Study on HSK Chinese Text Grading Based on Deep Hybrid Neural Network Model

Anonymous ACL submission

Abstract

This paper proposes a deep hybrid neural network model for HSK text grading, aiming to accurately assess the difficulty levels of Chinese texts. The model adopts a hierarchical processing architecture, including a character embedding layer, a multi-scale feature extraction module, a bidirectional LSTM sequence encoding module, a multi-head self-attention mechanism module, and a classification output module. By organically integrating scale convolutional feature extraction, sequence modeling, and attention mechanisms, the model effectively captures complex linguistic characteristics for proficiency assessment. We construct a dataset of 4,474 texts based on the latest HSK standard textbooks and sample exams, aligned with the International Chinese Education Chinese Proficiency Level Standard (HSK9), using a combined "HSK9 + sentence length" and "maximum distribution density segmentation" method. Experiments show that the model achieves an accuracy of 94.17% and a weighted F1 score of 94.17%, significantly outperforming baseline models including BERT (92.32%), MLF-BERT (91.05%), and KGMNN (80.36%). Ablation studies confirm the contributions of each component, with the CNN module contributing 4.85 percentage points to accuracy and the attention mechanism contributing 4.59 percentage points.

1 Introduction

With the rapid development of international Chinese education, accurately assessing the difficulty level of Chinese texts has become increasingly important for textbook compilation, instructional design, and language testing. Traditional text grading methods often rely on manual annotation and statistical formulas, which are subjective, inefficient, and poorly transferable across corpora.

In recent years, deep learning techniques have been widely applied in natural language processing, and neural network-based text grading models have gradually become a research focus. These models automatically learn deep linguistic features and can more accurately estimate text difficulty. However, existing studies still suffer from three major limitations:

- **Lack of deep hybrid models for language proficiency assessment**: Most current models are designed for general text classification or short-text tasks, rather than for language proficiency grading, which involves more complex dimensions such as lexical sophistication, syntactic diversity, and discourse coherence.
- **Insufficient semantic depth in Chinese grading tasks**: Existing models often focus on shallow linguistic markers or entity boundaries. However, advanced HSK texts exhibit high syntactic and semantic complexity, requiring modeling of long-distance dependencies.
- **Limitations in data and evaluation**: There is a lack of widely recognized datasets aligned with the latest HSK Level 9 standard. Most studies use closed-domain data, and there is a shortage of publicly available baseline models. Evaluation metrics are often single, failing to capture the multidimensional nature of language proficiency.

To address these issues, this paper proposes a deep hybrid neural network model for HSK text grading. The main contributions are:

- **Dataset Construction**: We build a dataset based on the latest HSK standard textbooks

and sample exams, aligned with the International Chinese Education Chinese Proficiency Level Standard (HSK9).

- **Deep Semantic Modeling**: The proposed hierarchical hybrid model extracts features at multiple levels. CNNs capture local lexical-syntactic patterns, LSTMs model discourse-level coherence, and attention mechanisms focus on key proficiency dimensions.
- **Multi-Granularity Fusion**: The model integrates n-gram features from CNNs and sequence features from LSTMs, surpassing single-level feature extraction.
- **Empirical Validation**: Experiments on a large-scale HSK dataset show that the model outperforms strong baselines in accuracy and F1-score.

2 Related Work

2.1 Linear Regression Formulas

Early work adapted English readability formulas (e.g., Flesch) to Chinese. Wang Lei (2005) used average sentence length, HSK3 vocabulary coverage, and clause count to build a linear model with 80.3% explained variance (Du et al., 2022). Guo Wanghao (2010) introduced "character count" as an independent variable, achieving $R^2 = 0.917$ (Guo, 2010). However, linear models fail to capture feature interactions and suffer from threshold dependency and corpus drift (Zuo and Zhu, 2014).

2.2 Machine Learning Models

Wu Siyuan (2020) proposed a 104-dimensional feature pool spanning "character, vocabulary, syntax, and discourse" levels (Wu et al., 2020). Yang Wendi (2020) constructed an 86-dimensional feature set across four dimensions (Yang, 2020). Wu Jifeng et al. (2025) performed a large-scale analysis of 700 reading texts, systematically validating the discriminatory power of 12 syntactic complexity indicators (wu2).

2.3 Deep Learning Models

Xue Xingrong and Jin Qibing (2022) integrated contextual and local features via a BiLSTM-CNN-Attention model (Xue and Jin, 2022). Cao Xiang et al. (2019) validated hybrid neural networks for short text classification (**?**). Chen Kejin et al. (2018) combined CNN and RNN with an attention mechanism for question classification (**?**).

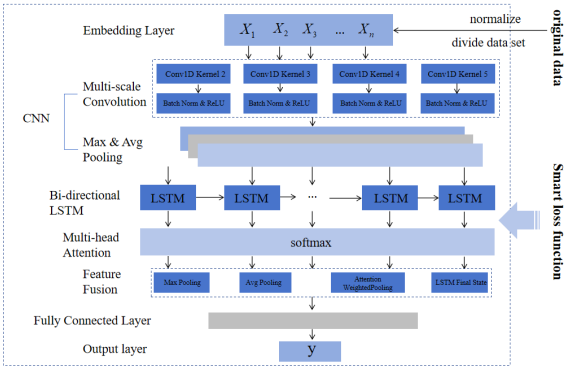

Figure 1: Architecture of the proposed deep hybrid neural network for HSK text grading.

2.4 Large Language Models

Yin Xiaojun (2024) utilized ChatGPT to generate simplified texts for elementary levels (**?**). Han Xinxin (2025) found that the variance of log probabilities of texts generated by DeepSeek significantly correlated with human-assigned grades ($r = 0.63$) (**?**).

3 HSK Text Grading Model

3.1 Overall Architecture Design

This paper proposes an automatic HSK text grading model based on a deep hybrid neural network. By integrating convolutional neural networks, bidirectional long short-term memory networks, and a multi-head self-attention mechanism, the model constructs an end-to-end classifier aimed at accurately assessing the language complexity of Chinese texts.

The deep hybrid model follows a hierarchical feature extraction and fusion pipeline. The input text is first converted into a character-level index sequence and mapped to dense vector representations via a trainable embedding layer. Subsequently, this sequential representation is processed in parallel by three core components:

1. Multi-scale Convolutional Neural Network
2. Bidirectional Long Short-Term Memory Network
3. Multi-head Self-Attention Mechanism

3.2 Detailed Explanation of Core Hybrid Components

3.2.1 Multi-scale Convolutional Feature Extractor

For an embedded sequence $\mathbf{Z} \in \mathbb{R}^{B \times L \times d_{\text{embed}}}$, we transpose it and feed it into multiple parallel one-dimensional convolutional layers:

$$\mathbf{C}_k = \text{GELU}(\text{BatchNorm1D}(\text{Conv1D}_k(\mathbf{Z}^\top))) \quad \text{for } k \in \{2, 3, 4\} \tag{1}$$

where each Conv1D_k produces $C_{\text{conv}} = 128$ feature maps.

3.2.2 Sequence Context Encoder (BiLSTM)

For an $N_{\text{layer}} = 3$ L-layer BiLSTM:

$$\begin{aligned} \overrightarrow{\mathbf{h}}_t^{(l)}, \overrightarrow{\mathbf{c}}_t^{(l)} &= \text{LSTM}^{(l)}(\mathbf{z}_t^{(l-1)}, \overrightarrow{\mathbf{h}}_{t-1}^{(l)}, \overrightarrow{\mathbf{c}}_{t-1}^{(l)}) \\ \overleftarrow{\mathbf{h}}_t^{(l)}, \overleftarrow{\mathbf{c}}_t^{(l)} &= \text{LSTM}^{(l)}(\mathbf{z}_t^{(l-1)}, \overleftarrow{\mathbf{h}}_{t+1}^{(l)}, \overleftarrow{\mathbf{c}}_{t+1}^{(l)}) \\ \mathbf{h}_t^{(l)} &= [\overrightarrow{\mathbf{h}}_t^{(l)}; \overleftarrow{\mathbf{h}}_t^{(l)}] \end{aligned} \tag{2}$$

3.2.3 Global Semantic Weight Re-estimator (Multi-Head Attention)

For the i-th head:

$$\text{head}_i = \text{softmax}\left(\frac{(\mathbf{H}_{\text{lstm}}\mathbf{W}_i^Q)(\mathbf{H}_{\text{lstm}}\mathbf{W}_i^K)^\top}{\sqrt{d_k}} + \mathcal{M}\right) \times (\mathbf{H}_{\text{lstm}}\mathbf{W}_i^V) \tag{3}$$

where $d_k = d_v = 48$.

3.3 Feature Fusion and Decision Making

Features from different modalities are concatenated:

$$\mathbf{F}_{\text{total}} = [\mathbf{F}_{\text{conv}}; \mathbf{F}_{\text{lstm}}; \mathbf{F}_{\text{attn}}] \tag{4}$$

This hybrid feature is then passed through a three-layer feature fusion network ($1920 \to 768 \to 384 \to 192$).

3.4 Optimization Objective

The model's total loss function combines a modified focal loss with multi-task learning:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda \cdot \mathcal{L}_{\text{aux}} \tag{5}$$

The main loss incorporates label smoothing:

$$\mathcal{L}_{\text{main}} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{7} \alpha_c (1 - y_{\text{main},b,c})^\gamma \times y_{b,c}^{\text{LS}} \log(y_{\text{main},b,c}) \tag{6}$$

with $\gamma = 2.0$, $\epsilon = 0.1$.

Difficulty Level	Number of Texts
Level 1	540
Level 2	723
Level 3	1045
Level 4	784
Level 5	1011
Level 6	371
Total	4474

Table 1: Number of Articles per Level in the Final Filtered Dataset

Model	Text_Acc/%	Text_F1/%
KGMNN	80.36	80.35
MLF-BERT	91.05	91.06
BERT	92.32	87.20
CNN+LSTM+Attention	94.17	94.17

Table 2: Experimental Comparison Results with Other Models

4 Experiments and Analysis

4.1 Experimental Dataset

We compiled 6,269 texts from standard HSK textbooks and tests. After filtering using the "HSK9 + Sentence Length" and "maximum distribution density segmentation" method, we obtained a final dataset of 4,474 texts across HSK Levels 1-6.

4.2 Experimental Setup

The dataset was split into training, validation, and test sets in an 8:1:1 ratio. Key hyperparameters: word vector dimension=256, max sequence length=200, BiLSTM layers=3, attention heads=8, optimizer=AdamW (lr=8e-5).

4.3 Baseline Model Comparison

We compare with three strong baselines: BERT, MLF-BERT , and KGMNN .

4.4 Results and Analysis

Our model achieves an accuracy and weighted F1 score of 94.17% on the test set, outperforming all baselines (Table 2).

4.5 Ablation Study and Results

Ablation studies confirm the contributions of each component (Table 3).

The ablation experiment results indicate that the full model achieved an accuracy of 94.17% and an

Model Variant	Text_Acc/%	ΔAcc
CNN+LSTM+Attention	94.17	-
-CNN	89.32	-4.85
-Attention	89.58	-4.59
-CNN & -Attention	85.42	-8.75

Table 3: Ablation Study Results

F1 score of 94.17%, significantly outperforming all simplified variants. Removing the CNN module led to a 4.85 percentage point decrease, indicating its key role in local feature extraction. Removing the attention mechanism led to a 4.59 percentage point decrease, demonstrating the necessity of adaptive weight allocation.

5 Conclusion

This paper proposes an HSK text leveling model based on a deep hybrid neural network. By organically combining multi-scale convolutional feature extraction, Bi-directional LSTM sequence encoding, and multi-head self-attention mechanisms, accurate evaluation of Chinese text difficulty is achieved. Experimental results show that the model achieves an accuracy of 94.17% and a weighted F1 value of 94.17% on the test set, significantly outperforming baseline models. Ablation experiments further verify the contribution of each component. This study provides an efficient and accurate deep learning method for HSK text leveling.

Future work will further explore model optimization and the application potential of this model in other language text leveling tasks.

Limitations

Currently, there are relatively few texts for HSK7-9 difficulty levels. In the future, we will increase the scanning of HSK7-9 difficulty level texts, expand the dataset, and enrich research.

Acknowledgments

I would like to express my deepest gratitude to my research supervisor, whose profound expertise in computational linguistics and unwavering support have been instrumental throughout this research. His insightful critiques, patient guidance, and constant encouragement were invaluable in navigating the challenges of this work on deep hybrid models for HSK-level classification. His rigorous academic standards have profoundly shaped my growth as a researcher.

My sincere thanks also go to the members of the Intelligent Language Processing Lab at my University for their stimulating discussions, collaborative spirit, and constructive feedback during our regular seminars. The intellectually rich environment they fostered was crucial for refining the ideas presented in this paper.

Finally, my heartfelt appreciation goes to my family and friends for their unconditional love, understanding, and endless patience during the demanding phases of this work. Their support has been my anchor and inspiration.

References

Syntactic and phraseological complexity in chinese as a second language adapted teaching materials.

Yueming Du and 1 others. 2022. Research on automatic readability assessment of hsk reading texts. *Applied Linguistics*, 2022(3):73–86.

Wanghao Guo. 2010. *A Study on the Readability Formula for Chinese as a Foreign Language Texts*. Ph.D. thesis, Shanghai Jiao Tong University. [Online; accessed 2025-04-25].

Siyuan Wu, Dong Yu, and Xin Jiang. 2020. Construction and validity verification of a chinese text readability feature system. *World Chinese Teaching*, 34(1):81–97.

Xingrong Xue and Qibing Jin. 2022. A deep learning-based chinese text classification algorithm. *Computer and Digital Engineering*, 50(1):111–115.

Wendi Yang. 2020. *Readability Assessment of Chinese as a Foreign Language Texts Based on Multi-dimensional Features and Random Forest*. Ph.D. thesis, Central China Normal University. [Online; accessed 2025-04-25].

Hong Zuo and Yong Zhu. 2014. A readability formula for intermediate-level european and american learners of chinese. *World Chinese Teaching*, 28(2):263–276.