InterpretARA: Enhancing Hybrid Automatic Readability Assessment with Linguistic Feature Interpreter and Contrastive Learning

Jinshan Zeng[1], Xianchao Tong[1], Xianglong Yu[1], Wenyan Xiao[2], Qing Huang[1]*

[1]Jiangxi Normal University
[2]Jiangxi University of Science and Technology
jinshanzeng@jxnu.edu.cn, xianchaotong@jxnu.edu.cn, xianglongyu@jxnu.edu.cn,wy.xiao@jxust.edu.cn,qh@jxnu.edu.cn

Abstract

The hybrid automatic readability assessment (ARA) models that combine deep and linguistic features have recently received rising attention due to their impressive performance. However, the utilization of linguistic features is not fully realized, as ARA models frequently concentrate excessively on numerical values of these features, neglecting valuable structural information embedded within them. This leads to limited contribution of linguistic features in these hybrid ARA models, and in some cases, it may even result in counterproductive outcomes. In this paper, we propose a novel hybrid ARA model named **InterpretARA** through introducing a linguistic interpreter to better comprehend the structural information contained in linguistic features, and leveraging the contrastive learning that enables the model to understand relative difficulty relationships among texts and thus enhances deep representations. Both document-level and segment-level deep representations are extracted and used for the readability assessment. A series of experiments are conducted over four English corpora and one Chinese corpus to demonstrate the effectiveness of the proposed model. Experimental results show that InterpretARA outperforms state-of-the-art models in most corpora, and the introduced linguistic interpreter can provide more useful information than existing ways for ARA.

Introduction

Readability assessment (RA) quantifies the difficulty level of a text, and particularly measures how easy or difficult it is to read and comprehend the content (Laughlin 1969; Klare 2000). Research on automatic readability assessment (ARA) can be traced back to the last century (Lively and Pressey 1923; Klare 1963). ARA systems automatically assign difficulty levels to texts, making them widely applicable in various domains such as the enhancement of reading skills (Pera and Ng 2014), facilitation of second language learning (Qiu et al. 2021), and comprehension of clinical informed consent documents (Perni et al. 2019). As one of the earliest and systematically developed approaches in the field of linguistics, RA has developed various linguistic features and has been extensively studied using both traditional machine learning (Deutsch, Jasbi, and Shieber 2020; Hansen et al. 2021; Lee, Jang, and Lee 2021) and deep learning methods (Ma, Fosler-Lussier, and Lofthus 2012; Devlin et al. 2019; Li, Ziyang, and Wu 2022).

Linguistic features are indicators used to describe the linguistic properties of texts (Dell'Orletta, Montemagni, and Venturi 2011; Hancke, Vajjala, and Meurers 2012). In ARA, researchers often utilize various linguistic features to quantify the difficulty level of a text (Flesch 1948; Chall 1948). These features can capture different aspects of the linguistic nature of a text, including vocabulary, syntax, semantics, and discourse structures (Dell'Orletta, Montemagni, and Venturi 2011; Hancke, Vajjala, and Meurers 2012; Sung et al. 2015; Denning, Pera, and Ng 2016; Arfé, Mason, and Fajardo 2018; Jiang et al. 2019). Due to the richness of linguistic features, they are incorporated in both traditional machine learning and modern deep learning approaches. For example, a large number of linguistic features are designed in machine learning models to assess text readability (Deutsch, Jasbi, and Shieber 2020; Hansen et al. 2021; Lee, Jang, and Lee 2021). In particular, hybrid ARA models combining both linguistic and deep features for RA have recently attracted amounts of attention due to their impressive performance (Lee, Jang, and Lee 2021; Li, Ziyang, and Wu 2022).

Despite significant improvements achieved by current hybrid models over both traditional machine learning models and deep learning models solely using linguistic or deep features (Lee, Jang, and Lee 2021; Li, Ziyang, and Wu 2022), the potential of linguistic features has not yet been fully explored, as most of ARA models often focus on numerical values of these linguistic features, overlooking the valuable structural insights embedded within them. This limits the contribution of linguistic features in the kind of hybrid models. For instance, when a feature value of some linguistic feature is fed into the model, the model is hard to grasp the specific textual information that this numerical value represents. During the training process, the model can only judge the relationship between feature values and text readability based on numerical differences. This excessive emphasis on numerical values neglects the crucial structural information carried by these features. As a result, the contribution of linguistic features to the hybrid ARA models is limited, sometimes even resulting in counterproductive outcomes. Moreover, the difficulty correlation among different texts has not yet been fully investigated in existing models (Lee, Jang, and Lee 2021; Li, Ziyang, and Wu 2022).

*Corresponding author

To address these issues, we first introduce a linguistic interpreter to capture more structural information such as the semantic information, and thus enrich these features. This interpreter reinterprets these linguistic features in natural language forms that contain much valuable structural information of linguistic features, and facilitates to deal with linguistic features through a deep model. Besides, we employ contrastive learning (Khosla et al. 2020) to learn the difficulty correlation among texts, which is helpful for extracting deep representations. The major contributions of this paper can be summarized as follows.

- This paper introduces a novel hybrid ARA model called **InterpretARA**, which enriches linguistic representations by incorporating a linguistic interpreter and enhances deep representations through the utilization of contrastive learning. Additionally, we employ segment-level representations to capture local features of texts.
- The introduced linguistic interpreter in this paper utilizes natural language descriptions to achieve detailed explanations of linguistic features. These interpretations not only greatly enrich the structural and linguistic information of linguistic features but also provide possibilities for their processing through deep models.
- The effectiveness of the proposed model is validated on four English corpora and one Chinese corpus. Experimental results demonstrate that the proposed model outperforms state-of-the-art models in most corpora, and the introduced linguistic interpreter can more effectively leverage linguistic features than existing ways, leading to substantial improvements in model performance.

Related Work

Early studies in readability assessment primarily involved linguists defining various linguistic features and creating readability formulas through educational and psychological methods, including Flesch (Flesch 1948), Dale-Chall (Chall 1948), and SMOG (McLaughlin 1969). Although these formulas have advantages of simplicity and ease of interpretation, they fall short in representing certain complex features such as the structure and semantic complexity of a text. This limitation restricts the performance of these readability formulas.

Later, statistical machine learning methods have been applied to readability assessment. These methods incorporate some classical machine learning models such as the support vector machines (SVM) into statistical language models (LM) with parsing trees, vocabulary, and features related to semantics and syntax (Lu, Qiu, and Cai 2019). In Lu, Qiu, and Cai (2019), the authors conducted experiments analyzing the impact of 88 linguistic features on sentence complexity. The results in Lu, Qiu, and Cai (2019) showed that these statistical machine learning models using linguistic features significantly outperform existing readability formulas. However, the performance of these statistical machine learning models is limited by the capacity of the used linguistic features (Deutsch, Jasbi, and Shieber 2020).

With the development of deep learning, deep learning methods have shown remarkable performance in readability assessment tasks due to the powerful approximation ability of deep neural networks. Azpiazu and Pera (2019) proposed a hierarchical attention network (HAN) model for ARA based on multi-attention recurrent neural network architecture. Zeng et al. (2022) proposed a refined model through employing the pre-trained BERT in both word encoder and sentence encoder, and introducing soft labels to take advantage of the ordinal regression nature of ARA, as well as a novel pre-training task for initialization. Despite achieving commendable performance in text readability assessment, deep learning models generally overlook the role of linguistic features or do not fully explore their potential.

Motivated by advantages of linguistic features, such as their good interpretability for ARA, the kind of hybrid ARA models combining both linguistic and deep features has recently attracted rising attention (Lee, Jang, and Lee 2021; Li, Ziyang, and Wu 2022). In Lee, Jang, and Lee (2021), the authors constructed a correlation graph among features, representing linguistic features as nodes and their correlations as edges. In Li, Ziyang, and Wu (2022), the authors integrated difficulty knowledge to extract topic features with anchored correlation explanation, and fused linguistic feature with deep representations by projection filtering.

Unlike previous studies, we designed a linguistic feature interpreter to enable the model to better grasp the structural information embedded within linguistic features, and integrated contrastive learning to allow the model to comprehend the relative difficulty relationships among texts. By extracting deep features from both the document-level and segment-level perspectives, we acquired more enriched and comprehensive representations for ARA.

Proposed Model

The overall architecture of the proposed model is depicted in Figure 1. We yield enriched linguistic representations by introducing a linguistic feature interpreter, and extract segment-level representations embodying the local strucures of texts with pre-trained BERTs, and obtain enhanced document-level representations by a pre-trained BERT together with the contrastive learning, and finally fuse these representations for ARA according to the similar projection filtering way in Li, Ziyang, and Wu (2022).

Enriched Linguistic Representations

Linguistic features can quantify the difficulty level of a text and provide additional textual information to enhance the performance of deep learning models. To yield more structural information of linguistic features, we introduce a linguistic feature interpreter. For a given text, we firstly yield its several types of linguistic features such as discourse, syntactic, lexical and shallow features by associated linguistic feature extractors or formulas, then feed them into the introduced linguistic feature interpreter to yield their corresponding interpretations in natural language forms, and finally obtain enriched linguistic representations f_α with pre-trained BERTs followed by a mix pooling and a linear layer, where all pre-trained BERTs share the same parameters.

In the linguistic feature interpreter, we design some novel interpretation templates for linguistic features, where these

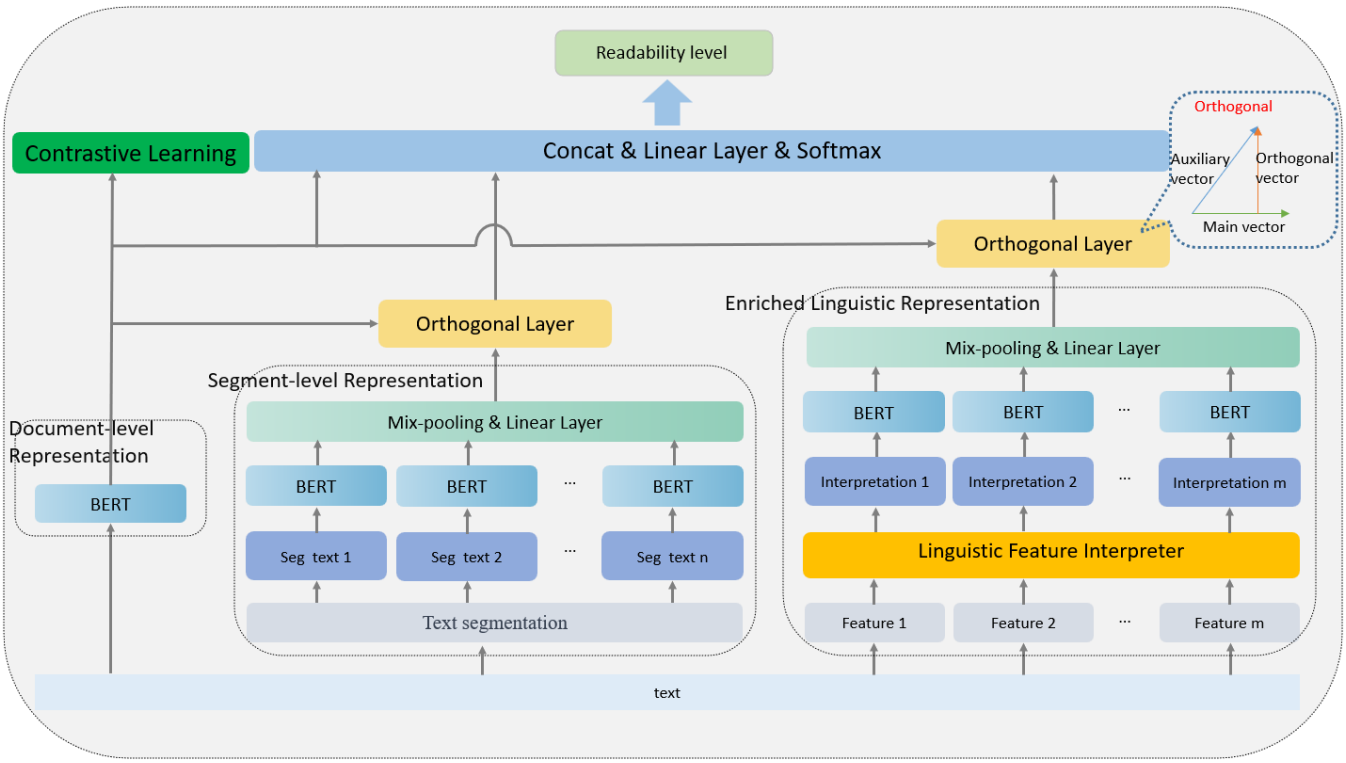


Figure 1: The overall structure of our proposed model for readability assessment.

Feature name	Template
average count of characters per sentence	The average count of characters per sentence is [value].
average count of syllables per sentence	The average count of syllables per sentence is [value].
average count of Content words per sentence	The average count of Content words per sentence is [value].
average count of noun phrases per sentence	The average count of noun phrases per sentence is [value].

Table 1: Examples of interpretation templates for linguistic features.

feature templates are generally designed to be the concatenations of feature names and values. Some examples of templates are presented in Table 1 and more are shown in Supplementary materials. From Table 1, we take the feature "average count of characters per sentence" as an example. Instead of the sole use of the numerical value of this linguistic feature, we reinterpret this feature with its feature name and value in natural language way. By leveraging such natural language description way, linguistic features can be exploited more thoroughly by deep models than existing ways.

Deep Representations

The deep representation plays a crucial role in text readability assessment. Deep representations are capable of capturing the semantics, structure, and contextual information embedded in the raw text data, thereby aiding the model in better understanding and processing text. To fully leverage the deep-level features of texts, we employ pre-trained BERTs (Devlin et al. 2019) to obtain both document-level and segment-level representations.

The document-level representations f_β are yielded by feeding the text into a pre-trained BERT, which enables better understandings of semantic and contextual relationships.

Although the document-level representations can provide global information for a text, some kinds of local information such as correlations among different sentences and words within a paragraph important for the readability assessment are ignored. To capture these kinds of local information of a text, we introduce the segment-level representations f_γ, inspired by the literature (Li, Ziyang, and Wu 2022). Specifically, given a text, we firstly divide this text into multiple segments with the same number of sentences, then feed them into pre-trained BERTs to extract their individual representations, where the pre-trained BERTs in this module share parameters of the pre-trained BERT used in the document-level representation module, and finally yield the segment-level representations f_γ by a mix-pooling operation followed by a linear layer.

Projection Filtering Fusion

Noticing that the document-level, segment-level and enriched linguistic representations are extracted from the same

text from three different perspectives, there is redundant information among them. In order to get rid of the redundancy, we exploit a simple orthogonal projection filtering scheme to fuse these kinds of representations. Specifically, we regard the document-level representations f_β as the base representations, and then yield two representations orthogonal to the document-level representations individually from both segment-level representations f_γ and enriched linguistic representations f_α. Specifically, the orthogonal projections of f_α and f_γ onto f_β, i.e., $f_\alpha^\perp$ and $f_\gamma^\perp$ can be yielded according to the following:

$$f_\alpha^\perp = f_\alpha - \frac{f_\alpha \cdot f_\beta}{|f_\beta|^2} f_\beta,\ f_\gamma^\perp = f_\gamma - \frac{f_\gamma \cdot f_\beta}{|f_\beta|^2} f_\beta.$$

With these projection filtering operations, the redundancy among these kinds of representations has been significantly removed. After these orthogonal projection operations, we yield the final text representations for readability assessment by concatenating $f_\alpha^\perp$, f_β and $f_\gamma^\perp$, followed by a linear layer and the softmax operation.

Contrastive Learning and Training Loss

Text readability assessment aims to understand the difficulty level of texts. The correlations of different texts within the same readability levels and among the different readability levels should be important for the readability assessment. In order to adaptively capture these correlations, we introduce the known contrastive learning (Khosla et al. 2020) to further enhance text representations through learning relative difficulty relationships among texts.

Specifically, the correlations among different texts are captured by maximizing the similarity between document-level representations within the same difficulty level and minimizing the similarity between document-level representations among different difficulty levels.

$$\mathcal{L}_{cl} = \frac{1}{N} \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} -\log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in \mathcal{A}_i} \exp(z_i \cdot z_a / \tau)},$$

where $\tau \in R^+$ is the temperature factor, $\mathcal{A}_i$ is the index set of contrasting samples, $\mathcal{P}_i \subseteq \{p \in \mathcal{A}_i : z_p = z_i\}$is the index set of positive samples, and $|\mathcal{P}_i|$ is the cardinality of $\mathcal{P}_i$.

With the contrastive learning, the total training loss of the proposed model can be described as follows:

$$\mathcal{L}_{overall} = (1 - \lambda)\, \mathcal{L}_{ce} + \lambda \mathcal{L}_{cl}, \tag{1}$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss for the prediction of readability levels, and λ is a hyperparameter providing tradeoff between these two losses.

Experimental Settings

In this section, we describe experimental settings in detail.

Corpora

We conducted a series of experiments on four English corpora and one Chinese corpus to demonstrate the effectiveness of the proposed model.[1] Some statistics of these corpora are presented in Table 2.

[1]The source codes, datasets, and supplementary materials are available on https://github.com/JinshanZeng/InterpretARA

Properties	Weebit	Cambridge	Newsela	CLEAR	CMT
No. classes	5	5	11	10	12
No. texts	3125	300	9565	4724	2621
Avg. length	288	510	747.37	171.96	674.72

Table 2: Statistics for the used corpora, where CMT is a Chinese corpus while the others are English corpora.

Weebit (Vajjala and Meurers 2012). This dataset is commonly regarded as a benchmark for English readability assessment. It is an extension of the Weekly Reader corpus and is categorized into five levels of difficulty. For each difficulty category, downsampling was performed on 625 texts.

Cambridge (Xia, Kochmar, and Briscoe 2016). The levels of the Cambridge English tests (KET, PET, FCE, CAE, CPE) are used to categorize articles. For each difficulty category, we performed downsampling on 60 texts.

Newsela (Xu, Callison-Burch, and Napoles 2015). The Newsela corpus consists of 10,786 texts, spanning across English and Spanish levels from grades 2 to 12. This dataset is a parallel corpus of original and simplified document alignments. We excluded the Spanish documents and focused solely on the readability assessment of English documents. Hence, the total number of Newsela samples used in our experiments is 9,565.

CLEAR (Crossley et al. 2022). The corpus comprises 4,724 text excerpts designed for readers in grades 3-12. It spans over 250 years of writing across two different genres and provides a unique criterion for readability based on teachers' evaluations of text difficulty for student readers. To adapt to the current task, we utilized Lexile levels, resulting in the creation of ten classes.

CMT (Zeng et al. 2022). The corpus is an extended one built upon (Cheng, Xu, and Dong 2020). It comprises a total of 2,621 texts, which are sourced from Chinese textbooks used in mainland China, covering a range from the first grade of primary school to the third grade of high school.

Baselines

We consider the following state-of-the-art models as baselines to verify the effectiveness of the proposed model.

BERT (Devlin et al. 2019) uses the default BERT model for fine-tuning and the default learning rate (2e-5).

HAN (Yang et al. 2016) employs two Bi-LSTM models, one for word-level and another for sentence-level attention mechanism, enabling differential focus on content importance during text representation construction. In our experiments, we adopted the same configuration as in the original work, with the word and sentence embedding dimensions set to 200 and 100, respectively.

BigBird (Zaheer et al. 2020) is a model based on the Transformer architecture with a sparse attention mechanism, capable of handling up to 4096 tokens and achieving enhanced performance. We utilized the default BigBird model with a learning rate of 1e-5 for our experiments.

Lee2021 (Lee, Jang, and Lee 2021) is a combined application of hand-crafted linguistic features and deep neural network models. It uses both deep and linguistic features to

Dataset	Weebit	Cambridge	Newsela	CLEAR	CMT
segments	3	3	3	3	4
λ	0.9	0.8	0.85	0.95	0.85
Epoch	30	30	30	30	10
Learn rate	1e-5	1e-5	1e-5	1e-5	1e-5

Table 3: Part of the hyperparameter settings.

train a statistical classifier and introduces three novel features in terms of high-level semantics.

DTRA (Zeng et al. 2022) is characterized by a HAN-like structure, with pre-trained BERT used in both the word encoder and sentence encoder. It learns the sequential information between texts by predicting the relative difficulty of paired texts and employing distance-dependent soft labels.

BERT-FP-LBL (Li, Ziyang, and Wu 2022) is a hybrid readability assessment model that integrates difficulty knowledge to extract topic features with anchored correlation explanation, and fused linguistic feature with deep representations by projection filtering.

Implementation Details

For English corpora, we extracted linguistic features at various levels, including discourse, syntactic, lexical, and surface levels, using the toolkit suggested by Lee, Jang, and Lee (2021). For the Chinese corpora, we extracted 67 linguistic features at different levels, including lexical, semantic, syntactic, and cohesion aspects.

We utilized the AdamW optimizer (Loshchilov and Hutter 2017) with a weight decay parameter of 0.01 and a warm-up ratio of 0.1 to train the proposed model. The other hyperparameters are presented in Table 3. All experiments were conducted on an RTX 3090 GPU and implemented using the PyTorch framework.

We evaluated the proposed model using four commonly used classification metrics: *accuracy* (Acc), *precision* (Pre), *macro F1 metric* (F1), and *Quadratic Weighted Kappa* (QWK). For each corpus, we divide it into training, validation, and testing sets in an 8:1:1 ratio. We recorded results by running three trials on average. As for the latest works of Lee, Jang, and Lee (2021) and Li, Ziyang, and Wu (2022), we directly utilize their reported results for comparison, due to the unavailability of reproducible source codes.

Experimental Results

We firstly compared the proposed **InterpretARA** with existing state-of-the-art methods over the concerned corpus, verifying the superiority of our model, then conducted ablation experiments to assess the impact of different modules within our model on performance, and finally conducted a comparison on ways of using linguistic features to demonstrate the effectiveness of the proposed linguistic interpreter for text readability.

Comparison with State-of-the-art Models

The comparison results are presented in Tables 4 and 5. It can be observed that the proposed **InterpretARA** model outperforms baselines on most datasets, achieving the best performance on three English corpora and the concerned Chinese corpus, and the second best performance on only one English corpus Newsela (slightly worse than BigBird).

As shown in Table 4, our approach outperformed the baseline BERT model across all metrics for the Weebit, Cambridge, Newsela, and CLEAR English datasets. Specifically, we achieved improvements of 1.69, 16.67, 8.78, and 2.12 in accuracy, highlighting the effectiveness of our segment-level, enriched linguistic representations, as well as contrastive learning in readability assessment.

As evidenced in Table 4, our model demonstrated remarkable superiority over the HAN and DTRA models. Notably, across the Weebit, Cambridge, Newsela, and CLEAR English datasets, the accuracy of our model exceeded that of DTRA by 7.83, 12.22, 4.08, and 9.73, respectively. The multi-level deep representations in HAN and DTRA are represented and fused hierarchically, which excel at preserving content in longer texts and perform well with such texts. In contrast, the deep and linguistic representations in our model are generated in parallel. We believe that segment-level representation, compared to hierarchical representations, not only preserves content in longer texts but also retains complete text structural information. Besides, our model incorporates linguistic feature representations, introducing extra textual information, which contributes to the performance enhancement of our model.

When compared to the model *Lee2021*, our model exhibited superior performance on the Weebit and Cambridge datasets, achieving accuracy improvements of 2.62 and 13.7, respectively. Although both our model and *Lee2021* employ a combination of deep features and linguistic features, the distinction lies in the methodology employed. *Lee2021* transforms deep features into shallow features before integrating them with linguistic features for traditional machine learning-based readability assessment. In contrast, our approach extracts the enriched linguistic representations in a deep way through reinterpreting linguistic features in a natural language form with the introduced linguistic interpreter, and then integrates them with deep features for text readability assessment. The way used in our model should be better to comprehend the structural information within the text, thereby contributing to the assessment of text readability.

When compared to the BERT-FP-LBL model, our proposed model achieved better results on the Weebit and Cambridge datasets, with improvements of 0.42 and 2.22 in the accuracy, respectively. BERT-FP-LBL is also a hybrid model of deep features and linguistic features. It employs a concatenation method to merge linguistic features and deep features, and introduces topic linguistic features to enhance model performance. These experiment results further demonstrates the effectiveness of our approach in text readability assessment.

When compared to the BigBird model, the proposed model exhibits better performance on the Weebit, Cambridge, and CLEAR datasets, achieving accuracy improvements of 0.32, 1.11, and 3.81 over state-of-the-art models, respectively, while provides competitive performance on the Newsela. Noticing that the BigBird model is specifically de-

Model		HAN	BERT	BigBird	Lee2021*	DTRA	BERT-FP_LBL*	InterpretARA
Weebit	Acc	82.54	91.43	*92.80*	90.50	85.29	92.70	**93.12**
	Pre	83.73	91.46	92.84	90.50	85.54	*92.89*	**93.46**
	F1	82.76	91.42	*92.82*	90.50	85.30	92.73	**93.17**
	QWK	94.48	96.84	97.41	96.80	95.65	*97.78*	**97.81**
Cambridge	Acc	76.67	73.33	*88.89*	76.30	77.78	87.78	**90.00**
	Pre	80.49	76.02	*89.48*	79.20	79.21	89.46	**91.29**
	F1	75.77	71.93	*88.46*	75.20	77.07	87.73	**89.88**
	QWK	92.64	92.29	**97.23**	91.90	92.62	96.87	*96.88*
Newsela	Acc	83.80	78.37	**87.34**	-	83.07	-	*87.15*
	Pre	83.86	78.65	**87.54**	-	82.96	-	*87.01*
	F1	83.70	78.05	**87.32**	-	82.82	-	*87.04*
	QWK	98.35	98.06	*98.75*	-	98.41	-	**98.81**
CLEAR	Acc	65.75	*81.18*	79.49	-	73.57	-	**83.30**
	Pre	64.69	*80.91*	78.85	-	70.35	-	**83.04**
	F1	64.41	*80.75*	79.08	-	71.57	-	**82.81**
	QWK	88.15	*94.77*	93.99	-	91.37	-	**95.54**

Table 4: Comparison results of InterpretARA and baselines over four English benchmark corpora. $*$ Experimental results are taken directly from the literature. The best and second best results are marked in bold and underlined, respectively.

Model		HAN	BERT	BigBird	DTRA	Our
CMT	Acc	42.53	39.74	37.18	*44.42*	**44.87**
	Pre	40.57	38.63	36.47	*44.24*	**45.65**
	F1	41.09	38.29	35.94	**43.87**	*43.22*
	QWK	88.00	88.52	83.33	*89.95*	**91.89**

Table 5: Comparison performance on Chinese corpus CMT.

signed for processing longer texts, it should be more effective on Newsela.

Moreover, as shown in Table 5, the proposed model outperforms these concerned baselines in terms of most evaluation metrics. Specifically, our approach outperforms the baseline BERT model on all metrics for the Chinese dataset CMT, achieving an improvement of 5.13 in accuracy. When compared to other models, except for a slightly lower F1 metric in our model compared to the DTRA model on the CMT dataset, the proposed model achieves state-of-the-art (sota) results. It is worth noting that the Chinese BigBird model undergoes insufficient training due to the scarcity of data, resulting in compromised performance on Chinese datasets. This indicates the continued effectiveness of our method in assessing the readability of Chinese text.

Ablation Studies

To illustrate contributions of modules in our model, we conducted ablation experiments on WeeBit, Cambridge, and CMT datasets and reported the results in Table 6.

When the linguistic feature (**LF**) module is removed, both F1 score and accuracy decrease for all corpora. Specifically, the decreases in F1 score are 0.87, 8.44 and 2.58 for all these three corpora, respectively, while the decreases in accuracy are 0.32, 7.78 and 1.83, respectively. The impact of linguistic features is more significant for the Cambridge corpus with the smallest size of data than the other two corpora. This

Model		Weebit	Cambridge	CMT
F1	InterpretARA	**93.17**	**89.88**	**43.22**
	w/o LF	92.30	81.44	40.64
	w/o (LF, CL)	91.96	79.24	38.35
	w/o (LF, CL, SL)	91.42	71.93	38.29
Acc	InterpretARA	**93.12**	**90.00**	**44.87**
	w/o LF	92.80	82.22	43.04
	w/o (LF, CL)	91.96	80.00	40.11
	w/o (LF, CL, SL)	91.43	73.33	39.74

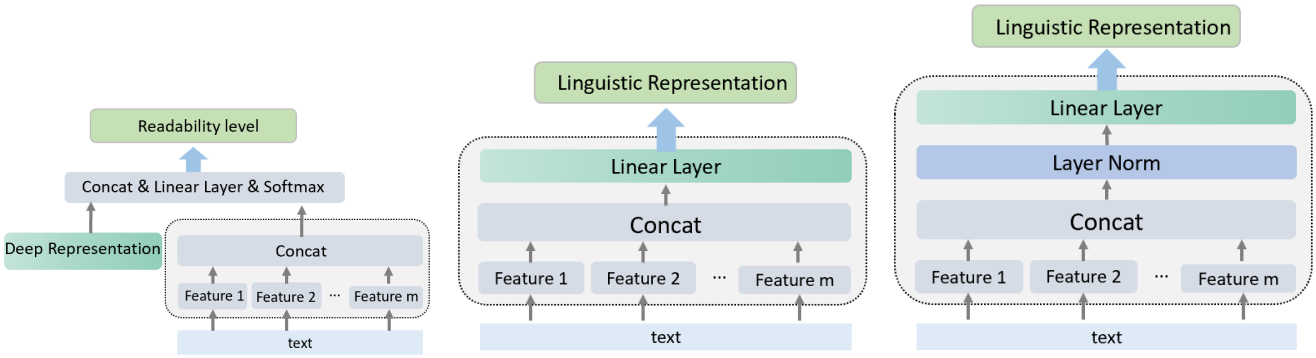
Table 6: Experimental results of ablation studies. LF: Linguistic features. CL: Contrastive learning. SL: Segment-level Representation.

indicates that linguistic features contribute to text readability assessment, especially for corpora with small training data.

When the contrastive learning (**CL**) module is further removed, the performance of models keeps degrading in terms of both F1 score and accuracy. This shows that **CL** can help the model to capture more useful structural information.

When further removing the segment-level representation (**SL**) module, it can be observed from Table 6 that the performance of models gets worse for all corpora in terms of both F1 score and accuracy. The decreases of F1 score are 0.54, 7.31, 0.06 for Weebit, Cambridge and CMT, respectively, while the decrease of accuracy are 0.53, 6.67 and 0.37, respectively. These show that the introduced segment-level representations can capture local structural information of texts and are useful for readability assessment.

When comparing impacts among these three concerned modules, the linguistic feature module and segment-level representation module gain the largest and second largest improvements on the performance, respectively. These show that both linguistic and segment-level representations provide important information for text readability assessment.



(a) way 1 (b) way 2 (c) way 3

Figure 2: Illustration of three linguistic feature processing ways commonly used in the literature.

Comparison on Ways of Using Linguistic Features

Metric	F1			Acc		
Dataset	Weebit	Cambridge	CMT	Weebit	Cambridge	CMT
Baseline	92.30	81.44	40.64	92.80	82.22	43.04
way 1	92.25	85.29	36.64	92.17	85.56	37.91
way 2	91.83	78.47	36.00	91.75	74.44	37.36
way 3	92.77	83.81	41.28	92.70	84.44	43.59
our	**93.17**	**89.88**	**43.22**	**93.12**	**90.00**	**44.87**

Table 7: Comparisons on the performance of different ways to utilize linguistic features. Baseline: Eliminating linguistic features based on the proposed InterpretARA model.

We compared the performance of the introduced linguistic interpreter with three existing ways of using linguistic features to demonstrate its effectiveness. The concerned three ways to utilize linguistic features are shown in Figure 2 and described in detail as follows.

- **way 1**: Building upon the baseline model, we adopted the most straightforward approach of concatenating the raw linguistic feature vectors with deep representations, and then performed text readability assessment.
- **way 2**: In contrast to direct concatenation as in **way 1**, we performed a linear transformation preprocessing on the linguistic features, aligning their dimensions with the document-level representations. This was followed by replacing the enriched linguistic representation module in our model for text readability assessment.
- **way 3**: It is similar to **way 2** with an additional layer normalization to remove variation among linguistic features.

Comparison results are presented in Table 7. It can observed from Table 7 that the ways of using linguistic features have significant impacts on performance. Specifically, when **way 1** and **way 2** are adopted, the performance of the associated hybrid models using both linguistic and deep features get worse than the base model only using deep features. This shows that the direct concatenation way and the simple linear mixing way of linguistic features should be ineffective for readability assessment, possibly due to the different representation spaces of deep and linguistic features. This also motivates us to design more dedicate ways to combine deep and linguistic features. When **way 3** with a layer normalization is adopted, the performance of the corresponding hybrid model gains slight improvements in most cases. This shows that the variation among different features should be harmful to the model.

Distinguished from these three ways that only linearly mix linguistic features in shallow representation spaces, we firstly suggest a linguistic feature interpreter to reinterpret these linguistic features into natural language descriptions, then yield their representations in a deep representation space, consistent with the representation space of deep features. By doing these, the performance of the proposed InterpretARA model gains substantial improvements for all corpora in terms of both F1 and accuracy. In particular, the proposed model achieves remarkable improvements on the Cambridge corpus. These clearly show the effectiveness of the introduced ways of using linguistic features.

Conclusion

How to effectively exploit linguistic features is crucial for the text readability assessment. This paper proposes an effective hybrid ARA model through introducing a novel linguistic feature interpreter to enrich the linguistic representations, and adopting the contrastive learning as well as the segment-level representation module to improve deep representations. The suggested linguistic feature interpreter provides a natural language description for a linguistic feature. Such natural language description way can provide more semantic information and bring the way to extract linguistic representations more thoroughly through a deep model. The effectiveness of the proposed model is demonstrated by a series of experiments. It is worth pointing out that this paper provides a novel way to use linguistic features and the suggested linguistic feature interpreter can be easily adapted to other text readability assessment models as well as other related applications. One future direction of this paper is to study more effective fusion scheme of deep and linguistic representations.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China [Grant Nos. 62376110, 61977038, 62262031], Thousand Talents Plan of Jiangxi Province [Grant No. jxsq2019201124], Jiangxi Provincial Natural Science Foundation for Distinguished Young Scholars (20224ACB212004), Jiangxi Provincial Educational Science Foundation during the "13th Five-Year Plan" (20YB080), Humanities and Social Science Foundation of Higher Education Institutions of Jiangxi Province (YY22209), and Graduate Innovation Fund of Jiangxi Provincial Department of Education (YC2023-S319).

References

Arfé, B.; Mason, L.; and Fajardo, I. 2018. Simplifying informational text structure for struggling readers. *Reading and Writing*, 31(9): 2191–2210.

Azpiazu, I. M.; and Pera, M. S. 2019. Multiattentive Recurrent Neural Network Architecture for Multilingual Readability Assessment. *Transactions of the Association for Computational Linguistics*, 7: 421–436.

Chall, D. J. S. 1948. A Formula for Predicting Readability: Instructions. *Educational Research Bulletin*, 27(2): 37–54.

Cheng, Y.; Xu, D.; and Dong, J. 2020. On Key Factors of Text Reading Difficulty Grading and Readability Formula Based on Chinese Textbook Corpus. *Applied Linguistics*, 1: 132–143.

Crossley, S.; Heintz, A.; Choi, J. S.; Batchelor, J.; Karimi, M.; and Malatinszky, A. 2022. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 1–17.

Dell'Orletta, F.; Montemagni, S.; and Venturi, G. 2011. READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification. *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, 73–83.

Denning, J.; Pera, M. S.; and Ng, Y.-K. 2016. A readability level prediction tool for K-12 books. *association for information science and technology*, 67(3): 550–565.

Deutsch, T.; Jasbi, M.; and Shieber, S. 2020. Linguistic Features for Readability Assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Flesch, R. F. 1948. A new readability yardstick. *The Journal of applied psychology*, 32 3: 221–33.

Hancke, J.; Vajjala, S.; and Meurers, D. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012: Technical Papers*, 1063–1080.

Hansen, H.; Widera, A.; Ponge, J.; and Hellingrath, B. 2021. Machine Learning for Readability Assessment and Text Simplification in Crisis Communication: A Systematic Review. In *Hawaii International Conference on System Sciences*.

Jiang, Z.; Gu, Q.; Yin, Y.; Wang, J.; and Chen, D. 2019. GRAW+: A two-view graph propagation method with word coupling for readability assessment. *Journal of the Association for Information Science and Technology*, 70(5): 433–447.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 18661–18673. Curran Associates, Inc.

Klare, G. R. 1963. *Measurement of readability*. Iowa, USA: Iowa State University Press.

Klare, G. R. 2000. The measurement of readability: useful information for communicators. *ACM Journal of Computer Documentation*, 24(3): 107–121.

Laughlin, G. H. M. 1969. SMOG Grading-a New Readability Formula. *Journal of Reading*, 12(8): 639–646.

Lee, B. W.; Jang, Y.; and Lee, J. H.-J. 2021. Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. In *Conference on Empirical Methods in Natural Language Processing*.

Li, W.; Ziyang, W.; and Wu, Y. 2022. A Unified Neural Network Model for Readability Assessment with Feature Projection and Length-Balanced Loss. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7446–7457. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Lively, B. A.; and Pressey, S. L. 1923. A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision*, 9(7): 389–398.

Loshchilov, I.; and Hutter, F. 2017. Fixing weight decay regularization in adam.

Lu, D.; Qiu, X.; and Cai, Y. 2019. Sentence-Level Readability Assessment for L2 Chinese Learning. In *Chinese Lexical Semantics*.

Ma, Y.; Fosler-Lussier, E.; and Lofthus, R. 2012. Ranking-Based Readability Assessment for Early Primary Children's Literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, 548–552. USA: Association for Computational Linguistics. ISBN 9781937284206.

McLaughlin, G. H. 1969. SMOG Grading - A New Readability Formula. *The Journal of Reading*.

Pera, M. S.; and Ng, Y.-K. 2014. Automating Readers' Advisory to Make Book Recommendations for K-12 Readers. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, 9–16. New York, NY, USA: Association for Computing Machinery. ISBN 9781450326681.

Perni, S.; Rooney, M. K.; Horowitz, D. P.; Golden, D. W.; McCall, A. R.; Einstein, A. J.; and Jagsi, R. 2019. Assessment of Use, Specificity, and Readability of Written Clinical Informed Consent Forms for Patients With Cancer Undergoing Radiotherapy. *JAMA Oncology*, 5(8): e190260–e190260.

Qiu, X.; Chen, Y.; Chen, H.; Nie, J.; Shen, Y.; and Lu, D. 2021. Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment. In *Annual Meeting of the Association for Computational Linguistics*.

Sung, Y.-T.; Lin, W.-C.; Dyson, S. B.; Chang, K.-E.; and Chen, Y.-C. 2015. Leveling L2 Texts Through Readability: Combining Multilevel Linguistic Features with the CEFR. *The Modern Language Journal*, 99(2): 371–391.

Vajjala, S.; and Meurers, D. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, 163–173.

Xia, M.; Kochmar, E.; and Briscoe, T. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 12–22. San Diego, CA: Association for Computational Linguistics.

Xu, W.; Callison-Burch, C.; and Napoles, C. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3: 283–297.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. San Diego, California: Association for Computational Linguistics.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33: 17283–17297.

Zeng, J.; Xie, Y.; Yu, X.; Lee, J.; and Zhou, D.-X. 2022. Enhancing Automatic Readability Assessment with Pretraining and Soft Labels for Ordinal Regression. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4557–4568. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.