Cross-Lingual Leveled Reading Based on Language-Invariant Features

Simin Rao[*], **Hua Zheng**[*], **Sujian Li**[†]
Department of Computer Science and Technology, Peking University
Key Lab of Computational Linguistics (MOE), Peking University
{raosimin, zhenghua,lisujian}@pku.edu.cn

Abstract

Leveled reading (LR) aims to automatically classify texts according to different reading capabilities and provide appropriate reading materials to readers. However, most state-of-the-art LR methods rely on the availability of copious annotated resources, which prevents their adaptation to low-resource languages like Chinese. In our work, to tackle Chinese LR, we explore to perform different language transfer methods on English-Chinese LR. Specifically, we focus on adversarial training and cross-lingual pre-training method to transfer the LR knowledge learned from annotated data in the rich-resource English language to Chinese. For evaluation, we introduce the age-based standard to align datasets with different leveling standards, and conduct experiments in both zero-shot and few-shot settings. Experiments show that the cross-lingual pre-training method can capture language-invariant features more effectively than adversarial training. We also conduct analysis to propose further improvement in cross-lingual LR.

1 Introduction

Imagine searching the appropriate reading materials for a 10-year-old child in the bookstore: the Tale of Peter Rabbit is a bit outdated; Animal Farm, though sounds suitable, is too allegorical; the Harry Potter series may be just right for the age. Leveled reading (LR) provides such selection guides by automatically classifying texts with regard to the reading level appropriate for readers, which has proven to be of importance in multiple fields, including education (Lennon and Burdick, 2004), health (Petkovic et al., 2015) and advertisement (Chebat et al., 2003). Different from the traditional readability assessment (Aluisio et al., 2010; Madrazo Azpiazu and Pera, 2020) which is formulated as a binary classification problem, LR can be regarded as a multi-class classification task that provides specific reading levels with regard to the cognitive reading level instead of text quality. This fine-grained leveling forms a fundamental component in downstream applications, since it is essential to label different levels even within the Harry Potter series when the stories get darker.

Most previous research focus on English by extracting language-specific features, ranging from traditional readability formulas to using machine learning methods. As the most widely studied language in LR, English holds mature LR standards with abundant reading materials, such as Lexile (Lennon and Burdick, 2004) and Accelerated Reader (Topping et al., 2008), and has recently developed a set of LR datasets for training automatic methods, such as the WeeBit (Vajjala and Meurers, 2012), NewSela (Xu et al., 2015) and OneStopEnglish (Vajjala and Lučić, 2018a) corpus. By contrast, low-resource languages like Chinese lack both established LR standards and training data, which results in only a few LR research conducted in Chinese (Sun et al., 2020). Can we use the existing resources of English to guide the cross-lingual LR of low-resource languages such as Chinese.

There has been a recent trend towards learning language-invariant features to ease the cross-lingual generalization from high-resource languages to low-resource languages (Litschko et al., 2018; Kondratyuk and Straka, 2019). We hypothesize that these language-invariant features also exist in LR, especially in the equivalent level of reading among different languages, which may be automatically extracted through deep learning methods.

For example, the reading materials in different languages in the equivalent level may talk about the similar story and express same thoughtsand they may have similar changes in text structure and vocabulary as the level changes.

Thus, to verify our hypothesis and transfer En-

[*] Equal Contribution
[†] Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2677–2682
November 7–11, 2021.

glish LR knowledge into Chinese, we explore both adversarial training and cross-lingual pre-training method to extract language-invariant features for English and Chinese LR corpora to guide LR in Chinese. Overall, our contributions are summarized as follows:

- We organize the available LR datasets and precess the new LR datasets, including three LR corpora for English, and a variety of textbooks across 12 grade levels and extracurricular books in Chinese. We re-classify the datasets according to age into a uniform standard of reading levels to map both Chinese and English datasets for transfer learning.
- We explore the performance of two transfer learning methods, adversarial training and multi-lingual pre-training, on our aligned LR datasets.

2 Related Works

2.1 Leveled Reading Methods

Early works on LR devised various readability formulas, such as the Gunning Fog Index (Gunning, 1952), Automated Readability Index (Senter and Smith, 1967) and Flesch Reading Ease (Kincaid et al., 1975), which mainly rely on shallow language features based on ratios of characters, phrases and sentences. Later work adopted statistical machine learning methods based on extensive feature engineering, which generally improved accuracy by capturing semantic and contextual features (Vajjala and Meurers, 2012; Xia et al., 2016; Vajjala and Lučić, 2018b). Recently, Martinc et al. (2019) and Deutsch et al. (2020) used deep neural networks to enhance LR and achieved the state-of-the-art performances. Due to resource limitations, only a few works study LR in Chinese (Liu et al., 2017; Sun et al., 2020), which does not have copious annotated data like English.

2.2 Cross-Lingual Methods

Cross-lingual methods transfer knowledge from high-resource languages with abundant annotated data to low-resource target languages with limited or even no annotated data. Some works trained cross-lingual representations based on bilingual parallel corpora (Mikolov et al., 2013; Gouws et al., 2015); Other works used direct transfer methods by employing self-training (Artetxe et al., 2017) or unsupervised models like adversarial training (Chen et al., 2018) and heuristic initialization (Artetxe et al.). Madrazo Azpiazu and Pera (2020) first proposed to use cross-lingual strategy for enhancing readability assessment as a binary classification problem, which shows improvement in accuracy for low-resource languages.

3 LR Datasets

In this section, we elaborate on the LR datasets collected which are classified by the standards of gradeletterand number. And we re-align these datasets using an age-based standard because these standards have been designated approximate age range.

English Datasets: We compile the English dataset based on both the WeeBit (Vajjala and Meurers, 2012) corpus and Reading A-Z (RAZ)[1] reading materials. The WeeBit corpus contains 3,125 texts of five classes, based on WeeklyReader and BBC-Bitesize for readers aged from 6 to 17. I'm Following (Deutsch et al., 2020), we apply additional preprocessing to remove extraneous materials in each text, such as copyright declaration and links. The RAZ reading materials originally contain books at 29 levels from level AA to Z2. And each level has a corresponding suitable age. Level AA to C are suitable for children aged from 4 to 6 which are not in our consideration. Among the materials available to the public, we select and compile 360 texts from level D to Z for readers aged from 6 to 17.

Chinese Datasets: There is no mature system and corpus for Chinese LR. For evaluations and few-shot training, we compile a dataset of Chinese school textbooks. Considering the needs of teaching are from simple to difficult, the difficulty of Chinese textbooks in the same grade is not the same difficulty. And the purpose of our study is to investigate the guiding significance of English LR for Chinese. This restricts direct use of the Chinese textbooks for supervised classification tasks. The dataset we have collected contains 2,903 texts in school textbooks for six grades of elementary schools, three grades of junior high schools and four grades of senior high schools. To ensure full coverage and fine consistency, we use textbooks of six local editions and deleted texts written in classic Chinese. In addition, we collected 21 extra extracurricular leveled Chinese books following the recommendation of (QianLei, 2015).

[1] https://www.learninga-z.com/

Level	Age	EN-Weebit (#)	EN-RAZ (#)	CN-textbooks (#)	CN-extra Books (#)
1	6-7	WRLevel2 (571)	D-P (128)	Grades 1-2 (551)	Grades 1-2 (1)
2	8-9	WRLevel3 (700)	Q-W (127)	Grades 3-4 (601)	Grades 3-4 (12)
3	10-11	WRLevel4 (726)	X-Z (120)	Grades 5-6 (642)	Grades 5-6 (18)
4	12-13	BitKS3 (579)	N/A	Grades 7-8 (673)	N/A
5	14-17	BitGCSE (908)	N/A	Grades 9-12 (453)	N/A

Table 1: LR standards for both Chinese and English datasets. N/A represents no corresponding data available in the dataset. Each level corresponds to a specific age range based on data distribution. # denotes the number of each level for both Chinese and English datasets.

Different grades have corresponding age ranges. For example, grades 3-4 are suitable for Level D to P are suitable for children aged from 8 to 9. These datasets we used in our models are all processed text data instead of the printed book. So the physical manifestations of the printed book like text structure, page layout and illustration have been lost, which plays an important part in our task even in many NLP tasks. In the future, we are supposed to use the data more comprehensively to get more information.

Standard Benchmarks: Since leveling standards vary across different datasets, previous methods are trained and evaluated independently on each dataset (Martinc et al., 2019; Deutsch et al., 2020). To align both English and Chinese datasets, we map each data sets into five reading levels with respect to different ages, as shown in Table 1. For example, the original standards of the WeeBit corpus overlap on the neighboring levels, and thus we take the lower boundary of each overlapping level as the standard level. And We re-classify the RAZ dataset according to age into three reading levels. For example, level D to P are suitable for children aged from 6 to 7.

4 Methodology

Adversarial training and pre-training are recently popular deep learning methods, which can better learn text representation, and have been applied to cross-lingual tasks to extract common features. Inspired by this, we also try to apply these two methods to our cross-lingual LR task.

4.1 Adversarial Model for LR

We extend the ADAN model in (Chen et al., 2018) to incorporate the language-invariant features, containing three main components in the network: a joint feature extractor F that maps the input to the shared feature space, a language discriminator D that predicts whether the input is from English or Chinese, and an LR classifier R that classifies the texts into its reading level,as shown in Figure 1. If the language discriminator can't distinguish between Chinese and English, then we can recognize that the model has learned language-independent features.

4.2 Pre-training Model for LR

Cross-lingual Language Model (XLM) (Conneau and Lample, 2019) is a transformer-based (Vaswani et al., 2017) model that has been pretrained on the Wikipedias of 104 languages using masked language model, achieving state-of-the-art results on multiple cross-lingual tasks (Ruder et al., 2019), especially for low-resource languages by training on the high-resource language. The model uses a shared vocabulary and adopt byte-pair encoding as the tokenizer. In our LR setting, we fine-tune XLM by adding a classification layer with softmax on top of XLM and for LR prediction. **Dataset**: We split the datasets described in 3 into training, validation and test set by 8:1:1. Specifically, Weebit and Raz are used as English training set during zero-shot training, and the CN-textbooks data is used for few-shot training. CN-extra books are only used as test datasets.

5 Experiments

5.1 Experimental Settings

XLM: We use the pretrained XLM-RoBERTa (XLM-R) downloaded from Hugging Face [2] unmodified. We run 20 epochs with a batch size of 32 during zero-shot and few-shot training. We adopt Adam (Kingma and Ba, 2015) as the optimizer with a learning rate of 1e-5. Since the limited length of the pre-training model and all our data is long text, we divide each article in our datasets into one piece of data according to paragraphs. And we take only the first 512 tokens of each data to reduce the effects of the length limit in XLM.

[2]https://huggingface.com/

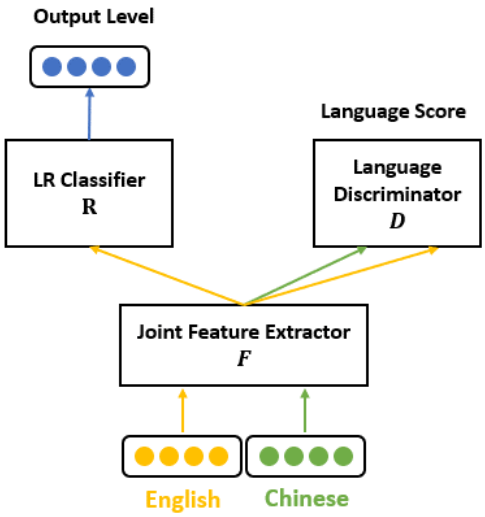


Figure 1: Illustration of the adversarial model for LR.

	Zero-shot	Few-shot
GFI	24.95	24.95
FRE	20.58	20.58
ADAN	32.19	45.62
XLM	**51.61**	**65.07**

Table 2: Evaluation results. Best results are in **Bold**.

ADAN: The feature extractor F, LR classifier R and Language discriminator D have three fully-connected layers with the ReLU activation. We adopt Adam (Kingma and Ba, 2015) as the optimizer with a learning rate of 5e-4.

Baselines: We use different existing English readability formulas to calculate the readability of Chinese text and we adopt two highly-recognized and more suitable readability formulas for comparison, the Gunning Fog Index (GFI) (Gunning, 1952) and the Flesch Reading Ease (FRE) (Kincaid et al., 1975). Since these readability formulas originated in English texts, we directly apply the formulas to the Chinese evaluation test set.

5.2 Results and Analysis

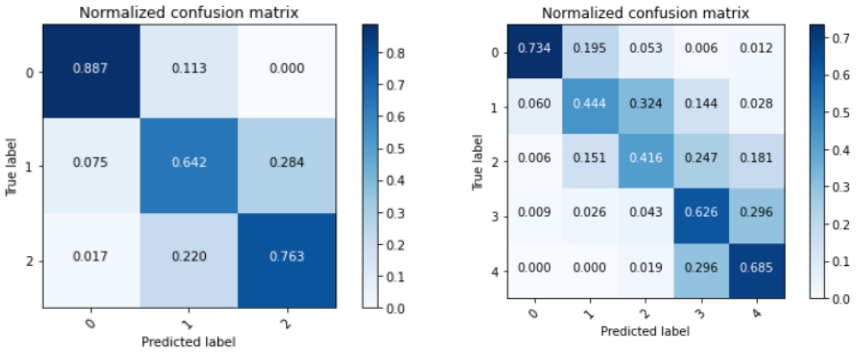
We show the experimental results of all methods in Table 2. From the table, we have the following three observations: (1) The readability formulas GFI and FRE perform the worst in both zero-shot and few-shot settings, which may result from the fact that word length is generally fixed for Chinese words, and thus is not an effective LR indicator. (2) For the better performing ADAN and XLM, the results in the few-shot setting are generally better than in the zero-shot setting. (3) XLM performs the best by 19.42 and 19.45 better than ADAN in the zero-shot and few-shot settings, respectively. The above results show that ADAN and XLM can indeed assist LR in low-resource languages. Concerning the advantage of XLM over ADAN, we speculate that XLM better captures high-level semantics like the topic and theme of the texts.

setting		Zero-shot		Few-shot	
level	data	textb	extrab	textb	extrab
3	**r**	**50.82**	**51.67**	**69.84**	61.35
3	**r+w**	43.55	51.62	68.28	**64.77**
5	**w**	36.30	N/A	60.27	N/A
5	**r+w**	**51.61**	N/A	**65.07**	N/A

Table 3: XLM evaluation results. **Bold** is the best.3 represents training on 1-3 levels and 5 represents training on 1-5 levels. r represents RAZ, w represents WeeBit. textb represents CN-textbooks and extrab represents CN-extra books.

Since this paper mainly aims to explore different transfer methods on Chinese LR, we leave the investigation of different high-level semantics to future work. To explore the impact of different datasets, we evaluate using the best performing XLM methods. Since the datasets differ in covered levels, we conduct experiments in two settings, one is based on three levels from 1 to 3 for readers aged from 6 to 11, and the other is based on five levels from 1 to 5 for readers aged from 6 to 17. As shown in Table 3, we can find that XLM trained on the RAZ performs the best in both settings, indicating that RAZ has greater guiding significance for cross-lingual LR.

As shown in Figure 2, the overall experimental results show a clear trend, the results on the edge level are better than those on the middle level. We speculate that one of the reason of the diversion may be due to the fact that not all textbooks of one grade have the same difficulty. Specifically, RAZ covers all aspects of human geography, cognition, fairy tales, legends and novels, which may assist LR regarding the difference in theme. In the future work, it is beneficial to analyze the impact of different text types on LR and consider combining vocabulary, grammar, and other relevant information, which will provides better guidance for cross-lingual LR. In addition to improving the quality of the corpus and expanding the corpus, we can explore more low-resource and cross-lingual methods to apply to our tasks in the future. Furthermore, maybe we can add some additional knowledge about LR like vocabulary difficulty and topic



(a) few shot based on RAZ

(b) few shot based on RAZ and WeeBit

Figure 2: Confusion matrix.

information to our model.

6 Conclusion

In our work, we explore two methods to tackle Chinese LR using deep neural networks without any extra features, the adversarial training model ADAN and cross-lingual pre-trained language model XLM. We organize and re-classify the LR datasets, including three LR corpora for English, and a variety of textbooks across 12 age levels and extracurricular books recommended in Chinese. To the best of our knowledge, this is the first attempt to integrate different corpora and leverage neural language models for cross-lingual LR. Experimental results show that cross-lingual Language model is more effective, and we can leverage only English corpus to predict the reading level of Chinese text, which solves the insufficient data problem in the low-resource Chinese language. After the summary of our experiment, there are still some flaws in both our datasets and methods, we have suggested some directions for future development.

Acknowledgments

This work was partially supported by National Key Research and Development Project (2019YFB1704002) and National Natural Science Foundation of China (61876009). The corresponding author of this paper is Sujian Li.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Jean-Charles Chebat, Claire Gelinas-Chebat, Sabrina Hombourger, and Arch G Woodside. 2003. Testing consumers' motivation and linguistic ability as moderators of advertising readability. *Psychology & Marketing*, 20(7):599–624.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Q. Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Trans. Assoc. Comput. Linguistics*, 6:557–570.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

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

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37, pages 748–756.

Robert Gunning. 1952. technique of clear writing.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*.

Colleen Lennon and Hal Burdick. 2004. The lexile framework as an approach for reading measurement and success. *electronic publication on www. lexile. com*.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1253–1256.

Hao Liu, Si Li, Jianbo Zhao, Zuyi Bao, and Xiaopeng Bai. 2017. Chinese teaching material readability assessment with contextual information. In *2017 International Conference on Asian Language Processing (IALP)*, pages 66–69. IEEE.

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

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2019. Supervised and unsupervised neural approaches to text readability. *arXiv preprint arXiv:1907.11779*.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Jennifer Petkovic, Jonathan Epstein, Rachelle Buchbinder, Vivian Welch, Tamara Rader, Anne Lyddiatt, Rosemary Clerehan, Robin Christensen, Annelies Boonen, Niti Goel, et al. 2015. Toward ensuring health equity: readability and cultural equivalence of omeract patient-reported outcome measures. *The Journal of rheumatology*, 42(12):2448–2459.

QianLei. 2015. The first teaching method of primary school chinese children's literature published in china[j]. In *Publisher*, page 000(010): p.8.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.*, 65:569–631.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Yuxuan Sun, Keying Chen, Lin Sun, and Chenlu Hu. 2020. Attention-based deep learning model for text readability evaluation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

KJ Topping, Jay Samuels, and Terry Paul. 2008. Independent reading: the relationship of challenge, nonfiction and gender to achievement. *British Educational Research Journal*, 34(4):505–524.

Sowmya Vajjala and Ivana Lučić. 2018a. 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 Ivana Lučić. 2018b. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. *arXiv preprint arXiv:1906.07580*.

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.