Proceedings of the 5th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/2026.30884

Readability Classification and Prediction of Machine Translation Texts Based on Machine Learning Algorithms

Yunhan Liu

Department of International Affairs, Shandong University, Jinan, China
ashley_liuyunhan@foxmail.com

Abstract. Against the backdrop of the continuous acceleration of global communication and cross-border dissemination of digital content, machine translation has become a core tool to break down language barriers and is widely applied in various scenarios such as cross-border business communication, academic literature sharing, and multilingual media dissemination. At present, most mainstream machine translation systems focus on ensuring semantic accuracy, but often neglect the key dimension of text readability, that is, whether the translation can conform to the reading habits, cognitive levels and specific scene requirements of the target language users. To make up for the deficiencies of the existing models, this paper proposes a BiLSTM classification algorithm based on the Red-billed blue magpie algorithm and the Attention mechanism optimization. The research first conducts violin graph analysis and correlation analysis, and then compares the performance of the proposed algorithm with that of various machine learning algorithms. The experimental results show that the RBMO-Attention-BiLSTM algorithm proposed in this paper exhibits the optimal performance in all evaluation indicators. Comprehensively surpassing comparison algorithms such as logistic regression, random forest, Naive Bayes, decision tree, XGBoost, BP neural network, ExtraTrees and CatBoost. This research provides a new technical solution for enhancing the readability and practical value of machine translation, which is of great significance for promoting the efficient application of machine translation systems in real-world scenarios and strengthening the accuracy and fluency of cross-language communication.

Keywords: Machine translation, machine learning, RBMO, BiLSTM

1. Introduction

With the acceleration of global communication and cross-border dissemination of digital content, machine translation has become a core tool to break down language barriers and is widely used in scenarios such as cross-border business communication, academic literature sharing, and multilingual media dissemination [1]. However, current machine translation systems mostly focus on semantic accuracy but often neglect text readability - that is, whether the translation conforms to the reading habits, cognitive levels and scene requirements of the target language users [2]. For instance, the excessive accumulation of complex sentence structures in the translation of scientific and technological literature, the stiff transplantation of terms in legal texts, or the improper

DOI: 10.54254/2753-8818/2026.30884

conversion of culturally loaded words in literary works can all lead to a significant decline in the readability of the translation. This not only increases the cost of user understanding but may also trigger information misunderstandings and communication risks [3]. Traditional readability assessment mostly relies on manual annotation or is based on fixed linguistic rules. It is not only time-consuming and labor-intensive, but also highly subjective. Moreover, it is difficult to adapt to diverse translation scenarios such as low-resource languages and professional fields, and cannot meet the real-time readability detection requirements of large-scale machine translation texts. There is an urgent need for efficient and intelligent automated assessment solutions.

Machine learning algorithms provide key technical support for the classification and prediction of the readability of machine translation texts, effectively addressing the limitations of traditional evaluation methods [4]. On the one hand, machine learning can automatically extract multi-dimensional features from massive translated texts - including average sentence length and syntactic complexity at the syntactic level, proportion of new words and term density at the lexical level, as well as fluency association features at the pragmatic level, without the need for manual presetting of rules, significantly enhancing the comprehensiveness and objectivity of feature extraction. On the other hand, different machine learning models can specifically address the core issues in readability classification: traditional machine learning models excel at handling structured features and are suitable for small-scale, domain-specific readability evaluations; Deep learning models can capture the context dependencies of text sequences and better model the interactive influence between syntax and semantics [5]. However, the existing models still have shortcomings: traditional models have a weak ability to capture the features of long text sequences, while BiLSTM is prone to getting stuck in local optimal solutions and pays insufficient attention to key linguistic features, resulting in limited classification accuracy in complex scenarios.

To make up for the deficiencies of the existing models, this paper proposes a BiLSTM classification algorithm based on the Red-billed Blue Magpie algorithm (RBMO) and the Attention mechanism optimization. The RBMO algorithm mimics the group foraging and information transmission behavior of the red-billed blue magpie. It has the advantages of strong global search ability and fast convergence speed. It can be used to optimize the initial weights and network structure parameters of BiLSTM, avoid the model falling into local optimum, and improve its adaptability to diverse translation texts. The Attention mechanism can dynamically allocate weights, enabling the model to focus on key features that have a significant impact on readability and enhancing the modeling ability of core information. The optimization algorithm formed by the combination of the two not only retains BiLSTM's ability to capture the context of text sequences, but also enhances the classification accuracy and generalization ability of the model through the parameter optimization of RBMO and the feature focusing of Attention, providing a new technical path for the automated and high-precision assessment of the readability of machine translation texts.

2. Data sources

The dataset used in this paper contains a total of 728 valid samples, covering 10 independent variables with clear linguistic significance and adaptability to actual translation scenarios, as well as 1 predictor variable. The independent variables comprehensively cover syntactic features (average sentence length, syntactic complexity), lexical features (proportion of unfamiliar words, average word length, term density), linguistic features (usage rate of logical conjunctions, grammatical differences between the source language and the target language), text attributes (domain type, source text difficulty), and translation quality-related features (fluency score). The predictor variable

DOI: 10.54254/2753-8818/2026.30884

is the readability classification label. It is divided into three categories: easy to read, medium to read and difficult to read.

Correlation analysis was conducted on each variable in the data and a correlation heat map was drawn, as shown in Figure 1. It can be known from the results of the correlation heat map that the positive correlation coefficient between source_diff_level and readability_label is as high as 0.899. Both conj_ratio and fluency_score are strongly negatively correlated with readability_label, and the coefficients are -0.897-0.896 respectively. syntax_complexity, unknown_word_ratio, avg_word_length, syntax_diff_score, avg_sent_len, term_density and readability_label All showed a strong positive correlation, and the coefficients were 0.883, 0.878, 0.877, 0.873, 0.870, and 0.854 in sequence.

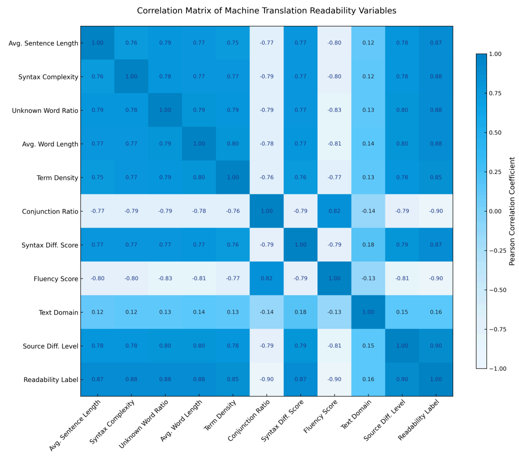

Figure 1. Correlation heat map

Draw the violin plots of each variable as shown in Figure 2. In the violin graph, the violin graphs of data distribution are respectively drawn according to the readability classification labels of each category. The distribution of the data can be observed.

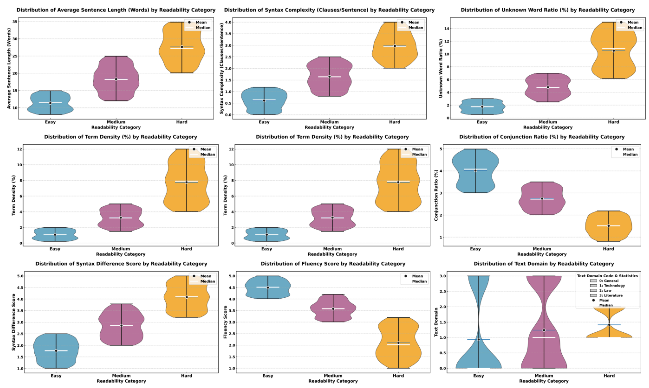

Figure 2. The violin plots of each variable

3. Method

3.1. RBMO

The Red-billed Blue Magpie algorithm is a bionic optimization algorithm that simulates the group behavior of red-billed blue magpies. Its core is to achieve global optimization through four

DOI: 10.54254/2753-8818/2026.30884

behavioral mechanisms. It has 2 to 5 individuals form collaborative groups to conduct local searches, and generates new solutions by using the average position of the groups and individual differences to enhance the exploration ability. When more than 10 individuals form a large group to conduct cluster search, the exploration and development are balanced based on the global average position [6]. The group attack behavior makes the individual move towards the optimal solution. By adjusting the step size through the adaptive coefficient, the dynamic switching between the initial exploration and the later development is achieved. The food storage mechanism retains historical optimal solutions, avoids the loss of global optimal information, and significantly improves the convergence speed and robustness of the algorithm in complex problems [7].

3.2. Attention mechanism

The Attention mechanism simulates the cognitive characteristics of human selective attention, with its core being the dynamic allocation of weights to different parts of the input sequence. It accomplishes the calculation by querying the vector Q, key vector K and value vector V: first, it calculates the similarity score between Q and K to measure the relevance of each input element to the current task; Then, the scores are normalized into weight distributions through the Softmax function, enabling important information to obtain higher weights. Finally, the weighted sum of V with weights is obtained to obtain the context vector focusing on the key information [8]. This mechanism resolves the problems of information overload in traditional models and the loss of information in fixed-length vectors, while enhancing the interpretability of the model. The network structure of the Attention mechanism is shown in Figure 3.

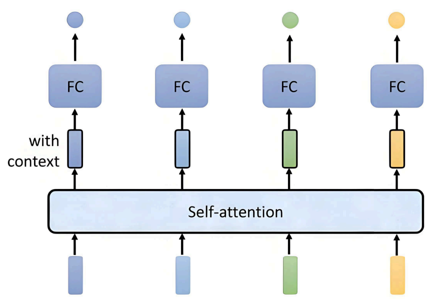

Figure 3. The network structure of the Attention mechanism

3.3. BiLSTM

BiLSTM, or Bidirectional Long Short-Term Memory Network, is an extended structure of LSTM, specifically designed to capture the bidirectional context of sequences [9]. It adds a reverse LSTM layer on the basis of the traditional LSTM. The forward layer processes data from the beginning to the end of the sequence and captures past information. The reverse layer processes data from the end to the beginning of the sequence and captures future information. The hidden state of each time step is generated by the fusion of forward and reverse outputs. Through the gating mechanism, the vanishing gradient problem is effectively suppressed, and the long-term dependency relationships in long sequences can be accurately extracted [10]. The network structure of BiLSTM is shown in Figure 4.

DOI: 10.54254/2753-8818/2026.30884

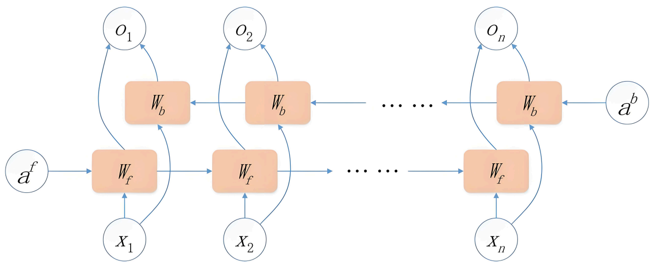

Figure 4. The network structure of BiLSTM

3.4. RBMO-Attention-BiLSTM

RBMO-Attention-BiLSTM is a classification model that integrates optimization algorithms and deep learning. The input data is first transformed into a low-dimensional vector by the embedding layer and sent to the BiLSTM layer to extract bidirectional context features, thereby obtaining the hidden state set of the entire sequence. The Attention layer calculates weights for these hidden states, focuses on the feature information crucial for classification, and generates weighted sentence-level feature vectors. The RBMO algorithm is used to optimize model parameters. Through its multi-strategy search mechanism, it prevents the model from falling into local optimum and balances the comprehensiveness of feature extraction with the focus of key information. Finally, the weighted feature vectors are sent to the fully connected layer, and the classification results are output through Softmax, significantly improving the classification accuracy of complex sequence data.

4. Result

The parameter Settings for this project are as follows: The population size is 12, the maximum number of iterations is 6, and the lower limits of the three parameters to be optimized (L2 regularization coefficient, initial learning rate, and the number of hidden layer nodes) are 1e-5, 0.0001, and 10 respectively, while the upper limits are 1e-2, 0.002, and 50 respectively. Among the model training parameters, the maximum training round is 500, the gradient threshold is 1, the learning rate adjustment period is 400, the learning rate adjustment factor is 0.2, the dropout rate is 0.01, the training environment is cpu, the optimization algorithm adopts adam, and the learning rate scheduling method is segmented.

In the comparison model, this paper uses Logistic regression, Random Forest, Naive Bayes, Decision tree, XGBoost and BP neural network, ExtraTrees and CatBoost are used as comparison models. The comparison results of the indicators of each model are shown in Table 1.

Table 1. The results of the comparative experiment

Model	Accuracy	Recall	Precision	F1	AUC
Logistic regression	0.95	0.95	0.951	0.95	0.957
Random Forest	0.95	0.95	0.952	0.95	0.977
Naive Bayes	0.959	0.959	0.959	0.959	0.976
Decision tree	0.922	0.922	0.93	0.924	0.942
XGBoost	0.963	0.963	0.965	0.964	0.987
BP neural network	0.973	0.973	0.974	0.973	0.983

DOI: 10.54254/2753-8818/2026.30884

Table 1. (continued)

ExtraTrees	0.968	0.968	0.968	0.968	0.975
CatBoost	0.963	0.963	0.964	0.964	0.977
Our model	0.982	0.982	0.983	0.982	0.989

Output the bar comparison charts of each indicator, as shown in Figure 5. It can be seen from the results that the RBMO-Attention-BiLSTM algorithm proposed in this paper demonstrates the best performance in all evaluation indicators, comprehensively surpassing the comparison algorithms such as logistic regression, random forest, Naive Bayes, decision tree, XGBoost, BP neural network, ExtraTrees and CatBoost. In terms of accuracy, the RBMO-Attention-BiLSTM algorithm reaches 98%, which is much higher than the 92% of decision trees, the 96% of Naive Bayes, XGBoost and CatBoost, as well as the 97% of ExtraTrees and BP neural networks. It is also superior to 95% of logistic regression and random forest. In terms of recall rate, precision rate and F1 value, the RBMO-Attention-BiLSTM algorithm is all 98%, significantly higher than other algorithms. Among them, the three indicators of ExtraTrees and BP neural network are all 97%. It performs best except for the RBMO-Attention-BiLSTM algorithm, but still lags behind the RBMO-Attention-BiLSTM algorithm. In terms of AUC value, the RBMO-Attention-BiLSTM algorithm reaches 99%, slightly higher than the 99% of XGBoost and far exceeding the 94% of decision trees. It is also higher than 96% of logistic regression, 98% of random forest, Naive Bayes, BP neural network, ExtraTrees and CatBoost, demonstrating stronger classification performance and generalization ability overall.

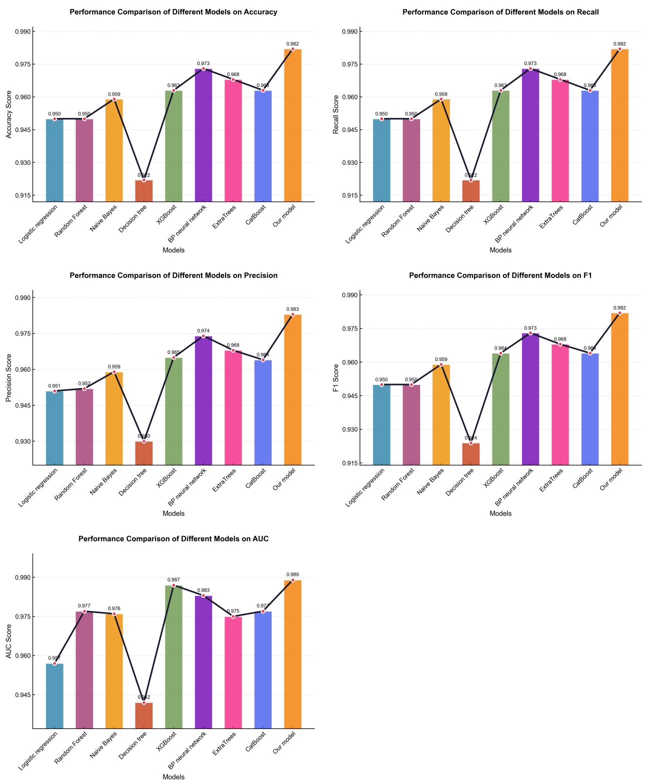

Figure 5. The bar comparison charts of each indicator

Output the prediction confusion matrix of the RBMO-Attention-BiLSTM model test set, as shown in Figure 6.

DOI: 10.54254/2753-8818/2026.30884

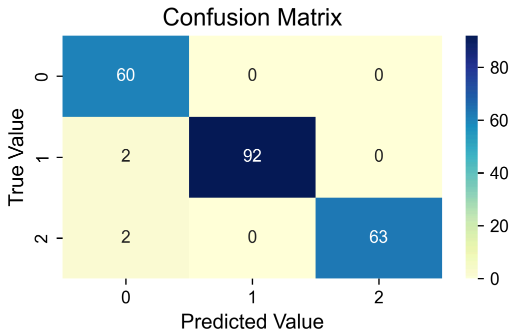

Figure 6. The RBMO-Attention-BiLSTM model test set

5. Conclusion

Against the backdrop of deepening global exchanges and the continuous acceleration of cross-border digital content dissemination, machine translation has become the core support for breaking down language barriers and is widely applied in various scenarios such as cross-border business communication, academic literature sharing, and multilingual media dissemination. However, most current machine translation systems often focus on semantic accuracy but frequently overlook the crucial dimension of text readability - whether the translation can align with the reading habits, cognitive levels, and specific scene requirements of the target language users. To fill this defect of the existing models, this paper proposes a BiLSTM classification algorithm based on the Red-billed blue magpie algorithm and the Attention mechanism optimization. During the research process, violin graph analysis and correlation analysis were first carried out. Subsequently, the performance of this algorithm was compared with multiple machine learning algorithms. The results showed that the RBMO-Attention-BiLSTM algorithm proposed in this paper exhibited the optimal performance in all evaluation indicators. Comprehensively surpassing comparison algorithms such as logistic regression, random forest, Naive Bayes, decision tree, XGBoost, BP neural network, ExtraTrees and CatBoost. This achievement provides an efficient algorithmic solution for enhancing the readability of machine-translated texts, which is of great practical significance for promoting better adaptation of machine translation to actual application scenarios and enhancing the effectiveness of cross-language communication.

References

[1] Wang, Qingran, and Jun Xu. "Neural machine translation in AVT teaching in China: An in-depth analysis from the readability perspective." Linguistica Antverpiensia, New Series–Themes in Translation Studies 22 (2023).

[2] Choi, HyunJung, et al. "Spoken-to-written text conversion for enhancement of Korean–English readability and machine translation." ETRI Journal 46.1 (2024): 127-136.

[3] Blaneck, Patrick Gustav, et al. "Automatic readability assessment of German sentences with transformer ensembles." arXiv preprint arXiv: 2209.04299 (2022).

[4] Park, Dojun, and Sebastian Padó. "Multi-dimensional machine translation evaluation: model evaluation and resource for Korean." arXiv preprint arXiv: 2403.12666 (2024).

[5] Kong, Marianna, et al. "Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis." BMJ quality & safety (2025).

[6] Dong, Jun. "Transfer learning-based neural machine translation for low-resource languages." ACM Transactions on Asian and Low-Resource Language Information Processing (2023).

[7] Elmadani, Khalid, Nizar Habash, and Hanada Taha. "A large and balanced corpus for fine-grained Arabic readability assessment." Findings of the Association for Computational Linguistics: ACL 2025. 2025.

DOI: 10.54254/2753-8818/2026.30884

[8] North, Kai, Marcos Zampieri, and Matthew Shardlow. "Lexical complexity prediction: An overview." ACM Computing Surveys 55.9 (2023): 1-42.
[9] Lan, Zhibin, et al. "Translatotron-v (ison): An end-to-end model for in-image machine translation." arXiv preprint arXiv: 2407.02894 (2024).
[10] Minutolo, Aniello, et al. "A multi-level methodology for the automated translation of a coreference resolution dataset: an application to the italian language." Neural Computing and Applications 34.24 (2022): 22493-22518.