(Not) Finding latent readability scores in binary annotated text using Genetic Search

Johan Falkenjack

19 January 2016

Disclaimer

This is a short and informal paper on a novel approach to finding latent readability scores written for the final project in the course MAI0083 Metaheuristics. It is not intended for publication and should not be read as a potential submission for publication. Most of the Background section is taken more or less verbatim from one of my previous published papers on readability classification.

Background

The problem of readability assessment is the problem of mapping from a text to some unit representing the text's degree of readability.

Measures of readability are mostly used to inform a reader how difficult a text is to read, either to give them a hint that they may try to find an easier to read text on the same topic or simply to inform them that a text may take some time to comprehend. This can be especially useful for persons with reading disabilities, but can also be used to, for instance, assist teachers with assessing the reading ability of a student. By measuring the reading abilities of a person, it might also be possible to automatically find texts that fits that persons reading ability.

Readability gives rise to a number of problems. For instance, readability is not a function of text only but a function of both text and reader, as defined by (Dale and Chall 1949): "[Readability is] the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at optimal speed, and find it interesting."

However, in this study we make the assumption that a function of text only can be a useful approximation. This assumption is supported by and related to the practice of American researchers to normalize their metrics to the U.S. grade level. In later years that normalization has been made easier by the existence of a corpus of texts tagged with suitable grade level, Weekly Reader, not available to us at this time. These grade levels can be used both as a basis for regression (Pitler and Nenkova 2008) and for creation of detectors, single-class classifiers (Petersen 2007).

Resources for such a normalization for Swedish are not yet readily available, other than as test data, and until they are we have been focusing on the problem of classifying texts as either easy-to-read or not. In this work, I use a novel approach to try to generalize this classification into a regression by looking for the underlying structures which give rise to the binary readability labels.

Existing measures for Swedish

Readability assessment for Swedish has mostly been done using metrics similar to metrics constructed for other languages. The most utilized readability metric for Swedish is LIX, Läsbarhetsindex (Readability index) (Björnsson 1968).

$$LIX = \frac{n(w)}{n(s)} + \left(\frac{n(lw)}{n(w)} \times 100\right)$$

Where $n(w)$ is the total number of words in the text, $n(lw)$ is the total number of long words (more than 6 characters) and $n(s)$ is the number of sentences. Today the LIX metric is basically the standard metric for readability in Swedish. However, in recent years new research has shown that the metric is not always sufficient (Mühlenbock and Johansson Kokkinakis 2009).

The OVIX metric, Ordvariationsindex (Word variation index), and Nominal Ratio metric (Hultman and Westman 1977) have been used in research to complement LIX as they are assumed to correlate with degree of readability viewed from other linguistic levels.

$$OVIX = \frac{log(n(w))}{log(2 - \frac{log(n(uw))}{log(n(w))})}$$

Where $n(w)$ is the total number of words in the text and $n(uw)$ is the number of unique words. OVIX is basically traditional type-token-ratio with some logarithmic magic to minimize the influence of text length.

$$NR = \frac{n(noun) + n(preposition) + n(participle)}{n(pronoun) + n(adverb) + n(verb)}$$

Nominal ratio is a measure on the formality of the text where a larger value implies a more formal text, for instance scientific literature, and a smaller value implies more informal text, such as fiction.

Text features used for readability assessment

In this work I have been using a small set of 14 text features previously used for easy-to-read classification for Swedish (Heimann Mühlenbock 2013). These text features are:

- Average sentence length
- Average word length
- Ratio of words in the SweVoc word list
- Ratio of category C words in the SweVoc word list
- Ratio of category D words in the SweVoc word list
- Ratio of category H words in the SweVoc word list
- OVIX
- Nominal Ratio
- Mean distance between two dependent words in a dependency grammar sense, on a sentence basis
- Mean distance between two dependent words in a dependency grammar sense, on a word by word basis
- Number of subordinate clauses per sentence
- Number of prenominal modifiers per sentence
- Number of postnominal modifiers per sentence
- Mean parse tree height

Datasets

I'm working with two sets of binary labeled texts. One set, the easy-to-read set, comes from the LäSBarT corpus of easy-to-read Swedish texts. The other set comes from the Stockholm-Umeå Corpus v.2 (SUC2), which is a general text corpus for Swedish.

I also work with three sets of graded texts. The readability grades are based on the average performance of students reading the texts and then taking comprehension tests based on them. Each set consist of 6 texts with some overlap between the sets but different student groups for each set (4th grade, 6th grade and 8th grade in the Swedish school system).

Statistical approach, Probit regression

A Probit regression is a frequentist statistical model related to Logistic regression. The purpose is to fit a regression model to a binary response variable. In some sense, this can be interpreted as a traditional linear regression but with a link function applied, mapping the response onto the interval [0,1], when deviation is calculated. The final response can be interpreted as the probability of a data point belonging to a class of objects or not. Thus, Probit regression is actually more of a classification model than a regression model.

However, under the hood there is still a traditional linear regression being performed and this linear regression is usually viewed as having a latent response variable. If a Probit regression would be applied to my binary labeled texts the latent variable could be interpreted as the actual readability of the text.

My approach, Genetic search

The Probit approach has one weakness in that it, as most classification algorithms, assumes that training data is well separated. I.e. that there are no data points with the wrong label. In my data I know that there are such data points, namely, the SUC2 data probably includes a number of texts which might be labeled easy-to-read, but their label implies that they are not. The Probit regression use these labels when fitting the model and thus these erroneous labels introduce unnecessary noise.

Instead of assuming the truth of these labels I use them only as a starting point for a search through the space of possible latent readability scores.

Initial population

For each text, a random readability score is assigned, normally distributed around 1 for easy-to-read texts and normally distributed around 2 for other texts. Each individual is represented as a vector of such randomly generated readability scores for each text. I work with a population size of 20 individuals.

Objective function

The objective function consist of the mean absolute error from a 2-fold cross-validated Ordinary Least Squares regression plus the standard deviation of the absolute error. This means that the texts are randomly split into two sets of the same size. One set is used to train the OLS model and the other one is used to test it, then the sets are switched around and the second set is used to train an OLS and the first to test it. For each data point, the absolute residual is calculated (the absolute difference between the OLS estimated value, $\hat{y}$, and the randomly generated "true" value, y). The mean of these residuals, from both folds, is called mean absolute error. The standard deviation of absolute errors is used to penalize any outliers, i.e. any estimated values lying far from the "true" value.

$$residual_i = y_i - \hat{y}_i$$

$$mean(|residual|) + sd(|residual|)$$

Stop condition

From the beginning I stopped the search when the value of the objective function drops below 0.2 with a maximum of 100 000 generations. Because of variations in the result from different runs I narrowed the stop condition to first 0.15 and then 0.1, I also increased the maximum number of generations to 1 000 000.

Creating the next generation

The 4 best individuals from the current generation is automatically added to the next generation. Another 4 individuals are randomly generated in the same way as the initial population. The last 12 individuals are created using uniform crossover. The parents for this uniform crossover is selected using roulette wheel selection, a pair of parents are selected randomly (with higher probability given to more fit individuals) for each new individual. Fitness is calculated as the inverse of the objective function. The same individual might end up as parent to more than one child and potentially, the same two individuals might even generate more than one child together.

Results

The performance of different readability assessments are measured using two types of correlation metrics. The Kendall τ metric measure ordinal correlation, that is, to what degree two sets of rankings of the same objects agree. For instance, Kendall τ can measure the agreement of the ranking of search results from two different search engines.

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

The Pearson ρ, often just called the "correlation coefficient", is the most well known correlation metric and measure to what degree two sets of values on the same objects are correlated. For instance, to what degree the maximum daily temperature in Linköping agree with the maximum daily temperature in Norrköping.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where *cov* is the covariance between X and Y and σ is standard deviation (for X and Y respectively).

Established metrics

Table 1: Correlations between LIX score and student performances, should be negative.

	Kendall	Pearson
4th grade	-0.4666667	-0.1521346
6th grade	-0.3333333	-0.5595922
8th grade	-0.0666667	-0.1229710

Table 2: Correlations between OVIX score and student performances, should be negative.

	Kendall	Pearson
4th grade	-0.8666667	-0.8230110
6th grade	-0.3333333	-0.5910558
8th grade	0.0666667	0.1690629

Table 3: Correlations between Nominal Ratio and student performances, should be positive.

	Kendall	Pearson
4th grade	0.3333333	0.4368544
6th grade	0.4666667	0.4264147
8th grade	-0.0666667	-0.0436688

Probit regression

Table 4: Correlations between Probit estimated latent readability and student performances, should be negative.

	Kendall	Pearson
4th grade	-0.2000000	-0.1988830
6th grade	-0.2000000	-0.3988525
8th grade	-0.0666667	0.1320185

Genetic Search

I ran the search 5 times in order to see whether I found the same structure each time. Sadly, the result seems to be that we do not find the same structure each time.

Table 5: Results on the three test sets with an error value < 0.2 as stop criterion, should be negative.

	V2	V3	V4	V5	V6
kendall4	-0.4666667	-0.3333333	-0.3333333	-0.3333333	-0.3333333
kendall6	-0.6000000	-0.4666667	-0.6000000	-0.2000000	-0.4666667
kendall8	0.0666667	0.0666667	0.2000000	0.0666667	0.0666667
pearson4	-0.5034420	-0.4289294	-0.4771944	-0.4462790	-0.4156996
pearson6	-0.4864176	-0.3727685	-0.4387453	-0.1963957	-0.4527257
pearson8	0.2002415	-0.0674682	0.0567243	-0.0659640	0.0650918

Narrowing down the stop criterion to 0.15 or 0.1 does not solve this problem, however, see Discussion on possible reasons behind this result.

Table 6: Results on the three test sets with an error value < 0.15 as stop criterion, should be negative.

	V2	V3	V4	V5	V6
kendall4	-0.3333333	-0.3333333	-0.3333333	-0.4666667	-0.3333333
kendall6	-0.6000000	-0.6000000	-0.6000000	-0.6000000	-0.6000000
kendall8	0.0666667	0.0666667	0.0666667	0.2000000	-0.0666667
pearson4	-0.3939642	-0.4786693	-0.4871690	-0.4538802	-0.4755078
pearson6	-0.4368117	-0.4491398	-0.4000921	-0.3686541	-0.3968864
pearson8	-0.0583339	0.1080639	-0.1400981	0.1718998	-0.0715083

Table 7: Results on the three test sets with an error value < 0.1 as stop criterion, should be negative.

	V2	V3	V4	V5	V6
kendall4	-0.2000000	-0.4666667	-0.3333333	-0.3333333	-0.2000000
kendall6	-0.4666667	-0.4666667	-0.4666667	-0.3333333	-0.6000000
kendall8	-0.2000000	0.2000000	0.2000000	0.0666667	-0.0666667
pearson4	-0.4271253	-0.5080367	-0.4537100	-0.4501437	-0.4649583
pearson6	-0.3270643	-0.3627628	-0.3675643	-0.3195483	-0.4186973
pearson8	-0.1920649	0.0477997	-0.0504460	0.1470260	-0.0223448

Discussion

The results are sadly not very impressive, neither for established metrics, the Probit regression nor the Genetic Search. While the Genetic Search seems to outperform Probit regression, it does not perform better than established metrics, and the results are not consistent between runs.

One reason for more complex models, trained on the binary labeled data, not working very well might be a fundamental mismatch between this training data and the graded test data. The training data cover a very wide span, from very easy-to-read to advanced text, while the test data cover only a narrow span of texts aimed at students in the 4th to 8th grade in the Swedish school system. This means that while the all the approaches might manage to find a rough readability structure over the wider span, this structure might not be fine grained enough to accurately sort texts within the narrower span of the test data. To illustrate this, we can look at the variances of the different features used in the model:

Table 8: Variances for all features in the data and the ratio between the variances in the training set and test sets.

	TrainingSet	TestSet	Ratio
avgSentenceLength	15.2791302	4.2020275	0.2750175
avgWordLength	0.2891401	0.0332219	0.1148990
ratioSweVocTotal	0.0080634	0.0027147	0.3366638
ratioSweVocC	0.0089754	0.0031228	0.3479325
ratioSweVocD	0.0002877	0.0000728	0.2531726
ratioSweVocH	0.0007306	0.0002174	0.2975351
ovixValue	130.1330187	29.2118841	0.2244771
meanDepDistanceSentence	240.8951445	44.7920302	0.1859399
meanDepDistanceDependent	0.1162303	0.0310235	0.2669138
dep_UA	0.0001116	0.0000337	0.3017899
dep_AT	0.0003462	0.0000444	0.1282076
dep_ET	0.0003939	0.0000726	0.1843322
dep_PT	0.0000040	0.0000019	0.4657963
avgSentenceDepth	1.1127102	0.3398851	0.3054570
nrValue	1.1983880	0.0211179	0.0176219

As we can see above, the variances in the training data are consistently smaller than the variances in the test data. The situation might actually be even more extreme for each individual test set as in this table I have lumped all texts in the three test sets together which increases the variance of the total set.

If we look at the structures found by 5 consecutive runs we can see that when we narrow down the stop criterion, we also increase the correlation between the found results.

$$C_{0.2} = \begin{bmatrix} 1 & 0.872501 & 0.8923633 & 0.8777622 & 0.8909092 \\ 0.872501 & 1 & 0.8653918 & 0.8675264 & 0.8756686 \\ 0.8923633 & 0.8653918 & 1 & 0.8840548 & 0.8922573 \\ 0.8777622 & 0.8675264 & 0.8840548 & 1 & 0.8937066 \\ 0.8909092 & 0.8756686 & 0.8922573 & 0.8937066 & 1 \end{bmatrix}$$

$$C_{0.15} = \begin{bmatrix} 1 & 0.9272674 & 0.9206292 & 0.9250845 & 0.9279204 \\ 0.9272674 & 1 & 0.9259983 & 0.9301824 & 0.9374219 \\ 0.9206292 & 0.9259983 & 1 & 0.9262317 & 0.9317028 \\ 0.9250845 & 0.9301824 & 0.9262317 & 1 & 0.9258696 \\ 0.9279204 & 0.9374219 & 0.9317028 & 0.9258696 & 1 \end{bmatrix}$$

$$C_{0.1} = \begin{bmatrix} 1 & 0.9541351 & 0.964385 & 0.9601578 & 0.9578075 \\ 0.9541351 & 1 & 0.9615688 & 0.9617344 & 0.9435537 \\ 0.964385 & 0.9615688 & 1 & 0.9623627 & 0.9618744 \\ 0.9601578 & 0.9617344 & 0.9623627 & 1 & 0.9534342 \\ 0.9578075 & 0.9435537 & 0.9618744 & 0.9534342 & 1 \end{bmatrix}$$

This implies that the problem is at least somewhat convex and that the structure found by the approach, though infeasible to find with very good precision, is a true latent structure. However, the fact that I still get different results on the test sets for subsequent runs strengthens the hypothesis that the main problem is the difference in variance between the training set and the test set.

Another problem with the test data is that the differences between the average performances within each group are small. Due to the nature of the data collection, these averages might also be noisy. It might be possible, using standard psychometric models such as the Rasch model, or established variable selection models from machine learning, such as AIC or LASSO, to clean up these averages by removing noise resulting from "bad questions" from the comprehension tests. This cleaning would not be dependent on the training data in any way and would not introduce any circular reasoning or "teaching to the test".

Another approach would be to drop the 14 text feature model and start looking, among the ~120 features I can extract at this time, for any features which work better. It might also be possible to combine these features and map them onto a smaller space, for instance by using Principal Component Analysis or by applying Genetic Programming on the features rather than just a simple Genetic Search on the latent readability scores.

At this time though, I can only draw the conclusion that either none of the approaches in this paper work very well or the training and test sets are too fundamentally different.

References

Björnsson, Carl Hugo. 1968. *Läsbarhet*. Stockholm: Liber.

Dale, Edgar, and Jeanne S. Chall. 1949. "The concept of readability." *Elementary English* 26 (23).

Heimann Mühlenbock, Katarina. 2013. "I see what you mean. Assessing readability for specific target groups." PhD thesis, Språkbanken, Dept of Swedish, University of Gothenburg. http://hdl.handle.net/2077/32472.

Hultman, Tor G., and Margareta Westman. 1977. *Gymnasistsvenska*. Lund: LiberLäromedel.

Mühlenbock, Katarina, and Sofie Johansson Kokkinakis. 2009. "LIX 68 revisited - An extended readability measure." Edited by Michaela Mahlberg, Victorina González-Díaz, and Catherine Smith. *Proceedings of the Corpus Linguistics Conference CL2009*. Liverpool, UK.

Petersen, Sarah. 2007. "Natural language processing tools for reading level assessment and text simplification for bilingual education." PhD thesis, Seattle, WA: University of Washington.

Pitler, Emily, and Ani Nenkova. 2008. "Revisiting Readability: A Unified Framework for Predicting Text Quality." In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 186–95. Honolulu, HI.