Natural Language Processing : Research Article Reproducibility

Experiments with Universal CEFR classifications

Ariel Nataf, Adnene Khalbous, Jean-Baptiste Chanier
Université de Paris, Université de Paris, Université de Paris
ariel.nataf@etu.u-paris.fr, adnene.khalbous@etu.u-paris.fr, jean-baptiste.chanier@etu.u-paris.fr

Abstract

As part of the Master 2 "*Machine Learning pour la Science des Données*" at the *Université de Paris*, our group of students was interested in studying, analysing and reproducing results from a research article published in 2018 entitled "*Experiments with Universal CEFR classifications*" (Vajjala and Rama, 2018). The paper builds on the codes used by the original authors of the paper on the github platform. It offers a solution to a classification problem known as *Automated Essay Scoring* (AES), the automatic assignment of grades to student essays based on their language proficiency. The different levels of expression are based on the *Common European Framework of Reference* (CEFR). The paper presents a comparison of the ratings obtained with three languages (German, Italian and Czech), using monolingual, cross-lingual and multi-lingual techniques.

Keywords: Machine Learning, NLP, AES, CEFR

1. INTRODUCTION

The reproducibility of a research paper is a primordial indicator of the degree of interest to be given to the results it gives. Indeed, a result can only be considered statistically significant if it can be obtained under similar experimental conditions.

This idea was brought to the *12th Edition of the Language Resources and Evaluation Conference*, or LREC, held in May 2020, through the *Shared Task on the Reproduction of Research Results in Science and Technology of Language*, or *REPROLANG 2020*. This initiative, supported by ELRA, *the European Language Resources Association*, who brought together teams seeking to replicate experiments from research publications and to record the results, thereby validating the reproducibility of the values presented to be measured.

As part of this approach, this paper presents the conclusions reached by our team in attempting to replicate the experiment presented in the article "*Experiments with Universal CEFR Classification*" (Vajjala and Rama, 2018).

Regarding the data used in the paper, we will present our remarks with respect to the FAIR principle: be Findability, Accessibility, Interoperability, and Reuse of digital assets.

We also wanted to take into consideration the degree of ease of use of the implementation of the algorithms for people outside the initial team of researchers, as well as the potential discrepancies between our results and the original values presented in the paper, along with hypotheses on the causes of occurrence of these phenomena.

The issues addressed in this paper are therefore the following: Are the results presented in the article reproducible? Does the publication meet the requirements of the FAIR paradigm?

2. STATE OF THE ART

In the paper we reviewed, the authors mentioned other work that could be related to theirs:
The publication "*Exploring CEFR classification for German based on rich linguistic modeling*" (Julia Hancke, 2013), also deals with an AES work on the MERLIN dataset, but this time only monolingual since only the German portion was selected for the study. J. Hancke and D. Meurers thus sought to quantify the contribution of features to AES work, by comparing the performances obtained with different types of variables:

- **Lexical Features**: Features related to the use of a lexicon, i.e. the set of possible words of the studied language. The authors used indicators such as the "depth of lexical knowledge", which is based on lexical frequency scores for each word, or "shallow measures", i.e. the number of textual errors in the text or the observed word length.
- **Syntactic Features**: Features based on syntax, i.e. the set of rules that describes the rules according to which linguistic units are combined into sentences. Examples include the use of statistical language models or analyses related to the use of Stanford's POS TaggerNLP
- **Morphological Features**: Features based on morphology, i.e. taking into account word shapes. For example, the authors used indicators such as "the ration of nominal suffixes", or the study of inflections (person, verb-form, mood for a verb and case for a noun). The researchers also took into account the frequency of ratio of verbal tense features

In the article "*A Neural Approach to Automated Essay Scoring*" (Taghipour and Ng, 2016), a neural net-

work based approach was experimented. The objective of the study was to create end-to-end models, thus requiring no prior feature extraction step, the models were trained directly on the raw text. The metric used is the quadratic weighted Kappa measure, which measures the agreement between the exact score and that proposed by the model. Taking into account the sequential nature of texts (ordered sequences of words), the authors chose to focus on recurring neural networks. They have chosen to experiment with many types of architectures, among which we can mention convolutional neural networks, various types of RNN (GRU, basic RNN, LSTM) bidirectional or not, attention mechanisms. The LSTM layer-based network in particular has provided a clear improvement in baseline for this procedure, as this was previously calculated using SVRs, counterpart of SVMs for regression.

In 2014, researchers Sowmya Vajjala and Kaidi Lõo published their work as the paper "*Automatic CEFR Level Prediction for Estonian Learner Text*" (Vajjala and Lõo, 2014). The authors focused mainly on morphological features and information from the use of POS taggers, obtaining a set of 78 features in total, allowing to take into account the richness of the Estonian language. Moreover, in this study, the AES work was considered first as a classification problem (obtaining an accuracy score of 79%), then as a regression problem (R^2 indicator of 0.85). The conclusion of the authors was that the consideration of the problem from a classification point of view is more efficient in terms of exact error.

3. SUMMARY OF THE STUDY

The article studied in this paper deals with a problem of classifying student texts according to the level of language proficiency demonstrated by the evaluated author. This procedure is referred to as *Automated Essay Scoring* or AES. The different possible modalities for the target variable are set out in the *Common European Framework of Reference for Languages* (CEFR) and are defined from the most basic level (A1) to the most advanced as follows (C2) (elsafrenchteacher.com, 2020) :

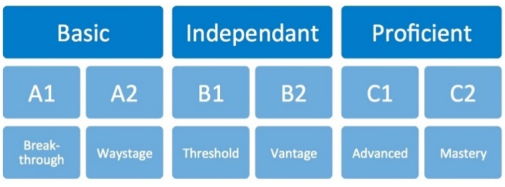

Figure 1: CERF levels

In the study, the dataset used is the MERLIN corpus (Boyd et al., 2014), gathering 2286 German, Italian or Czech texts. The authors of the paper chose to pre process the dataset (Vajjala and Rama, 2018) remove categories for a given language if they had less than 10 usable texts. After cleaning, the corpus contains only 2266 texts, and can be considered as significantly unbalanced, considering the number of representatives per language and per category of the CEFR. This information is summarized in the following table, which comes from the original article:

CEFR level	DE	IT	CZ
A1	57	29	0
A2	306	381	188
B1	331	393	165
B2	293	0	81
C1	42	0	0
Total	1029	803	434

Figure 2: CERF Results

3.1. Features

In the original paper, the AES models are trained on features extracted directly from the available corpora. The authors chose to use features specific to the language of the text, such as: n-grams of words and part-of-speech categories, embedding obtained from a single-layer neural network on the words and characters encountered, "*dependency n-grams*" which list both the preceding and following words but also their POS label. They also took into account language-independent features such as the length of the text, the richness of the lexicon used in the text and the total number of spelling errors observed.

3.2. Models

The model architectures used by the authors are the following: Logistic Regression, Support Vector Machine, RandomForest, Multi-Layer Perceptron and more complex neural networks directed towards more specific tasks. These models have been implemented in the Python language using the SciKit and Keras libraries. For the sake of conciseness, only the results concerning the best performing models have been presented, *i.e.* the RandomForest and the Logistic Regression models (referred as L in the tables).

3.3. Metrics

There is a debate in the research world about the consideration of the type of problem in SEA. Indeed, the task can be seen as either a regression or a classification problem. In the article studied, the authors have chosen to consider the problem as a classification problem, which leads to the use of specific metrics. Here, taking into account the imbalanced nature of the corpus, the authors have chosen to use the weighted F1-score, i.e. the weighted average of the F1-scores of each class according to their respective numbers.

4. RESULTS

The article presents results from three procedures:

- **Monolingual Classification**: Training, validation and testing on data corresponding to a single language. This procedure is the most classical in AES and allows to set a consistent baseline of achievable results.

  We obtain:

	DE	IT	CZ
Baseline	0.50L	.6	.61L
Word ngrams (1)	.66	.83	.73L
POS ngrams (2)	.67	.81	.70
Dep. ngrams (3)	.65	0.8	.72
Domain Features	.49L	.6	.65
Domain + (1)	.68	.82	.74
Domain + (2)	.66	.82	.70
Domain + (3)	.67	.79	.73
Word embeddings	.63	.8	.66

Table 1: Monolingual Replicated Results

- **Multilingual Classification**: Training and validation on all the data at once and then testing on the texts of each language independently.

	lang (-)	lang (+)
Baseline	.40L	
Word ngrams (1)	.72	.72L
POS ngrams (2)	.73	.73
Dep. ngrams (3)	.71	.71
Domain Features	.49L	.43L
Word + Char embeddings	.68	.68

Table 2: Multilingual Replicated Results

- **Crosslingual Classification**: Training and validation on the data of one language, then testing on the texts corresponding to another language. The model was trained on German language data because, according to the authors, "it has examples for all categories in our corpus", and then tested on Italian and Czech samples.

	IT	CZ
Baseline	.55L	.45L
POS ngrams (2)	.74	.67
Dep. ngrams (3)	.63	.64
Domain Features	.61SVC	.47

Table 3: Cross Replicated Results

For each of these procedures, the authors chose to present a baseline result, which corresponds to the use of a Logistic Regression model based only on the size of the text to be classified.

5. ISSUES

During this project we were confronted with a few problems:

- The code lacked a proper documentation and an interface to conveniently obtain the desired results from the different .py files. It required trials and error to use the proper functions and tune the different parameters. We also updated the output format to be more readable.
- The code was not formatted to be used by someone else. Path files were hardcoded and absolute. It only required a few fixes to be updated as relatives paths.
- Depreciated packages versions on python (scikit, scipy, numpy): To solve this we looked for the suitable version for each package.
- Lines of code blocking the execution because they were depreciated but not used. We went through all of a some scripts to understand how it worked and updating the lines causing problems.
- A package (a python wrapper for LanguageTool) required java but has not been updated, we used a fork working with java 8[1].

6. DISCUSSION

The results are close to what was expected from reading the original paper.
When the results are followed by a L, they are obtain using a Logistic Regression; followed by a SVC, they are from SVC ; followed by nothing they are obtained using a Random forest.
Sometimes results from two different methods are close to each other and it can lead to a best Classification method to vary from the initial paper and our reproduction.

7. CONCLUSION

During this study we tried to reproduce the approaches presented in the article "Experiments with Universal CEFR classifications" (Vajjala and Ruma, 2018). The article presented a study on the use of multilingual approaches in the context of an "Automated Essay Scoring" task, taken as a classification problem. Firstly, the access to the data respected the FAIR paradigm as the dataset was freely available and easily handled.
Furthermore, the preprocessing of the data and the extraction of features can be considered as reproducible from the codes provided by the authors of the article. Similarly, the model building and training phase does not pose any particular problem, each of the procedures (Monolingual, multilingual, crosslingual) being correctly represented. When it comes to results, slight

[1]Fork for Language Tool: language-tool-python

variations due to the differing hardware used were observed, but the scores are roughly similar to those stated in the article. In this reproduction work, the first problems we encountered concerned directly the execution of the codes since we had to recreate from scratch and without prior indication an environment with the necessary libraries in their adequate version. In a second step, we had to modify the codes slightly so that the program could run normally. We also noted that some lines, sometimes without comments, were not used and seemed to block the proper functioning of the code. As a final statement, we can conclude that the reproduction of the results of the original article is feasible with the available codes, with only slight programming adaptation required.

8. Bibliographical References

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., and Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

elsafrenchteacher.com. (2020). Cerf levels graph. `https://elsafrenchteacher.com/cefr-a1-c2-levels`.

Julia Hancke, D. M. (2013). Exploring cefr classification for german based on rich linguistic modeling. In *Conference: Learner Corpus Research 2013. Book of Abstracts*.

Taghipour, K. and Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas, November. Association for Computational Linguistics.

Vajjala, S. and Lõo, K. (2014). Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127, Uppsala, Sweden, November. LiU Electronic Press.

Vajjala, S. and Rama, T. (2018). Experiments with universal CEFR classification. *CoRR*, abs/1804.06636.

9. Language Resource References

Sowmya Vajjala and Taraka Rama. (2018). *nishkalavallabhi/UniversalCEFRScoring: Bea 2018 paper code*. Zenodo.