Experiments with Universal CEFR Classification

Sowmya Vajjala
Applied Linguistics and Technology Program
Iowa State University, USA
sowmya@iastate.edu

Taraka Rama
Department of Informatics
University of Oslo, Norway
tarakark@ifi.uio.no

Abstract

The Common European Framework of Reference (CEFR) guidelines describe language proficiency of learners on a scale of 6 levels. While the description of CEFR guidelines is generic across languages, the development of automated proficiency classification systems for different languages follow different approaches. In this paper, we explore universal CEFR classification using domain-specific and domain-agnostic, theory-guided as well as data-driven features. We report the results of our preliminary experiments in monolingual, cross-lingual, and multilingual classification with three languages: German, Czech, and Italian. Our results show that both monolingual and multilingual models achieve similar performance, and cross-lingual classification yields lower, but comparable results to monolingual classification.

1 Introduction

Automated Essay Scoring (AES) refers to the task of automatically grading student essays written in response to some prompt. Different approaches for AES have been proposed in literature, where it is modeled as a regression, ranking or a classification problem (cf. Yannakoudakis et al., 2011; Taghipour and Ng, 2016; Pilán et al., 2016). To our knowledge, all the previous work described approaches that work with a single language (mostly English). Feature representations that can work for multiple languages and those that support cross-lingual AES have not been explored.

At first thought, using an essay scoring model developed for one language to test on another language seems unacceptable. However, CEFR guidelines are not developed for a specific language. This leads us to hypothesize about a common model of "proficiency" that can work across languages. The existence of such a model would also be beneficial for quick prototyping of AES systems for languages that do not have readily available training data.

In this paper, we explore this hypothesis by exploring CEFR-classification for three languages-German, Italian, and Czech, for which CEFR graded data is publicly available. Apart from constructing individual models using generic text classification and AES specific features, we also looked into cross-lingual (i.e., training a model on one language and testing on another) and multilingual classification approaches (i.e., building a single classification model trained on all the three languages at once).

Testing our universal CEFR hypothesis would require a common feature representation across languages. We developed such a representation, by employing features based on part-of-speech tags and dependency relations from the Universal Dependencies (UD)(Nivre et al., 2016) project which provides treebanks for over 60 languages.[1] Therefore, this approach can be easily extended to other languages with available CEFR graded texts and UD treebanks.

In short, the contributions of this paper are as follows:

1. We study AES for multiple languages for the *first* time using CEFR scale.

2. We explore, for the *first* time, the possibility of a Universal CEFR classifier by training a single model consisting of three languages.

3. We also report *first* results on cross-lingual AES.

The rest of this paper is organized as follows: Section 2 describes related work. Section 3 describes our data and methods. Section 4 discuss

[1] http://universaldependencies.org/

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 147–153
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

our experiments and results in detail. Section 5 concludes the paper with pointers to future work.

2 Related Work

AES is a well studied research problem and AES systems are used to automatically grade essays in exams such as GRE® and TOEFL® (Attali and Burstein, 2004). There is a considerable amount of work that explored various aspects of AES research such as: dataset development, feature engineering, multi-corpus studies and the role of prompt and task information (Yannakoudakis et al., 2011; Phandi et al., 2015; Zesch et al., 2015; Alikaniotis et al., 2016; Taghipour and Ng, 2016; Vajjala, 2018).

AES models developed for non-English languages, primarily using the CEFR scale (Hancke 2013 for German, Pilán et al. 2016 for Swedish, Vajjala and Lõo 2014 for Estonian) employ several language specific features and show their relevance for the task. However, to the best of our knowledge, there is no previous work on developing common models and feature representations that work across languages. Against this background, we set out to address the question: "Is there a universal model for language proficiency classification?"

3 Approach

3.1 Dataset

To test our hypotheses, we need corpora graded with CEFR scale for multiple languages. One such multi-lingual corpus is the freely available MERLIN (Boyd et al., 2014) corpus.[2] This corpus consists of 2286 manually graded texts written by second language learners of German (DE), Italian (IT), and Czech (CZ) as a part of written examinations at authorized test institutions. The aim of these examinations is to test the knowledge of the learners on the CEFR scale which consists of six categories – A1, A2, B1, B2, C1, C2 – which indicate improving language abilities. The writing tasks primarily consisted of writing formal/informal letters/emails and essays. MERLIN corpus has a multi-dimensional annotation of language proficiency covering aspects such as grammatical accuracy, vocabulary range, sociolinguistic awareness etc., and we used the "Overall CEFR rating" as the label for our experiments in this paper. Other information provided about the authors included- age, gender, and native language, and information about the task such as topic, and the CEFR level of the test itself. We did not use these information in the experiments reported in this paper. Further, we removed all Language-CEFR Category combinations that had less than 10 examples in the corpus (German had 5 examples for level C2 and Italian had 2 examples for B2 which were removed from the data). We also removed all the unrated texts from the original corpus. The final corpus had 2266 documents covering three languages, and Table 1 shows the distribution of labels in the final corpus.

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

Table 1: Composition of MERLIN Corpus

3.2 Features

Our feature set consists of features that are commonly used in AES systems, as well as others that can be generalized across languages. They are described below:

1. Word and POS n-grams, which were commonly used in AES models in the past (Yannakoudakis et al., 2011).

2. Task-specific word and character embeddings trained through a softmax layer. Although word embeddings were used in recent neural AES models(Alikaniotis et al., 2016), this paper is the first to explore character embeddings as a cross-linguistic feature for AES model.

3. Dependency n-grams where each unigram is a triplet consisting of dependency relation, POS tag of the dependent, POS tag of the head. To our knowledge, these features were not used in any of the previous work on AES.

4. Linguistic features specific to AES literature:

   (a) Document length: The number of words in a document which is a common feature used in AES literature.

[2] http://merlin-platform.eu/

(b) Lexical richness features: Lu (2012) described several lexical richness and language proficiency for English, which were used in previous AES systems (Hancke, 2013). In this paper, we used lexical density, lexical variation, and lexical diversity features that are commonly used in the AES literature.

(c) Error features: Total number of errors and total spelling errors are obtained for German and Italian from an open-source, rule based spelling and grammar checker.[3] To the best of our knowledge, there is no existing tool for Czech grammar check, and hence we did not extract error features for Czech.

We will refer to these as domain features in this paper.

We extracted all n-gram features where $n \in [1, 5]$ and excluded those n-grams that appeared less than 10 times in the corpus. All the POS and dependency relation based features are extracted using the UDPipe parser (Straka et al., 2016) trained on Universal Dependencies treebanks (Nivre et al., 2016).

Feature Combinations: In addition to the above mentioned features, we also explored the effectiveness of combining n-gram features with domain features. The n-gram features are sparse whereas the domain features are dense; therefore, we combined them by training a n-gram feature classifier and using the probability distribution over its cross-validated predictions with domain features to train the final classifier.

3.3 Classification and Evaluation

We compared logistic regression, random forests, multi-layer perceptron, and support vector machines for experiments with non-embedding features and Neural Network models trained on task-specific embedding representations for other experiments. Word embeddings for each language were task-specific are trained only using the MERLIN corpus. The embeddings are stacked with a softmax layer and trained with categorical cross-entropy loss and Adadelta algorithm. We also experimented by training a softmax classifier with character and word embeddings as input and found that the combined model does not perform as well as a stand-alone word embeddings model.

Considering the space restrictions, we report only the best performing systems in this paper. Due to the unbalanced class distribution across all the three languages in the data, we employed weighted-F1 score to evaluate the performance of our trained models. Weighted F1 is computed as the weighted average of the F1 score for each label, taking label support (i.e., number of instances for each label in the data) into account. For both monolingual and multilingual settings, we report results with 10-fold cross validation. For cross-lingual evaluation, we report results on the test language's data.

All our neural network models are implemented using Keras (Chollet et al., 2015) with TensorFlow as the backend (Abadi et al., 2015) and other models were implemented using scikit-learn (Pedregosa et al., 2011; Buitinck et al., 2013).[4]

While it is also possible to model AES as a regression task, we report classification results which is common in CEFR classification tasks. Our initial experiments with linear regression gave Pearson and Spearman correlation in the range of $0.7 - 0.9$ with gold standard scores, which is comparable with previous results on English AES task obtained using regression models (Alikaniotis et al., 2016).

4 Experiments and Results

For all the experiments, we considered a classifier using only document length (number of words per document) as the feature as the baseline. Unless explicitly stated, all the reported results for non-embedding features are based on Random Forest classifier, which was the best performing classifier in our experiments. Numbers with superscript ^{L} indicate performance of results with a Logistic Regression model.

4.1 Monolingual classification

Our classification results with different feature sets for the three languages are summarized in table 2.

All feature representations perform better than the document length baseline, resulting in close to 25% improvement in the macro F1 score in some cases. All the three sets of n-gram features per-

[3] https://languagetool.org/

[4] Relevant code, generated results and the parameter settings are available at: https://github.com/nishkalavallabhi/UniversalCEFRScoring

Features	**DE**	**IT**	**CZ**
Baseline	0.497	0.578^{L}	0.587^{L}
Word ngrams (1)	0.666	0.827	0.721
POS ngrams (2)	0.663	0.825	0.699
Dep. ngrams (3)	0.663	0.813	0.704
Domain features	0.533^{L}	0.653^{L}	0.663
(1) + Domain	**0.686**	**0.837**	**0.734**
(2) + Domain	**0.686**	0.816	0.709
(3) + Domain	0.682	0.806	0.712
Word embeddings	0.646	0.794	0.625

Table 2: Weighted F1 scores for Monolingual Classification

form comparably in the case of German and Italian. In the case of Czech, word n-grams turn out be a better predictor of CEFR scale than syntactic features. The domain features, by themselves, do not perform well for any of the languages. However, concatenating the domain features with n-gram features yield slightly better classification results. Word embeddings perform poorly for Czech compared to other non-embedding features, and come close to lexical and syntactic features in the case of German and Italian. Whether using embeddings pre-trained on a larger corpus will give us better scores is something that needs to be explored in future.

To our knowledge, Hancke (2013) is the only comparable work which explored CEFR classification for German using the same dataset, but with several language specific morphological and syntactic features. Our results are comparable to the reported results of Hancke (2013), although we primarily rely on data-driven features. To our knowledge, there are no existing results for Czech and Italian.

German, which has a larger dataset, seems to perform poorer than the other two languages. One possible explanation for this could be that we are dealing with a 5 class classification for German, where as it is only a 3 class problem for Czech and Italian. It is also possible that these feature representations are not sufficient to model German language proficiency labeling task. Further experiments (and possibly with other existing CEFR datasets) are needed to understand why the classification results differ between different languages.

4.2 Multilingual classification

In this setup, we combined all the language texts and trained a single universal CEFR classifier. Table 3 shows the results. For the non-neural models, we experimented with and without considering language information as a categorical feature. The neural network model is a multitasking model (Çöltekin and Rama, 2016) that consists of character and word embeddings as input. The model learns to predict both the language of the text (language identification) and the CEFR category simultaneously. The model is trained using categorical cross-entropy and Adadelta algorithm. The table shows results with and without language identification for neural models.

Features	**lang (-)**	**lang (+)**
Baseline	0.428^{L}	-
Word n-grams	0.721	0.719
POS n-grams	**0.726**	**0.724**
Dependency n-grams	0.703	0.693
Domain features	0.449^{L}	0.471^{L}
Word + Char embeddings	0.693	0.689

Table 3: Weighted F1 scores for multilingual classification with models trained on combined datasets.

We observe that the document length baseline seems to perform poorer than monolingual models in this case. Further, we can see that the average result on monolingual model as close to the multilingual model in case of POS n-grams, dependency n-grams, and embeddings. However, domain features clearly perform poorly compared to monolingual case. While one could argue that the better performance multilingual model over some monolingual models is due to more training data, this does not seem to be true for some feature groups (baseline, domain features). One inference we can draw is that some feature groups have similarities in terms of proficiency categories assigned for different languages, which lends support to our hypothesis. Although we did not perform a qualitative language specific evaluation yet, the results so far indicate that efforts to build such a universal scoring model is a worthwhile effort.

4.3 Cross-lingual classification

In this setup, we trained a CEFR model on one language and tested it on others. We trained the cross-lingual model only on German data since it has examples for all categories in our corpus. Table 4 summarizes our results. We did not train with word n-grams and word embeddings here as they are lexical and are language specific and are not suitable for this scenario. Table 4 presents the results of the experiments in this setup. The results show a drop in performance when compared to monolingual models, which is not surprising as the feature weights are tuned to German syntactic features. However, it is interesting to note that the drop is less than 10% in both cases. In the case of Italian, the domain features yield similar results to monolingual results suggesting that there are some possible universal patterns of language use in the progression towards language proficiency. All feature groups perform better than the document length baseline for Italian, and domain features perform poorer than the baseline for Czech. The confusion matrices for these experiments (cf. tables 5a and 5b) suggest that most of the misclassification occurs only between adjacent levels of proficiency.

Features	**Test:IT**	**Test:CZ**
Baseline	0.553^{L}	0.487^{L}
POS n-grams	**0.758**	0.649
Dependency n-grams	0.624	**0.653**
Domain features	0.63^{L}	0.475

Table 4: Weighted F1 scores for cross-lingual classification model trained on German.

The results of this experiment indicate that while cross-lingual classification results in a drop in performance, it still captures the proficiency scale meaningfully. So, the next step in this direction would be to explore better representations of the data, and better modeling methods.

5 Conclusion

In this paper, we reported the results of first experiments conducted with the aim of exploring a "universal CEFR classifier". The results so far indicate that cross-lingual and multilingual classifiers yield comparable performance to individual language models. These results provide some evidence for a universal notion of language proficiency and leave open many questions which need to be explored further in future. Our immediate future plans include a systematic exploration of feature representations which are meaningful for the AES context while being portable across languages. Modeling proficiency classification as a domain adaptation problem (where the domain is another language), and doing multi-task learning by considering other annotation dimensions are other interesting directions to pursue in future. Considering that we have publicly available CEFR graded corpora for other languages such as Estonian, it would be interesting to extend this approach to new languages. This would enable us to investigate questions such as the relationship between genetic/typological similarities between languages and cross/multi-lingual CEFR classification task in future.

→ Pred	A1	A2	B1	B2	C1
A1	5	24	0	0	0
A2	9	311	56	5	0
B1	1	70	279	44	0

(a) DE-Train:IT-Test setup with POS n-gram features

→ Pred	A1	A2	B1	B2	C1
A2	0	129	57	2	0
B1	0	23	101	41	0
B2	0	5	25	51	0

(b) DE-Train:CZ-Test setup with Dependency features

Table 5: Confusion matrices for cross-lingual scoring with Random Forests by training on German data (DE-train).

When it comes to using such methods in real world language testing applications, researchers express concerns about the validity of the chosen feature constructs, and bias and fairness in models. Some recent research (Madnani et al., 2017) in this direction leaves us with some pointers to incorporate these aspects in future research.

Acknowledgments

The second author is supported by BIGMED[5], a Norwegian Research Council funded Lighthouse project, which is gratefully acknowledged.

[5] https://bigmed.no

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. https://www.tensorflow.org/.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 715–725. http://www.aclweb.org/anthology/P16-1068.

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

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.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. pages 108–122.

François Chollet et al. 2015. Keras. https://github.com/keras-team/keras.

Çağrı Çöltekin and Taraka Rama. 2016. Discriminating similar languages with linear svms and neural networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*. pages 15–24.

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

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal* 96(2):190–208.

Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. 2017. Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. pages 41–52.

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.

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*.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *EMNLP*. pages 1882–1891.

Sowmya Vajjala. 2018. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education* 28(1):79–105.

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*. Linköping University Electronic Press, 107.

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*. Association for Computational Linguistics, Portland, Oregon, USA, pages 180–189. http://www.aclweb.org/anthology/P11-1019.

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 relevant for our experiments are available at: `https://zenodo.org/badge/latestdoi/108113378`.