Automatic proficiency classification in L2 Portuguese

Iria del Río
University of Lisbon, Center of Linguistics-CLUL
igayo@letras.ulisboa.pt

Abstract: We present the first experiments on automatic proficiency classification for L2 Portuguese. For the experiments, we take advantage of a new version of the NLI-PT dataset, a compilation of L2 Portuguese texts written by learners. We use supervised learning and we approach the task as a classification problem, using the CEFR scale. Different linguistic features are tested, combined with different algorithms. With the best model, we get an accuracy of 72%, a result in line with previous experiments with other languages.

Keywords: Proficiency level, CEFR, L2 Portuguese, Supervised Learning

1 Introduction

This work has two main contributions. First, we present a larger and better version of the NLI-PT dataset, a compilation of L2 Portuguese texts with different types of linguistic annotations. Secondly, we describe the first experiments in automatic proficiency classification for L2 Portuguese, where we got similar results to previous works on the field.

The availability of data with linguistic annotations benefits different types of research, from theoretical analysis to statistical approaches like Machine Learning. Learner data is particularly difficult to gather, because of the specific context where this data is produced. For the English language there are big collections of learner data available, like the Cambridge Learner Corpus (16 millions of words) (Nicholls, 1999), but such type of collections are not common for other languages. The NLI-PT dataset aims to solve this gap for European Portuguese. We present a bigger and improved version, with more texts, better annotations and a different and more intuitive organization of the data.

As an example of the usefulness of the dataset, we present the first experiment for automatic proficiency classification of L2 Portuguese. Proficiency classification is a common task in second language learning. The development of the learner is usually defined in relation to a specific scale with different levels of linguistic complexity. One of the most common scales is the one described in the Common European Framework of Reference for Languages (CEFR) (Europe et al., 2009). The CEFR defines 3 broad divisions: A, basic user; B, independent user; C, proficient user, which are subdivided into 6...
opment levels: A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), C1 (advanced) and C2 (proficient). Each level is related to specific linguistic features and skills, establishing a progression from a very rudimentary language to a performance close to a native production. In this context, it is common that learners of a second language perform placement tests that define their proficiency level. The interest of an automatic system that can perform this task is, therefore, evident.

Automatic proficiency classification is commonly considered as a type of Automatic Essay Scoring (AES) task. AES systems are primarily developed for English (Burstein, 2003; Burstein and Chodorow, 2012; Yannakoudakis and Loo 2013), but in recent years systems for other languages have begun to emerge (Vajjala and Loo, 2013). AES has been modeled in different ways, as a regression (Yannakoudakis, Briscoe, and Medlock, 2011), ranking (Taghipour and Ng, 2016) or a classification problem (Pil´an, Vajjala, and Volodina, 2016). The features used are di-verse, from Bag-of-words (BOW) to more ab-stract representations that use higher levels of linguistic information (morphological, syn-tactic or even discursive). It is also very com-mon the use of descriptive metrics of the text related to word or sentence length, like average syllable length, which have been con-nected to proficiency development in the area of Second Language Acquisition (SLA) (Lu, 2012). In general, AES is seen as a monolingu-al task, but recent works like (Vajjala and Rama, 2018) have explored multi and cross-lingual approaches.

Usually, the term AES is used as a general term for referring to different tasks: from proficiency classification of learner texts to readability assessment of teaching materials (Pil´an, Vajjala, and Volodina, 2016). In fact, for the Portuguese language, and to the best of our knowledge, AES works have focused only on readability assessment (Branco et al., 2014) and (Curto, Mamede, and Baptista., 2015). We would like to differentiate the nature of these tasks, that is, readability assessment of input materials for the student, and proficiency classification of learners’ texts because the linguistic parameters they involve are different. Readability assessment tasks usually focus on evaluating the linguistic complexity of potential input mate-
ship between proficiency in L2 English and several lexical dimensions, concluding that
the features linked to lexical variation (like Type-Token ratio) are the most correlated
to the quality of an L2 essay. Several features identified as relevant in this work have
been used by automatic approaches afterward. (Kyle and A. Crossley, 2014) explored
lexical features too and showed that 47.5% of the variance in holistic scores of lexical
proficiency in English as second language can be explained using a range of lexical sophis-
tication indices. Other characteristics like syntactic complexity or error patterns have
been studied too, mainly for English (Tono, 2000),(Lu, 2012), (Vyatkina, 2012), but also
for other languages (Gyllstad et al., 2014).

Concerning automatic proficiency classification, (Yannakoudakis et al., 2018) is one of
the most recent works for the English lan-
guage. The authors used a subset of the
Cambridge Learner Corpus with human pro-
ciciency annotations (levels A1 to C2), con-
taining a total of 2,312 texts. They model the
task as a ranking function and evaluate the
quality of the predicted score by calculating
Pearson’s product-moment and Spearman’s
rank correlation coefficient against the scores
assigned by a human expert. The features
used include character sequences, POS, hy-
brid word, and POS sequences, phrase struc-
ture rules and error rates. The best model
gets a Pearson $r$ of 0.765 and a Spearman $\rho$
of 0.773, with a $\kappa$ of 0.738 (the standard error
is 0.026) that indicates high agreement be-
tween the predicted CEFR scores and those
assigned by humans.

In another recent study, (Vajjala and Rama, 2018) present the first multi and
cross-lingual approach for proficiency classi-
fication. The authors use 2,286 manually
graded texts (five levels, A1 to C1) from
the MERLIN learner corpus (Boyd et al.,
2014). It is an unbalanced dataset, with the
following distribution of learner texts: Ger-
man, 1,029 texts; Italian, 803 texts, and
Czech, 434 texts. The authors compare dif-
f erent algorithms: logistic regression, ran-
dom forests, multi-layer perceptron, and sup-
port vector machines for experiments with
non-embedding features, and Neural Network
models trained on task-specific embedding
representations for other experiments. For
non-embedding features, the best algorithm
is Random Forests in most of the scenarios.

They use a wide range of features: word and
POS n-grams; task-specific word and char-
acter embeddings trained through a softmax
layer; dependency n-grams (not used before);
domain features mainly linked to lexical as-
psects (Lu, 2012); and error features. In
their experiments, monolingual and multiling-
ual models achieve similar performance, and
cross-lingual classification yields lower, but
comparable results to monolingual classification.

3 Dataset
3.1 Corpus
For our experiments, we use an updated
version of the NLI-PT dataset (del Río,
Zampieri, and Malmasi, 2018). The goal of this resource is to make available annotated data produced by L2 Portuguese learners. NLI-PT was originally created for running Native Language Identification (NLI) experiments, and it contains written texts compiled from different learner corpora of L2 Portuguese. Those texts are presented in a clean TXT version, together with versions annotated at two linguistic levels: morphological (POS) and syntactic. The annotation of the dataset was performed with freely available tools. For POS there is a simple POS representation, that is, only type of word, and a fine-grained POS, which is the type of word plus its morphological features. The annotations were performed using the LX Parser (Silva et al., 2010) for the simple POS and the Portuguese morphological module of Freeling (Padró and Stanilovsky, 2012) for detailed POS. Concerning syntactic annotations, NLI-PT includes constituency (from LX Parser) and dependency (DepPattern toolkit (Otero and González, 2012)) annotations.

The new version of the dataset is bigger and contains several improvements. We have corrected some tokenization issues and improved the constituency annotations. Besides this, we have enlarged the dataset with texts from the CAL2 learner corpus 2 (930 new texts). Additionally, we have modified the structure of the dataset. In the new version, the name of the files contains three types of information: the source corpus, the L1 and the proficiency level of the text. For example, for the file “ara_A_008CVETF cops.txt”, the prefix ara corresponds to the native language, Arabic; the A corresponds to the CFR proficiency level and the suffix cop refers to the source of the file, the COPLE2 corpus (Mendes et al., 2016). The CEFR proficiency levels considered in the original learner corpora are not the same: two corpora consider five levels, A1 to C1, while other two consider only three major levels, A, B and C. For this reason, in NLI-PT we have homogenized the levels to three: A, B, and C. The final dataset contains a total of 3,069 texts, corresponding to 15 different native languages. The distribution by proficiency level is presented in Table 1.

As we can see, the distribution of texts by proficiency level is not balanced. For this reason, in our experiments we have used two different datasets: one containing the whole dataset and one with a balanced distribution of 466 texts per class. For the experiments, we split both corpora in training (80%) and testing (20%) sets.

3.2 Features
We were interested in investigating the impact of different linguistic features in the classification task. For this reason, we have tested different types of features extracted from NLI-PT:

1. **Bag of words**, with different variations: using the word form, tokens and lemmas. We performed some initial experiments with the training set to check which representation produced the best results. We got similar results for word forms and tokens, being the word form representation slightly better. For this reason, for further experiments we kept only the word form representation.

2. **POS n-grams**: we used the fine-grained POS representation of NLI-PT, which contains the main POS and also morphological information, like gender or number. We consider that this information could be especially interesting because Portuguese has a rich morphology, and this feature is problematic for certain learners, especially at the initial stages. Agreement errors like areia branca (white-MasculineSingular sand-FeminineSingular) can be captured with a POS n-gram representation, and we wanted to measure the impact of this feature. We evaluated n-grams of different sizes in the experiments.

3. **Dependency triplets n-grams**: we extracted dependency triplets with the form head, relation, dependent generated with DepPattern. Dependency re-

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>Number of Texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>A - Beginner</td>
<td>1,388</td>
</tr>
<tr>
<td>B - Intermediate</td>
<td>1,215</td>
</tr>
<tr>
<td>C - Advanced</td>
<td>466</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3,069</strong></td>
</tr>
</tbody>
</table>

Table 1: Distribution of texts by CEFR proficiency level in the NLI-PT dataset

---

2http://clunl.fsh.unl.pt/en/online-resources/corpus-de-aquisicao-de-l2/
lations are not common in proficiency classification, and we were interested in checking their impact. We also evaluated different types of sizes for the dependency n-grams.

4. **Descriptive and lexical features of the text**: set of 39 features that the studies of SLA have proved as linked with proficiency. Those features are not present in NLI-PT, and therefore we extracted them using the software Pylinguistics (Woloszyn et al., 2016). The features include different types of measures:

- **Lexical features**: number of nouns, number of verbs, number of connectives, lexical diversity, content diversity...

- **Descriptive measures**: average syllables per word, syllable count, word count, etc. We also used the Portuguese adaptation of the Flesch reading index (Martins et al., 1996).

4 **Experiments**

As we have seen, the task of proficiency evaluation can be considered as a classification or a regression problem, depending on the way we consider the proficiency levels, that is, as discrete or continuous scales. For this first attempt, we explore the task as classification, considering that this model obtained better results in previous works (Vajjala and Lőo, 2014).

We used the scikit-learn package (Pedregosa et al., 2011) for training and testing the models and for feature selection. We divided both datasets into training and test sets. We performed some initial tests for feature selection (see above) and for evaluating different algorithms. In these previous experiments, we performed 10-fold cross-validation with the training set and the different sets of features, and we trained a different classifier for each type of features to support a comparison of them. We evaluated Logistic Regression, Linear Discriminant Analysis, Support Vector Machines, Random Forests, and LogitBoost. In general, we had the best results with three algorithms: Logistic Regression (LR), Random Forests (RF) and LogitBoost (LB). For this reason, we only used the models generated with these three algorithms against the test set.

We employed accuracy as the main measure to evaluate the performance of our trained models. We also report weighted-F1 score because the whole dataset is unbalanced. Weighted-F1 score 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. As a baseline, we used text length, extracted with Pylingeistics.

4.1 **Results and Discussion**

Due to space restrictions, we report only the best-performing systems for each combination of features.³

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline_LR</td>
<td>0.58</td>
<td>0.54</td>
</tr>
<tr>
<td>BOW_LB</td>
<td>0.70</td>
<td>0.7</td>
</tr>
<tr>
<td>POS_LB</td>
<td>0.66</td>
<td>0.65</td>
</tr>
<tr>
<td>Dep_RF</td>
<td>0.64</td>
<td>0.59</td>
</tr>
<tr>
<td>Desc_RF</td>
<td>0.63</td>
<td>0.59</td>
</tr>
<tr>
<td>ALL(noBOW)_RF</td>
<td>0.67</td>
<td>0.66</td>
</tr>
<tr>
<td>ALL_LR</td>
<td>0.72</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 2: Results for the whole dataset

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline_RF</td>
<td>0.48</td>
<td>0.46</td>
</tr>
<tr>
<td>BOW_LB</td>
<td>0.66</td>
<td>0.67</td>
</tr>
<tr>
<td>POS_RF</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>Dep_RF</td>
<td>0.54</td>
<td>0.53</td>
</tr>
<tr>
<td>Desc_RF</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>ALL(noBOW)_RF</td>
<td>0.59</td>
<td>0.58</td>
</tr>
<tr>
<td>ALL_RF</td>
<td>0.65</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Table 3: Results for the balanced dataset

<table>
<thead>
<tr>
<th>Features</th>
<th>A-F1</th>
<th>B-F1</th>
<th>C-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline_LR</td>
<td>0.67</td>
<td>0.58</td>
<td>0</td>
</tr>
<tr>
<td>BOW_LB</td>
<td>0.8</td>
<td>0.70</td>
<td>0.43</td>
</tr>
<tr>
<td>POS_LB</td>
<td>0.77</td>
<td>0.66</td>
<td>0.28</td>
</tr>
<tr>
<td>Dep_RF</td>
<td>0.74</td>
<td>0.64</td>
<td>0.02</td>
</tr>
<tr>
<td>Desc_RF</td>
<td>0.72</td>
<td>0.63</td>
<td>0.13</td>
</tr>
<tr>
<td>ALL(noBOW)_LR</td>
<td>0.77</td>
<td>0.67</td>
<td>0.3</td>
</tr>
<tr>
<td>ALL_LR</td>
<td>0.8</td>
<td>0.72</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Table 4: Results per class for the whole dataset

³For each set of features, the abbreviation after the underscore indicates the name of the algorithm employed: LR for Logistic Regression; RF for Random Forests; LB for LogitBoost.
Table 5: Results per class for the balanced dataset

For all the models, the results obtained are better in the whole dataset than in the balanced one. The best result we got is 0.72 accuracy using an ensemble combination of all the features (ALL) with LR, although this value is very close to the one using a BOW representation, 0.7. Interestingly, in the balanced dataset the ensemble combination with all the features had worse results than the best model, BOW\_LB, which uses only one feature. For the ensemble combination that does not use lexical information (it does not include the word forms), POS+Dep+Desc., the results are slightly better than the POS n-gram representation for the whole dataset and worse for the balanced dataset. Both results seem to indicate that adding more linguistic features to the best single-feature models (BOW and POS n-grams) implies only a small gain for the whole dataset and a drop in accuracy for the balanced dataset. In this case, simpler models work generally better.

If we compare the models that use only one type of feature, the best results are for the BOW representation in both datasets, followed by the POS n-gram representation. One of the reasons why the BOW representation captures better the proficiency can be the fact that it keeps the information concerning orthographic problems. A comparison between the results for each type of feature shows similar behavior for both datasets, with the only difference that the Descriptive feature set performs better than the Dep one for the balanced dataset. The algorithms with the best results differ among datasets: LR for the whole dataset; RF for the balanced one.

Concerning the results by class, we can see clear differences between datasets. In the whole corpus, the C class (the one with fewer texts) performs clearly worse than the other two, being the best result 0.43 (BOW\_LB). In the balanced dataset, the F1 score is more equalized between classes, being the results for the B and the C classes pretty similar. In fact, in general, the C class gets better results than the B class. For both datasets and for all models, the best results are always for the A class. This pattern suggests that A texts exhibit certain specific features that make them easy to identify, in comparison with the other two levels.

Due to the lack of space, we cannot include the confusion matrix for each model, therefore, we include table 6 as a reference. False negatives for A are more frequent in the adjacent class, B, and the same happens with the B class, where more texts are classified as C than as A. For the C class, false negatives are more frequent in the previous class, B. This picture seems to show the expected progression of the learners as the proficiency level increases. Interestingly, B texts are more often confused with the adjacent class, instead of with the previous one.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>66</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>12</td>
<td>60</td>
<td>21</td>
</tr>
<tr>
<td>C</td>
<td>12</td>
<td>29</td>
<td>52</td>
</tr>
</tbody>
</table>

Table 6: Confusion matrix for the best model (Words\_LB) in the balanced dataset

5 Conclusions and future work

We present a new improved version of the NLI-PT dataset, and we use it to perform the first experiments on proficiency classification for L2 Portuguese, using the CEFR scale. We modeled the task as classification, and, with the best model, we obtained an accuracy of 72%.

We were interested in answering the question: What defines the proficiency level of an L2 Portuguese text? With this goal, we tested the contribution of different linguistic features combined with different algorithms to the classification task. Additionally, we wanted to test the influence of the distribution of texts by class, and therefore we used two datasets: the whole NLI-PT corpus and a balanced set. We found that an ensemble model, with all the features combined, had the best accuracy (for the whole dataset), but
also that a BOW model achieves a very similar performance (70% accuracy) in both corpora. A POS n-gram model, that does not use lexical information, gets a close result, 66%. This finding is particularly interesting because a POS n-gram model is a more abstract representation that can be less biased by topic or task variables and that can be applied to other L2 Portuguese corpora or even to other similar languages, like Spanish. The two ensemble combinations of features, one with all the features and the other without BOW, get slightly better results than the single-feature models in the whole dataset, but not in the balanced one. This result seems to indicate that simpler models work better in our datasets. A hypothesis that can explain the dominance of the BOW model is the fact that it may capture the orthographic particularities of the learners’ writing, but further analyses are needed to prove this. For both datasets and all models, the class with the best F1 score is the basic user level. This fact seems to indicate that this is the proficiency level with the most characteristic traits.

We would like to investigate several aspects in future work. Since variables like task or textual genre have been proved to influence the linguistic complexity and accuracy of L2 texts (Alexopoulou et al., 2017), we would like to test our best models against L2 Portuguese texts from different sources with different topics and tasks, to check the influence of these variables. We also would like to run a cross-lingual experiment with a close language, like Spanish, following the approach of (Vajjala and Rama, 2018). Concerning the machine learning techniques we employed, we are curious about changing the approach and conceive the task as regression, as in (Yannakoudakis et al., 2018). Finally, we would like to test word embeddings and a neural network model, as in (Vajjala and Rama, 2018).

Aknowledgements
This work was supported by national funds through FCT - Fundação para a Ciência e a Tecnologia postdoctoral program.

References
Alexopoulou, D., M. Michel, A. Murakami, and D. Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques: Task effects in a large-scale learner corpus. Language Learning, 03.


Gyllstad, H., J. Granfeldt, P. Bernardini, and M. Källkvist. 2014. Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity


