Automatic Assessment of English CEFR Levels Using BERT Embeddings

Veronica Juliana Schmalz[1,3], **Alessio Brutti**[1,2]
1. Free University of Bozen-Bolzano, Bolzano, Italy
2. Fondazione Bruno Kessler, Trento, Italy
3. KU Leuven, *imec* research group *itec*, Kortrijk, Belgium
veronicajuliana.schmalz@kuleuven.be, brutti@fbk.it

Abstract

The automatic assessment of language learners' competences represents an increasingly promising task thanks to recent developments in NLP and deep learning technologies. In this paper, we propose the use of neural models for classifying English written exams into one of the Common European Framework of Reference for Languages (CEFR) competence levels. We employ pre-trained Bidirectional Encoder Representations from Transformers (BERT) models which provide efficient and rapid language processing on account of attention-based mechanisms and the capacity of capturing long-range sequence features. In particular, we investigate on augmenting the original learner's text with corrections provided by an automatic tool or by human evaluators. We consider different architectures where the texts and corrections are combined at an early stage, via concatenation before the BERT network, or as late fusion of the BERT embeddings. The proposed approach is evaluated on two open-source datasets: the English First Cambridge open language Database (EFCAMDAT) and the Cambridge Learner Corpus for the First Certificate in English (CLC-FCE). The experimental results show that the proposed approach can predict the learner's competence level with remarkably high accuracy, in particular when large labelled corpora are available. In addition, we observed that augmenting the input text with corrections provides further improvement in the automatic language assessment task.

1 Introduction

Finding a system which objectively evaluates language learners' competences is a daunting task. Several aspects need to be considered, including both subjective factors, like age, native language, cognitive capacities of the learner, and learning-related factors, for example the amount and type of received linguistic input (James, 2005; Chapelle and Voss, 2008; Jang, 2017). Indeed, language competences are not holistic, but concern different domains, so that considering the mere formal correctness of learners' language has been shown not to represent a proper assessment procedure (Roever and McNamara, 2006; Harding and McNamara, 2017; Chapelle, 2017). Moreover, human evaluators, despite having to adhere to a predefined scale and guidelines, such as the CEFR (Council of Europe, 2001), have proved to be biased (Karami, 2013) and inaccurate (Figueras, 2012). For these reasons, new language testing methods and tools have been developed. Current state-of-the-art models, such as Transformers, allow to process numerous and complex linguistic data efficiently and rapidly, by means of attention-based mechanisms and deep neural networks that capture the relevant features for the targeted task. However, the creation and access to necessary language examination resources including annotations and metadata appear to date limited. In this paper, we propose using a series of BERT-base models to automatically assign CEFR levels to language learners' exams.

Our aim is examining the possibility of providing the system with previously generated corrections, either by humans or automatically with a language checker. Additionally, we want to analyse the impact of the amount of data on the accuracy of the model in the classification of written exams taken from the English First Cambridge Open Language Database (EFCAMDAT)

(Geertzen et al., 2013) and the Cambridge Learner Corpus for the First Certificate in English (CLC-FCE) (Yannakoudakis et al., 2011). In this way, a significant turning point could be made both in improving the functioning of these automatic systems and in the future collection of data from other languages.

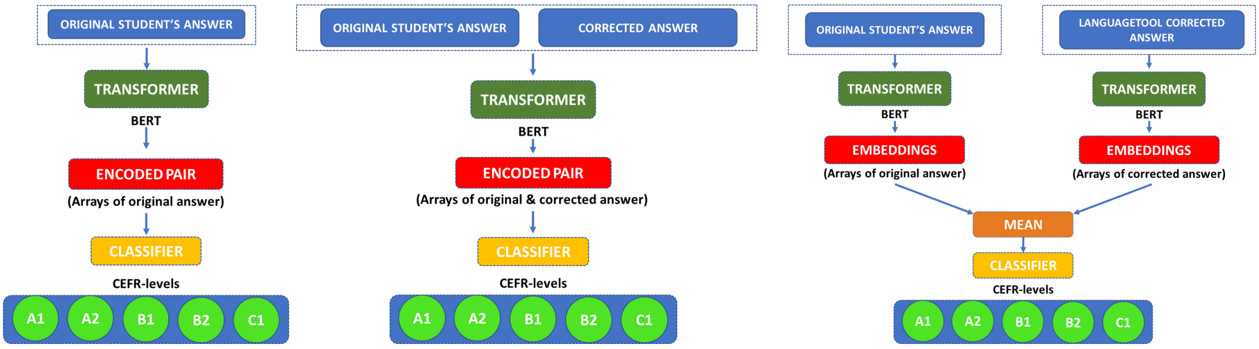
2 Related Works

Automatic language assessment methods concern the creation of fast, effective, unbiased and cross-linguistically valid systems that can both simplify assessment and render it objective. However, achieving such results represents a complex task that researchers have been addressing for years while experimenting with several methodologies and techniques. The first developed tools used to mainly deal with written texts and exploited Parts-of-Speech (PoS) tagging to grade students' essays (Burstein et al., 2013), and latent semantic analysis to evaluate the content, providing also short feedback (Landauer, 2003). Advances in AI, NLP and Automatic Speech Recognition (ASR) led to the additional emergence of systems that assess spoken language skills, such as the *SpeechRater* (Xi et al., 2008), which considers clarity of expression, pronunciation and fluency. To date, several other automatic language assessment tools are applied in the domain of large scale testing, for example *Criterion* (Attali, 2004), *Project Essay Grade* (Wilson and Roscoe, 2020), *MyAccess!* (Chen and Cheng, 2008) and *Pigai* (Zhu, 2019). The first can detect grammatical and usage-based errors, as well as punctuation mistakes, providing also feedback. However, it requires being trained on the specific topics to assess. The second system exploits a training set of human-scored essays to score unseen texts, evaluating diction, grammar and complexity from statistical and linguistic models. Similarly, *MyAccess!*, calibrated with a large number of essays, can score learners' texts and measure advanced features such as syntactic and lexical complexity, content development and word choice, providing detailed feedback. On the contrary, *Pigai*, exploits NLP to compare the essays submitted by students with those contained in its corpora, measuring the distance between the two (Zhu, 2019). Despite the extreme efficiency of these tools, to perform accurately they generally need large amounts of labelled and human-corrected training data. Furthermore, a standard scale is needed, which can be extended between different groups of learners. In addition, powerful computational resources, and in certain cases, significant memory, are required. All these elements together constitute fundamental pre-requisites which can be difficultly fulfilled. For this reason, we present a distinct approach to the previous ones which, starting from different amounts of students' original texts, provides a classification within the different CEFR levels exploiting BERT-base models and subsidiary corrections.

3 Proposed Approach

The approach we propose for the automatic assessment of the language competences of adult English language learners is based on the use of Transformer-type architectures performing multi-class classification. Among these, BERT-based models, characterised by efficient parallel training and the capacity of capturing long-range sequence features, distinguish themselves for their size and amount of training data (Vaswani et al., 2017). Being pre-trained on generic large corpora, with Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) strategies, they can be conveniently employed in a wide range of tasks, including text classification, language understanding and machine translation.

The models we use for our experiments are grounded on the *BERT-base-uncased* architecture, part of the Hugging Face Transformers Library released in 2019 (Wolf et al., 2020) and inspired by BERT (Devlin et al., 2018) from Google Research, that encodes input texts into low-dimensional embeddings. Our baseline model maps these compact representations into the CEFR levels using a network with two fully connected layers. Fig. 1(a) graphically represents the architecture. Note that this approach requires training the final classifier only. Retraining or fine-tuning the BERT model would probably require very large datasets which are not always available for this task. In order to augment the input text with corrections (either automatic or human) we investigate two possible directions. The first one (Fig. 1(b)) concatenates the two texts and applies the pre-trained BERT model. The resulting embeddings are expected to encode the information related to both texts. Conversely, the second architecture extracts individual embeddings for the original texts and the corrected ones.


ORIGINAL STUDENT'S ANSWER
TRANSFORMER
BERT
ENCODED PAIR
(Arrays of original answer)
CLASSIFIER
CEFR-levels
A1
A2
B1
B2
C1
ORIGINAL STUDENT'S ANSWER
CORRECTED ANSWER
TRANSFORMER
BERT
ENCODED PAIR
(Arrays of original & corrected answer)
CLASSIFIER
CEFR-levels
A1
A2
B1
B2
C1
ORIGINAL STUDENT'S ANSWER
LANGUAGETOOL CORRECTED ANSWER
TRANSFORMER
BERT
EMBEDDINGS
(Arrays of original answer)
TRANSFORMER
BERT
EMBEDDINGS
(Arrays of corrected answer)
MEAN
CLASSIFIER
CEFR-levels
A1
A2
B1
B2
C1

(a) (b) (c)

Figure 1: Proposed architectures for CEFR prediction. a) **Baseline**: original learners' texts as input; b) **Concatenation**: model taking the original learners' texts and the corrections concatenated; c) **Two-streams**: model processing the original learners' texts and the corrections with separate streams.

These are then merged and processed by the classifier, as shown in Fig. 1(c).

We resort to these types of models to be able to efficiently process texts capturing long-range sequence features thanks to parallel word-processing and self-attention mechanisms. Regardless of the length of the texts, the architecture should be, indeed, able to accurately categorise the examinations according to the CEFR A1, A2, B1, B2 and C1 levels of competence. These, in fact, are fed to the model as labels during the training together with single contextual embeddings, or concatenated ones if corrections are included. Note that we do not provide the model with any indication about the types of errors in the original text. This information is directly extracted by the model when processing the original text together with its corrected version.

4 Experimental Analysis

We evaluate the architectures described above, using both automatic and human corrections, on two English open-source datasets: EFCAMDAT and CLC-FCE. We also experiment varying the amount of training material. The performance of the models is measured in terms of weighted classification accuracy.

4.1 EFCAMDAT Dataset

The EFCAMDAT dataset constitutes one of the largest language learners datasets currently available (Geertzen et al., 2013). The version we use contains 1,180,310 essays submitted by adult English learners from more than 172 different nationalities, covering 16 distinct levels compliant with the CEFR proficiency ones. Each essay has been corrected and evaluated by language instructors; in addition to the original texts, their corrected versions and annotated errors are also included.

We considered a sub-set of the dataset comprising 100,000 tests. Table 1 reports the distribution of the exams across the different CEFR levels, including also the average numbers of violations identified by both humans evaluators and the automatic tool, normalized by the average text length. Note that the average errors per word decrease as the level of competence increases. Observe also that the automatic errors tend to be more numerous than the human ones, in particular for low competence levels. We use the official test partition composed of 1,447 essays. The development set is a 20% subset of the training set.

4.2 CLC-FCE Dataset

The CLC-FCE dataset is a collection of texts produced by adult learners for English as a Second or Other Language (ESOL) examinations from the First Certificate in English (FCE) written exam to attest a B2 CEFR level (Yannakoudakis et al., 2011). The learners' productions, consisting of two texts, have been evaluated with a score between 0 and 5.3 and the errors have been classified in 77 classes. Following the guidelines of the authors, the average score of the two texts has been mapped to CEFR levels, as shown in Table 2. Note that only 4 levels are available in this dataset and that the labels do not uniformly match the ones present in EFCAMDAT. Table 2 reports also the distributions of the texts across the 4 classes with the error partitions. We notice that, in this case,

levels	n. exams	average length	manual errors per word	automatic errors per word
A1	37,290	40	$4 \cdot 10^{-2}$	$10 \cdot 10^{-2}$
A2	36,618	67	$4 \cdot 10^{-2}$	$6 \cdot 10^{-2}$
B1	18,119	92	$4 \cdot 10^{-2}$	$5 \cdot 10^{-2}$
B2	6,042	129	$3 \cdot 10^{-2}$	$4 \cdot 10^{-2}$
C1	1,732	170	$2 \cdot 10^{-2}$	$3 \cdot 10^{-2}$

Table 1: EFCAMDAT dataset (sample of 100,000 exams): number of exams per CEFR level, mean text length (in tokens), mean number of manually and automatically annotated errors per word.

scores	levels	N. exams	average length	manual errors per word	automatic errors per word
0.0 - 1.1	A2	10	220	$16 \cdot 10^{-2}$	$7 \cdot 10^{-2}$
1.2 - 2.3	B1	417	205	$14 \cdot 10^{-2}$	$7 \cdot 10^{-2}$
3.1 - 4.3	B2	1,414	212	$9 \cdot 10^{-2}$	$6 \cdot 10^{-2}$
5.1 - 5.3	C1	265	234	$6 \cdot 10^{-2}$	$4 \cdot 10^{-2}$

Table 2: CLC-FCE dataset: assigned scores and number of exams per CEFR level, mean text length (in tokens), mean number of manually and automatically annotated errors per word.

manual errors have been annotated more in detail and they are indeed more numerous than the automatic ones. In general, the number of errors is higher than what observed in EFCAMDAT. Also for this corpus the average amount of errors per word, both automatic and manual, decreases as the level increases. The total number of texts within the corpus is 2,469. We employed a data partition according to which 2,017 examinations constituted the training set, whereas the remaining 194 constituted the test set. Differently, 10% of the training material represented the validation set. From the entire corpus we had to exclude 10 texts since they were not provided with an assigned score. Despite its small size, CLC-FCE represents an important resource given its systematic analysis of errors and the human corrections provided.

4.3 LanguageTool

In both datasets, the content written by language learners varies according to the levels of competence they were supposed to demonstrate. In addition to the human corrections provided with the data, we have generated automatic corrections using LanguageTool (Miłkowski, 2010), a language checker capable of detecting grammatical, syntactical, orthographic and stylistic errors to automatically correct texts of different nature and length (Naber and others, 2003). The automatic checker is based on surface text processing, does not use a deep parser and does not require a fully formalised grammar. By means of this, we have applied the pre-defined rules for the English language to the learners' essays, generating new correct texts for EFCAMDAT and for CLC-FCE. These were used as additional input data for the experiments.

4.4 Implementation Details

Our models have been implemented using Keras and Hugging-Face's pre-trained *BERT-base-uncased* architecture (Wolf et al., 2020). The models' encoder module, consisting of a Multi-Head Attention and Feed Forward component, receives as inputs the original learners' exams, together with additional possible human or automatic corrections. The transformed contextual embeddings are obtained applying *Global Average Pooling* to the outputs of the pre-trained frozen BERT Head. The classifier consists of a Dense layer of 768 units, with activation function *ReLu* and a Dropout rate of 0.2, followed by another Dense layer with less units, 128, and the same activation function and Dropout rate[1].

Lastly, the output layer consists of a Dense layer with *Softmax* as activation function and the models' final logits correspond to the different CEFR levels within which the texts are respectively clas-

[1] https://www.kaggle.com/akensert/bert-base-tf2-0-now-huggingface-transformer

N. Exams	text only	concatenation		two-streams	
		manual	automatic	manual	automatic
10K	95.2%	95.0%	**95.4%**	94.3%	94.4%
50K	97.1%	97.1%	97.0%	97.1%	97.0%
100K	97.4%	**97.7%**	97.3%	97.4%	97.2%

Table 3: Classification accuracy on EFCAMDAT using different amounts of training data, different inputs and different architectures.

sified. The selected loss is the *Sparse Categorical Cross-entropy* and the evaluation metric is the *accuracy*. The model is trained using *Adam* as optimizer with learning rate 10^{-5} for EFCAMDAT and 10^{-4} for CLC-FCE. The batch size is 32 and the input text maximum length is set to 450 for EFCAMDAT and 512 for CLC-FCE. These hyperparameters were optimized on the related development sets.

5 Experimental Results

Table 3 reports the classification accuracy on the EFCAMDAT test set using the proposed architectures in Fig. 1. Note that although EFCAMDAT features more than 1 million samples, we limit our analysis to 100K texts, due to memory issues and performance saturation. The results include also variations in the amount of training material, considering 10K and 50K training exams. These subsets have been obtained sampling in a uniform way the training set, therefore the distribution of exams per class does not change.

First of all, it is worth noting that the best approach reaches an extremely high classification accuracy (almost 98%). In addition, performance almost saturates with 50K essays, while with only 10K training samples the accuracy is well above 95%. The use of corrections, concatenated with the original text, provides some improvements over the model with original texts only. Automatic corrections seem to be more effective with less training data, while manual annotations outperform the baseline with larger training sets. The latter can, indeed, be more accurate, in particular for high proficiency levels, but their inherited variability makes the learning task more difficult. As a consequence, more training samples are needed to properly learn how to classify the input text. This is evident in Table 3 where the manual corrections are the worst for 10K samples, aligned with the baseline with 50K training samples, and the best performing when the 100K training texts are used. Finally, the two-stream approach averaging the BERT embeddings of the two texts, seems to be less performing, although by a small margin. Probably, the averaging operation does not represent the most suitable one in this context as it tends to generate embedding representations which are somehow intermediate between those of the original texts and those of the corrections and, hence, less discriminative.

Table 4 reports the results obtained on the CLC-FCE corpus. With respect to EFCAMDAT, this corpus is characterized by a smaller amount of training material and by a less consistent evaluation of the input text. These two facts lead to a clear reduction of the classification accuracy, as reported in the table. Due to the lower accuracy and smaller size of the training set, the final performance of each model has a certain degree of variability, which dependents on the model initialization and on the other random number generations in the training process. Therefore, we performed several runs varying the seed of the random number generator. The average accuracy, as well as the standard deviation, are also reported in Table 4.

model	accuracy
text only	61.5% ± 2.0
manual corr.	60.7% ± 1.8
autom. corr.	61.7% ± 1.8
two-streams	61.5% ± 1.3

Table 4: Classification accuracy on CLC-FCE using different architectures and types of corrections. The two-streams model uses automatic corrections. Results are averaged over multiple runs.

Given the limited size of the training set, it is not surprising to find rather similar results across all the models. As expected, the manual corrections are the worst performing, since they would require large training sets to learn how to handle human evaluations. It is worth pointing out that the amount of errors per word in CLC-FCE is much larger than in EFCAMDAT, which makes the learning task even more complex. Nevertheless, considering also the standard deviations, the models based on automatic corrections are slightly better than the model using the original texts only. The two-streams model appears extremely close to the concatenation model, but this could be related to the fact that the overall accuracy is not that high.

6 Conclusions

In this paper we presented an alternative approach for the efficient and unbiased assessment of the competences of English language learners using pre-trained BERT-base models. We structured a multi-class classification task to map the BERT embeddings of written exams from the EFCAMDAT and CLC-FCE open-source corpora to five different levels of the CEFR scale. Alongside the students' original texts and the provided manual corrections, we automatically generated additional corrected versions with LanguageTool, a multifaceted and versatile language checker . Thus, we conducted several experiments varying both the type and quantities of the models' input, as well as the typologies of models. Our results proved that BERT-based architectures remarkably succeed in classifying CEFR proficiency levels starting from original texts, especially with numerically significant data. Moreover, we observed that adding automatic and manual corrections can contribute to improve the quality of results.

References

Yigal Attali. 2004. Exploring the feedback and revision features of criterion. *Journal of Second Language Writing*, 14:191–205.

Jill Burstein, Joel Tetreault, and Nitin Madnani. 2013. The e-rater® automated essay scoring system. In *Handbook of automated essay evaluation*, pages 77–89. Routledge.

Carol A Chapelle and Erik Voss. 2008. Utilizing technology in language assessment. *Encyclopedia of language and education*, 7:123–134.

Carol A Chapelle. 2017. Evaluation of technology and language learning. *The handbook of technology and second language teaching and learning*, pages 378–392.

Chi-Fen Emily Chen and Wei-Yuan Eugene Cheng Cheng. 2008. Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in efl writing classes. *Language Learning & Technology*, 12(2):94–112.

Education Committee Council of Europe, Council for Cultural Co-operation. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Neus Figueras. 2012. The impact of the cefr. *ELT journal*, 66(4):477–485.

Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, et al. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254. Citeseer.

Luke William Harding and Tim McNamara. 2017. Language assessment: The challenge of elf. In *Routledge Handbook of English as a Lingua Franca*. Routledge.

Carl James. 2005. Contrastive analysis and the language learner. *Linguistics, language teaching and language learning*, 120.

Eunice Eunhee Jang. 2017. Cognitive aspects of language assessment. *Language Testing and Assessment,*, pages 163–177.

Hossein Karami. 2013. The quest for fairness in language testing. *Educational Research and Evaluation*, 19(2-3):158–169.

Thomas K Landauer. 2003. Automatic essay assessment. *Assessment in education: Principles, policy & practice*, 10(3):295–308.

Marcin Miłkowski. 2010. Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience*, 40(7):543–566.

Daniel Naber et al. 2003. A rule-based style and grammar checker.

Carsten Roever and Tim McNamara. 2006. Language testing: The social dimension. *International Journal of Applied Linguistics*, 16(2):242–258.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Joshua Wilson and Rod D Roscoe. 2020. Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1):87–125.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Xiaoming Xi, Derrick Higgins, Klaus Zechner, and David M Williamson. 2008. Automated scoring of spontaneous speech using speechratersm v1. 0. *ETS Research Report Series*, 2008(2):i–102.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Wenxin Zhu. 2019. A study on the application of automated essay scoring in college english writing based on pigai. In *2019 5th International conference on social science and higher education (ICSSHE 2019)*, pages 451–454.