Automated Essay Scoring in Foreign Language Students Based on Deep Contextualised Word Representations

Bojana Ranković
EPFL (Switzerland)
bojana.rankovic@epfl.ch

Sarah Smirnow
University of Zurich (Switzerland)
sarah.smirnow@ibe.uzh.ch

Martin Jaggi
EPFL (Switzerland)
martin.jaggi@epfl.ch

Martin J. Tomasik
University of Zurich (Switzerland) and University of Witten-Herdecke (Germany)
martin.tomasik@ibe.uzh.ch

ABSTRACT: We introduce a method for automated grading of handwritten essays written by foreign language learners of French. The handwriting recognition system allows digitising the essays for further processing and functions at a low character error rate. The transcriptions are then vectorised using embeddings from state-of-the-art pre-trained natural language processing models. On top of the extracted word-level features, a deep recurrent network was trained for grade predictions for essays, using the nine different grading criteria as target variables. Scores on these criteria were previously obtained from human expert raters for more than 6'000 student essays. We present preliminary findings on prediction accuracy and discuss possible future developments and applications of the system.

Keywords: Assessing writing, automated grading, handwriting recognition, natural language processing

1 BACKGROUND

Providing students, teachers and schools with objective and reliable information about students' writing competency as well as providing an evaluation related to their reference groups has become particularly relevant. This is in response to the so-called PISA shock in 2000, when, against all expectations, Switzerland was ranked just above average (e.g. Buschor, Gilomen, & McClusky, 2003). This disappointing result spurred efforts to improve Switzerland's educational system. One of the measures implemented was regular assessments of students' competencies. In our presentation, we will focus on the assessment of writing competencies in an initiative of four German-speaking cantons. More specifically, we will outline the assessment and scoring of handwritten, paper-based writing assignments in French as a foreign language as part of a set of compulsory standardised, large-scale assessments that are administered in grade eight (*N* > 12'000). We will explore digital automated scoring to better understand text features that differentiate between competence levels and to define a model for evaluating texts to a specific prompt to support human raters.

2 WRITING ASSESSMENTS

Each test consists of two open writing tasks that require students to write different types of texts (e.g. letters, messages, stories, reports) in French. These texts are analytically scored by expert readers, using a standardised grid in which different elements of the text are evaluated according to verbal gradations (Weigle, 2002). Essays are rated against nine criteria within two dimensions. The content dimension is operationalised within five criteria (task fulfillment, comprehensibility, creativity, coherence and greetings in the case of letter writing), and the language dimension is operationalised within four criteria (syntax, linguistic range, and grammatical and orthographic competence). The selection of these criteria is guided by the communicative and linguistic abilities reflected in the written product (e.g. CEFR, 2001). A major challenge in testing writing competences with open tasks is the time and expense needed for scoring (Page, 1968). Training for raters takes place during the first two days of a scoring period, and various procedures are used throughout the whole period to ensure accurate and consistent scoring. After this extensive training, raters achieve interrater consistencies of $.96 < r_{\text{ICC}} < .97$ (computed as one-way, multiple rater consistency), according to McGraw and Wong (1996), depending on the respective dimension.

3 TECHNOLOGICAL INNOVATION

Our proposed system consists of two phases. The first is digital handwritten text recognition (HTR) of more than 6’000 essays, and the second is the prediction of the annotated scores and the highlighting of interpretable features (such as keywords or other patterns) that contribute and correlate with each score or level of competence.

3.1 Handwriting recognition

In an interdisciplinary effort, the handwritten texts were digitised in order to apply natural language processing (NLP) techniques to analyse the prescored essays on each dimension. The HTR system used a neural network architecture with convolutional and recurrent layers to achieve a reliable performance of 8% character error rate on average. Despite the high heterogeneity in students’ handwriting, this error rate is comparable to state-of-the-art methods in the field (e.g. Slimane, Mazzei, Topalov, Verzi, & Kaplan, 2017). Taken together, this initial part of the study served to transcribe handwritten essays into a digitised form for further evaluation using NLP.

3.2 Essay Feature Representation

In the main part of the study, the resulting transcribed texts from the HTR phase were vectorised using embeddings from state-of-the-art pretrained NLP models (BERT; Devlin et al., 2018). The model was chosen for being capable of modeling semantic and syntactic characteristics of word use while also being able to distinguish between different linguistic contexts and thereby successfully modeling polysemy. We employed a model pre-trained on French (CamemBERT; Martin et al., 2019). On top of the extracted word-level features, a deep recurrent network was trained for predicting grades on entire essays, using the nine different grading criteria as target variables. The initial training was carried out using a subset of 100 essays split into test and train examples through 5-fold cross-validation.

3.3 Representation Depth and Scoring Predictability

The results are satisfying on all of the nine criteria showing that the system can be used to support human raters. An example is presented in Table 1. Here we can also observe how different encoding layers of BERT architecture influence prediction results. The numbers show mean squared error evaluated on separate cross-validation folds. Lower-level encoding layers are able to capture syntactic rules which, in turn, enables the model to achieve lower mean squared error using these layers. In the case of originality, however, higher-level layers capture semantics and present a better choice for this content dimension.

	Bert encoding layers		
Criteria	Low-level	Middle-level	Top-level
Syntax	0.39±0.02	0.35±0.01	0.45±0.04
Originality	0.55±0.08	0.53±0.03	0.49±0.04

Table 1: MSE results for different choice of layers

4 SUMMARY AND OUTLOOK

In our presentation, we will discuss the detailed prediction accuracies in all criteria, as well as the features that were identified as the most predictive for human ratings and discuss future developments of the system. The reliable HTR system, in combination with the automated evaluation of an essay for different criteria makes it possible to develop systems that not only grade students' essays but also provide formative feedback, thereby helping to significantly improve writing style (e.g. Wingate, 2010, Malik et al., 2019). We can assume that such systems would be particularly effective if embedded in more comprehensive feedback or tutoring systems that collect and process broader data on students' competencies. One could also imagine teachers using end-to-end systems on handheld devices that can support them in evaluation and feedback. Future research is needed to test how well the models generalise to (a) different essay topics and (b) to different languages.

REFERENCES

Buschor, E., Gilomen, H., & McCluskey, H. (2003). *PISA 2000: Synthese und Empfehlungen.* Neuchâtel, Switzerland: Bundesamt für Statistik.

CEFR, Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching and assessment.* Cambridge, England: Cambridge University Press.

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*.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., ... & Sagot, B. (2019). CamemBERT: a Tasty French Language Model. *arXiv preprint arXiv:1911.03894*.

Page, E. B. (1968). The use of computers in analyzing student essays. *International Review of Education, 14*, 201–225.

Slimane, F., Mazzei, A., Topalov, O., Verzi, G., & Kaplan, F. (2017). A web-based tool for segmentation and automatic transcription of historical documents. *Proceedings of the 2017 International Joint Conference on Neural Networks*, 2730–2737.

Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Wingate, U. (2010). The impact of formative feedback on the development of academic writing. *Assessment and Evaluation in Higher Education, 35*, 519–533.