Building a Spanish Readability Classifier

Deekshitha Garrepelly, Kevin Lee, Brandon Luton

Abstract

Teachers and learners of Spanish face difficulty selecting texts appropriate for the level of the student. It would be beneficial to have a corpus where teachers and learners can easily find texts suitable for the language learning goals of themselves or their students. Here we investigate various features of a given Spanish text and the impact they have on text readability. We found that of the features analyzed, the average number of words per sentence, the mean number of letters per word, stem overlap, and the content word overlap across the entire text were the most influential in predicting text readability.

1 Introduction

Readability classification aims to classify texts based on their complexity. Traditionally, readability formulas such as those by Flesch and Dale and Chall have been used to mathematically capture the complexity of the text. However, these formulas only account for a few simple linguistic features. Natural Language Processing (NLP) tools can be used to consider more complex linguistic features and provide much faster calculations than before.

Specifically for language learners, accessing texts that are appropriate for their skill level can be difficult. Having a tool to give an objective view of the complexity of a text can be useful for language learners and teachers to select appropriately difficult texts for themselves or their students and enhance their language learning capabilities.

Here we investigate what features contribute to text readability as it pertains to the Spanish language by creating various classification models based on different feature inputs. Our hope is that this work can be used to further develop a model to help classify Spanish texts for Spanish language learners and teachers.

2 Proposed Solution

To align with the goals of the B.E.A.R.D lab to assess the readability of Spanish texts, we propose a classification model which classifies a given document as an advanced/beginner text based on certain features extracted from the document. The features that we extract from the documents are a combination of descriptive features such as the number of words, number of sentences, etc., and comprehensibility features such as lexical richness, connectives incidence, the incidence of various parts of speech tags, and polysemous word incidence. The model could be trained using Stochastic gradient descent to learn the weights of various features while minimizing the cross-entropy loss. We use a max entropy model to predict the output and which outputs are the most probable output among all the possible outputs. This model's approach is very close to Coh-Metrix-Esp's approach in taking both descriptive and comprehensibility features to assess readability.

3 Related Work

Early readability formulas used simple measures to predict text readability. One of the most popular formulas is the Flesch-Kincaid which measures the readability of a text based on the number of syllables per word and the number of words per sentence. It assigns a score to a text based on these measures, with higher scores indicating that texts are easier to read.

Most models and formulas, including the Flesch-Kincaid formula, focus on readability as it relates to Native Learners and in particular the English language. However, likely due to the lack of large, annotated datasets geared towards second language (L2) learners, research concerning readability with a focus on L2 learners, and non-English languages, is relatively recent.

One paper by Quispesaravia et al details the creation of a Spanish complexity analysis tool called Coh-Metrix-Esp. This method took into consideration various Coh-Metrix indices, based on the original English version (Graesser et al.), to determine quantitatively the readability of the document. These indices include – Descriptives, Referential Cohesion, Lexical Diversity, Connectives, Syntactic Complexity, Syntactic pattern density, Word Information, and Readability. The model is trained by presenting the features of individual documents as the indices described before with two classes of documents – simple and complex.

Many traditional readability classification models make use of surface features such as average sentence length and average word length in characters or syllables. Other models create lists of "difficult" words based on the frequency counts. (Vajalla and Meurers, 2012). Vajalla builds upon this by grouping features into lexical, syntactic, and traditional categories in order to classify the difficulty of a text.

4 Methodology

4.1 The Data

There are limited resources freely available for readability classification. The majority of the data comes from web scraping articles from kwiziq and Hablacultura. Additionally, a few articles came from lingua.com which offers some free sample articles. These articles tend to be short and target Spanish learners trying to improve their reading comprehension as opposed to native Spanish speakers.

The articles cover a wide range of topics including entertainment, culture, literature, and science. Some articles were fiction. Additionally, while most articles were written in a basic paragraph format, a few articles were lists, songs, or dialogues, which have a far different format. Most articles remained unchanged with the only changes being the deletion of some subheadings.

The articles cover a wide range of topics including entertainment, culture, literature, and science. Some articles were fiction. Additionally, while most articles were written in a basic paragraph format, a few articles were lists, songs, or dialogues, which have a far different format. Most articles remained unchanged with the only changes being the deletion of some subheadings.

All of the articles from these sites contain information regarding the readability of the text based on the Common European Framework of Reference for Languages (CEFR). CEFR divides learners into classes based on their proficiency in the language. A1 and A2 represent a basic understanding of a language. B1 and B2 are the intermediate levels and C1 and C2 are the advanced levels.

The texts are processed and annotated with the open-source natural language processing library spaCy (Honnibal & Montani, 2017). The texts are annotated with the following labels: (ID, Form, Lemma, UPOS, XPOS, Head, and DEPREL).

In total there were about 400 texts in the corpus. 75% of the texts were split into the training data (300 texts) and the rest served as the testing data. To simplify the problem, texts were divided into 2 classes: basic (A1, A2, B1) and advanced (B2, C1, C2).

4.2 The Features

Our experimental methodology is to extract different combinations of the following features from individual documents to train the model and determine which features hold the most weight in classification. We combine the various features specified in our proposed solution.

The descriptive features are based on the number of words, number of sentences, and the average number of words per sentence in the document. Easily readable documents may have less words, fewer sentences, and less words per sentence in comparison to advanced documents.

We measured referential cohesion based on the local and global overlap of content words between sentences, as well as the overlap of the lemma in adjacent sentences. Easily readable documents typically have a higher overlap of these entities compared to difficult documents.

Lexical diversity is broken into the type-token ratio (TTR) of content words and the TTR of all words in the document. The TTR is measured by the (number of unique word types) / (number of tokens for these word types). The lexical diversity is inferred to be lower for an easier document in comparison to a difficult document. Thus, lower TTR may be associated with easier documents.

Connective incidence is measured by the number of connectives per n words. We created a list of common Spanish connectives to measure the number of connectives. As connectives increase the coherence of a document, easily readable documents may have higher connective incidence than difficult documents.

Incidence scores for various part-of-speech tags are measured using the number of words with specific POS tags per length n. Easily readable documents are less diverse in terms of their tokens and vocabulary, which implies a lower incidence score for beginner texts.

As for polysemous incidence, we measure the proportion of ambiguous words per certain length in the document. This score tends to be lower for easily comprehensible documents than for more difficult ones.

4.3 Experiment 1

The first experiment analyzed basic descriptive features of the text such as the number of words, the mean number of words per sentence, and the type-token ratio of content words. For the type-token ratio, a unique word is defined as any word whose part of speech is either a noun, verb, adverb, or adjective.

We observed that the model associates higher values of these features with higher text difficulty. This aligns with basic intuition as easy texts tend to use shorter sentences, have overall less content (number of words), and tend to repeat words and phrases resulting in less lexical diversity (lower type-token ratio). Notably, the type-token ratio had the least effect on the classification with the other two features having approximately equal weight in the model's predictions.

4.4 Experiment 2

The second experiment analyzed three new features including the incidence of connectives, the overall type-token ratio, and the incidence of nouns. The incidence of connectives is a measure of the number of connective words/phrases appearing within 30 words. The overall type-token ratio does not consider only content words in its calculations. The incidence of nouns is a measure of the number of appearances of a noun within 30 words. This feature proved to have the most influence on the model's predictions

In this experiment, the model provided some unexpected results. Firstly, the model associates beginner texts with a lower connective incidence than advanced texts. The reason could be that the list of connectives that are used in featurization is not exhaustive. Additionally, many of the beginner texts are quite short and may not necessarily use connectives.

Secondly, the model associates beginner texts with more lexical diversity than the advanced texts. A potential cause of this is that the beginner texts may have greater lexical diversity to help introduce new vocabulary to its audience, but this new vocabulary may be very simple.

4.5 Experiment 3

Three new features were analyzed for the third experiment. These were the number of sentences, content word overlap, and polysemous incidence. Content word overlap measures the overlap of content words between adjacent sentences. Polysemous incidence is a measurement of the frequency of polysemous words, or words that have multiple meanings or senses, in a text.

In this experiment, we observed the model to associate low weights for the number of sentences, with high weights for content word overlap and low weights for polysemous incidence for beginner texts compared to advanced texts. These observations indicate that the number of sentences and the polysemous index are less of an indicator as to whether a text is beginner or advanced compared to content word overlap.

4.6 Experiment 4

Experiment 4 combined all the features from the previous experiments into a single model. For this experiment, it was observed that the average number of words per sentence, the mean number of letters per word, stem overlap, and the content word overlap across the entire text were the largest factors in determining text readability. Beginner texts tend to have shorter sentences, more content word overlap, and higher stem overlap measure than advanced texts. Surprisingly, beginner texts had longer words (higher mean number of letters per word). One possible explanation for this is that beginner texts may have more proper nouns which tend to be longer in length.

4.7 Results

SNo	Exp - 1	Exp - 2	Exp - 3	Exp - 4
Accuracy	0.59	0.61	0.47	0.63
Precision (Macro)	0.58	0.60	0.50	0.64

Precision (Micro)	0.59	0.61	0.47	0.63
Recall (Macro)	0.58	0.60	0.50	0.64
Recall (Micro)	0.59	0.61	0.47	0.63
F1 (Macro)	0.58	0.60	0.47	0.63
F1 (Micro)	0.59	0.61	0.47	0.63

Table 1- Metrics for various experiments on Test split (Exp refers experiment)

Table 1 shows all the statistics for the four different models used in this investigation. Model four had the best performance, which is expected since it used all the features of the previous models. The third model had the worst performance. Since content overlap and the number of sentences were poor indicators of text readability, model three only had one major feature, polysemous incidence, to base the text classifications off.

5 Limitations of Work

The first major limitation lies in the dataset itself. The dataset lacked enough articles for each of the CEFR levels, so the articles had to be combined into two categories: basic and advanced. The articles' CEFR levels were determined by different humans so there could be errors in the classification of the articles themselves. CEFR levels also do not have strict boundaries. For example, multiple articles suggested a range of levels such as A2/B1, showcasing that the distinction between CEFR levels is not very distinct at all. The articles were primarily pulled from two sources whose target audience is Spanish-language learners. This lack in variety of styles and authors could also limit the performance of the model when given articles from other sources.

Two features we wanted to include but could not, due to a lack of resources, were latent semantic analysis and an analysis of the number of syllables per word. Latent semantic analysis is important to readability as it analyzes how concepts and ideas flow from one sentence to the next. The cohesion of ideas from one sentence to the next helps a reader follow the ideas of the text and improves the readability of the text. Syllables can also be a good predictor of text readability since harder words are typically longer and have more syllables. Syllable analysis has also been widely used in classic readability formulas. Both of these features could help predict text readability.

6 Future Work

Work should be done to attempt to improve the overall performance of the model presented here. One way is to improve the corpus itself by adding more texts, more variety of texts, and better-quality texts. The corpus currently contains 402 texts but has no C2-level texts. Future iterations of the corpus can include higher-level texts such that the model has more diverse texts to classify from.

A potential method of increasing the number of C2-level texts if human-tagged texts are not readily available is to take excerpts from advanced literary texts since these texts will almost certainly be advanced C2-level texts. A limitation to this method, as noted by Dr. Beard, is that the inclusion of difficult texts from literary works may skew the readability levels of intermediate texts to appear easier.

One distinction in higher CEFR levels is a higher complexity due to figurative interpretation. The language in a C1/C2 text may not be more advanced than a B1/B2 text in terms of lexical scope. However, figurative interpretation of the text requires a better understanding of the language and the connection between vocabulary and intangible ideas. Figurative analysis was not a part of our study but is an area that should be explored in future work.

A shift away from classifying based on CEFR levels towards a different scale may also improve performance. One potential idea is to classify texts based on the U.S. grade levels. This could allow for a linear regression model and texts at a higher level (beyond C2) could be classified as a higher grade level (say 12th grade).

7 Conclusion

This project provides some analysis on the readability level of Spanish texts based on a combination of surface-level features of the text such as the number of words, number of sentences etc., and the comprehensibility features such as the incidence of connectives, the incidence of various parts of speech tags, and the lexical diversity of the document's content. It was determined that the most influential features when predicting text readability were the average number of words per sentence, the mean number of letters per word, stem overlap, and the content word overlap across the entire text. More work is needed to verify these results and explore other features that potentially have a large impact on text readability

8 Thanks

Special thanks to Dr. David Beard who provided guidance throughout the development of this project.

9 See Also

Refer to the link below to see all the resources used for this project including the code, this paper, and the data used.

https://drive.google.com/drive/u/1/folders/1se8B6-RgsnfXy9FutI_cjhLtIMRw7yAu

References

Andre Quispesaravia, Walter Perez, Marco Sobrevilla Cabezudo, and Fernando Alva-Manchego. 2016. Coh-Metrix-Esp: A Complexity Analysis Tool for Documents Written in Spanish. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 4694–4698, Portorož, Slovenia. European Language Resources Association (ELRA).

Bengoetxea, Kepa, and Itziar Gonzalez-Dios. "MultiAzterTest: a Multilingual Analyzer on Multiple Levels of Language for Readability Assessment." *arXiv preprint arXiv:2109.04870* (2021).

Crossley, Scott A., Jerry Greenfield, and Danielle S. McNamara. "Assessing text readability using cognitively based indices." *Tesol Quarterly* 42.3 (2008): 475-493.

Dale, Edgar, and Jeanne S. Chall. "A formula for predicting readability: Instructions." Educational research bulletin (1948): 37-54.

Flesch, Rudolph. "A New Readability Yardstick." *Journal of Applied Psychology*, vol. 32, no. 3, June 1948, pp. 221–33. EBSCOhost, https://doi.org/10.1037/h0057532.

Graesser, Arthur C., et al. "Coh-Metrix: Analysis of text on cohesion and language." *Behavior research methods, instruments, & computers* 36.2 (2004): 193-202.

Heilman, Michael, et al. "Combining lexical and grammatical features to improve readability measures for first and second language texts." *Human language technologies 2007: the conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*. 2007.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 460–467, Rochester, New York. Association for Computational Linguistics.

Quispesaravia, Andre, et al. "Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish." *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016.

Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H. Sanchez Gutierrez, and Kenji Sagae. 2020. Developing NLP Tools with a New Corpus of Learner Spanish. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 7238–7243, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Vajjala, Sowmya, and Detmar Meurers. "On improving the accuracy of readability classification using insights from second language acquisition." *Proceedings of the seventh workshop on building educational applications using NLP*. 2012.

Vásquez-Rodríguez, Laura, et al. "Texts Readability Analysis for Spanish." *W&B*, 29 Mar. 2022, https://wandb.ai/readability-es/readability-es/reports/Texts-Readability-Analysis-for-Spanish--VmlldzoxNzU2MDUx.