Automatic Readability Classifier for Portuguese Scholar Books

Gonçalo André Ramos Carvalho Pinto
goncaloapinto@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2018

Abstract

Studies show that Portuguese students, compared to the other OECD countries, have below-average reading literacy rates. This can be problematic since the ability to read plays a pivotal role not only in the student’s academic performance but also in his professional success.

Based on the assumption that reading and, consequently, comprehension and interpretation skills are acquired throughout school life, it would be interesting to verify if a possible reason for this low literacy rate could be found in the textbooks themselves, more concretely, in the absence of correspondence between the year of schooling and the level of difficulty of the texts of the corresponding textbooks.

The main objective of this project was to classify a set of texts according to their degree of difficulty using a set of linguistic parameters. These texts are part of a corpus of Portuguese textbooks from the 5th to the 12th year of schooling, kindly made available by Porto Editora.

In this way, we developed a system with the ability to evaluate the intelligibility of a text. In the first phase of this project, a study was done to evaluate the approaches and tools that are used in this context. Subsequently, a natural language processing system was developed, able to extract a set of linguistic parameters, representative of the difficulty of a given text.

Four different classifiers were developed, trained with the parameters and different combinations of scales in the corpus of school textbooks. The best classifier, on the 8 level scale (grades 5-12), obtained an accuracy of 29.16%, while for the simplified scale of 3 levels (2nd and 3rd Cycles and High School), it reached an accuracy of 59.37%.

Keywords: Readability, Readability metrics, Automatic Readability Classifier, Machine Learning.

1. Introduction

Statistics obtained from a national and international study show that Portuguese students have certain difficulties on reading domain. These skills are acquired throughout life, but it is at academic level that they are most exercised with a focus on the ability to understand, interpret and reflect upon texts. Because of this, some experts [11] claim that one way to improve this capabilities is by providing adequated materials to the students, gradually increasing the degree of difficulty as the learning process develops as well as acquiring those skills.

In this way, selecting appropriate reading materials for students is an extremely important task, which involves measuring the text readability of the texte – a measure of difficulty that distinguishes an easy from a difficult text to read, understand and interpret. Currently, this task is performed mostly manually.

This paper presents an automatic classifier for European Portuguese teaching texts based on a set of supervised learning algorithms and on a set of lexical and syntactic features, that defines the global difficulty of a text. To accomplish this, the system uses existing and new Natural Language Processing (NLP) tools, a parser, an hyphenator, a tool that extracts lexical information, and, finally, another tool, developed under this research, that extracts the word readability based on the number of senses of each word. Currently, the system extracts 55 features, grouped in 7 groups: parts-of-speech (POS), syllables, words, chunks and phrases, averages and frequencies, and some extra features.

The corpus used for the extraction of the linguistic features and the training of the classifiers is composed by the texts belonging to the scholar books of Porto Editora, regarding of 5th to 12th grades.

Two classifiers were developed to carry the classification task: one based on a eight-levels scale taken from each learning grade level, and a second based in a simplified three-levels scale representi...
ling the 2nd and 3rd cycles of the Basic education and the Secondary education.

The structure of the paper is as follows: first, in Section 2, is presented some of the related work. In Section 3, the NLP tools used in this project are described, followed by the features extracted from the text (Section 4). In Section 5, are introduced and described the automatic readability classifier herein developed, and, finally, in Section 6 are summarized the obtained results, followed by the conclusions and future work.

2. Related Work

This area of research is in constant development, since there are always aspects to improve in order to achieve the best system, the one that allows, more reliably, to identify the level of intelligibility of a text. For this reason, different approaches to this end have been developed over the years. Initially, these approaches were based on the idealization and construction of mathematical formulas, composed by simple linguistic variables, such as average sentence length, average number of syllables per word, number of words, number of different words, etc. Examples of this metrics are: the Flesch Reading Ease [8], despite being one of the first, remains the best known and one that will always have to be taken into account, the Dale-Chall formula [5] [6], the Fog index [13], the Fry Graph [9][10], the Simple Measure of Gobbledygook (SMOG) [17] and the Flesch-Kincaid formula [15]. More complex approaches followed, such as the Lexile framework, which uses a corpus of more than 600 million words, in the process of identifying the difficulty of reading texts, and the system developed by Heilman, Thompson, Callan e Eskenazi [14] in 2007, which uses language and grammatical models, based on unigrams, which after being trained are able to predict the difficulty of a text.

More recently, with the emergence and development of artificial intelligence area, namely automatic learning algorithms, the first systems based on automatic classifiers, trained from previously classified corporas and a set of languages features that allows the evaluation of non-classified texts, began to appear. An example is the REAP.PT [3] thought for European Portuguese, which has been developed from the REAP system. This system uses the method of 10-fold cross validation achieved an adjacent accuracy of 87.6% and a root mean square error (RMSE) of 0.68. Another example is the Coh-Metrix Port [19], which also has an English version [12], that uses a set of 41 linguistic metrics to train the Sequential Minimal Optimization algorithm. The evaluation of the classifier was made by determining the F-measure, whose value was 0.97.

Finally, and the most important work for this solution, since this project was a continuation of his master thesis dissertation, is the system designed by Curto [4], which system is based on a set of PLN tools, responsible for the extraction of 52 linguistic parameters. He developed two distinct classifiers, the first one predicts the readability on a five-levels scale (A1, A2, B1, B2, C1), and the second in a simplified three-levels scale (A, B, C). For the five-levels scale the best performance was achieved by the boosting algorithm, LogitBoost, with an accuracy of 75.11%, an adjacent accuracy of 91.98% and a RMSE of 0.27. In the other hand, with an accuracy of 81.44% and a RMSE of 0.346 the C4.5 grafted algorithm performed the best classification for the three-levels scale.

3. Natural Language Processing Tools

The process of gathering the 55 linguistic parameters is accomplished by PLN tools. The main tool that extracts the majority of this statistical information about the texts is a natural language processing chain named Statistical and Rule-Based Natural Language Processing chain (STRING)2 [16]. This tool is a rule-based system, composed by several modules, developed for European Portuguese, by L2F-Laboratory of Languages, at INESC-ID Lisboa, where each of the modules provides different functionalities, namely tokenization and text segmentation, POS tagging, morphosyntactic disambiguation, shallow parsing (chunking), and deep parsing (dependency extraction).

The second PLN tool is Yet Another Hyphenator (YAH) [7] which is responsible for extracting information related to the number of syllables in a word, since it has the ability to proceed to syllabic division, identifying and delimiting the syllables of a specific word. It is important to say that this procedure is done according to a set of rules.

The third one is a JAVA tool implemented by Curto [4] that extracts the lexical information of a text, namely the number of words, the number of different words and the words frequency, after the text normalization, removing all special and numeric characters.

The last tool, also developed within this project, evaluates the difficulty of a word according to its number of senses / meanings. The solution developed is divided into three distinct components:

- **Senses Database**: data model composed by the senses of the words of each lexical re-

---

1http://call.12f.inesc-id.pt/reap.public

2https://string.12f.inesc-id.pt/w/index.php
source. For each lemma present in the lexical resources a line is inserted in the dataset, once different resources can define in different ways the same word;

- **Lemmas Processing**: responsible for the filtering of lemmas, because not all should be considered for the text readability, since some of them are common in the language, such as stop words and auxiliary verbs, and for that, its senses, are irrelevant to distinguish texts with different difficulties;

- **Senses Extractor**: this module, as long as the words are in the database, is able to calculate the number of senses for a random word given as input.

### 4. Linguistic Features

The set of tools and systems described in the previous Section, allowed to extract 55 features that were important for assessing the difficulty level of a text. These parameters can be grouped taking into account the type of information that each one offers, as follows:

- **Syntagmas**: linguistic segments that express relationships between the words of a sentence. Coordination and subordination are typical examples of relationships between words. According to the literature [4], the use of coordinated rather than subordinate structures makes the process of reading and understanding simpler. The system extracts 13 different types of syntagmas;

- **Phrases and words counts**: includes measures such as the number of words, the number of different words, the number of sentences and the frequency of words. The length of the text was relevant in some previous systems, since texts with long phrases are generally more complex, making it difficult to understand;

- **Verbs features**: comprehends parameters such as the number of different verbs, the number of auxiliary verbs, the number of main verbs, and the length of verbal chains. The counting of verbs is done taking into account the different flexed forms of verbs, since the difficulty of a text is directly related to the use of different tenses and verbal forms.

- **Averages and frequencies metrics**: this features are responsible for group the averages and the frequencies of all previous statistics. The average number of verbal phrases per phrase, the average length of sentences, the average length of syllables per word and the frequency of verbs are examples of this type of features. All these parameters were relevant for the calculation of readability in previous work. In the Pitler and Nenkova approach [18] it is indicated that the longer the sentence and / or the more verbs it contains, the harder is its comprehension. On the other hand, in the Coh-Metrix-Port system [19] the frequency of verbs and nouns also proved to be important in distinguishing between complex and simple texts.

- **Syllables metrics**: the counting of the number of syllables was possible through the use of the YAH hyphenator. This class of resources was taken into account because it was very relevant and influent in the identification of text difficulty in researchs like the formulas of Fry Graph [10], Fog Index [13] and SMOG [17];

- **Word senses metrics**: the meanings of words were introduced in the final classifier through the creation of 5 new linguistic parameters, each of which characterizes the total number of meanings of a text per lexical resource. The choice of separating the number of senses by lexical resource was taken in the sense of being able to understand the separate contribution of each lexical resource to the final evaluation of the texts.

### 5. Readability Classifier

The corpus used to train the classifier consists in a set of 1142, belonging to the scholarbooks of Porto Editora, regarding of 5th to 12th grades.

In turn, the classifiers were developed for evaluation on two distinct scales: the first one of 8 levels (corresponding to a 5th - 12th schoolyear) and a second and simplified scale of 3 levels representative of 2nd and 3rd cycles (basic school) and high school.

### 6. Evaluation

The evaluation of the system was divided into two parts. The first one is responsible for identifying the most important parameters, that is, those that contributed most to the training of the classifiers. This identification was performed through the attribute selection algorithm InfoGainAttributeEval³, available in WEKA [1] toolkit which allows to evaluate the value of an attribute by measuring the gain of information relative to a given class.

The second one is responsible for measuring the performance of the developed classifiers in order to determine how accurate they are. To do that, a 10-fold cross-validation technique was used, supported by a set of performance metrics, such as accuracy (percentage of correctly classified instances), RMSE (percentage of error), ROC area, Kappa statistics (index of agreement) anf F-measure.

The following Subsections addresses the two developed classifiers and for each one is presented the above measures.

### 6.1. Readability Classifier

#### 6.1.1 Eight-levels Classifier

The best-performing learning algorithm was the Logistic Regression (Table 1).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>20.14%</td>
<td>0.43</td>
</tr>
<tr>
<td>Random Forest</td>
<td>26.71%</td>
<td>0.318</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>28.02%</td>
<td>0.38</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>27.67%</td>
<td>0.32</td>
</tr>
<tr>
<td>Decision Strumps</td>
<td>20.67%</td>
<td>0.33</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>25.83%</td>
<td>0.33</td>
</tr>
<tr>
<td>K-NN learner</td>
<td>22.68%</td>
<td>0.44</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>20.67%</td>
<td>0.33</td>
</tr>
<tr>
<td>Holte’s OneR</td>
<td>21.02%</td>
<td>0.44</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>29.16%</td>
<td>0.32</td>
</tr>
<tr>
<td>K*</td>
<td>19.44%</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 1: Algorithms comparison results (eight-levels classifier).

On the other hand, the performance measures obtained in the algorithm evaluation process are represented in Table 2. In this eight-level classifier we also considered the adjacent accuracy within 1 grade level as a useful evaluation metrics, because for classifiers with large classification scales, the normal accuracy is a very rigorous measure, since the difficulty in manually assigning a readability to a given text is commonly know among linguists, sometimes causing disagreements between them.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>RMSE</th>
<th>AUC</th>
<th>Kappa</th>
<th>Adj. Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.16%</td>
<td>0.32</td>
<td>0.73</td>
<td>0.19</td>
<td>58.23%</td>
</tr>
</tbody>
</table>

Table 2: Evaluation of the readability classifier (eight-levels).

In this scenario, the classifier correctly classified 333 of 1142 texts, reached an accuracy of 29.16% and a RMSE of 0.32. Regarding the value of AUC (0.73) obtained, the system is considered a fair classifier [2].

The confusion matrix resulting from the classification is shown in Table 3.

#### 6.1.2 Three-levels Classifier

In this scenario, the algorithm that performed the best classification was the Support Vector Machines, followed by Random Forest with slightly worse values of accuracy (Table 4).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>50.18%</td>
<td>0.56</td>
</tr>
<tr>
<td>Random Forest</td>
<td>59.11%</td>
<td>0.41</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>55.25%</td>
<td>0.52</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>59.37%</td>
<td>0.43</td>
</tr>
<tr>
<td>Decision Strumps</td>
<td>51.40%</td>
<td>0.44</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>57.88%</td>
<td>0.41</td>
</tr>
<tr>
<td>K-NN learner</td>
<td>47.64%</td>
<td>0.59</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>51.40%</td>
<td>0.45</td>
</tr>
<tr>
<td>Holte’s OneR</td>
<td>51.14%</td>
<td>0.57</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>57.88%</td>
<td>0.41</td>
</tr>
<tr>
<td>K*</td>
<td>20.23%</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Table 4: Algorithms comparison results (three-levels classifier).

In this case, the best-performing learning algorithm classified correctly 678 texts. Compared with the previous one, it was expected that the results would improve, because the classification scale was considerably reduced in five-levels. With this idea in mind, the accuracy obtained was 59.37%, an RMSE of 0.43, an AUC of 0.72 and a kappa statistic of 0.38. In experiments of this type, where the classification scale is only three-levels, the adjacent accuracy within 1 grade level is not calculated, since if we consider the intermediate level (2nd cycle of the basic education, in the context of this work), it would always obtain the maximum value for performance, since it would include the remaining two levels of difficulty. This values are represented in the Table 4.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>RMSE</th>
<th>AUC</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>59.37%</td>
<td>0.43</td>
<td>0.72</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Table 5: Evaluation of the readability classifier (three-levels).
<table>
<thead>
<tr>
<th>Actual</th>
<th>Observed level</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>148</td>
<td>103</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>94</td>
<td>207</td>
<td>125</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>29</td>
<td>106</td>
<td>312</td>
</tr>
</tbody>
</table>

Table 5: Confusion Matrix (three-levels).

### 6.2. Feature Contribution

The most significant parameters for the evaluation are represented in Figure 1 and Figure 5.

#### 6.2.1. Feature Contribution Chart

![Feature contribution chart for the eight-levels scale classification.](chart1.png)

Figure 1: Feature contribution for the eight-levels scale classification.

![Feature contribution chart for the eight-levels scale classification.](chart2.png)

Figure 2: Feature contribution for the eight-levels scale classification.

Observing the figures, the parameters that contributed most to the classification process were repeated in both classifiers, especially the Flech Reading Ease formula, the average syllables per word, the average sentence length, the number of nodes and the number of sentences. The main differences between these classifiers and the others were the use of the parameters related to the number of word senses, which allows to validate, that the degree of polysemy of the words can be related, in a relevant way, with the general difficulty of the texts. The parameter that counted the number of senses using the lexical resource ViPER was the one that had the greatest impact, followed by, despite the slightest less influence, the parameters related to Infopedia and TemaNet. PAPER and MWN.PT had no influence on classification.

Additionally, for a more detailed analysis of the contribution of these features, boxplots were built, by educational level, in order to observe the variation of each one, along the different readability levels.

![Boxplot of number of senses (ViPER) value variations between the different readability levels.](boxplot1.png)

Figure 3: Number of senses (ViPER) value variations between the different readability levels.

![Boxplot of number of senses (TemaNet) value variations between the different readability levels.](boxplot2.png)

Figure 4: Number of senses (TemaNet) value variations between the different readability levels.

![Boxplot of number of senses (Infopedia) value variations between the different readability levels.](boxplot3.png)

Figure 5: Number of senses (Infopedia) value variations between the different readability levels.

### 7. Conclusions and Future Work

The main purpose of this paper was to present and describe a valid and reliable solution for the automatic readability classification to assist the selection of adequate reading materials for teaching Eu-
European Portuguese.

The architecture presented here was divided into three main components: the text analysis module; the training module; and the classification module.

The text analysis module was responsible for extracting the linguistic parameters used by the automatic classifier. To do this, we used a set of PLN tools that, in total, extracted 55 linguistic parameters. Of these, it was possible to verify that the most significant linguistic factors for the classification process are those obtained through the most complex PLN systems, emphasizing their importance in relation to the others, which are obtained through simple counts of words and phrases. Another idea that was validated was the hypothesis stated in the objectives of the study, which stated that the degree of polysemy of the words could be related, in a relevant way, to the general difficulty of the texts, since three of the parameters that related to this type of characteristics were among the 15 most significant parameters for the classification of the difficulty of the texts.

In both scenarios, with eight readability levels or with three-levels, the classifiers here developed achieved worse results compared to previous studies. If we look at the results obtained by Curto [4], this differences can be explained by the large variance in the total number of texts used for the training of the two classifiers, although the texts belong to the same sample corpus.

In order to develop a system that allows with high reliability evaluate the level of readability of a text it is necessary that the research work continues to be done, looking for new and different solutions that bring significant improvements to this area of PLN. In this sense, the inclusion of new tools of PLN that allow to gather other type of characteristics, that are shown to be significant for the process of evaluation is something that can be taken into account.

References


