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.

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