Going Beyond Passages: Readability Assessment for Book-level Long Texts

Wenbiao Li[1,2], **Rui Sun**[1,2], **Tianyi Zhang**[1,2], **Yunfang Wu** [1,3*]
[1]MOE Key Laboratory of Computational Linguistics, Peking University
[2]School of Software & Microelectronics, Peking University
[3]School of Computer Science, Peking University, Beijing, China
{2001210322,2101210366,2301210557}@stu.pku.edu.cn, wuyf@pku.edu.cn

Abstract

Readability assessment for book-level long text is widely needed in real educational applications. However, most of the current researches focus on passage-level readability assessment and little work has been done to process ultra-long texts. In order to process the long sequence of book texts better and to enhance pretrained models with difficulty knowledge, we propose a novel model DSDR, difficulty-aware segment pre-training and difficulty multi-view representation. Specifically, we split all books into multiple fixed-length segments and employ unsupervised clustering to obtain difficulty-aware segments, which are used to re-train the pretrained model to learn difficulty knowledge. Accordingly, a long text is represented by averaging multiple vectors of segments with varying difficulty levels. We construct a new dataset of Graded Children's Books to evaluate model performance. Our proposed model achieves promising results, outperforming both the traditional SVM classifier and several popular pretrained models. In addition, our work establishes a new prototype for book-level readability assessment, which provides an important benchmark for related research in future work.

1 Introduction

The readability assessment is to automatically predict the difficulty level of an input text, which has a wide range of applications, such as automating reader advisory (Pera and Ng, 2014), clinical informed consent forms (Perni et al., 2019) and patient education based on the Internet (Sare et al., 2020).

The readability assessment has a long history of research. In early days, researchers focus on scoring the reading difficulty of texts by designing various readability formulas. Later, many works treat it as a classification task, and various linguistic features are exploited to perform machine learning classification (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009; Qiu et al., 2017). In recent years, traditional linguistic features are combined with the pre-trained model to predict the difficulty level of a text (Qiu et al., 2021; Li et al., 2022). In addition, there are works regarding readability assessment as a regression task (Sheehan et al., 2010) or a ranking task (Ma et al., 2012; Lee and Vajjala, 2022).

Today, reading a whole book plays a vital role to improve children's reading comprehension ability, thus followed by an increasing demand for the book readability assessment. However, most of the existing work processes readability assessment with a passage as input, which is usually an extracted part of a book. These passage-level models have limitations when it comes to the assessment of book-level long texts.

Going from passage-level to book-level readability assessment, a challenging task is how to deal with the long input sequence. As is illustrated in Figure 1, the length of a book text is typically tens times more than that of a passage text. To handle the long sequence simultaneously is a hard work. Taking the pre-trained model BERT as an example, when the mini-batch is 1, using a 24G video memory GPU, the maximum length of slices should be 16 * 512 = 8192. In our children's books data, only 23% [74/320] of the samples were below this value in length. Moreover, the widely applied strategies to process long

*Corresponding author.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1298-1309, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

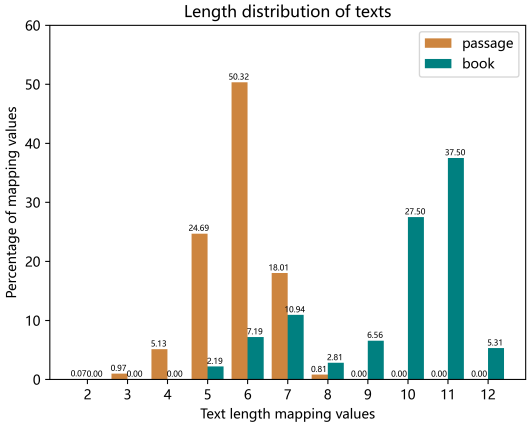

Figure 1: Statistics of length distribution comparing passages and books, derived from textbooks and children's books respectively. Let the text length be x, the mapping function is $f(x) = \lfloor ln(x) \rfloor$.

sequence in passage-level text applications, such as truncating, sliding window and pooling, could not work well for book-level long texts because of too much loss of information.

Nowadays, pre-trained models have made remarkable achievements in many NLP tasks. But these models, such as the pioneer BERT (Devlin et al., 2018) and the long-sequence-specialized BigBird (Zaheer et al., 2020), are mostly based on masked language modeling and next sentence prediction, and thus can not express difficulty knowledge. Moreover, the computational complexity of the multi-head self-attention mechanism in the model structure is proportional to the square of the length of the input sequence. Consequently, directly applying pre-trained models to book readability assessment cannot obtain satisfying results.

In this paper, Firstly, we investigate whether the passage-level readability assessment model can be applied to predict the reading difficulty of a book, and we get a negative answer, indicating that there exists a large gap between passage-level and book-level readability assessment.

Therefore, we propose a two-stage model with **d**ifficulty-aware **s**egment pre-training and multi-view **d**ifficulty **r**epresentation (DSDR) for this challenging task. Specifically, we split all book texts into fixed-length segments, and perform unsupervised clustering to assign each segment with a difficulty label according to linguistic features. The labeled segment difficulty data is exploited to finetune the pre-trained model with a difficulty classification task, in order to enhance the pre-trained model with the ability to recognize and represent difficulty knowledge. In the next stage, the difficulty-aware pre-trained model is employed to extract difficulty features of segments, and multiple difficulty vectors are fused to form the overall representation of a whole book.

We construct a Graded Children's Books (GCB) for book-level readability assessment, including 320 popular stories and famous novels in Chinese. We conduct detailed experiments on this dataset. The experimental results demonstrate that our proposed model achieves an accuracy of 73.125, which significantly outperforms the traditional machine learning method SVM, neural network classifiers and pretrained models. We will make our data and code publicly available for future research.

To sum up, our contributions are as follows:

- We address the readability assessment of book-level long text with pre-trained language models, and construct a Chinese dataset of children's books labelled with difficulty levels.
- We propose a novel DSDR model to deal with the long sequence effectively and enrich the pre-trained models with difficulty knowledge.
- Our proposed model achieves promising results for long text readability assessment, outperforming both the traditional classifier and popular pre-trained models.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1298–1309, Taiyuan, China, July 25 – 28, 2024.
Volume 1: Main Conference Papers

2 Related Work

2.1 Readability Assessment

In the early days, researchers developed a variety of readability formulas, including Flesch (Flesch, 1948), Dale-Chall (Dale and Chall, 1948) and SMOG (Mc Laughlin, 1969). In the field of NLP, readability assessment is often regarded as a classification problem.

For traditional machine learning methods, researchers have used various linguistic features defined by linguists to construct models (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009; Qiu et al., 2017; Lu et al., 2019). Some works (Imperial, 2021; Deutsch et al., 2020; Lee et al., 2021) combined features extracted from neural networks with linguistic features and use traditional machine learning algorithms to classify.

For deep learning methods, (Jiang et al., 2018) purposes knowledge-enriched word embedding (KEWE), which encodes the knowledge of reading difficulty into the representation of words. (Azpiazu and Pera, 2019) present a multi-attentive recurrent neural network architecture for automatic multilingual readability assessment. There are also some works (Qiu et al., 2021; Li et al., 2022) that fuse neural network features with artificial linguistic features.

Besides English and Chinese texts, (Reyes et al., 2022) develop a readability assessment model for Cebuano. However, the existing works don't specialize for long text.

2.2 Long Sequence Processing

For book-level readability assessment, the process of the long sequence input is a challenging task because of the excessive computational complexity of full attention in long sequence processing. Current work on reducing the computational complexity exploits the following methods.

Sparse attention. It reduces the computational complexity by keeping only the connections of surrounding tokens (Beltagy et al., 2020; Zaheer et al., 2020).

Approximation methods. Reformer (Kitaev et al., 2020) replaces dot-product attention by one that uses locality-sensitive hashing and uses reversible residual layers instead of the standard residuals. Performers (Choromanski et al., 2020) use a Fast Attention Via positive Orthogonal Random features approach (FAVOR+) to approximate softmax attention kernels. Linformer (Wang et al., 2020) demonstrates that the self-attention mechanism can be approximated by a low-rank matrix.

Alternative methods. FNet (Lee-Thorp et al., 2021) replaces the self-attention sublayers with simple linear transformations PoNet (Tan et al., 2021) uses a Pooling Network for token mixing in long sequences with a linear complexity.

Compression methods. Pyraformer (Liu et al., 2021) introduces the pyramidal attention module (PAM). Triformer (Cirstea et al., 2022) uses a triangular, variable-specific attention which contains Linear complexity and Variable-specific parameters.

In our work, we choose BigBird with sparse attention as a baseline model and borrow the compression methods to implement PyramidFormer.

2.3 Multi-representation Modeling

In the task of pairwise comparison between sequences, there are usually two strategies, Cross-encoder and Bi-encoder, the former works well but is too slow. Poly-encoder (Humeau et al., 2019) is proposed as a compromise between the two methods. Drawing on the idea of Poly-encoder, some multi-representation work was born. (Kong et al., 2022) design a lightweight network to fuse the aspect embeddings for representing queries and documents. (Xu et al., 2022) introduce a novel multi-task model called Mixture of Virtual-Kernel Experts (MVKE) to learn user preferences on various actions and topics unitedly.

3 Can Passage-level Readability Assessment Model Apply to Long Texts?

Comparing with passage-level readability assessment, book-level readability assessment is less addressed and more challenging. For both levels, our experimental datasets come from the educational

field, and these texts are selected by experts according to children's cognition. Therefore, children who master the classroom content (passage-level data) are expected to be able to read the extended books at the same level. In order to investigate whether the trained passage-level readability assessment models can be directly applied to books, we conduct experiments by training on passages and then perform transfer inference on books.

3.1 Datasets

Passage-level Data Following previous work on passage-level readability assessment, we collected texts in different grades from the textbooks of primary school (TPS) of more than ten publishers, where we deleted poetry and traditional Chinese texts. According to the specification in the *Chinese Curriculum Standards for Compulsory Education* by Ministry of Education, texts of Grade 1-2, Grade 3-4, Grade 5-6 are assigned with the difficulty level 1, 2, 3, respectively.

Book-level Data For book-level readability assessment, we constructed a dataset Graded Children's Books (GCB) in Chinese. We collected 320 books including popular stories and famous novels, with a total of 18 million characters. Each book is assigned with a suitable grade for children to read, according to the *Reading Guidance for School Students by Ministry of Education* [0] and *Basic Bibliography of the Library of Aiyue Primary School* [1]. If there is a conflict for a the same book, the standards of the Ministry of Education shall prevail. The difficulty is also classified into 3 levels, corresponding to Grade 1-2, 3-4, 5-6, respectively.

The statistics of both datasets is shown in Table 1. There are two notable differences between two datasets: 1) The number of book samples is much smaller than that of passage samples. 2) The average length of book texts is much longer than that of passage texts. These two properties pose more challenges for book readability assessment.

Data	Level	Items	Avg.Len	Min.Len	Max.Len
TPS	1	814	266	15	900
	2	1063	679	110	3903
	3	1104	1140	109	6054
	All	2981	737	15	6054
GCB	1	66	12935	175	154517
	2	100	48498	355	203041
	3	154	84391	339	317476
	All	320	58437	175	317476

Table 1: Statistics of the passage-level dataset TPS and the book-level dataset GCB.

Model	Dataset	Precision	Recall	F1
SVM	TPS	87.231	87.333	87.140
	TPS-GCB	52.707	57.812	50.377
TextCNN	TPS	78.378	78.333	78.338
	TPS-GCB	53.469	49.375	38.965
BERT	TPS	83.852	83.667	83.236
	TPS-GCB	56.863	52.812	44.030

Table 2: Transfer inference results from passage-level TPS to book-level GCB.

3.2 Methods

We implement several classical classification models to predict the difficulty level of passages and books. We train models based on passage data and then test on both passage-level and book-level data. Specifically, on the passage data, the training, validation and testing set is split by 8:1:1. And for book-level transfer evaluation, all 320 books are considered as the test data.

SVM. The linguistic difficulty features are extracted using zhfeat (Li et al., 2022) which are normalized using MinMaxScaler (range -1 to 1). We adopt the libsvm [2] framework for our experiments.

[0] http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202004/W020200422556593462993.pdf
[1] http://www.cptoday.cn/news/detail/8309
[2] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1298-1309, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

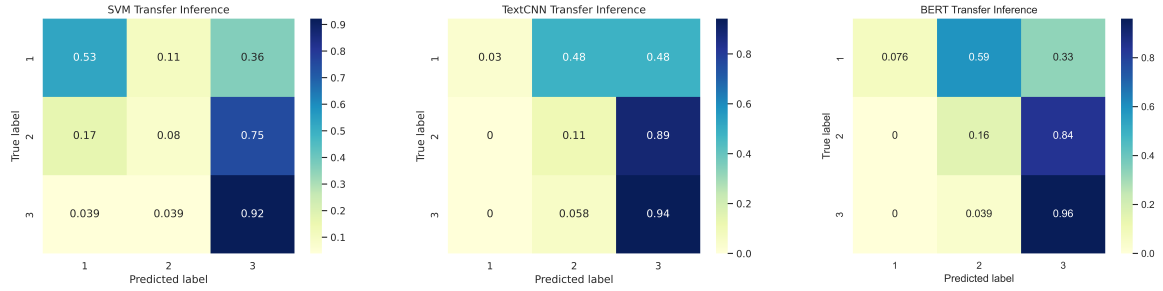

Figure 2: The confusion matrix of transfer inference from TPS to GCB, where the values on the diagonal are the recall of each category.

TextCNN (Kim, 2014). The maximum truncated length of the text is set to 2000. We adopt a random initialized character embedding. The number of convolution kernels is 128, and the size list is [2,3,4].

BERT. The maximum truncated length of the text is set to 2000 tokens, which is cut into 4 segments of 500. We use a sliding window to obtain the $[CLS]$ vector of each segment, and utilize the maximum pooling to obtain the final text representation.

3.3 Results

The experimental results are shown in Table 2. In terms of F1, SVM, TextCNN and BERT have a drop of 36.763 points (87.140-50.377), 39.373 points (78.338-38.965) and 39.206 points (83.236-44.030), respectively. These results demonstrate that the passage-level readability assessment models can't be directly applied to predict the difficulty level of a book. It is necessary to train a model specialized for book-level texts.

To display how the models perform predictions on different difficulty levels, we construct confusion matrices of transfer inference, as shown in Figure 2. Compared with SVM, TextCNN and BERT are sensitive to length and tend to predict a book as a higher difficulty level. We think the bias comes from the direct transfer inference. When we apply a model trained on passages that are relatively short to long texts, the model tends to predict the text as a higher difficulty level. As a result, the model obtains a quite high recall on the highest level but a low recall for the other levels.

4 Readability Assessment of Long Texts

To handle the book-level long text readability assessment, we propose a DSDR model and the overall structure is shown in Figure 3. Based on the Transformer structure, DSDR adds two modules: difficulty-aware segment pre-training and difficulty multi-view representation.

4.1 Difficulty-aware Segment Pre-training

Due to hardware limitations, we cannot directly utilize pre-trained models to process ultra long text. To make full use of the long text without loss of information, we compress a long text into multiple parts that are shorter and can be input into pre-trained models.

A book text is split into several segments, and the reading difficulty of each segment is different, which is also not consistent with the label of the whole book. To obtain a difficulty label for each segment, we classify the segments by clustering according to their linguistic features. Further, we leverage the automatic clustered data to do supervise training, through which the pre-trained model is enhanced with difficulty knowledge.

Concretely, given a book B, we split it into several fixed-length segments $B = (S_1, S_2, ..., S_N)$ to facilitate processing. We perform this segmentation on all books in the data, with G denoting the total number of segments of all books.

Then for each segment S_i, we extract its linguistic difficulty features F_i:

$$F_i = \text{Extractor}(S_i), i = 1, 2, ..., G \quad (1)$$

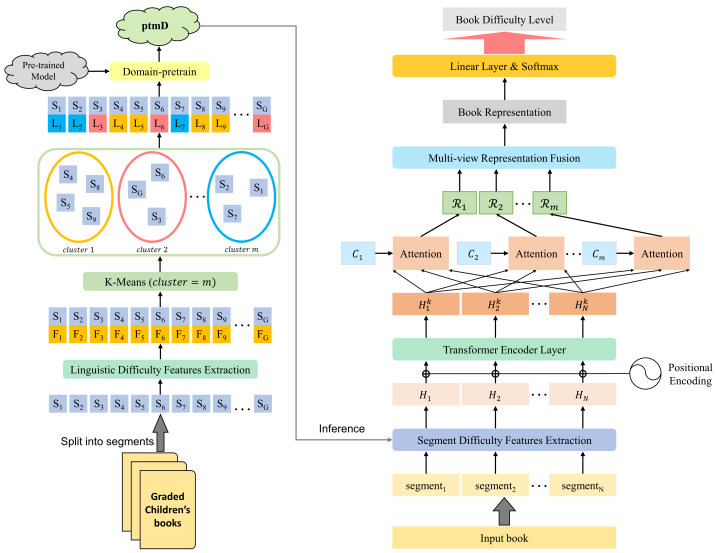

Figure 3: The overall architecture of our proposed DSDR model for book readability assessment.

where Extractor is the feature extractor. Following the previous work (Li et al., 2022), we adopt its Chinese linguistic features, including character, word and sentence features.

By extracting difficulty features in each segment of all the books, we get a series of pairs (S_i, F_i) . Then we exploit the clustering algorithm KMeans to classify segments into m categories:

$$L_i = \text{KMeans}(\text{clusters} = m, F_i), i = 1, ..., G \tag{2}$$

where L_i is the category label of Segment S_i.

In this way, each segment is assigned with a unique label. Gathering all the segments, we construct a dataset $D = \{(S_1, L_1), (S_2, L_2), ..., (S_G, L_G)\}$, in an unsupervised way.

Further, on the dataset D, we employ the pre-trained model BERT followed by a linear layer to do a supervised training, where the training objective is to predict the difficulty label of each segment. Consequently, we obtain a **p**re-**t**rained **m**odel enhanced with **d**ifficulty knowledge:

$$\textbf{ptmD} = \text{BERT}(\text{train} = D) \tag{3}$$

4.2 Difficulty Multi-view Representation

Given an input book, we split it into N segments $B = (S_1, S_2, ..., S_N)$, and then do inference on each segment with **ptmD** to get its text representation:

$$H_i = \textbf{ptmD}(S_i), i = 2, ..., N \tag{4}$$

where H_i is the [CLS] token representation of the segment S_i.

Thus, for a book B, its semantic representation with difficulty knowledge is $H = (H_1, H_2, ..., H_N)$. In this way, a long text is compressed and represented with H without loss of difficulty information.

Since the semantics of each segment is extracted independently, context information is lacking. Therefore, we apply the Transformer (Vaswani et al., 2017) encoder to supplement the contextual representation:

$$H^0 = H + P \tag{5}$$

$$H^t = \text{Transformer}(H^{t-1}), t = 1, 2, ..., k \tag{6}$$

where $P = (P_1, P_2, ..., P_N)$ is the positional encoding sequence, P_i denoting the position of the i-th segment in the book. k is the number of encoder layers. The calculation method is consistent with the paper (Vaswani et al., 2017).

Intuitively, a long text might contain information of various difficulty levels. To further exploit the difficulty knowledge, we set multiple difficulty representation vectors $C \in \mathbb{R}^{m \times d}$, where d is the hidden size, and m is the number of difficulty levels that is the same as the number of clusters set in KMeans.

Each difficulty representation vector $C_i \in \mathbb{R}^{1 \times d}$ is trained to extract the information of the corresponding difficulty level. The difficulty representation vectors C are utilized to extract information from H^k using cross-attention. The extracted information $R \in \mathbb{R}^{m \times d}$ are representations of m difficulty levels:

$$R = \text{Attention}(C\text{W}^{\text{Q}}, H^k\text{W}^{\text{K}}, H^k\text{W}^{\text{V}}) \tag{7}$$

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{Softmax}(\frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}}})\text{V} \tag{8}$$

where W^{Q}, W^{K} and W^{V} are trainable parameters. $R = (R_1, R_2, R_3, ..., R_m) \in \mathbb{R}^{m \times d}$ is the difficulty multi-view representation, where R_i is the singe-view representation of segment difficult Level i, which is not directly mapped to the book's difficulty level.

4.3 Difficulty Level Prediction

Based on multi-view representations, we employ mean pooling to obtain the overall representation:

$$T = \text{MeanPooling}(R_1, R_2, R_3, ..., R_m) \tag{9}$$

Finally, we compute the probability that a book belongs to the i-th difficulty category by:

$$p_i = \text{Softmax}(T\text{W} + \text{b}) \tag{10}$$

where W are trainable parameters, and b are scalar biases.

5 Experimental Setup

5.1 Data and Metrics

We conduct experiments on our new data GCB, consisting of 320 children's books labelled with difficulty levels. The statistics of GCB is listed in Table 1. We apply stratified k-fold (k=5, train=0.8, test=0.2) and report the average results.

For evaluation, we calculate the weighted precision, recall, F1 score and accuracy.

5.2 Comparing Methods

We provide a comprehensive comparison with previous approaches, including the traditional SVM, the neural network classifiers (TextCNN, DPCNN) and the pre-trained models (BERT, BigBird).

SVM-L. Only the length features are input into the SVM classifier.

SVM. **TextCNN**. **BERT**. The details are the same as in Section 3.2, except for the maximum length in TextCNN and BERT.

DPCNN (Johnson and Zhang, 2017). It increases the number of layers of CNN to extract long-range text dependencies. The number of convolution kernels is 128.

BERT-FP-LBL (Li et al., 2022). We apply the most recent passage-level model on Chinese data, where the adjustment factor ρ is 0 and other parameters remain the same as in the original paper.

BigBird (Zaheer et al., 2020). It uses the sparse attention to handle sequences that are eight times longer than usual.

PyramidFormer. Inspired by the work (Cirstea et al., 2022), we designed a model for long sequence compression, in which we replace the self-attention in BERT with the linear patch attention. We use random initialized character embeddings.

DS+TextCNN. **DS+DPCNN**. We apply our **ptmD** model to get the difficulty-aware text representation H, and use H as the input sequence of TextCNN and DPCNN, respectively. Other settings remain unchanged.

5.3 Training Details

We do experiments using the Pytorch (Paszke et al., 2019) framework. For training, we use the AdamW optimizer, the weight decay is 0.02 and the warm-up ratio is 0.1. For the difficulty-aware segment pre-training in Section 4.1, the number of difficulty categories m is 14, the segment length is 500, the batch size is 16, the epoch is 10 and the learning rate is 3e-5. For the difficulty multi-view representation in Section 4.2, the batch size is 16, the number of Transformer encoder layers k is 3, the epoch is 100 and the learning rate is 3e-5. We choose the length as 180000 which covers 95% of the books. Accordingly, the book is represented as 360 (180000/500) segments, as shown in Table 3.

6 Results and Analysis

Model	Max.Len	Embedding	Precision	Recall	F1	Accuracy
SVM-L	Unlimited	-	62.908	66.250	63.002	66.250
SVM	Unlimited	-	68.587	68.750	68.228	68.750
TextCNN	180000	180000 × 64	58.488	55.937	46.401	55.937
DPCNN	180000	180000 × 64	68.860	68.750	68.142	68.750
BERT	8192	16 × 512 × 768	50.258	61.875	54.309	61.875
BERT-FP-LBL	4000	8 × 500 × 768	61.420	63.437	57.178	63.437
BigBird	4096	4096 × 768	53.710	61.250	51.146	61.250
PyramidFormer	180000	180000 × 64	69.061	69.062	68.253	69.062
DS + TextCNN	180000	360 × 768	69.306	69.062	67.997	69.062
DS + DPCNN	180000	360 × 768	72.338	70.625	70.831	70.625
DSDR	180000	360 × 768	**73.266**	**73.125**	**72.466**	**73.125**

Table 3: Overall experimental results on children's books for readability assessment.

6.1 Overall Results

The overall experimental results are summarized in Table 3. Our proposed DSDR model achieves the best performance on all metrics. Compared with the traditional SVM classifier, our model improves by 4.238 on F1. The simple SVM classifier with only the length feature (SVM-L) obtains an F1 score of 63.002, which is much better than TextCNN, BERT and BigBird, demonstrating that the length feature plays a vital role for readability assessment. But it is about 10 points lower than our model DSDR.

Though the pre-trained models are becoming mainstream for various NLP tasks, it is not feasible to directly apply them to our task. Using the pretrained model BERT only obtains an F1 score of 54.309. Even for the pre-trained model specialized for long text like BigBird, it's too long of our book text to be compressed as an input, resulting in a poor result of 51.146 F1 score. In contrast, our model DSDR processes the long input sequence effectively and at the same time incorporates difficulty knowledge, outperforming BigBird by more than 20 points on F1 and more than 10 points on accuracy.

Compared with TextCNN and DPCNN, both DS+TextCNN and DS+DPCNN bring big improvements by applying our trained **ptmD** to get the difficulty-aware representation as model input. It improves F1 by 21.596 points over TextCNN, and outperforms the advanced DPCNN by 2.689 F1 points. This validates that our difficulty-aware segment pre-training is flexible, and can be easily applied to other neural network models to improve readability assessment performance.

6.2 Ablation Study

We explore the contribution of each component in our DSDR model by conducting ablation studies with the following settings:

-DPT: Do not apply difficultty-aware segment pre-training, instead, we use the original pretrained model to extract difficulty features from segments.

-MVDR: Do not use multi-view representations.

The results are shown in Table 4. Without applying difficulty-aware segment pre-training leads to an obvious drop of performance, with about 5 points on both F1 and accuracy. In contrast, after exploiting difficulty-aware segment pre-training on our automatically constructed data, the pre-trained model learns to extract segment difficulty features.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1298-1309, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

Model	F1	Acc
DSDR	72.466	73.125
-DPT	67.402	68.125
-MVDR	71.142	71.563

Table 4: Ablation study on the model structure.

Without the use of difficulty multi-view representations (MVDR) also leads to a drop of performance. This indicates that, for a book-level long text with varying difficulty levels, it's important to set up multiple difficulty representations.

6.3 Confusion Matrix

To clearly display the difficulty levels predicted by different models, we construct confusion matrices based on the predicted results of SVM, BERT, BigBird and DSDR. The results are shown in Figure 4. Compared with SVM, our DSDR model obtains an obvious performance gain on Level 1, and compared with BERT and BigBird, our model achieves significant improvements on Level 2.

As shown in Figure 4, BERT and BigBird obtain low recall value on Level 2 and high recall value on Level 3. For BERT and Bigbird, we need to truncate the input text into a certain length to feed into the pre-trained model, which loses length information.

After truncation, the texts of Level 2 and Level 3 have similar lengths. As a result, the model is likely to predict texts of Level 2 as Level 3, which causes the low recall in Level 2 and the false high recall in Level 3. In contrast, our model represents the long text with multiple difficulty-enriched segments, avoiding the loss of information.

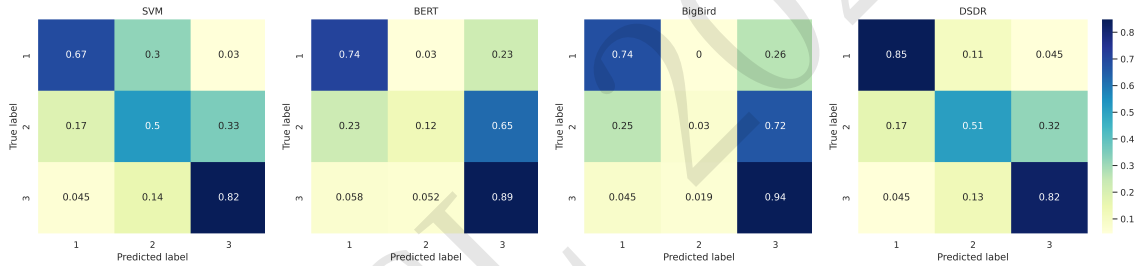

Figure 4: Comparative analysis of confusion matrix.

6.4 Analysis on the Number of Clusters

During difficulty-aware segment pre-training, we exploit the clustering algorithm KMeans to classify different segments into multiple categories. To explore the effect of the number of clusters on the results, we set up different m and report the experimental results in Figure 5.

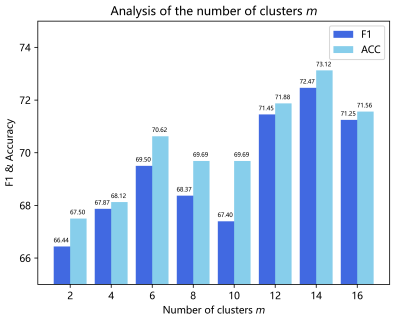

Figure 5: Effect of the the number of clusters m.

When m is small, the model's ability to distinguish difficulty levels is weak. When m is relatively big, more difficulty fragments are extracted, and the overall performance of modeling as multiple representations is better. When m=14, we obtain the best results on both F1 and accuracy.

6.5 Experiments on Different Multi-view Representation Fusion Methods

With reference to the work (Kong et al., 2022), we design several different ways of difficulty multi-view representation fusion.

Average Weight (AW). The method used in our paper.

Random Weight (RW). We set a learnable parameter $\alpha \in \mathbb{R}^{1\times m}$, and compute the weight of each difficulty representation:

$$T = \sum_{i=1}^{m} \frac{e^{\alpha_i}}{e^{\alpha_1} + \cdots + e^{\alpha_m}} R_i \tag{11}$$

Transforming Weight (TW). We set a linear layer (LL) to transform each difficulty representation to 1 dimension, and compute the weights by:

$$T = \sum_{i=1}^{m} \frac{\mathrm{LL}(R_i)}{\mathrm{LL}(R_1) + \cdots + \mathrm{LL}(R_m)} R_i \tag{12}$$

The experimental results are shown in Table 5. The fusion method does not have a significant effect on the results. We believe this is because our difficulty vector C is learned as a parameter and is able to adapt itself to changes in the weights of individual difficulty representations, and thus it is not sensitive to fusion methods.

Method	P	R	F1	Acc
AW	73.266	73.125	72.466	73.125
RW	72.738	73.125	72.147	73.125
TW	73.464	72.500	72.541	72.500

Table 5: Experiments on different fusion methods.

6.6 Case Study

To display our interesting task of readability assessment for children's books, we list some examples in Table 6, reporting the predicted results by DSDR and some baseline models.

Title	Length	True level	Predict Label			
			SVM	BERT	BigBird	DSDR
绿山是一个谜(Green Mountain is a mystery)	11191	1	1	3	3	1
小飞人卡尔松 (Peter Pan Karlsson)	154340	2	3	3	3	2
帽子的秘密 (The Secret of the Hat)	50805	2	3	3	3	2
小兵张嘎 (Little Soldier Zhang Ga)	54901	3	2	3	3	3
布鲁克林有棵树 (A Tree Grows in Brooklyn)	277362	3	1	3	3	2

Table 6: Case study

7 Conclusion

In this paper, we propose a novel model DSDR to assess the reading difficulty of book-level long texts. Through difficulty-aware segment pre-training, the pre-trained model is enhanced with the ability to represent difficulty knowledge, and the difficulty multi-view representations help to compress a long text into a representation without loss of information. In addition, we construct a new dataset Graded Children's Books in which each book is labelled with a difficulty level. Experimental results show that our approach achieves superior performance against traditional methods and popular pre-trained language models.

Acknowledgements

This work is supported by the Key Project of Natural Science Foundation of China (61936012) and the National Natural Science Foundation of China (62076008).

References

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Razvan-Gabriel Cirstea, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, and Shirui Pan. 2022. Triformer: Triangular, variable-specific attentions for long sequence multivariate time series forecasting–full version. *arXiv preprint arXiv:2204.13767*.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Joseph Marvin Imperial. 2021. Bert embeddings for automatic readability assessment. *arXiv preprint arXiv:2106.07935*.

Zhiwei Jiang, Qing Gu, Yafeng Yin, and Daoxu Chen. 2018. Enriching word embeddings with domain knowledge for readability assessment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 366–378.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing*.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Weize Kong, Swaraj Khadanga, Cheng Li, Shaleen Kumar Gupta, Mingyang Zhang, Wensong Xu, and Michael Bendersky. 2022. Multi-aspect dense retrieval. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3178–3186.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. *arXiv preprint arXiv:2203.07450*.

Bruce W Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *arXiv preprint arXiv:2109.12258*.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*.

Wenbiao Li, Ziyang Wang, and Yunfang Wu. 2022. A unified neural network model for readability assessment with feature projection and length-balanced loss. *arXiv preprint arXiv:2210.10305*.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. 2021. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*.

Dawei Lu, Xinying Qiu, and Yi Cai. 2019. Sentence-level readability assessment for l2 chinese learning. In *Workshop on Chinese Lexical Semantics*, pages 381–392. Springer.

Yi Ma, Eric Fosler-Lussier, and Robert Lofthus. 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*, pages 548–552.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Maria Soledad Pera and Yiu-Kai Ng. 2014. Automating readers' advisory to make book recommendations for k-12 readers. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 9–16.

Subha Perni, Michael K Rooney, David P Horowitz, Daniel W Golden, Anne R McCall, Andrew J Einstein, and Reshma Jagsi. 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.

Sarah E Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.

Xinying Qiu, Kebin Deng, Likun Qiu, and Xin Wang. 2017. Exploring the impact of linguistic features for chinese readability assessment. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 771–783. Springer.

Xinying Qiu, Yuan Chen, Hanwu Chen, Jian-Yun Nie, Yuming Shen, and Dawei Lu. 2021. Learning syntactic dense embedding with correlation graph for automatic readability assessment. *arXiv preprint arXiv:2107.04268*.

Lloyd Lois Antonie Reyes, Michael Antonio Ibañez, Ranz Sapinit, Mohammed Hussien, and Joseph Marvin Imperial. 2022. A baseline readability model for cebuano. *arXiv preprint arXiv:2203.17225*.

Antony Sare, Aesha Patel, Pankti Kothari, Abhishek Kumar, Nitin Patel, and Pratik A Shukla. 2020. Readability assessment of internet-based patient education materials related to treatment options for benign prostatic hyperplasia. *Academic Radiology*, 27(11):1549–1554.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

Kathleen M Sheehan, Irene Kostin, Yoko Futagi, and Michael Flor. 2010. Generating automated text complexity classifications that are aligned with targeted text complexity standards. *ETS Research Report Series*, 2010(2):i–44.

Chao-Hong Tan, Qian Chen, Wen Wang, Qinglin Zhang, Siqi Zheng, and Zhen-Hua Ling. 2021. Ponet: Pooling network for efficient token mixing in long sequences. *arXiv preprint arXiv:2110.02442*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Zhenhui Xu, Meng Zhao, Liqun Liu, Lei Xiao, Xiaopeng Zhang, and Bifeng Zhang. 2022. Mixture of virtual-kernel experts for multi-objective user profile modeling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4257–4267.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.