

with different levels. Therefore, we are currently expanding CEFR-SP to make it parallel through the manual rewriting of sentences in the corpus. Our sentence-level assessment model helps this process. We can complement sentences of scarce levels by adding additional rewriting tasks.

We suspect that the proposed method is directly applicable to other label-imbalanced classification problems. The empirical investigation of this is out of the scope of the present paper and is left for future work.

Ethics Statement

Ethics in Annotation Process The sentences in CEFR-SP were sampled from Newsela-Auto (news articles), Wiki-Auto (Wikipedia articles), and SCoRE (sentences written for an educational application of an academic project). We believe them to be free from harmful content that insults annotators.

We contracted with a commercial company that provides data annotation services for academia, including the management of annotators. We paid annotators \$0.50 per sentence, *i.e.*, approximately \$44/h. This was significantly higher than the minimum wage in the place where this study was conducted, reflecting our respect for the expertise required.

License Compliance We comply with the licenses of the original data sources of CEFR-SP. Specifically, we separate CEFR-SP sentences by data source and distribute them with the same license as the original datasets from which they were sampled.

Wiki-Auto CC BY-SA 3.0

SCoRE CC BY-NC-SA 4.0

Newsela-Auto We ask people first to obtain Newsela corpus (<https://newsela.com/data/>) and then contact us, following the distribution policy of Newsela-Auto.

For the reproducibility of the study, the training-, validation-, and test-set splits are maintained.

Acknowledgements

We appreciate the anonymous reviewers for their insightful comments and suggestions to improve the paper. This work was supported by JSPS KAKENHI Grant Number JP21H03564.

References

- Sweta Agrawal, Weijia Xu, and Marine Carpuat. 2021. [A non-autoregressive edit-based approach to controllable text simplification](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 3757–3769.
- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. [Optuna: A next-generation hyperparameter optimization framework](#). In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2623–2631.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. [The \(un\)suitability of automatic evaluation metrics for text simplification](#). *Computational Linguistics*, 47(4):861–889.
- Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. [Assessing relative sentence complexity using an incremental CCG parser](#). In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1051–1057.
- Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. [Multiattentive recurrent neural network architecture for multilingual readability assessment](#). *Transactions of the Association of Computational Linguistics (TACL)*, 7:421–436.
- Roger Bakeman and John M. Gottman. 1997. *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press.
- Dominique Brunato, Lorenzo De Mattei, Felice Dell’Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. [Is this sentence difficult? Do you agree?](#) In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2690–2699.
- Kiyomi Chujo, Kathryn Oghigan, and Shiro Akasegawa. 2015. [A corpus and grammatical browsing system for remedial EFL learners](#). In Agnieszka Leńko-Szymańska and Alex Boulton, editors, *Multiple Affordances of Language Corpora for Data-driven Learning*, Studies in Corpus Linguistics 69, pages 109–128. John Benjamins.
- Kevyn Collins-Thompson and James P. Callan. 2004. [A language modeling approach to predicting reading difficulty](#). In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 193–200.
- Alexis Conneau and Guillaume Lample. 2019. [Cross-lingual language model pretraining](#). In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*.

- Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.
- 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 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 Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 1–17.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding.** In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.
- Evelyn Fix and J. L. Hodges, Jr. 1989. Discriminatory analysis - Nonparametric discrimination: Consistency properties. *International Statistical Review*, 57(3):238–247.
- Thomas François. 2009. Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. In *Proceedings of the Student Research Workshop at EACL*, pages 19–27.
- Cristina Garbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. 2021. **Explainable prediction of text complexity: The missing preliminaries for text simplification.** In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1086–1097.
- 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 *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 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 Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 71–79.
- David M. Howcroft and Vera Demberg. 2017. **Psycholinguistic models of sentence processing improve sentence readability ranking.** In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 958–968.
- Joseph Marvin Imperial. 2021. **BERT embeddings for automatic readability assessment.** In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 611–618.
- Yasutake Ishii and Yukio Tono. 2018. Investigating Japanese EFL learners’ overuse/underuse of English grammar categories and their relevance to CEFR levels. In *Proceedings of Asia Pacific Corpus Linguistics Conference*, pages 160–165.
- Patrick Jacob and Alexandra Uitdenbogerd. 2019. **Readability of Twitter tweets for second language learners.** In *Proceedings of the Annual Workshop of the Australasian Language Technology Association*, pages 19–27.
- Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. **Neural CRF model for sentence alignment in text simplification.** In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7943–7960.
- Rohit Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond Mooney, Salim Roukos, and Chris Welty. 2010. **Learning to predict readability using diverse linguistic features.** In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 546–554.
- Peter J. 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 56, Institute for Simulation and Training.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. **RoBERTa: A robustly optimized BERT pretraining approach.** *arXiv*, 1907.11692.
- Ilya Loshchilov and Frank Hutter. 2019. **Decoupled weight decay regularization.** In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. **Supervised and unsupervised neural approaches to text readability.** *Computational Linguistics*, 47(1):141–179.
- Ryan McDonald, Joakim Nivre, Yvonne Quirkbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. **Universal Dependency annotation for multilingual parsing.** In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 92–97.

- Changping Meng, Muham Chen, Jie Mao, and Jennifer Neville. 2020. *ReadNet: A hierarchical transformer framework for web article readability analysis*. In *Proceedings of European Conference on IR Research (ECIR)*, pages 33–49.
- Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. *Subjective assessment of text complexity: A dataset for German language*. *arXiv*, 1904.07733.
- Akifumi Nakamachi, Tomoyuki Kajiwara, and Yuki Arase. 2020. *Text simplification with reinforcement learning using supervised rewards on grammaticality, meaning preservation, and simplicity*. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 153–159.
- Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. *Controllable text simplification with lexical constraint loss*. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266.
- Toshiaki Ozasa, George R. S. Weir, and Masayasu Fukui. 2007. *Measuring readability for Japanese learners of English*. In *Proceedings of the Conference of Pan-Pacific Association of Applied Linguistics (PAAL)*, pages 122–125.
- Gustavo Henrique Paetzold. 2016. *Lexical Simplification for Non-Native English Speakers*. Ph.D. thesis, University of Sheffield.
- David Pellow and Maxine Eskenazi. 2014. *An open corpus of everyday documents for simplification tasks*. In *Proceedings of the Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 84–93.
- Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: A corpus analysis. In *Proceedings of Speech and Language Technology in Education (SLaTE)*, pages 69–72.
- Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. *Rule-based and machine learning approaches for second language sentence-level readability*. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 174–184.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. *Stanza: A python natural language processing toolkit for many human languages*. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 101–108.
- Taraka Rama and Sowmya Vajjala. 2021. Are pre-trained text representations useful for multilingual and multi-dimensional language proficiency modeling? *arXiv*, 2102.12971.
- Nils Reimers and Iryna Gurevych. 2019. *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Angeliki Salamoura and Nick Saville. 2010. Exemplifying the CEFR : Criterial features of written learner English from the English profile programme. In Inge Bartning, Maisa Martin, and Ineke Vedder, editors, *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, Eurosla Monographs Series 1, pages 101–132. Eurosla.
- Carolina Scarton and Lucia Specia. 2018. *Learning simplifications for specific target audiences*. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 712–718.
- Jake Snell, Kevin Swersky, and Richard Zemel. 2017. *Prototypical networks for few-shot learning*. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*.
- Sanja Stajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. *Automatic assessment of absolute sentence complexity*. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4096–4102.
- Sanja Štajner, Maja Popović, Horacio Saggion, Lucia Specia, and Mark Fishel. 2016. *Shared task on quality assessment for text simplification*. In *Proceedings of Workshop & Shared Task on Quality Assessment for Text Simplification (QATS)*, pages 22–31.
- Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. *Hierarchical attention prototypical networks for few-shot text classification*. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 476–485.
- Satoru Uchida and Masashi Negishi. 2018. *Assigning CEFR-J levels to English texts based on textual features*. In *Proceedings of Proceedings of Asia Pacific Corpus Linguistics Conference*, pages 463–467.
- Sowmya Vajjala and Ivana Lučić. 2018. *On eStopEnglish corpus: A new corpus for automatic readability assessment and text simplification*. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 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 Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 163–173.