TJHSST Computer Systems Lab Senior Research Project
Automatic Readability Evaluation Using a Neural Network
2009-2010

Vivaek Shivakumar

October 30, 2009

Abstract

Many formulas and methods for assessing the readability of a text or determining the appropriate grade level are inaccurate and based only on surface features of text. In automatic assessment of a text's reading level, computers can easily run more sophisticated models than simple algebraic formulas; the goal of this project is to create such a model. Indexes and statistics will be computed with respect to various features of a text such as sentence length and lexical density. A combination of these textual features is necessary to accurately capture the readability of a text. A neural network will be used to implement a model for readability using said features as inputs. After being trained the model will be useful for determining approximately what U.S. grade level corresponds to a given text, for use in educational or other settings to assess writing for a certain audience.

Keywords: readability, reading level, understandability, text classification, neural network, semantic, syntactic

1 Introduction

Readability classification is a valuable tool in educational, academic, and business situations dealing with writing for a certain audience. In order to check whether a text is written at the appropriate level for the target audience or whether an already-written text is grade level-appropriate, several factors are taken into account. Formulas exist to estimate the reading level of a text. Due to the rapid growth of information online and of the number of electronic texts, it is becoming more and more important to be able to automatically perform tasks such as assigning a reading level to a text; and such formulas, although they are widely used in computer applications, are not optimal for computer use. These formulas frequently use only primitive criteria to determine a readability score for a text. A more accurate model calls for both more sophisticated criteria and a more suitable way to use those criteria to develop a model for readability.

The purpose of this project is to implement such a model, going further than traditional readability formulas by using textual features that are not just syntactic or orthgographic in nature, and by using a flexible and more accurate model by way of machine learning. Several different textual features will be used in the model, ranging from simple ones such as word lengths to the more sophisticated such as the number of parsed dependencies of a certain type within a sentence. A neural network will be used to implement machine learning, using the text criteria as inputs to result in a readability score.

To train such a neural network, or to work with any such sort of automatic tool performing a somewhat subjective real-life task, a multitude of data is necessary in the form of text corpora and preassigned grade level values. However, the amount of such resources will be limited so the implementation of this readability classifier is merely a beginning to be improved upon. The project will investigate the relative effectiveness of various features of prose in determining how difficult the text is to read. The complexity and difficulty of analysis of certain textual features will limit the scope of this project in that respect as well. However, the main objective is to develop an efficient model of text readability using more sophisticated, modern computational techinques that is more accurate than widely used traditional formulas.

2 Background

2.1 Readability Formulas

Three widely-used readability formulas that are compared and evaluated in the preliminary portion of this project are the Flesch-Kincaid Grade Level (used, e.g., by Microsoft Word [1]), the Dale-Chall Index, and the SMOG (Simple Measure of Gobbledygook) index. All three are based on surface linguistic features, although the Dale-Chall index utilizes a simple index of the semantic (word meaning) difficulty of a text [2]. The formulas are as follows:

$$Flesch_Kincaid = 0.39 * \frac{Words}{Sentences} + 11.8 * \frac{Syllables}{Words} - 15.59 \tag{1}$$

This is for the Flesch-Kincaid Grade Level, derived from the Flesch-Kincaid Readability Index [3].

$$Dale_Chall = 0.1579 * PDW + 0.0496 * ASL + 3.6365 \tag{2}$$

PDW is percentage of difficult words, i.e., ones that are not in a certain list of 3000 common English words. ASL is the average sentence length, in words. [2]

$$SMOG = 1.043 * \sqrt{30 * \frac{Complex_Words}{Sentences}} + 3.1291 \tag{3}$$

A complex word is one with three or more syllables [4].

2.2 Textual Features

To fully evaluate the readability of a text, several different types of features must be taken into account: orthographic, phonological, semantic, syntactic, and if permissible by the scope of the project (not so in this case), pragmatic. These features have to do with letters, sounds, word meanings, sentence structure, and contextual meaning, respectively. The basis of most readability formulas such as Flesch-Kincaid Grade is on orthographic, phonological and syntactic features, at the most primitive level. However, the difficulty of reading a text obviously has somewhat to do with the meanings of the words, e.g., how specialized the text is or how many obscure words are present, and also with other syntactic features such as the type or complexity of a sentence. For example, two sentences may be the same length but one may contain several subordinate clauses while the other does not. Such factors must be considered in a readability model. For example, one study concluded that the often-used surface linguistic feature of word length by syllables was not a good indicator of grade level of text on elementary and middle school science websites [5]. Another study [6] mentions several possible textual features to be analyzed and in particular parse tree height, indicating that parsing sentences is important to automatic evaluation of text readability; sentence parsing requires a computer, indicating that it is already a more sophisticated criteria than those used by traditional readability formulas. Therefore there are many possible criteria that determine the difficulty of reading and comprehending a text, and an accurate model of readability would ideally factor in all or almost all of these criteria.

2.3 Machine Learning

Almost all other projects that deal with readability analysis involve machine learning of some sort [6][7]. A machine learning method would "learn" to output an appropriate reading level score for a text based on its features as described above, after being trained on a training set of data. One implementation of machine learning is a neural network. A neural network uses inputs (in this case scores or indices based on the text features and criteria) and manipulates them in a model using a web of connections and weights to output one or more values, which would in this case be the readability score. Neural networks can either be supervised or unsupervised. A supervised network is preliminarily trained by a training set, e.g., a corpus of text with predetermined grade level values, by attempting to modify the model after each trial. On the other hand, an unsupervised network learns on its own as it goes along receiving input, by identifying patterns. The nature of a neural network, including the number of layers of feeding forward and number of nodes, depends on the nature and relationships between the various types of input that will be used.

3 Methodology

The programs for this project will be written in Python and Perl - the former for the neural network and training program, and the latter for other programs such as obtaining counts.

3.1 Neural Network and Corpus Texts

To implement the idea of machine learning to classify a text for readability, a neural network will be used with the text features as inputs. The neural network will be supervised, meaning that it will be trained with a set of texts of “known grade levels”. The texts in this training set will consist of mostly reading passages of national or state-administered standardized tests available online. Effort has been taken to collect passages at each U.S. grade level and from a range of states as to not induce bias as to what the typical standard is for each grade.

3.2 Textual Features

A simple program has been written to obtain syllable, word, and sentence counts of a text. The characters that are identified as end-of-sentence marks are the same as those used by Talburt (1985) [3]. These are “. ? ! : ;”. Since syllable demarcation can be very irregular in the English language due to a slightly unphonetic orthography, the method for determining syllables is similar to what was used by Talburt (1985) [3] in that implementation of the Flesch-Kincaid index. The method is as follows: Each group of consecutive vowels (a,e,i,o,u) counts towards a syllable, with the following exceptions:

1. Final -ES, -ED, and -E are not counted as syllables (besides -LE, which is).

2. The letter “y” is a vowel unless it starts a word or follows another vowel.

3. Any word of three letters or less counts as one syllable.

For semantic features, word lists of the English language are readily available online, and so are frequency lists of English. These will be used to compute the ratio of frequency of a word in the given text to the frequency in English in general, to compute the “lexical density.” Another method is simply to calculate the percentage of words in the text that do not appear in a given list of common English words. A third method, especially to indicate text specialization, although it is limited, is the use of frequency word lists for an array of specific subjects, computing the percentage of words that appear in any of those lists to indicate a specialized text.

Syntactic features can be analyzed through parsing programs. Since the goal of this project has nothing to do with parsing, it will suffice to use an existing parser rather than create one from scratch. Currently the parser being used is the Stanford Parser which can produce both parse trees (showing a tree diagram of a sentence) and dependency trees (a list of word pairs of various dependency types, e.g., verb and direct object). For each sentence's output, either or both trees can be analyzed for anything from the height of the tree to the occurrence of a specific type of clause, etc.

3.3 Preliminary Evaluation of Existing Readability Formulas and Parsing Criteria

Before the main phase of the project, data was collected on the texts of the inital corpus to analyze the effectiveness of the three readability formulas mentioned above in determining the U.S. grade level for a text. The scores for each formula, along with the average dependency tree height and average parse tree height from the Stanford Parser output were obtained for 92 texts at various grade levels predetermined by those who designed the various standardized tests or practice tests from which the samples were obtained.

The plots of all five values vs. actual grade level can be found in Appendix A. The values given by the three readability formulas in the preliminary evaluation show a positive linear association with the actual predetermined sample text grade levels. However, there is much variability and some systematic inaccuracy, so none of the formulas are reliable as accurate indicators of text grade level. Therefore it can be concluded that while surface linguistic features such as average sentence length do help in determining readability, they cannot be used alone.

Dependency and parse tree sizes show a slight positive linear association with reading grade level, indicating that they are indeed factors in readability. However, due to large variability in sentence sizes within a text, the average parse tree sizes for a text may be similar accross a wide span of reading levels. In cases where a difficult text happens to have short or relatively simple-structured sentences, the distinguishing factors are semantic in nature, again indicating that the multitude of readability factors must be used collectively rather than separately. A plot of average dependency tree size vs. average parse tree size show that the two factors are highly correlated, indicating that only one of the two need be included as a criterion for readability. The decision between the two will probably be made based on speed.

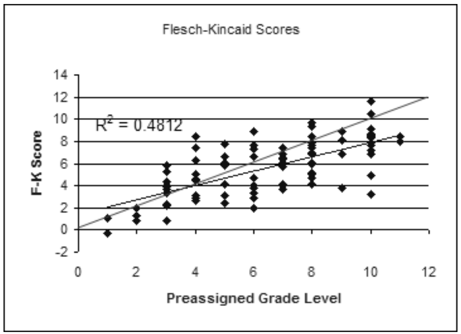
4 Results and Analysis

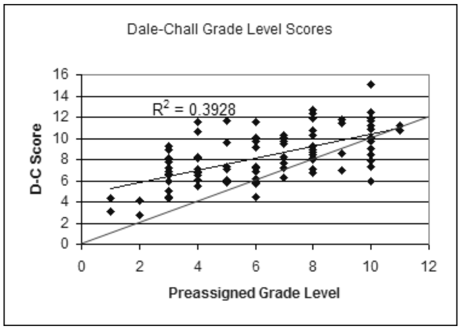
4.1 Expected Results

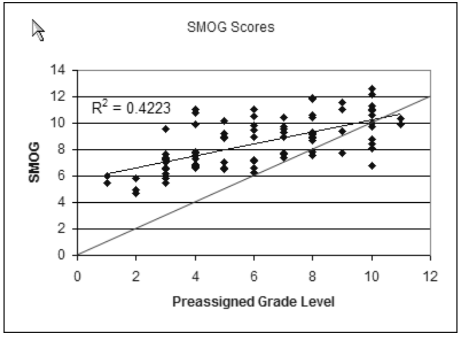
The ideal outcome of this project is that the neural network model successfully learns, by training, the patterns based on the different text features analyzed that give certain texts certain reading level scores. Due to the visible difference between the texts of standardized test reading passages for different grade levels, the model would be expected to somewhat accurately assign a grade level score to a given text. Although no automatic readability evaluator can be perfect, this model should turn out to be more accurate than the algebraic formulas like SMOG in estimating readability. When the program has been completed, it will have valuable application in an educational, or even personal setting to check the level of a text to that of a desired audience or to compare the readability of different texts.

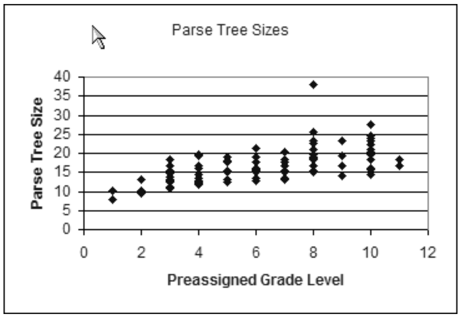
Appendix A. Charts of Preliminary Evaluation

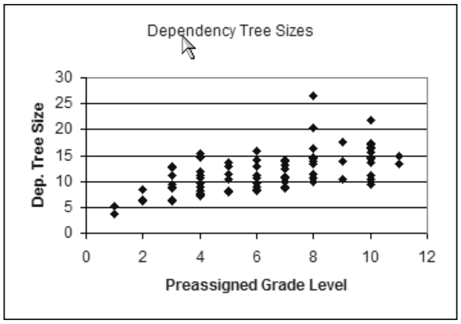
The charts of the formula results contain both a regression line and the line that marks perfect grade level prediction.

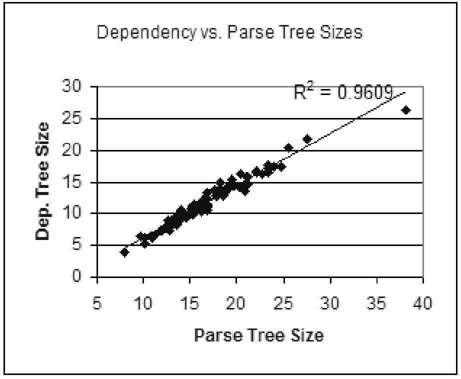












References

[1] "Test your document's readability." *Microsoft Office Online*. Microsoft, n.d. Web. 30 Oct. 2009.

[2] "The New Dale-Chall Readability Formula." *ReadabilityFormulas.com.* My Byline Media, n.d. Web. 30 Oct. 2009.

[3] Talburt, John. "The Flesch index: An easily programmable readability analysis algorithm." In: *SIGDOC 85: Proceedings of the 4th annual international conference on Systems documentation, ACM Press* (1985): 114-122.

[4] McLaughlin, G. Harry. "SMOG grading a new readability formula." *Journal of Reading* 22 (1962): 639-646.

[5] Si, Luo and Jamie Callan. "A Statistical Model for Scientific Readability." In: *Proceedings of the tenth international conference on Information and knowledge management* (2001): 574-576.

[6] Feng, Lijun. "Automatic Readability Assessment for People with Intellectual Disabilities." *ACM SIGACCESS Accessibility and Computing* 93 (2009): 84-91.

[7] vor der Brück, Tim, Sven Hartrumpf and Hermann Helbig. "A Readability Checker with Supervised Learning Using Deep Indicators." *Informatica* 32 (2008): 429-435.

[8] Kane, Lorna, Joe Carthy and John Dunnion. "Readability Applied to Information Retrieval." *Advances in Information Retrieval. Lecture Notes in Computer Science* 3936 (2006): 523-526.

[9] Miltsakaki, E. and Audrey Troutt. "Read-X: Automatic Evaluation of Reading Difficulty of Web Text." In: *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (2007): 7280-7286.