MEASUREMENT, EVALUATION, AND DATA SCIENCE

Automated Essay Scoring for Multiple Languages: An Application of Large Language Models

Tahereh Firoozi, Mark J. Gierl

Centre for Research in Applied Measurement and Evaluation (CRAME)

UNIVERSITY OF ALBERTA

What is Automated Essay Scoring?

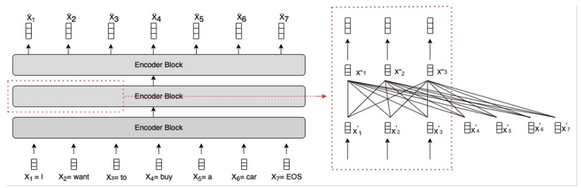
- An automated essay scoring (AES) system is a computer program designed to evaluate student responses so that the program yields test scores that are similar to those of trained human raters (Shermis, 2014).
- AES is a classification method where linguistic input in the text is mapped to specific output like an essay score so that the input and output are related to one another statistically.
- The mapping function is called a scoring model and it can be used to classify new instances of the input text onto the output score. The use of a scoring model allows educators to scale the assessment because, instead of a human, the computer can be used to score students' written tasks.
- To emulate human scoring, the AES program builds the model using techniques and procedures from the fields of natural language processing (NLP) and computational linguistics where features are extracted from the example instances—called the training dataset—that have been scored by human raters.

Multilingual AES

- The vast majority of the published AES studies have focused on essays written in English.
- The remaining languages, also know as low-resource languages, are less studied, less digitized, less privileged, less commonly taught, and less accessible compared to English (Magueresse et al., 2019).
- Studies that implement AES using low-resource languages are rare.
- Multilingual AES refers to AES as it applies to languages other than English. Multilingual AES is a critically important research area because the language of assessment for many students is not English.
- In this study, we describe and evaluate a multilingual transformer-based AES system for scoring essays using four different languages, including Persian, Italian, Czech and German.

Transformer-Based AES System

- Transformers are encoder-decoder-based neural networks used to solve sequence-to-sequence problems by finding a mapping function f from an input sequence (e.g., word or sentence) of n vectors $X_{1:n}$to a sequence of m target vectors $Y_{1:m}$.
- The application of transformers in NLP tasks was accelerated with the advent of a specific transformer-based language model called the Bidirectional Encoder Representation for Transformers or BERT (Devlin et al., 2018).
- We use the multilingual version of BERT—mBERT—to score essays written in Persian. mBERT is a transformer-based approach created by the Google AI Research team.

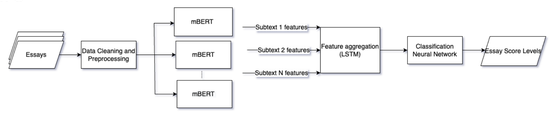


Method

Dataset

- The data in this study consisted of 2000 essays written in Persian, 1029 essays in German, 804 essays in Italian, and 434 essays in Czech.
- The training set consisted of 60% of the data. The remaining 40% was used as the testing set.

Architecture



Analysis

- The data were imported and read in Google Colab Pro and then Python 2.10.0 was used for data pre-processing.
- The essay scores were predicted using the above model consisting of mBERT, a recurrent neural layer, and a classifier.

Results

- The mBERT model performed with high classification consistency (QWK=0.84 vs. Baseline QWK=0.75; Kappa = 0.93 vs. Baseline Kappa = 0.82) reaching a level described by Koch and Landis (1977) as "almost perfect agreement".

Accuracy of the Transformer mBERT Model

Score Level	Precision	Recall	F1-Score
Elementary	82%	86%	84%
Pre-Intermediate	78%	75%	76%
Intermediate	70%	72%	71%
Upper-Intermediate	73%	70%	71%
Advanced	60%	64%	62%

- The result from the accuracy measures shows that 73% of the total number of the essays were correctly scored by the mBERT model.
- The Precision measure reveals that among all the essays assigned to each levels by the mBERT model, 70% or more of the essays were correctly classified.

Future Work

- The advancement in the large language models (LLMs), such as GPT and LLAMA, and PALM, paved the way for developing automated feedback models that can generate personalized feedback on student writings.

References

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. Assessing Writing, 20, 53-76.

Magueresse, A., Carles, V., & Heetderks, E. (2019). Low-resource languages: A review of past work and future challenges. arXiv preprint arXiv:2006.07264.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159-174.