How Hard Can it Be? The E-Score - A Scoring Metric to Assess the Complexity of Text

Sandeep Mathias, Pushpak Bhattacharyya
Department of Computer Science and Engineering
IIT Bombay, India
{sam,pb}@cse.iitb.ac.in

Abstract

In this paper, we present an evaluation metric, the E-Score, to calculate the complexity of text, that utilizes structural complexity of sentences and language modelling of simple and normal English to come up with a score that tells us how simple / complex the document is. We gather gold standard human data by having human participants take a comprehension test, in which they read articles from the English and Simple English Wikipedias. We use this data to evaluate our metric against a pair of popular existing metrics - the Flesch Reading Ease Score, and the Lexile Framework.

Keywords: text complexity, structural complexity, lexical complexity

1. Introduction

Today, there are many readability formulae that are used for evaluating the readability / complexity of text. Some of them, like the Flesch Reading Ease (Flesch, 1948) score (FRES) are based on surface values, like average words per sentence and average syllables per word. Others, like the Lexile Framework (Stenner, 1996) make use of the fact that rarer words are more complex than words that occur more commonly in a general corpus. The C-Score (Temnikova and Maneva, 2013) is yet another means of evaluating the difficulty of a text. However, unlike the Lexile Framework and Flesch Reading Ease, the C-Score is calculated using human readers produce the data necessary to calculate it. Yet, because the data used to calculated C-Score is gotten manually, it is one of the best metrics for getting gold standard data about the complexity of text. Because of this, we use the C-Score as the gold standard for estimating the complexity of the texts used in our experiment.

The Flesch Reading Ease is one of the earliest readability tests. In the Flesch Reading Ease score, higher valued texts are said to be simpler to read. This readability formula takes into account only the average number of syllables per word, and the average number of words in a sentence of the document.

The Lexile Framework (Stenner, 1996) makes use of the frequency of words in a training corpus, as well as the number of words in a sentence. It takes into account the mean of the log of the frequency of the word in a corpus, as well as the log of the mean sentence length to calculate the score. Currently, it is being used in the United States to provide reading suggestions to schoolchildren, as well as assess their reading ability as part of the Common Core Standards for English[1]. Using the training corpus, each word in a test passage is assigned a particular score - the log of their frequency in the training corpus. A value, the theoretical logit for a passage, is calculated using the mean log frequencies of the words in the passage, as well as the log of the mean sentence length.

Despite the fact that Lexile is a data-driven formula, it still suffers from criticism. Certain books, like *The Library Mouse* by Daniel Kirk have an abnormally high Lexile rating[2] despite being a children's book, as compared to a young adult book, *Twilight* by Stephanie Meyer[3].

More recently, (Schwarm and Ostendorf, 2005) demonstrated a means of classifying texts based on their complexity into appropriate grade levels. (Schwarm and Ostendorf, 2005) made use of support vector machines and language models and showed that it performed significantly better than FRES and Lexile when it came to assigning a grade-level for a document. While our work also makes use of language models, it differs from (Schwarm and Ostendorf, 2005) as it gives a raw score to the difficulty of the document, rather than the grade-level it is meant for.

2. The E-Score - Our Complexity Metric

To calculate the E-Score, we make use of two types of complexity, namely:

1. Structural complexity; and
2. Lexical complexity

2.1. Structural Complexity

Structural complexity is a measure of how complex the sentence is, based on its parse tree. There are many measures of defining structural complexity. We define structural complexity as follows for calculating the E-Score. For a given sentence S, we define the structural complexity S_c, as the number of factual statements extracted from Michael Heilman's factual statement extractor[4] (Heilman and Smith, 2010). A factual statement is a simple sentence that contains a single fact. For example, in the sentence

[1] http://www.corestandards.org/wp-content/uploads/Appendix-A-New-Research-on-Text-Complexity.pdf

[2] https://lexile.com/book/details/9780810993464/

[3] https://lexile.com/book/details/9780316015844/

[4] The system can be downloaded from www.cs.cmu.edu/˜ark/mheilman/qg-2010-workshop

"Bernie Sanders, the Senator from Vermont, is campaigning against Hillary Clinton, the wife of former President Bill Clinton, to become the President of the United States."

gives rise to the following factual statements:

- Bernie Sanders is the Senator from Vermont. (Appositive)
- Hillary Clinton is the wife of former President Bill Clinton. (Appositive)
- Bernie Sanders is campaigning against Hillary Clinton to become the President of the United States. (Main Clause)

The different types of simplified factual statements we extract from an input sentence are:

1. Main clause sentences
2. Factual statements from relative clauses
3. Factual statements from appositives
4. Factual statements from noun and verb participial phrases
5. Factual statements from other subordinate clauses

We use this definition of structural complexity because a sentence that is more complex would have more clauses in it that can be extracted into simpler factual statements.

2.2. Lexical Complexity

Lexical complexity is the complexity of the text based on its vocabulary. It is based on the complexity of the words and phrases used in the text. We use a unigram and bigram language model of a Simple English - English parallel corpus to calculate the lexical complexity of each n-gram. The complexity of an n-gram is comprised of 2 parts, namely the corpus complexity and the syllable count.

1. **Corpus complexity** For each n-gram (g) of the sentence, we calculate its corpus complexity (Biran et al., 2011), $C_c(g)$, defined as the ratio of the log likelihood of g in the English corpus to the log likelihood of g in the Simple English corpus. In other words,

   $$C_c(g) = \frac{LL(g|normal)}{LL(g|simple)}$$

   Here, we assume that every n-gram in the Simple English corpus has to occur at least once in the English corpus. Section 4 contains more details about the corpus used.

2. **Syllable count** We consider that readers read words one syllable at a time. The syllable count, $s(g)$, of an n-gram (g) is defined as the sum of syllables of the words in that n-gram.

With these two ideas, we go ahead and calculate the lexical complexity of an n-gram (g) as:

$$Lc(g) = s(g) \times C_c(g)$$

Hence, for a given sentence S, and an n-gram size, the lexical complexity is given by

$$Lc(S, n) = \sum_g s(g) \times C_c(g),$$

where g is an n-gram of size n.
In addition to this, we also attach a weight W_n to the lexical complexity calculated for a particular n-gram. For a given n-gram size of n, the weight is $\frac{1}{n}$. This is because of the unigrams in the n-gram are added n-times. For example, if n is 2, and we have an n-gram sequence "a b c d e f g ...", unigrams like b, c, d, e, f, etc. get added twice.
Therefore, we can say that the lexical complexity of a sentence is given by

$$L_c(S) = \sum_n W_n \sum_g s(g) \times C_c(g),$$

2.3. Calculating the E-Score

Both the structural complexity and the lexical complexity contribute to the overall complexity of the text. Hence, the formula used to calculate the E-Score is:

$$E = \sum_{s \epsilon S} \frac{S_c(s)+L_c(s)}{|S|}$$

where S is the set of sentences in the text, and S_c and L_c are the structural and lexical complexities respectively.

3. Data

3.1. The C-Score

The C-Score (Temnikova and Maneva, 2013), unlike the earlier readability formulae is calculated using manual data. It is calculated based on participants taking a multiple choice comprehension test. It takes into account factors like number of correct answers that the participants got, the amount of time they took to read the passage, and the amount of time they took to solve the individual questions. Due to the vast differences in size of the individual articles (ranging from 84 words to 939 words), we allowed the participants to take as much time as they needed to read the articles (unlike (Temnikova and Maneva, 2013) which required participants to read them in a limited time), and normalized the C-Score based on reading time.
Like the Flesch Reading Ease Score, the C-Score is also a measure of simplicity. The higher the value, the simpler the text is. The formula for C-Score of a passage is

$$C - Score = \frac{P_r T_s}{T_r} \sum_{q=1}^{N_q} \frac{Q_s(q)}{t_{mean}(q)},$$

where $C - Score$ is the C-Score of the passage, P_r is the percentage of correct answers, T_s is the size of the text, T_r is the mean time taken to read the text, N_q is the number of questions in the text, $Q_s(q)$ is the size of question q, and $t_{mean}(q)$ is the mean time spent in answering question q. The question size is given by

$$Q_s(q) = N_a(q) \times (L_q(q) + L_a(q)),$$

where $N_a(q)$ is the number of options for question q and L_q and L_a are the lengths of the question and answers respectively.

3.2. Getting the Data

We set up a reading comprehension test in which participants had to read a set of 8 passages, alternating between Simple English[5] and English[6] Wikipedia articles. Since a few of the articles in the English Wikipedia were too long, only a small part was provided to the participants for reading. The topics of the passages chosen were generic in nature, such as art, culture, history, film, music, sports, science and world[7]. Table 1 shows the sizes of various passages.

Passage	Simple	Normal
Art	320	939
Culture	235	705
History	196	342
Film	275	538
Music	373	284
Sports	174	381
Science	131	253
World	84	223

Table 1: Lengths of various passages

A total of 30 people took part in the experiment. Their educational qualifications ranged from high school graduates to PhD graduates. 19 of the participants were L2 English learners, while the rest were L1 English learners. 10 of them had won prizes in either the inter-school or intra-college level in literary activities like creative writing, quizzing, word games, scrabble, etc.

Each participant read 8 articles, alternating between Simple English and English Wikipedia articles. After reading each article, they had to answer 5 multiple choice questions (with 4 options each) on that passage. We measured the time taken to read the passages, as well as attempt each question for calculating the C-Score for various passages.

The results of the C-Score test are as shown in Table 2. In most cases, the normal shows a lower score than the simple (in Art, the ratio between simple to normal is more than 2). However, in a few cases, the C-Score of the simple article is lower than that of the normal. Film has the largest desparity, but so also does World. Film has a very high normal value and a lower simple value because of the fact that many respondents claimed to have knowledge of films, as compared to other fields (the number was nearly as much as Science). The Sports simple passage had a very long sentence at the end of it, that while structurally simple, had over 50 words. The World passage also showed the simple being harder than the normal. One of the main reasons is the fact that the size of the World "simple" passage was by far, the shortest passage.

Passage	Simple Score	Normal Score
Art	45.81	22.68
Culture	43.09	49.11
History	59.13	38.13
Film	52	110.92
Music	55.18	37.07
Sports	38.02	68.26
Science	49.73	46.72
World	47.41	79.95

Table 2: C-Score values of different passages

4. Experimental Setup

In the previous section, we described how to get the data against which we will be comparing our metric, as well as the FRES and Lexile scores. We make use of the English Wikipedia - Simple English Wikipedia (Kauchak, 2013) parallel corpus for calculating the corpus complexity of the n-grams. Since the corpus provides a sentence-aligned and a document-aligned corpus, we make use of the document-aligned corpus only for calculating the corpus complexity. The Simple English Wikipedia has around 60,000 articles, each with a corresponding English Wikipedia entry. The document-aligned corpus has all these Simple English Wikipedia articles as well as all their corresponding articles in the English Wikipedia. For each of the 16 articles (8 Simple English Wikipedia and 8 English Wikipedia articles), for which we calculated the C-Score, we calculate the E-Score, using:

1. Michael Heilman's factual statement extractor (Heilman and Smith, 2010)
2. The unigram and bigram lexical complexities from the English Wikipedia - Simple English Wikipedia parallel corpus
3. MorphAdorner[8] to count the syllables in each unigram and bigram

We also calculate the FRES and Lexile scores for each of the articles.

5. Results and Analysis

Table 3 shows the comparison of our metric, the E-score, with other metrics, such as Flesch Reading Ease Score and the Lexile Framework. The values in the table are to show how much more complex the English Wikipedia article is, with respect to the Simple English Wikipedia article.

We use the ratios, rather than the individual text values, because each of the different metrics give different ranges and directions for their scores. Flesch Readability Ease Score (Flesch, 1948) has a range between 0 and 120 (although individual sentences can have a negative value) and has simpler text getting a higher score. Lexile (Stenner, 1996) has a range between 0 and over 2000 and has more complex texts getting a higher score, unlike the C-Score. The E-Score has a range between 0 and about 2, also with more complex

[5] http://simple.wikipedia.org

[6] http://en.wikipedia.org

[7] The Simple English article for art would be from http://simple.wikipedia.org/wiki/Art while that for the English Wikipedia article would be an extract from http://en.wikipedia.org/wiki/Art

[8] http://morphadorner.northwestern.edu

Passage	C-Score	Flesch	Lexile	E-Score
Art	2.02	**1.41**	1.25	0.91
Culture	0.88	2.96	1.51	**0.65**
History	1.55	2.23	**1.33**	1.13
Film	0.47	1.51	**0.95**	1.04
Music	1.49	1.78	**1.35**	1.11
Sports	0.56	1.04	0.96	**0.90**
Science	1.06	1.59	1.56	**0.90**
World	0.59	1.85	1.76	**1.09**

Table 3: Comparison of complexity ratios of different passages with different metrics. Ratios in bold are those closest to the ratio got from the data we got using the C-Score

texts getting a higher score. Therefore, in order to normalize the values for comparison, we take the ratio of complexity (i.e. how complex the English Wikipedia article is compared to the equivalent Simple English Wikipedia article). To see how close we are to the gold-standard ratios (ratio of the article's Simple English Wikipedia C-Score value to that of its corresponding English Wikipedia C-Score value) that we got from our C-Score experiment, we use the following error metrics (lower is better).

1. $S0 = \frac{\sum_{i=1}^{n} x_i}{n}.x_i = 0$ if the metric is closest to the gold standard ratio and $x_i = 1$ otherwise. This measures percentage of the metric's ratio not agreeing with that of the gold standard ratio.

2. $S1 = \frac{\sum_{i=1}^{n} |metric_i - gold_i|}{n}$ is the mean absolute error between the metric's ratio and the gold standard ratio.

3. $S2 = \frac{\sum_{i=1}^{n} (metric_i - gold_i)^2}{n}$ is the mean square error between the metric's ratio and the gold standard ratio.

Evaluation Metric	S0	S1	S2
Lexile	0.63	0.54	0.38
Flesch	0.88	0.87	1.05
E-Score	**0.50**	**0.47**	**0.29**

Table 4: Results of error analysis. Bold denotes the evaluation metric with the least error

The Flesch Reading Ease Score assumes that the complexity of the text is dependent only on the sentence length and the number of syllables per word. It considers words like "automobile" and "procrastinate" to be of same complexity because both words have 4 syllables. With the use of data though, it can be shown that "automobile" is far more easier as compared to "procrastinate" (because "automobile" is on the Dale Chall Word List[9], while "procrastinate" is not).

The Lexile Score makes use of a corpus, in which it assumes that the frequency of a word determines its simplicity / complexity. More frequent the word is, simpler it is. While this is probably true in most cases, one of the issues is that it is corpus dependent. For instance, a medical corpus would have terms like disease names, drugs, etc. being as common / more common than common everyday phrases like "traffic light".

The E-Score outperforms the other two because it takes into account factors like corpus complexity, and syllable count. Corpus complexity gives a more precise measure than just frequency, of how complex an n-gram is, by measuring how much more probable it is in a parallel simplified corpus. Our metric's measure of structural complexity also measures the fact a complex sentence is shown by having more information in it, as compared to just the number of words in it.

6. Conclusions

Using the document aligned English - Simple English Wikipedia Corpus, we are able to assign weights (i.e. the corpus complexity) to n-grams that occur in text, unlike FRES. We also look at quantities like corpus complexity (Biran et al., 2011) while assigning the complexity of a word, as well as the number of syllables, unlike Lexile, which only looks at the frequency of words in a given corpus. Our language modelling approach, in which we measure lexical complexity using n-grams, rather than just words is also an improvement over Lexile and FRES. If we were to, say, reorder the phrases of the sentence (so that we still end up with the same structural complexity), FRES and Lexile would give the same score, but our approach would give a different score, showing that the reordered sentence may be harder than the original. For example, the sentence, "Join the Dark Side, the boy will"[10] will give a different E-Score value, compared to the "The boy will join the Dark Side". The FRES and Lexile scores for both sentences though will remain the same. Using structural complexity in our calculation of complexity is also better than that of the FRES and Lexile scores which take into account only the the number of words of the sentence and not its structure.

7. Bibliographical References

Biran, O., Samuel, B., and Elhadad, N. (2011). Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, volume 2, pages 496–501. Association for Computational Linguistics.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.

Heilman, M. and Smith, N. A. (2010). Extracting simplified statements for factual question generation. In *Proceedings of QG2010: The Third Workshop on Ques-tion Generation*, page 11.

Kauchak, D. (2013). Improving text simplification language modelling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1537–1546. Association for Computational Linguistics.

[9] http://www.rfp-templates.com/Research-Articles/Dale-Chall-3000-Simple-Word-List

[10] Quote from Yodha in *Star Wars Episode I: The Phantom Menace*

Schwarm, S. E. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.

Stenner, A. J. (1996). Measuring Reading Comprehension with the Lexile Framework. *ERIC*.

Temnikova, I. and Maneva, G. (2013). The C-Score–Proposing a Reading Comprehension Metric as a Common Evaluation Measure for Text Simplification. In *Proceedings of the Second Wrokshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 20–29.