Are pre-trained text representations useful for multilingual and multi-dimensional language proficiency modeling?

Anonymous EACL submission

Abstract

Development of language proficiency models for non-native learners has been an active area of interest in NLP research for the past few years. Although language proficiency is multidimensional in nature, existing research typically considers a single "overall proficiency" while building models. Further, existing approaches also considers only one language at a time. This paper describes our experiments and observations about the role of pre-trained and fine-tuned multilingual embeddings in performing multi-dimensional, multilingual language proficiency classification. We report experiments with three languages – German, Italian, and Czech – and model seven dimensions of proficiency ranging from vocabulary control to sociolinguistic appropriateness. Our results indicate that while fine-tuned embeddings are useful for multilingual proficiency modeling, none of the features achieve consistently best performance for all dimensions of language proficiency.

1 Introduction

Automated Essay Scoring (AES) is the task of grading test taker writing using computer programs. It has been an active area of research in NLP for the past 15 years. Although most of the existing research focused on English, recent years saw the development of AES models for second language proficiency assessment for non-English languages, typically modeled using the Common European Framework of Reference (CEFR) reference scale (of Europe, 2002) in Europe.

Most of the past research focused on monolingual AES models. However, the notion of language proficiency is not limited to any one language. As a matter of fact, CEFR (of Europe, 2002) provides language agnostic guidelines to describe different levels of language proficiency, from A1 (beginner) to C2 (advanced). Hence, a universal, multilingual language proficiency model is an interesting possibility to explore. From an application perspective, it will be useful to know if one can achieve cross-lingual transfer and build an AES system for a new language without or with little training data. Vajjala and Rama (2018) explored these ideas with basic features such as n-grams and POS tag ratios. The usefulness of large, pre-trained multilingual models (with or without fine-tuning) from recent NLP research has not been studied for this task, especially for non-English languages.

Further, AES research generally considers language proficiency as a single construct. However, proficiency encompasses multiple dimensions such as vocabulary richness, grammatical accuracy, coherence/cohesion, usage of idioms etc. (Attali and Burstein, 2004). CEFR guidelines also provide language proficiency rubrics for individual dimensions along with overall proficiency for A1–C2. Modeling multiple dimensions instead of a single "overall proficiency" could result in a more fine-grained assessment for offering specific feedback.

Given this background, we explore the usefulness of multilingual pre-trained embeddings for training multi-dimensional language proficiency scoring models for three languages – German, Czech and Italian. The main contributions of our paper are listed below:

- We address the problem of multi-dimensional modeling of language proficiency for three (non-English) languages.
- We explore whether large pre-trained, multilingual embeddings are useful as feature representations for this task with and without fine-tuning.
- We investigate the possibility of a universal multilingual language proficiency model and

zero-shot cross lingual transfer using embedding representations.

The paper is organized as follows. Section 2 briefly surveys the related work. Section 3 describes our corpus, features, and experimental settings. Section 4 discusses our results in detail. Section 5 concludes the paper with pointers to future work.

2 Related Work

Automated Essay Scoring (AES) is a well-researched problem in NLP and has been applied to real-world language assessment scenarios for English (Attali and Burstein, 2004). A wide range of features such as document length, lexical/syntactic n-grams, and features capturing linguistic aspects such as vocabulary, syntax and discourse are commonly used (Klebanov and Flor, 2013; Phandi et al., 2015; Zesch et al., 2015). In the recent past, different forms of text embeddings and pre-trained language models have also been explored (Alikaniotis et al., 2016; Dong and Zhang, 2016; Mayfield and Black, 2020) along with approaches to combine linguistic features with neural networks (Shin, 2018; Liu et al., 2019). Ke and Ng (2019) and Klebanov and Madnani (2020) present the most recent surveys on the state of the art in AES (focusing on English).

In terms of modeling, AES has been modeled as a classification, regression, and ranking problem, with approaches ranging from linear regression to deep learning models. Some of the recent work explored the usefulness of multi-task learning (Cummins and Rei, 2018; Berggren et al., 2019) and transfer learning (Jin et al., 2018; Ballier et al., 2020). Going beyond approaches that work for a single language, Vajjala and Rama (2018) reported on developing methods for multi- and cross-lingual AES.

Much of the existing AES research has been focused on English, but there is a growing body of research on other European languages: German (Hancke and Meurers, 2013), Estonian (Vajjala and Lõo, 2014), Swedish (Pilán et al., 2016), Norwegian (Berggren et al., 2019) which explored both language specific (e.g., case markers in Estonian) as well as language agnostic (e.g., POS n-grams) features (Vajjala and Rama, 2018) for this task. However, to our knowledge, the use of large pre-trained language models such as BERT (Devlin et al., 2018) has not been explored yet for AES in non-English languages.

Further, most of the approaches focused on modeling language proficiency as a single variable. Although there is some research focusing on multiple dimensions of language proficiency (Lee et al., 2009; Attali and Sinharay, 2015; Agejev and Šnajder, 2017; Mathias and Bhattacharyya, 2020), none of them focused on non-English languages or used recent multilingual pre-trained models such as BERT. In this paper, we focus on this problem of multi-dimensional modeling of language proficiency for three languages—German, Italian, and Czech—and explore whether recent research on multilingual embeddings can be useful for non-English AES.

3 Datasets and Methods

In this section, we describe the corpus, features, models, and implementation details. We modeled the task as a classification problem and trained individual models for each of the 7 dimensions of language proficiency. The rest of this section describes the different steps involved in our approach in detail.

3.1 Corpus

In this paper, we employed the publicly available MERLIN corpus (Boyd et al., 2014), which was also used in the experiments reported in some past research (Hancke, 2013; Vajjala and Rama, 2018) and in the recently conducted REPROLANG challenge (Branco et al., 2020). The MERLIN corpus[1] contains CEFR scale based language proficiency annotations for texts produced by non-native learners in three languages: German, Czech, and Italian in seven dimensions which are described below:

1. **Overall proficiency** is the generic label expected to summarize the language proficiency across different dimensions.

2. **Grammatical accuracy** refers to the usage and control over the language's grammar.

3. **Orthographic control** refers to the aspects of language connected with writing such as punctuation, spelling mistakes etc.

4. **Vocabulary range** refers to the breadth of vocabulary use, including phrases, idiomatic expressions, colloquialisms etc.

[1] Available here for download: https://merlin-platform.eu/C_download.php

5. **Vocabulary control** refers to the correct and appropriate use of vocabulary.

6. **Coherence and Cohesion** refers to the ability to connect different parts of the text using appropriate vocabulary (e.g., connecting words) and creating a smoothly flowing text.

7. **Sociolinguistic appropriateness** refers to the awareness of language use in different social contexts. For example, using proper form of introduction, ability to express both in formal as well as informal language, understanding the sociocultural aspects of language use etc.

Detailed description of a dimension at each CEFR level is provided in the structured overview of CEFR scales document (of Europe, 2002). In the MERLIN corpus, these annotations were prepared by human graders who were trained on these well defined rubrics. More details on the examination setup, grade annotation guidelines, rating procedure, inter-rater reliability and reliability of rating measures can be found in the project documentation (Bärenfänger, 2013). We used the texts and their universal dependency parsed versions – shared by Vajjala and Rama (2018) – consisting of 2266 documents in total (1029 German, 803 Italian, 434 for Czech).

The German corpus had A1–C1, Italian corpus had A2–B1, and Czech had A2–B2 levels for the overall proficiency category. More CEFR levels were represented in the corpus for other proficiency dimensions.[2] In this paper, we treat the annotated labels in the corpus as the gold standard labels.

Missing labels Less than 10 documents had an annotation of 1 instead of the CEFR scale (A1–C2) for some of the dimensions. The documentation did not provide any reason behind this label assignment and we removed them from our experiments. In the case of German and Italian, for less than ten documents, some individual dimensions had a score of "0" while the overall rating was A1. For these documents, we treated "0" score as A1 rating for that dimension. In the case of Czech, about half of the documents (247) for the sociolinguistic appropriateness dimension had a score of "0". The corpus manual does not provide any explanation for the missing annotation, therefore, we excluded this dimension from all experiments involving Czech data.

Inter-dimensional correlations Bärenfänger (2013)'s analyses on MERLIN corpus show that correlations among the different dimensions (including overall proficiency) range from $0.2-0.8$ in different languages.[3] In general, correlations between any two dimensions and specifically with overall proficiency dimension are higher for German and Italian than for Czech. There is no consistent high correlation of overall proficiency with any single dimension. The variations show that these individual dimensions as are indeed different from each other as well as overall proficiency dimension, and we could expect that a model trained on one dimension need not necessarily reflect on the language proficiency of the test taker in another dimension. This further motivates our decision to explore a multi dimensional proficiency assessment approach in this paper.

3.2 Features

One of the goals of the paper is to examine if text representations computed from large, pre-trained, multilingual models such as mBERT (Devlin et al., 2018) and LASER (Artetxe and Schwenk, 2019) are useful for the AES task. We trained classifiers based on these two pre-trained models and compare them with two previously used features—document length baseline and the n-gram features used in Vajjala and Rama (2018). All the features are described below:

- **Baseline**: Document length (number of tokens) is a standard feature in all AES approaches (Attali and Burstein, 2004).

- **Lexical and syntactic features**: n-grams ($1 \leq n \leq 5$) of Word, Universal POS tags (UPOS) from the Universal Dependencies project (Nivre et al., 2016), dependency triplets consisting of the head POS label, dependent POS label, and the dependency label extracted using UDPipe (Straka et al., 2016). While word n-grams are useful only for monolingual setting, the syntactic level n-grams were used in multi/cross-lingual scenarios as

[2] More details on the corpus distribution can be found in MERLIN documentation, and the result files we shared as supplementary material contain CEFR level distributions for all the classification scenarios, for all languages.

[3] For details: Refer to Table 4 for Czech, Table 11 for German, Table 17 for Italian in Bärenfänger (2013).

well, as they are all derived from the same tagset/dependency relations.

- **LASER embeddings** map a sentence in a source language to a fixed dimension (1024) vector in a common cross-lingual space allowing us to map the vectors from different languages into a single space. Since the number of sentences in an essay is variable, we map each sentence in the segmented text to a vector and then compute the average of the vectors to yield a 1024 dimension representation as our feature vector.

- **mBERT**: We apply the 12-layer pre-trained multilingual BERT (trained on Wikipedias of 104 languages with shared word-piece vocabulary) for mapping an essay (truncated to 400 tokens which is the upper bound of the length for 93% of the documents) into a 768 dimension vector. Specifically, we use the vector for the `CLS` token from the final layer as the feature vector for non-finetuned classification experiments. We used the MERLIN corpus texts to do task specific fine-tuning of mBERT.

It is possible to use other representations such as using average of the tokens' embeddings of the last layer instead of using CLS token for mBERT, or explore other recent pre-trained mono-/multilingual representations. Our goal is not to explore the best representation but rather test if a representative approach could be used for this problem. To our knowledge, only Mayfield and Black (2020) studied the application of BERT for AES in English, and its utility in the context of non-English and multilingual AES models has not been explored.

Although, it is possible to use the "domain" features such as spelling/grammar errors which are commonly seen in AES systems, our goal in this paper is to explore how far we can go with the representations without any language specific resources for this task. Considering that such representations are expected to capture different aspects of language (Jawahar et al., 2019; Edmiston, 2020), we could hypothesize that some of the domain specific features are already captured by them.

3.3 Models and Evaluation

As discussed in Section 1, our motivation in this paper is to evaluate whether pre-trained multilingual embedding representations are useful for performing multidimensional AES, whether they can be used to achieve a universal representation for this task (multilingual), as well as transfer from one language to another (cross-lingual) and if the pre-trained embedding representations can be transferred to the AES task (*fine-tuning*). To explore this, we trained mono/multi/cross lingual classification models using each of the features described in Section 3.2, for each of the 7 dimensions.

All the traditional classification models based on n-grams, LASER and mBERT were tested using traditional classification algorithms: Logistic Regression, Random Forests, and Linear SVM. The mBERT fine-tuned model consists of a softmax classification layer on top of the `CLS` token's embedding. We used the MERLIN corpus texts to fine-tune mBERT for this task in all the three classification scenarios.

We evaluate the classifiers in monolingual and multilingual scenarios through stratified five-fold validation where the distribution of the labels is preserved across the folds. Owing to the nature of the corpus and the presence of unbalanced classes in all the languages and dimensions in the dataset, we used weighted F_1 score to compare model performance, as was done in the REPROLANG challenge (Branco et al., 2020). In the cross-lingual scenario, we trained on the German dataset and tested separately on Czech and Italian languages.

3.4 Implementation

All the POS and dependency n-gram features were computed using UDPipe (Straka et al., 2016). All the traditional classification models were implemented using the Python library `scikit-learn` (Pedregosa et al., 2011) with the default settings. LASER embeddings were extracted using the python package `laserembeddings`.[4] The extraction of mBERT embeddings and fine-tuning was performed using the `Hugging Face` library and PyTorch.[5] The code, processed dataset, and detailed result files are uploaded as supplementary material with this paper.

4 Experiments and Results

As mentioned earlier, we trained monolingual, multilingual and cross-lingual classification models for all the seven proficiency dimensions. We report

[4] https://pypi.org/project/laserembeddings/

[5] https://huggingface.co/transformers/v2.2.0/model_doc/bert.html#bertforsequenceclassification

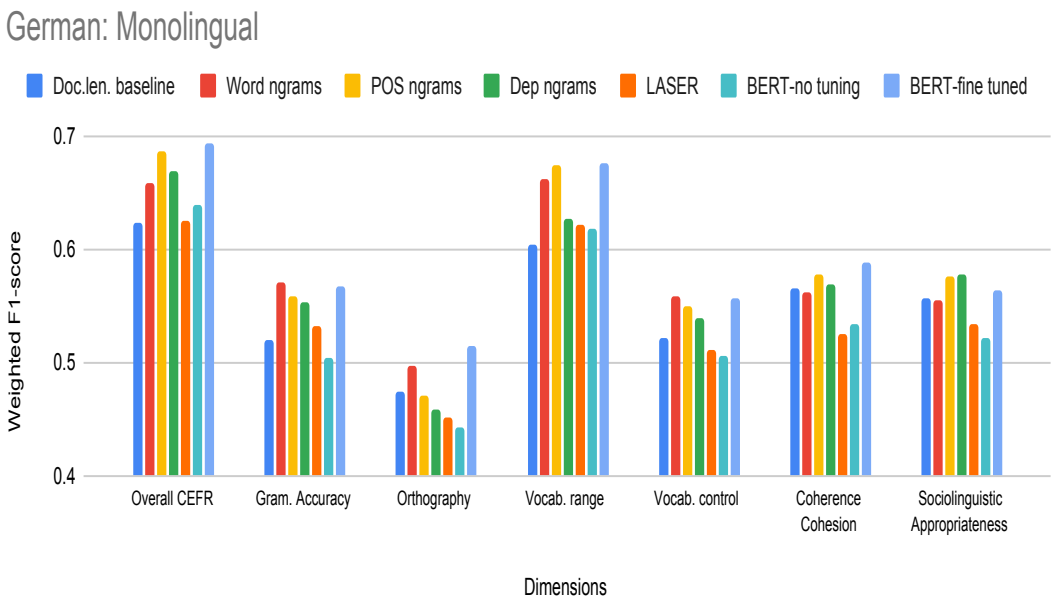

Figure 1: German monolingual five-fold validation results. All *POS* and *Dep* n-grams are based on Universal Dependencies framework.

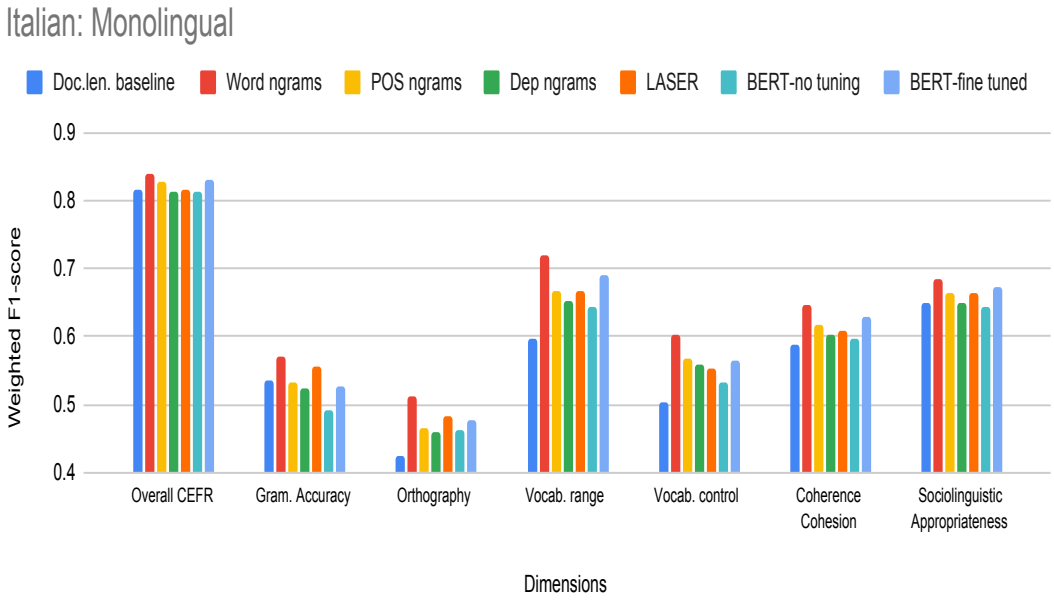

Figure 2: Italian monolingual five-fold validation results

results with logistic regression, which performed the best in most of the cases. The results for the other classifiers such as Random Forest and Linear SVM are provided in the supplementary material.

4.1 Monolingual Classification

Figures 1, 2 and 3 show the results of monolingual classification for German, Italian and Czech respectively for all the feature sets and proficiency dimensions.

German The fine-tuned BERT model performs the best (from Figure 1) for the Overall CEFR proficiency prediction dimension closely followed by POS n-grams. Except for the Vocabulary Range dimension, none of the other dimensions seem to perform on par with Overall proficiency in terms

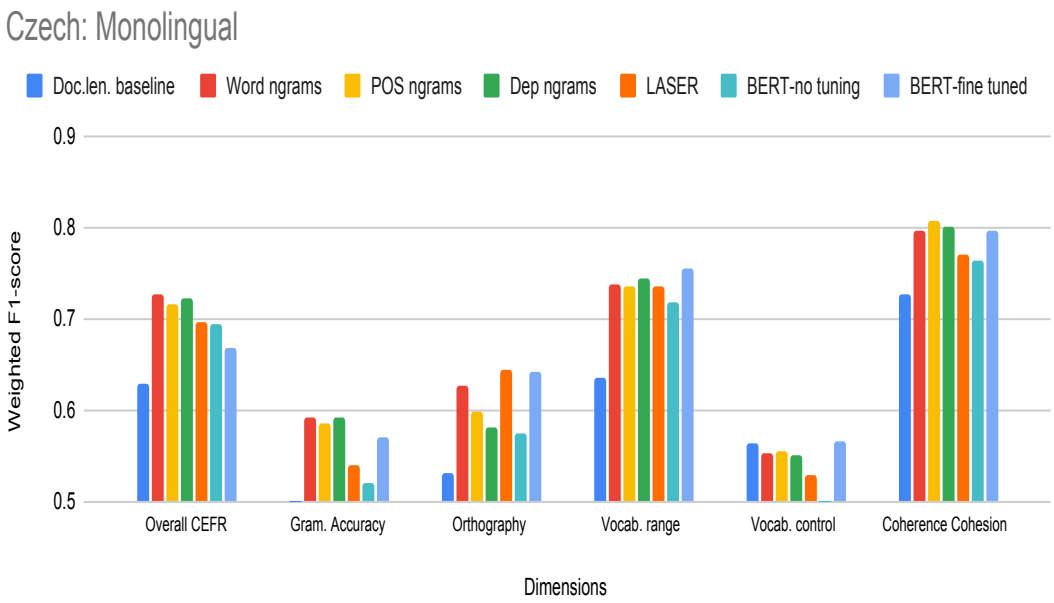

Figure 3: Czech monolingual five-fold validation results

of absolute numbers, though. Fine-tuned mBERT performs the best for Orthographic control dimension, where the rest of the feature sets performed rather worse. Overall, these results seem to indicate that all our features only capture the 'Overall proficiency' dimension well, and to some extent the 'Vocabulary Range' dimension. All features perform rather poorly at the task of prediction of orthographic control.

Italian The word n-grams perform the best for Overall Proficiency closely followed by POS n-grams and fine-tuned mBERT model. There is not much variation among the features, with little improvement over the strong document length baseline for any feature group. Further, the performance on other dimensions seems far worse than Overall Proficiency, compared to German. Orthographic control is the worst performing dimension even for Italian. Word n-grams are the best feature representation across all dimensions for Italian. Although mBERT fine-tuning improved the performance over non fine-tuned version, both LASER and mBERT based models don't perform better than word or POS n-grams in any dimension. Thus, while there are some similarities between German and Italian classification, we also observe some differences.

Czech Across all the dimensions, the results (Figure 3) for Czech are different from German and Italian. The performance of the different systems on Coherence/Cohesion dimension is much better than the Overall Proficiency. Orthographic Control, which seemed to be the worst modeled dimension for German and Italian, does better than grammatical accuracy and vocabulary control. There is a larger difference between the baseline performance and the best performance for most of the dimensions, than it was for German and Italian.

The main conclusions from the monolingual classification experiments are as follows:

- The feature groups don't capture multiple dimensions of proficiency well and there is no single feature group that works equally well across all languages.
- Pre-trained and fine-tuned text representations seem to perform comparably to traditional n-gram features in several language-dimension combinations.

One possible reason for the variation across dimensions could be that the corpus consists of texts written by language learners, coming from various native language backgrounds. It is possible that there are no consistent n-gram patterns in various dimensions to capture due to this characteristic. Further, models such as LASER and mBERT are pre-trained on well formed texts, and may not be able to capture the potentially erroroneous lan-

guage patterns in MERLIN texts. We can hypothesize that the overall proficiency label potentially captures some percent of each dimension, and is probably easier to model than others. However, even this hypothesis does not hold for the case of Czech, where Coherence/Cohesion dimension perhaps much better than the overall proficiency. Clearly, more analysis and experiments are needed to understand these aspects. The current set of experiments indicate that it is a worthwhile future direction to pursue.

4.2 Multilingual Classification

In multilingual classification, we work with a single dataset formed by combining the essays from all the three languages. We trained and tested classifiers for all combinations of feature sets and dimensions on the single large dataset. Since CEFR guidelines for language proficiency are not specific to any one language, we would expect multilingual models to perform on par with individual monolingual models. The results of our multilingual experiments are given in Figure 4.

Our results show that the fine-tuned mBERT model performs the best on most of the dimensions, closely followed by the UPOS n-grams features. To understand the relation between the multilingual model and its constituents, we looked at how each language fared in this model. For overall proficiency dimension, for example, the best result is achieved with fine-tuned classifier based on mBERT (0.745), which is closer to the average of the results from the three monolingual models. While German (0.693 in monolingual vs 0.683 in multilingual) and Italian (0.829 vs 0.826) saw a slight dip in the multilingual setup, Czech (0.669 vs 0.718) saw a 5 point increase due to multilingual classification.

Clearly, multilingual classification is a beneficial setup for languages with lower monolingual performance or less data, without compromising on those languages that had better performance. However, there is still a lot of performance variation in terms of absolute numbers across dimensions. As with the case of monolingual models, we can potentially attribute this to the fact that we are dealing with a relatively smaller sized dataset in MERLIN, with texts written by people with diverse native language backgrounds, although more experiments are needed in this direction to confirm this.

4.3 Crosslingual Classification

Here, we train different classification models on German, and test them on Italian and Czech. We chose German for the training side since German has the largest number of essays in MERLIN corpus. In the case of mBERT, we performed fine-tuning on German part of the corpus, and tested the models on Italian and Czech texts respectively. The goal of this experiment is to test if there are any universal patterns in proficiency across the languages and understand if zero-shot cross-lingual transfer is possible for this task.

UPOS n-grams consistently performed better than other features for most of the dimensions, in both the cross lingual setups. There is more performance variation among the different dimensions for Italian compared to Czech. In the case of Czech, similar to the monolingual case, the Coherence/cohesion dimension achieved superior performance than others, even with the baseline document length feature. This is a result worth considering further qualitative analysis in future. More details on the results of this experiment can be found in the Figures folder in the supplementary material. Our cross-lingual experiments seem to indicate that the embedding representations we chose are not useful for zero-shot learning, and that UPOS n-grams may serve as a strong baseline for building AES systems with new languages.

4.4 Error Analysis

We observed substantial performance differences across features/dimensions/languages in various experimental settings. While we don't have a procedure to understand the exact reasons for this yet, examining the confusion matrices (provided in the supplementary material) may give us some insights into the nature of some of these differences. Therefore, we manually inspected a few confusion matrices, posing ourselves three questions:

1. How does a given feature set perform across different dimensions for a given language?
2. How do different features perform for a single dimension for a given language?
3. How does a given feature set perform for a given dimension among the three languages?

In all these cases, we did not notice any major differences, and the confusion matrices followed the expected trend (observed in previous research)

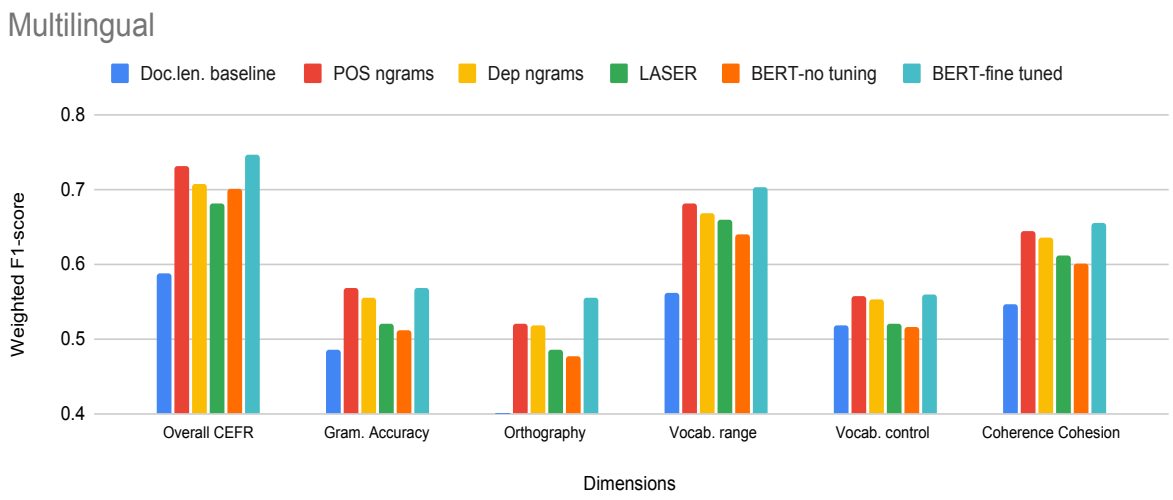

Figure 4: Multi-dimensional, Multilingual language proficiency classification. The `doclen` baseline for Orthography domain is 0.3956 which is less than the minimum threshold of 0.4.

– immediately proximal levels such as A2/B1 or A1/A2 are harder to distinguish accurately as compared to distant levels such as A1/B2 along with the expected observation that levels with larger representation have better results. It is neither possible to cover all possible combinations nor is it sufficient to gain more insights into the models just by looking at confusion matrices alone. Carefully planned interpretable analyses should be conducted in future to understand these differences further.

5 Summary and Conclusions

In this paper, we reported several experiments exploring multi-dimensional CEFR scale based language proficiency classification for three languages. Our main conclusions from these experiments can be summarized as follows:

1. UPOS n-gram features perform consistently well for all languages in monolingual classification scenarios for modeling "overall proficiency", closely followed by embedding features in most language-dimension combinations.

2. Fine-tuned large pre-trained models such as mBERT are useful language representations for multilingual classification, and languages with low monolingual performance benefit from a multilingual setup.

3. UPOS features seem to provide a strong baseline for zero-shot cross lingual transfer, and fine-tuning was very not useful in this case.

4. None of the feature groups consistently perform well across all dimensions/languages/classification setups.

The first conclusion is similar to (Mayfield and Black, 2020)'s conclusion on using BERT for English AES. However, these results need not be interpreted as a "no" to pre-trained models. Considering that they are closely behind n-grams in many cases and were slightly better than them for German, we believe they are useful to this task and more research needs to be done in this direction exploring other language models/fine-tuning options.

Pre-trained and fine-tuned models are clearly useful in a multilingual classification setup, and it would be an interesting new direction to pursue for this task. As a continuation of these experiments, one can look for a larger CEFR annotated corpus for a language such as English, and explore multilingual learning for languages with lesser data.

The results from the experiments presented in this paper highlight the inherent difficulty in capturing multiple dimensions of language proficiency through existing methods, and the need for more future research in this direction. An important direction for future work is to develop better feature representations that capture specific dimensions of language proficiency, which can potentially work for many languages. Considering that all the dimensions share some commonalities and differences with each other, multi-task learning is another useful direction to explore.

References

Tamara Sladoljev Agejev and Jan Šnajder. 2017. Using analytic scoring rubrics in the automatic assessment of college-level summary writing tasks in l2. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 181–186.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Yigal Attali and Jill Burstein. 2004. Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, 2004(2).

Yigal Attali and Sandip Sinharay. 2015. Automated trait scores for toefl® writing tasks. *ETS Research Report Series*, 2015(1):1–14.

Nicolas Ballier, Stéphane Canu, Caroline Petitjean, Gilles Gasso, Carlos Balhana, Theodora Alexopoulou, and Thomas Gaillat. 2020. Machine learning for learner english: A plea for creating learner data challenges. *International Journal of Learner Corpus Research*, 6(1):72–103.

Olaf Bärenfänger. 2013. Assessing the reliability and scale functionality of the merlin written speech sample ratings. Technical report, European Academy, Bolzano, Italy.

Stig Johan Berggren, Taraka Rama, and Lilja Øvrelid. 2019. Regression or classification? automated essay scoring for norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–102.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The merlin corpus: Learner language and the cefr. In *LREC*, pages 1281–1288.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with reprolang2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545.

Ronan Cummins and Marek Rei. 2018. Neural multi-task learning in automated assessment. *arXiv preprint arXiv:1801.06830*.

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*.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077.

Daniel Edmiston. 2020. A systematic analysis of morphological content in bert models for multiple languages. *arXiv preprint arXiv:2004.03032*.

Council of Europe. 2002. Common european framework of reference for languages: Learning, teaching, assessment. structured overview of all cefr scales.

Julia Hancke. 2013. Automatic prediction of cefr proficiency levels based on linguistic features of learner language. *Master's thesis, University of Tübingen*.

Julia Hancke and Detmar Meurers. 2013. Exploring cefr classification for german based on rich linguistic modeling. pages 54–56.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: a survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6300–6308. AAAI Press.

Beata Beigman Klebanov and Michael Flor. 2013. Word association profiles and their use for automated scoring of essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1148–1158.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Yong-Won Lee, Claudia Gentile, and Robert Kantor. 2009. Toward automated multi-trait scoring

of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, 31(3):391–417.

Jiawei Liu, Yang Xu, and Lingzhe Zhao. 2019. Automated essay scoring based on two-stage learning. *CoRR*, abs/1901.07744.

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91, Seattle, WA, USA Online. Association for Computational Linguistics.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439.

Ildikó Pilán, David Alfter, and Elena Volodina. 2016. Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. *CL4LC 2016*, page 120.

Eunjin Shin. 2018. A neural network approach to automated essay scoring: A comparison with the method of integrating deep language features using coh-metrix.

Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *LREC*.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic cefr level prediction for estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, 107. Linköping University Electronic Press.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal cefr classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *BEA@ NAACL-HLT*, pages 224–232.

A Supplemental Material

The code and data for these experiments are uploaded together with the paper.