Predicting CEFRL levels in learner English on the basis of metrics and full texts

Taylor Arnold[1], Nicolas Ballier[2], Thomas Gaillat[3], and Paula Lissón[4]

[1]University of Richmond, USA*
[2]Université Paris Diderot, France
[3]The Insight Centre for Data Analytics NUI Galway, Ireland
[4]Université Paris Diderot, France

Abstract

This paper analyses the contribution of language metrics and, potentially, of linguistic structures, to classify French learners of English according to levels of the Common European Framework of Reference for Languages (CEFRL). The purpose is to build a model for the prediction of learner levels as a function of language complexity features. We used the EFCAMDAT corpus [GAK13], a database of one million written assignments by learners. After applying language complexity metrics on the texts, we built a representation matching the language metrics of the texts to their assigned CEFRL levels. Lexical and syntactic metrics were computed with LCA and LSA [Lu14] and koRpus [mm17]. Several supervised learning models were built by using Gradient Boosted Trees and Keras Neural Network methods and by contrasting pairs of CEFRL levels. Results show that it is possible to implement pairwise distinctions, especially for levels ranging from A1 to B1 (A1=>A2: 0.916 AUC and A2=>B1: 0.904 AUC). Model explanation reveals significant linguistic features for the predictiveness in the corpus. Word tokens and word types appear to play a significant role in determining levels. This shows that levels are highly dependent on specific semantic profiles.

Keywords: Learner corpora, criterial features, lexical complexity, syntactic complexity, automatic language scoring, NLP, supervised learning.

*taylor.arnold@acm.org

1 Introduction

This paper focuses on the detection of language levels in Second Language Acquisition. Foreign language education centers provide courses that are tailored according to the different levels of their learners and this leads to two requirements. In the process of learning, it is paramount to provide regular evaluations to both learners and teachers so as to help them focus on specific areas to train upon. There is also a growing demand to group learners homogeneously in order to set adequate teaching objectives and methods. These two requirements rely on language assessment tests whose design and organization are labor intensive and thus costly. Currently, language centers rely on instructors to design and manually correct tests. Alternatively, they use specifically designed short-context and rule-based on-line exercises in which a discrete number of specific language errors is used as a paradigm for level assignment. This creates a bias to use errors as the sole criterion for assessment. Recent research has shown that metrics used in the domain of text mining and NLP can help characterize complexity and thus levels. Consequently, there is a need to use error independent tools to compute levels.

The objective of our experiment is to show that a supervised learning approach is possible. By creating a vector representation of texts matched with language complexity metrics, we build a predictive model for language levels. We use the EFCAMDAT [1] corpus [GAK13] in which learner texts are classified according to the six levels of the Common European Framework

[1]The training and test data were selected and manipulated independently of direct involvement from the EF and Cambridge research teams.This corpus is publicly available at https://corpus.mml.cam.ac.uk/efcamdat2/public_html/

of Reference for Languages (CEFRL). The corpus is used to create a dataset made up of complexity metrics. These complexity metrics are used as input for a Neural Network (NN) whose output layer consists of CEFRL levels. The rest of this paper is organized as follows. Section 2 covers previous work in the domain of automatic language proficiency assignment. In Section 3, we present the corpus and the method used to build the NN model. Section 4 describes the models and we discuss results and conclude in Section 5

2 Related Work

There is a large body of research in automatic language scoring starting with [Pag68]. Over the last four decades, there has been various methods to provide language level analysis. Methods evolved from rule-based approaches focused on pattern matching to Machine Learning (ML) approaches including unsupervised and supervised learning methods.

Rule-based systems have included analyses relying on error detection [LC03, Mit02] while other systems identify features assumed to be to be indicative of proficiency [HXZW11]. With the advent of ML techniques, probabilistic models have appeared. Some approaches use unsupervised learning [Ton13] but, since the task of score assignment relies on human annotated corpora, most ML strategies rely on supervised learning with SVM [YB12] or linear regression and decision tree [CZ11]. Our proposal falls into this category but it uses a Neural Network including several LTSM and dense layers. The input layer includes a multi-dimensional feature representation of written essays and the output corresponds to language levels.

This area of research is closely linked to the research on criterial features that define levels. All the aforementioned studies include a large number of features based on morpho-syntactic patterns, word counts, text and readability metrics [Vaj18]. Authors tested their significance in terms of correlation or classification performance. Another perspective is to specifically focus on weighing feature significance, e.g. by applying strategies based on entropy [FH15], errors [Ton13] or lexical metrics [BG16]. In our experiment, we use lexical and syntactic complexity features [Lu10, Lu12]. Our approach supports two perspectives, i.e. classification according to features and modeling for feature significance analysis with a gradient boost tree model.

Few studies make use of the CERFL as a standard for level description. In English, many studies use other language level scales such as [FH15] with the NICT-JLE corpus levels, [CSMJ11] with TOEFL levels and [YBM11] for the Cambridge Learner Corpus (CLC). We built our data set with texts from the EF-Cambridge Open Language Database (EFCamDat) as a Gold Standard and conducted experiments on English texts classified according to the CEFRL.

3 Method

3.1 Data extraction

All the available raw texts from French learners were downloaded from the EFCAMDAT database in separate XML files. Each XML file corresponded to the EFCAMDAT levels associated with each one of the CEFRL levels. In total, 41,626 texts (approx. 3,298,343 tokens), corresponding to 128 units and 7,695 French learners were downloaded.

3.2 Lexical Diversity Metrics

Most of the lexical diversity metrics are based on the relationship between the numbers of types and tokens within a given text. That is, for example, the case of the **TTR:** (types/tokens), and some of its mathematical transformations, such as the **MSTTR** (types/tokens, with fragments of n tokens), **Herdan's C** (logTypes/logTokens), **Guiraud's RTTR** (types/$\sqrt{tokens}$), **Carrol's CTTR** (types/ $2\sqrt{tokens}$), **Dugast's Uber Index** ($logTokens^2$ / logTypes - log tokens), **Summer's Index (S)** (log(logTypes) / log(logTokens)), **Maas a:** $a^2 = (logTypes/logTokens)/logTokens^2$, **Maas log** ($logTypes_0 = logTypes/\sqrt{1 - logTypes^2/logTokens}$), and **Yule's K** ($10^4(\sum(fX * X^2) - tokens)/tokens^2$, where X is a vector with the frequencies of each type, and f_X is the frequency for each X).

However, most of these metrics are said to be unreliable because they are highly dependent on text length [TB98]. As a consequence, a second generation of metrics, with more complex mathematical transformations, has been developed: **MTLD:** (types/factors, where factors are segments that have reached the stabilization point of TTR; **MATTR** (mean of moving TTR, generated through a 'window' technique of variable sizes that computes TTR of samples of the text); **MTLD-MA**, which combines both factors and the window technique, or the **HDD-D** metric, which computes, for each type, the probability of finding any of its tokens in a random sample of 42 words taken from the text.

3.3 Complexity Metrics

Complexity metrics were generated using Xiaofei Lu's software [Lu12], the L2 Syntactic Complexity Analyzer (L2SCA). There are nine syntactic measures : the number of words (W), the number of sentences (S), of verbal phrases (VP), of clauses (C), of T-units (T), dependent clauses (DC), complex T-units (CT), coordinated phrases (CP) and complex noun phrases (CN). The central unit for these metrics is the T-unit, defined as 'one main clause plus any subordinate clause or nonclausal structure that is attached to or embedded in it' (Hunt 1970:4).

Fourteen indices of syntactic complexity: the mean length of the sentence (MLS), the mean length of the T-unit (MLT), the mean length of the clause (MLC), the number of clauses per sentences (C/S), the number of verbal phrases per T-units (VP/T), the number of clauses per T-units (C/T), the number of dependent clauses per clauses (DC/C), the number of dependent clauses per T-units (DC/T).

3.4 Readability metrics

Readability metrics have traditionally been used to assess the difficulty of texts, i.e, how comprehensible or "readable" a text is for a particular audience. In that sense, for example, some of the metrics were initially designed to determine whether texts were suitable for particular school or college years, such as the Dale and Chall formula or the Bormuth's formulae.

Most of the metrics take into account the average number of words per sentence as well as the average word length, such as the Automatic Readability Index (**ARI** = words per sentence + 9 · word length), or the **LIX** (number of words per sentence + percentage of words with more of 6 characters) and **RIX** (number of long words / number of sentences) formulae; whereas other metrics also include syllable count, such as the well-known **Flesch-Kincaid** (0.39 · average sentence length + 11.8 · average number of syllables per word - 15.59) formula , the **Fog** (0.4 · (average sentence length + number of words with more than two syllabes), the **FORCAST** (20 - number of one-syllable words / 10), and the **Linsear Write** (number of one-syllable words + 3 · number of sentences) indexes. A couple of metrics also take into account the 'complexity' of the vocabulary deployed in the text by including parameters related to the use of 'difficult' or 'hard' words as defined by word lists previously created on the basis of native use, such as the **Dale and Chall** formula (0.1579 · percentage of difficult words + (0.496 · average sentence length) + 3.6365), the **Spache** grade (0.141 · average sentence length + (0.086 · percentage of unfamiliar words) + 0.839), or **Bormuth**'s four formulae.

4 Learner classification models

4.1 Prediction task set-up

The language learner ability levels in the available training data are highly skewed towards beginners, from 17,605 samples of learners at the A1 level to only 76 at the C2 level. One way to account for this in building a predictive classifier is to only compare adjacent classes. Rather than a single multinomial model, therefore, we use 5 pairwise models. While motivated by computational necessity, this approach makes sense from a linguistic perspective as well. The most interesting distinguishing marks between learners should come from adjacent classes. Also, in most practical applications the utility of a predictive model comes from distinguishing between subtle differences in proficiency.

A second challenge in establishing a reliable classification algorithm comes from differences in the tasks given as prompts in the EFCAMDAT dataset. Each of the essays is a response to a given question prompt from a chosen topic such as *Attending a robotics conference* or *Covering a news story*. Each of the 128 topics is provided as a prompt to only one specific learner level. If a modeling dataset was built by randomly assigning documents to training and testing sets, it is likely that models could be predicting the topics rather than language proficiency (particularly if specific word counts are used). For example, we may find that use of the word 'robot' is a good predictor of a learner achieving C2 proficiency because only those learners were asked to write about a robotics conference. In order to avoid these spurious predictors, we split the data into training and testing sets such that all of the essays from specific topics are grouped together. Similar strategies are commonly used in authorship prediction tasks to separate the effects of topical and stylistic predictors.

4.2 Readability metrics

The first set of models use only the computed readability metrics. Gradient boosted trees are used for this task because they are generally very good general-purpose algorithms and, unlike linear models, are particularly robust to highly correlated inputs. Trees are also ideal for metric inputs defined on a scale, such as *readability*, that is primarily designed as an ordinal measurement. Experimentation of the training set led