An Analysis of Statistical Models and Features for Reading Difficulty Prediction

Michael Heilman, Kevyn Collins-Thompson and **Maxine Eskenazi**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{mheilman,kct,max}@cs.cmu.edu

Abstract

A reading difficulty measure can be described as a function or model that maps a text to a numerical value corresponding to a difficulty or grade level. We describe a measure of readability that uses a combination of lexical features and grammatical features that are derived from subtrees of syntactic parses. We also tested statistical models for nominal, ordinal, and interval scales of measurement. The results indicate that a model for ordinal regression, such as the proportional odds model, using a combination of grammatical and lexical features is most effective at predicting reading difficulty.

1 Introduction

A reading difficulty, or readability, measure can be described as a function or model that maps a text to a numerical value corresponding to a difficulty or grade level. Inputs to this function are usually statistics for various lexical and grammatical features of the text. The output is one of a set of ordered difficulty levels, usually corresponding to grade levels for elementary school through high school. As such, reading difficulty prediction can be viewed as a regression of grade level on a set of textual features.

Early work on readability measures employed simple proxies for grammatical and lexical complexity, including sentence length and the number of syllables in a word. Fairly simple features were often employed because of a lack of computational power. Such features exhibit high bias because they rely on strong assumptions about what makes a text difficult to read. For example, the use of sentence length as a measure of grammatical complexity assumes that a longer sentence is more grammatically complex than a shorter one, which is often but not always the case. In one early model, the Dale-Chall model (Dale and Chall, 1948; Chall and Dale, 1995), reading difficulty is a linear function of the mean sentence length and the percentage of rare words, as defined by a list of 3,000 words commonly known by 4th grade. In this paper, sentence length is defined as the mean number of words in the sentences of a text.

Many early measures did not employ direct estimates of word frequency due to computational limitations (e.g., (Gunning, 1952; McLaughlin, 1969; Kincaid et al., 1975)). Instead, these measures relied on the strong relationship between the frequency of and the number of syllables in a word. More frequent words are more likely to have fewer syllables (e.g., "the") than less frequent words (e.g., "vocabulary"), an association that is related to Zipf's Law (Zipf, 1935). The Flesch-Kincaid measure (Kincaid et al., 1975) is probably the most common reading difficulty measure in use. It is implemented in common word processing programs. This measure is a linear function of the mean number of syllables per word and the mean number of words per sentence. Klare (1974) provides a summary of other early work on readability.

More recent approaches to reading difficulty employ more sophisticated models that make use of the growth in computational power. The Lexile Framework (e.g., (Stenner, 1996)) uses individual word frequency estimates as a measure of lexical difficulty. The word frequency estimates are derived

Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications, pages 71–79,
Columbus, Ohio, USA, June 2008.

from a large, varied corpus of text. Lexile uses a Rasch model (Rasch, 1980) with the mean log word frequency as a lexical feature and the log of the mean sentence length as a grammatical feature. The Rasch model, related to logistic regression, is used to estimate the level of a student that would comprehend 75% of a given text. The converted log odds ratio called a "Lexile" that is used as part of this measure can be easily mapped to grade school levels.

A reading difficulty measure developed by Collins-Thompson and Callan (2005) uses smoothed unigram language modeling to capture the predictive ability of individual words based on their frequency at each reading difficulty level. Collins-Thompson and Callan found that certain words were very predictive of certain levels. For example, "grownup" was very predictive of grade 1, and "essay" was very predictive of grade 12. For a given text, this measure estimates the likelihood that the text was generated by each level's language model. The prediction is the level of the model with the highest likelihood of generating the text. There are no grammatical features.

Natural language processing techniques enable more sophisticated grammatical analysis for reading difficulty measures. Rather than using sentence length as a proxy, measures can employ tools for automatic analysis of the syntactic structure of texts (e.g., (Charniak, 2000)). A measure by Schwarm and Ostendorf (2005) incorporates syntactic analyses, among a variety of other types of features. It includes four grammatical features derived from syntactic parses of text: the mean parse tree height, the mean number of noun phrases, mean number of verb phrases, and mean number of "SBARs." "SBARs" are non-terminal nodes that are associated with subordinate clauses. Their system led to better predictions than the Flesch-Kincaid and Lexile measures, but the predictive value of the grammatical features is not entirely clear. In initial experiments using such course-grain grammatical features alone, rather than in conjunction with language modeling and other features as in Schwarm and Ostendorf's system, we found relatively poor prediction performance. Our final approach using subtrees of syntactic parses allows for a finer level of discrimination that may support the detection of differences in grade levels between texts that exhibit the same high level features.

A reading difficulty measure developed by Heilman, Collins-Thompson, Callan, and Eskenazi (2007) uses the frequency of grammatical constructions as a measure of grammatical difficulty. A set of approximately twenty constructions were selected from English as a Second Language grammar textbooks. This set includes grammatical constructions such as the passive voice, relative clauses, and various verb tenses. The frequencies are used as features for a nearest neighbor classification algorithm. The unigram language modeling approach of Collins-Thompson and Callan (2005) is used to estimate lexical difficulty in this measure. The final prediction is a linear function of the lexical and grammatical components. That model assumes that grammatical difficulty is adequately captured by a small number of constructions chosen according to detailed knowledge of English grammar. In that work, the constructions were selected from an English as a Second Language grammar textbook, a labor- and knowledge-intensive task that may be less practical for other languages.

We aim to identify the appropriate scale of measurement for reading difficulty–nominal, ordinal, or interval–by comparing the effectiveness of statistical models for each type of data. We also extend previous work combining lexical and grammatical features (Heilman et al., 2007) by making it possible to include a large number of grammatical features derived from syntactic structures without requiring significant linguistic or pedagogical content knowledge, such as a reference guide for the grammar of the language of interest.

2 Types of Features

2.1 Lexical Features

This section and the following section describe the lexical and grammatical features used in our reading difficulty models. The lexical features are the relative frequencies of word unigrams. The use of word unigrams is a standard approach in text classification (Yang and Pedersen, 1997), and has also been successfully used to predict reading difficulty (Collins-Thompson and Callan, 2005). Higher order n-grams such as bigrams and trigrams were not used as features because they did not improve predictions

in preliminary tests. The specific set of lexical features was chosen based on the frequencies of words in the training corpus. The system performs morphological stemming and stopword removal. The remaining 5000 most common words comprised the lexical feature set.

2.2 Grammatical Features

Grammatical features are extracted from automatic context-free grammar parses of sentences. The system computes relative frequencies of partial syntactic derivations, which will be called 'subtrees' hereafter. The approach extends (Heilman et al., 2007), where frequencies of manually defined syntactic patterns were extracted from syntactic structures. In that approach, the features are defined manually using linguistic knowledge of the target language to implement tree search patterns, a labor- and knowledge-intensive process. The approach advocated in this paper, however, extracts frequencies for an automatically defined set of subtree patterns. The system considers all subtrees up to a given depth that occur in the training corpus. Examples of grammatical features at levels 0 through 2 are shown in Figure 1. The sentence for the parse tree shown was taken from a third grade text.

For depth 0, the system includes all subtrees consisting of just nonterminal nodes. This includes all parts of speech, as well as non-terminal nodes for noun phrases, adjective phrases, clauses, etc. For depth 1, the system includes subtrees corresponding to the application of a single context free grammar rule in the derivation of the tree. An example of a feature at this level would be a sentence node that dominates nodes for noun phrases and verb phrases. For deeper levels, the system includes subtrees corresponding to the successive application of rules on non-terminals symbols until either a terminal symbol is reached or the given depth is reached. An example feature for level 2 is a subtree in which a prepositional phrase node dominates a preposition node and noun phrase node, and the preposition node in turn dominates a preposition, and the noun phrase dominates determiner, adjective, and noun nodes.

We used a maximum depth of 3 in our experiments. Features of deeper levels occur less frequently in general, and deeper levels were avoided due to data sparseness. A depth first search algorithm extracts candidate grammatical features from the training corpus. First, a context-free grammar parser (Klein and Manning, 2003) derives parse trees for all texts in the training corpus. The algorithm traverses these parses, at each node counting all subtree features up to the given depth that are rooted at that node. The subtree features are sorted by their overall counts in the corpus. In our experiments, frequencies of the most common 1000 subtrees were chosen as the final features. These included 64 level 0 features corresponding to nonterminal symbols, 334 level 1 features, 461 level 2 features, and 141 level 3 features. Deeper levels have more possible features, but sparsity at level 3 resulted in fewer level 3 features being selected.

In our experiments, the subtrees included terminal symbols for stopwords. However, the system effectively removed content word terminals from parses before extracting features. The system could be modified to include terminal symbols for content words, or even to ignore all nodes for terminal symbols. Subtree features including terminal symbols for content words would, of course, occur with low frequency and not likely be included in the final feature set. Terminal symbols for content words were omitted so that lexical information was not included in the set of grammatical features. Similar to leaving higher order n-grams out of the lexical feature set, omitting terminal symbols for content words avoids confounding grammatical and lexical information in the grammatical feature set. Subtree counts are normalized by the number of words in a text to compute the relative frequencies. Normalization by the number of sentences in a text is also possible, but did not perform as well in preliminary tests. The Stanford Parser (Klein and Manning, 2003) version 1.5.1 was used to derive tree structures for sentences. We used the unlexicalized model included in the distribution which was trained on Wall Street Journal texts.

3 Statistical Models

3.1 Scales of Measurement for Reading Difficulty

Several statistical models were tested for effectiveness at predicting reading difficulty. The appropriateness of these models depends on the nature of

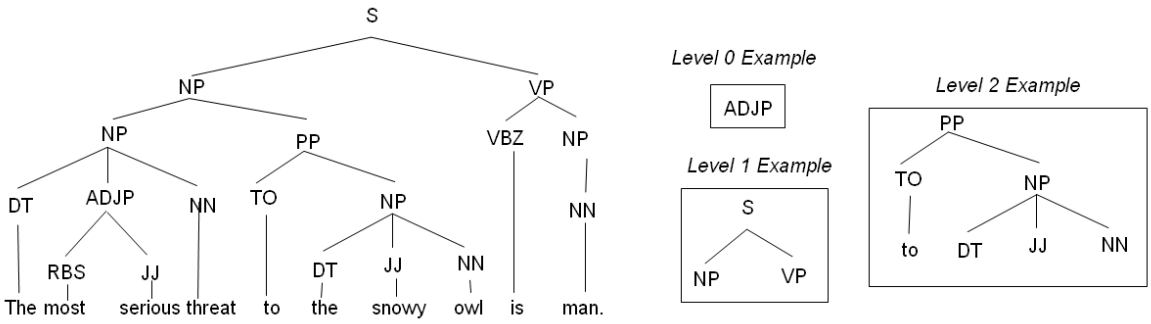

Figure 1: Parse Tree for Sentence from Third Grade Text with Example Subtree Features.

reading difficulty data, particularly the scale of measurement. The standard unit for reading difficulty is the grade level. First through twelfth grade levels in American schools have been used in previous work (e.g., (Heilman et al., 2007; Collins-Thompson and Callan, 2005)). English as a Second Language levels have also been used (Heilman et al., 2007), as well as grade levels for other languages such as French (Collins-Thompson and Callan, 2005). While these grades are assigned evenly spaced integers, the ranges of reading difficulty corresponding to these grades are not necessarily evenly spaced. It is possible, of course, that assuming even spacing between levels might produce more parsimonious and accurate statistical models. A more reasonable assumption is that the grade numbers assigned to each difficulty level denote an ordering: for example, that grade 1 is in some sense less than grade 2, which is less than grade 3, etc. Different statistical models handle this assumption more or less well.

Statistics generally distinguish four scales of measurement, which are, ordered by increasing assumptions about the relationships between values: nominal, ordinal, interval, and ratio (Stevens, 1946; Cohen et al., 2003). *Nominal* data involve no relationships between the labels or classes of the data. An example would be types of fruits, where a model might be used to make decisions between apples and oranges. This type of prediction is generally called classification in machine learning and related fields. *Ordinal* data have a natural ordering, but the values are not necessarily evenly spaced. For example, data about the severity of illnesses might have labels such as mild, moderate, severe, deceased, in which the transitions between consecutive classes all have the same direction but not the same magnitude. Making predictions about such data is generally called ordinal regression (McCullagh, 1980). *Interval data*, however, are both ordered and evenly spaced. An example would be temperature as measured in Fahrenheit degrees. Such data have an arbitrary zero point, and negative values may occur. *Ratio* data, of which annual income is an example, do have a meaningful zero point. We will not discuss ratio data further since its distinction from interval data is not relevant to this paper. It is not clear to which scale reading difficulty corresponds. The assumption of an interval scale allows for simpler models with fewer parameters. However, models for ordinal or even nominal data might be more appropriate if the strong assumption of an interval scale does not hold.

We experimented with three linear and log-linear models corresponding to interval, ordinal, and nominal data. Parameters were estimated using L_2 regularization, which corresponds to a Gaussian prior distribution with zero mean and a user-specified variance over the parameters. We chose these models because they are commonly used in the statistics, machine learning, and behavioral science communities, and aimed to set up meaningful comparisons among the scales of measurement. Other machine learning algorithms might also be employed. In fact, we briefly tested the maximum margin (Vapnik, 1995) approach, which led to comparable results and might be worth exploring in future work.

3.2 Linear Regression

Linear Regression (LIN) produces a linear model in which the dependent, or outcome, variable is a linear function of the values for predictor variables, or features. A prediction for a given text is the inner product of a vector of feature values for the text and a vector of regression coefficients estimated from training data. For the case of reading difficulty, the grade level is a linear combination of the lexical and/or grammatical feature values. LIN provides continuous estimates of reading difficulty, such that a prediction might fall between grade levels. The estimates were not rounded to whole numbers in the experiments. For rare cases of an LIN prediction falling outside the appropriate range of grade levels, the value was set to the maximum or minimum grade level. LIN implicitly assumes that the data fall on an interval scale, meaning that the levels are evenly spaced. The LIN model has relatively few parameters but makes strong assumptions about the scale of measurement. For details, see (Hastie et al., 2001).

3.3 Proportional Odds Model

The Proportional Odds (PO) model, also called the parallel regression model and the cumulative logit model, is a form of log-linear, or exponential, model for ordinal data (McCullagh, 1980). Given a new unlabeled instance as input, the model provides estimates of the probability that the instance belongs to a class at or above a particular level. In Equation (1), $P(y \geq j)$ is this estimated probability, α_j is an intercept parameter for the given level j, β is vector of regression coefficients, X_i is the vector of feature values for instance i, and y_i is the predicted reading difficulty level.

$$P(y_i \geq j) = \frac{exp(\alpha_j + \beta^T X_i)}{1 + exp(\alpha_j + \beta^T X_i)} \quad (1)$$

$$ln \frac{P(y_i \geq j)}{1 - P(y_i \geq j)} = \alpha_j + \beta^T X_i \quad (2)$$

The PO model has a parameter α_j for the threshold, or intercept, at each level j, but only a single set β of parameters for the features. These two types of parameters correspond to an implicit assumption of ordinality. Having a single set of parameters for features across the levels means that changes in feature values proportionally affect the odds of transitioning from any one class to another.

The estimated probability of an instance belonging to a particular class is the difference between estimates for that class and the next highest class. For example, the estimated probability of a text being at the eighth grade level would be the estimate for being at or above eighth grade minus the estimate for being at or above ninth grade. As in binary logistic regression, the PO model estimates log odds ratios based on the values of features or predictor variables. The numerator of the odds ratio is the probability of being at or above a level, and the denominator is the probability of being below a level. Equation (2) shows the form of the model that is linear in the parameters.

3.4 Multi-class Logistic Regression

Multi-class Logistic Regression (LOG), or multinomial logit regression, is a log-linear model for nominal data. In contrast to the simpler PO model, the model maintains parameters for all of the features for every class except one category, which is used for comparison. Thus, for reading difficulty, there are about 11 times as many parameters to estimate compared to LIN and PO. The increased difficulty of parameter estimation for this model is offset for domains in which assumptions of ordinality or linearity do not hold. For more details, see (Hastie et al., 2001).

4 Evaluation

4.1 Web Corpus

The corpus of materials used for training and testing the models consists of the content text extracted from Web pages with reading difficulty level labels. Web pages were used because the system for predicting reading difficulty is being used as part of the REAP tutoring system, which finds authentic and appropriate Web pages for English vocabulary practice (Brown and Eskenazi, 2004; Heilman et al., 2006). Approximately half of these texts were authored by students at the particular grade level, and half were authored by teachers or writers and aimed at readers at a particular grade level. Texts were found for grade levels 1 through 12. The twelfth grade level also included some post-secondary level

texts. Various genres and subjects were represented. In all cases, either the text itself or a link to it identified it as having a certain level. The content text was manually extracted from these Web pages so that noisy information such as navigation menus and advertisements were not included. Automatic content extraction may, however, be able to remove such noisy information without human intervention (e.g., (Gupta et al., 2003)). This Web corpus is adapted from the corpora used in prior work on reading difficulty predication (Collins-Thompson and Callan, 2005; Heilman et al., 2007). We modified that corpus because it contained a number of documents pertaining to mathematics and vocabulary practice. The majority of tokens in these texts were not part of well-formed, grammatical sentences suitable for reading practice. Since our goal is to measure the difficulty of reading passages, we removed these documents and added additional texts consisting of more suitable reading material. The corpus consisted of approximately 150,000 words, distributed among 289 texts. The number of texts for each grade level was approximately the same, with at least 28 texts at each level. The mean length in words of the texts was approximately 500 words, which corresponds to about a page. Texts for lower grades were necessarily shorter. We extracted excerpts for higher level texts so that texts were otherwise roughly equal in length across levels. For these excerpts, the first 500 or so words of text were extracted, while respecting sentence and paragraph boundaries.

4.2 Evaluation Metrics

Root mean square error (RMSE), Pearson's correlation coefficient, and accuracy within 1 grade level served as metrics for evaluating the performance of reading difficulty predictions. Multiple statistics were used because it is not entirely clear what the best measure of prediction quality is for reading difficulty. RMSE is the square root of the empirical mean of the squared error of predictions. It more strongly penalizes those errors that are further away from the true value. It can be interpreted as the average number of grade levels that predictions measure deviate from human-assigned labels.

Pearson's correlation coefficient measures the strength of the linear relationship, or similarity of trends, between two random variables. A high correlation would indicate that difficult texts would more likely receive high predicted difficulty values, and easier texts would be more likely to receive low predicted difficulty values. Correlations do not, however, measure the degree to which values match in absolute terms.

Adjacent accuracy is the proportion of predictions that were within one grade level of the human-assigned label for the given text. Exact accuracy is too stringent a measure because the human-assigned reading levels are not always perfect and consistent. For example, one school might read "Romeo and Juliet" in 9th grade while another school might read it in 10th grade. The drawback of this accuracy metric is that predictions that are two levels off are treated the same as predictions that are ten levels off.

4.3 Baselines

The performance of other algorithms for estimating reading difficulty was estimated using the same data. These comparison include Collins-Thompson and Callan's implementation of their language modeling approach (2005), an implementation of the Flesch-Kincaid reading level measure (Kincaid et al., 1975), and a measure using word frequency and sentence length similar to Lexile (Stenner et al., 1983). We did not directly test the approach described by (Heilman et al., 2007). We observe that its reported results for first language texts were not significantly different in terms of correlation and only slightly better in terms of mean squared error than the language modeling approach. Finally, a simple uniform baseline, which always chose the middle value of 6.5, was tested.

The Lexile-like measure (LX) used the same two features as the Lexile measure: mean log frequency or words and log mean sentence length. Instead of using a Rasch model and converting scores to "Lexiles," however, the PO model was used to directly predict grade levels. The log frequency values for words were estimated from the second release of the American National Corpus (Reppen et al., 2005), a 20 million word corpus with texts in American English from different genres on a variety of subjects. Using the proportional odds models is effectively equivalent to using Lexile's Rasch model and mapping its output to grade levels. The major difference between the Lexile measure and the implemen-

tation used in these experiments is the training data sets used to estimated word frequencies and model parameters.

4.4 Procedure

The Web Corpus was randomly split into training and test sets. The test set consisted of 25% of the individual texts at each level, a total of 84 texts. Ten-fold stratified cross-validation on the training set was employed to estimate the prediction performance according to the evaluation metrics. In cross-validation, data are partitioned randomly into a given number of folds, and each fold is used for testing while all others are used for training. For more details and a discussion of validation methods, see (Hastie et al., 2001). The regularization hyper-parameters were tuned on the training set during cross-validation by a simple grid search. After cross-validation, models were trained on the entire training set, and then evaluated using the held-out test data.

We tested whether each feature-set, algorithm pair or baseline performed significantly differently than our hypothesized best model, the PO model with the combined feature set. We employed the bias-corrected and accelerated (BC_a) Bootstrap (Efron and Tibshirani, 1993) with 50,000 replications of the held-out test data to generate confidence intervals for differences in evaluation results. If the $(1-\alpha)\%$ confidence intervals for the difference do not contain zero, which is the value corresponding to the null hypothesis, then that difference is significant at the α level. For example, the 99% confidence interval for the difference in adjacent accuracy between the language modeling baseline and the PO model with the combined feature set was (-1.86, -0.336), indicating that this difference is significant at the .01 level since it does not contain zero.

5 Results

Table 1 presents correlation coefficients, RMSE values, and accuracy values for cross-validation and held-out test data. Statistical significance was tested only for the held-out test data since the hyper-parameters were tuned during cross-validation. Our discussion of the results pertains mostly to the evaluation on the test-set.

Of the various statistical models, the PO model for ordinal data appears to provide superior performance over the LIN and LOG models. Compared to the LOG model, the PO model performs significantly better in terms of correlation and RMSE and comparably well in terms of adjacent accuracy. Compared to the LIN model, the PO model performs almost as well in terms of correlation, comparably well in terms of RMSE, and far better in terms of accuracy.

The performance of the methods when using different feature sets does not clearly indicate a best set of features to use for predicting reading difficulty. For the PO model, none of the feature sets lead to significant gains over the others in terms of any of the metrics. However, the combined feature set led to the best performance in terms of correlation and adjacent accuracy during cross-validation as well as RMSE on the test set, suggesting at the very least that including the extra features does not degrade performance.

The PO model with the combined feature set outperformed most of the baseline measures. LX had the same accuracy value on the test set. The LX method appears to perform the best in general of the baselines models. Interestingly, LX uses proportional odds logistic regression like PO, and thus assumes an ordinal but not interval scale of measurement. RMSE values were significantly lower for the PO model than for LX and the language modeling approach.

No statistically significant advantages are seen for PO model when compared to Flesch-Kincaid. We observe however, that for the sample of web pages which constitutes the evaluation corpus the PO model produced superior results across evaluation metrics. That is, PO performed better in terms of adjacent accuracy, RMSE, and correlation coefficients, both in cross-validation and testing with held-out data.

6 Discussion

In our tests, the PO model, which assumes ordinal data, lead to the most effective predictions of reading difficulty in general. This result indicates that the reading difficulty of texts, according to grade level, lies on an ordinal scale of measurement. That is,

Method	Features	Cross-Validation			Held-Out Test Set		
		Correl.	RMSE	Adj. Acc.	Correl.	RMSE	Adj. Acc.
LIN	Lexical	.629	2.73	.242	.779	2.42	.167**
	Grammatical	.767	2.26	.294	.753	2.33	.274*
	Combined	.679	2.57	.284	**.819****	**2.21**	.226**
PO	Lexical	.713	2.57	.498	.780	2.29	.464
	Grammatical	.762	**2.22**	.505	.734	2.42	**.560**
	Combined	**.773**	2.24	**.519**	.767	2.23	.440
LOG	Lexical	.517	3.24	.443	.619*	2.83*	.548
	Grammatical	.632	2.87	.443	.506**	3.38**	.464
	Combined	.582	2.94	.446	.652*	2.71*	.556
LX	-	.659	2.77	.467	.731	2.67*	.464
Lang. Modeling	-	.590	2.74	.370	.630	2.70**	.381
Flesch-Kincaid	-	.697	2.66	.388	.718	2.54	.369
Uniform	-	.000	3.39	.170	.000**	3.45**	.167**

Table 1: Results from Cross-Validation and Test Set Evaluations, as measured by Correlation Coefficients (Correl.), Root Mean Square Error (RMSE), and Adjacent Accuracy. The best result for each metric for each evaluation is given in bold. Asterisks indicate significant differences compared to the PO model with a Combined Feature Set. * = $p < .05$, ** = $p < .01$.

reading difficulty appears to increase steadily but not linearly with grade level. As such, the LIN approach that produces linear models was less effective, particularly in terms of adjacent accuracy. The LOG model, for nominal data, also led to inferior performance compared to the PO model, which can be attributed to the difficulty of accurately estimating a more complex model with many parameters for each level.

Our tests found that grammatical features alone can be effective predictors of readability. This finding disagrees with a previous result that found that a model using a combination of lexical and manually defined grammatical features (Heilman et al., 2007) outperformed a model using grammatical features alone. The superior predictive ability of the models we describe that use grammatical features can be attributed to the automatic derivation of a grammatical feature set that is more than an order of magnitude larger than in the previous approach. Our approach enables the use of much larger grammatical feature sets because it does not require the extensive linguistic knowledge and effort to manually define the grammatical features. The automatic approach also enables an easier transition to other languages, assuming a parser is available. Using the combined feature set did not hurt performance, however, and since regularized statistical models can avoid overfitting large numbers of parameters, a combined feature set still seems appropriate.

Acknowledgments

We thank Jamie Callan for his comments and suggestions. This research was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040063 to Carnegie Mellon University; Dept. of Education grant R305G03123; the Pittsburgh Science of Learning Center which is funded by the National Science Foundation, award number SBE-0354420; and a National Science Foundation Graduate Research Fellowship awarded to the first author. Any opinions, findings, conclusions, or recommendations expressed in this material are the authors, and do not necessarily reflect those of the sponsors.

References

Jon Brown and Maxine Eskenazi. 2004. Retrieval of authentic documents for reader-specific lexical practice. *Proceedings of InSTIL/ICALL Symposium 2004*.

J. S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books. Cambridge, MA.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. *Proceedings of the NAACL.*

J. Cohen, P. Cohen, S. G. West, and L. S. Aiken. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd Edition*. Lawrence Erlbaum Associates, Inc.

Michael Collins and Nigel Duffy. 2002. Convolution Kernels for Natural Language. *Advances in Neural Information Processing Systems.*.

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology, 56(13). pp. 1448-1462.*.

E. Dale and J. S. Chall. 1948. A Formula for Predicting Readability. *Educational Research Bulletin Vol. 27, No. 1*.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.

R. Gunning. 1952. *The technique of clear writing.*. McGraw-Hill, New York.

S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm. 2003. *DOM-based content extraction of HTML documents*. ACM Press, New York.

Trevor Hastie, Robert Tibshirani, Jerome Friedman. 2003. *The Elements of Statistical Learning:Data Mining, Inference, and Prediction*. Springer.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Proceedings of the Human Language Technology Conference*. Rochester, NY.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2006. Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. *Proceedings of the Ninth International Conference on Spoken Language Processing*. Pittsburgh, PA.

J. Kincaid, R. Fishburne, R. Rodgers, and B. Chissom. 1975. Derivation of new readability formulas for navy enlisted personnel. *Branch Report 8-75*. Chief of Naval Training, Millington, TN.

G. R. Klare. 1974. Assessing Readability. *Reading Research Quarterly, Vol. 10, No. 1. pp. 62-102.*.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430*.

G. H. McLaughlin. 1969. SMOG grading: A new readability formula. *Journal of Reading*.

P. McCullagh. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 42, No. 2. pp. 109-142*.

G. Rasch. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press, Chicago, IL.

G. Rasch. 2005. *American National Corpus (ANC) Second Release.*. Linguistic Data Consortium. Philadelphia, PA.

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.

A. J. Stenner, M. Smith, and D. S. Burdick. 1983. Toward a Theory of Construct Definition. *Journal of Educational Measurement, Vol. 20, No. 4. pp. 305-316*.

A. J. Stenner. 1996. Measuring reading comprehension with the Lexile framework. *Fourth North American Conference on Adolescent/Adult Literacy*.

S. S. Stevens. 1946. On the theory of scales of measurement. *Science, 103, pp. 677-680*.

V. N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.

Y. Yang and J. P. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp. 412-420*.

G. K. Zipf. 1935. *The Psychobiology off Language*. Houghton Mifflin, Boston, MA.