Automatic Construction of Large Readability Corpora

Jorge Alberto Wagner Filho, Rodrigo Wilkens and Aline Villavicencio
Institute of Informatics, Federal University of Rio Grande do Sul
Av. Bento Gonçalves, 9500, 91501-970, Porto Alegre, RS, Brazil
{jawfilho,rodrigo.wilkens,avillavicencio}@inf.ufrgs.br

Abstract

This work presents a framework for the automatic construction of large Web corpora classified by readability level. We compare different Machine Learning classifiers for the task of readability assessment focusing on Portuguese and English texts, analysing the impact of variables like the feature inventory used in the resulting corpus. In a comparison between shallow and deeper features, the former already produce F-measures of over 0.75 for Portuguese texts, but the use of additional features results in even better results, in most cases. For English, shallow features also perform well as do classic readability formulas. Comparing different classifiers for the task, logistic regression obtained, in general, the best results, but with considerable differences between the results for two and those for three-classes, especially regarding the intermediary class. Given the large scale of the resulting corpus, for evaluation we adopt the agreement between different classifiers as an indication of readability assessment certainty. As a result of this work, a large corpus for Brazilian Portuguese was built[1], including 1.7 million documents and about 1.6 billion tokens, already parsed and annotated with 134 different textual attributes, along with the agreement among the various classifiers.

1 Introduction

Text readability assessment refers to measuring how easy it is for a reader to read and understand a given text. In this context methods for automatic readability assessment have received considerable attention from the research community (DuBay, 2004). The task of attributing a readability level to a text has a wide range of applications, including support for student reading material selection (Petersen and Ostendorf, 2009) or help for clinical patients (Feng et al., 2009). It can also be used for ensuring that instructions and policies are written in an easily comprehensible way even for readers with low education (McClure, 1987). It can also contribute to the task of text simplification, evaluating the obtained version to indicate if further simplification is needed (Aluisio et al., 2010). Recently, authors such as Petersen and Ostendorf (2009), Vajjala and Meurers (2014) and Scarton et al. (2010) have started treating this task as one of text classification, using corpora manually annotated with readability classifications to train automatic learning models, based on a large set of text metrics, including deeper features, for example derived from n-gram language models and parse trees. However, an important limitation to this approach is the small availability of reliably annotated train data. Moreover, this task is known to be very subjective, and even human annotators present a high disagreement rate in their evaluations (Petersen and Ostendorf, 2009).

In this work, we aim to develop large corpora classified by readability levels. To achieve this objective we present a study of different Machine Learning approaches to the task of readability assessment of texts, focusing on Portuguese, and apply the relatively recent concept of building corpora from the Web (Bernardini et al., 2006) to automatically generate large corpora classified by readability levels. For that, we follow the framework proposed by Wagner Filho et al. (2016), where a readability classifier

[1] http://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC

Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity,
pages 164–173, Osaka, Japan, December 11-17 2016.

was incorporated into a crawler, but changing the classifier position in the pipeline so that we can work with both low and high-cost complexity features. We also experiment with learning models trained on several different reference corpora, in both Portuguese and English, and investigate the relevance of the agreement between them.

We focus our study on two hypothesis: (H1) a learning model trained in a reference annotated corpus is able to classify a new corpus so that its classes present significant linguistic differences and (H2) the use of syntactic attributes contributes to a better classification. As a result of this work, a new Portuguese corpus of around 1.6 billion tokens was built and annotated with four different readability classifiers. This paper is structured as follows. In Section 2, we discuss some relevant work in the literature and, in Section 3, we present our materials and methods, especially our training data, features and classification algorithms and our Web corpus collection framework. In Section 4, we apply our methodology and validate our hypothesis trough a series of experiments. Finally, in Section 5 we present our conclusions and ideas for future work.

2 Related Work

Some traditional works in readability classification include, for example, Flesch and others (1946), Dale and Chall (1948), Gunning (1952), which were based in shallow textual measures. A very known example, the Flesch Reading Ease index (Flesch and others, 1946), uses the number of syllables per word and words per sentence to determine lexical and syntactic complexity. For Portuguese, Martins et al. (1996) adapted the Flesch Reading Ease index to account for language differences, and Érica Sapgnolo and Finatto (2014) analysed Brazilian news texts to generate lists of simple words.

The classical readability measures have been criticized for applying a superficial analysis of textual characteristics, ignoring, for example, that larger sentences may be clearer and more explicative than a smaller equivalent (Williams, 2004). These formulas are not able to capture several elements of cohesion and textual difficulty, according to McNamara et al. (2002), who also point that these tools force editors to modify the text to increase the calculated readability, but actually reducing cohesion. Recent studies tried to apply automatic approaches that better approximate the complexity of a text, for example using n-gram language models to identify reading ease. Petersen and Ostendorf (2009) trained Support Vector Machines using a corpus created from an educational newspaper, Weekly Reader, with different versions for four different grade levels, completed with articles for adults from the Associated Press. They worked with lexical and syntactic features and also with traditional formulas. Investigating the contribution of syntactic features, it was observed that they were not good enough separately, but contributed to the general performance. Complementarily, Vajjala and Meurers (2014) applied 152 lexical and syntactic attributes to classify a corpus of subtitles from different BBC channels for children and adults, also using SVMs. The most predictive attribute was shown to be the age of acquisition. Similar approaches were applied in multiple other languages, including Italian (Dell'Orletta et al., 2011), German (Hancke et al., 2012) and Basque (Gonzalez-Dios et al., 2014). In Portuguese, Scarton and Aluísio (2010) classified articles for children and adults from local newspapers, using a SVM trained on 48 psycholinguistic features and obtaining an F-measure of 0.944.

Wagner Filho et al. (2016) proposed a framework to take advantage of the increasing availability of language content in the Web to create large repositories of text suitable for different reading levels, incorporating a readability classifier to a pipeline similar to the one used by Baroni et al. (2009) to build a series of large Web corpora, such as the ukWaC (Baroni et al., 2009), composed by 1.91 billion tokens and 3.8 million types. This pipeline is composed by four steps: (1) identification of an appropriate set of seed URLs, (2) post crawling cleaning, (3) detection and removal of near duplicate content and (4) annotation. Considering the large corpora size made possible by these approaches and the current limitations in readability assessment, in this paper, we build on these works, experimenting with different training corpora in both English and Portuguese, and applying the resultant learning models to a large collection of documents obtained from the Web.

3 Materials and Methods

In this section, we describe the series of resources and the methodology that were applied to achieve our objective of developing large corpora classified by readability levels. Sections 3.1, 3.2 and 3.3 present, respectively, our corpora, classification models and readability features, while Section 3.4 presents the proposed Web corpora collection framework.

3.1 Corpora

We selected a series of corpora (described in Table 1) to represent different levels of readability. Aiming to avoid language specific readability issues, we explored corpora for both Portuguese (Wikilivros, ESOC, PSFL, ZH, BrEscola) and English (Wikibooks, SW, BB). *Wikilivros* (Elementary School, High School and College levels) was constructed in Wagner Filho et al. (2016). The É Só o Começo (ESOC) corpus contrasts classic literature works in Portuguese with adapted versions for modern language use. *Para o Seu Filho Ler (PSFL)* e *Zero Hora (ZH)*, corpora of news articles, were constructed by Aluisio et al. (2010), the former comparing articles for children with articles for adults, and the latter comparing original articles for adults with two different levels of simplification (natural and strong). *Brasil Escola* (BrEscola), a corpus of educational materials for children and teenagers, *Wikibooks*, a corpus of virtual books for readers of different proficiency levels (Beginner, Intermediary, Advanced and Professional), and *Britannica Biographies* (BB), a corpus of biographies with versions in three different readability levels (Elementary, Medium and High), were collected especially for this study, crawling different sections of the websites of the same names[2]. The *Simple Wikipedia* (SW) corpus was compiled by Coster and Kauchak (2011), pairing articles from the English and Simple English versions of Wikipedia[3].

In the case of the corpora Wikilivros, ZH, Wikibooks and BB, which consider more than two readability levels, tests were also done with adapted binary versions, in order to verify the impact of the number of classes in the classifier performance. For that, in Wikilivros and BB, the most simple and most difficult levels were selected. In ZH, we used the original and the natural simplification class, since the strong simplification was exaggerated for our classification purposes, and, in Wikibooks, we discarded the Beginner class, which was too small, and grouped the Advanced and Professional classes.

Language	Corpus	Classes	Documents	Sentences	Types	Tokens
PT	Wikilivros	3	78	38,865	54,462	636,309
	ESOC	2	130	21,667	32,180	442,391
	PSFL	2	259	3,075	8,628	51,963
	ZH	3	279	7,127	8,511	107,930
	BrEscola	2	9,083	200,132	95,928	3,516,097
EN	Wikibooks	4	35	65,704	24,638	897,971
	SW	2	4,480	515,230	183,824	10,384,518
	BB	3	2,385	101,149	45,687	1,747,733

Table 1: Description of the readability corpora

3.2 Classification models

We worked with the Weka Machine Learning tool (Hall et al., 2009) for generating classification models, especially with its implementations SMO, from the Sequential Minimal Optimization algorithm for SVM training (Platt, 1998), SimpleLogistic, for construction of linear logistic regression models (Landwehr et al., 2005), DecisionStump, for the one level decision tree (Iba and Langley, 1992), and RandomForest, for construction of a forest of decision trees (Breiman, 2001). All models performances were evaluated using F-measure and 10 fold cross-validation. These models represent a variety of approaches to text classification, and also allow us to evaluate any possible algorithm bias in the task.

[2]http://brasilescola.uol.com.br, https://en.wikibooks.org and http://school.eb.com

[3]We used here only articles that presented more than 30 sentences in both versions.

3.3 Readability features

Given our intention of assessing the contribution of different categories of language features, a large set of 134 different language attributes for Portuguese and 89 for English was selected. From basic counts, we worked with numbers of sentences, words, syllables, letters and types. We also calculated the average number of words per sentence (WPS), syllables per word (SPW) and the Type-Token Ratio, a measure of lexical diversity. The average and standard deviation of letters per word (AWL) were also used, based on the hypothesis that more complex texts are more prone to present larger words, given the more frequent presence of prefixes and suffixes, which aggregate new meaning to words.

From classical readability metrics, we used the Flesch Reading Ease (English and Portuguese versions) (Flesch and others, 1946; Martins et al., 1996), the Coleman-Liau Index (Coleman and Liau, 1975), the Flesch Grade Level, the Automated Readability Index (Senter and Smith, 1967), Fog (Gunning, 1952), SMOG (Mc Laughlin, 1969) and, for English, the Dale-Chall Formula (Dale and Chall, 1948) as well.[4]

In order to account for word ambiguity, a metric based on the hypothesis that more commonly used words, and therefore easier to understand (Vajjala and Meurers, 2014), tend to present multiple meanings in a language, we used the average number of senses from BabelNet (Navigli and Ponzetto, 2010), for Portuguese, and WordNet (Miller, 1995), for English. Moreover, following Si and Callan (2001), we worked with the average frequency in a general corpus (AFGC) and standard deviation as frequency measures, based on the hypothesis that words with higher frequencies in a general corpus tend to be more knownand, therefore, included in more levels of texts, while rarer words are more inclined to be restricted to more complex levels.

We also worked with a series of closed word lists to count word classes (*stopwords*, prepositions, articles, pronouns, personal and possessive pronouns (*PP*), conjunctions and functional words), particles ("*e*", "*ou*" and "*se*", in Portuguese, and their respective equivalents in English "*and*", "*or*" and "*if*") (McNamara et al., 2002) and simple words. For this last category, we used the lists DG and CB[5], DG+CB, CHILDES (MacWhinney, 2000) and the concatenation of all, in Portuguese. In English, we used the lists Oxford 3000, Dale-Chall (Dale and Chall, 1948), CHILDES (MacWhinney, 2000) and once again the concatenation of all of them. Simple word lists are a traditional resource in text difficulty assessment, having been notably used by Dale and Chall (1948) and also by Petersen and Ostendorf (2009), who used lists of frequent words in the lower class for a similar purpose. Finally, the incidence of unknown words (Unknown)[6] was used as a indicative of rarer, more complex vocabulary, possibly domain-specific.

Finally, we worked with counts based based in syntactic analysis, including part of speeches (18 for Portuguese and 20 for English) and dependency tags (72 for Portuguese and 27 for English), besides 7 measures of verb analysis, including verb transitivity, passive voice, average number of modifiers and average sub-categorization frame length. In Portuguese, we also analysed the incidence of verbs in the imperative mood. All these counts are frequently used as indicators of syntactic complexity, according to the online tool Coh-Metrix (McNamara et al., 2002). The parsers Palavras (Bick, 2000) and Rasp (Briscoe et al., 2006) were used to obtain these features for Portuguese and English, respectively.

Since we want to assess the contribution of different feature categories, a few specific groups were defined. The selected groups were: *sub-categorization* (transitivity, average number of modifiers, average sub-categorization frame length), classical *readability* formulas, *descriptors* (counts of sentences, words, syllables, letters and types and TTR) and *corpora-based* (incidence of unknown words, average frequency in a general corpus, lists of simple words). Moreover, we also divided our complete feature sets in three categories according to their computational costs: *shallow* (counts and lists), *medium* (part-of-speech tagging dependent) and *deep* (parsing or WordNet dependent).

[4]For the different versions of the Flesch formulas we computed the number of syllables in a word using, for English, an approximation based on the number of vowels and, for Portuguese, a rule-based syllabification tool (Neto et al., 2015).

[5]Available at `http://www.ufrgs.br/textecc/porlexbras/porpopular/massafiles/Lista_FINAL_MASSA.pdf` (Érica Sapgnolo and Finatto, 2014)

[6]We consider unknown words all words not present in a list (3 million words for Portuguese and 840 thousand for English).

3.4 Web corpora collection framework

In order to build a large corpus from Web content, we followed the pipeline approach from Bernardini et al. (2006), incorporating a readability classifier as in Wagner Filho et al. (2016). We followed the approach from Boos et al. (2014), submitting pairs of words with medium frequency in a general corpus to a search engine API, and obtaining the ten first results for each query. These results were then expanded collecting all the links contained in them as well. For the collection, cleaning and near duplicate content removal phases, we used the implementation provided by Ziai and Ott (2005), with some adaptations, such as the adoption of a more efficient text cleaning tool, *jusText* (Pomikálek, 2013). The Palavras (Bick, 2000) parser was used to enrich the corpus with syntactic annotation.

4 Experiments

In this section we present a comparison of different features categories (Section 4.1); evaluation of classification models (Section 4.2); and assess the generalisation of these models (Section 4.3). Finally, in Section 4.4, we create an large corpus of Web content, which we classify with our models in Section 4.5.

4.1 Feature analysis

In order to determine the most relevant features and also observe the effect of varying the training data used in our models, we worked with the entropy-based algorithm information gain[7]. We used the feature groups (defined in Section 3.3) to perform a more detailed evaluation, based in the average rank of the features when ordered decreasingly according to information gain. The results are presented in Table 2 , which corroborated to our previous observation that classical formulas have a great relevance in English but not in Portuguese, while textual descriptors presented a good classificatory power in both languages. Another noticeable pattern was that in both languages shallow features outperformed deep ones.

It was observed that most of the training corpora, in both studied languages, exhibited a large quantity of shallow attributes (e.g. basic counts) and readability formulas amongst the most predictive. Shallow attributes are indeed known to be good indicators, being this the reason for the creation of the classical formulas. However, this was especially observed in the corpora built trough manual simplification of text content (ESCO, PSFL, ZH, SW and BB), what supports McNamara et al. (2002) claims of the excessive influence of these metrics in authors of simple texts. Another possible interpretation, contrasting English and Portuguese corpora, is that the classical formulas, having been created focusing on the former, are good classifiers for it, but in Portuguese present a poorer performance. A noticeable exception to this behaviour was the pair of corpora Wikilivros and Wikibooks, which did not present almost any shallow metric amongst the twenty more relevant. This may be attributed to the collaborative approach adopted by the websites which gave origin to these corpora, resulting in a classification produced by regular users, who may consider other factors besides the language complexity in their assessments, such as the nature of the content. The great relevance of simple word lists in classifying Wikibooks also indicates the attention of users to the vocabulary. Therefore, the use of these corpora to train readability classifiers may lead to over-fitting on not necessarily relevant textual characteristics. Another factor to be considered in these corpora is that they present closer classes with fuzzy borders (more advanced High School texts and simpler College texts, for example, may be very similar), what may indicate that deep attributes are relevant to a more accurate classification.

4.2 Model performance analysis

We trained classifiers (discussed in Section 3.2), and the linear logistic regression algorithm, SimpleLogistic, presented, overall, the best results for both languages, possibly due to its built-in feature selection (Table 3 presents the F-measures for this algorithm[8]). The DecisionStump algorithm, which constructs a one-level decision tree with the most significant feature, offered an interesting baseline, achieving good results in some corpora, but not necessarily generalisable. The algorithms RandomForest, which constructs a forest of complete decision trees, and SMO, also achieved good results. As expected, the

[7] We used the InfoGainAttributeEval implementation available in the Weka toolkit.

[8] For reasons of space, we omit the results for the remaining ones.

	Wikil.	PSFL	ZH	BrEsc	ESOC	Avg Pt	Wikib.	SW	BB	Avg En
Shallow	49,98	39,68	41,91	43,01	43,99	43,71	25,53	33,21	32,23	30,32
Medium	24,51	54,90	28,43	40,03	71,17	43,80	73,41	55,84	53,67	60,97
Deep	83,58	82,17	86,36	83,61	77,11	82,56	50,41	51,55	53,68	51,88
Subcat.	46,53	84,00	29,60	82,90	38,62	56,33	26,18	64,45	62,45	51,03
Formulas	48,03	21,88	6,83	65,22	88,12	46,01	30,40	7,03	5,50	14,31
Descrip.	63,12	3,53	32,92	25,52	8,12	26,64	39,92	11,42	17,28	22,87
C. based	49,73	54,06	60,01	40,30	29,29	46,67	21,91	51,76	50,30	41,32

Table 2: Average rank of feature classes in the different training corpora (smaller values indicate a bigger relevance to a given class)

intermediary classes were always the most difficult to classify correctly, but still presented reasonable performance. Comparing the tests with the two versions of the ZH corpus, a great negative impact in performance was observed when considering three classes. Comparing shallow and deep attributes, we observed that the former tend to present a good classificatory power with a low computational cost but, in five out of eight scenarios, the performance was enhanced with the combination of both categories, confirming the results of François and Miltsakaki (2012) that shallow attributes are great indicators of readability while the combination with deeper attributes is positive, and also our hypothesis H2.

Lang.	Corpus	All	Shallow	Medium	Deep	Formulas	Descriptors
PT	Wikilivros	*0.71 (0.24)*	**0.75 (0.15)**	*0.67 (0.24)*	*0.69 (0.23)*	*0.59 (0.23)*	*0.59 (0.26)*
	ESOC	*0.98 (0.03)*	**0.99 (0.02)**	*0.96 (0.03)*	*0.98 (0.03)*	0.69 (0.10)	0.90 (0.07)
	PSFL	**0.99 (0.01)**	*0.98 (0.01)*	0.81 (0.10)	**0.99 (0.01)**	0.80 (0.09)	*0.98 (0.01)*
	$ZH_{2\ levels}$	**0.89 (0.08)**	*0.82 (0.13)*	*0.82 (0.06)*	*0.83 (0.04)*	*0.80 (0.10)*	*0.83 (0.12)*
	$ZH_{3\ levels}$	**0.63 (0.04)**	0.55 (0.04)	*0.56 (0.08)*	0.53 (0.07)	*0.58 (0.11)*	*0.61 (0.08)*
	BrEscola	**0.81 (0.01)**	0.77 (0.01)	0.65 (0.02)	0.67 (0.01)	0.66 (0.03)	0.67 (0.03)
EN	Wikibooks	*0.48 (0.25)*	*0.51 (0.26)*	*0.54 (0.33)*	*0.49 (0.15)*	**0.75 (0.24)**	*0.49 (0.28)*
	SW	**0.92 (0.01)**	*0.91 (0.01)*	0.82 (0.02)	0.88 (0.01)	0.88 (0.01)	0.89 (0.01)
	BB	**0.86 (0.02)**	*0.83 (0.03)*	0.62 (0.02)	0.80 (0.02)	0.80 (0.02)	0.79 (0.02)

Table 3: Average F-measures and standard deviations for the regression classifiers trained in different sets of features (best results are bold, and the italic if for not statistically different from best result)

4.3 Generalisation analysis

An important concern when training a classification model is how much this model will be able to be generalized beyond the training data. This is especially relevant in this context, where the training data are, by definition, very limited in volume, while the quantity of data we want to classify is very large. Initially, we analysed the compatibility between the different models. Using the lists of features ordered by information gain obtained in Section 4.1, we assessed the Spearman rank correlation between the different corpora. These results were very weak, indicating very little similarity between them. The closest corpora were Wikilivros and ZH, with a correlation of 0.62.

In a complementary analysis, we implemented projection tests, testing in a corpus a simple logistic model trained in another, in all possible combinations of our corpora in Portuguese[9]. For this tests, we worked with models trained in the binary versions of all corpora. The results indicted once again little agreement between the models, but were coherent with their individual characteristics. For example, classifiers trained in corpora with a higher complexity threshold between the classes (Wikilivros and ZH) classified most documents of corpora for children/teenagers (PSFL, BrEscola) as simple, and vice versa.

[9]We excluded the ESOC corpus, considering that this corpus presents language differences that are not resultant only from different complexity but also from differences between language use nowadays and the time of the original works.

The only exception for this behaviour was the projection of the model trained in BrEscola onto Wikilivros, which presented null performance in the upper level, opposed to the expected. Considering the weak agreement observed between the different models, we decided to employ the agreement between all of them in order to obtain a more generalisable classification (Enríquez et al., 2013).

4.4 Web corpus collection

Following the framework established in Section 3.4, we started with six thousand random pairs of average frequency words from a frequency list[10] obtained from several corpora from the Linguateca repository, after the removal of stopwords. These pairs were submitted to the Microsoft Bing API, and the resultant sixty thousand URLs were expanded by breadth first recursion in two levels, producing around twenty four million seeds of ".br" extension, as in Boos et al. (2014). These were then processed, resulting in corpus with 1.56 billion tokens, 4.15 million types (TTR 0.0026). All the documents were annotated with Palavras parser and the 134 different features described in Section 3.1.

4.5 Large corpus classification

In this section, we present the results of the classification of the collected Web corpus with the learning models trained in Section 4.2. We chose to apply here the models generated with the SimpleLogistic algorithm, since, besides having presented a better performance, they are easily implementable through a series of regression equations operating with the attribute values calculated by the last module of our pipeline. We work with the models trained in the whole set of attributes, and only those trained in two class scenarios, given the performance limitations observed in three class classifiers. The results are presented in Table 4, and as we already expected, considering the projection tests, the agreement between our models was small. Only 126,245 (7.5%) documents were classified as simple by all models, while only 17,634 (1%) were unanimously difficult. Discarding the BrEscola classifier, the remaining three agreed in 210,879 (12.5%) as simple and 149,279 (8.8%) as difficult[11]. This classifier presented an unexpected behaviour, since, considering its low complexity threshold (texts for children against texts for teenagers), we expected most documents to be classified as difficult. Moreover, it had already presented the poorest performance in the cross-validation tests in Section 4.2 and an unexpected behaviour in the projection tests in Section 4.3.

Model	Simple documents	Difficult documents
SimpleLogistic$_{PSFL}$	613,877 (36.4%)	1,076,173 (63.6%)
SimpleLogistic$_{ZH}$	448,199 (26.5%)	1,241,851 (73.5%)
SimpleLogistic$_{Wikilivros}$	1,413,211 (83.7%)	276,839 (16.3%)
SimpleLogistic$_{Brasil\ Escola}$	1,417,339 (83.9%)	272,711 (16.1%)

Table 4: Behaviour of the different classifiers in the collected Web corpus

Due to the size of the resulting corpus a qualitative analysis of the classification was done in terms of the behaviour of the different readability indicator features in the corpus classified with the different models presented in Table 4. Additionally we also analysed the agreement between all models. For that, all feature values were normalized to enable the comparison per feature category, as seen in Table 5. We also show the differences observed in the PSFL training data, for reference purposes. This corpus was selected due to its performance in the cross-validation tests, and because produced the most balanced document distribution. All classifiers resulted in significant differences between simple and complex documents in all categories of features (114 out of the 134 features presented statistically significant differences in all models with $p<0.01$, and 121 with $p<0.05$), confirming our hypothesis H1. It is important to note that, even though these indicators were part of our initial set of features, they were not necessarily the ones on which the model was trained, since only 10 features were selected by the simple

[10] `http://dinis2.linguateca.pt/acesso/tokens/formas.totalbr.txt`

[11] PSFL and ZH agreed in 245,431 (14%) documents as simple and 873,405 (52%) as difficult, while PSFL and Wikilivros agreed in 519,870 (31%) simple and 182,832 (11%) difficult and ZH and Wikilivros in 380,094 (22%) and 208,734 (12%).

logistic built-in feature selection. While in the individual models, the differences between the final classes were, more subtle than in the training data, considering the agreement between all classifiers, the observed differences were consistently larger. This result indicates that using more than one model can be a relevant approach for Web content classification, offering a more strict classification, less prone to over-fitting to the characteristics of a given corpus, especially considering the frequent small size of the available readability-annotated training data.

Category	PSFL train data	PSFL	ZH	Wikilivros	3-model agreement
Shallow	0.15	0.15	0.08	0.11	0.27
Medium	0.30	0.06	0.09	0.09	0.17
Deep	0.29	0.10	0.10	0.16	0.23
Sub-categorization	0.07	0.03	0.06	0.05	0.10
Formulas	0.19	0.06	0.09	0.08	0.20
Descriptors	0.16	0.62	0.27	0.11	0.82
Corpora-based	0.09	0.04	0.01	0.12	0.09

Table 5: Average difference per feature category between the simple and complex classes in the PSFL train corpus and in different classifications of our Web corpus

5 Conclusion

In this work, we presented a comparative study of different machine learning approaches to the task of readability assessment of texts in Portuguese and English, working with a framework for automatic generation of large corpora classified by readability from the Web. We observed that, as previously found in the literature for English (François and Miltsakaki, 2012), shallow, low computational cost features present a very good classificatory performance, although the complete set including deeper features outperforms them in most cases, validating our hypothesis (H2), that complex text attributes contribute to the classification according to readability levels in Portuguese. Nonetheless, in a comparison with English, we observed that classical formulas, based in these shallow features, tend to present more relevance in that language, which is explainable since they were developed focusing on its characteristics. The logistic regression presented the best classification results overall, although there was a great performance difference between classifiers for two and three levels, especially when it comes to the intermediary class, showing the difficulty of this task in a non-binary context. Finally, regarding the generalization of the classifiers, there was disagreement between the models trained in different reference corpora, reflecting the connection between the model training and the desired classification in the final corpus.

We applied the proposed methodology and the generated models in a large scale, observing significant differences between the classes in the collected Web corpus, for several indicators of readability. These differences were even greater when we considered only the documents in which three different models agreed in the classification, demonstrating the benefits of applying multiple models simultaneously to improve text classification (this will also contribute for the next phase of the project, a manual assessment by linguists). This confirmed our hypothesis (H1) that a learning model trained in a reference annotated corpus is capable of classifying a new corpus satisfactorily. The contributions of this work also include the large Web corpus produced and classified by four different learning models with different characteristics, which can contribute to further studies. As future work, new analysis must be done over the characteristics of the different documents classes in the classified corpus, including the manual sample assessment by linguists and a more fine-grained assessment of documents. Moreover, this approach can be straightforwardly expanded to develop large readability corpora for other languages.

Acknowledgements

This research was partially developed in the context of the project *Text Simplification of Complex Expressions*, sponsored by Samsung Eletrônica da Amazônia Ltda., in the terms of the Brazilian law n. 8.248/91, and by CNPq (400715/2014-7 and 312114/2015-0).

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Silvia Bernardini, Marco Baroni, and Stefan Evert. 2006. A wacky introduction. *WaCky*, pages 9–40.

Eckhard Bick. 2000. *The parsing system" Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework*. Aarhus Universitetsforlag.

Rodrigo Boos, Kassius Prestes, Aline Villavicencio, and Muntsa Padró. 2014. brWaC: a WaCky corpus for Brazilian Portuguese. In *Computational Processing of the Portuguese Language*, pages 201–206. Springer.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the rasp system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80. Association for Computational Linguistics.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

William Coster and David Kauchak. 2011. Simple English Wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 665–669. Association for Computational Linguistics.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability. *Educational research bulletin*, pages 37–54.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83. Association for Computational Linguistics.

WH DuBay. 2004. The principles of readability. *Costa Mesa, CA: Impact Information (2004)*.

Fernando Enríquez, Fermín L Cruz, F Javier Ortega, Carlos G Vallejo, and José A Troyano. 2013. A comparative study of classifier combination applied to nlp tasks. *Information Fusion*, 14(3):255–267.

Érica Sapgnolo and Maria José B. Finatto. 2014. Buscando delinear um vocabulário básico: comparaço de duas listas de frequencia de palavras - jornais populares e linguagem geral. *Available at http://www.ufrgs.br/textecc/porlexbras/porpopular/massafiles/Lista_FINAL_MASSA.pdf*.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237. Association for Computational Linguistics.

Rudolf Franz Flesch et al. 1946. *Art of Plain Talk*. Harper.

Thomas François and Eleni Miltsakaki. 2012. Do nlp and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57. Association for Computational Linguistics.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza Sánchez, and Haritz Salaberri. 2014. Simple or complex? assessing the readability of basque texts. In *COLING*, pages 334–344.

Robert Gunning. 1952. The technique of clear writing.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *COLING*, pages 1063–1080.

Wayne Iba and Pat Langley. 1992. Induction of one-level decision trees. In *Proceedings of the ninth international conference on machine learning*, pages 233–240.

Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic model trees. *Machine Learning*, 59(1-2):161–205.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Teresa BF Martins, Claudete M Ghiraldelo, Maria das Graças Volpe Nunes, and Osvaldo Novais de Oliveira Junior. 1996. *Readability formulas applied to textbooks in brazilian portuguese*. Icmsc-Usp.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Glenda M McClure. 1987. Readability formulas: Useful or useless? *Professional Communication, IEEE Transactions on*, (1):12–15.

Danielle S McNamara, Max M Louwerse, and Arthur C Graesser. 2002. Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Technical report, Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Nelson Neto, Willian Rocha, and Gleidson Sousa. 2015. An open-source rule-based syllabification tool for brazilian portuguese. *Journal of the Brazilian Computer Society*, 21(1):1–10.

Sarah E Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.

J. Platt. 1998. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

J Pomikálek. 2013. justext: Heuristic based boilerplate removal tool. *Available: Google code, online http://code. google. com/p/justext*.

Carolina Evaristo Scarton and Sandra Maria Aluísio. 2010. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61.

Carolina Scarton, Caroline Gasperin, and Sandra Aluisio. 2010. Revisiting the readability assessment of texts in portuguese. In *Advances in Artificial Intelligence–IBERAMIA 2010*, pages 306–315. Springer.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, DTIC Document.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM.

Sowmya Vajjala and Detmar Meurers. 2014. Exploring measures of "readability" for spoken language: Analyzing linguistic features of subtitles to identify age-specific tv programs. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, pages 21–29.

Jorge Wagner Filho, Rodrigo Wilkens, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. Crawling by readability level. In *Proceedings of 12th International Conference on the Computational Processing of Portuguese (PROPOR)*.

Sandra Williams. 2004. *Natural Language Generation (NLG) of discourse relations for different reading levels*. Ph.D. thesis, University of Aberdeen.

Ramon Ziai and Niels Ott. 2005. Web as corpus toolkit: User's and hacker's manual. *Lexical Computing Ltd., Brighton, UK*.