Automatic Readability Evaluation Using a Neural Network

Vivaek Shivakumar
Computer Systems Lab 2009-2010

Abstract

Measures of text readability using simple algebraic formulas are much outdated but still widely used, including for classifying texts by U.S. grade level for reading. However, a good measure of the grade level of a text must take into account a variety of features of text including semantic and syntactic ones to form a good readability model. This project attempts to create a working, learning model by use of a neural network to assign a reading level to any text by collecting various input factors more than just the primitive ones encountered in traditional formulas for readability. The product will be useful in both educational and personal settings in telling whether a certain text is written at the appropriate level for an intended audience.

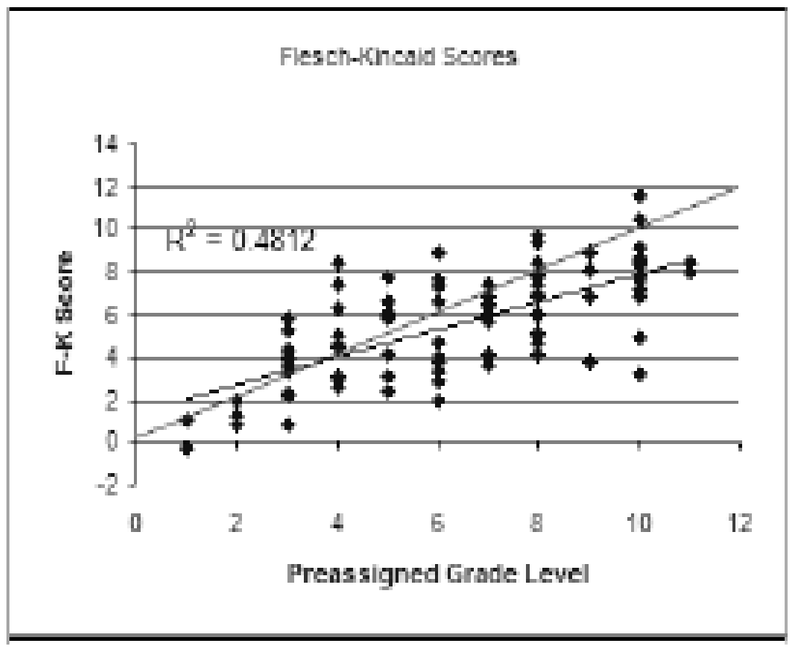

Figure 1: Evaluation of the Flesch-Kincaid Grade Level

Background

Three widely-used readability formulas that are compared and evaluated in the preliminary portion of this project are the Flesch-Kincaid Grade Level, the Dale-Chall Index, and the SMOG index. All three are based on primitive surface linguistic features, e.g., here is the formula for the Flesch-Kincaid Grade Level:

$$\text{Flesch-Kincaid} = 0.39 * \frac{\text{Words}}{\text{Sentences}} + 11.8 * \frac{\text{Syllables}}{\text{Words}} - 15.59$$

This is derived from the Flesch-Kincaid Readability Index [3].

Textual Features

To fully evaluate the readability of a text, several different types of features must be taken into account, dealing with such aspects as word lengths, word meanings, and sentence structure. Most readability formulas do not consider semantics as criteria for readability. However, the difficulty of reading a text obviously is related with the meanings of the words, e.g., how specialized the text is or how many obscure words it contains, and also with other syntactic features such as the complexity of a sentence's constructions. One 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.

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 to output one or more values, which would in this case be the readability score.

Procedure

At this stage, I have done a preliminary evaluation of existing readability formulas. Implementation of the simple algebraic formulas is trivial. The sample texts used were taken from various state grade level standards assessments as to be representative of the “actual” grade level of a text.

The same texts were also analyzed for two syntactic criteria, the average parse tree size and average dependency tree size. The parser used was the Stanford parser.

Current Results and Conclusions

The results of the preliminary analysis show that all three readability formulas are positively associated with the actual grade level, but there is considerable variability. There is enough inconsistency to conclude that none of the formulas are reliable in assessing the reading level of text.
However, the criteria used by the formulas, such as average sentence length, do give an indication of readability, but are not to be used alone.

The results for the parse and dependency tree sizes show that both criteria are somewhat associated with readability, although less so than the surface features like word length. However, parse tree size and dependency tree size are highly correlated (r=.96), so it can be concluded that using just one as a criteria in a readability model will suffice.