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

Vivaek Shivakumar

October 29, 2009

Abstract

Measures of text readability using simple formulas are much outdated, yet still widely used including for classifying texts by U.S. Grade level for reading. A good measures of the grade level of a text must take into account primitive as well as semantic and syntactic features of text to form a model. This project attempts to create a working model to assign a reading level to text using machine learning with various input factors more than just the primitive ones encountered in traditional formulas for readability. The product will be useful in telling whether a certain text is written at the appropriate level for an intended audience, especially in an academic setting from elementary to high school.

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

1 Purpose and Scope

The majority of readability tests in use today are in reality simple formulas based on counts and averages of letters, syllables, words, and/or sentences,

with each having its variations and modifications. Although the primary purpose of these simple algebraic formulas is to be able to calculate the readability of a text by hand (say, by taking a sample of sentences from an article) they are being implemented in word processing programs and in online applications where there is room for much more sophisticated methods of readability evaluation.

Although countless factors affect how difficult a certain piece of text is to read or comprehend, certain characteristics and statistics can be analyzed by a computer for use in a model of text reading level. These characteristics range from the simplest, e.g., the number of characters per discrete word or token, to the most complicated to handle in a computer program, language pragmatics. The scope of this project will allow a limited number of such factors in computing a model, and features used will span most of the said range, from individual letters to sentence syntax and word meanings, i.e., semantics.

To create a model a neural network using these inputs from a text will be implemented, at first with supervised learning and possibly unsupervised learning afterwards if the scope of this project allows. The obvious question is how to train the model, specifically, from where may one obtain texts labeled with the "correct" reading level. The answer to this is that the most reliable sources are objective measures by educational (and psychological and linguistic) experts who work in the making of standardized testing for reading. Therefore, reading passages that are used as benchmarks in tests for various U.S. grade levels will provide the source of training.

2 Background and Research

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 focus of most readability formulas such as Flesch-Kincaid are on orthographic, phonological and most of all syntactic (mainly length of sentences). 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.

Feng (2009) 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. Almost all other projects that deal with this topic involve machine learning of some sort (Feng, 2009; vor der Brck, Hartrumpf and Hermann Helbig, 2008). A machine learning method would "learn" to output an appropriate reading level score for a text based its features as described above, after being trained on a training set.

3 Methodology

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", most probably from reading passages of national and state standardized tests available online. Counts and indices based on the features to be used as inputs in the neural network must somehow be obtained. A simple program can be written to count words, sentences, etc. For more semantic features, word lists of the English language are readily available online. 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.

The programs for this project will be written in Python and Perl. Since the model is a neural network, any "testing" of it is just further machine learning.

4 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 different texts.