Writing Automarker Evaluation

1. Task Overview

This task asks you to evaluate the performance of a bespoke writing automarker.

More specifically, you are asked to use the provided **evaluation dataset** to undertake a **high-level analysis** of this automarker and make some **summary recommendations**.

The focus is on the automarker's scoring accuracy against a set of "ground truth" human scores, as generated for a set of writing produced by students learning English as a second language. The resulting accuracy can be seen as one evaluation gate for deciding if the automarker is good enough to be used in a live writing exam.

A simple approach here might involve calculating a basic set of machine learning and agreement statistics for the scores as a whole (e.g. overall RMSE). However, you are encouraged to explore a more nuanced evaluation, to the extent permitted by the data.

Importantly, there is no predefined benchmark or single correct result for the task. Rather, we are interested in your overall approach, clarity of reporting, and ability to thoughtfully engage with the data. In this sense, the limited nature of the dataset is itself intended as a prompt to consider what else is needed to better evidence the automarker.

Evaluation Criteria

- **Report Quality:** Clarity, structure, and suitability of recommendations
- **Analysis Approach:** Appropriacy of metrics and method
- **Conceptual Understanding:** Ability to explain decisions and thoughtfully engage in discussions relevant to our modelling context

Submitting your Solution

Once you have completed the task, please submit both **your report** and any **associated code** by uploading it to the same **Kiteworks** folder that you used to access the original material. We specifically ask that you submit:

1. Your report in pdf form.
2. Your code as a compressed package that can be run without any modifications.
    a. Please ensure that any required library packages are clearly specified
    b. Please include a clear README.md file explaining how to set up and run your code

If you have any questions at all regarding the submission process or the task, please do not hesitate to reach out before the deadline.

2. Task Guidelines

Coding

- Please use **Python or R** – either is fine.
- Please also note:
  - You are welcome to use any relevant libraries, public tools, &c that you have access to
  - Although not required, it is preferable that any code be managed via Git

Report

- Please submit a **1–2-page draft report** summarising:
  a. Your high-level analysis of the automarker's **scoring accuracy**, detailing any metrics and notable methodological decisions.
  b. Your **overall recommendation** as to whether the automarker is suitable for use in a live exam context, indicating your reasoning. For context, you should assume that this automarker would be used in a "high stakes" writing exam, such as for university entrance or a visa application.
  c. Any **key limitations** of the evaluation that you have identified, indicating how you would address these limitations in order to better evidence the quality of the automarker. In doing so, please indicate at least one way in which the data should be improved and one additional study that you would recommend conducting.
- Please also note that
  - The suggested length **does not include any tables or plots** – you may include as many of these as you wish. You may also produce a longer report if you think this is necessary to justify your analysis.
  - The report itself **is not expected to be a thorough formal report.** For example, it is fine to use bullet points, rather than extended paragraphs and extensive formatting. The key consideration is that the structure and analysis is clear and easy to follow.

Follow Up Discussion (During the Interview)

You should expect that a substantive portion of the interview will provide the opportunity to talk in greater depth about your evaluation and your thoughts around it. To enable this, please:

- Have your code ready for demonstration and discussion
- Be prepared to discuss your report and its contents.

3. Dataset Information

The evaluation dataset is a modified subset of the Write & Improve Corpus (W&I2024), a corpus of scored writing produced by learners of English as a second language.

For the present task, we have made the following modifications:

- converted the original scoring categories to an equivalent 0-11 **raw score** scale;
- mapped the scores to **CEFR levels**. These can be understood as standardised categories that represent a learner's overall level of proficiency in English;
- created a hypothetical mapping from the raw scores to a binary **pass/fail** outcome. This can be understood as representing whether or not the learner would have passed the exam, had such an outcome been possible;
- generated a new set of **automarker scores**. These scores are generated by a completely bespoke machine learning model which has been trained to predict writing scores using learner writing from the same domain. The automarker is a BERT-based model that incorporates a separate regression head architecture. This architecture incorporates a number of linguistic features designed to explicitly capture various aspects of the writing construct which the automarker is expected to measure (e.g. VOCD as a measure of lexical diversity).

Please note that the full W&I2024 is available for educational research purposes. However, access to this specific version is confined to yourself & limited to the analysis necessary for this task. You do not therefore have permission to share the data beyond yourself. You must also delete any data or derived materials on completion of the interview process.[1]

Provided Files

To enable your analysis, you are provided with the following files:

- An **evaluation_set.csv** file. This is the core dataset containing the automarker scores to be analysed.
- A **score_mappings.csv** file. This contains the mappings from raw scores to the two types of categorial outcome mentioned above: the resulting **CEFR** levels and **pass/fail** outcomes.
- A **train_dev_set.csv** file. This comprises the underlying set of scores for the dataset used to train the automarker used here.

[1] You are still very welcome to make subsequent use of any corpus data, however. You simply need to request independent access under the general licensing conditions. This can be done via the website, here.

Metadata Information

***** Indicates information specifically created for the interview task and which is therefore not part of the official CUP&A Write & Improve Corpus***

VARIABLE	DESCRIPTION
public_essay_id	The unique ID for the learner's written response. Each piece of writing was produced in response to 1 of a possible 50 distinct writing prompts.
split	What type of dataset a response was sampled into at training: **train, dev, test.** Automarker scores are not provided for either the **train** or **dev** set.
language	The first language of the learner e.g. "Arabic", "Spanish".
**** rater_score**	The original raw scores assigned by the human raters, as mapped here to a 0-11 scale. Each score represents the score of a single rater. These constitute the ground truth scores for the evaluation.
**** rater_cefr_level**	The **CEFR level** achieved by the learner according to the human rater, derived from the **rater_score**. CEFR levels are a categorial measure of language proficiency, often used for reporting results. Please see the **score_mappings.csv** file for the mapping.
**** rater_pass_fail**	Whether the candidate "passed" or "failed" according to the human rater, again derived from the **rater_score**. In an exam context, this is often the key reason why a learner will be taking a particular test. Again, please see the **score_mappings.csv** file for the mapping. *Please note that, in the present context, this is purely a hypothetical outcome, as the notions of "pass" and "fail" do not apply to the W&I Corpus data strictly speaking.*
**** automarker_score**	The raw score produced by our bespoke automarker. This is a BERT-based model with a distinct regression head architecture. As noted above, this incorporates a number of linguistic features designed to more explicitly capture various aspects of the writing construct.
**** automarker_confidence**	An associated confidence score for our automarker. This predicts the likelihood of the candidate achieving the same CEFR level from the automarker as they did from the human rater. "1" represents maximum confidence; "0" represents the lowest level of confidence.
**** automarker_cefr_level**	The equivalent automarker CEFR level, as derived from the **automarker_score**. The mapping here is the same as for the **rater_cefr_level**. However, automarker scores are rounded before any CEFR mapping is determined
**** automarker_pass_fail**	The equivalent automarker pass-fail outcome, as derived from the **automarker_score**. Again, the mapping here is the same as for **rater_pass_fail**, with automarker scores rounded prior to mapping.