Unsupervised Readability Assessment via Learning from Weak Readability Signals

Yuliang Liu
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
mf21330057@smail.nju.edu.cn

Zhiwei Jiang*
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
jzw@nju.edu.cn

Yafeng Yin
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
yafeng@nju.edu.cn

Cong Wang, Sheng Chen
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
{cw,mg21330006}@smail.nju.edu.cn

Zhaoling Chen
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
zhaolingchen@smail.nju.edu.cn

Qing Gu
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
guq@nju.edu.cn

ABSTRACT

Unsupervised readability assessment aims to evaluate the reading difficulty of text without any manually-labeled data for model training. This is a challenging task because the absence of labeled data makes it difficult for the model to understand what readability is. In this paper, we propose a novel framework to Learn a neural model from Weak Readability Signals (LWRS). Instead of relying on labeled data, LWRS utilizes a set of heuristic signals that specialize in describing text readability from different aspects to guide the model in outputting readability scores for ranking. Specifically, to effectively use multiple heuristic weak signals for model training, we build a multi-signal learning model that ranks the unlabeled texts from multiple readability-related aspects based on intra- and inter-signal learning. We also adopt the pairwise ranking paradigm to reduce the cascade coupling among partial-order pairs. Furthermore, we propose identifying the most representative signal based on the batch-level consensus distribution of all signals. This strategy helps identify the predicted signal that is most correlated with readability in the absence of ground-truth labels. We conduct experiments on three public readability assessment datasets. The experimental results demonstrate that our LWRS outperforms each heuristic signal and their combinations significantly, and can even perform comparably with some supervised methods. Additionally, our LWRS trained on one dataset can be effectively transferred to other datasets, including those in other languages, which indicates its good generalization and potential for wide application.

*Corresponding author.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591695

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Information systems** → **Content analysis and feature selection**.

KEYWORDS

Readability Assessment, Unsupervised Ranking, Multi-Signal Learning, Pairwise Ranking

ACM Reference Format:
Yuliang Liu, Zhiwei Jiang, Yafeng Yin, Cong Wang, Sheng Chen, Zhaoling Chen, and Qing Gu. 2023. Unsupervised Readability Assessment via Learning from Weak Readability Signals. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591695

1 INTRODUCTION

Readability assessment aims to evaluate the reading difficulty of a given text according to its linguistic factors such as sentence length, lexical difficulty, and grammatical complexity. It is helpful and commercially valuable in many fields, such as educational applications [12, 26, 30], information retrieval [19, 35], and recommender systems [2, 3, 43].

Research on readability assessment has spanned the last century [30]. As the field has developed, the mainstream approach has shifted from readability formulas to traditional machine learning methods and, more recently, to state-of-the-art deep learning methods. Compared with readability formulas and handcrafted-features-based machine learning methods, deep learning methods have been demonstrated to be more effective due to their ability to representation learning. Nevertheless, despite their effectiveness, most of them treat the task as a supervised learning problem and cannot be well-trained without labeled data.

As the process of collecting manually-scored texts for model training is both time-consuming and labor-intensive, the development of Unsupervised Readability Assessment (URA) method is crucial. The most commonly used URA method is reusing existing

readability formulas. However, it often fails to achieve satisfactory performance due to the issues of inconsistent data distribution and a limited number of employed features. To this end, researchers have subsequently developed many heuristic URA methods in terms of specific perspectives, such as the coherence of semantic concepts [1, 19], the vocabulary difficulty [11], or the Perplexity (PPL) output by pre-trained language models [32]. The major limitations of these methods are their dependence on static heuristic calculation algorithms and consideration of only a specific perspective of readability, which cannot provide a dynamic and comprehensive assessment of readability for texts with different data distributions.

In this paper, we propose a novel framework to Learn a neural model from Weak Readability Signals (LWRS) for unsupervised readability assessment. By introducing a set of heuristic readability signals as supervision, this framework allows us to train a neural model that can dynamically adapt to new data and can comprehensively take multiple readability-related aspects into consideration. Compared to previous URA methods, our LWRS can benefit from information from both raw text and the consensus of multiple weak readability signals and thus is promising to achieve better performance. Specifically, to effectively use multiple heuristic weak signals for model training, we build a multi-signal learning model to rank the unlabeled texts from multiple readability-related aspects and enhance its ranking ability based on intra- and inter-signal pairwise ranking. Furthermore, we propose identifying the most representative signal based on the batch-level consensus distribution of all signals. This strategy helps identify the predicted signal that is most correlated with readability in the absence of ground-truth labels.

The major contributions of this paper can be summarized as follows:

- We propose an unsupervised readability assessment framework based on Learning from Weak Readability Signals (LWRS), which can get rid of the requirement of ground-truth labels by utilizing a set of heuristic signals as supervision.
- We build a multi-signal learning model to predict enhanced signals based on intra- and inter-signal pairwise ranking, and can well identify the predicted signal that is highly correlated with readability in an unsupervised way.
- Experimental results on three public datasets demonstrate the effectiveness and good transferability of our LWRS under the unsupervised setting, indicating the feasibility of training neural model with multiple weak readability signals for unsupervised readability assessment.

2 RELATED WORK

In this section, we briefly introduce the following two aspects relevant to readability assessment.

2.1 Supervised Readability Assessment

Research on readability assessment has lasted for about a century [30] and most of them focused on supervised learning methods. Early work revolved around designing readability formulas, which are typically structured as the linear regression of several easy-to-compute surface-level statistics of texts, such as average sentence length (ASL) and average word length (AWL). Some of the famous formulas include the Gunning Fog Index [15], Flesch Reading Ease [24]. Until the begin of this century, to take more statistics into consideration, researchers thereafter explore various linguistic features [6, 9, 16, 23] along with various classification, regression, and ranking models [22, 38, 41]. Such complicated feature engineering and machine learning algorithms brought great performance improvement. In recent years, researchers have turned their attention to deep learning techniques. With the help of representation learning, the performance of readability assessment has been further improved [18, 21, 25, 34, 37].

As most of these studies exploring text representation and model design mainly focus on English, some research has investigated readability in languages other than English, such as German [36], Filipino [17] and Spanish [13]. In addition to these studies on a single language, there are also studies devoted to the issue of multilingual readability assessment [5, 27, 31]. For example, Lee et al. [27] explored zero-shot cross-lingual evaluation for English to Spanish and English to French tasks.

2.2 Unsupervised Readability Assessment

Unsupervised Readability Assessment (URA) is a good complement to supervised readability assessment, as it can handle scenarios where labeled data is unavailable. The most commonly used URA method is reusing existing readability formulas, but it often fails to achieve satisfactory performance. Thereafter, researchers have developed many other URA methods [1, 11, 19, 32]. Jameel et al. [19, 20] evaluate the readability cost of texts in terms of the sequential n-gram cohesion. Ehara [11] proposes performing readability assessment based on vocabulary tests and accurately estimated word difficulty. Martinc et al. [32] design a heuristic ranked sentence readability score based on the Perplexity (PPL) output by pre-trained language models.

3 TASK DEFINITION

We first introduce some notations and formalize the Unsupervised Readability Assessment (URA) task. Let $\boldsymbol{X} = \{x_i\}_{i=1}^{N}$ denote a set of texts, $\boldsymbol{Y} = \{1, 2, ..., K\}$ denote a set of readability levels at ordinal scale, and $(x \in \boldsymbol{X}, y \in \boldsymbol{Y})$ denote a text and its corresponding ground-truth readability level respectively. In URA task, we assume that only unlabeled texts $\boldsymbol{X}_0 = \{x_i\}_{i=1}^{N_0} \in \boldsymbol{X}$ without corresponding $\boldsymbol{Y}_0$ are given for training. Besides, a test set $\boldsymbol{X}_u = \{x_i\}_{i=1}^{N_u} \in \boldsymbol{X}$ with corresponding ground-truth $\boldsymbol{Y}_u$ is set for testing, where $\boldsymbol{X}_0 \cap \boldsymbol{X}_u = \emptyset$. The objective of URA is to learn a function F based on $\boldsymbol{X}_0$ and predict the readability score $\hat{\boldsymbol{Y}}_u$ of texts in $\boldsymbol{X}_u$, where the predicted $\hat{\boldsymbol{Y}}_u$ is expected to be close to $\boldsymbol{Y}_u$. With the help of label information $\boldsymbol{D}$ and unlabeled texts $\boldsymbol{X}_0$, F can be defined as:

$$\hat{y} = F(x; \boldsymbol{D}, \boldsymbol{X}_0) \tag{1}$$

For typical readability assessment tasks under the setting of supervised learning, $\boldsymbol{D}$ in Eq.1 can be replaced with $\boldsymbol{Y}_0$. However, under the unsupervised readability assessment setting, $\boldsymbol{Y}_0$ is not available. To address this problem, we propose to use weak readability signals as D instead for the learning of F.

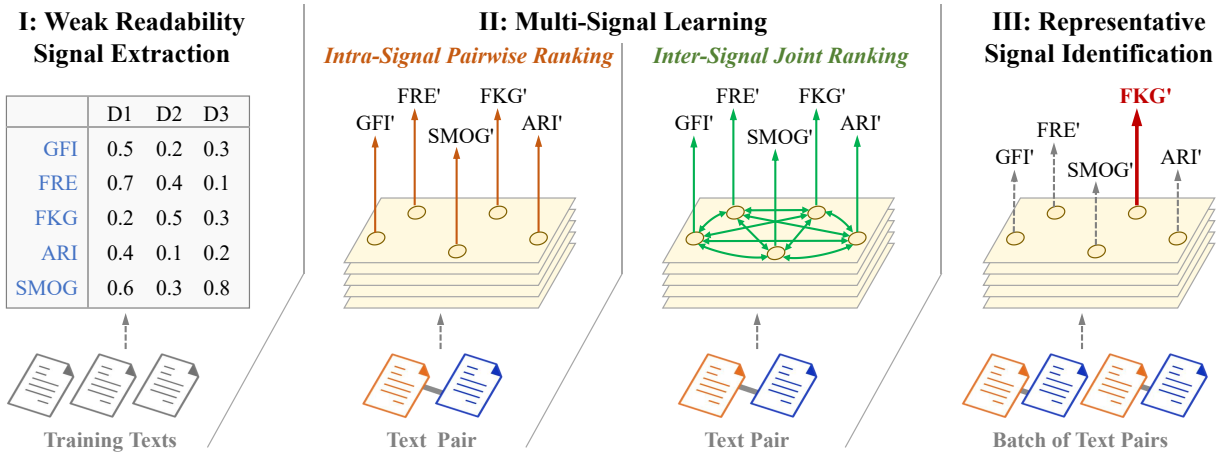

Figure 1: Illustration of the LWRS framework for the unsupervised readability assessment task.

4 THE PROPOSED FRAMEWORK

In this section, we introduce the proposed LWRS framework, followed by its technical details.

4.1 An Overview of LWRS

For the URA task, we propose an LWRS framework to Learn the neural URA model based on Weak Readability Signals. LWRS is designed based on the multi-task learning paradigm, treating the prediction of each readability signal as a task. It enables the prediction of multiple signals to benefit from each other, achieving better performance than predicting each signal separately. Since multiple weak readability signals may exhibit different values, but similar partial order, we consider viewing the prediction of each signal as a ranking problem rather than a regression problem. To alleviate the conflicts among signals, we adopt the pairwise ranking scheme, which can reduce the cascade effect of partial-order pairs in the rank list. Compared with the manually extracted signals, the predicted signals can benefit from the information of both text input and other signals, resulting in improved readability assessment. To determine the final readability of a text, we further design an algorithm to identify the most representative signal among all predicted signals.

Specifically, as shown in Figure 1, LWRS comprises three main components: *Weak Readability Signal Extraction* which provides the supervision information $\boldsymbol{D}$ for model training; *Multi-Signal Learning* which learns multiple readability assessment functions $\boldsymbol{F}$ based on multi-task learning and pairwise ranking; *Representative Signal Identification* which identifies the most representative function $\boldsymbol{F}$ for testing based on the consensus of multiple predicted signals. In the following, we introduce these components of LWRS along with their technical details.

4.2 Weak Signals Extraction

According to previous studies on supervised readability assessment [15, 24], one or several surface signals are not enough for comprehensive readability assessment and more signals often bring better performance. This implies that the commonly-used surface signals are weak but correlated with readability to some extent, which provides us a chance to perform unsupervised readability assessment based on these weak signals.

4.2.1 ***Weak Readability Signals.*** To avoid introducing excessive complexity, we only consider extracting the surface signals and the readability formulas based on surface signals. Examples of these surface signals are the average number of syllables per word, the average number of characters per word, the number of words, and so on. Famous readability formulas are also taken into account, which include the Gunning Fog Index [15], the Flesch-Kincaid Grade [24], and so on.

Specifically, we totally employ N_s weak signals and define the set of manually-extracted original signals as $\boldsymbol{O} = \{O_1, O_2, ..., O_{N_s}\}$. Then, as shown in part I of Figure 1, for each text $x_i \in \boldsymbol{X}_0$ of the unlabeled training set, we can extract N_s weak signals from it, where the j-th signal of x_i is denoted as $O_j(x_i)$.

4.2.2 ***Signal Normalization.*** Although these signals are all correlated with readability, they often have different ranges of signal values. To prepare for later *Inter-Signal Joint Ranking*, we normalize the values of all signals into the same range.

Specifically, for a specific j-th signal, we denote its raw and normalized vector of extracted signal values as $O_j \in \mathbb{R}^{N_0}$ and $R_j \in \mathbb{R}^{N_0}$, respectively. Then, we can restrict $R_j(x_i) \in [-1, 1]$ by performing the normalization operation for each signal:

$$R_j = \frac{O_j}{max(\|O_j\|_2, \epsilon)}, j = \{1, 2, ..., N_s\}, \tag{2}$$

where ϵ is a small number to avoid zero division. This operation does not change the partial order of texts defined by the signals.

4.3 Multi-Signal Learning

Motivated by multi-task learning, we build a multi-signal learning model, expecting the model can produce better readability predictions than each original weak signal, with the help of information from both input text and multiple weak signals. In the following, we first describe the architecture of the multi-signal learning model and then introduce two loss functions for model training, which

include the intra-signal pairwise ranking loss and the inter-signal joint ranking loss. Finally, we summarize the overall training loss.

4.3.1 ***Model Architecture.*** We employ an encoder $f_\phi(\cdot)$ to extract features of an input x_i, where $f_\phi(x_i;\phi)$ refers to the embedding of x_i and ϕ indicates the parameters of the encoder. Then, we employ N_s linear layers $f_{\theta_j}(\cdot)$ to predict the readability score of each text $x_i \in \boldsymbol{X}_0$, where $S_j(x_i) = f_{\theta_j}(f_\phi(x_i;\phi);\theta_j)$ refers to the j-th predicted signal for i-th text and θ_j denotes the parameter of the corresponding scoring layer. Here, the encoder $f_\phi(\cdot)$ is a shared encoder of N_s linear layers and can be a text encoder such as the pre-trained BERT [10] or RoBERTa [29].

4.3.2 ***Intra-Signal Pairwise Ranking.*** To train the multi-signal learning model, we first design an intra-signal loss for each output branch of signal prediction. Instead of using the conventional MSE loss for signal regression, we introduce a pairwise ranking loss to capture the partial order relationship among texts, which helps avoid overfitting to unreliable values of weak signals. Considering that the texts with similar signal values may have the same readability level, we propose a three-class pairwise ranking loss, which sets an extra class for the 'equal' relation in addition to the relations of 'greater than' and 'less than'.

Specifically, as shown in Figure 2, given a text pair (d_a, d_b) randomly sampled from $\boldsymbol{X}_0$, N_s original signals can be manually extracted for each of d_a and d_b, where the i-th original signal of d_a and d_b are denoted as $R_i(d_a)$ and $R_i(d_b)$, respectively. For each text pair (d_a, d_b), we assign N_s pair-level labels $\{z^i_{ab}\}_{i=1}^{N_s}$ to it corresponding to N_s weak signals. We categorize the pair-level labels z^i_{ab} into three classes $\{1, -1, 0\}$ according to the difference of values between $R_i(d_a)$ and $R_i(d_b)$:

$$z^i_{ab} = \begin{cases} 1, & \text{if } R_i(d_a) - R_i(d_b) > \lambda_i; \\ -1, & \text{if } R_i(d_a) - R_i(d_b) < -\lambda_i; \\ 0, & \text{otherwise;} \end{cases} \tag{3}$$

where λ_i is a pre-defined threshold for the i-th original signal and is used to define whether two original signal values are distinguishable.

Since the predicted signals for each text in a text pair (d_a, d_b) are $S_i(d_a)$ and $S_i(d_b)$ respectively, the predicted pair-level label of (d_a, d_b) can also be defined as:

$$\hat{z}^i_{ab} = \begin{cases} 1, & \text{if } S_i(d_a) - S_i(d_b) > \Lambda_i; \\ -1, & \text{if } S_i(d_a) - S_i(d_b) < -\Lambda_i; \\ 0, & \text{otherwise;} \end{cases} \tag{4}$$

where Λ_i is a learnable threshold for the i-th predicted signal and is used to define whether two predicted signal values are distinguishable.

Based on the above definition, we can estimate the probabilities of $\hat{z}^i_{ab} = 1$, -1, and 0, respectively. For convenience, we define a intermediate variable Δ^i_{ab} and a sigmoid function $Q(\cdot)$:

$$\Delta^i_{ab} = S_i(d_a) - S_i(d_b), \tag{5}$$

$$Q(x) = \frac{1}{1+e^{-x}}. \tag{6}$$

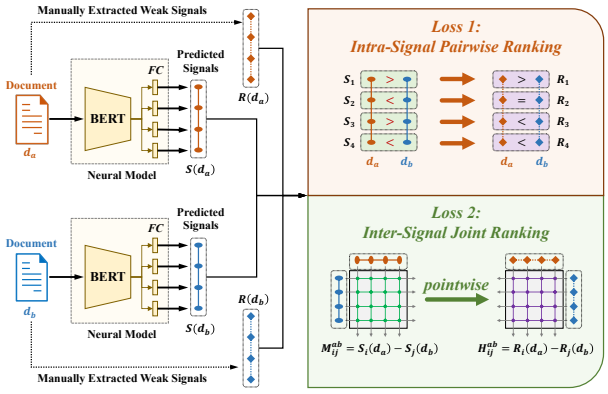

Figure 2: Illustration of multi-signal learning.

Based on Δ^i_{ab} and $Q(\cdot)$, we have the probability of $\hat{z}^i_{ab} = 1, -1, 0$:

$$\begin{aligned} P(\hat{z}^i_{ab} = \ 1 \mid \phi, \theta_i) &= P(\Delta^i_{ab} - \Lambda_j > 0 \mid \phi, \theta_i) \\ &= Q(\Delta^i_{ab} - \Lambda_j), \\ P(\hat{z}^i_{ab} = -1 \mid \phi, \theta_i) &= P(-\Lambda_j - \Delta^i_{ab} > 0 \mid \phi, \theta_i) \\ &= Q(-\Lambda_j - \Delta^i_{ab}), \\ P(\hat{z}^i_{ab} = \ 0 \mid \phi, \theta_i) &= P(-\Lambda_j \le \Delta^i_{ab} \le \Lambda_j \mid \phi, \theta_i) \\ &= 1 - Q(\Delta^i_{ab} - \Lambda_j) - Q(-\Lambda_j - \Delta^i_{ab}), \end{aligned} \tag{7}$$

Then we could have the negative log-likelihood function for the i-th signal prediction task:

$$\begin{aligned} & -log(P(\hat{z}^i_{ab} = z^i_{ab} \mid \phi, \theta_i)) \\ &= \sum_{q \in \{-1,1,0\}} -[z^i_{ab} = q] log(P(\hat{z}^i_{ab} = q \mid \phi, \theta_i)) \end{aligned} \tag{8}$$

where $[\mathcal{B}] = 1$ if event $\mathcal{B}$ happens, otherwise $[\mathcal{B}] = 0$.

By accumulating the negative log-likelihood function of N_s signal prediction tasks, we get the intra-signal pairwise loss:

$$\mathcal{L}_{intra} = -\sum_{i=1}^{N_s} log(P(\hat{z}^i_{ab} = z^i_{ab} \mid \phi, \theta_i)) \tag{9}$$

4.3.3 ***Inter-Signal Joint Ranking.*** To enable these signals to benefit from each other, we build a signal interaction matrix $H^{ab} \in \mathbb{R}^{N_s \times N_s}$ for each text pair (d_a, d_b) according to the difference between original signals R_i and R_j:

$$H^{ab}_{ij} = \begin{cases} 1, & \text{if } R_i(d_a) >= R_j(d_b), \\ 0, & \text{if } R_i(d_a) < R_j(d_b). \end{cases} \tag{10}$$

Furthermore, for the predicted signals, we can build another signal interaction matrix $M^{ab} \in \mathbb{R}^{N_s \times N_s}$ for each text pair (d_a, d_b) according to the difference between predicted signal S_i and S_j:

$$M^{ab}_{ij} = \begin{cases} 1, & \text{if } S_i(d_a) >= S_j(d_b), \\ 0, & \text{if } S_i(d_a) < S_j(d_b). \end{cases} \tag{11}$$

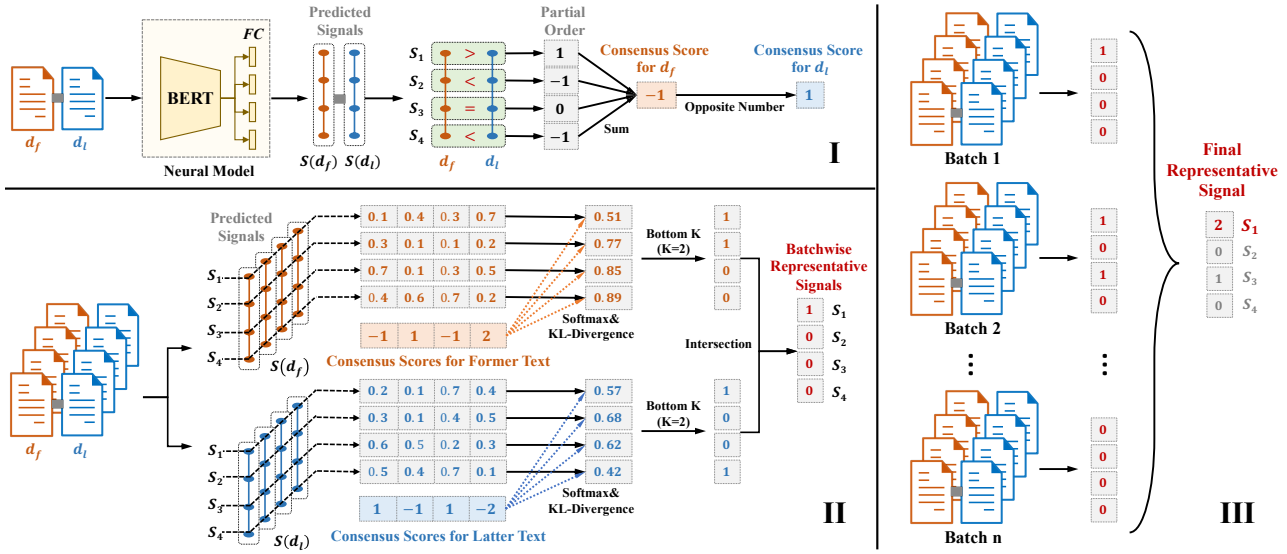

Figure 3: Illustration of Representative Signal Identification. I: Calculating consensus scores for a text pair. II: Selecting representative signals for a batch. III: Identifying the final representative signal.

Based on the above definition, we could have the negative log-likelihood function for the i-th and j-th signal prediction tasks:

$$\begin{aligned} & -log(P(M_{ij}^{ab} = H_{ij}^{ab} \mid \phi, \theta_i, \theta_j)) \\ = & -\sum_{q\in\{1,0\}} [H_{ij}^{ab} = q] log(P(M_{ij}^{ab} = q \mid \phi, \theta_i, \theta_j)) \\ = & -\sum_{q\in\{1,0\}} \Big([H_{ij}^{ab} = q] log(Q(S_i(d_a) - S_j(d_b))) \\ & + [H_{ij}^{ab} = q] log(1 - Q(S_i(d_a) - S_j(d_b))) \Big) \end{aligned} \tag{12}$$

By accumulating the negative log-likelihood function of N_s signal prediction tasks, we get the inter-signal pairwise loss:

$$\mathcal{L}_{inter} = -\sum_{i=1}^{N_s}\sum_{j=1}^{N_s} log(P(M_{ij}^{ab} = H_{ij}^{ab} \mid \phi, \theta_i, \theta_j)) \tag{13}$$

4.3.4 ***Overall Training Loss.*** Finally, we have the overall training loss of the multi-signal learning model by summing the losses of both intra-signal pairwise ranking and inter-signal joint ranking:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{intra} + (1-\alpha)\mathcal{L}_{inter}, \tag{14}$$

where α is a tradeoff parameter.

4.4 Representative Signal Identification

Given the lack of ground-truth readability labels to guide the selection of the optimal predicted signal, we develop an unsupervised strategy to identify the most representative signal among all predicted signals. To establish criteria for identifying the representative signal, we mine the pair-level consensus of signals for texts and leverage it to construct a batch-level consensus distribution. By measuring the divergence between the distribution of each predicted signal and the consensus distribution, we can identify the most similar signal as the final representative signal. This strategy allows us to obtain consensus of all signals while avoiding the negative influence of low-quality signals. Specifically, as shown in Figure 3, the strategy consists of three main steps: pairwise consensus mining, batchwise signal selection, and final signal identification. This strategy is conducted concurrently during the training of the multi-signal learning model.

4.4.1 ***Pairwise Consensus Mining.*** We suggest using the consensus of predicted signals to identify which predicted signal is most representative to indicate the readability of a text. To achieve this goal, we propose calculating the consensus scores for each pair of texts based on partial order. The underlying rationale is that for a given pair, if the majority of predicted signals for the former text are greater than their corresponding signals for the latter text, then we consider that the former text is likely to be difficult in terms of readability while the latter text is likely to be easy. Using the pair-level partial order to measure consensus is advantageous over using the average value of all predicted signals, as it eliminates the negative impact of inconsistent score distribution of signals.

Specifically, as shown in Figure 3-I, given a text pair (d_f, d_l) randomly sampled from $\boldsymbol{X}_0$, N_s predicted signals can be predicted by the neural model for each of d_f and d_l, where the i-th predicted signal of d_f and d_l is denoted as $S_i(d_f)$ and $S_i(d_l)$, respectively. Here, f and l denote the former and latter text in a pair, respectively. According to the difference of values between $S_i(d_f)$ and $S_i(d_l)$, we can calculate a predicted pair-level label $\hat{z}_{fl}^i \in \{1, -1, 0\}$ corresponding to the i-th predicted signals based on Eq. 4. By summing up N_s predicted pair-level labels, we can calculate two consensus scores $c_f = \sum_{i=1}^{N_s} \hat{z}_{fl}^i$ and $c_l = -c_f$ for the former and latter text to indicate their readability, respectively.

4.4.2 ***Batchwise Signals Selection.*** Based on the pair-level consensus scores, we can derive batch-level consensus distributions (i.e., distribution of consensus scores) as the criteria to identify the

representative signal. The underlying rationale posits that, for a given batch of text pairs $\{(d_f^i, d_l^i)\}_{i=1}^{N_b}$ with N_b pairs, we first construct two consensus distributions $\overline{D}_f$ and $\overline{D}_l$ corresponding to the former texts $\{d_f^i\}_{i=1}^{N_b}$ and latter texts $\{d_l^i\}_{i=1}^{N_b}$, respectively.

$$\begin{aligned} \overline{D}_f &= \text{Softmax}([c_f^1, \cdots, c_f^i, \cdots, c_f^{N_b}]) \\ \overline{D}_l &= \text{Softmax}([c_l^1, \cdots, c_l^i, \cdots, c_l^{N_b}]) \end{aligned} \tag{15}$$

where c_f^i and c_l^i are the consensus score of d_f^i and d_l^i, respectively.

Then, for each predicted signal S_j, we also respectively construct two distributions D_f^j and D_l^j:

$$\begin{aligned} D_f^j &= \text{Softmax}([S_j(d_f^1), \cdots, S_j(d_f^i), \cdots, S_j(d_f^{N_b})]) \\ D_l^j &= \text{Softmax}([S_j(d_l^1), \cdots, S_j(d_l^i), \cdots, S_j(d_l^{N_b})]) \end{aligned} \tag{16}$$

Finally, to select the batch-level representative signal, we measure the difference between the distributions of each predicted signal S_j and the consensus distributions based on the Kullback-Leibler divergence:

$$\begin{aligned} L_f &= \left[KL(\overline{D}_f||D_f^1), \cdots, KL(\overline{D}_f||D_f^j), \cdots, KL(\overline{D}_f||D_f^{N_s})\right] \\ L_l &= \left[KL(\overline{D}_l||D_l^1), \cdots, KL(\overline{D}_l||D_l^j), \cdots, KL(\overline{D}_l||D_l^{N_s})\right] \end{aligned} \tag{17}$$

Subsequently, by applying Bottom-K selection on L_f and L_l, we can get two sets of candidate representative signals, of which the intersection is regarded as the batch-level representative signals.

4.4.3 ***Final Signal Identification***. During the training process of the multi-signal learning model, we concurrently record all representative signals of each batch. Once the training is complete, we determine the final representative signal by selecting the predicted signal with the highest frequency of being the batch-level representative signal.

Specifically, we initialize a record vector, denoted as $m = [0]^{N_s}$, with N_s dimensions at the beginning of each training epoch. Each batch-level representative signal is recorded by incrementing its corresponding value in m. At the end of the epoch, the signal with the maximum value in m is identified as the final representative signal for that epoch.

5 EXPERIMENTS

In this section, we present the results of performance comparison conducted on three datasets, followed by the ablation study and model analysis, to verify the effectiveness of our proposed approach.

5.1 Datasets and Evaluation Metrics

We conduct experiments on three widely-used public datasets for evaluation, which are:

Table 1: Spearman's rank correlation coefficients between the selected signals and the ground-truth labels.

Signals	OSE	CAM	NSL
Syllable Count (SC)	.7349	.8250	.7462
Lexicon Count (LC)	.6853	.8008	.6940
Flesch Reading Ease (FRE) [14]	.5049(-)	.6784(-)	.8133(-)
Flesch-Kincaid Grade (FKG) [24]	.5810	.6942	.8966
SMOG Index (SMOG) [33]	.5019	.7044	.8605
Gunning Fog Index (GFI) [15]	.6432	.7471	.9105
Automated Readability Index (ARI) [39]	.5831	.6815	.9087
Dale-Chall Readability Score (DC) [8]	.5859	.6834	.8112
Average Sentences (AS)	.6824	.6583	.9005
Difficult Words (DW)	.8015	.8553	.8342
Characters Per Word (CPW)	.3428	.5569	.5648
Type Token Ratio (TTR)	.0659(0)	.5828(-)	.0827(0)
Characters (C)	.7332	.8203	.7372
Syllables Per Word (CPW)	.3894	.6173	.6373
Words (W)	.6852	.7971	.6933
Wordtypes (WT)	.7526	.8195	.7266
Long Words (LW)	.7561	.8356	.7817
Complex Words (CW)	.7365	.8607	.8109
Complex Words DC (CWD)	.7390	.8166	.7751
LIX (LIX) [4]	.6933	.8012	.7001
Coleman Liau (CL) [7]	.3490	.5754	.5697

- **OneStopEnglish (OSE)** [40] is a meticulously curated dataset developed for readability assessment and text simplification tasks. This dataset comprises 567 texts that have been rewritten from the original 189 texts, where each of the 189 texts is crafted into three texts corresponding to three levels of difficulty.
- **Cambridge (CAM)** [42] is a dataset collected from Cambridge English Exam and it is divided into 5 categories: KET, PET, FCE, CAE, CPE. Each category has more than 60 texts, resulting in a total of 326 texts.
- **Newsela (NSL)** [44] has 11 levels in total, covering grades 2 to 12. It has 1911 original English texts and 243 original Spanish texts. Each text has up to four simplified versions, resulting in a total of 9565 texts in English and 1221 texts in Spanish.

To evaluate the performance of methods, we conduct five-fold cross-validation, where the proportion of the training set, validation set, and test set is 3:1:1. For each time of the cross-validation, we measure the correlation between the predicted readability scores of all test texts and corresponding ground-truth labels based on three metrics: Spearman's Rank Correlation coefficient (SRC), Normalized Discounted Cumulative Gain (NDCG), Pearson Correlation Coefficient (PCC). The average results are reported.

5.2 Implementation Details

We extract a total of 21 heuristic readability signals for experiments based on the textstat package[1] and the readability package[2]. These signals include some commonly-used readability formulas and several easy-to-compute surface-level statistics. The Spearman's rank correlation coefficients between the ground-truth label and each extracted signal are shown in Table 1, indicating the quality of these signals. The signals with low or negative correlation with the ground-truth label are marked as (0) and (-), respectively.

[1]https://pypi.org/project/textstatll/
[2]https://pypi.org/project/readability/

Table 2: Performance of all comparison methods on three datasets. Italic and bold are the best performance of supervised and unsupervised methods, respectively. The *Best predicted signal* in LWRS is identified by ground-truth labels.

Settings	Methods		OSE			CAM			NSL		
			SRC	NDCG	PCC	SRC	NDCG	PCC	SRC	NDCG	PCC
Supervised	BERT-Large [11]		0.866	-	0.864	-	-	-	-	-	-
	BERT-Large-half [11]		0.751	-	0.747	-	-	-	-	-	-
	RoBERTa-RF-T1 [25] (Re-Implement)		***0.972***	0.992	***0.972***	***0.923***	0.948	***0.922***	***0.988***	0.998	***0.986***
	BERT-Base		0.9115	0.9356	0.8897	0.9180	***0.9598***	0.8755	0.9575	***0.9990***	0.9540
	TFIDF-SVM		0.8326	0.9827	0.8287	0.8311	0.9437	0.8432	0.9202	0.9643	0.8817
	TFIDF-MLP		0.8449	0.9616	0.8464	0.8023	0.9286	0.8063	0.8508	0.8964	0.8018
	TFIDF-Linear-Regression		0.9050	***0.9925***	0.8858	0.8678	0.9468	0.8454	0.9017	0.9702	0.8618
Unsupervised	RSRS [32]		-	-	0.615	-	-	-	-	-	0.894
	RSRS [32] (Re-Implement)		0.6710	0.9826	0.6597	0.7539	**0.9781**	0.7578	0.9318	**0.9941**	0.8925
	LURAT [11]		0.730	-	0.715	-	-	-	-	-	-
	LWRS (Ours)	**Identified representative signal**	**0.8886**	**0.9846**	**0.8810**	**0.8992**	0.9715	**0.8858**	**0.9405**	0.9708	**0.9318**
		Best predicted signal	0.8891	0.9848	0.8825	0.9037	0.9767	0.8918	0.9411	0.9741	0.9382
		Average metric of predicted signals	0.7105	0.9513	0.7021	0.8093	0.9427	0.7856	0.8947	0.9401	0.8732
	Baseline	**Mean of original signals**	0.5305	0.9427	0.5551	0.5989	0.8974	0.5789	0.6778	0.9303	0.6762
		Best original signal	0.8015	0.9562	0.7938	0.8553	0.9464	0.8488	0.9105	0.9602	0.8863
		Average metric of original signals	0.6304	0.9427	0.6232	0.7328	0.8974	0.7229	0.8376	0.9303	0.7806
		Best predicted signals using regression	0.7680	0.9552	0.7659	0.7958	0.9417	0.8734	0.9033	0.9478	0.8565

For our LWRS, the first ten signals in Table 1 are used for model training by default. For the mode architecture, we use the pre-training model bert-base-uncased[3] as an encoder. For the intra-signal pairwise ranking, we sample 2000 pairs for model training and set λ of each signal to a number that can make 40% of pairs fall into zero class. For the representative signal identification, we set K to be $\frac{1}{2}N_S$.

5.3 Comparison Methods

We compare LWRS with existing unsupervised methods and some supervised methods under our setting of five-fold cross-validation.

- **BERT-Large and BERT-Large-half** [11] are two supervised readability assessment methods based on the bert-large-cased-whole-word-masking pre-trained model, which correspond to using total and half of training set, respectively.
- **RoBERTa-RF-T1** [25] is a supervised readability assessment method based on RoBERTa pre-trained model and Radom Forest classifier. We re-implement it for comparison.
- **RSRS** [32] is an unsupervised readability assessment method based on the output perplexity of the pre-trained language model. We also re-implement it for comparison.
- **LURAT** [11] is an unsupervised readability assessment method based on vocabulary tests and estimated word difficulty.

Furthermore, we also implement many supervised baselines, such as methods based on TFIDF and three supervised models, and methods based on BERT-base. The variant methods related to original signals are used as unsupervised baselines.

[3]https://huggingface.co/BERT-base-uncased

5.4 Performance Comparison

As shown in Table 2, among all unsupervised methods, the best performance is mostly achieved by our LWRS. LWRS shows superior performance over RSRS and LURAT and achieves more stable performance than the re-implemented RSRS. This demonstrates the effectiveness of our LWRS framework.

By observing the variants of our method, we can find that the *Average Metric of Predicted Signals* can be improved over the *Average Metric of Original Signals* by 0.0801/0.0765/0.0571 in the SRC metric, on the OSE/CAM/NSL datasets, respectively. Besides, our *Identified Representative Signal* outperforms the *Best Original Signal* significantly, and can even achieve a comparable performance with the *Best Predicted Signal*. This indicates that our designed multi-signal learning and representative signal identification are effective.

By observing the supervised methods, we can find that LWRS also performs well. LWRS can perform better than the TFIDF-based methods in most cases. Besides, as to the fine-tuned pre-trained language model, LWRS can outperform BERT in terms of NDCG on OSE and CAM datasets. The good performance of LWRS is mainly attributed to the readability information brought by heuristic signals and our framework's adequate utilization of these signals.

5.5 Ablation Study

We explore the effects of the components designed in LWRS by removing or replacing each of them individually. Table 3 shows Spearman's rank correlation coefficients of the Identified representative signal, the Best predicted signal, and the Average metric of predicted signals.

For the multi-signal learning module, we study the effects of encoder, loss, and supervision. Firstly, regarding the structure of the encoder, we can find that the pre-trained language models generally

Table 3: Ablation study of LWRS. Spearman's rank correlation coefficients are reported.

Model Setting		OSE			CAM			NSL		
		Identified	Best	Average	Identified	Best	Average	Identified	Best	Average
LWRS		0.8886	0.8891	0.7105	0.8992	0.9037	0.8093	0.9405	0.9411	0.8947
Encoder	**TFIDF-MLP**	0.7937	0.8275	0.6914	0.8762	0.8762	0.8001	0.855	0.8666	0.8333
	BERT-tiny	0.8952	0.8952	0.7671	0.8003	0.8207	0.7343	0.9190	0.9228	0.9166
	BERT-mini	0.8855	0.8882	0.7179	0.8222	0.8402	0.7557	0.9381	0.9381	0.9143
	RoBERTa	0.8962	0.9100	0.7927	0.7821	0.7877	0.7532	0.9172	0.9177	0.9171
Loss	**MAE (replacing overall loss)**	0.7669	0.7770	0.6019	0.8836	0.8865	0.7582	0.7058	0.7277	0.5360
	No Loss (original signal as prediction)	0.8015	0.8015	0.6304	0.8473	0.8553	0.7328	0.8372	0.9105	0.8376
Supervision	**Rank Index (replacing signal value)**	0.7170	0.8224	0.6996	0.8210	0.8768	0.8042	0.9234	0.9298	0.9115
Bottom-K	**K=1**	0.8739	0.8891	-	0.8663	0.9037	-	0.8940	0.9411	-
	K=1/4 N_s	0.8735	0.8891	-	0.8987	0.9037	-	0.9399	0.9411	-
	K=3/4 N_s	0.8252	0.8891	-	0.8589	0.9037	-	0.9363	0.9411	-
Set Operation	**Union**	0.8850	0.8891	-	0.8662	0.9037	-	0.9411	0.9411	-
	No Operation (using two separate sets)	0.8748	0.8891	-	0.8589	0.9037	-	0.9117	0.9411	-

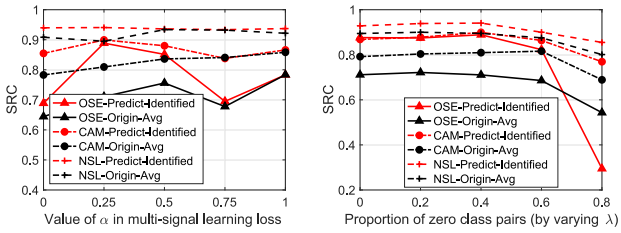

Figure 4: Effects of α and λ on multi-signal learning.

outperform TFIDF. However, there are exceptions, such as on the CAM dataset, where TFIDF performs better than other language models. Overall, our employed BERT base achieves the best and most stable performance and the size of the encoder is not the determining factor of the performance. Secondly, regarding the loss, we can find that replacing our overall loss with MAE loss or directly using original signals as a prediction without the loss function will lead to a decrease in performance. Besides, Figure 4 shows that both intra-signal loss and inter-signal loss are indispensable (left figure), and a proper proportion of zero-class pairs contributes to the improvement of performance (right figure). All of these phenomena indicate that the loss function we designed is effective. Thirdly, regarding the supervision, we can find that using signals' rank index instead of using signals' numerical value as supervision is less effective.

For the representative signal identification module, we study the effects of candidate selection and set operation. Firstly, regarding the candidate selection, we can find that selecting too many (i.e., 3/4 N_s) or too few (i.e., 1) signals as candidate signals do not achieve the best performance. The best performance is achieved by selecting approximately half of the signals (i.e., 1/2 N_s) as candidates. Secondly, regarding the set operation, we can find that other set operations might also perform well. *Union* means that the two sets selected from the former and latter parts are merged into one set based on the union operation, and *No Operation* means that no additional processing is done on sets corresponding to the former and latter parts of the pair. It can be found that taking both the union and intersection of the two parts leads to better results, but not when the two parts are recorded separately.

5.6 Model Analysis

In this part, we analyze the effect of each component of the model.

5.6.1 ***Effect of the quality of signals.*** For the signal learning part, we first analyze the effect of signal quality on signal learning by removing a signal from the first ten signals in Table 1 one at a time and measuring the average quality of the remaining signals before and after training. By observing Figure 5(a), we can find that *Predict-Identified* represented using the dashed line and *Origin-Avg* represented using the dotted line exhibit the same trend in most cases and have similar degrees of undulation. This indicates that the higher the average quality of the weak readability signals, the better the multi-signal learning and representative signal identification.

5.6.2 ***Effect of the number of signals.*** We then explore the effect of the signal number on the performance, using top 5, 10, 15, and 21 signals listed in Table 1. From Figure 5(b), we can find that the lines for *Predict-Identified* and *Predict-Best* almost overlap, showing a trend of firstly increasing and then stabilizing with the increase of signal number. This indicates that our representative signal identification strategy can almost always find the best predicted signal and is not significantly affected by signal number. Besides, we can observe that the lines for *Origin-Average* and *Predict-Average* both decrease when the signal number is 15, while the line for *Predict-Identified* does not decrease significantly. This implies that some noisy signals are introduced and lead to a decrease in the average quality of the signals. However, our representative signal identification strategy is still able to overcome the negative impact of noisy signals on signal selection, demonstrating its robustness.

5.6.3 ***Effect of batch size.*** Figure 5(c) shows the performance of the best predicted signal and the identified representative signal on three datasets. We set up the experiments with batch size (BS) of 2, 4, 8, 12, 16, and 20, where BS=2 indicates there are 2 pairs in each batch. It shows that the accuracy of identification is well for BS=2 (dichotomous case) and BS=8, whereas the training process

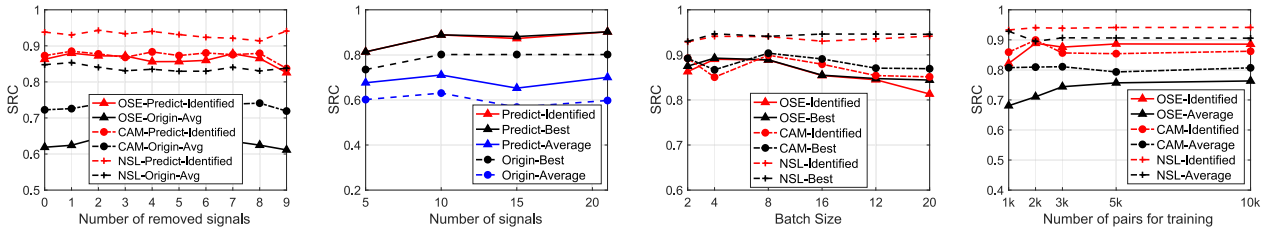

Figure 5: Effects of quality of signals, number of signals, batch size, and number of pairs on the performance of LWRS.

Table 4: Performance under cross-dataset transfer setting.

Method	Source Dataset	Target Dataset		
		OSE	CAM	NSL
LWRS	**OSE**	0.8886	0.8497	0.8680
	CAM	0.8447	0.8992	0.8751
	NSL	0.8638	0.8566	0.9405
TFIDF-MLP	**OSE**	0.8449	0.6534	0.6657
	CAM	0.8389	0.8023	0.7678
	NSL	0.7774	0.7315	0.8508
BERT [28] (Re-Implement)	**OSE**	0.9115	0.8194	0.7208
	CAM	0.7573	0.9180	0.7289
	NSL	0.8074	0.7419	0.9575

Table 5: Performance under cross-lingual transfer setting.

Method		NSL (En→En)		NSL (En→Es)	
		SRC	NDCG	SRC	NDCG
LWRS	**Identified**	0.9428	0.9771	0.9337	0.9692
	Best	0.9435	-	0.9357	-
	Average	0.9415	-	0.9331	-
NPRM [28]		-	-	0.985	0.996
mBERT [28]		-	-	0.957	0.992

takes longer time for BS=2. Though the constructed consensus distribution is often finer when the BS is bigger, the model does not work well in identifying the representative signal under the setting of big batch size. Instead, the rough distribution composed of small batch size is better for identifying the representative one.

5.6.4 ***Effect of number of pairs.*** Finally, we analyze the effect of the number of pairs on the performance. We use 1000, 2000, 3000, 5000, and 10000 pairs for the experiments. As shown in Figure 5(d), we can find that on both NSL and CAM datasets, both of the metrics show an increasing trend as the number of pairs increased. At more than 5000 pairs, the metrics of NSL and CAM datasets tend to be stable. The performance of the OSE data set fluctuates greatly with the number of pairs. Combined with the original signal quality, we believe that the reason is that the quality of the OSE signals varies greatly, and the quality of the signals themselves is also poorer than other two datasets. This indicates that the text of OSE varies more compared to other datasets and thus the bias in the results due to random sampling is large.

5.7 Experiments on Model Transferability

In this part, we investigate the transferability of the neural model trained by our LWRS.

5.7.1 ***Cross-dataset setting.*** Table 4 shows the cross-dataset performance of LWRS. For comparison, we also conduct experiments using TFIDF-MLP and BERT Re-Implement models trained under a supervised setting. We can find that the performance of BERT Re-Implement decreases by more than ten points in the cross-dataset experiment, which we believe is because the fine-tuned pre-trained model incorporates too much semantic information from the original dataset. The average decrease in the six migration metrics of LWRS is 0.0498, which indicates that LWRS has transferability and performs well.

5.7.2 ***Cross-lingual setting.*** Table 5 shows the cross-lingual performance of LWRS. We use multilingual BERT[4] (mBERT) as the encoder and compare our results with previous supervised methods. The pre-trained language model mBERT is trained using a multilingual corpus, including English and Spanish, which we used for training and testing, respectively. From Table 5, we can find that the LWRS based on mBERT achieves a correlation of 0.9337 with ground-truth labels under the cross-lingual setting, approaching some supervised methods. This indicates that LWRS performs well as an unsupervised method under the cross-lingual setting.

6 CONCLUSION

In this paper, we aim to perform readability assessment under an unsupervised setting. To this end, we propose the LRWS framework to train a neural model with the help of multiple weak readability signals. To enable the predicted signals to benefit from both raw text and original signals, we propose the multi-signal learning paradigm. Moreover, we propose to identify representative signal based on batch-level consensus distribution of signals. Experimental results demonstrate the effectiveness of the proposed LWRS framework for URA tasks, as well as its good transferability.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grants Nos. 61972192, 62172208, 61906085, 41972111. This work is partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization.

[4]https://huggingface.co/bert-base-multilingual-uncased

REFERENCES

[1] Hélder Antunes and Carla Teixeira Lopes. 2020. Proposal and Comparison of Health Specific Features for the Automatic Assessment of Readability. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1973–1976. https://doi.org/10.1145/3397271.3401187

[2] Oghenemaro Anuyah, Ion Madrazo Azpiazu, David McNeill, and Maria Soledad Pera. 2017. Can readability enhance recommendations on community question answering sites? *CEUR Workshop Proceedings* 1905 (2017). 2017 Poster Track of the 11th ACM Conference on Recommender Systems, Poster-Recsys 2017 ; Conference date: 28-08-2017 Through 28-08-2017.

[3] Ion Madrazo Azpiazu and Maria Soledad Pera. 2016. Is Readability a Valuable Signal for Hashtag Recommendations?. In *Proceedings of the Poster Track of the 10th ACM Conference on Recommender Systems (RecSys 2016), Boston, USA, September 17, 2016 (CEUR Workshop Proceedings, Vol. 1688)*, Ido Guy and Amit Sharma (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-1688/paper-21.pdf

[4] Richard Bamberger and Annette T. Rabin. 1984. New Approaches to Readability: Austrian Research. *The Reading Teacher, vol. 37, no. 6* (1984), 512–519.

[5] Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. MultiAzterTest: A Multilingual Analyzer on Multiple Levels of Language for Readability Assessment. arXiv:2109.04870 [cs]

[6] Tianyuan Cai, Ho Hung Lim, John S. Y. Lee, and Meichun Liu. 2022. Enhancing Automatic Readability Assessment with Verb Frame Features. In *International Conference on Asian Language Processing, IALP 2022, Singapore, October 27-28, 2022*, Rong Tong, Yanfeng Lu, Minghui Dong, Wengao Gong, and Haizhou Li (Eds.). IEEE, 413–418. https://doi.org/10.1109/IALP57159.2022.9961289

[7] Meri; Coleman and T. L Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology, Vol. 60, pp* (1975), 283–284.

[8] E. Dale and J.S Chall. 1948. A Formula for Predicting Readability. *Educational Research Bulletin* (1948), 37–54.

[9] Tovly Deutsch, Masoud Jasbi, and Stuart M. Shieber. 2020. Linguistic Features for Readability Assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2020, Online, July 10, 2020*, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch (Eds.). Association for Computational Linguistics, 1–17. https://doi.org/10.18653/v1/2020.bea-1.1

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs]

[11] Yo Ehara. 2021. LURAT: A Lightweight Unsupervised Automatic Readability Assessment Toolkit for Second Language Learners. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. 806–814. https://doi.org/10.1109/ICTAI52525.2021.00129

[12] Yo Ehara. 2022. Uncertainty-aware Personalized Readability Assessment Framework for Second Language Learners. *J. Inf. Process.* 30 (2022), 352–360. https://doi.org/10.2197/ipsjjip.30.352

[13] Mohamed El-Madkouri and Beatriz Soto Aranda. 2022. Readability and Communication in Machine Translation of Arabic Phraseologisms into Spanish. In *Computational and Corpus-Based Phraseology - 4th International Conference, Europhras 2022, Malaga, Spain, 28-30 September, 2022, Proceedings (Lecture Notes in Computer Science, Vol. 13528)*, Gloria Corpas Pastor and Ruslan Mitkov (Eds.). Springer, 78–89. https://doi.org/10.1007/978-3-031-15925-1_6

[14] R Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology 32(3* (1948), 221–233.

[15] Robert Gunning et al. 1952. Technique of clear writing. (1952).

[16] Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability Classification for German using Lexical, Syntactic, and Morphological Features. In *Proceedings of the 24th International Conference on Computational Linguistics*. 1063–1080.

[17] Michael Ibañez, Lloyd Lois Antonie Reyes, Ranz Sapinit, Mohammed Hussien, and Joseph Marvin Imperial. 2022. On Applicability of Neural Language Models for Readability Assessment in Filipino. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium - 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13356)*, Maria Mercedes T. Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova (Eds.). Springer, 573–576. https://doi.org/10.1007/978-3-031-11647-6_118

[18] Joseph Marvin Imperial. 2021. BERT Embeddings for Automatic Readability Assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3September, 2021*, Galia Angelova, Maria Kunilovskaya, Ruslan Mitkov, and Ivelina Nikolova-Koleva (Eds.). INCOMA Ltd., 611–618. https://aclanthology.org/2021.ranlp-1.69

[19] Shoaib Jameel and Xiaojun Qian. 2012. An Unsupervised Technical Readability Ranking Model by Building a Conceptual Terrain in LSI. In *2012 Eighth International Conference on Semantics, Knowledge and Grids*. 39–46. https://doi.org/10.1109/SKG.2012.20

[20] Shoaib Jameel, Xiaojun Qian, and Wai Lam. 2012. N-Gram Fragment Sequence Based Unsupervised Domain-Specific Document Readability. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, 1309–1326.

[21] 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, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, 366–378. https://aclanthology.org/C18-1031/

[22] Zhiwei Jiang, Qing Gu, Yafeng Yin, Jianxiang Wang, and Daoxu Chen. 2019. GRAW+: A two-view graph propagation method with word coupling for readability assessment. *J. Assoc. Inf. Sci. Technol.* 70, 5 (2019), 433–447. https://doi.org/10.1002/asi.24123

[23] Zhiwei Jiang, Gang Sun, Qing Gu, Tao Bai, and Daoxu Chen. 2015. A Graph-based Readability Assessment Method using Word Coupling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 411–420. https://doi.org/10.18653/v1/d15-1047

[24] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 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.

[25] Bruce W. Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. 2021. Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 10669–10686. https://doi.org/10.18653/v1/2021.emnlp-main.834

[26] Bruce W. Lee and Jason Hyung-Jong Lee. 2020. LXPER Index 2.0: Improving Text Readability Assessment for L2 English Learners in South Korea. *CoRR* abs/2010.13374 (2020). arXiv:2010.13374 https://arxiv.org/abs/2010.13374

[27] Justin Lee and Sowmya Vajjala. 2022. A Neural Pairwise Ranking Model for Readability Assessment. *arXiv:2203.07450 [cs]* (March 2022). arXiv:2203.07450 [cs]

[28] Justin Lee and Sowmya Vajjala. 2022. A Neural Pairwise Ranking Model for Readability Assessment. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 3802–3813. https://doi.org/10.18653/v1/2022.findings-acl.300

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs]

[30] Bertha A Lively and Sidney L Pressey. 1923. A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision* 9, 7 (1923), 389–398.

[31] Ion Madrazo Azpiazu and Maria Soledad Pera. 2020. An Analysis of Transfer Learning Methods for Multilingual Readability Assessment. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 95–100. https://doi.org/10.1145/3386392.3397605

[32] Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics* 47, 1 (April 2021), 141–179. https://doi.org/10.1162/coli_a_00398

[33] G.H McLaughlin. 1969. SMOG grading-a new readability formula. *Journal of Reading, 12(8)* (1969), 639–646.

[34] Hamid Mohammadi and Seyed Hossein Khasteh. 2019. Text as Environment: A Deep Reinforcement Learning Text Readability Assessment Model. *CoRR* abs/1912.05957 (2019). arXiv:1912.05957 http://arxiv.org/abs/1912.05957

[35] Neil Newbold, Harry McLaughlin, and Lee Gillam. 2010. Rank by Readability: Document Weighting for Information Retrieval. In *Advances in Multidisciplinary Retrieval*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Hamish Cunningham, Allan Hanbury, and Stefan Rüger (Eds.). Vol. 6107. Springer Berlin Heidelberg, Berlin, Heidelberg, 20–30. https://doi.org/10.1007/978-3-642-13084-7_3

[36] Florian Pickelmann, Michael Färber, and Adam Jatowt. 2023. Ablesbarkeitsmesser: A System for Assessing the Readability of German Text. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 13982)*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, 288–293. https://doi.org/10.1007/978-3-031-28241-6_28

[37] 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. In *Proceedings of the 59th Annual Meeting*

of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, 3013–3025. https://doi.org/10.18653/v1/2021.acl-long.235

[38] Andreas Schlapbach, Frank Wettstein, and Horst Bunke. 2008. Estimating the readability of handwritten text - a Support Vector Regression based approach. In *19th International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA*. IEEE Computer Society, 1–4. https://doi.org/10.1109/ICPR.2008.4761907

[39] E A Smith and R J Senter. 1967. Automated Readability Index. *AMRL-TR. Aerospace Medical Research Laboratories (U.S.)* (1967), 1–14.

[40] Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*. 297–304.

[41] Tim vor der Brück. 2009. Approximation of the Parameters of a Readability Formula by Robust Regression. In *Machine Learning and Data Mining in Pattern Recognition, 6th International Conference, MLDM 2009, Leipzig, Germany, July 2009, Poster Proceedings*, Petra Perner (Ed.). ibai Publishing, 115–125.

[42] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, San Diego, CA, 12–22. https://doi.org/10.18653/v1/W16-0502

[43] Haoran Xie, Minhong Wang, Di Zou, and Fu Lee Wang. 2019. A Personalized Task Recommendation System for Vocabulary Learning Based on Readability and Diversity. In *Blended Learning: Educational Innovation for Personalized Learning*, Simon K. S. Cheung, Lap-Kei Lee, Ivana Simonova, Tomas Kozel, and Lam-For Kwok (Eds.). Vol. 11546. Springer International Publishing, Cham, 82–92. https://doi.org/10.1007/978-3-030-21562-0_7

[44] Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics* 3 (05 2015), 283–297. https://doi.org/10.1162/tacl_a_00139 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00139/1566780/tacl_a_00139.pdf