Approximation of the Parameters of a Readability Formula by Robust Regression

Tim vor der Brück

Intelligent Information and Communication Systems (IICS)
FernUniversität in Hagen
58084 Hagen, Germany
tim.vorderbrueck@fernuni-hagen.de

Abstract. Most readability formulas calculate a global readability score by combining several indicator values by a linear combination. Typical indicators are *Average sentence length*, *Average number of syllables per word*, etc. Usually the parameters of the linear combination are determined by a linear OLS (ordinary least square estimation) minimizing the sum of the squared residuals in comparison with human ratings for a given set of texts. The usage of OLS leads to several drawbacks. First, the parameters are not constraint in any way and are therefore not intuitive and difficult to interpret. Second, if the number of parameters become large, the effect of overfitting easily occurs. Finally, OLS is quite sensitive to outliers. Therefore, an alternative method is presented which avoids these drawbacks and is based on robust regression.

1 Introduction

Estimating the readability of texts has a long tradition. Normally the readability is described numerically and calculated by a readability formula. According to George Klare [**?**, p.34], "a readability formula is a method of estimating the probable success a reader will have in reading and understanding a piece of writing". More formally, a readability formula $R : A^* \rightarrow \Re^+$ assigns a text t over an alphabet A a numerical value $R(t)$ where usually a large value relates to good readability, a low value to poor readability. The first readability formulas were developed in the middle of the 20th century[**?**].

Usually a readability formula combines several indicator values by a linear combination [**?**]. An example of such a readability formula is the Amstad readability index [**?**] which is defined as follows:

$$\begin{aligned} R_{\text{Amstad}} = 180 &- (1.0 \cdot \text{ASL}) \\ &- (58.5 \cdot \text{ASW}) \end{aligned} \tag{1}$$

where ASL denotes the average sentence length in words and ASW denotes the average number of syllables per word.

The parameters of such a linear combination are normally determined using OLS by minimizing the (R)MSE[1] to human ratings for a given set of texts. One major drawback of the usage of OLS consists in the fact that OLS is quite sensitive to outliers. Also, if

[1] (R)MSE is the abbreviation of (root) mean squared error

a large number of indicators is used, the effect of overfitting can easily occur. Furthermore, it is quite difficult for a human expert to estimate and compare the importances of the indicators from the parameter values.

Thus, in our readability checker DeLite an alternative approach is used. First all indicator values are normalized to an interval from zero to one. Afterwards, the normalized indicators are combined by a weighted sum, where all weights are positive and sum up to one (convex combination). To determine the weights a robust regression method is used which is quite insensitive to outliers. Non-relevant indicators are assigned a weight of zero and can therefore be removed from the model. Furthermore, indicators with low importance are assigned a small weight and their influence is therefore strongly limited. Finally, the importance of each indicator is immediately obvious from the indicator weight.

2 Robust Regression Techniques

Least square estimation has the major drawback that, due to the minimization of the mean squared error, this approach is quite sensitive to outliers. Often, such outliers cannot be filtered out by outlier detection algorithms since the associated coordinates of the outliers might not differ from the other data points in absolute terms but only regarding the fit to some function. Hence, several methods were proposed to overcome this problem [**?**] which are called robust regression methods. Quite popular robust regression techniques are:

Least Median of Squares: Instead of the squared sum of residuals, the median of the squared residuals is minimized. For that, the residuals are sorted from small to large. The index of the regarded residual in this sorted collection is then determined by $\mathrm{round}(n/2 + (p+1)/2)$ where n denotes the sampling size and p the number of parameters. The parameters determined by this approach describe the center of the smallest plane covering the majority of the data where the distance is measured along the coordinate of the explained variable. Note that all data vectors outside this plane are completely ignored by this algorithm.

Least Trimmed Squares: The least trimmed squares estimator minimize the sum of the h smallest squared residuals where h is usually set to $\mathrm{round}(n(1-\alpha)+1)$ for some α between zero and one.

Minimizing the sum of the absolute residuals: Instead of the sum of the squared residuals the sum of the absolute residuals is minimized. Such a minimization problem can be solved by linear optimization. Linear optimization has the further advantage over OLS that it allows the definition of inequality constraints with minimal overhead. Such constraints are required to ensure that all indicator weights are nonnegative. Thus, we decided to use this method to determine the weights and normalization parameters of our readability function.

3 Checking Readability

3.1 Readability Indicators

We employed robust regression analysis on the parameters of the DeLite readability checker. This readability checker investigates the readability of German texts on several linguistic levels:

- Morphological level
  - Example indicator: Number of components in a compound word
  - Example: *Donaudampfschifffahrtsgesellschaft* (Donau-dampf-schiff-fahrts-gesellschaft)
  - Translation: *'Donau streamship company'*
  - Value: 5
- Lexical level
  - Example indicator: Number of different readings of a word
  - Example: *Raum*
  - Translation: *'space'*, *'room'*, *'scope'*
  - Value: 3
- Syntactic level
  - Example indicator: Embedding Depth
  - Example: *Er verließ das Haus, in dem die Frau, die er liebte, wohnte, sofort.*
  - Translation (literally): *'He left the house where the woman he loved lived immediately.'*
  - Value: 3 for *liebte 'loved'*
- Semantic level
  - Example indicator: Large number of propositions per sentence
  - Example: *Das könnte bewirken, dass der Fahrer aus Angst vor den Nachbarn die Geschwindigkeit reduziert.*
  - Translation: *'This could achieve that the driver reduces the speed for fear of the neighbors.'*
  - Value: 3
- Discourse level
  - Example indicator: More than one potential reference for a pronoun
  - Example: *Dr. Peters lädt Herrn Müller zum Essen ein, da heute sein Geburtstag ist.*
  - Translation: *'Dr. Peters invites Mr. Müller for dinner since it is his birthday today.'*
  - Value: 2 (the word *sein 'his'* can either refer to *Mr. Müller* or *Dr. Peters*)

In order to calculate the indicator values described above, a semantic network, a dependency tree and a list of tokens are derived for each sentence by a deep syntactico-semantic analysis [**?**].

3.2 Calculating the Readability Score

The global readability score of DeLite is calculated in the following steps (see Fig. 1 for an illustrative example with the two indicators *Average sentence length* and *Average number of characters per word*)

- Segmentation: First, the whole document is split up in sentences, phrases and words.
- Calculation: In the next step the indicator values are determined for each segment the associated indicator is applicable to. There are indicators which operate on word, sentence, phrase or document level.
- Aggregation: Now the indicator values are aggregated by averaging. This means that, for each indicator, there exists as a result of this step a single aggregated value for the entire document. Consider for example a text containing two sentences with length 8 and 10 in the text. Then the aggregated value for the indicator *Average sentence length* is 9 if arithmetic averaging is used.
- Normalization: The indicators are normalized by applying a normalization function.
- Combination: Finally, a global readability score is determined by combining the indicator values.

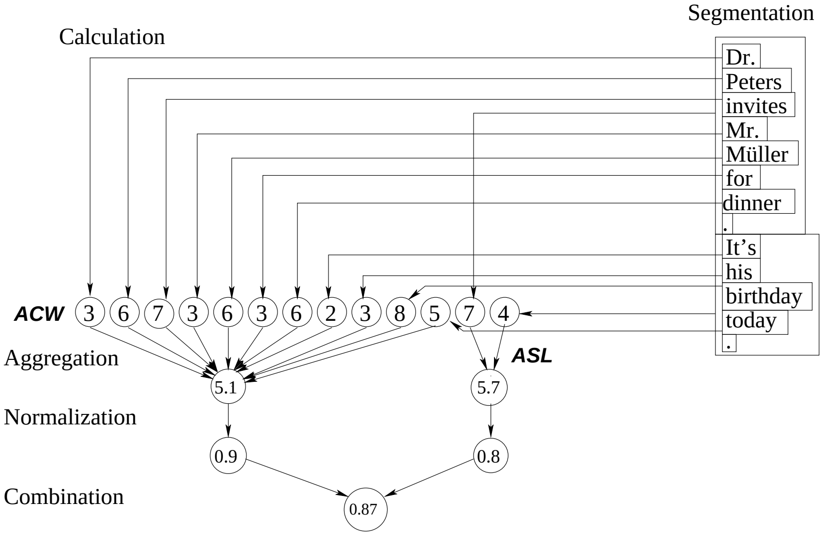

Fig. 1. Steps for calculating a global readability score in DeLite. For better illustration only the two indicators *Average sentence length*(ASL) and *Average number of characters per word*(ACW) are displayed.

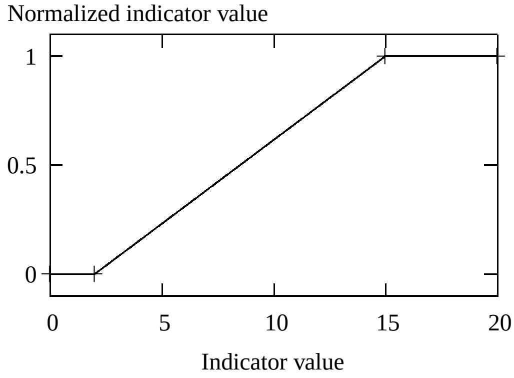

Fig. 2. Typical normalization with a piecewise linear function for data with the minimum 2 and maximum 15.

4 Normalizing Indicator Values

Before the data is combined, it is normalized to the interval [0,1]. Such a normalization is often done by a piecewise linear normalization function (depicted in Fig. 2). However, such a normalization has several drawbacks. First, the normalization function is not differentiable at the two locations where the slope changes. This makes it difficult to apply optimization techniques which employ the derivative. Second, a smooth transition seems to be more natural. Therefore we decided to use a derivation of the Fermi-Function which is defined as follows:

$$N(x, \mu, \delta) = 1 - \frac{1}{1 + e^{-\frac{x-\mu}{\delta}}}$$

where

- x: indicator value
- μ, δ: constants which are associated to a certain indicator where the parameter μ is the location of the 0.5-intercept ($N(\mu) = 0.5$) and δ specifies the incline of the function.

The graph of this function is displayed in Fig. 3.

In our readability checker, all indicator values are non-negative and high unnormalized indicator values correspond to less readability.

In an initial estimation, μ_j is set to the *mean value* of the distribution. The parameter δ_j was obtained by computing the arithmetic mean for solutions of $N(x, \mu_j, \delta_j)$ for given values of μ and maximum and minimum values of the indicator value x under consideration. This initial estimation does not make use of the user ratings at all.

5 Determining Parameter Weights with Linear Optimization

The robust regression used here estimates the weights by minimizing the sum of the absolute residuals instead of the sum of the squared residuals. The minimization is

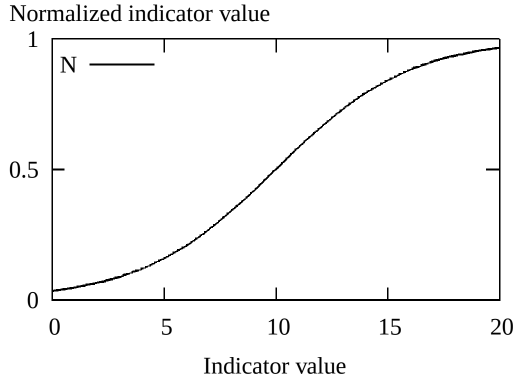

Fig. 3. Normalization with a variant of the Fermi-Function.

achieved via linear optimization, applying the Simplex method [**?**]. This kind of regression is called robust since it is not as sensitive to outliers as OLS.

The minimization problem for determining the weights of our readability function is defined as follows:

$$\mathbf{w}_{\text{opt}} = \arg\min_{\mathbf{w}}(\sum_{i=1}^{n} |y_i - \mathbf{X_i w}|) \tag{2}$$

where

- y_i: average user rating for text t_i
- $\mathbf{w} = (w_1, \ldots, w_m)^T$: vector of indicator weights
- $\mathbf{X_i} = (x_{i1}, \ldots, x_{im})$: vector of indicator values for text t_i

In Equation 2, $|y_i - \mathbf{X_i w}|$ can be replaced by variables z_i, if the constraints

$$z_i \geq |y_i - \mathbf{X_i w}| \text{ with } 1 \leq i \leq n \tag{3}$$

are added [**?**]. Using the equivalence in Equation 4, the optimization problem can be rewritten as shown in Equation 5.

$$z_i \geq |y_i - \mathbf{X_i w}| \Leftrightarrow z_i \geq (y_i - \mathbf{X_i w}) \wedge z_i \geq -(y_i - \mathbf{X_i w}) \tag{4}$$

$$\begin{array}{c} \arg\min_{\mathbf{w}} z_1 + \ldots + z_n \text{ , with} \\ z_i \geq y_i - x_{i1}w_1 - \ldots - x_{im}w_m \text{ ,} \\ z_i \geq x_{i1}w_1 + \ldots + x_{im}w_m - y_i \end{array} \tag{5}$$

and $i = 1, ..., n$.

This problem consists of linear equations only and can therefore be solved by traditional linear optimization algorithms like the Simplex method.

6 Further Error Reduction

Currently, the user ratings are not used at all for the determination of the normalization parameters. However, for an optimal setting, i.e., a setting which minimizes the error to the user ratings, this can be an important factor. Thus, in this section a method is described to incorporate these ratings but additionally avoiding the complexity of a full-blown non-linear robust optimization. We actually succeeded in further reducing the MAE (mean absolute error) using an iterative approach. For that, only the parameters δ and μ of the normalization function for a single indicator are modified, leaving all others parameters and weights constant. This optimization process is iterated for each indicator. In each iteration step the error decreases if evaluated on the training data. Naturally, this does not hold necessarily for the error determined by cross-validation. In contrast to a nonlinear robust optimization, this approach is quite efficient and easy to realize.

Let us now investigate how the parameters δ_k and μ_k are determined for a single indicator I_k.

We look for parameters μ_k, δ_k which lead to a small sum of the absolute residuals $|\epsilon_1| + \ldots + |\epsilon_n|$ with

$$y_i = \sum_{j=1}^{m} w_j N(x_{ij}, \mu_j, \delta_j) + \epsilon_i \tag{6}$$

assuming δ_j, μ_j have a constant value with $j = 1, \ldots, m$ and $j \neq k$ and all w_j including w_k are constant.

In the first step, all summands except the kth are moved to the left side:

$$y'_i := y_i - \sum_{1 \leq j \leq m, j \neq k} N(x_{ij}, \mu_j, \delta_j) = w_k N(x_{ik}, \mu_k, \delta_k) + \epsilon_i \tag{7}$$

Variables a, b are introduced with $a := -1/\mu_k$ and $b := \frac{\mu_k}{\delta_k}$. y'_i can be rewritten as shown in Equation 8.

$$y'_i = w_k(1 - \frac{1}{1 + e^{ax_{ik}+b}}) + \epsilon_i. \tag{8}$$

$ax_{ik} + b$ can be isolated [**?**] by making some equivalence conversions.

$$ax_{ik} + b = \ln(\frac{1}{1 - \frac{y'_i - \epsilon_i}{w_k}} - 1) \tag{9}$$

Instead of solving this problem directly we look for the solution a, b of the related problem[2]

$$ax_{ik} + b + \gamma_i = \ln(\frac{1}{1 - \frac{y'_i}{w_k}} - 1) \tag{10}$$

[2] An optimal solution of the second problem creates a nearly optimal solution for the first problem.

by minimizing $|\gamma_1| + \ldots + |\gamma_n|$. Since the expression on the right side is considered to have a constant value, Equation 10 represents a linear model. a, b can be determined by solving the following minimization problem:

$$\arg\min_{a,b} \sum_{i=1}^{n} |y_i'' - ax_{ik} - b| \tag{11}$$

where y_i'' is given as:

$$\ln\left(\frac{1}{1 - \frac{y_i'}{w_k}} - 1\right) \tag{12}$$

Equation 11 can be solved for a and b by linear optimization. Note that the standard Simplex method requires the optimization variables a, b to be nonnegative. This can be achieved by making further replacements:

$$ax_{ik} = a_+ x_{ik} - a_- x_{ik} \tag{13}$$

$$b = b_+ - b_- \tag{14}$$

A negative factor for x_{ik} is obtained if $a_+ < a_-$. Analogously for b. The original parameters μ and δ can then easily be determined from the solutions for a and b.

After all normalization parameters are obtained in the indicated way, the weights of the indicators can be recalculated as described in Sect. 5 on the basis of these newly calculated parameters. Afterwards, the normalization parameters can again be recalculated using the new weights which means that, theoretically, this process can be repeated indefinitely. The entire optimization process is illustrated using pseudocode in Fig. 4.

```
for v=1;v≤num_iterations;v++
 w=determine_weights($\mu_{1\ldots m}$,$\delta_{1\ldots m}$,$\mathbf{X}$,y);
 for (k=1;k≤m;k++)
  ($\mu_k, \delta_k$)=determine_pars($\mu_{1\ldots(k-1),(k+1)\ldots m}$,$\delta_{1\ldots(k-1),(k+1)\ldots m}$,
                 w,$\mathbf{X}$,y,k);
```

Fig. 4. Pseudocode for determining indicator weights and normalization parameters. $\mu_{1\ldots m} := (\mu_1, \ldots, \mu_m)$, $\delta_{1\ldots m}$ is defined analogously.

7 Evaluation

The evaluation was based on an online user study conducted with more than 300 participants rating the readability of 500 German texts. In total, the data consist of about 2 800 readability judgments. The participants rated the readability of each text on a seven point Likert scale [**?**].

43.1 % of the participants were female and 56.9 % male. 91.4 % of them were German native speakers. Four people were not native speakers and their German language skills were, according to their own judgment, worse than “Good”. Since the aim of this experiment was to test the readability for German native speakers their ratings were filtered out. Readability experiments for non-native speakers were not carried out and left for future work.

Almost 70 % of the participants were between 20 and 40 years old; the number of participants over 60 was very small (ca. 3 %). The participants were mainly well-educated. 58 % of them owned a university or college degree. There is none who had no school graduation at all. The participants of the evaluation belonged to a large variety of professions, e.g., software-developers, scientists, physicians, linguists, pharmacists, administrators, psychologists, and musicians.

The texts were automatically extracted from Web pages of local administrations. For that, we looked for PDF texts employing a search engine with typical keywords for this domain. All PDF texts were converted into plain texts with pdftotext and processed by the deep syntactico-semantic analyzer WOCADI[**?**] which creates a semantic network, a dependency tree and a list of tokens. This information is then employed by the DeLite readability indicators. 53 readability indicators are used in total.

The parameters of the normalization function and the weights were determined by the algorithm described in Sect. 4–6 using robust regression. RMSE and MAE were determined between the average user rating of a text and the readability score employing the learned weights and parameters by a 10-fold cross-evaluation and after several iterations of the parameter/weight learning process.

The results are displayed in Table 1 and Fig. 5. MAE and RMSE were considerably reduced in comparison to a normalization which does not take into account the user ratings (number of iterations equals to zero). Furthermore, no considerable improvement was observed using more than two iterations.

Table 1. MAE and RMSE between the DeLite score and the readability ratings.

Type of Error	Iterations				
	0	1	2	3	10
MAE	0.140	0.125	0.119	0.119	0.120
RMSE	0.169	0.153	0.148	0.149	0.150

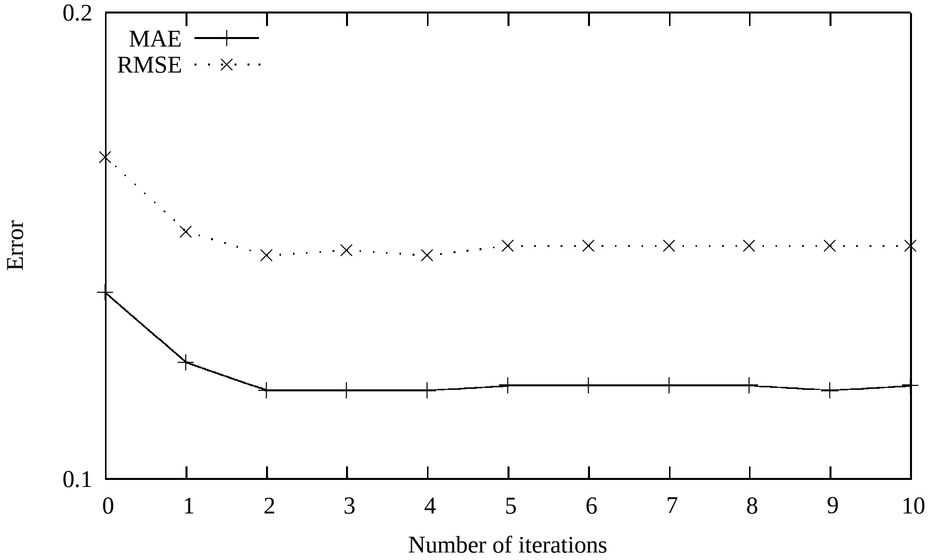

Fig. 5. MAE and RMSE between the user ratings and the DeLite readability score for the given number of iterations.

8 Conclusion and Future Work

This work describes a robust regression approach to determine the parameters and weights of a readability function. It avoids the computational complexity of a full-blown nonlinear robust optimization. Furthermore, the MAE and RMSE are considerably lower than the errors by following a naïve approach which does not use any optimization technique at all for determining the normalization parameters. However, by not modifying all parameters and weights simultaneously, it is not guaranteed that the actual minimum can ever be reached. But still, the parameters determined by this approach can be used as a first guess (or as one of the first guesses) for a nonlinear optimization algorithm.

One aspect, which needs more attention, is the sequence in which the parameters are processed by the optimization algorithm described in this work which can have an impact on the result. Furthermore, a comparison with a nonlinear robust optimization algorithm which determines the exact solution as well as with evolutionary algorithms [**?**] or optimization methods using swarm intelligence [**?**] would be interesting.

9 Acknowledgements

We wish to thank our colleagues at the FernUniversität in Hagen for their support, especially Ingo Glöckner for proof-reading this document.