Robustness of Hybrid Models in Cross-domain Readability Assessment

Ho Hung Lim, Tianyuan Cai, John S. Y. Lee, Meichun Liu
Department of Linguistics and Translation
City University of Hong Kong
{hhlim3,jsylee,meichliu}@cityu.edu.hk, tianycai-c@my.cityu.edu.hk

Abstract

Recent studies in automatic readability assessment have shown that hybrid models — models that leverage both linguistically motivated features and neural models — can outperform neural models. However, most evaluations on hybrid models have been based on in-domain data in English. This paper provides further evidence on the contribution of linguistic features by reporting the first direct comparison between hybrid, neural and linguistic models on cross-domain data. In experiments on a Chinese dataset, the hybrid model outperforms the neural model on both in-domain and cross-domain data. Importantly, the hybrid model exhibits much smaller performance degradation in the cross-domain setting, suggesting that the linguistic features are more robust and can better capture salient indicators of text difficulty.

1 Introduction

Automatic Readability Assessment (ARA) predicts how difficult it is for the reader to understand a text. Traditional machine learning approaches for ARA typically train statistical classifiers with handcrafted features (Pitler and Nenkova, 2008; Sung et al., 2015). Similar to other NLP tasks, neural approaches have recently achieved superior performance (Tseng et al., 2019; Azpiazu and Pera, 2019; Martinc et al., 2021). Combining linguistic features and neural models has been found to benefit a variety of NLP tasks (Lei et al., 2018; Strubell et al., 2018). While these 'hybrid' models have also been applied in ARA, there have been varying results ranging from no effect (Deutsch et al., 2020), marginal improvement (Filighera et al., 2019), to significant improvement (Lee et al., 2021).

Past studies comparing hybrid and neural models have mostly been conducted in an in-domain setting, with the training and test data drawn from the same source. However, real-word applications of ARA models are often targeted at cross-domain or cross-corpus data. Consider the task of retrieving extra-curricular reading materials for language learning from web texts, which likely diverge in style and content from the training data. In-domain evaluation therefore may not accurately reflect the actual performance on such tasks.

This paper focuses on the task of predicting the grade level of an input text. We present the first comparison between hybrid, neural and linguistic models on this task in the cross-domain setting. Our contribution is two-fold. First, we show that the hybrid model outperforms the neural model both in-domain and cross-domain in Chinese datasets, providing further evidence on the contribution of linguistic features. Second, the hybrid model exhibits much smaller performance degradation on cross-domain data, suggesting their robustness and ability to capture more salient indicators of text difficulty.[1]

After a review of previous work (Section 2), we present our datasets (Section 3). We then outline our approach (Section 4) and report experimental results (Section 5).

2 Background

2.1 Hybrid model design

Statistical classifiers can be trained on a variety of features, capturing lexical, syntactic and semantic characteristics of a text, to determine its readability or grade level (Dell'Orletta et al., 2011; François and Fairon, 2012; Sung et al., 2015). While these classifiers lend themselves to more explainable and linguistically-motivated results, neural models can achieve superior performance and do not require feature engineering (Tseng et al., 2019; Martinc et al., 2021).

Various methods for combining these approaches have been investigated. For example,

[1] Our implementation is available at https://github.com/hhlim333/ALTA2022Readability

a Bi-LSTM can incorporate part-of-speech information (Azpiazu and Pera, 2019). A statistical classifier can directly use sentence embeddings as features (2021). It can also incorporate the decision of the neural model as a single numeric feature (Deutsch et al., 2020), or 'soft' labels expressing the probabilities of each grade as predicted by the neural model (Lee et al., 2021). Our experiments will directly compare the performance of these three approaches.

2.2 In-domain vs. cross-domain evaluation

There can be a mismatch between ARA training datasets and the texts on which the ARA model is deployed. Domain adaptation techniques can be applied to address differences between native and non-native texts. For example, scores from an ARA ranking model trained on graded texts for native speakers can help estimate the CEFR level of a text for non-native learners (Xia et al., 2016).

Another type of mismatch is caused by cross-domain or cross-corpus data, which has been investigated in the ranking task in ARA. When ranking models are trained on Newsela, they suffered a performance degradation when tested on OneStopEnglish and Vikidia (Lee and Vajjala, 2020). For the grade prediction task, however, cross-domain evaluation has been reported mainly in terms of correlation (Chen and Meurers, 2016). This may be due to the fact that different grade scales are adopted in the major benchmarks, such as Newsela, OneStopEnglish and WeeBit. In this work we leverage two comparable datasets in Chinese (Section 3) to conduct cross-domain evaluation on hybrid models to assess the contribution of linguistic features in the grade prediction task.

3 Data

Since the benchmark ARA corpora adopt different grade scales (Section 2.2), we utilize two datasets of Chinese-language textbook materials, graded under comparable scales but drawn from different sources.

Mainland China texts (in-domain): Drawn from textbooks for Chinese language used in Mainland China (Lee et al., 2020), this dataset consists of 7.15M characters distributed in 4,831 passages in 12 grades (Cheng et al., 2020).

Hong Kong texts (cross-domain): Chinese-language textbooks in Hong Kong follow similar language proficiency standards as those in the Mainland. They are however compiled independently from different sources and use traditional rather than simplified characters, thus providing a challenging cross-domain scenario. We constructed a corpus of 298K characters distributed in 502 passages in 12 grades, all taken from current textbooks in Hong Kong.

Grade	# text	# char
1	50	4793
2	50	9042
3	50	15107
4	50	22191
5	50	28345
6	50	32776
7	50	35957
8	42	32859
9	46	44906
10	35	31179
11	13	22703
12	16	18686

Table 1: Statistics on the corpus of Hong Kong texts

4 Approach

We compared the following ARA models for predicting the grade (1-12) of an input text.

4.1 Baseline: Neural Model

We fine-tuned[2] MacBERT (Cui et al., 2020), RoBERTa (Cui et al., 2020), BERT (Devlin et al., 2019) and BERT-wwm (Cui et al., 2020) on the Mainland dataset for grade prediction.[3]

4.2 Baseline: Linguistic Model

We trained a statistical classifier on the 221 linguistic features provided by ChiLingFeat [4], an open-source toolkit that extracts most features used in previous Chinese ARA studies (Sung et al., 2015; Lu et al., 2020). We evaluated SVM, Random Forest (RF), and XGBoost (XGB) using the implementation in scikit-learn (Pedregosa et al., 2011).

[2]We used the code by Lee et al. (2021) in default parameters for fine-tuning, accessed from https://github.com/yjang43/pushingonreadability_transformers

[3]We used macbert-large, chinese-roberta-wwm-ext, bert-base-chinese, and chinese-bert-wwm, respectively.

[4]https://github.com/ffliu6/ChiLingFeat

Trans-former	Hybrid model type	In-domain	Cross-domain
BERT	Hard labels	0.312	0.288
	Soft labels	**0.342**	**0.290**
	Sent. Embed.	0.322	0.269
BERT-wwm	Hard labels	0.295	**0.283**
	Soft labels	**0.341**	0.278
	Sent Embed.	0.318	**0.283**
RoBERTa	Hard labels	0.301	0.285
	Soft labels	**0.341**	**0.301**
	Sent Embed.	0.318	0.287
MacBERT	Hard labels	0.305	0.283
	Soft labels	**0.353**	**0.309**
	Sent. Embed.	0.329	0.269

Table 2: Accuracy of the three hybrid model types (Section 4.3)

We applied Variance Threshold algorithm in scikit-learn for feature selection, but obtained the best result with the full feature set.

4.3 Hybrid Model

Following Lee et al. (2021), we adopted the simple approach of wrapping linguistic features and neural model output in a non-neural, statistical classifier. We evaluated three types of hybrid models:

Hard labels (Deutsch et al., 2020): The grade of the input text, as predicted by the neural model (Section 4.1) serves as an additional feature in the classifier.

Soft labels (Lee et al., 2021): The probabilities of each grade, as predicted by the neural model (Section 4.1), serve as additional features.

Sentence Embeddings (Imperial, 2021): The sentence vectors, produced by SBERT (Reimers and Gurevych, 2019) from the sentences in the input text, serve as additional features.

5 Experiments

In-domain evaluation used stratified 5-fold cross-validation on the Mainland Chinese dataset, based on a train:dev:test split of 8:1:1. Cross-domain evaluation used the entire Mainland China corpus as training data, and the entire Hong Kong corpus as test data. Among the three classifiers, RF outperformed SVM and XGB in most settings and metrics. The rest of the paper will refer to results based on RF.

5.1 Metrics

We use accuracy, F1, adjacent accuracy and quadratic weighted kappa (QWK) as our metric for the experiment. For adjacent accuracy, the system is considered correct if the predicted label is within one grade higher or lower than the gold grade. QWK also helps capture the distance between gold and predicted grades. These metrics give a comprehensive evaluation of model performance from different perspectives.

5.2 Hybrid model types

Table 2 reports the performance of the three hybrid model types (Section 4.3). For in-domain evaluation, hybrid models with soft labels outperformed those with hard labels and sentence embeddings, regardless of the transformer. For cross-domain evaluation, that was also the case for BERT, RoBERTa and MacBERT. The only exception was BERT-wwm, for which hard labels and embeddings performed slightly better (0.283), but still less accurate than the other transformers. The results presented below will be based on soft labels.

5.3 In-domain evaluation

Baselines. As shown in Table 3, the Linguistic Model yielded 0.276 accuracy in the in-domain setting. It was outperformed by the Neural Model regardless of the transformer used. MacBERT achieved the best performance for the Neural Model on accuracy (0.333) and all other metrics.

Hybrid Model. The Hybrid Model trained on MacBERT attained the highest accuracy (0.353) and F1, while RoBERTa led to the best adjacency accuracy and QWK (tied with BERT). Regardless of the choice of transformer or metric, the Hybrid Model outperformed both baselines. The absolute accuracy gains over the Neural Model ranged from 2.0% (MacBERT) to 4.8% (RoBERTa).[5] Consistent with previous results on English datasets (Lee et al., 2021), linguistic features enhance the performance of neural models on the Chinese datasets.

5.4 Cross-domain evaluation

Baselines. As expected, model performance degraded in the cross-domain setting. MacBERT produced the best-performing Neural Model in terms of all four metrics. Unlike the in-domain evaluation, the Linguistic Model outperformed the Neural

[5]The improvement is statistically significant for all four models at $p < 0.01$ according to McNemar's Test with continuity correction.

Metric	Linguistic Model (RF)		Neural Model			Hybrid Model	
	In-domain	Cross-domain	Transformer	In-domain	Cross-domain	In-domain	Cross-domain
Acc.	0.276	0.263 (-0.013)	BERT	0.303	0.197 (-0.106)	0.342	0.290 (-0.052)
			BERT-wwm	0.308	0.196 (-0.112)	0.341	0.278 (-0.063)
			RoBERTa	0.293	0.196 (-0.097)	0.341	0.301 (-0.040)
			MacBERT	0.333	0.239 (-0.094)	**0.353**	**0.309** (-0.044)
Adj. Acc.	0.596	0.561 (-0.035)	BERT	0.618	0.504 (-0.114)	0.690	0.656 (-0.034)
			BERT-wwm	0.627	0.485 (-0.142)	0.688	0.639 (-0.049)
			RoBERTa	0.599	0.488 (-0.111)	**0.699**	**0.683** (-0.016)
			MacBERT	0.644	0.563 (-0.081)	0.685	0.677 (-0.008)
F1	0.259	0.221 (-0.038)	BERT	0.273	0.154 (-0.119)	0.338	0.262 (0.076)
			BERT-wwm	0.280	0.154 (-0.126)	0.337	0.249 (-0.088)
			RoBERTa	0.256	0.147 (-0.109)	0.335	0.273 (-0.062)
			MacBERT	0.307	0.198 (-0.109)	**0.348**	**0.276** (-0.072)
QWK	0.739	0.475 (-0.264)	BERT	0.759	0.633 (-0.126)	**0.841**	0.817 (-0.024)
			BERT-wwm	0.755	0.612 (-0.143)	0.833	0.782 (-0.051)
			RoBERTa	0.731	0.597 (-0.134)	0.841	0.822 (-0.019)
			MacBERT	0.768	0.712 (-0.056)	0.829	**0.832** (+0.003)

Table 3: Performance of the Hybrid Model and the two baselines. The gap between in-domain and cross-domain performance is shown in brackets

Training dataset size	Linguistic Model (RF)		Neural Model		Hybrid Model	
	In-domain	Cross-domain	In-domain	Cross-domain	In-domain	Cross-domain
20%	0.281	0.247 (-0.034)	0.267	0.231 (-0.036)	0.325	0.294 (-0.031)
60%	0.286	0.259 (-0.027)	0.307	0.236 (-0.071)	0.337	0.299 (-0.036)
100%	0.276	0.263 (-0.013)	0.333	0.239 (-0.106)	0.353	0.309 (-0.044)

Table 4: Model accuracy at different training dataset size, expressed in percentage of the full dataset. The gap between in-domain and cross-domain performance is shown in brackets

Model in terms of accuracy (0.263 vs. 0.239) and F1, though worse in terms of adjacent accuracy and QWK. Its competitive performance can be attributed to the robustness of linguistic features in the face of dissimilar materials. While the Linguistic Model degraded only slightly (-0.013) in accuracy on cross-domain data, the Neural Model suffered a much more substantial drop (-0.094).

Hybrid Model. The Hybrid Model outperformed both baselines in all metrics and all transformers.[6] MacBERT again led to the best performance in terms of accuracy (0.309), F1 and QWK, but was slightly worse than RoBERTa in adjacent accuracy.

The superior performance of the Hybrid Model resulted from its smaller degradation on cross-domain data. This can be seen by the gap between in-domain and cross-domain performance, shown in brackets in the "Cross-domain" column in Table 3). For all transformers and all metrics, the gap was substantially smaller with the Hybrid Model. For example, the gap was only 0.044 cross-domain but more than doubled (0.094) in-domain for MacBERT. This suggests that some textual characteristics learned by the Neural Model may be only accidentally correlated with readability in the training corpus, while the Hybrid Model benefits from linguistic features that are more generally relevant to readability and therefore transferable to new domain.

Our hypothesis can be corroborated with the analysis on various dataset sizes in Table 4. When trained on only 20% of the dataset, all three models exhibited a similar gap between in-domain and cross-domain performance. With additional training data, the Neural Model became more accurate

[6]The improvement of the hybrid model over the neural model is statistically significant for BERT, BERT-wwm and RoBERTa at $p < 0.00001$ according to McNemar's Test.

in-domain (0.267 to 0.333). However, the improvement hardly carried over cross-domain, leading to a growing performance gap (-0.036 to -0.106), possibly indicating overfit to corpus-specific textual characteristics. In contrast, the gap shrank for the Linguistic Model, and remained relatively stable for the Hybrid Model, even as it improved steadily in accuracy.

6 Conclusions

We have presented the first cross-domain comparison of hybrid, neural and linguistic models for ARA. Results on a Chinese dataset show that the hybrid model outperforms the neural model both in-domain and cross-domain. Analyses on the gap between in-domain and cross-domain performance further demonstrate the robustness of linguistic features. While the gap grows for the neural model as more training data becomes available, it remained more stable for the hybrid model. These results are expected to inform future ARA research by showing that linguistic features can help neural models capture more generalizable characteristics for text difficulty, especially in the cross-domain context.

Acknowledgements

We thank Prof. Dekuan Xu for providing access to the corpus of textbooks from Mainland China. This work was partly supported by the Language Fund from the Standing Committee on Language Education and Research (project EDB(LE)/P&R/EL/203/14) and by the General Research Fund (project 11207320).

References

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

Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 84–94, San Diego, CA. Association for Computational Linguistics.

Yong Cheng, Dekuan Xu, and Jun Dong. 2020. On key factors of text reading difficulty grading and readability formula based on chinese textbook corpus [in chinese] 基于语文教材语料库的文本阅读难度分级关键因素分析与易读性公式研究. *Applied Linguistics* 语言文字应用, 1:132–143.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Findings of EMNLP*. Association for Computational Linguistics.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proc. 2nd Workshop on Speech and Language Processing for Assistive Technologies*.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In *Proc. North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*, page 335–348. Springer.

Thomas François and Cédric Fairon. 2012. An "AI Readability" Formula for French as a Foreign Language. In *Proc. EMNLP-CONLL*.

Joseph Marvin Imperial. 2021. BERT Embeddings for Automatic Readability Assessment. In *Proc. Recent Advances in Natural Language Processing*, page 611–618.

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*.

John Lee, Meichun Liu, and Tianyuan Cai. 2020. Using Verb Frames for Text Difficulty Assessment. In *Proc. International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*.

Justin Lee and Sowmya Vajjala. 2020. A Neural Pairwise Ranking Model for Readability Assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813.

Wenqiang Lei, Yuanxin Xiang, YuweiWang, Qian Zhong, Meichun Liu, and Min-Yen Kan. 2018. Linguistic Properties Matter for Implicit Discourse Relation Recognition: Combining Semantic Interaction, Topic Continuity and Attribution. In *Proc. AAAI*, pages 4849–4855.

Dawei Lu, Xinying Qiu, and Yi Cai. 2020. Sentence-level readability assessment for l2 chinese learning. *CLSW 2019, LNAI*, 11831:381–392.

Matej Martinc, Senja Pollak, Marko, and Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and O. Grisel. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: a Unified Framework for Predicting Text Quality. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. EMNLP-IJCNLP*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Yao-Ting Sung, Wei-Chun Lin, Scott Benjamin Dyson, Kuo-En Chang, and Yu-Chia Chen. 2015. Leveling L2 Texts Through Readability: Combining Multilevel Linguistic Features with the CEFR. *The Modern Language Journal*, 99(2):371–391.

Hou-Chiang Tseng, Hsueh-Chih Chen, Kuo-En Chang, Yao-Ting Sung, and Berlin Chen. 2019. An Innovative BERT-Based Readability Model. In *Lecture Notes in Computer Science, vol 11937*.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proc. 11th Workshop on Innovative Use of NLP for Building Educational Applications*, page 12–22.