Self-Supervised Collaborative Information Bottleneck for Text Readability Assessment

Jinshan Zeng[1]*, Xianglong Yu[1], Xianchao Tong[1], Wenyan Xiao[2]
[1]Jiangxi Normal University
[2]Jiangxi University of Science and Technology
jinshanzeng@jxnu.edu.cn, xianglongyu@jxnu.edu.cn, xianchaotong@jxnu.edu.cn,wy.xiao@jxust.edu.cn

Abstract

Text readability assessment involves categorizing texts based on readers' comprehension levels. Hybrid automatic readability assessment (ARA) models, combining deep and linguistic features, have recently attracted rising attention due to their impressive performance. However, existing hybrid ARA models generally ignore the specific-intrinsic information of deep and linguistic representations, and cannot fully explore their common-intrinsic information. In this paper, we introduce a self-supervised collaborative information bottleneck (SCIB) module for ARA to address these issues. Specifically, we collaboratively consider both specific-intrinsic and common-intrinsic information of the linguistic representation and various levels of deep representations including the document-, sentence- and word-level deep representations, and yield their refined representations via a self-supervised information bottleneck scheme. Extensive experiments are conducted on four English and two Chinese corpora to demonstrate the effectiveness of the proposed model. Experimental results show that the proposed model outperforms state-of-the-art models in terms of four important evaluation metrics, and the suggested SCIB module can effectively capture the specific- and common-intrinsic information.

Introduction

The assessment of text readability is designed to determine the complexity of a given text, helping readers select materials that match their comprehension level (McLaughlin 1969; Klare 2000). This not only caters to individual language skills, cognitive abilities, and developmental stages but also extends its utility to diverse sectors such as text recommendations, the crafting of clinical informed consents, and the publishing of books.

With the continuous advancement of readability research, a series of readability assessment methods have been developed, including readability formulas (Dale and Chall 1948; Kincaid et al. 1975) , statistical machine learning methods (Dell'Orletta, Montemagni, and Venturi 2011; Sung et al. 2015), and neural language modeling methods (Deutsch, Jasbi, and Shieber 2020; Tseng et al. 2019; Zeng et al. 2022). The kind of readability formula methods mainly yields the readability level according to empirical formulas with respect to some linguistic features, while the kind of statistical machine learning methods mainly focuses on yielding the readability level by training some statistical models such as support vector machines (SVM) with handcrafted linguistic features (Sung et al. 2015). Due to the limited representation capability of these linguistic features (also called surface features), both kinds of readability formula and statistical machine learning methods cannot yield satisfactory performance, especially for some complex texts.

Inspired by the powerful representation ability of neural networks, neural language models have been widely favored by researchers for readability assessment (Azpiazu and Pera 2019; Deutsch, Jasbi, and Shieber 2020; Tseng et al. 2019; Zeng et al. 2022). Azpiazu and Pera (2019) proposed a multilingual readability assessment model (Vec2Read) based on the hierarchical attention network (HAN) (Yang et al. 2016). This method combines word part of speech (POS) and morphological tags, and utilizes hierarchical information to generate word- and sentence-level attention scores for creating text representations. Later, some novel deep models based on the pre-trained model have been suggested in the literature (Deutsch, Jasbi, and Shieber 2020; Tseng et al. 2019; Zeng et al. 2022), with the emergence of pre-trained models such as Transformer (Vaswani et al. 2017) and BERT (Devlin et al. 2019). Tseng et al. (2019) adopted the pre-trained BERT for the text readability assessment. Zeng et al. (2022) integrated the pre-trained BERT into the HAN architecture and proposed a novel deep model for readability assessment by introdcing soft labels for ordinal regression.

Despite the impressive performance of the kind of neural models, their performance is generally limited by the insufficient training corpora. To alleviate this issue, recent research turns to the direction of combining linguistic features and deep features for readability assessment, i.e., hybrid automatic readability assessment (ARA) models called in the literature (Deutsch, Jasbi, and Shieber 2020; Lee, Jang, and Lee 2021; Li, Wang, and Wu 2022; Zeng et al. 2023, 2024). There are two key ingredients for the kind of hybrid ARA models, i.e., feature extraction and fusion. The literature (Deutsch, Jasbi, and Shieber 2020; Lee, Jang, and Lee 2021) exploited the pre-trained language models such as BERT (Devlin et al. 2019) to extract the deep feature, and fed deep and linguistic features simultaneously into a statistical ma-

*Corresponding author

chine learning model to yield readability levels without fusion. Li, Wang, and Wu (2022) extracted various levels of deep features including the word-, sentence- and document-level deep features, and fused them with the linguistic feature by an orthogonal projection scheme. Motivated by (Li, Wang, and Wu 2022), Zeng et al. (2023) incorporated the prompt learning (Lee and Lee 2023) into the deep feature extraction to fully explore deep representations, and utilized a hierarchical orthogonal projection fusion scheme to fuse these various levels of deep and linguistic features. Later, Zeng et al. (2024) introduced a linguistic interpreter and used the contrastive learning (Khosla et al. 2020) to fully explore linguistic and deep representations respectively, and fused them via a similar hierarchical orthogonal projection fusion scheme.

Notice that existing orthogonal projection fusion schemes do not consider the specific-intrinsic information of various levels of features important for the readability assessment, and cannot yield their refined representations. Moreover, the performance of the orthogonal projection fusion scheme is sensitive to the choice of the orthogonal base feature. To address these issues, we propose a sophisticated fusion scheme for ARA, motivated by the well-known information bottleneck principle (Tishby and Zaslavsky 2015). The major contributions of this paper can be summarized as follows.

- We propose a novel hybrid ARA model through introducing a self-supervised collaborative information bottleneck (SCIB) module, which collaboratively considers both specific-intrinsic and common-intrinsic information of various levels of representations including the linguistic-level representation and the word-, sentence-, and document-level deep representations, and yields their refined representations via a self-supervised information bottleneck scheme. By leveraging the SCIB module, the specific-intrinsic and common-intrinsic information of various levels of deep and linguistic features are fully explored and fused for the readability assessment.
- Extensive experiments are conducted over four English and two Chinese corpora to demonstrate the effectiveness of the proposed model. Experimental results show that the proposed model outperforms competing neural models on most datasets, and the suggested SCIB module can effectively capture the specific- and common-intrinsic information of these deep and linguistic features.

Related Work

Linguistic Representation

Linguistic features defined in linguistics generally reflect some surface structures of texts and can provide important information for the readability assessment. Existing linguistic features can be mainly divided into four categories, i.e., the lexical, semantic, syntactic and cohesion level features (Lee, Jang, and Lee 2021). Zeng et al. (2023) used these linguistic feature to improve the performance of readability assessment, where the linguistic representation was yielded from these linguistic features by implementing the layer normalization and linear projection operations. To enrich the linguistic representation, Zeng et al. (2024) introduced a linguistic interpreter to transfer linguistic features into some natural language sentences according to some templates, which are facilitate to representing them in a deep way.

Deep Representation

In the early stage, deep representations were directly extracted by some pre-trained language models such as BERT (Tseng et al. 2019). Later, multi-level deep representations were suggested in the literature (Zeng et al. 2022; Li, Wang, and Wu 2022; Zeng et al. 2023, 2024). Motivated by HAN (Yang et al. 2016), Zeng et al. (2022) suggested using the hierarchical multi-level deep representations for readability assessment. Li, Wang, and Wu (2022) suggested using the parallel multi-level deep representations including the document-, sentence/paragraph-, and word-level for readability assessment, since such kind of multi-level deep representations generally contains richer information than the kind of hierarchical multi-level deep representations. To further improve deep representations, the prompt learning (Lee and Lee 2023) and contrastive learning (Khosla et al. 2020) were adopted in the literature (Zeng et al. 2023) and (Zeng et al. 2024), respectively.

Fusion Scheme

It is crucial to fuse these various levels of deep representations and the linguistic representation. To reduce the redundancy among these representations, the orthogonal projection fusion scheme was widely used in the literature (Li, Wang, and Wu 2022; Zeng et al. 2023, 2024). However, existing orthogonal projection fusion scheme does not consider the specific-intrinsic information of these various-level representations, and is sensitive to the choice of the orthogonal base representation. To address these issues, we suggest a sophisticated fusion scheme called self-supervised collaborative information bottleneck scheme inspired by the information bottleneck principle (Tishby and Zaslavsky 2015). The suggested scheme collaboratively considers both specific- and common-intrinsic information, and yields refined representations of these deep and linguistic representations via a self-supervised information bottleneck scheme.

Proposed Model

In this section, we describe the proposed model dubbed *SCIB-ARA* in detail. As shown in Figure 1, the proposed model mainly consists of the feature representation module and the suggested SCIB module, where the feature representation module aims to yield various levels of deep and linguistic representations, and the SCIB module is introduced to effectively fuse them. In the feature representation module, besides the linguistic-level representation, three levels of deep representations including the word-, sentence- and document-level deep representations are yielded. In the SCIB module, both the specific-intrinsic and common-intrinsic information of these four various levels of representations are collaboratively considered and the associated refined representations are yielded by the self-supervised information bottleneck principle, while these refined representations are finally concatenated for assessment.

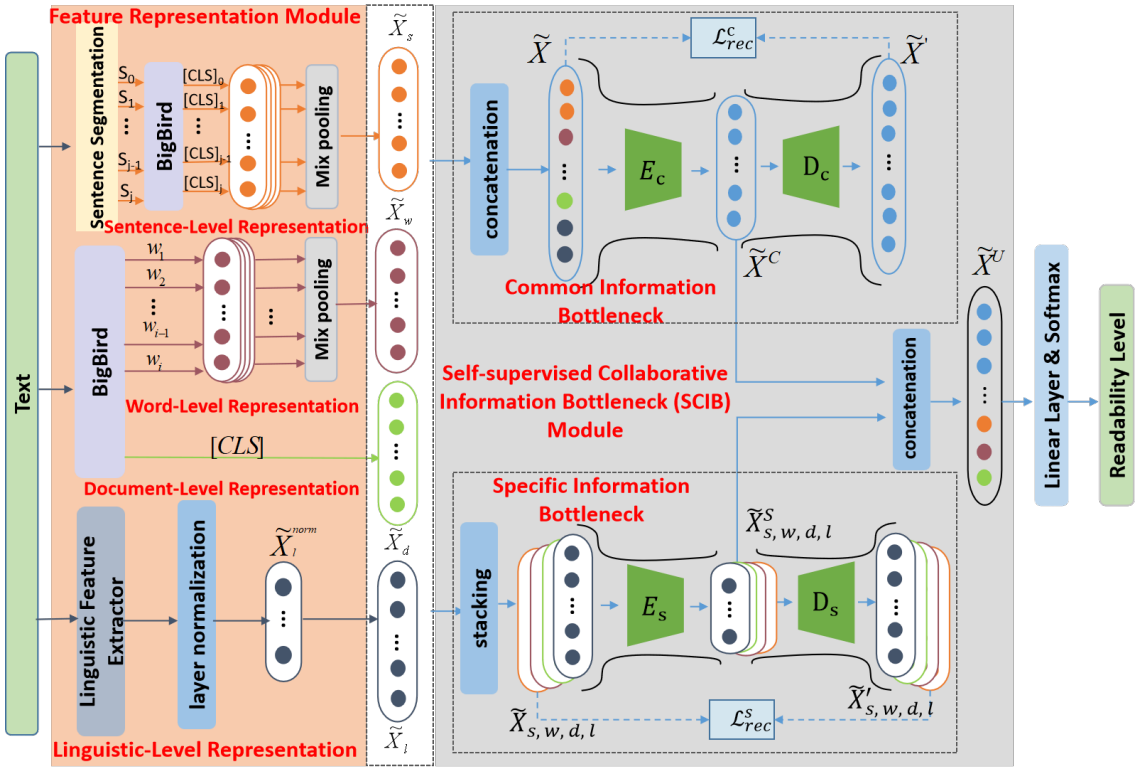


Figure 1: The pipeline of the proposed SCIB-ARA model for text readability assessment.

Feature Representation Module

The source of feature representations mainly comprises two parts: deep representations and linguistic representations. Three levels (i.e., document-, sentence- and word-level) of deep representations are derived from features within deep models, while linguistic representations are the fundamental properties manually extracted from texts.

Specifically, for a given text, its document-level representation $\widetilde{X}_d$ is directly taken as the embedding from BigBird (Zaheer et al. 2020) with the whole text as the input. For the word-level deep representation $\widetilde{X}_w$, we feed the text into the BigBird and produce embeddings for each word, and then yield the word-level deep representation with a mixing pooling layer, which is a hybrid of the max-pooling and mean-pooling operations designed to extract both the maximum value and the average value (Yu et al. 2014). The sentence-level deep representation $\widetilde{X}_s$ is yielded in a similar way through feeding the text to BigBird sentence by sentence.

Since linguistic features characterize some important textual natures, and provide many additional insights and contextual information for ARA, linguistic features should be paid particular attention in ARA. For linguistic representations, we firstly extract various linguistic features, then perform the layer normalization (Ba, Kiros, and Hinton 2016) to eliminate the influence of changes between different features and obtain a normalized linguistic representations $\widetilde{X}_l^{norm}$, finally yield the linguistic representation $\widetilde{X}_l$ with the same dimension of deep representations by a linear layer.

Self-supervised Collaborative Information Bottleneck Module

With these levels of deep and linguistic representations, we collaboratively consider both the common-intrinsic information and specific-intrinsic information of these representations in the SCIB module to yield refined representations for readability assessment, where the specific information bottleneck sub-module and the common information bottleneck sub-module are designed to capture the specific-intrinsic information and common-intrinsic information of these four levels of representations, respectively. The common-intrinsic information contains the consistent information of these four levels of representations, while the specific-intrinsic information can provide important individual information important for readability assessment.

Specifically, in the common information bottleneck sub-module, we firstly concatenate these four levels of text representations, i.e., $\widetilde{X}_s$, $\widetilde{X}_w$, $\widetilde{X}_d$, and $\widetilde{X}_l$, and then feed the concatenated representation $\widetilde{X}$ into an encoder E_c to generate a common representation $\widetilde{X}^C$. To guide the encoder E_c to learn the consistent information of these four various levels of representations, we introduce a decoder D_c, motivated by the information bottleneck principle. We feed $\widetilde{X}^C$ to the decoder D_c and yield an estimate $\widetilde{X}'$ of the consistent rep-

resentation of these four various levels of representations as the supervision of training. To achieve this, we impose the following self-supervised loss on the proposed model:

$$\mathcal{L}_{rec}^{c} = \sum_{i \in \{s,w,d,l\}} \|\widetilde{X}_i - \widetilde{X}'\|^2$$

Notice that the common text representation $\widetilde{X}^C$ may overlook unique information specific to individual-level representations, which can provide important supplementary information for the common representation. To achieve this, we collaboratively consider the specific-intrinsic and common-intrinsic information, and introduce the specific information bottleneck sub-module to capture the unique information specific to these four levels of representations.

Specifically, we feed the sentence-level representation $\widetilde{X}_s$ into the associated encoder E_s to yield its specific representation $\widetilde{X}_s^S$, then feed $\widetilde{X}_s^S$ into the associated decoder D_s to yield an estimate $\widetilde{X}_s'$ of the sentence-level representation $\widetilde{X}_s$, which is used to form the reconstruction loss to supervise the learning procedure. Following the similar procedures, we can yield the specific representations and reconstruction loss functions for the other two levels of deep representations. For the linguistic representation, we use the normalization $\widetilde{X}_l^{norm}$ of the linguistic representation $\widetilde{X}^l$ instead of itself to form the self-supervised scheme for the learning of linguistic specific representation, where an additional linear layer is included to perform the normalization compared to self-supervised procedures for deep representations. Thus, we impose the following loss on the training of specific information bottleneck sub-module:

$$\mathcal{L}_{rec}^{s} = \sum_{i \in \{s,w,d\}} \|\widetilde{X}_i - \widetilde{X}_i'\|^2 + \|\widetilde{X}_l^{norm} - \widetilde{X}_l'\|^2.$$

With the common representation $\widetilde{X}^C$ and specific representations $\widetilde{X}_{s,w,d,l}^S := \{\widetilde{X}_s^S, \widetilde{X}_w^S, \widetilde{X}_d^S, \widetilde{X}_l^S\}$, we concatenate them to yield the unified representation $\widetilde{X}^U$, and then feed $\widetilde{X}^U$ into a neural predictor consisting of a linear layer and softmax to yield the prediction of the readability level.

According to the above description, the training loss of the proposed model can be formulated as follows:

$$\mathcal{L}_{\text{SCIB-ARA}} = \mathcal{L}_{CE} + \lambda \left(\mathcal{L}_{rec}^{c} + \mathcal{L}_{rec}^{s}\right), \tag{1}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss for prediction, λ is a tunable hyper-parameter.

Experiments

In this section, a series of comparative experiments were conducted on four English and two Chinese corpora to compare the effectiveness of the proposed model with existing state-of-the-art models. We also conducted ablation experiments on various components of the proposed model.

Experimental Settings

We describe experimental settings in detail.

A. Corpora. We evaluated the proposed model through experiments over four English corpora (i.e., WeeBit (Vajjala and Meurers 2012), Cambridge[1], Newsela[2] and CLEAR), and two Chinese corpora (i.e., CMT (Lee, Liu, and Cai 2020; Zeng et al. 2022) and CMER). Some statistics of these corpora are presented in Table 1.

Corpus	Weebit	Cambridge	Newsela	CLEAR	CMT	CMER
No. classes	5	5	11	10	12	12
No. texts	3125	300	9565	4724	2621	2260
Ave.length	288	510	747	172	927	675

Table 1: Statistics for the used six corpora including four English and two Chinese corpora.

B. Baselines. We considered eight state-of-the-art models as baselines to verify the effectiveness of the proposed model. These include two typical pre-trained language models (i.e., BERT (Devlin et al. 2019) and BigBird (Zaheer et al. 2020)), two representative deep models based on hierarchical attention networks (i.e., HAN (Yang et al. 2016) and DTRA (Zeng et al. 2022)), and four hybrid ARA models (i.e., Lee-2021 (Lee, Jang, and Lee 2021), BERT-FP-LBL (Li, Wang, and Wu 2022), PromptARA (Zeng et al. 2023) and InterpretARA (Zeng et al. 2024)).

C. Implementation details. For the proposed model, we used AdamW (Loshchilov and Hutter 2017) as the optimizer with a weight decay parameter of 0.01 and a warmup ratio of 0.1. The used linguistic features for the proposed model are presented in Supplementary Materials.

For these two baselines, i.e., Lee-2021 (Lee, Jang, and Lee 2021) and BERT-FP-LBL (Li, Wang, and Wu 2022), we directly took experimental results reported in the literature for comparison, since we cannot access their reproducible codes. For a fair comparison, we followed the similar setups in (Lee, Jang, and Lee 2021) and (Li, Wang, and Wu 2022) for other baselines. For each corpus, we split the data into training, validation, and test sets in a ratio of 8:1:1, and reported the average results of three trails. All experiments were implemented on RTX 3090 and A40 GPUs, and in the PyTorch framework.

Comparison with State-of-the-art Models

We conducted a series of experiments to evaluate the performance of the proposed model through comparing with the state-of-the-art models, in terms of four commonly used metrics for classification, i.e., *accuracy* (Acc), *precision* (Pre), *macro F1-metric* (F1), and *quadratic weighted kappa* (QWK). The comparison results are presented in Table 2.

From Table 2, we can observe that the proposed model achieves the best or suboptimal performance over most of corpora in terms of all four evaluation metrics. Specifically, in comparison with these two pre-trained deep models, i.e., BERT (Devlin et al. 2019) and BigBird (Zaheer et al. 2020), the proposed model is superior to these two models over all corpora, in particular over the Cambridge, CLEAR, CMT and CMER corpora. When comparing the performance between BERT and BigBird, we can observe that BigBird

[1]http://www.cambridgeenglish.org

[2]https://newsela.com

Datasets	Metrics	BERT	BigBird	HAN	DTRA	Lee-2021*	BERT-FP-LBL*	PromptARA	InterpretARA	**Our**
Weebit	Acc	91.53	92.70	82.54	85.29	90.50	92.70	*93.12*	*93.12*	**94.07**
	Pre	91.56	92.73	83.73	85.54	90.50	92.89	93.19	*93.46*	**94.26**
	F1	91.51	92.70	82.76	85.30	90.50	92.73	93.09	*93.17*	**94.10**
	QWK	97.10	97.17	94.48	95.65	96.80	97.78	97.43	*97.81*	**98.03**
Cambridge	Acc	75.56	87.78	76.67	77.78	76.30	87.78	*91.11*	90.00	**95.56**
	Pre	72.75	88.29	80.49	79.71	79.20	89.46	*92.24*	91.29	**95.87**
	F1	72.95	87.53	75.55	77.07	75.20	87.73	*90.88*	89.88	**95.54**
	QWK	91.59	97.04	92.64	92.62	91.90	96.97	*97.82*	96.88	**98.99**
CLEAR	Acc	76.74	78.86	67.87	72.09	-	-	82.03	*83.30*	**83.72**
	Pre	76.23	79.01	66.22	70.75	-	-	81.99	*83.04*	**83.49**
	F1	76.05	78.34	66.43	70.81	-	-	81.75	*82.81*	**83.49**
	QWK	92.86	94.03	89.29	91.85	-	-	**95.54**	**95.54**	*95.45*
Newsela	Acc	77.12	87.15	83.80	83.07	-	-	**88.40**	87.15	*87.57*
	Pre	78.45	87.14	83.86	82.96	-	-	**88.41**	87.01	*87.44*
	F1	76.59	87.05	83.70	82.82	-	-	**88.24**	87.04	*87.47*
	QWK	97.67	98.79	98.55	98.41	-	-	**98.88**	*98.81*	*98.81*
CMT	Acc	38.46	39.19	42.53	44.42	-	-	43.96	*44.87*	**45.79**
	Pre	38.79	40.60	40.57	44.24	-	-	43.17	**45.65**	*45.24*
	F1	37.17	35.97	41.09	**43.87**	-	-	41.60	*43.22*	42.82
	QWK	88.09	88.97	88.00	89.95	-	-	*91.20*	**91.89**	90.98
CMER	Acc	22.30	24.06	23.40	*26.50*	-	-	*26.50*	25.39	**31.79**
	Pre	20.13	25.55	15.47	*25.36*	-	-	24.24	24.60	**30.55**
	F1	13.49	22.58	18.48	*25.16*	-	-	23.92	23.40	**27.51**
	QWK	65.39	70.86	72.10	70.53	-	-	68.74	*73.95*	**75.33**

Table 2: Comparison results of the proposed SCIB-ARA model and baselines over four English and two Chinese benchmark corpora. $*$ Experimental results were directly taken from the literature since we cannot access their reproducible source codes. The best and second best results are marked in bold and underlined, respectively.

has substantial improvement on the performance than BERT over all corpora. This also motivates us to utilize BigBird to extract deep features in the proposed model.

Compared to the kind of hierarchical attention network based models such as HAN (Yang et al. 2016) and DTRA (Zeng et al. 2022), the proposed model significantly outperforms them over most of corpora. Specifically, as compared to HAN, the proposed model achieves improvements of 11.53%, 18.89%, 15.85%, 3.77%, 8.39% and 3.26% in accuracy over six concerned corpora respectively, while yields the associated improvements of 8.78%, 17.78%, 11.63%, 4.50%, 5.29% and 1.37% in accuracy compared to DTRA.

When compared to existing hybrid models like Lee-2021 (Lee, Jang, and Lee 2021) and BERT-FP-LBL (Li, Wang, and Wu 2022), the proposed model also yields substantial improvements over Weebit and Cambridge corpora. In terms of accuracy, the proposed model achieves improvements of 3.57% and 19.26% over Weebit and Cambridge, respectively when compared to Lee-2021, while yields the associated improvements of 1.37% and 7.78% as compared to BERT-FP-LBL. In Lee-2021, deep features are firstly utilized to yield certain soft labels, which together with the linguistic features are simultaneously fed to a statistical classifier for the final prediction without fusion. On the other hand, in BERT-FP-LBL, only the document-level deep representation and linguistic representation are fused by a simple orthogonal projection layer. It can be observed that the fusion way used in this paper through learning consistent and specific representations of both deep and linguistic representations can yield refined representations incorporating both deep and linguistic information. Moreover, we extract multi-level deep representations instead of the sole document-level representation to yield deep representations of higher quality and abundance. The comparison results with these two hybrid models demonstrate the superiority of using multi-level deep representations methods.

When Compared to the recent two hybrid ARA models, i.e., PromptARA (Zeng et al. 2023) and InterpretARA (Zeng et al. 2024), our model also demonstrates the superior performance. Specifically, compared with PromptARA using prompts and orthogonal projection fusion, the proposed model improves by 0.95%, 4.45%, 1.69%, 5.29% and 1.83% in accuracy over Weebit, Cambridge, CLEAR, CMER and CMT, respectively. While compared with InterpretARA using the linguistic interpreter and orthogonal projection fusion scheme, the proposed model improves by

0.95%, 5.56%, 0.42%, 0.42%, 6.40% and 0.92% in accuracy over the concerned six corpora, respectively. These results show the superiority of the proposed model over state-of-the-art models. Moreover, these numerical results demonstrate that the suggested SCIB fusion scheme is superior to the commonly used fusion scheme based on the orthogonal projection, though the recent PromptARA and InterpretARA models use other advanced techniques such as prompts and linguistic interpreter to enrich representations.

Regarding the performance of the proposed model over different corpora, we can observe from Table 2 that the proposed model achieves high accuracy for Weebit and Cambridge with the accuracy over 94%, and yields the moderately high accuracy for CLEAR and Newsela with the accuracy over 83%, and achieves relatively low accuracy for these two Chinese corpora with the accuracy 31.79% and 45.79%, respectively. These results show that it is generally more challenging for the readability assessment of Chinese texts due to the semantic ambiguity of Chinese texts.

Dataset	Model	Acc	Pre	F1	QWK
Weebit	SCIB-ARA	**94.07**	**94.26**	**94.10**	**98.03**
	w/o C	93.55	93.58	93.54	97.75
	w/o S	93.44	93.48	93.43	97.81
	w/o (C, S)	92.06	92.18	92.05	96.87
Cambridge	SCIB-ARA	**95.56**	**95.87**	**95.54**	**98.99**
	w/o C	94.44	94.92	94.30	98.61
	w/o S	94.45	94.86	94.41	98.59
	w/o (C, S)	88.89	89.81	88.62	96.30
CLEAR	SCIB-ARA	**83.72**	**83.49**	**83.49**	**95.45**
	w/o C	81.61	81.28	81.14	94.63
	w/o S	81.61	81.13	80.95	94.63
	w/o (C, S)	81.40	80.75	80.95	94.57
CMER	SCIB-ARA	**31.79**	**30.55**	**27.51**	**75.33**
	w/o C	26.50	29.65	25.44	74.43
	w/o S	26.93	25.38	21.99	73.67
	w/o (C, S)	24.28	19.43	19.85	74.15
CMT	SCIB-ARA	**45.97**	**45.24**	**42.82**	**90.98**
	w/o C	45.79	43.55	41.93	90.79
	w/o S	44.51	41.31	40.32	89.78
	w/o (C, S)	41.58	39.87	39.23	85.25

Table 3: On effectiveness of the introduced SCIB moodule.

Ablation Study

In this subsection, we conducted a series of ablation studies on three English corpora (Weebit, Cambridge and CLEAR) and two Chinese corpora (CMT and CMER) to investigate the feasibility and effectiveness of our proposed ideas.

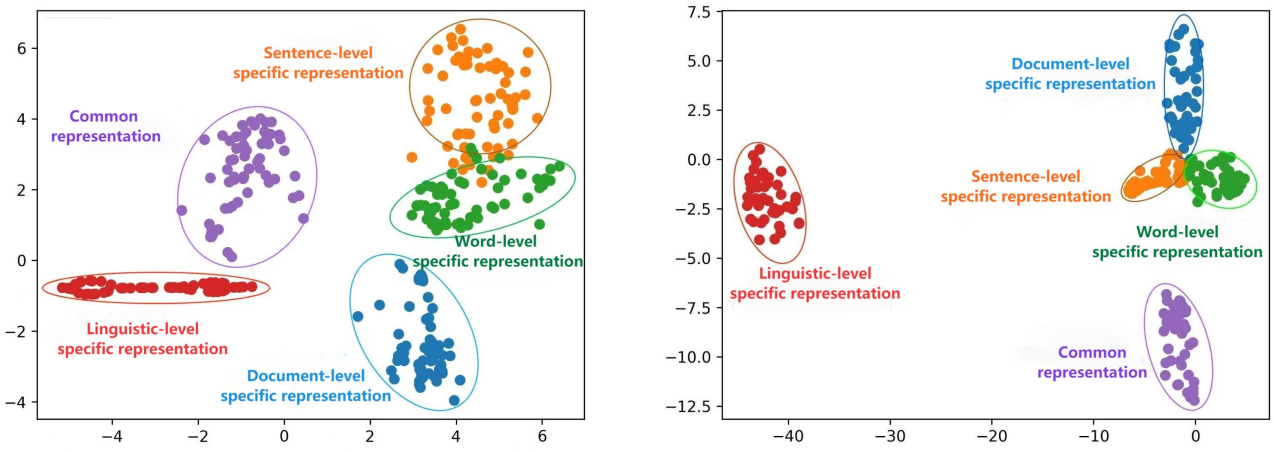
A. On effectiveness of SCIB module. We verified the effectiveness and feasibility of the introduced common and specific information bottleneck sub-modules by comparing the performance of the proposed model with that of models by removing the common representation (**w/o C**), the specific representations (**w/o S**), and both (**w/o (C, S)**). The comparison results are presented in Table 3.

It can be observed from Table 3 that by removing different representation modules, the performance of the proposed model has undergone varying degrees of changes in terms of all evaluation metrics. Specifically, in terms of accuracy, removing common, and specific information bottleneck sub-modules respectively, as well as simultaneously removing them, result in an average performance decrease of 1.84%, 2.03%, and 4.58%, respectively. It shows that removing both common and specific information bottleneck sub-modules simultaneously results in a cumulative decrease that is similar to removing individual common and specific sub-modules separately, and the specific information bottleneck sub-module gains slightly more improvement than the common information bottleneck sub-module on average, which shows the importance of the considered specific-intrinsic information for the readability assessment.

When concerning different corpora, the introduced SCIB module achieves improvements of 2.01%, 6.67%, 2.32%, 7.51%, and 4.39% in accuracy over these five concerned corpora, respectively. These results show that the introduced SCIB module yields significant improvements to Cambridge, CMER and CMT, and achieves substantial improvements to Weebit and CLEAR. When concerning the impacts of the individual common and specific information bottleneck sub-modules, it can be observed that the specific sub-module gain more improvements overall than the common sub-module over these two Chinese corpora. These indicate that both common and specific sub-modules have learnt important information in text readability evaluation. Similar claims can also be drawn based on the other three evaluation indicators. These results clearly show the effectiveness of the introduced SCIB module.

In order to demonstrate the feasibility of the SCIB module, we visualized the common and specific representations yielded by the proposed model over the English corpus Weebit and the Chinese corpus CMER in Figure 2. It can be observed from Figure 2 that the proposed model can learn the common and specific representations over both English corpus Weebit and Chinese corpus CMER, where the common and specific representations are explicitly decomposed into different subspaces. When concerning different specific representations, the concerned four types of specific representations generally lie in different subspaces, since they generally characterize different levels of information of texts. It can be observed that these three deep specific representations generally lie in three relatively close subspaces because they all focus on deep representations of texts, while the linguistic specific representation lies in a subspace that is different to these deep specific representations, since linguistic features focus on the surface structures of texts and are generally different to deep features. These results verify the feasibility of the proposed idea.

B. On effectiveness of the self-supervised scheme. As shown in Figure 1, we adopted a self-supervised scheme in the SCIB module to supervise the representation refinement. In the following, we conducted a series of experiments over five concerned corpora to demonstrate the effectiveness of the introduced self-supervised scheme. The comparison experimental results are presented in Table 4. It can be ob-



(a) Weebit (b) CMER

Figure 2: Visualization of common and specific representations yielded by the proposed SCIB-ARA model over the English corpus Weebit and the Chinese corpus CMER.

served from Table 4 that the performance of the proposed model is degraded much by removing the self-supervised scheme. Specifically, the performance decreases by 2.22%, 6.67%, 3.59%, 6.84%, and 4.58% in accuracy over these five concerned corpora, respectively, by removing the self-supervised scheme. These results clearly show the effectiveness of the proposed self-supervised scheme.

Dataset	Model	Acc	Pre	F1	QWK
Weebit	SCIB-ARA	**94.07**	**94.26**	**94.10**	**98.03**
	w/o self-supervised	91.85	92.00	91.87	96.92
Cambridge	SCIB-ARA	**95.56**	**95.87**	**95.54**	**98.99**
	w/o self-supervised	88.89	90.41	88.51	97.31
CLEAR	SCIB-ARA	**83.72**	**83.49**	**83.49**	**95.45**
	w/o self-supervised	80.13	79.43	79.59	94.15
CMER	SCIB-ARA	**31.79**	**30.55**	**27.51**	**75.33**
	w/o self-supervised	24.95	23.69	23.29	73.61
CMT	SCIB-ARA	**45.97**	**45.24**	**42.82**	**90.98**
	w/o self-supervised	41.39	35.89	37.70	87.88

Table 4: On effectiveness of the self-supervised scheme.

C. On importance of the linguistic representation. We implemented a series of experiments to demonstrate the importance of the linguistic representation for the proposed model. Experimental results are presented in Table 5, where **w/o L** represents the model without using the linguistic representation. It can be observed from Table 5 that the performance of the proposed model degrades when getting rid of the linguistic representation. Moreover, as shown in Figure 2, the linguistic representation lies in a very different subspace in comparison to deep representations. These show the importance of the linguistic representation for ARA, and in particular that the linguistic representation can provide important supplementary information for deep representations.

Dataset	Model	Acc	Pre	F1	QWK
Weebit	SCIB-ARA	**94.07**	**94.26**	**94.10**	**98.03**
	w/o L	94.01	94.18	94.02	97.97
Cambridge	SCIB-ARA	**95.56**	**95.87**	**95.54**	**98.99**
	w/o L	94.45	94.86	94.41	98.59
CLEAR	SCIB-ARA	**83.72**	**83.49**	**83.49**	**95.45**
	w/o L	82.24	82.11	81.71	94.84
CMER	SCIB-ARA	**31.79**	**30.55**	**27.51**	**75.33**
	w/o L	28.26	29.38	27.12	74.48
CMT	SCIB-ARA	**45.97**	**45.24**	**42.82**	**90.98**
	w/o L	43.04	42.23	40.39	90.07

Table 5: On importance of the linguistic representation.

Conclusion

This paper proposes a novel hybrid ARA model to effectively fuse the deep and linguistic representations, motivated by the concept of information bottleneck. A self-supervised collaborative information bottleneck module is suggested to collaboratively consider the common-intrinsic and specific-intrinsic representations for various levels of deep and linguistic representations. The effectiveness of the proposed model is demonstrated by extensive experiments over four English and two Chinese corpora. Experimental results show that the proposed model outperforms the state-of-the-art models in terms of four important evaluation metrics. Numerous ablation studies are also conducted to demonstrate the feasibility and efficiency of proposed ideas.

Acknowledgements

This work of Jinshan Zeng is supported in part by the National Natural Science Foundation of China [Grant No. 62376110], Double Thousand Plan of Jiangxi Province [Grant No. jxsq2019201124], the Science Fund for Distinguished Young Scholars of Jiangxi Province [Grant No. 20224ACB212004], Jiangxi Provincial Key Laboratory of High Performance Computing [Grant No. 2024SSY03101], the Humanities and Social Science Foundation of Higher Education Institutions of Jiangxi Province [Grant No. YY22209], and the Graduate Innovation Fund of Jiangxi Provincial Department of Education [Grant No. YC2023-S319].

References

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.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv:1607.06450*.

Dale, E.; and Chall, J. S. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, 37–54.

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.

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*, 1–17. Seattle, USA.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019*, 4171–4186.

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.

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

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

Lee, B. W.; Jang, Y. S.; and Lee, J. 2021. Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. In *EMNLP 2021*, 10669–10686.

Lee, B. W.; and Lee, J. 2023. Prompt-based Learning for Text Readability Assessment. In Vlachos, A.; and Augenstein, I., eds., *Findings of the Association for Computational Linguistics: EACL 2023*, 1819–1824. Dubrovnik, Croatia: Association for Computational Linguistics.

Lee, J.; Liu, M.; and Cai, T. 2020. Using Verb Frames for Text Difficulty Assessment. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, 56–62.

Li, W.; Wang, Z.; and Wu, Y. 2022. A Unified Neural Network Model for Readability Assessment with Feature Projection and Length-Balanced Loss. In *EMNLP 2022*, 7446–7457.

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

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

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.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *Information Theory Workshop*. IEEE.

Tseng, H.-C.; Chen, H.-C.; Chang, K.-E.; Sung, Y.-T.; and Chen, B. 2019. An innovative BERT-based readability model. In *International Conference on Innovative Technologies and Learning (ICITL): Innovative Technologies and Learning*, 301–308.

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.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS 2017*, 5998–6008.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical Attention Networks for Document Classification. In *NAACL-HLT 2016*, 1480–1489. San Diego, California.

Yu, D.; Wang, H.; Chen, P.; and Wei, Z. 2014. Mixed pooling for convolutional neural networks. In *Rough Sets and Knowledge Technology: 9th International Conference*, 364–375.

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. In *NeurIPS 2020*, 17283–17297.

Zeng, J.; Xie, Y.; Yu, X.; Lee, J.; and Zhou, D.-X. 2022. Enhancing Automatic Readability Assessment with Pre-training and Soft Labels for Ordinal Regression. In *Findings of EMNLP 2022*, 4557–4568.

Zeng, J.; Yu, X.; Tong, X.; and Xiao, W. 2023. PromptARA: Improving Deep Representation in Hybrid Automatic Readability Assessment with Prompt and Orthogonal Projection.

In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of EMNLP 2023*, 15360–15371.

Zeng, J.; Yu, X.; Tong, X.; and Xiao, W. 2024. InterpretARA: Enhancing Hybrid Automatic Readability Assessment with Linguistic Feature Interpreter and Contrastive Learning. In *AAAI 2024*.