduced to quantify how effectively the fusion model reduces errors relative to the two single-channel variants. Fusion Gain is 6.4% ± 0.8% for WMT and 3.5% ± 0.7% for UNPC, indicating that lexical and syntactic cues are more complementary in news text, while their complementary effect is weaker in formal text. In addition to classification performance, model size and computational efficiency are also analyzed, as shown in Table 3.

Table 3. Efficiency Comparison

Method	Parameters (M)	FLOPs (G)	Training Time (min/epoch)	Inference Time (ms/sample)
BiLSTM	18.6	6.2	2.9	1.8
RoBERTa (fine-tuned)	102.3	22.4	6.5	3.9
Syntax-Aware Transformer	128.7	31.6	8.8	5.3
Proposed Method	121.5	28.9	7.9	4.7

Source: created by the authors

As shown in Table 3, the proposed method does not primarily aim at computational efficiency. BiLSTM and RoBERTa are lighter than the proposed method, but they also show lower classification performance. Compared with the Syntax-Aware Transformer, the proposed method uses fewer parameters and FLOPs and achieves shorter training and inference time. Together with Table 2 (Appendix), this indicates that the proposed method improves WMT performance while remaining less costly than the strongest syntax-enhanced baseline.

Figure 5 shows the evolution of Macro-F1 across training epochs on the WMT validation set for two strong baselines and the proposed method. The proposed method does not show an early lead in the initial training phase, but begins to build a clearer advantage in the middle of training and then remains relatively stable. This trend suggests that cross-channel interaction and local auxiliary learning mainly improve representation refinement during the later stages of optimization.

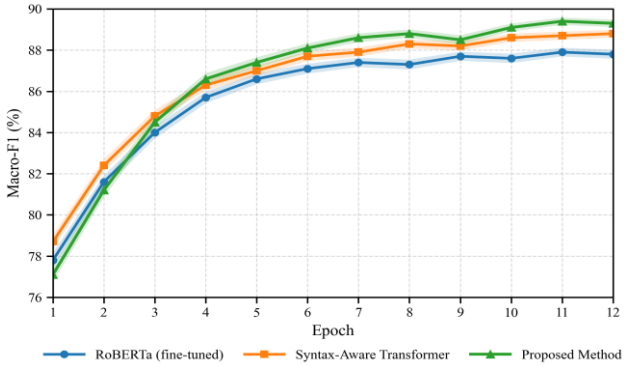


Fig. 5: Macro-F1 Curves of Different Models on the WMT Validation Set

Source: created by the authors

4.4 Case Study

To analyze the basis for the model's decisions, this paper selects a correctly predicted sample and an incorrectly classified sample for presentation. After visualizing the input sentences, the prediction results, and the local high-response segments, the role of different language segments in the prediction process could be observed (Figure 6).

In the cases of correct predictions, the samples were taken from the WMT Chinese news test set, and their true labels were the translated texts. The model also correctly identified their categories. The example sentence was "该协议将为双方未来合作奠定基础，并在多个领域发挥重要作用。" The visualization results showed that segments including "奠定基础" (lay the foundation), "发挥重要作用" (play an important role), and "多个领域" (multiple domains) received high response scores. The model also noticed the parallel structure in the sentence. This result indicates that even if the overall expression of the translation is relatively smooth, the model can still identify the translation phenomena present through local expression methods and syntactic organization features.

In the erroneous cases, the samples were taken from the UNPC Chinese formal text corpus. The ground-truth label is the original text, but the model incorrectly classifies it as translated text. The example sentence is: "有关各方应继续在相互尊重和协商一致的基础上推进相关工作。" The visualization indicates high responses for segments such as "有关各方" (relevant parties), "相互尊重" (mutual respect), "协商一致" (consensus), and "推进相关工作" (advance related work). This suggests that standardized formal expressions may be confused with translationese-like signals in formal written contexts.

In short, this case analysis is consistent with the quantitative results. The proposed method can use local expression patterns and syntactic structures effectively in news-style text. In formal texts, standardized expressions may overlap with translationese-like patterns, making the boundary between formal register and translationese less clear. Therefore, more fine-grained modelling of register variation is still needed for formal-domain translationese detection.

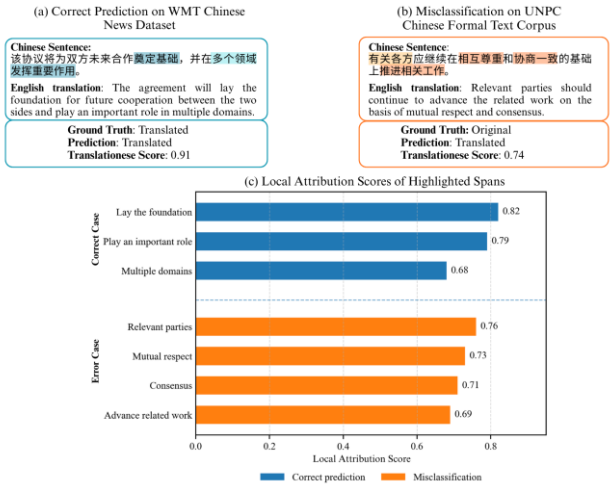


Fig. 6: Case Visualization of the Proposed Method
Source: created by the authors

4.5 Robustness Analysis

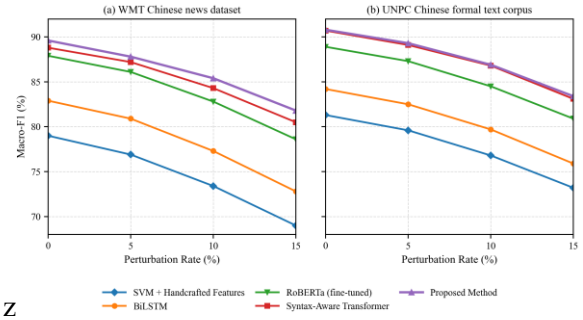
To assess model robustness under non-ideal input conditions, we construct perturbed test sets using the WMT Chinese news dataset and the UNPC Chinese formal text corpus. This study applied random perturbations at 0%, 5%, 10%, and 15% proportions to the input text, introducing word-level perturbations such as substitution, deletion, and slight reordering, and used macro F1 as the main evaluation metric. This experiment was designed to simulate possible local errors or expression changes in the text, such as word omission, word order variations, and part-of-speech deviations, and to observe performance changes across different models when input quality deteriorates. According to the stability analysis in Section 3.5, input variations may affect lexical representations, syntactic information extraction, and subsequent classification processes. Therefore, this experiment focuses on evaluating whether the proposed model can maintain relatively stable prediction capabilities under perturbation conditions.

As shown in Figure 7, as the perturbation ratio increases, the performance of all models declines to varying degrees, but the decline ranges significantly. On the WMT dataset, the SVM and BiLSTM based on manually designed features show a faster decline in performance, indicating that these methods are more sensitive to local word-order changes and word-segmentation perturbations; in contrast, RoBERTa and Syntax-Aware Transformer demonstrate better stability. The proposed method maintains a high macro F1 value under different perturbation levels and still shows good stability at 10% and 15% perturbation conditions. This result indicates that cross-channel consistency and local-global consistency design can help the model

maintain more stable representations, thereby reducing the performance decline caused by input perturbations.

On the UNPC dataset, the performance decline of each model is relatively slow, and the gap between Syntax-Aware Transformer and the proposed method also narrows further. This shows that in formal text scenarios, grammar-enhanced methods already have good stability, while this method mainly demonstrates its role in mitigating performance decline under medium-to-high perturbation conditions. Therefore, improvements in model robustness are related to the characteristics of the data domain, and when the text itself has a relatively stable syntactic structure, the improvement space is relatively limited.

Overall, the experimental results are consistent with the stability analysis in Section 3.5. The proposed method demonstrates stronger robustness to perturbations in news-type texts, while the improvement in formal texts is relatively limited. The experimental results indicate that the consistency fusion mechanism can enhance the model's adaptability to input variations to a certain extent, but its effect is still influenced by the type of perturbation and the text domain.



z

Fig. 7: Macro F1 under Different Disturbance Rates
Source: created by the authors

4.6 Ablation Study

To analyze the impact of different modules on the model's performance, this paper conducted ablation experiments on the WMT Chinese news test set and the UNPC Chinese official text corpus. The experiments used the complete model as the control group, and successively removed the dual-channel heterogeneous representation, cross-channel interaction and fusion module, local auxiliary discrimination module, cross-channel consistency term L_c, and local and global consistency term L_{lg}.

As shown in Table 4, after removing different components, model performance showed a certain degree of decline, indicating that each module played a role in the overall framework. Among them, the performance decline caused by removing the dual-channel representation was the most obvious. The accuracy and macro F1 on the WMT dataset decreased by 1.2 and 1.5 percentage points, respectively. This indicates that modeling lexical and syntactic information separately is highly significant for translation quality detection. In contrast, using simple concatenation instead of cross-channel interaction also causes performance degradation, indicating that simply inputting the two types of information in parallel is not sufficient to fully exploit their interrelationships.

These components therefore serve as complementary constraints. After removing the local auxiliary module L_c or the local-global consistency term $L L_{lg}$, the model performance all showed a certain decline. After removing Lc, the macro F1 score on the WMT dataset decreased from 89.6 to 89.0; after removing Llg, the macro F1 scores on the WMT and UNPC datasets dropped to 89.1 and 90.5, respectively. Although the decline was smaller than that caused by the dual-channel representation and interaction module, the results still indicate that local information and the connection between local predictions and overall judgments can help the model perform more accurate translation detection.

Table 4. Ablation Results

Model Variant	WMT Accuracy	WMT Macro-F1	UNPC Accuracy	UNPC Macro-F1
Full Model	90.0 ± 0.5	89.6 ± 0.5	90.8 ± 0.8	90.8 ± 0.7
w/o dual-channel (lexical only)	88.8 ± 0.6 (-1.2)	88.1 ± 0.6 (-1.5)	90.0 ± 0.8 (-0.8)	89.8 ± 0.8 (-1.0)
w/o interaction (concatenation)	89.2 ± 0.5 (-0.8)	88.6 ± 0.6 (-1.0)	90.3 ± 0.8 (-0.5)	90.1 ± 0.7 (-0.7)
w/o local auxiliary module	89.6 ± 0.6 (-0.4)	89.0 ± 0.5 (-0.6)	90.4 ± 0.8 (-0.4)	90.3 ± 0.8 (-0.5)
w/o consistency term (L_c)	89.5 ± 0.5 (-0.5)	89.0 ± 0.6 (-0.6)	90.5 ± 0.8 (-0.3)	90.4 ± 0.7 (-0.4)
w/o local-global consistency term (L_{lg})	89.7 ± 0.6 (-0.3)	89.1 ± 0.6 (-0.5)	90.4 ± 0.9 (-0.4)	90.5 ± 0.7 (-0.3)
w/o dual-channel + interaction	88.4 ± 0.7 (-1.6)	87.8 ± 0.7 (-1.8)	89.8 ± 0.9 (-1.0)	89.6 ± 0.8 (-1.2)

Source: created by the authors

When the dual-channel representation and the interaction module are removed simultaneously, the model performance declines most significantly. The WMT Macro-F1 drops by 1.8 percentage points. This indicates that the two modules have complementary

roles in the model: the dual-channel representation can separately model lexical and syntactic information, while the interaction module further promotes the integration of the two types of information. When lexical and syntactic information can be fully combined, the model can achieve better detection results.

In summary, the improvement in model performance mainly comes from the dual-channel representation and cross-channel interaction module. The local auxiliary mechanisms, L_c, and L_{lg} further utilize local information and strengthen the connection between local prediction results and sentence-level judgments.

4.7 Parameter Sensitivity Analysis

To analyze the impact of auxiliary target weights on the model performance, this paper further conducted a parameter sensitivity analysis on λ_1, λ_2, λ_3, and ρ. These parameters are respectively used to regulate local auxiliary supervision, cross-channel consistency, local and global consistency, and the similarity between lexical and syntactic representations. During the experiment, only one parameter was adjusted at a time, while the other parameters remained unchanged. The specific results are shown in Table 5.

As shown in Table 5, the selected setting $\lambda_1 = 0.5$, $\lambda_2 = 0.1$, $\lambda_3 = 0.2$, and $\rho = 0.5$ obtains the highest Macro-F1 among the tested settings on both datasets. The experimental results show that different auxiliary weights can have a certain impact on the model performance, but the overall change range is relatively limited. For λ_1, a lower value weakens the information supplementation provided by local auxiliary supervision, while a higher value may cause the model to overly focus on local patterns in weak labels, increasing noise interference.

The changes of λ_2 and λ_3 follow a similar pattern. When the cross-channel consistency weight is too low, the connection between lexical and syntactic information is weakened; when the weight is too high, the two channels may tend to be overly consistent. Changes in local and global consistency weights also affect the extent to which the model uses local information. When it is too low, it is difficult to exert the auxiliary role of local predictions, and when it is too high, the model may tend to rely on local evidence.

From the experimental results, it can be seen that when the target similarity $\rho = 0.5$, the model performs the best. At this point, the lexical representation and syntactic representation have a certain correlation while retaining their respective characteristics. Overall, these results support the parameter setting used in the main experiments, while also indicating that the auxiliary constraints should remain supportive rather than dominant in the overall objective.

Table 5. Parameter Sensitivity Analysis

Setting	WMT Macro-F1	UNPC Macro-F1
$\lambda_1 = 0.3, \lambda_2 = 0.1, \lambda_3 = 0.2, \rho = 0.5$	89.2 ± 0.7	90.5 ± 0.8
$\lambda_1 = 0.5, \lambda_2 = 0.1, \lambda_3 = 0.2, \rho = 0.5$	89.6 ± 0.5	90.8 ± 0.7
$\lambda_1 = 0.7, \lambda_2 = 0.1, \lambda_3 = 0.2, \rho = 0.5$	89.1 ± 0.6	90.4 ± 0.9
$\lambda_1 = 0.5, \lambda_2 = 0.05, \lambda_3 = 0.2, \rho = 0.5$	89.3 ± 0.6	90.4 ± 0.7
$\lambda_1 = 0.5, \lambda_2 = 0.2, \lambda_3 = 0.2, \rho = 0.5$	89.0 ± 0.8	90.3 ± 0.8
$\lambda_1 = 0.5, \lambda_2 = 0.1, \lambda_3 = 0.1, \rho = 0.5$	89.2 ± 0.5	90.6 ± 0.7
$\lambda_1 = 0.5, \lambda_2 = 0.1, \lambda_3 = 0.4, \rho = 0.5$	88.8 ± 0.7	90.2 ± 0.9
$\lambda_1 = 0.5, \lambda_2 = 0.1, \lambda_3 = 0.2, \rho = 0.3$	89.1 ± 0.6	90.5 ± 0.8
$\lambda_1 = 0.5, \lambda_2 = 0.1, \lambda_3 = 0.2, \rho = 0.7$	88.9 ± 0.7	90.3 ± 0.8

Source: created by the authors

5 Discussion

The results in Table 2 (Appendix), Table 4, Figure 5, and Figure 6 suggest that the proposed method benefits from the joint use of lexical and syntactic cues. The WMT news domain is a typical case where dual-channel representation and cross-channel interaction produce clearer gains. These results indicate that translationese is associated with syntactic transfer as well as lexical traces. Compared with models based on unified textual representations, the explicit separation and interaction of the two signals improve discrimination in the present experimental setting, which is consistent with recent studies that have proposed the role of structural evidence in translated-text analysis [14], [20].

Based on the ablation results and training curves, the dual-channel representation and interaction modules are mainly responsible for the performance improvement. The local auxiliary mechanism, local-global consistency constraint, and cross-channel consistency constraint also provide relatively small but stable gains. These auxiliary constraints appear to refine the representation space during optimization by linking local evidence, global prediction, and lexical-syntactic alignment. These findings suggest that translationese signals vary across lexical, syntactic, and local levels [6].

At the same time, the results show that one challenge in translationese detection is the fuzzy boundary between translationese and formal written language. The proposed method shows limited improvement over the strong syntax-enhanced baseline on the UNPC dataset. Based on the error analysis, the model may mistakenly identify some standardized formal expressions as translationese features. Therefore, when conducting translationese detection in formal text scenarios, it is necessary to further consider the expression differences between different language domains. Existing studies have also pointed out that the degree of manifestation of translationese phenomena is influenced by the system, text type, and language level [16], [19].

The robustness experiments show that the proposed method maintains stable performance on unperturbed data and only experiences a certain degree of performance decline after adding input perturbations. When some lexical or syntactic information is affected, the other channel and the fusion module can still provide auxiliary information. However, this method mainly mitigates performance loss caused by noise and cannot completely eliminate the impact of input changes.

This study still has certain limitations. The existing experiments rely on public language corpora, and different datasets vary in terms of text sources, writing styles, and category distribution. These factors may affect the comparison of model performance and the interpretation of results. The local auxiliary module has not yet been further verified through the segment-level data annotated by humans. Additionally, in formal texts, there may still be confusion between the features of translationese and the standardized expressions. Differences in text structure at different levels may also affect the model performance analysis.

The current efficiency assessment has not fully considered the additional computational cost brought about by the preprocessing of syntactic information. Future research can further expand the data range of different types of texts, introduce more granular local annotations, and explore more suitable detection methods for formal language domains.

6 Conclusion

This paper focuses on the detection problem of Chinese translationese. It focuses on identifying the expression features left during the translation process by using different language information in the text. In response to the problem that existing methods have difficulty in simultaneously utilizing lexical, syntactic, and local expression information, this paper constructs a dual-channel Transformer model, enabling different types of language features to work together in the detection process.

Experiments are conducted on the WMT Chinese news test set and the UNPC Chinese formal text corpus. The results show that the proposed method achieves better performance on the main WMT dataset, outperforming the strongest syntax-enhanced baseline in Accuracy, Macro-F1, and ECE. In the UNPC official text data, the model's performance was close to existing better methods, but the improvement was not significant. It may be because official texts are usually processed with stricter language norms, and some expressions are similar to those in translated texts, increasing the difficulty of judgment. In contrast, differences in word choice and sentence expression are more obvious in news texts, making it easier for the model to detect translation features.

Further analysis revealed that considering both word usage and sentence structure information simultaneously is an important reason for the model's better performance. The model also incorporates local expression features within sentences for judgment, making the prediction results more stable. However, experiments also showed that when official texts use relatively fixed and standardized expressions, the model may still misidentify them as translated texts. This indicates that differences between different writing styles remain an important factor affecting the detection results.

This paper verifies the feasibility of detecting translationese by integrating lexical, sentence structure, and local expression information. The experimental results show that considering different levels of language characteristics comprehensively can provide more basis for the identification of translationese. This method has certain application value in tasks such as screening machine translation results, assisting post-editing, and generating text style analysis.

Future research can further expand the data scope to cover more fields and text types, and enhance the annotation of local language features. At the same time, it is necessary to optimize the processing method of syntactic information and reduce additional computational costs during model operation.

References:

[1] Church, K., Li, B., Vickers, P., Dudy, S., & Yue, R. (2025). Emerging trends: translationese. Natural Language Processing, 31(3), 965-981. https://doi.org/10.1017/nlp.2025.1

[2] Doshi, M., Dabre, R., & Bhattacharyya, P. (2024). Pretraining language models using translationese. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Miami, Florida, USA) (pp. 5843-5862). DOI: 10.18653/v1/2024.emnlp-main.334. https://aclanthology.org/2024.emnlp-main.334/

[3] Bouamor, H., Pino, J., & Bali, K. (2023). Proceedings of the 2023 conference on empirical methods in natural language processing. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Singapore) (Accessed Date: 10 Dec 2023). https://aclanthology.org/2023.emnlp-main.0/

[4] Fang, T., Liu, X., Wong, D. F., Zhan, R., Ding, L., Chao, L. S., ... & Zhang, M. (2023). Transgec: Improving grammatical error correction with translationese. In Findings of the association for computational linguistics: ACL 2023 (Toronto, Canada) (pp. 3614-3633). DOI: 10.18653/v1/2023.findings-acl.223. https://aclanthology.org/2023.findings-acl.223/

[5] Li, Y., Zhang, R., Wang, Z., Zhang, H., Cui, L., Yin, Y., ... & Zhang, Y. (2025). Lost in literalism: How supervised training shapes translationese in LLMs. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vienna, Austria) (pp. 12875-12894). DOI: 10.18653/v1/2025.acl-long.630. https://aclanthology.org/2025.acl-long.630/

[6] Liu, Y., Zhang, W., Wang, Y., Tang, J., Zhang, P., Yang, B., ... & Hu, H. (2025). Translationese-index: Using Likelihood Ratios for Graded and Generalizable Measurement of Translationese. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (Suzhou, China) (pp. 12532-12549). DOI: 10.18653/v1/2025.emnlp-main.633. https://aclanthology.org/2025.emnlp-main.633/

[7] Jalota, R., Chowdhury, K., España-Bonet, C., & van Genabith, J. (2023). Translating away translationese without parallel data. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Singapore) (pp. 7086-7100). DOI: 10.18653/v1/2023.emnlp-main.438. https://aclanthology.org/2023.emnlp-main.438/

[8] Wastl, M., Vamvas, J., & Sennrich, R. (2025). Machine Translation Models are Zero-Shot Detectors of Translation Direction. In Findings of the Association for Computational Linguistics: ACL 2025 (Vienna, Austria) (pp. 1054-1074). DOI: 10.18653/v1/2025.findings-acl.59. https://aclanthology.org/2025.findings-acl.59/

[9] Kunilovskaya, M., Chowdhury, K. D., Przybyl, H., España-Bonet, C., & Genabith, J. (2024). Mitigating translationese with gpt-4: Strategies and performance. In Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1) (Sheffield, UK) (pp. 411-430). https://aclanthology.org/2024.eamt-1.35/

[10] Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., ... & Arnold, T. (2024). Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024) (Mexico City, Mexico) (pp. 2057-2079). DOI: 10.18653/v1/2024.semeval-1.279. https://aclanthology.org/2024.semeval-1.279/

[11] Freitag, M., Mathur, N., Deutsch, D., Lo, C. K., Avramidis, E., Rei, R., ... & Lavie, A. (2024). Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In Proceedings of the Ninth Conference on Machine Translation (Miami, Florida, USA) (pp. 47-81). DOI: 10.18653/v1/2024.wmt-1.2. https://aclanthology.org/2024.wmt-1.2/

[12] Macko, D., Moro, R., Uchendu, A., Lucas, J., Yamashita, M., Pikuliak, M., ... & Bielikova, M. (2023). MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Singapore) (pp. 9960-9987). DOI: 10.18653/v1/2023.emnlp-main.616. https://aclanthology.org/2023.emnlp-main.616/

[13] Ahmad, Z., Torres-Ruiz, M., Mahmood, A., Quintero, R., Ameer, I., & Bölücü, N. (2026). Human or Machine? A Survey on Machine-Generated Text Detection. IEEE Access, 14, 34113-34136. https://doi.org/10.1109/ACCESS.2026.3666781

[14] Wang, Z., Cheung, A. K., & Liu, K. (2024). Entropy-based syntactic tree analysis for text classification: a novel approach to

distinguishing between original and translated Chinese texts. Digital Scholarship in the Humanities, 39(3), 984-1000. https://doi.org/10.1093/llc/fqae030

[15] Zhang, S., & Zhao, C. (2026). Machine translationese of large language models: Dependency triplets, text classification, and SHAP analysis. PLoS One, 21(1), e0339769. https://doi.org/10.1371/journal.pone.0339769

[16] Sizov, F., España-Bonet, C., Van Genabith, J., Xie, R., & Chowdhury, K. D. (2024). Analysing translation artifacts: A comparative study of LLMs, NMTs, and human translations. In Proceedings of the Ninth Conference on Machine Translation (Miami, Florida, USA) (pp. 1183-1199). DOI: 10.18653/v1/2024.wmt-1.116. https://aclanthology.org/2024.wmt-1.116/

[17] Xu, B. (2026). Translationese: Can AI Refinement Techniques Completely Eradicate Translational Distortions?. In AI Through LLMs: Transforming Translation Practice and Teaching (pp. 95-120). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-95-4798-2_4

[18] Wein, S., & Schneider, N. (2024). Lost in translationese? reducing translation effect using Abstract Meaning Representation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (St. Julian's, Malta) (pp. 753-765). DOI: 10.18653/v1/2024.eacl-long.45. https://aclanthology.org/2024.eacl-long.45/

[19] Kong, D., & Macken, L. (2025). Decoding machine translationese in English-Chinese news: LLMs vs. NMTs. In Proceedings of Machine Translation Summit XX (Geneva, Switzerland): Volume 1 (pp. 99-112). https://aclanthology.org/2025.mtsummit-1.8/

[20] Jiang, Z. (2025). Does LLM translation align with translation universals? A cross-genre simplification study on English-Chinese translation based on dependency grammar. PLoS One, 20(6), e0324830. https://doi.org/10.1371/journal.pone.0324830

[21] Wang, L., & Jiang, Y. (2024). Do translation universals exist at the syntactic-semantic level? A study using semantic role labeling and textual entailment analysis of English-Chinese translations. Humanities and Social Sciences Communications, 11(1), 1-12. https://doi.org/10.1057/s41599-024-03317-6

[22] Koppel, M., & Ordan, N. (2011). Translationese and its dialects. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (Portland, Oregon, USA) (pp. 1318-1326). https://aclanthology.org/P11-1132/

[23] Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., & Xu, B. (2016). Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (Osaka, Japan) (pp. 3485-3495). https://aclanthology.org/C16-1329/

[24] Wang, M., Zhang, Z., Li, H., & Li, B. (2025). Modeling Sentence Classification: Considering Syntactic Analysis and Position Embedding. In International Conference on Innovation in Artificial Intelligence (pp. 399-410). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-95-0887-7_30

[25] Devianti, R. S., & Miyao, Y. (2024). Transferability of Syntax-Aware Graph Neural Networks in Zero-Shot Cross-Lingual Semantic Role Labeling. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 20-42) (Miami, Florida, USA). DOI: 10.18653/v1/2024.findings-emnlp.2. https://aclanthology.org/2024.findings-emnlp.2/

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Qianqian Tian is the only author of this paper.

Sources of Funding for Research Presented in a Scientific Article or the Scientific Article Itself

The research received no funding sources.

Conflict of Interest

The authors have no conflicts of interest to declare.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0 https://creativecommons.org/licenses/by/4.0/deed.en_US

APPENDIX

Table 2. Main Performance Comparison

Method	WMT Accuracy	WMT Macro-F1	WMT ECE↓	UNPC Accuracy	UNPC Macro-F1	UNPC ECE↓
SVM + Handcrafted Features	79.8 ± 1.0	79.0 ± 1.2	0.088 ± 0.008	82.1 ± 1.1	81.3 ± 1.0	0.096 ± 0.007
BiLSTM	83.6 ± 0.9	82.9 ± 1.0	0.070 ± 0.006	84.8 ± 1.0	84.2 ± 1.1	0.075 ± 0.006
RoBERTa (fine-tuned)	88.3 ± 0.6	87.9 ± 0.7	0.044 ± 0.004	89.4 ± 0.8	88.9 ± 0.8	0.048 ± 0.004
Syntax-Aware Transformer	89.4 ± 0.7	88.8 ± 0.6	0.037 ± 0.003	91.0 ± 0.8	90.7 ± 0.8	0.040 ± 0.004
Proposed Method	90.0 ± 0.5	89.6 ± 0.5	0.034 ± 0.003	90.8 ± 0.8	90.8 ± 0.7	0.038 ± 0.004

Note: Lower ECE indicates better calibration. The p-values (two-tailed t-test) comparing the proposed method with the Syntax-Aware Transformer are as follows: WMT Accuracy 0.041, WMT Macro-F1 0.014, WMT ECE 0.052; UNPC Accuracy 0.184, UNPC Macro-F1 0.087, UNPC ECE 0.061.
Source: created by the authors