Automatic Language Proficiency Assessment of Written Texts: Training a CEFR classifier in L2-Finnish

Jenny Tarvainen
University of Jyväskylä, Finland
jenny.h.tarvainen@jyu.fi

Ida Toivanen
University of Jyväskylä, Finland
ida.m.toivanen@jyu.fi

Ari Huhta
University of Jyväskylä, Finland
ari.huhta@jyu.fi

Presentation Abstract

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), is a framework commonly used to assess the proficiency level of language learners (e.g. Martiyk & Noijons, 2007). It is also utilized for the assessment of language proficiency for citizenship purposes in Finland (Rocca et al.., 2020). To study the suitability of using a deep learning model for the CEFR classification task, we develop and present a language proficiency classifier for Finnish as a second language (F2) written texts. The classifier has been trained to recognize the six Common European Framework of Reference (CEFR) proficiency levels from A1 (basic user) to C2 (proficient user). During the development process, we seek answers to the following questions:

1. Is there enough Finnish learner language data for training deep learning models?

2. Training with existing datasets, how well can a deep learning model detect different CEFR levels?

3. How does the model compare to other CEFR-models?

The FinBERT (Virtanen et al., 2019) language model has been further trained with the datasets of (1) the International Corpus of Learner Finnish (ICLFI), (2) The Advanced Finnish Learner's Corpus (LAS2), (3) subcorpus of young Finnish learners in the Cefling project, and (4) The Finnish Subcorpus of Topling - Paths in Second Language Acquisition. These datasets provide a volume of c. 8000 texts, combining to c. 1.5 million tokens.

After training of models, the best accuracy we obtain is a score of 76.8 %. Accuracy is calculated by dividing the number of correctly classified samples by the total number of samples (see, e.g., Tharwat, 2020), so for instance, 100 % accuracy would mean that the model classified all the data samples correctly. The results indicate that there is room for improvement in model performance and a need for more CEFR-annotated Finnish learner language training data. However, the wrong classifications were mostly only off by one proficiency level (e.g., if the annotation should have been A2, the classifier should label it as A1 or B1). One should also keep in mind that even human assessors do not always agree, and a similar one-off phenomenon could also present itself (see, e.g. Yancey et al., 2023).

Keyword: Automatic Writing Assessment, CEFR, Classifier, L2 Proficiency, LLM

REFERENCES

Martinyuk, W. & Noijons, J. (2007). The use of the CEFR at national level in the Council of Europe Member States. *The Common European Framework of Reference for Languges (CEFR) and the development of language policies: challenges and responsibilities, Strasbourg, 6-8 February 2007.* Council of Europe.

Rocca, L., Carlsen, C. H, & Deygers, B. (2020). *Linguistic Integration of adult migrants: requirements and learning opportunities. Report on the 2018 Council of Europe and ALTE survey on language and knowledge of society policies for migrants.* Council of Europe.

Tharwat, A. (2020). Classification assessment methods. *Applied computing and informatics 17*(1),168-192. DOI:10.1016/j.aci.2018.08.003

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint* arXiv:1912.07076.

Yancey, K., Laflair, P., Verardi, G. A., & Burstein, J. (2023). Rating Short L2 Essays on the CEFR Scale With GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576-584). Association for Computational Linguistics.