A Statistical Model for Scientific Readability

Luo Si and Jamie Callan
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lsi@cs.cmu.edu, callan@cs.cmu.edu

ABSTRACT

This paper presents a new method of using statistical models to estimate the reading difficulty of Web pages. Language Models are used to represent the content typically associated with different readability levels. Reading level classifiers are created as linear combinations of a language model and surface linguistic features. Experiments show that this new method is more accurate than the widely used Flesch-Kincaid readability formula

KEYWORDS

Readability, Flesch-Kincaid, Unigram Language Model, EM.

1. INTRODUCTION

Readability metrics are intended to identify the difficulty of understanding a passage of text. Readability metrics are often based on features such as the average number of syllables per word, and words per sentence. These features ignore concept difficulty and are based on assumptions about writing style that may not hold in all environments. Web documents, for example, have very different characteristics than newspaper articles or pages in a textbook, which can cause traditional readability metrics to produce misleading estimates of readability.

Our hypothesis is that readability estimates would be more accurate on a wider range of document types if the readability metric incorporated information about the document content. In particular, we were interested in educational Web pages, which traditional readability metrics often misjudged because the pages are short and more likely to include text fragments that are not organized into traditional sentences and paragraphs.

This paper presents a new method of estimating readability that combines traditional readability features with statistical models. The EM algorithm is used with a set of training data to combine different types of information. Experiments demonstrate that this new approach to readability produces more accurate readability estimates for K-8 science Web pages than the widely used Flesch-Kincaid readability formula.

CIKM '01, Nov 5-10, 2001, Atlanta, GA.

2. READABILITY METRICS

There are many readability metrics. FOG, SMOG and Flesch-Kinciad are three of the most widely used readability metrics. They all estimate the educational grade level necessary to understand a document.

The FOG readability metric [2] is defined as:

$$\text{GradeLevel} = 3.0680 + 0.877 * \text{AverageSentenceLength} + 0.984 * \text{PercentageOfMonosyllables}.$$

The SMOG readability metric [3] is defined as:

$$\text{GradeLevel} = 3 + \text{Sqrt (Number of Polysyllable Words in 30 Sentences)}.$$

If the document is longer than 30 sentences, the first 10 sentences, the middle 10 sentences, and the last 10 sentences are used. If the document has fewer than 30 sentences, some rules and a conversion table are used to calculate the grade level [3].

The Flesch-Kincaid readability metric [4] is defined as:

$$\text{GradeLevel} = 0.39 * \text{AvgNumberWordsPerSentence} + 11.80 * \text{AvgNumberSyllablesPerWord} - 15.59$$

The FOG metric is considered suitable for secondary and older primary age groups. The SMOG measure tends to give higher values than other readability metrics [5]. Flesch-Kincaid is used more often than the FOG and SMOG metrics. It is a U.S. Department of Defense standard [4].

3. STATISTICAL LANGUAGE MODELS

Most readability metrics ignore document content and only consider surface linguistic features. However, some surface linguistic features, such as average number of words per sentence, are influenced by presentation style, and some monosyllable words, such as "quark", represent concepts that are not easy to understand. The parameters in existing readability metrics are assigned manually, and require adjustment when the domain changes. These weaknesses suggest a new approach.

Our hypothesis was that readability measures should be sensitive to content as well as to surface linguistic features. We also believe that model parameters should be learned from actual corpora.

Statistical Language Models are widely used to capture the regularities of natural language in order to improve the

performance of natural language applications [6]. Our second hypothesis was that statistical language models could capture the content information related to reading difficulty.
Unigram language models assume that the probability of generating a word is independent of its context. Although unigram language models are weak models of human language, they are adequate for many applications, and have the advantage that they can be trained from smaller amounts of data. We chose unigram language models for the work reported here.

We treated readability estimating as a kind of text categorization problem, as follows:

$$\begin{aligned} G^* &= \arg\max_g \; P(g \mid d) \\ &= \arg\max_g \frac{P(d \mid g) * P(g)}{P(d)} = \\ &= \arg\max_g \sum_{w \in d} \log(P(w \mid g)) \end{aligned}$$

g represents different readability level, d represents a specific document, P(d|g) is the unigram language model for different readability grade level, and P(g) is the prior probability of different levels. P(w|g) represents the probability of generating a single word from a specific readability level. A uniform distribution of the prior probability is assigned and P(d) is treated as a constant.

A combination of surface linguistic features and content information may be better than either alone. Features based on sentence length and number of syllables per word are common in readability measures such as the FOG, SMOG and Flesch-Kincaid metrics, presumably based on the hypotheses that longer sentences, and sentences containing longer words, are more difficult to read. Our first task was to investigate the reliability of these types of features in our application domain.

A sample corpus of 91 Web documents was created (see Section 4). The sentence length distribution for three readability levels in this corpus is shown below.

Categories	**Mean**	**Variance**
k-2	10.6	6.84
3-5	13.9	9.47
6-8	14.7	8.95

The mean values of sentence lengths in the three different levels increase monotonically. We conclude that sentence length is a useful feature for this problem.

The distribution of word lengths in a sentence, measured in syllables, is shown below.

Categories	**Mean (Syllables)**	**% Words With 3 or More Syllables**
k-2	1.5	10.9
3-5	1.6	17.7
6-8	1.6	15.0

The mean number of syllables per word is not a useful feature for this dataset. The mean values do not increase monotonically with grade level (the average syllable number of the grade level 6-8 is less than that of 3-5). Some readability metrics, such as SMOG, use the percentage of polysyllable words in a document as an indicator of readability level. That feature is not a reliable indicator of readability in this corpus. Web pages written for grades 3-5 had more polysyllable words than Web pages written for grades 6-8.

Our third hypothesis was that the normal distribution can be used to model the sentence length distribution. This means that a normal distribution with a specific mean and variance can be used to model the sentence length distribution of each readability grade level. The normal distribution may not perfectly model the sentence length distribution, but our hypothesis was that it would be close enough. It was modeled as shown below.

$$\begin{aligned} G^* &= \arg\max_g \; P(g \mid d) \\ &= \arg\max_g \sum_{s \in d} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(s_l - s_{\lg})^2}{2\sigma^2}}\right) \end{aligned}$$

s represents each sentence in this document. The s_l represents the length of each sentence, $s_{\lg}$ is the mean value of sentence length in each readability level, σ is the standard deviation value of sentence lengths.

Linear combination is a common and intuitive choice for combining different models, and the EM algorithm [7] is often used to calculate the optimal parameters in the linear model. A linear interpolation formula was chosen to combine the language model and the sentence length model in our application.

$$P_c(g \mid d_i) = \lambda * P_a(g \mid d_i) + (1-\lambda) * P_b(g \mid d_i)$$

In this formula, $P_a(g \mid d_i)$ represents the unigram language model and $P_b(g \mid d_i)$ represents the sentence length distribution model. λ is the weight value between this two models.

The EM algorithm was used to calculate the optimal λ. Our goal is a lambda value that maximizes the following criterion:

$$\lambda = \arg\max_\lambda \sum_d \log(\lambda * P_a(c = correct \mid d) + (1-\lambda) P_b(c = correct \mid d))$$

0. Set lambda an initial value:
1. Calculate the probability of generating the correct readability level by the language model and sentence distribution model.
2. Update the lambda parameter according to the following formula:

$$\lambda^{j+1} = \frac{1}{n} \sum_d \frac{\lambda^j P_a(c = correct \mid d)}{\lambda^j P_a(c = correct \mid d) + (1-\lambda^j) P_b(c = correct \mid d)}$$

3. Test whether the algorithm has converged; if not, go to step 1.

4. EXPERIMENTAL RESULTS

Our goal was a readability metric for educational Web pages. A set of representative Web pages was collected by searching for pages related to science education and then selecting pages that were written by students and included a grade-level or age, or were written by adults and indicated the grade-level or the age of the intended audience. Less than one hundred pages were collected in this manner. Children have varying writing abilities, and adults write for their intended audience with varying levels of success, so the collected pages were grouped into bins covering grade levels, for example K-2, 3-5, and 6-8.

Preliminary experiments produced mixed results, in part because of the small amount of training data and the large amount of variation in the training data.

A second approach to acquiring training data is to use the syllabi of elementary and middle school science courses. Three sets of syllabi (one per readability level) were collected from different Web sites. Initially we used only three scientific syllabi to train the language models. We found that the content of scientific Web pages has a strong relationship with some other topics, especially mathematics. For the remainder of the experiments mathematics syllabi were combined with the original scientific syllabi to build language models.

A total of 91 Web pages downloaded from the web. The reading level of each page was indicated by the source Web or was inferred based on the age of the author. Pages were grouped into three readability levels: Kindergarten-Grade2, Grade3-Grade5, and Grade6-Grade8. The EM algorithm was used to calculate the lamda parameter in the combined model. 10 training pages from each category were used as training data. 61 Web pages were used for testing.

The EM algorithm converged quickly. The final value of lambda was 0.91, indicating that the unigram language model was the most reliable form of evidence, but that the surface linguistic features were also helpful. The test result is summarized below.

Model	**Accuracy**
Lambda=1.00	70.5%
Lambda=0.91	75.4%
Lambda=0.00	42.6%
Flesch-Kincaid	21.3%

The Flesch-Kincaid metric gives a numerical value that ranges from 0 to 12. This numerical value represents a readability level from Kindergarten to Grade 12. The accuracy of this metric is calculated as the number of Web pages identified with appropriate grade category (bin) divided by the number of test Web pages. In this experiment, Flesh-Kincaid did worse than random selection.

One problem with the Flesch-Kincaid metric is that it is only allowed to be numerical value from 0 to 12. Many documents' grade levels are calculated as more than 12, but are cut off at 12, so that there is no difference in the readability levels of these documents. Our statistical method gives a probability to each readability category. The Web page can not only be identified with a single grade level as a hard metric but also the probability value of each grade level can be deemed as a soft metric. That is a very useful property in many applications.

From this experiment, we conclude that the language model is a more important factor for determining readability than sentence length. The sentence length distribution feature is of some help. The combined model has a better performance than each single model.

5. CONCLUSION

Most existing readability metrics are based on surface linguistic features of the text but ignore the content information. This paper presents a novel statistical model to identify the readability level of K-8 science Web pages. Our method uses unigram language models to represent the content typically associated with each readability level.

Sentence length and features based on the number of syllables per word are surface linguistic features often used in existing readability formulae. Our experiments with K-8 science Web pages show that sentence length is a good feature but the mean number of syllables per word is not.

A linear model was used to combine content-based and surface linguistic features into a single classifier. The EM algorithm was used to find the optimal parameters in this combined model, instead of setting them manually as done in existing readability metrics. This approach allows parameters to be set or tuned easily and automatically for particular domains or applications, even domains or applications where the features are not all equally useful. Our experiments showed that the combined model is much more accurate on K-8 science Web pages than the widely used Flesch-Kincaid readability metric.

6. ACKNOWLEDGEMENTS

This material is based on work supported by NSF grant IIS-0096139. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] Gilliland, J. *Readability.* University of London Press, 1972.

[2] Gunning, R. *The Technique of Clear Writing.* McGraw-Hill, 1952.

[3] McLaughlin, H. "SMOG grading – a new readability formula." *Journal of Reading*, 22, pp. 639-646. 1962.

[4] http://www.itl.nist.gov/iaui/ovrt/people/sressler/Persp/Views.html

[5] http://www.timetabler.com/reading.html

[6] Ronald Rosenfeld. "Two decades of statistical language modeling: Where do we go from here?" *Proceedings of the IEEE*, 88(8), 2000.

[7] A. P. Dempster, N. M. Laird and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, volume 39, number 1, pages 1-38,1977.