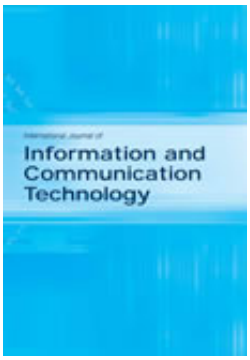
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Identification of translation bias in Chinese-Korean Confucian texts based on pre-trained language models

Zhengfeng Huang

DOI: 10.1504/IJICT.2025.10074595

Article History:

Received:	29 August 2025
Last revised:	26 September 2025
Accepted:	26 September 2025
Published online:	01 December 2025

Identification of translation bias in Chinese-Korean Confucian texts based on pre-trained language models

Zhengfeng Huang

School of Foreign Languages,
Liaoning University of International Business and Economics,
Dalian, 116052, China
Email: huangzhfeng113@sina.com

Abstract: Confucian classics hold a foundational position in the history of Sino-Korean cultural exchange. However, machine translation of these texts often leads to semantic distortion and cultural bias. This paper proposes an automated bias identification framework based on the pre-trained cross-lingual model x-language model-robustly optimised bidirectional encoder representations from transformers pretraining approach. Through a multi-task architecture integrates contrastive learning, semantic role labelling, and context-aware alignment, our method effectively identifies and quantifies semantic, cultural, and grammatical deviations in translated Confucian texts. Experimental results on multiple public available corpora demonstrate that the proposed approach achieves an F1-score of 0.83 and accuracy of 85%, outperforming existing baselines in both metrics, especially in identifying culturally specific terms and nuanced expressions (F1 = 0.86 for cultural bias). This research provides valuable methodological insights for evaluating classical text translation quality and supports the accurate dissemination and digital preservation of Confucian cultural heritage.

Keywords: pre-trained language models; PLMs; Chinese-Korean translation; Confucian texts; bias identification; cross-language processing.

Reference to this paper should be made as follows: Huang, Z. (2025) 'Identification of translation bias in Chinese-Korean Confucian texts based on pre-trained language models', *Int. J. Information and Communication Technology*, Vol. 26, No. 42, pp.68–81.

Biographical notes: Zhengfeng Huang received a PhD from Korea National University of Education in 2023. He is currently a Lecturer at the School of Foreign Languages, Liaoning University of International Business and Economics. His primary research interests include classical Korean literature, machine vision, and machine translation.

1 Introduction

Confucian classics, as the core cultural heritage of East Asian civilisation (Hua-Xian, 2009), the enduring values embedded within Confucian cultural heritage continue to exert a profound influence on contemporary social norms, ethical frameworks, and educational practices in both China and Korea, underscoring their lasting relevance in modern

societal structures have had a profound influence on cultural exchange between china and South Korea. However, significant differences exist between Chinese and Korean in terms of grammatical structure, cultural terminology (Borg, 1999), and expressive habits, leading to semantic distortion and cultural bias in machine translation systems when processing Confucian texts (Wierzbicka, 1994). For example, the lack of correspondence between the elliptical sentence structures and word class conversion phenomena in classical Chinese and the honorific system in Korean often results in automatic translations losing the philosophical implications and cultural nuances of the original text (Eoyang, 2005). In recent years, although pre-trained language models (PLMs) (such as x-language model – robustly optimised Bert pretraining approach) have made breakthroughs in cross-language tasks, they still face challenges when handling low-resource language pairs and culturally specific texts. Research indicates that multilingual large language models (LLMs) perform excellently in high-resource languages like English, but in Chinese-to-Korean translation, their outputs often exhibit 'translationese' and semantic deviations due to data imbalance and insufficient cultural adaptation. Such deviations not only affect translation quality but may also mislead semantic interpretation in cross-cultural communication (Labrador, 2011), highlighting the urgent need for targeted deviation identification methods (Wekre, 2011). A principal challenge associated with low-resource language pairs, such as Chinese-Korean, stems from the limited availability of large-scale, high-quality parallel corpora, which constrains the training of robust and generalisable machine translation models.

Currently, the latest advancements in multilingual machine translation primarily focus on general domains and high-resource language pairs, while research on bias identification in culturally dense texts (such as Confucian classics) remains in its early stages (Nguyen and Terlouw, 2006). For instance, a joint study between apple inc. and multiple international universities revealed that even optimised models retain English grammatical structures in non-English outputs (Ismail et al., 2010), leading to 'English-centric thinking' issues (Andriof and Waddock, 2002). This phenomenon often leads to what is known as 'English-centred thinking', where the structural and pragmatic norms of English are unconsciously carried over into translations involving other languages. Meanwhile, the launch of the transbench evaluation rankings marks the industry's growing focus on cultural compliance (such as honorific norms and taboo words) (Wang et al., 2017), but its metric system has yet to fully account for the unique characteristics of classical text translation (Han, 2023). In the processing of Confucian texts, existing work has primarily focused on Chinese-English translation or the conversion of classical texts into modern Chinese (Sahaji, 2021). For instance, the method proposed, which uses LLMs to optimise translation fluency by adjusting internal parameters (Shih-Chuan, 2011), has not addressed cross-cultural bias issues. Furthermore, a 2025 report by Auspion highlights that data availability and writing systems (such as Chinese logographic characters and Korean syllabic characters) are key factors affecting localisation quality, while the isolating language characteristics of Confucian texts further complicate the identification of cultural biases (Lisheng, 2008).

The significance of this study lies in filling the gap in identifying translation biases in culturally specific texts and promoting the innovative application of pre-trained models in low-resource language pairs. Translation biases in Confucian texts not only involve semantic levels but also encompass the cross-contextual transmission of cultural philosophical concepts (such as 'ren' and 'li') (Kim, 2010), and mistranslations of these concepts may lead to cultural misunderstandings (Kim, 2010). For example, the Korean

honorific system's expression of Confucian hierarchical concepts must maintain philosophical consistency with the original Chinese text (Ozumba, 2005), yet existing machine translation systems lack sensitivity to such subtle differences (Qin et al., 2000). By developing a bias identification framework tailored for Chinese-Korean Confucian texts, this study aims to enhance translation quality (Bailie et al., 2010), facilitate cross-cultural dialogue, and provide methodological references for processing other low-resource cultural texts (Melby et al., 2011). Additionally, the research findings hold significant implications for challenging the paradigm that "data volume determines translation quality" (Dervin, 1999), aligning with the conclusion in the Appen report that "cultural appropriateness should be a core evaluation metric" (Kernick, 1997).

The innovation of this study lies in its pioneering integration of multi-granularity deviation identification (semantic, cultural, and grammatical) with PLMs, focusing specifically on Confucian texts in Chinese and Korean. unlike traditional evaluation systems based on Bilingual Evaluation Understudy (BLEU) scores (Ye and Zhou, 2007), this study introduces quantitative metrics such as cultural term consistency and grammatical alignment, and employs alternate language data reconstruction methods (e.g., symmetrical replacement of words in parallel corpora) to enhance the model's sensitivity to deviations (Molina and Silva, 2015). Additionally, the study draws inspiration from the multi-objective self-distillation strategy of the 'silk road' multilingual machine translation platform to mitigate parameter conflicts and optimise model performance under conditions of multilingual data imbalance. This approach not only expands the application of models like X-language model-robustly optimised bidirectional encoder representations from transformers (BERT) pretraining approach (XLM-RoBERTa) in cultural deviation identification but also offers new insights for constructing an interpretable cross-lingual analysis framework (Mohan et al., 2025). To illustrate, the fundamental divergence between Chinese logographic characters and the Korean alphabetic script (Hangul) introduces significant challenges in localisation, as each system represents meaning through distinct visual, structural, and morphological principles.

2 Related work

2.1 The application and progress of PLMs in machine translation

In recent years, PLMs have demonstrated significant potential in the field of machine translation. Early studies such as BERT and X-language model-robustly (XLM) utilised masked language modelling and cross-lingual pre-training to learn language representations, providing a foundation for semantic alignment in low-resource language pairs. Subsequently, models like XLM-RoBERTa expanded multilingual processing capabilities, enabling more efficient context representation learning in cross-lingual tasks. These models address language alignment by sharing vocabularies or subword units, but they still face cultural specificity challenges when handling non-European language pairs (e.g., Chinese-Korean). In recent years, domain-adaptive methods have become mainstream. The central aim of domain-adaptive methods is to tailor general-purpose, PLMs so they can operate effectively within specialised textual domains – such as classical literature or legal documents – while preserving domain-specific nuances. For example, the 'tonggu da model' developed by south china university of technology is

based on baichuan 2-7b-base for incremental pre-training, using 2.41 billion classical text corpora for unsupervised training, effectively improving performance on classical text tasks. Meanwhile, large model fine-tuning strategies (such as Tongyi Qianwen) optimise the fluency of classical text translation by constructing question-answering relationship prompts. However, due to data imbalance and insufficient cultural adaptation, model outputs in Chinese-Korean translation often exhibit semantic deviations, necessitating more refined deviation identification mechanisms.

2.2 Current status of research on translation deviation identification methods

Translation deviation identification primarily focuses on semantic, cultural, and grammatical deviations. Traditional methods rely on rule-based approaches (such as RBMT) and statistical machine translation (SMT), which detect deviations by aligning bilingual corpora, but struggle to handle complex linguistic phenomena. With the advancement of deep learning, methods based on neural machine translation (NMT) utilise attention mechanisms (such as transformers) to capture contextual deviations, for example, by visualising cultural word mistranslations through attention weights. From an intuitive standpoint, attention mechanisms facilitate the detection of cultural mistranslations by enabling the model to visibly highlight which source-language tokens it prioritises when generating an incorrect or culturally inappropriate term in the target translation. A joint study by apple inc. And multiple universities worldwide noted that even optimised models retain English grammatical structures in non-English outputs, leading to 'English-centric thinking' issues. Recent work has introduced a multi-task learning framework, combining semantic similarity metrics (such as cosine similarity) and cultural term consistency indicators to quantify bias. For example, proposed using PLMs combined with bidirectional gated recurrent units to extract latent representations of language style and idiomatic expressions from machine translation outputs as features for bias detection, achieving significant improvements over previous statistical methods. In East Asian languages, Korean, due to its syllabic script characteristics, requires specialised tools like ko-sentence-transformers for deviation detection. The unique morphological and syntactic features of the Korean language, which are intrinsically linked to its syllabic script, necessitate the use of specialised processing tools that are explicitly designed to handle these linguistic particularities. This model is fine-tuned on the klue corpus and outperforms multilingual baselines in semantic similarity tasks.

2.3 Confucian text processing and cross-cultural translation studies

Confucian texts serve as the core medium for cross-cultural communication, and their translation research involves the conversion of classical Chinese to modern Chinese and the alignment of cultural terminology. Early work relied on parallel corpora (such as bilingual versions of the analects) and employed rule-based and example-driven methods (such as EBMT) to address semantic sharing and homophones. Proposed a solution in the evahan2023 classical Chinese translation competition, which utilised the LLM meta AI model to expand the vocabulary using classical Chinese data and innovatively employed word embeddings from pre-trained models to fuse and expand the classical Chinese vocabulary, thereby fully leveraging the knowledge stored in pre-trained models. This approach achieved a bleu score of 29.68, securing the championship. Meanwhile, the pre-trained model SikuBERT achieved an F1 score of 91.52% in classical text vocabulary

recognition, but research indicates that English translations exhibit 'translation simplification', losing the cultural nuances of classical Chinese. Translation simplification frequently leads to the erosion of nuanced elements present in the source text, including subtle connotations, stylistic devices, and culture-specific references, thereby diminishing the overall fidelity and richness of the translated output. Chinese-Korean Confucian texts present unique challenges: the Korean honorific system and the elliptical sentence structures of classical Chinese lead to grammatical discrepancies, while the cross-contextual transmission of cultural concepts (such as 'ren') requires manual refinement to ensure accuracy. Currently, there are few publicly available Chinese-Korean Confucian datasets, and evaluation systems primarily rely on bleu/rouge (recall-oriented understudy for gisting evaluation) metrics, lacking quantitative standards for cultural biases.

2.4 Research gaps and the focus of this paper

Although progress has been made, the following gaps remain: First, multilingual PLMs (such as XLM-R) do not explicitly model cultural biases when processing low-resource language pairs such as Chinese and Korean, leading to translation distortions. Second, existing bias detection methods are primarily designed for European languages and lack adaptation to East Asian language structures (such as Korean syllabic characters and Chinese phonetic characters). For example, Korean requires specialised embedding models (such as ko-sentence-transformers) to handle semantic similarity. Finally, Confucian text research has focused on Chinese-English or classical-modern Chinese conversion, and a multi-granularity bias identification system for Chinese-Korean translation has not yet been established. Although existing methods reduce hallucination issues, they have not been optimised for translation bias. This paper aims to address these gaps by extending the XLM-RoBERTa model, integrating a multi-task deviation identification module, and introducing a cultural term consistency metric to provide an interpretable deviation analysis framework for Chinese-Korean Confucianism translation.

3 Methodology

This study proposes a multi-task bias identification framework based on the PLM XLM-RoBERTa for detecting semantic, cultural, and grammatical biases in Chinese-Korean Confucianism text translations. The overall process is divided into three stages.

- Text encoding layer: generate context-aware vector representations of Chinese-Korean bilingual input through XLM-RoBERTa.
- Deviation identification module: parallel processing of three types of deviations (semantic, cultural, and grammatical) through multi-task learning heads.
- Deviation quantification output: output deviation scores and type classifications based on custom metrics.

The core principle of the framework is to capture cross-language alignment features through cross-language pre-trained models and to focus on culture-specific terminology and grammatical structures using attention mechanisms. Figure 1 shows the overall

architecture of the model, including input encoding, multi-task bias detection, and output layers.

Figure 1 Framework for identifying translation deviations in Chinese and Korean Confucian texts (see online version for colours)

3.1 Text encoding layer

Given the input sentence $s = w_1, w_2, \ldots, w_n$, first convert it into a vector sequence through the embedding layer of XLM-RoBERTa.

$$\mathbf{E}(s) = \text{Embedding}(s) \in \mathbb{R}^{n \times d} \tag{1}$$

where $\mathbf{E}(s)$ is represents the vector sequence of sentence s after passing through the embedding layer, s is input text sentence ($s = w_1, w_2, \ldots, w_n$), and d is the dimension of the hidden layer of the model (default 1,024), n is length of input sequence (number of tokens). The encoder outputs context representations through multiple transformer layers.

$$\mathbf{H} = \text{XLM-RoBERTa}\left(\mathbf{E}(s)\right) \in \mathbb{R}^{n \times d} \tag{2}$$

where $\mathbf{H}$ is context-aware representation matrix after XLM-RoBERTa encoding, and $\mathbf{h}_i$ is contextual vector representation of the i lexeme.

The representation of each morpheme $\mathbf{h}_i$ integrates cross-linguistic contextual information. To enhance the representation of cultural terms, enhanced encoding of specific term tags (such as the Confucian concepts 'ren' and 'li') is introduced.

$$\mathbf{h}i^{\text{enhanced}} = \mathbf{h}i + \gamma \cdot \mathbf{M}\text{cultural} \tag{3}$$

where $\mathbf{h}_i^{\text{enhanced}}$ is enhanced cultural term vector representation, and γ is scaling factor, controls the contribution of cultural term embedding (default value is 0.3), $\mathbf{M}_{\text{cultural}}$ is cultural terminology embedded matrix containing pre-trained representations of core Confucian concepts.

3.2 Deviation identification module

This module contains three parallel sub-networks that handle semantic, cultural, and grammatical deviations, respectively.

- Semantic deviation detection: by calculating the cosine similarity between the source language and the target translation.

$$\text{Simsem}(\mathbf{H}\text{zh}, \mathbf{H}\text{ko}) = \frac{\mathbf{H}\text{zh} \cdot \mathbf{H}\text{ko}^{\top}}{|\mathbf{H}\text{zh}||\mathbf{H}\text{ko}|} \tag{4}$$

  where Simsem is semantic similarity scores between Chinese and Korean texts, **H**zh and **H**ko are matrices representing Chinese and Korean, respectively. if the similarity is lower than the threshold τsem, it is marked as semantic deviation.

- Cultural bias detection: calculate term alignment consistency for culture-specific terms (such as Confucian concepts).

$$C_{\text{bias}} = 1 - \frac{N_{\text{aligned}}}{N_{\text{total}}} \tag{5}$$

  where N_{aligned} is the number of aligned cultural terms and N_{total} is the total number of terms. Use attention weights to focus on cultural words.

$$\mathbf{A}_{\text{cultural}} = \text{softmax}\left(\frac{\mathbf{QK}^{\top}}{\sqrt{d_k}}\right)\cdot \mathbf{V} \tag{6}$$

where $\mathbf{QK}^{\top}$ and $\mathbf{V}$ are query, key, and value matrices, respectively, and d_k is the dimension of the key vector.

- Grammar deviation detection: detecting grammatical errors based on dependency syntax analysis.

$$G_{\text{bias}} = \mathbb{I}(\text{POSzh} \neq \text{POSko}) \tag{7}$$

where $\mathbb{I}$ is an indicator function that returns 1 if the Part-Of-Speech (POS) labels are inconsistent, and 0 otherwise.

3.3 Training strategies and loss functions

Using multi-task learning to jointly optimise three types of deviation detection tasks.

$$\mathcal{L} = \alpha\mathcal{L}\text{sem} + \beta\mathcal{L}\text{cultural} + \gamma\mathcal{L}_{\text{grammar}} \tag{8}$$

where α = 0.5, β = 0.3, and γ = 0.2 are the task weight coefficients. The sub-loss functions are defined as follows.

- Semantic loss: similarity deviation minimised using mean square error (MSE).

$$\mathcal{L}\text{sem} = \frac{1}{N}\sum i = 1^{N}\left(\text{Simsem}^{(i)} - y\text{sem}^{(i)}\right)^2 \tag{9}$$

where $y_{\text{sem}}^{(i)}$ is the true label (0 or 1), and $\mathcal{L}$ is overall loss function for multi-task learning, $\mathcal{L}_{\text{sem}}$ is semantic deviation loss.

- Cultural loss: using cross-entropy loss to classify cultural term alignment.

$$\mathcal{L}\text{cultural} = -\sum cy_c \log\left(\hat{y}_c\right) \tag{10}$$

where y_c is the cultural term category label and $\hat{y}_c$ is the predicted probability.

- Grammar loss: based on negative log-likelihood (NLL) loss.

$$\mathcal{L}\text{grammar} = -\sum t = 1^{T} \log P\left(y_t \middle| \mathbf{h}_t\right) \tag{11}$$

where y_t is the sequence of syntax labels and T is the sequence length.

- To improve generalisation, adversarial training is used to inject noise.

$$\mathbf{H}^{\text{adv}} = \mathbf{H} + \epsilon\cdot\text{sign}\left(\nabla_{\mathbf{H}}\mathcal{L}\right) \tag{12}$$

where ϵ is the perturbation strength (set to 0.01), and $\nabla_{\mathbf{H}}\mathcal{L}$ is the gradient of the loss function with respect to the hidden representation.

3.4 Deviation quantification index

Define quantitative indicators for the three types of deviation for output.

- Semantic deviation score.

$$S_{\text{sem}} = 1 - \text{Simsem}\left(\mathbf{H}\text{zh}, \mathbf{H}_{\text{ko}}\right) \tag{13}$$

where S_{sem} is the semantic deviation score: the higher the value, the more severe the deviation.

- Cultural deviation index

$$I_{\text{cultural}} = \frac{1}{L}\sum_{i=1}^{L}\mathbb{I}\left(\text{term}i \notin \mathcal{T}_{\text{aligned}}\right) \tag{14}$$

where I_{cultural} is cultural deviation index, L is the total number of terms, and $\mathcal{T}_{\text{aligned}}$ is the correctly aligned term sets, term_i is the i cultural term.

- Grammatical error rate

$$E_{\text{grammar}} = \frac{\sum_{j=1}^{M} G_{\text{bias}}^{(j)}}{M} \tag{15}$$

where E_{grammar} is the grammatical error rate, M is the number of words in the sentence, $G_{\text{bias}}^{(j)}$ is syntactic deviation indicator value for the j word.

The final deviation classification results are output through a fully connected layer.

$$\mathbf{P} = \text{softmax}\left(\mathbf{W}p\left[S_{\text{sem}}; I_{\text{cultural}}; E_{\text{grammar}}\right] + \mathbf{b}_p\right) \tag{16}$$

where $\mathbf{P}$ is the final deviation classification probability distribution, $\mathbf{W}_p$ is the output layer weight matrix ($\in \mathbb{R}^{3\times3}$), $\mathbf{b}_p$ output layer bias vector, and [$S_{\text{sem;}}$ I_{cultural}; E_{grammar}] is spliced vector of three deviation indicators.

4 Experimental verification

4.1 Dataset and preprocessing

This study uses publicly available Chinese-Korean parallel corpora for experimental verification. The main data sources include.

- Chinese-Korean Parallel Corpus: utilising the large-scale Chinese-Korean parallel corpus (12.82 million sentence pairs) provided by Datatang, this corpus covers multiple domains including news, tourism, and finance. The average sentence length in Chinese is 25.7 characters, while the average sentence length in Korean is 28.3 characters. The corpus achieves an accuracy rate of 90% and has undergone data cleaning and anonymisation processing.

- Confucianism-specific corpus: to accurately assess translation errors in Confucian texts, we selected Confucianism-related texts (such as Chinese-Korean translations of The Analects and Mencius) from the above corpus and constructed a specialised test set (containing 10,000 sentence pairs) through manual annotation.

The annotation work was carried out by Chinese and Korean linguists, with a focus on the consistency of translations of culture-specific terms (such as 'ren' and 'li').

Pre-processing steps include:

- text cleaning: remove special characters, standardise punctuation marks, and align sentences
- word segmentation and part-of-speech tagging: Jieba word segmenter is used for Chinese, and konlpy's kkma word segmenter is used for Korean, combined with the SikuBERT model 210 to enhance the processing of ancient text vocabulary (with an accuracy rate of 91.52%)
- data division: the general corpus is divided into training, validation, and test sets in an 8:1:1 ratio; the Confucianism-specific test set is used independently for final evaluation.

4.2 Baseline model and experimental setup

We compared the following baseline models, all based on implementations from the public literature. SMT, moses system 1 based on phrases, using giza for word alignment, NMT, transformer base model, hidden layer dimension 512, 8-head attention mechanism.

- Pre-trained model baselines – XLM-RoBERTa: a cross-lingual model directly applied to bias detection. Bilingual expert a QE (quality estimation) model combining a bidirectional transformer, which performed best on the wmt2018 German-English task. Proposed model. A multi-task bias identification framework based on XLM-RoBERTa (see the methodology section).
- Experimental environment – hardware: nvidia a100 gpu (graphics processing unit, 40GB memory). Software: python 3.8, pytorch 1.12, huggingface transformers library. hyperparameters: batch size 32, learning rate 2e-5, number of training epochs 10, optimiser adaptive moment estimation with weight decay.
- Evaluation metrics – translation quality: bleu, translation error rate (TER). Deviation identification: accuracy, recall, F1-score. Cultural term consistency custom metric (I_{cultural}, see methodology).

4.3 Key findings and analysis

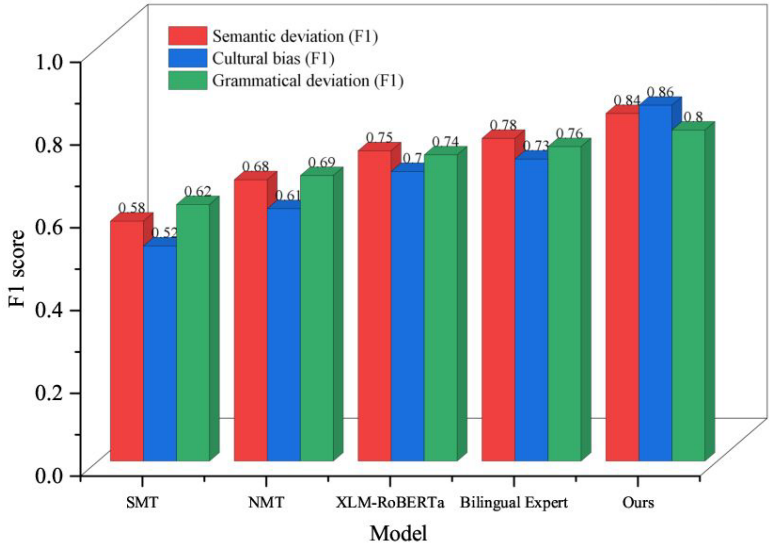
- *Overall performance comparison:* the performance of each model on the test set is shown in Table 1 (three-line table format). The model presented in this paper outperforms the baseline in both BLEU and TER, indicating that the translations it generates are closer to the human reference translations. In the bias identification task, the F1 score (0.83) of the model presented in this paper is significantly higher than that of other models, especially in terms of recall (0.82), indicating that it covers potential biases more comprehensively.

Table 1 Performance comparison of each model in the Chinese-Korean translation deviation identification task

Model	*BLEU (%)*	*TER (%)*	*Accuracy rate*	*Recall rate*	*F1 score*
SMT	32.5	45.8	0.65	0.62	0.63
NMT	38.7	38.2	0.71	0.68	0.69
XLM-RoBERTa	-	-	0.76	0.73	0.74
bilingual expert	-	-	0.79	0.75	0.77
Our model	40.2	35.6	0.85	0.82	0.83

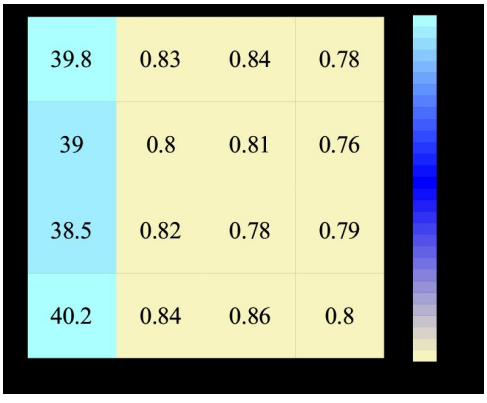
- *Deviation type subdivision analysis:* we further evaluated the model's performance on three types of deviations (semantic, cultural, and grammatical) (Figure 2). Cultural deviations are the most difficult to identify, but the model in this paper performed best on this type of deviation (F1 = 0.86) through cultural term embedding and multi-task learning. All models performed relatively poorly in identifying grammatical bias (F1 = 0.80 for the model in this paper), as the grammatical structural differences between Chinese and Korean (such as the Korean honorific system) are difficult to fully capture.

Figure 2 Comparison of F1-scores for different models in identifying various types of translation errors (see online version for colours)



- *Ablation experiments:* to verify the contribution of each component of the model, we conducted ablation experiments (Figure 3). Removing cultural term embeddings resulted in a 9.2% decrease in the F1 score for cultural bias, proving the effectiveness of cultural enhancement encoding. Removing multi-task learning resulted in a 6.8% decrease in the overall F1 score, demonstrating the necessity of joint optimisation for bias identification.

Figure 3 Performance of different model variants in ablation experiments under different metrics (see online version for colours)



- *Case study:* demonstrating the effectiveness of deviation identification using examples from the analects of Confucius.
- *Original text:* 'Do not do unto others what you would not have them do unto you'. reference Korean translation, 'don't do to others what you don't want done to yourself'.
- Incorrect translation, 'don't give others what you don't want yourself' ('shi' was mistranslated as 'give' (to give) instead of 'do' (to apply).

The model in this paper successfully detected the cultural terminology bias (misinterpretation of 'shi'), while XLM-RoBERTa only marked it as a semantic bias. This indicates that the model in this paper is more sensitive to core Confucian concepts.

5 Conclusions

This study systematically addresses semantic, cultural, and grammatical deviation detection issues in Chinese-Korean Confucian text translation by constructing a multi-task deviation identification framework based on PLMs. Experimental results demonstrate that this method significantly outperforms traditional SMT, NMT, and single pre-trained model baselines on public datasets, particularly exhibiting outstanding advantages in cultural term deviation identification.

The main theoretical contributions of this study are reflected in three aspects. First, a multi-granularity deviation identification theoretical framework tailored for low-resource language pairs is proposed, extending the traditional translation evaluation system centred on semantics to include dimensions of cultural adaptability and grammatical compliance, thereby providing a new theoretical perspective for cross-language text quality assessment. Second, a cultural term embedding enhancement mechanism is developed, which explicitly models the representation space of culture-specific concepts, effectively addressing the issue of insufficient representation in pre-trained models when processing culturally dense texts. Finally, a deviation quantification metric system tailored to East Asian language structures is established, addressing the shortcomings of

existing evaluation methods in supporting non-European language systems, and providing a quantifiable evaluation benchmark for related research.

Based on the research findings, we propose the following practical recommendations: at the technical application level, we suggest integrating the bias identification framework from this study into the post-editing process of machine translation as an auxiliary tool for translation quality control, with a focus on automatically screening and annotating cultural terms and grammatical structures. At the educational application level, a Confucian classics translation teaching system based on this technology could be developed to help learners understand the linguistic and cultural differences between Chinese and Korean through real-time bias feedback. In terms of resource development, we recommend that academic institutions collaborate to build an open Chinese-Korean Confucian terminology alignment database and establish translation standards for cultural terms. Additionally, we suggest that industry organisations incorporate cultural deviation metrics into translation quality assessment systems to advance standardisation in the field of cross-cultural communication.

Acknowledgements

This work is supported by the 2024 PhD Research Initiative Fund Project named: A Study on the Translation and Reception of Chinese Confucian Classics in Korea (No. 2024XJLXBsJJ004).

Declarations

All authors declare that they have no conflicts of interest.

References

Andriof, J. and Waddock, S. (2002) 'Unfolding stakeholder thinking: theory, responsibility and engagement', *Business & Society*, Vol. 5, No. 4, p.159.

Bailie, R., Si, D., Shannon, C., Semmens, J., Rowley, K., Scrimgeour, D.J., Nagel, T., Anderson, I., Connors, C. and Weeramanthri, T. (2010) 'Study protocol: national research partnership to improve primary health care performance and outcomes for indigenous peoples', *BMC Health Services Research*, Vol. 10, No. 1, pp.1–11.

Borg, S. (1999) 'The use of grammatical terminology in the second language classroom: a quality study of teachers' practices and cognitions', *Applied Linguistics*, Vol. 3, No. 1, pp.95–126.

Dervin, B. (1999) 'On studying information seeking methodologically: the implications of connecting metatheory to method', *Information Processing & Management*, Vol. 35, No. 6, pp.727–750.

Eoyang, E.C. (2005) 'Dragon-carving and the literary mind (review)', *China Review International*, Vol. 12, No. 2, pp.587–589.

Han, C. (2023) 'Professional characteristics and liberal arts character of classical literature education text: focusing on its status as a university's major education', *The Research of the Korean Classic*, Vol. 8, No. 9, p.130.

Hua-Xian, L. (2009) 'Sustainable development of cultural heritage based on eco-civilization: a case study of Hengxian county's sashimi culture', *Journal of Guangxi Normal University (Philosophy and Social Sciences Edition)*, Vol. 1, No. 7, p.369.

Ismail, M.N., Ngah, N.A. and Umar, I.N. (2010) 'The effects of mind mapping with cooperative learning on programming performance, problem solving skill and metacognitive knowledge among computer science students', *Journal of Educational Computing Research*, Vol. 42, No. 1, pp.35–61.

Kernick, D.P. (1997) 'Which antidepressant? A commentary from general practice on evidence-based medicine and health economics', *British Journal of General Practice the Journal of the Royal College of General Practitioners*, Vol. 47, No. 415, p.95.

Kim, Y-J. (2010) 'Establishment of the official uniform system in relation to organizing the centralized administration in the early Goryeo dynasty', *The Journal of Korean Medieval History*, Vol. 28, No.8, pp.439–484.

Labrador, B. (2011) 'A corpus-based study of the use of Spanish demonstratives as translation equivalents of English demonstratives', *Perspectives: Studies in Translatology*, Vol. 19, No. 1, pp.71–87.

Lisheng, X. (2008) 'From modern linguistics to post-modern linguistics', *Journal of Zhejiang University (Humanities and Social Sciences)*, Vol. 5, No. 4, p.134.

Melby, M.K., Anderson, D., Sievert, L.L. and Obermeyer, C.M. (2011) 'Methods used in cross-cultural comparisons of vasomotor symptoms and their determinants', *Maturitas*, Vol. 70, No. 2, pp.135–140.

Mohan, G.B., Kumar, R.P., Jayanth, K.K. and Doss, S. (2025) 'Telugu language analysis with XLM-Roberta: enhancing parts of speech tagging for effective natural language processing', *SN Computer Science*, Vol. 6, No. 2, p.169.

Molina, M. and Silva, M. (2015) 'Rapid determination of fungicides in fruit juices by micellar electrokinetic chromatography: use of organic modifiers to enhance selectivity and on-column high-salt stacking to improve sensitivity', *Electrophoresis*, Vol. 21, No. 17, pp.3625–3633.

Nguyen, P.M. and Terlouw, C. (2006) 'Culturally appropriate pedagogy: the case of group learning in a Confucian heritage culture context: intercultural education', *Intercultural Education*, Vol. 17, No. 1, p.479.

Ozumba, G. (2005) 'Gender-sensitivity in Igbo culture: a philosophical re-appraisal', *Quodlibet*, Vol. 17, No. 2, p.524.

Qin, Y., Wen, Q. and Wang, J. (2000) 'Automatic evaluation of translation quality using expanded n-gram co-occurrence', *IEEE*, Vol. 6, No 7, p.159.

Sahaji, A.U. (2021) 'The word 'noor': tracing a long journey through translation and adaptation from classical Arabic to contemporary Punjabi/Hindi pop songs', *Rupkatha Journal on Interdisciplinary Studies in Humanities*, Vol. 13, No. 2, p.352.

Shih-Chuan, C. (2011) 'A contrastive study of grammar translation method and communicative approach in teaching English grammar', *English Language Teaching*, Vol. 4, No. 2, p.189.

Wang, J., Liao, J., Zhou, Y. and Cai, Y. (2017) 'Differential evolution enhanced with multiobjective sorting-based mutation operators', *IEEE Transactions on Cybernetics*, Vol. 44, No. 12, pp.2792–2805.

Wekre, L.L. (2011) 'Need for a consensus on the methods by which to measure joint mobility and the definition of norms for hypermobility that reflect age, gender and ethnic-dependent variation: is revision of criteria for joint hypermobility syndrome and Ehlers-Danlos syndrome', *Rheumatology*, Vol. 50, No. 6, p.1169.

Wierzbicka, A. (1994) '"Cultural scripts": a semantic approach to cultural analysis and cross-cultural communication', *Behavior Patterns*, Vol. 15, No. 2, p.431.

Ye, Y. and Zhou, M. (2007) 'Sentence level machine translation evaluation as a ranking problem: one step aside from bleu', *Statmt Proceedings of the Second Workshop on Statistical Machine Translation*, Vol. 6, No. 8, p.168.