A Neural Pairwise Ranking Model for Readability Assessment

Justin Lee*
University of Toronto
CSA Group
Toronto, Canada
chunhin.lee@mail.utoronto.ca

Sowmya Vajjala
National Research Council
Ottawa, Canada
sowmya.vajjala@nrc-cnrc.gc.ca

Abstract

Automatic Readability Assessment (ARA), the task of assigning a reading level to a text, is traditionally treated as a classification problem in NLP research. In this paper, we propose the first neural, pairwise ranking approach to ARA and compare it with existing classification, regression, and (non-neural) ranking methods. We establish the performance of our model by conducting experiments with three English, one French and one Spanish datasets. We demonstrate that our approach performs well in monolingual single/cross corpus testing scenarios and achieves a zero-shot cross-lingual ranking accuracy of over 80% for both French and Spanish when trained on English data. Additionally, we also release a new parallel bilingual readability dataset in English and French. To our knowledge, this paper proposes the first neural pairwise ranking model for ARA, and shows the first results of cross-lingual, zero-shot evaluation of ARA with neural models.

1 Introduction

Automatic Readability Assessment is the task of assigning a reading level for a given text. It is useful in many applications from selecting age appropriate texts in classrooms (Sheehan et al., 2014) to assessment of patient education materials (Sare et al., 2020) and clinical informed consent forms (Perni et al., 2019). Contemporary NLP approaches treat it primarily as a classification problem which makes it non-transferable to situations where the reading level scale is different from the model. Applying learning to rank methods has been seen as a potential solution to this problem in the past. Ranking texts by readability is also useful in a range of application scenarios, from ranking search results based on readability (Kim et al., 2012; Fourney et al., 2018) to controlling the reading level of machine translation output (Agrawal and Carpuat, 2019; Marchisio et al., 2019). However, exploration of ranking methods has not been a prominent direction for ARA research. Further, recent developments in neural ranking approaches haven't been explored for this task yet, to our knowledge.

ARA typically relies on the presence of large amounts of data labeled by reading level. Further, although linguistic features are common in ARA research, it is challenging to calculate them for several languages, due to lack of available software support. Though there is a lot of recent interest in neural network based cross-lingual transfer learning approaches for various NLP tasks, there hasn't been much research in this direction for ARA yet.

With this context, we address two research questions in this paper:

1. Is neural, pairwise ranking a better approach than classification or regression for ARA?
2. Is zero-shot, cross-lingual transfer possible for ARA models through ranking?

The main contributions of this paper are:

1. A new neural pairwise ranking model with an application to automatic readability assessment.
2. Demonstration of the use pairwise ranking to achieve cross-corpus compatibility in ARA.
3. First evidence of zero shot, neural cross-lingual transfer in ARA.
4. A new parallel readability dataset, *Vikidia En/Fr*, the first of its kind in ARA research.

The rest of this paper is organized as follows: Section 2 gives an overview of related research and Section 3 describes the proposed neural pairwise ranking model. The next two Sections (4 5) describe our experimental setup and discuss the results. Section 6 summarizes our findings and discusses the limitations of this approach.

*Work done during an internship at National Research Council, Canada

Findings of the Association for Computational Linguistics: ACL 2022, pages 3802 - 3813

2 Related Work

Readability Assessment has been an active area in educational research for almost a century. Early research on this topic focused on the creation of readability "formulae", which relied on easy to calculate measures such as word and sentence length, and presence of words from some standard word list (Lively and Pressey, 1923; Flesch, 1948; Stenner, 1996). More than 200 such formulae were proposed in the past few decades DuBay (2007). The advent of NLP and machine learning resulted in more data driven research on ARA over the past two decades. Starting from statistical language models (Si and Callan, 2001), a range of lexical and syntactic features (Heilman et al., 2007; Petersen and Ostendorf, 2009; Ambati et al., 2016) as well as inter-sentential features (Pitler and Nenkova, 2008; Todirascu et al., 2013; Xia et al., 2016) were developed in the past. Features motivated by related disciplines such as psycholinguistics (Howcroft and Demberg, 2017), second language acquisition (Vajjala and Meurers, 2012) and cognitive science (Feng et al., 2009) were also explored for this task.

In the past few years, ARA research has been primarily focused on textual embeddings and deep learning based architectures. Word embeddings in combination with other attributes such as domain knowledge or language modeling (Cha et al., 2017; Jiang et al., 2018) and a range of neural architectures, from multi attentive RNN (Azpiazu and Pera, 2019) to deep reinforcement learning (Mohammadi and Khasteh, 2019) were proposed. Recent research explored combining transformers with linguistic features (Deutsch et al., 2020; Meng et al., 2020; Lee et al., 2021; Imperial, 2021).

Although a lot of this research evolved on English, the past decade saw ARA research in other languages such as German (Hancke et al., 2012), French (François and Fairon, 2012), Italian (Dell'Orletta et al., 2011), Bangla (Sinha et al., 2012) etc., which employed language specific feature sets. While most ARA research modeled one language at a time, some research created language agnostic feature sets and architectures and experimented with 2 to 7 languages (Shen et al., 2013; Azpiazu and Pera, 2019; Madrazo Azpiazu and Pera, 2020a,b; Martinc et al., 2021; Weiss et al., 2021). Although only one language is considered per model in all this research, there are two important exceptions. Madrazo Azpiazu and Pera (2020b) explored whether combining texts from related languages during training improves ARA performance for low resource languages. Weiss et al. (2021) used a model trained on English texts on German, based on a common, broad set of handcrafted linguistic features. However, to our knowledge, zero-shot cross lingual transfer of neural network architecture based approaches, without any handcrafted features, was not explored for this task in the past.

ARA is traditionally treated as a classification problem in NLP research, although there are some exceptions. Heilman et al. (2008) compared linear, ordinal, and logistic regression and concluded that ordinal regression with a combination of lexical and grammatical features worked the best for ARA, although classification approaches still dominated subsequent research on the topic. There is some past work that considered ARA as a pairwise ranking problem, using SVM/SVM^{rank} and hand crafted linguistic features (Pitler and Nenkova, 2008; Tanaka-Ishii et al., 2010; Ma et al., 2012; Mesgar and Strube, 2015; Ambati et al., 2016; Howcroft and Demberg, 2017). While Tanaka-Ishii et al. (2010) and Ma et al. (2012) showed that ranking performs better than traditional features and classification/regression respectively, Xia et al. (2016) did not find ranking to be consistently better across the board. Given this background, we take a fresh look at the application of ranking for ARA, by proposing a new neural pairwise ranking model.

3 Neural Pairwise Ranking Model

The data for our pairwise ranking model takes the form of (document, reading level) pairs. Let $X = [(x_1, y_1), ..., (x_n, y_n)]$ be n such pairs, where x_i is the vector representation for document i and y_i is the corresponding reading level. We then construct m pairwise permutations from X to form X'. The members of X' are constructed as follows: if a pair of documents and reading levels (x_i, y_i) and (x_j, y_j) are chosen, then both permutations $((x_i, x_j), (y_i, y_j))$ and $((x_j, x_i), (y_j, y_i))$ are added to X'.

The neural pairwise ranking model ($NPRM$) aims to maximize

$$P(y_i > y_j | x_i, x_j)$$

Formally, this is parametrized as

$$P(y_i > y_j|x_i, x_j) \triangleq NPRM(x_i, x_j)$$
$$= softmax(\psi(f(x_i, x_j)))$$
$$= [s_{ij1}, s_{ij2}]$$

where f is a neural model, ψ is a flexible function, s_{ij1} represents the predicted score of $P(y_i > y_j|x_i, x_j)$ and s_{ij2} represents the predicted score of 1 - $P(y_i > y_j|x_i, x_j)$. Training labels are created as

$$y'_{ij} = \begin{cases} [1, 0] & \text{if } y_i \geq y_j \\ [0, 1] & \text{if } y_i < y_j \end{cases}$$

We then calculate the loss function as

$$L = -y'_{ij1} \cdot log(s_{ijk1}) - y'_{ij2} \cdot log(s_{ijk2})$$

and back-propagate the errors with stochastic gradient descent. This loss function is known as the Pairwise Logistic Loss (Han et al., 2020).

Implementation Our neural pairwise ranking model ($NPRM$) consists of a BERT (Devlin et al., 2018) model as f and a fully connected layer as ψ. We evaluate the performance of the pairwise ranking approach as follows: for a list of texts to be ranked of size S and each text x_a within the list, $1 \leq a \leq S$ we compute

$$Score(x_a) = \sum_{b \neq a} NPRM(x_a, x_b)$$

We then rank each text x_a by $Score(x_a)$ in descending order.

This pairwise ranking framework allows for $NPRM$ to model relative reading difficulties between texts. While other neural methods have been proposed with more sophisticated learning objectives for ranking problems in the past (Wang et al., 2018; Ai et al., 2019), these methods require fixed-size inputs to rank. The $NPRM$ only needs a minimum of two texts to form a ranking for each, and the aggregation process of scores between pairwise permutations of texts can easily produce rankings for an arbitrary list size larger than two. The aggregation process also produces a bounded (by the list size), but continuous score for each document, which results in a ranking with no ties, as long as the input documents are different.

Additionally, the choice of f in the $NPRM$ framework can allow for multi-lingual predictability, improved performance, or improved efficiency. Due to the flexible modeling structure that the $NPRM$ maintains, we hypothesize that with the aid of a multilingual language model, zero-shot cross lingual ARA assessment may also be possible with $NPRM$. We demonstrate this possibility later in the paper[1].

4 Experimental Setup

We describe our experimental setup in terms of the datasets used, modeling and evaluation procedures, in this section.

4.1 Datasets

We experimented with three English, one Spanish and one French datasets, which are described below. All the datasets contain texts in multiple reading level versions (in a given language). We call such grouping of a given text in multiple reading levels a *slug*. We used the first two English datasets for training and testing our models, and the remaining three datasets only as test sets.

NewsEla-English (NewsEla-En): NewsEla[2] provides leveled reading content, which is aligned with the common core educational standards (Porter et al., 2011), and contains texts covering grade 2 to grade 12. It follows the Lexile (Stenner, 1996) framework to create such leveled texts. It was first used in NLP research by Xu et al. (2015) and has been a commonly used corpus for ARA and text simplification in the recent past. The English subset of the NewsEla dataset contains 9565 texts distributed across 1911 slugs. Slugs may or may not contain texts for the full range of reading levels available i.e., each text does not have all reading level versions.

OneStopEnglish (OSE): This consists of articles sourced from The Guardian newspaper, rewritten by teachers into three reading levels (beginner, intermediate, advanced) (Vajjala and Lučić, 2018) and has been used as a bench marking dataset for ARA in the recent past. This dataset contains 189 slugs and 3 reading levels, summing to a total of 567 texts (each slug has one text in three versions).

NewsEla-Spanish (NewsEla-Es): This is the Spanish subset within the existing NewsEla dataset and contains 1221 texts distributed across 243 slugs

[1] All code for the model and experiments is at: https://github.com/jlee118/NPRM/

[2] NewsEla corpus can be requested from: https://NewsEla.com/data/

and 10 reading levels. Similar to NewsEla-En, each slug does not have all 10 levels in it.

Vikidia-En/Fr: Vikidia.org[3] is a children's encyclopedia, with content targeting 8-13 year old children, in several European languages. Our dataset contains 24660 texts distributed across 6165 slugs and 2 reading levels, for English (*Vikidia-En*) and French (*Vikidia-Fr*) respectively i.e., each text in the corpus has four versions: en, en-simple, fr and fr-simple, and there are 6165 slugs in total. Azpiazu and Pera (2019)'s experiments used data from this source. However, the data itself is not publicly available. The uniqueness of the current dataset is that these are parallel, document level aligned texts in four versions - en, en-simple, fr, fr-simple. While we did not create paragraph/sentence level alignments on the corpus, we hope that this will be a useful dataset for future English and French research on ARA and Automatic Text Simplification. This is the first such dataset in ARA, and perhaps the first readily available French readability dataset. It can be accessed at: https://zenodo.org/record/6327828

4.2 Classification, Regression and Ranking models

Our primary focus in this paper is on the pairwise ranking model. However, we also compared the performance of other classification, regression, and ranking approaches with our pairwise ranking model to establish strong points of comparison.

Feature representation: While the use of linguistic features, and more recently, contexual embeddings, have been explored in ARA, non-contextual embeddings were not explored much. Hence, in this paper, we employ three non-contextual embeddings (GloVe (Pennington et al., 2014), Word2vec (Mikolov et al., 2013a), fastText (Bojanowski et al., 2017)) for training classification/regression/ranking models. Document-level embeddings are obtained by aggregating and averaging word-level embeddings for each token in the text. In addition, we also used a BERT (Devlin et al., 2018) based classifier.

Classification The following models were used for formulating baselines and comparisons for classification. Reading levels are treated as class labels, and evaluation is done via 5-Fold cross validation.

- Non-contextual embeddings fed into an SVM (Boser et al., 1992) classifier
- Non-contextual embeddings fed into a Hierarchical Attention Network (HAN) (Yang et al., 2016). This model was used with and without linguistic features in the past, for reading level classification (Deutsch et al. (2020) and Martinc et al. (2021) respectively).
- 110-M parameter, 12-layer, BERT model with a fully connected layer and a softmax output. The model is then fine-tuned on the classification task with categorical cross-entropy loss.

Regression The following models were used for formulating baselines for regression. Reading levels are treated as continuous outputs, and results are obtained through 5-Fold cross validation.

- Non-contextual word-level embeddings as input into an Ordinary Linear Regression (OLS) model.
- 110-M parameter BERT model with a fully connected layer. The model is then fine-tuned on the regression task with the mean squared error loss and will be referred to as *regBERT* in this paper.

(non-neural) Pairwise Ranking We employ an SVMRank model with a pairwise ranking framework similar to $NPRM$, but using the non-contextual word embeddings for feature extraction. Input features for the SVM model are obtained by differencing the obtained embeddings in the following manner: for any text representations x_i, x_j, with reading levels y_i, y_j, form training examples as $x'_i = x_i - x_j$ and $x'_j = x_j - x_i$, and training labels as:

$$y'_i = \begin{cases} 1 & y_i \geq y_j \\ 0 & y_i < y_j \end{cases}$$

$$y'_j = \begin{cases} 1 & y_j \geq y_i \\ 0 & y_j < y_i \end{cases}$$

Predicted scores are aggregated in the same manner as in $NPRM$ to form rankings. Results are obtained through 5-Fold cross validation.

4.3 Pairwise Ranking Training

To control for the variation in the text introduced by different topical content, the training and prediction process for the SVMRank and $NPRM$ aggregates the text by their slug designations before forming

[3] https://www.vikidia.org/

pairwise permutations. As a result, the pairwise permutations are constructed from the text within a slug. Note that slug is used for training and testing the model, but isn't really required while using the model for prediction. The trained model only takes a list of texts as inputs and returns a ranked list based on readability.

For controlling the computation time, we fixed the number of pairwise comparisons per slug (m in NPRM) to 3 levels. i.e., In datasets with more than 3 levels per slug (NewsEla-En and NewsEla-Es) we choose texts with the highest and lowest reading levels within a slug, and sample the third text from a reading level in between. Note that this will not affect the ability of the model to rank a list of texts where m is higher than 3. As with all baselines, results from $NPRM$ are obtained through 5-Fold cross validation.

4.4 Evaluation

Accuracy and F1-score are reported for classification and mean-absolute error (MAE) and mean-squared error (MSE) are reported for regression. To evaluate ranking performance, we calculate the Normalized Discounted Cumulative Gain (NDCG), Spearman's Rank Correlation(SRC), Kendall's Tau Correlation Coefficient (KTCC), and the percentage of slugs ranked completely correct, which we denote as Ranking Accuracy (RA). There is some work on evaluating ranking in NLP (Lapata, 2006; Katerenchuk and Rosenberg, 2016), without any consensus on the most suited metric. Hence, we chose to report multiple metrics instead of one, based on the commonly reported measures for such tasks.

We compare classification and regression predictions too using ranking metrics, in addition to traditional measures. To examine the ranking performance, the texts from each dataset are first grouped by their slugs. Then, ground-truth ranking of the texts within the slugs are compared against the rankings formed from the predicted scores of the models. For NDCG, we used the ground-truth reading levels as the relevance score. For all the metrics, We took the model predictions as is, and did not employ specific means to address ties (which can happen in classification). The metrics themselves address ties in different ways. NDCG averages ties in predicted scores, KTCC penalizes ties in ground truth and predicted scores, and SRC calculates the average rank of ties. Ranking accuracy does not handle ties.

4.5 Statistical Significance Testing

We used Wilcoxon's signed rank test (Conover, 1999), a non-parametric statistical hypothesis test to examine whether the performance differences between $NPRM$ and other methods are statistically significant, when the metrics are close to each other. Ranking metrics per slug from a sample per model are aggregated, and are then compared for any two models.

4.6 Technical Implementation

Non-neural machine methods used the sklearn (Pedregosa et al., 2011) library. The HAN model is a Keras implementation[4]. Transformers library (Wolf et al., 2020) was used for accessing and fine-tuning BERT and mBERT based models (*bert-base-uncased* and *bert-base-multilingual-uncased* models were used). TF-Ranking library [5] (Pasumarthi et al., 2019) was used for accessing the Keras-compatible Pairwise Logistic Loss function. SciPy(Virtanen et al., 2020) was used for statistical significance testing.

The Word2vec embeddings are pre-trained on English Google News (Mikolov et al., 2013b). The fasttext embeddings contain 1-million word vectors and are trained on subword information from Wikipedia 2017 (Bojanowski et al., 2017). The GloVe embeddings are trained on the Wikipedia 2014 and Gigaword 5 corpus (Pennington et al., 2014). All three are accessed through gensim[6].

5 Results

We performed within corpus evaluation for classification, and within/cross corpus as well as cross-lingual evaluation for regression and ranking. We did not employ classification approaches in the last two evaluation settings as there is no way of resolving ties with classifier predictions. Further, regression and ranking gave better performance than classification in monolingual, within-corpus settings.

5.1 Classification

We trained models using *Newsela-En* and *OSE* datasets respectively in a five fold CV setup, for classification. Table 1 shows the performance of

[4] https://github.com/tomcataa/HAN-keras
[5] https://github.com/tensorflow/ranking
[6] https://radimrehurek.com/gensim/

our best performing model in terms of traditional classification metrics, comparing with the state of the art.

Model	weighted-F1
NewsEla-En	
HAN (Martinc et al., 2021)	**0.81**
BERT	0.74
OSE	
HAN (Martinc et al., 2021)	0.79
BERT	0.93
BART (Lewis et al., 2020)+Linguistic features (Lee et al., 2021)	**0.97**

Table 1: Weighted-F1 for classification

In terms of traditional classification metrics, our approach achieves a lower performance than Martinc et al. (2021) for NewsEla-En corpus, but higher performance on the OSE corpus. A more recent paper by Lee et al. (2021) reported further improvement with OSE, with an extensive set of linguistic features. Table 2 shows the performance of all models in terms of the ranking metrics.

Model	NDCG	SRC	KTCC	RA
NewsEla-En				
BERT	**0.999**	**0.992**	**0.985**	0.927
GloVe + HAN	0.991	0.985	0.971	0.971
GloVe + SVM	0.947	0.866	0.796	0.981
fasttext + HAN	0.991	0.985	0.972	0.971
fasttext + SVM	0.996	0.939	0.892	**0.990**
OSE				
BERT	0.963	0.808	0.808	0.825
GloVe + HAN	0.938	0.741	0.741	0.841
Glove + SVM	0.875	**0.931**	**0.930**	**0.963**
fasttext + HAN	**0.964**	0.857	0.854	0.899
fasttext + SVM	0.867	0.763	0.763	0.884

Table 2: Ranking Metrics for Classification Evaluation

When evaluating the classification models in terms of ranking metrics, we notice some differences among the models evaluated using NewsEla-En and OSE. There is relatively less variation among different Newsela-En models for NDCG, compared to SRC, KTCC, and RA. We see larger variations across OSE models for all the metrics. It is interesting to note that the non-contextual embeddings perform competitively with BERT in terms of the ranking metrics and are all better than BERT in terms of ranking accuracy. Overall, The NewsEla-En + BERT classifier achieves the highest average NDCG, SRC, and KTCC, and the NewsEla + fasttext + SVM combination achieves the highest ranking accuracy. For OSE, the Glove+SVM classifier achieves the highest SRC, KTCC, and RA while fastText+HAN and BERT models achieve better scores in terms of NDCG.

All the ranking metrics in general seem to have higher scores with NewsEla-En trained models, than OSE models. This could potentially be due to the larger dataset size, as well as the fact that NewsEla-En covers a broader reading level scale. Although the classification models seem to generally perform well on ranking metrics too, it has to be noted that there is no inherent means within classification to distinguish between ties, where the model predicts the same class for two documents of different reading levels. Hence, it is not feasible to continue to use classifiers as rankers. This evaluation is to be seen only means of comparing classification, regression, and ranking with a common set of metrics.

5.2 Regression

Table 3 shows the performance of all the regression models using both regression and ranking metrics.

Although there are no other reported results of applying regression models on these datasets to our knowledge, the low MAE/MSE for both datasets indicate that regression models perform well for this problem. Like with classification, we notice that there is no huge difference among the contextual and non-contextual embeddings in terms of the ranking metrics. However, we notice some general differences between classification and regression approaches. In contrast to the classification models, when holding the training data and the regressor constant, models with GloVe embeddings perform worse than models using Word2vec or fasttext in regression specific metrics. When evaluating on ranking metrics, the regression models generally exhibit higher average NDCG, SRC and KTCC than the classification models. Again, like with classification evaluation, the differences across models in terms of the ranking metrics is larger for OSE compared to NewsEla-En. Overall, though, the neural regressor (regBERT) consistently performs better than the OLS regressor in terms of regression metrics, and is either comparable or better than OLS regressor in terms of all the ranking metrics.

Model	MSE	MAE	NDCG	SRC	KTCC	RA
NewsEla-En						
regBERT	**0.434**	**0.460**	0.999	0.997	0.994	**0.977**
gloVe+OLS	2.310	1.212	0.999	0.988	0.978	0.900
word2Vec+OLS	1.734	1.056	0.999	0.996	0.992	0.961
fasttext + OLS	1.766	1.058	0.999	0.997	0.994	0.971
OSE						
regBERT	**0.260**	**0.376**	0.986	**0.944**	**0.929**	**0.905**
gloVe + OLS	2.143	1.122	0.989	0.857	0.834	0.794
word2vec + OLS	1.888	1.076	0.988	0.873	0.855	0.852
fasttext + OLS	1.561	0.953	**0.995**	0.926	0.912	0.899

Table 3: Performance of Regression approaches

Model	Avg. NDCG	Avg. SRC	Avg. KTCC	RA
NewsEla-En				
NPRM BERT	**0.999**	0.995	0.990	0.948
word2vec + SVMRank	0.997	**0.997**	**0.997**	**0.979**
fasttext + SVM-Rank	0.998	0.995	0.991	0.957
GloVe + SVM-Rank	0.998	0.992	0.985	0.932
OSE				
NPRM BERT	**0.997**	**0.981**	**0.979**	**0.979**
word2vec + SVMRank	0.972	0.966	0.962	0.958
fasttext + SVM-Rank	0.991	0.947	0.940	0.931
GloVe + SVM-Rank	0.994	0.971	0.968	0.968

Table 4: Pairwise Ranking Evaluation

5.3 Pairwise Ranking

Table 4 shows the performance of pairwise ranking approaches on both the training datasets. When training on the NewsEla-En dataset, we observe that $NPRM$ outperforms at least one word-embedding + SVMRank combination in the ranking metrics, but only achieves the top score in NDCG when compared with word-embedding SVMRank methods. When training on the OSE dataset, $NPRM$ achieves the top score against the word-embedding + SVMRank combinations, but only NDCG was found to be statistically significant across all models. Comparisons between $NPRM$ and the word-embedding + SVMRank combinations had p-values < 0.05 for NDCG. For SRR, KTCC and RA, only the difference in scores between $NPRM$ and fasttext + SVMRank were found to be statistically significant. GloVe + SVM-Rank method produces the statistically equivalent scores in SRC, KTCC, and RA as $NPRM$.

Overall, while there is no single approach that ranked as the best uniformly across all the three model settings (Tables 2- 4), BERT based models perform competitively with most of the ranking metrics. Table 5 presents a summary of the performance of BERT in classification, regression and ranking setups.

Model	Avg. NDCG	Avg. SRC	Avg. KTCC	RA
NewsEla-En				
BERT-Class.	0.999	0.992	0.985	0.927
regBERT	0.999	0.997	0.994	0.977
NPRM BERT	0.999	0.995	0.990	0.948
OSE				
BERT-Class.	0.963	0.808	0.808	0.825
regBERT	0.986	0.944	0.929	0.905
NPRM BERT	0.997	0.981	0.979	0.979

Table 5: Classification vs Regression vs Ranking

For Newsela-En, all methods reported a high score of 0.999 for NDCG and regBERT is better with the other metrics. Testing for statistical significance between $NPRM$, regBERT and BERT classification showed that $NPRM$ is significantly better than BERT classifier ($p<0.05$) and there is no significant difference between $NPRM$ and regBERT. For OSE, NPRM BERT achieves a better performance for all metrics. We did not perform statistical significance testing in this case as the differences are larger.

To conclude, when training and testing from the same distribution, regBERT and NPRM BERT per-

form better than BERT based classifier in terms of the ranking metrics. Since the performance is generally expected to degrade slightly in a cross-corpus setting compared to a within corpus evaluation, the rest of our experiments will only focus on regBERT and NPRM, and we don't report further experiments with a BERT classifier.

5.4 Cross-corpus Pair-wise Ranking

In this experiment, we evaluated the performance of an ARA model trained with one English dataset, on other English datasets. Since NewsEla-en is the larger dataset with more diverse reading levels, we used that for training, and used OSE and Vikidia-En as test sets. Since regression scores can also be used to directly rank predictions, we compared the performance of NPRM with BERT based regression model. Table 6 shows the results.

NPRM				
Testset	NDCG	SRC	KTCC	RA
OSE	0.983	0.931	0.912	0.878
Vikidia-En	0.991	0.950	0.950	0.975
regBERT				
OSE	0.929	0.706	0.651	0.561
Vikidia-En	0.982	0.904	0.904	0.952

Table 6: Cross-Corpus Pairwise Ranking (Trained on Newsela-En)

NPRM model, trained on Newsela-En, does well with ranking both OSE and Vikidia-En texts by their reading level, and is more robust to variation among the corpora, compared to the regBERT model. All measures achieve performance > 0.87 for both the datasets with $NPRM$. The regBERT performs comparably on Vikidia-En, but does poorly on OSE. While the results for $NPRM$ are still somewhat lower in the cross-corpus evaluation than in within corpus evaluation setups, it has to be noted that this evaluation is done without any additional fine-tuning on the target datasets. We did not test for statistical significance in this case as the numbers have large differences between regBERT and $NPRM$ in most cases. This experiment leads us to a conclusion that $NPRM$ can successfully be used to rank documents on a different reading level scale too.

5.5 Zero shot, cross-lingual pair-wise ranking

Zero-shot cross-lingual scenario aims to evaluate whether a model trained on one language can be effectively used to rank texts from another language correctly without explicitly training on the target language. We evaluated NPRM and regBERT models trained with a multilingual BERT (mBERT) model as the base for this task. Both the models were trained on Newsela-En dataset and evaluated on Newsela-Es and Vikidia-Fr datasets. The mBERT[7] model is pre-trained on a corpus of multilingual data from 104 languages, including all the three languages in our experiment: English, French and Spanish. Table 7 shows the results of this experiment.

$NPRM$(mBERT)				
Testset	NDCG	SRC	KTCC	RA
NewsEla-Es	0.996	0.985	0.971	0.864
Vikidia-Fr	0.930	0.622	0.622	0.811
regBERT (mBERT)				
NewsEla-Es	0.992	0.957	0.931	0.741
Vikidia-Fr	0.913	0.527	0.527	0.764

Table 7: Zero-shot, cross-lingual Evaluation (Trained on Newsela-En)

We observe that the $NPRM$ with mBERT either performs comparably or outperforms a regression mBERT model on all metrics, for both the datasets. Specifically, the $NPRM$ has a performance increase of 12.3% in RA for Newsela-Es over Vikidia-Fr. Thus, we can conclude that our pairwise ranking approach performs well even in cross-lingual scenarios, and zero-shot, cross-lingual transfer can be useful to setup strong baseline models for new languages.

We can notice a lower performance on Vikidia-Fr compared to Newsela-ES. Apart from the fact that they are different languages, it can potentially also be due to the fact that Newsela-ES has content from the same domain as Newsela-EN, whereas Vikidia-Fr has more diverse content. It is also possible that the ranking metrics penalize Vikidia-Fr predictions more, as there are only two reading levels. A ranking can still be scored well if most of the ranking order is correct. However, in the case of Vikidia-Fr, an incorrect ranking levels would result in a completely reversed list, which is heavily penalized in SRC and KTCC. Thus, small number of completely incorrectly ranked slugs can result in low SRC and KTCC scores for Vikidia-Fr, but can still result in high SRC and KTCC scores for

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

NewsEla-ES. More future experiments, with additional languages, would lead us towards a better understanding of what works well across languages and datasets.

Ranking Metrics : We reported four ranking metrics in these experiments. While they all get high numbers in some experimental settings, none of them consistently seem like a better choice than others. We observe that the large majority of the methods score close to 1.0 on NDCG. In comparison, the SRC and KTCC, while generally quite high, appear more susceptible to poor ranking performance. We notice that RA is lower than SRC and KTCC for OSE (Table 6) and NewsEla-Es (Table 6), but SRC and KTCC lag behind RA for Vikidia-Fr (Table 7). We hypothesize that this could be because of the number of reading levels in the datasets. SRC and KTCC seem more forgiving when number of reading levels are more.

Clearly, each metric addresses the evaluation of ranking differently, and as the results show, there is no single model that consistently does well across all metrics, in all the evaluations. We hope that this illustrates the value of reporting multiple metrics while benchmarking a new model. Future work in this direction should also focus on the evaluation of the evaluation metrics themselves for this task.

6 Conclusions and Discussion

In this paper we proposed a neural pairwise ranking model for ARA ($NPRM$). We performed within corpus, cross-corpus and cross-lingual evaluations to benchmark $NPRM$. Our results in the context of the research questions we started with (Section 1) are discussed below:

1. *Is neural, pairwise ranking a better approach than classification or regression for ARA, to achieve cross-corpus compatibility?* - While regression, classification, and pairwise-ranking models all achieve comparable performance in a within-corpus scenario, pairwise ranking performs better in cross-corpus and cross-lingual evaluation scenarios.

2. *Is zero-shot, cross-lingual transfer possible for ARA models through such a ranking approach?* - Our experiments show that zero-shot cross-lingual ARA is possible with pairwise ranking. Our proposed model, $NPRM$, trained with English texts achieved $> 80\%$ ranking accuracy on both NewsEla-Es and Vikidia-Fr datasets.

Limitations of NPRM: $NPRM$ models the relative reading difficulty level between texts. While this approach has performed well for our generalizability experiments, there is a general lack of interpretability with $NPRM$. For example, the NewsEla-en dataset contains reading level designations that align with the common core educational standards (Porter et al., 2011), and are interpreted to match the school grades of U.S students from kindergarten to high school. Since the aggregation process of $NPRM$ sums the predicted scores between pairwise comparisons of an intended ranking, the aggregated score is bounded above by the input list size, is unlikely to correspond to the original reading level scale. Further, $NPRM$ takes a list of texts as input and the model forces the constraint of having at least two texts to be ranked as input. Hence, $NPRM$ is suitable only for scenarios where ranking by reading level is useful (e.g., ranking of search results by their reading level).

Outlook: All the five datasets used in these experiments come primarily from news and wikipedia articles. However, the ARA is also studied in other domains (e.g., financial disclosures (Loughran and McDonald, 2014)). Future work can test the validity of the conclusions of this paper in new domains. Further, all the datasets are human created texts. It would be interesting to explore how the model works for applications like text simplification and machine translation, which can support existing research on evaluating machine generated text.

Ethics Statement

In this paper, we report on the creation of a new dataset for readability assessment. The data collection process did not involve any human participants. So, no ethics board approval was necessary. Both the websites are available under permissive licenses that allow sharing and redistributing. The released dataset will follow the same procedures and is freely available at `https://zenodo.org/record/6327828`. An important point to note in the use of the dataset is that the length of texts is much shorter in the "simple" versions compared to regular Wikipedia articles, which may affect the quality of results in some use cases of the dataset.

Acknowledgements

We thank the four anonymous reviewers for their helpful comments on the submitted manuscript. We also thank Gabriel Bernier-Colborne and Taraka Rama for their comments on the initial draft of this paper, and Michel Simard and Yunli Wang for the early discussions on this topic.

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564.

Qingyao Ai, Xuanhui Wang, Sebastian Bruch, Nadav Golbandi, Michael Bendersky, and Marc Najork. 2019. Learning groupwise multivariate scoring functions using deep neural networks. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 85–92.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental ccg parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057.

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Miriam Cha, Youngjune Gwon, and HT Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006.

William Jay Conover. 1999. *Practical nonparametric statistics*, volume 350. john wiley & sons.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

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*.

William H DuBay. 2007. *Unlocking language: The classic readability studies*. Impact Information.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237.

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

Adam Fourney, Meredith Ringel Morris, Abdullah Ali, and Laura Vonessen. 2018. Assessing the readability of web search results for searchers with dyslexia. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1069–1072.

Thomas François and Cédrick Fairon. 2012. An "ai readability" formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477.

Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. 2020. Learning-to-rank with bert in tf-ranking. Technical report, arXiv.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 71–79.

David M Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 958–968.

Joseph Marvin Imperial. 2021. Knowledge-rich bert embeddings for 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.

Denys Katerenchuk and Andrew Rosenberg. 2016. RankDCG: Rank-ordering evaluation measure. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3675–3680, Portorož, Slovenia. European Language Resources Association (ELRA).

Jin Young Kim, Kevyn Collins-Thompson, Paul N Bennett, and Susan T Dumais. 2012. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 213–222.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

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*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Tim Loughran and Bill McDonald. 2014. Measuring readability in financial disclosures. *The Journal of Finance*, 69(4):1643–1671.

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.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020a. An analysis of transfer learning methods for multilingual readability assessment. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 95–100.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020b. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 193–203.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Changping Meng, Muhao Chen, Jie Mao, and Jennifer Neville. 2020. Readnet: A hierarchical transformer framework for web article readability analysis. *Advances in Information Retrieval*, 12035:33.

Mohsen Mesgar and Michael Strube. 2015. Graph-based coherence modeling for assessing readability. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 309–318, Denver, Colorado. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Hamid Mohammadi and Seyed Hossein Khasteh. 2019. Text as environment: A deep reinforcement learning text readability assessment model. *arXiv preprint arXiv:1912.05957*.

Rama Kumar Pasumarthi, Sebastian Bruch, Xuanhui Wang, Cheng Li, Michael Bendersky, Marc Najork, Jan Pfeifer, Nadav Golbandi, Rohan Anil, and Stephan Wolf. 2019. Tf-ranking: Scalable tensorflow library for learning-to-rank. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2970–2978.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Andrew Porter, Jennifer McMaken, Jun Hwang, and Rui Yang. 2011. Common core standards: The new us intended curriculum. *Educational researcher*, 40(3):103–116.

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*.

Kathleen M Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. 2014. The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, 115(2):184–209.

Wade Shen, Jennifer Williams, Tamas Marius, and Elizabeth Salesky. 2013. A language-independent approach to automatic text difficulty assessment for second-language learners. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 30–38.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.

Manjira Sinha, Sakshi Sharma, Tirthankar Dasgupta, and Anupam Basu. 2012. New readability measures for bangla and hindi texts. In *Proceedings of COLING 2012: Posters*, pages 1141–1150.

A Jackson Stenner. 1996. Measuring reading comprehension with the lexile framework.

Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting texts by readability. *Computational linguistics*, 36(2):203–227.

Amalia Todirascu, Thomas François, Nuria Gala, Cédric Fairon, Anne-Laure Ligozat, and Delphine Bernhard. 2013. Coherence and cohesion for the assessment of text readability. *Natural Language Processing and Cognitive Science*, 11:11–19.

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*, pages 297–304.

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

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Xuanhui Wang, Cheng Li, Nadav Golbandi, Mike Bendersky, and Marc Najork. 2018. The lambdaloss framework for ranking metric optimization. In *Proceedings of The 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*, pages 1313–1322.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

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*, pages 12–22.

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*, 3:283–297.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.