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.

→ 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).

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.

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`.